AN INVESTIGATION OF GENETICALLY DETERMINED TELOMERE LENGTH AND ITS IMPACT ON DISEASE RISK IN UK BIOBANK

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Svetlana Stoma

Department of Cardiovascular Sciences

University of Leicester

March 2021

An investigation of genetically determined telomere length and its impact on disease in UK Biobank

Svetlana Stoma

Abstract

Telomere length has been proposed as a marker of biological age, and numerous studies have shown associations between directly measured telomere length and age-related disease risk. The study into underlying genetic contribution to telomere length have recently begun to emerge and bring an understanding of the genetic effect of telomere length on human health.

This study investigates genetically determined telomere length and its association with age-related diseases. A genetic risk score was built within UK Biobank for each participant using the genetic determinants of telomere length identified in our large-scale genome-wide meta-analysis study and tested against a curated list of 127 diseases. Some of these associations were confirmed as causal using mendelian randomisation and taken further to model the effects on time to disease onset using survival analysis.

Shorter genetically determined telomere length was causally associated with an increased risk of cardiovascular, endocrine, and immune disease phenotypes, and a decreased risk of diseases with high proliferative capacity. This study provides evidence that genetically determined telomere length is involved in the aetiology of age-related disease and influences time to disease onset, which highlights the primary function of telomere length in limiting cell division. A genetic predisposition to shorter telomere length may contribute to an accelerated loss of telomeric repeats during cell division. Whilst a genetic predisposition to longer telomere length may contribute to increased telomere length maintenance that accumulates mutations that may lead to malignancies.

Acknowledgements

I would like to express my gratitude towards my supervisors Dr Christopher P. Nelson and Dr Veryan Codd for their support and guidance during my PhD. They have been supportive throughout all of my academic achievements, and I am very lucky to have them. I would like to thank the TeloTeam as a whole, who made it a pleasure to study at the Cardiovascular Research Centre of University of Leicester.

Publications related to this thesis

Li, C. *et al*. (2020) 'Genome-Wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length', The American Journal of Human Genetics, 106(Mar), pp. 1–16.

Codd, V. *et al*. (2021) 'A major population resource of 474,074 participants in UK Biobank to investigate determinants and biomedical consequences of leukocyte telomere length', medRxiv pre-print.

Codd, V. *et al*. (2021) 'Polygenic basis and biomedical consequences of telomere length variation', medRxiv pre-print.

Table of Contents

List of Tables

Table of Figures

List of Abbreviations

TL – Telomere Length

LTL – Leucocyte Telomere Length

GDTL – Genetically Determined Telomere Length

SNP – Single Nucleotide Polymorphism

DNA - Deoxyribonucleic acid

ALT - Alternative Lengthening of Telomeres

SASP - Senescence-Associated Secretory Phenotype

TRF - Terminal Restriction Fragments

qPCR - Quantitative Polymerase Chain Reaction

FISH - Fluorescent in situ hybridisation

STELA - Single telomere length analysis

UKB – UK Biobank

HRC - Haplotype Reference Consortium

KING - Kinship-based INference for Genome-wide association studies

HES - Hospital Episode Statistics

ICD - international classification of diseases

OPCS-4 - Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures (4th revision)

GWAS – Genome-wide association study

LD – Linkage disequilibrium

PCA - Principal Component Analysis

HWE - Hardy-Weinberg equilibrium

MAF - Minor Allele Frequency

QC - Quality Control

GRS – Genetic Risk Score

TL GRS – Telomere Length Genetic Risk Score

PRS – Polygenic Risk Score

FDR – False Discover Rate

FIQT – False Discovery Rate Inverse Quantile Transformation

GCTA-COJO - Genome-wide Complex Trait Analysis, Conditional and Joint analysis

GRScojo52 – Genetic Risk Score of 52 genetic determinants of telomere length

GRSclump234 – Genetic Risk Score of 234 genetic determinants of telomere length

MR – Mendelian Randomisation

RCT – Randomised Controlled Trial

MR-IVW – Mendelian Randomisation Inverse-Variance Weighted

MR-RAPS – Mendelian Randomisation Robust Adjusted Profile Score

KM - Kaplan-Meier Estimator

PH - Proportional Hazard

OR – Odds Ratio

HR – Hazard Ratio

SD – Standard Deviation

RR - Relative Risk

CAD - Coronary Artery Disease

CVD - Cardiovascular Disease

MI - Myocardial Infarction

COPD - Chronic Obstructive Pulmonary Diseases

BMI – Body mass index

ENGAGE - European Network for Genetic and Genomic Epidemiology

EPIC - European Prospective Investigation into Cancer and Nutrition

BiLEVE - Biobank Lung Exome Variant Evaluation

CARDIoGRAM - Coronary ARtery DIsease Genome-wide Replication and Meta-analysis

CARDIoGRAMplusC4D - Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics consortium

ARIC - Atherosclerosis in Communities Study

BIOSTAT-CHF - BIOlogy Study to TAilored Treatment in Chronic Heart Failure

GRAPHIC - Genetic Regulation of Arterial Pressure of Humans in the Community

WTCCC-CAD - Wellcome Trust Case Control Consortium Coronary Artery Disease

Chapter 1. Telomere - a marker of ageing

In this study I aim to investigate the relationship between telomere length, defined by genetic determinants, and age-related diseases. Numerous observational studies have associated shorter telomere length with an increased risk of cardiovascular diseases, and longer telomere length with an increased risk of cancers, although these results were inconsistent. While directly measured telomere length may reflect an individual's biological age and health status, a genetically determined telomere length (GDTL) may estimate the underlying risks of experiencing age-related diseases, be involved in the aetiology of disease and predict the time to disease onset. To investigate these hypotheses and show that genetic telomere length is a potential driver of age-related diseases I employ three approaches, a Genetic Risk Score, Mendelian Randomisation and Survival Analysis, respectively.

In chapter one I introduce some basic genomics, telomeres, and telomere length. In chapter two I give an up-to-date review about current epidemiological research into telomere length and the potential practical use of telomere length in the prediction of age-related disease risks. In chapter three I cover genome-wide association studies, specifically those used to identify common genetic determinants of TL that I use as determinants to investigate genetic TL effects on the development of age-related diseases in chapters four, five and six. In the final chapter I summarise the success of this study and describe possible future work.

1.1. Basic genomics

The genome or genetic information is the source of complex rules that build and enable an organism to live. Genetic information is encoded with a molecule called Deoxyribonucleic acid (DNA). The DNA sequence contains four types of nucleotides, where each is composed of a nucleobase (A – adenine, C – cytosine, G – guanosine or T – thymine), a deoxyribose sugar and a phosphate group. The sequence of nucleotides forms a polynucleotide chain, and two chains coiled around each other form a double helix. Both chains, also called strands, store the same genetic information. Each long

DNA double helix is organised into chromosomes. Human genome consists of 22 pairs of autosomes and one pair of sex chromosomes that are present in two copies, one inherited from each parent.

Most of the human genome is identical between all people but there are millions of sites where the genetic code differs. Such sites of common genetic variation are called Single Nucleotide Polymorphisms (SNPs), and nucleotides that may differ between two sister chromosomes are called alleles. We can denote alleles as A and B and write them in three states (AA: carrying 2 A alleles, or A homozygote, AB: carrying an A and B alleles, or heterozygote, and BB: carrying 2 B alleles, or B homozygote). We then refer to this as the genotype (Donaldson *et al.*, 2016; Dorak, 2017; Evangelou, 2018).

1.2. Telomere biology

1.2.1. What is a telomere?

Telomeres are a protective structure that caps the end of each chromosome, maintaining chromosomal integrity and ensuring stability of the genome. Chromosome ends, capped and protected by telomeric structures, are not incorrectly recognised as DNA double-strand breaks, which prevents DNA damage signalling and unnecessary DNA repair. Telomeres also suppress deleterious processes such as DNA degradation, DNA end-joining and DNA recombination that may lead to unstable chromosomes (Blackburn, 2005; Palm *et al.*, 2008; Aubert, 2014; Lazzerini-Denchi *et al.*, 2016; Erdel *et al.*, 2017).

Human telomeres are composed of double stranded telomeric TTAGGG repeats that are ~2-10 kilobases long and contain a single-stranded 3' overhang that is 50-300 nucleotides in length (Palm *et al.*, 2008; Maciejowski *et al.*, 2017). This is bound by protein complexes, including shelterin (**Figure 1.1 A** and **B**). Together they form a stable T-loop while the single-stranded terminal G-overhang at the 3' end invades the double-stranded telomeric DNA and forms a D-loop (Palm *et al.*, 2008; Aubert, 2014; Rivera *et al.*, 2017) (**Figure 1.1 C**). Shelterin is a complex of proteins that protects telomeres from the DNA repair machinery and regulates access of telomerase, a protein that elongates telomeres, to telomeric DNA (Cooper *et al.*, 2017). The shelterin complex has six subunits: Telomeric Repeat binding Factor 1 and 2 (*TRF1* and *TRF2*) that recognise

telomeric repeats and bind double-stranded DNA, Protection Of Telomeres 1 (*POT1*) that binds telomeric single-stranded DNA at the 3' overhang, the *TRF2*- and *TRF1*-Interacting Nuclear Protein 2 (*TIN2*) and *TIN2*-interacting protein 1 (*TPP1*) that connect the three DNA-binding proteins, and Repressor/Activator Protein 1 (*RAP1*) that is associated with *TRF2* (**Figure 1.1 B**) (Palm *et al.*, 2008; Nandakumar *et al.*, 2013; Lazzerini-Denchi *et al.*, 2016; Erdel *et al.*, 2017; Turner *et al.*, 2019).

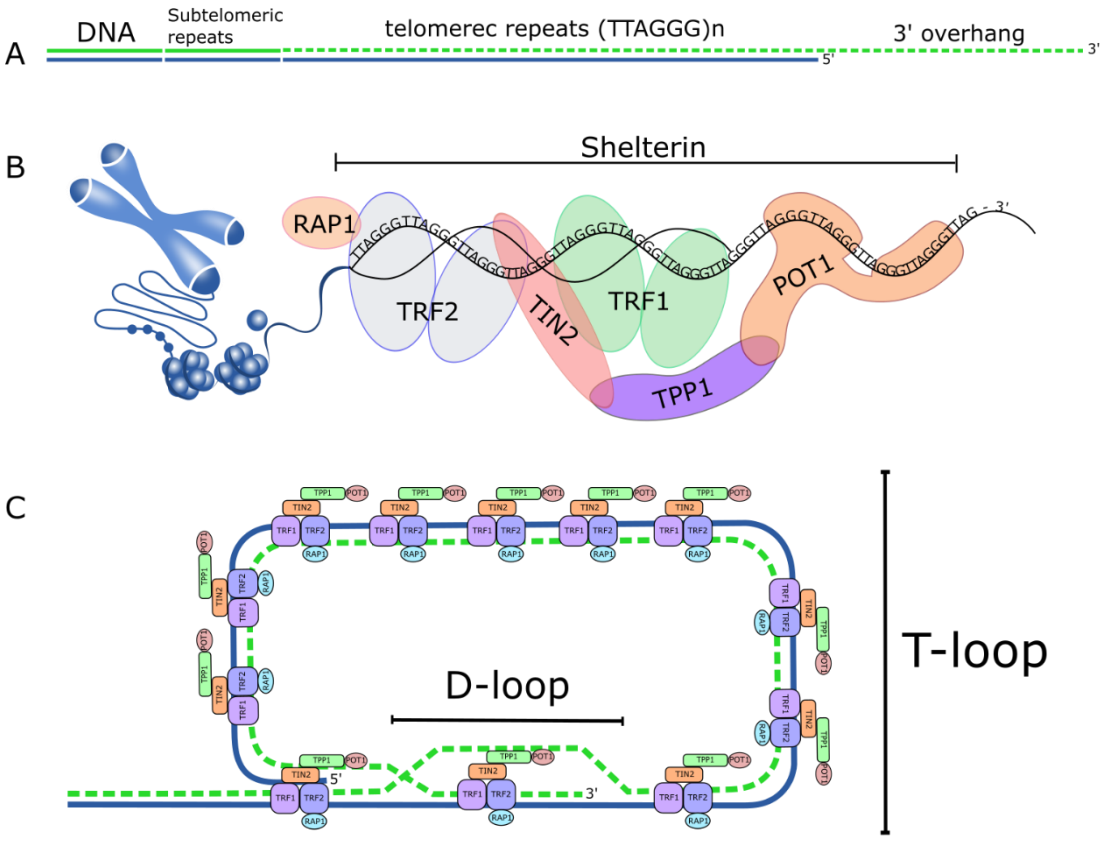


Figure 1.1. Telomere structure. A) telomeric repeats and chromosome end, B) shelterin complex of telomere binding proteins, C) formation of T- and D-loops by shelterins covering telomeric repeats.

1.2.2. Telomere length – a biological marker of cellular age

Telomeres were first recognised as important functional structures by Muller in 1938 (Muller, 1938) and by McClintock in 1941 (McClintock, 1941). The finite replicative capacity of somatic cells was reported by Hayflick in 1961, and now is often referred as Hayflick's limit (Hayflick *et al.*, 1961). The association between limited replicative capacity and 'end replication problem' of chromosomal ends was proposed by Olovnikov

in 1973 (Olovnikov, 1973). Since then, telomere attrition has become a subject for many studies on ageing (Olovnikov, 1996; Turner *et al.*, 2019).

Somatic cells have a lifespan, set by a limited number of divisions. The process of cell division requires genetic information to be duplicated to provide each daughter cell with a complete set of chromosomes. However, DNA polymerase is unable to fully replicate the 3' end of the DNA (Watson, 1972; Olovnikov, 1973; Allsopp *et al.*, 1992), and with each division the cell loses around 20-200 nucleotides of telomeric repeats. When at least one critically short telomere is reached, the cell is triggered to enter cell cycle arrest (senescence) and subsequently cellular death (apoptosis) (Allsopp *et al.*, 1992; Dekker *et al.*, 2011; Lazzerini-Denchi *et al.*, 2016). It is this finite limit of telomeres, promoting senescence once a telomere is critically short, that highlights their role as protective caps preventing DNA damage. Thus, telomere length (TL) reflects cellular lifespan and the capacity to divide, as TL shortens over time through cell division. For this reason, TL was proposed as a biological marker of cellular age, and in turn a marker of biological age (Von Zglinicki, 2002; Samani *et al.*, 2008).

Telomere length changes during cell division and is maintained by the ribonucleoprotein telomerase, the main components of which are a reverse transcriptase (*TERT*) and RNA template for telomeric repeats (*TERC*) (Blackburn, 1992, 2005; Palm *et al.*, 2008). Telomerase is inactive in somatic human cells, which prevents them from maintaining TL and dividing limitlessly. Telomerase has low activity in adult stem cells, which allows telomeres the ability to maintain and renew the tissue, if somatic cells are lost. Telomerase is active in stem and germ cells, where TL is maintained approximately at the same length throughout the life of the cell.

Telomere length can also be maintained by specific mechanisms, such as Alternative Lengthening of Telomeres (ALT), which is present in somatic cells (Neumann *et al.*, 2013) and employed by a subset of cancers (Apte *et al.*, 2017). ALT involves excessive telomeric DNA copying on another chromosome that gives rise to homologous recombination, a process of DNA exchange between two similar DNA sequences (Zhdanova *et al.*, 2016; Apte *et al.*, 2017).

Although telomere repeats are lost naturally in each cell division there are other mechanisms that may also lead to shortening, especially when telomeres become too long. Excessively long telomeres are trimmed, and cut-out T-loops form T-circles that

are degraded. Long telomeres may also lose sequence by the formation of C-circles under replicative stress (Rivera *et al.*, 2017).

1.2.3. Telomere length: inherited, genetic and modified by environmental factors

In the previous chapter I described how telomeres and telomere length are maintained at the cellular level. At the organismal level telomere length is a complex trait and can be considered as a measure of three important components. These are, 1) the inherited telomere length at birth, 2) the genetic effect on attrition and 3) the environmental effect on telomere attrition (Dugdale *et al.*, 2018; Entringer *et al.*, 2018).

1.2.3.1. Inherited telomere length

Telomere length inherited at birth is thought to explain the largest proportion of its effects on health, age-related diseases and longevity (Hjelmborg *et al.*, 2015; Factor-Litvak *et al.*, 2016; Entringer *et al.*, 2018; Lazarides *et al.*, 2019). TL at birth is a starting point, and longer TL allows cells to have a higher number of divisions, leading to longer tissue functioning.

Direct transmission of telomeres is thought to impact the initial inherited TL of offspring (Holohan *et al.*, 2015; Delgado *et al.*, 2019). Many factors are associated with determining TL at birth. Both paternal age at conception and environmental influences from the mother during fetal development were reported as significant factors and were studied in detail, as described below.

Observational epidemiological studies reported the association of older fathers having offspring with longer telomeres. This association was explained by the presence of telomerase activity in male germ cells that maintain TL at maximum levels during spermatogenesis and compensate for telomere attrition due to DNA replication (De Meyer *et al.*, 2007; Kimura *et al.*, 2008; Broer *et al.*, 2013; Ozturk, 2015). However, in a study by de Frutos *et al.*, 2016, it was shown that older male mice do not produce offspring with longer TL (De Frutos *et al.*, 2016). Also, the investigation by Fice *et al*., 2019, found that relative TL for paternal ageing in rats showed no difference in germ cell telomeres. It was pointed out that there may be large variability in TL within parental gametes due to their susceptibility to factors such as oxidative stress, exposure to toxic

substances and ageing (Delgado *et al.*, 2019; Fice *et al.*, 2019). Stindl in his review suggested that the association between paternal age and offspring TL is a result of confounding by birth cohort. The author explained that female germline TL decreases with age and longer TL is observed in first oocytes at younger ages and proposes that current older men have longer sperm TL because they belong to a generation that was born to younger mothers (Stindl, 2016). Either way, both a TL increase in sperms and a TL decrease in oocytes support the hypothesis of parental effect contribution to TL inheritance (Delgado *et al.*, 2019).

The conditions of fetal development were reported to have a great effect on an offspring's initial telomere biology setting (Entringer *et al.*, 2013; Ravlić *et al.*, 2018). Maternal stress during pregnancy was associated with shorter newborn TL (Entringer *et al.*, 2013, 2015; Lazarides *et al.*, 2019). Moreover, maternal health was reported to have an impact on offspring TL and their health outcomes in later life through association with shorter TL. For example, an increase in the perceived level of stress experienced by the mother (Entringer *et al.*, 2013; Send *et al.*, 2017; Lazarides *et al.*, 2019), an increase in maternal body mass index before pregnancy (Martens *et al.*, 2016), smoking during pregnancy (Osorio-Yáñez *et al.*, 2020) and increased exposure to air pollution (Song *et al.*, 2019) were all associated with shorter TL in offspring.

Inherited TL can be taken as setting the initial telomere biology for the offspring, that is highly dependent on parental factors. However, the parental role in offspring TL is greater still as the genomic inheritance from both parents also provides the genetic telomere setting that encodes the rules of response to external factors and controls TL accordingly.

1.2.3.2. Genetic telomere length

Although many genes have now been found to be implicated with telomere maintenance, those with the most established evidence are those that function within telomerase or key telomere binding complexes. It has been demonstrated that telomeres are compromised when these genes are experimentally deleted. The cells with such deletions exhibit shortened telomeres and accelerated ageing. The same outcomes have been observed in humans, where mendelian mutations in telomere-related genes compromise telomere length and function, which results in diseases of

early ageing also known as telomere syndromes or telomeropathies (discussed further in chapter 2.5. *Genetic telomere length as a driver of age-related disease*) (Armanios *et al.*, 2012; Blackburn *et al.*, 2015).

Mutations that lead to telomere syndromes are rare. In the general population TL is considered a complex phenotypic trait with multiple genetic variants associated with TL shown to have a small effect. These variants are thought to contribute to TL via multiple pathways, such as altering the biological system of TL during fetal development, providing individual resistance to telomere attrition and controlling the level of telomerase expression (Dugdale *et al.*, 2018).

The genetic determinants of telomere length are identified using genome-wide association studies that utilise TL data measured at a single time point in a large group of individuals (chapter 3. *Genome-wide association studies of telomere length*). While measured TL is thought to provide a measure of biological age, the genetic TL may show a predisposition to telomere maintenance at specific length.

A genetic predisposition to shorter or longer telomeres may contribute to biological ageing and, thus, to age-related disease risk. Investigation of telomere length and its genetic determinants and the pathways through which it affects ageing and disease development may aid in disease diagnosis, management and treatment, understanding disease aetiology and suggest therapeutic strategies and targets (Armanios, 2013).

Genetically determined telomere length is the tool utilised in this project. Up-to-date genetic determinants of TL are covered in chapter three and their use in the construction of a genetic risk score in chapter four.

1.2.3.3. Telomere length modified by environmental factors

Even though there is no change in the genetic variation that underlies the TL as a phenotype, environmental factors influence telomere maintenance by affecting the rate of telomere attrition (Dugdale *et al.*, 2018). The amount of telomere shortening depends on the initial telomere length, and also on an individual's experience when interacting with the environment, which if stressful may require larger number of cell divisions to renew the damaged tissue and restore the system.

Psychosocial stress, exposure to severe psychological trauma or psychopathological conditions have all been linked to TL (Mathur *et al.*, 2016; Dugdale *et al.*, 2018; Gorenjak *et al.*, 2019). Lifestyle choices such as diet, physical activity, smoking, alcohol consumption and other external factors were also reported to influence telomere shortening (Blackburn *et al.*, 2015; Frej *et al.*, 2015; Ravlić *et al.*, 2018).

1.3. Telomere length – a marker of biological age

TL as a phenotype is influenced by multiple factors and changes though time. An individual's TL, measured at a certain point in time, is a result of three TL components, the inherited TL from parents, genetic determinants that control TL maintenance and external effects of interaction with environment. To illustrate how TL relates to biological age I am going to use an analogy with cars.

Consider a test, where cars need to go the longest distance possible. The conditions may vary. The cars can be assembled by different manufacturers and comprise of different models with varying specifications. The cars may receive various amounts of fuel for the test and drive on different roads, straight and smooth, bumpy and muddy, or mixed. The interest of the test is to identify the furthest distance the car can go within the set parameters.

The longest distance a car can possibly go represents the car's longevity or the longest possible survival. The car manufacturer represents the parental effects. The conditions the car was assembled in may affect the build quality as, for example, maternal smoking may affect TL of offspring. The car model and specifications represent TL genetics. Some cars are more fuel efficient and were built for longer distances. The amount of fuel for the test represents the TL at birth that approximates to a number of kilometres that the car can drive or to a number of somatic cell divisions that TL allows. The quality of the road represents the environmental influences and its impact on TL. Bumpy roads are going to be more stressful causing to work harder and not be as fuel efficient.

If we take two cars with identical parameters and drive them as long as possible, we expect them to go the same distance. If we were to change, for example, the route, where one car would go on the straight smooth road, the other on a bumpy one, we would expect the first one to cover a longer distance than the second one. Similarly,

with telomeres, if we had two identical twins, where one would, for example, had been exposed to more harmful external factors, we would expect them to have different TL after prolonged exposure. If we looked at these cars or twins at the same time point, i.e. at 50 kilometres or 50 years and compare their potential to go further, we would expect to see that the one that went on the straight smooth road potentially will go much further than the second one, because the second one has already used up more fuel than the first due to exposure to a poorer road condition or due to more stressful events or harmful external factors. We would then say that at this time point the first one has more potential to go further, and to live longer, and, thus, is younger in comparison to the second one.

This is the concept of biological age that is being used as a relative measure of human health and potential longevity. Telomere length is thought to represent individual biological age that tells the current health status and surviving potential of an individual (Blackburn *et al.*, 2017).

While pre-birth and birth telomere biology settings are usually unknown, and genetic TL requires genomic data and genetic risk score estimation, the TL of an individual at a specific time point can be measured in a quite straightforward and fast way.

1.4. How to measure telomere length?

Telomere length can be measured in leucocytes and referred to as a leucocyte telomere length (LTL). This requires that a sample of blood is drawn, from which the leucocytes are isolated, and DNA extracted. The TL measurement can be performed using various methods:

- Southern blot of terminal restriction fragments (TRFs) that measures both canonical and noncanonical telomere components and provides the average TL and some indication of the distribution of TL across all chromosomes and cells within the DNA sample (Kimura *et al.*, 2010).
- Quantitative polymerase chain reaction (qPCR) measures telomere (T) in relation to single copy gene (S) and yields a relative T/S ratio that is proportional to the average TL across all chromosomes and cells (Cawthon, 2009).

- Fluorescent in situ hybridisation (FISH) measures the average length of telomere repeats in cells. This method uses labelled peptide nucleic acid probes, specific to telomere repeats, and measures fluorescence using flow cytometry (Baerlocher *et al.*, 2006).
- Single telomere length analysis (STELA) is a PCR-based technique that measures TL for individual chromosomes, but is only possible for a subset of individual chromosomes (Baird *et al.*, 2003).
- TelSeq is a software that measures average TL by utilising telomere repeat sequence reads from whole genome or exome sequence data (Ding *et al.*, 2014).

Most TL measurement methods will provide the average TL across leucocytes (Sanders *et al.*, 2013), and the choice of method depends on the aim of the study. Epidemiological population-based research tends to benefit from the fast and inexpensive qPCR method, as it is widely used and well-described in the literature. Mean TL, returned from this method, serves as a TL surrogate. Although measured LTL might not perfectly represent TL in all tissues (Sanders *et al.*, 2013), it is correlated between tissues within an individual (Demanelis *et al.*, 2020) and thought to be a sufficient marker of individual biological age for use in epidemiological studies.

1.5. Commercial telomere length measurement and anti-ageing therapies

As TL has been proposed to provide information about an individual's biological age and in turn linked to health, commercial TL measurements have also become available. A number of biotechnological companies provide this service as part of a DNA test on how to limit ageing and prolong healthy ageing by introducing a proposed lifestyle, as well as general advice, that is associated with processes intended to slow the loss of telomere repeats.

The results from the test, usually given as a mean TL, are advised to be taken with caution. The initial individual TL is usually unknown, while telomere shortening depends on the inherited length at birth. The genetic contribution to TL is not usually calculated, and genetic predisposition to TL is not reported as these genetic scores are still being developed. Moreover, the TL measurement result may oscillate over time and vary at different measurements due to measurement errors or ongoing health conditions at the

time of sample collection. Environmental factors such as having a cold, stress or trauma may have biological effects on measured LTL through influencing the proportions of different white blood cells within the sample. Any measurement error may also be due to differences in the sample collection and storing blood, issues with DNA extraction and assays used (Epel *et al.*, 2017; Codd *et al.*, 2021).

Furthermore, most statistical estimates of TL association with disease risk are obtained from a large group of individuals. Interpretation of TL results at the individual level can be difficult and not meaningful, because the range of TL at each age is wide. Age-based norms for TL are being developed (Blauwkamp *et al.*, 2017), but the variation of TL in normal individuals (Aubert *et al.*, 2012) still imposes difficulties in comparing one individual to another. For example, overlapping TL ranges may result in the same value for individuals of 40 and 80 years old.

Nonetheless, individual TL measurement may detect telomeres of extreme length that would indicate a potentially high risk of age-related diseases. The detection may help to evaluate the interventions required to delay the onset of age-related conditions (Gorenjak *et al.*, 2018).

Anti-ageing therapies targeting telomeres were proposed after investigations into telomere syndromes and human ageing. Telomere shortening can be reversed by telomerase that adds telomeric repeats. While most somatic cells have telomerase deactivated, its activity may still be present on lower levels. Pathways that regulate telomerase activity and, thus, control TL, are being investigated as specific targets for anti-ageing therapy. These involve identification of genes, involved in telomere biology, screening for the effects of various mutations in the promoter regions of telomerase genes, and testing the effects of natural compounds on telomerase activity. The long term effect of telomerase reactivation are not understood and, as telomerase reactivation is one of cancer hallmarks, thorough investigations are required in order to avoid undesirable consequences (Jafri *et al.*, 2016; Tsoukalas *et al.*, 2019).

While TL measurement can be performed at the individual level they can be considered as an experiment and not be taken too seriously by the individual. It can provide certainty that the individual is free from conditions that exhibit extreme telomere lengths, and give advice on improving lifestyle and health, based on results from

association studies of TL and environmental factors that are in turn associated with disease-free life and longevity.

TL measured within a population is more useful in epidemiological research that provides us with population statistics, identification of impactful genetic and environmental factors, and allows for estimation of age-related risks. One such example is the ongoing investigation of telomere length in UK Biobank, a large cohort that is going to be covered in detail in the next chapter.

1.6. UK Biobank

UK Biobank (UKB) is a large population-based cohort, initiated to improve public health by providing an opportunity for researchers to explore and analyse genetic and non-genetic determinants of diseases on a large scale. UKB has detailed genetic data and extensive phenotypic information including medical history and lifestyle on approximately 500,000 individuals aged 40-69 years that were recruited between 2006 and 2010 (Marchini, 2015; Sudlow *et al.*, 2015; UK Biobank, 2015). UKB is the primary dataset of this project.

As mentioned previously, TL is a marker of biological age and consists of three components: initial inherited TL, genetic TL and TL modified by external factors. The initial TL is not within the scope of this project, as TL was not obtained from individuals in UKB at their birth. The TL of UKB participants has been measured using the qPCR method within our research group at the Department of Cardiovascular Sciences, University of Leicester. Unfortunately, at the time of this project the data was not available and, thus, not included here. This data is expected to provide a wider knowledge of genetic determinants of TL and the effects of genetic and measured TL, coupled with external factors, on human health, healthy ageing, and longevity. The aims of this project are met with the use of genetic data for TL and its effects on human ageing. The genetic and phenotypic data of UKB is now going to be described with a focus on the details required for the analyses of genetic TL.

1.6.1. UK Biobank genotypic data

1.6.1.1. Imputed genotype files

Most of the individuals within UKB were genotyped with a custom Affymetrix UK Biobank Axiom array with ~800,000 SNPs (Sudlow *et al.*, 2015) while 50,000 were genotyped with a custom UK Biobank Lung Exome Variant Evaluation (BiLEVE) Affymetrix Axiom array (UK Biobank, 2015; Wain *et al.*, 2015). The two arrays are very similar though an adjustment for array within any genetic analysis is recommended by UKB. The causal polymorphisms may not always be directly genotyped on the array. Whole genome imputation was performed to greatly increase the number of SNPs for each individual in order to analyse as much genetic information as possible. Whole genome imputation comprises of two steps: 1) pre-phasing that infers haplotype structures for each individual, and 2) imputation that uses a reference panel to fill in the gaps within the inferred haplotypes (Marchini *et al.*, 2010; Marchini, 2015).

The process is complex and as such an example of the process involved whole genome imputation is given in **Figure 1.2**. We can consider the example data to be similar to UKB genotypic data, which is used as the input data to be imputed. The reference panel provides haplotypes for inferring missing data by matching the sequence in the observed data. Each line in the reference panel is coloured and represents a haplotype. In pre-phasing each UKB haplotype with missing genomic information is mapped to a reference panel and the best haplotype match is found. The imputation phase fills the missing parts of within the UKB haplotypes by the most probably identified matches from the reference panel. For example, in **Figure 1.2** the first UKB individual's genotype matches to parts of both the red and light green haplotypes from the reference panel, and this was used to fill in missing genotypic information for this individual. The resulting imputed genotype file is transformed to reflect the uncertainty of each imputed genotype by reporting the genotype as a probability at each position in the genome (Marchini *et al.*, 2010; Marchini, 2015; Reed *et al.*, 2015).

UKB imputed the array data centrally using two reference panels, the Haplotype Reference Consortium (HRC) panel and UK10K + 1000 Genomes panel, and enlarged the number of genetic variants to ~96 million variants (Huang *et al.*, 2015; Marchini, 2015).

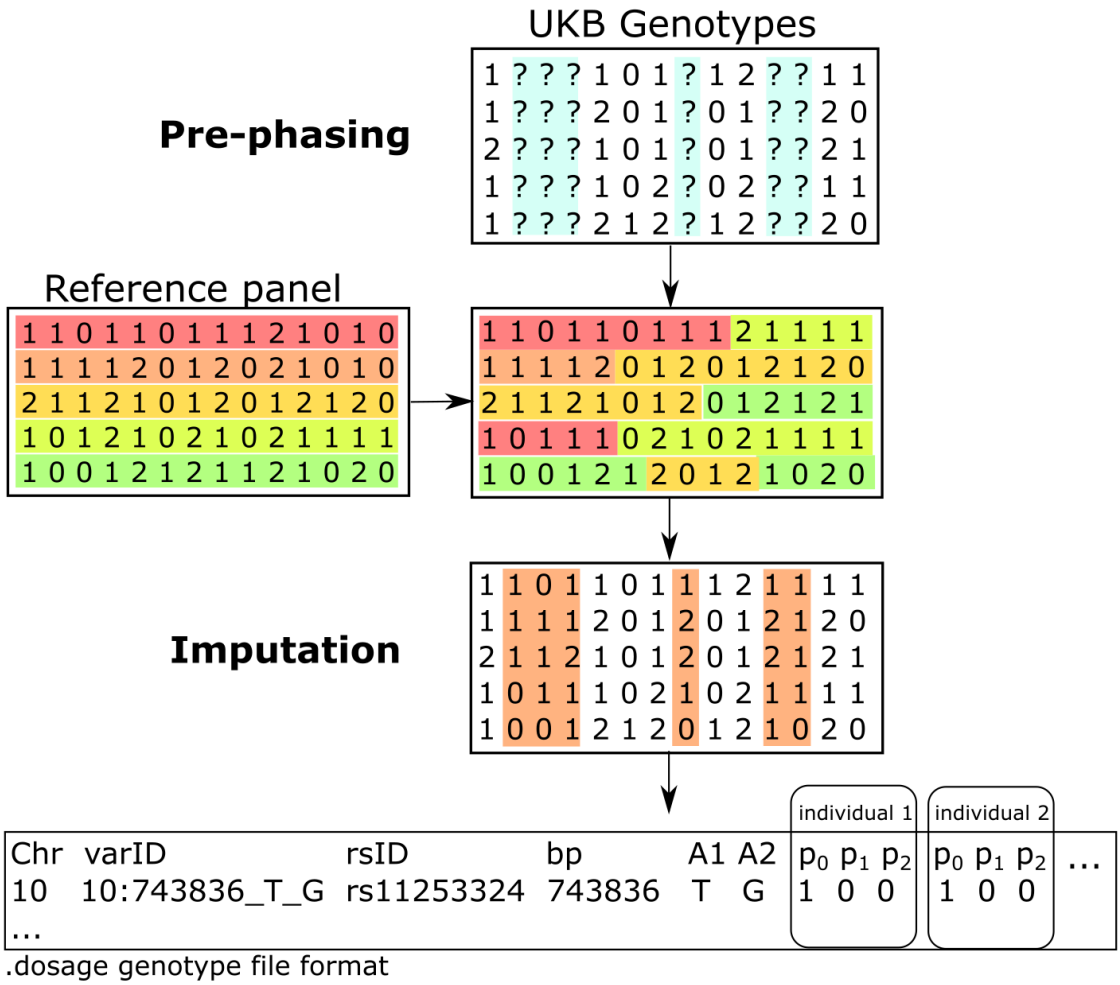


Figure 1.2. UK Biobank genotype imputation and the format of imputed genotype data. Haplotypes are inferred within UKB genotypes in pre-phasing, and gaps are filled in using reference panel in imputation step. The imputed genotype data is stored in .dosage file format with following columns: *Chr* – chromosome, *varID* – variant unique identifier, *rsID* – variant name, *bp* – base pair position, *A1* – allele T, *A2* – allele G, p_0 – probability of genotype TT of individual 1, p_1 – probability of genotype TG of individual 1, p_2 – probability of genotype GG of individual 1.

1.6.1.2. UK Biobank quality control

As with all genomic studies prior to imputation it is essential to perform some quality control (QC). For UKB this was performed at the SNP level and then at the individual level to remove poor quality data that may introduce bias in the imputation which could lead to false associations.

SNP level QC requires a variant to be excluded if:

1) The SNP has a low call rate, i.e., missing genotype call meaning that not all individuals have the specified SNP called through genotyping. UKB specified two cut-offs: 97% call rate for a primary subset of SNPs and 94% for a smaller subset of SNPs (UK Biobank, 2015).
2) There is a significant deviation from Hardy-Weinberg equilibrium (HWE) with a p-value<10^{-6}. HWE describes a state in which genotype frequencies in the population remain the same over generations. The disequilibrium in genotype frequencies may be a result of bias, genetic drift, immigration, selection or non-random mating (Dorak, 2017).
3) The number of carriers of the minor allele is too small. When the majority of individuals have two copies of the major allele, also referred as homozygosity at a given SNP, the minor allele frequency (MAF) is too small and is likely to be identified as a risk allele by GWAS (Kido *et al.*, 2018). It is common to remove markers with a low MAF≤0.01 to avoid misleading associations.

The sample level QC are then performed and UKB excluded any individual with insufficient data quality, sex mismatches or relatedness (Anderson *et al.*, 2011; Reed *et al.*, 2015; UK Biobank, 2015; Donaldson *et al.*, 2016; Evangelou, 2018). There are also additional QC measures that can be applied to the genetic data. UKB provides post-imputation per-individual QC-criteria in the phenotypic data set via dedicated variables covering four areas for QC:

1) Heterozygosity: A poor heterozygosity flag is given by UKB variable 22010. Having two alleles at a locus is called heterozygosity, and poor heterozygosity means that extreme or little genetic variability is present and deviation from HWE is likely. Flagging this with a value of 1 indicates that the genotype data is of insufficient quality and should be excluded from analyses such as GWAS (UK Biobank, 2015).
2) Sex mismatch: Genetic sex can be estimated using an appropriate algorithm to call SNPs on the sex chromosomes and differentiate between the X and Y chromosomes. The genetic sex is given by UKB variable 22001 and reported sex by UKB variable 31. A genetic sex value of 0 indicates that the sample is female, and 1 is male. A mismatch between genetic and reported sex may occur due to

clerical error, genetic sex not matching gender identity, or an abnormal number of sex chromosomes also referred to as sex chromosome aneuploidy. The mismatch indicates that these samples should be excluded.

3) BiLEVE study: Data analysed for the BiLEVE study (Wain *et al.*, 2015), that arrayed around 50,000 samples outside the UKB Biobank, have two UKB QC variables 22050 (*UKBiLEVE Affymetrix quality control for samples*) and 22051 (*UKBiLEVE genotype quality control for samples*). For both QC metrics a value of 0 indicates that the sample failed quality control in the UKBiLEVE project and should be excluded from genetic analyses.
4) Kinship: UKB did not purposefully collect related samples and yet one needs to account for correlated data. Individuals are assessed for genetic relatedness via the kinship coefficient, calculated using Kinship-based INference for Genome-wide association studies (KING) robust estimator, which is given in UKB variables 22011 (*Genetic relatedness pairing*) and 22012 (*Genetic relatedness factor*) (Manichaikul *et al.*, 2010; UK Biobank, 2015). The kinship coefficient shows the level of relatedness: < 0.044 – unrelated individuals, 0.044 – 0.088 – 3rd degree relatives, 0.088 – 0.177 – 2nd degree relatives, 0.177 – 0.354 – 1st degree relatives (parent-child or full siblings) and > 0.354 – monozygotic twins.

   A good practice is to remove one individual from a pair of closely related relatives in order to keep as many participants as possible for the analysis. For example, using a kinship coefficient threshold of 0.088, we allow only unrelated and 3rd degree relatives to be included into analyses. We keep only one individual of the genetically related pair with the least amount of missing genetic data that is given in UKB variable 22005 or if this is equal then a sample is randomly selected (**Figure 1.3**).

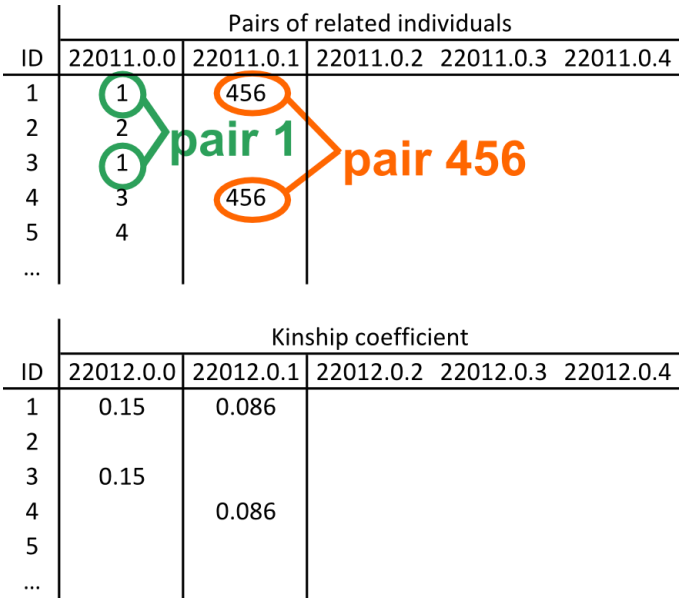


Figure 1.3. A genetic relatedness pairing in UK Biobank. The top table has individual IDs and columns that store the pairing identifier (22011.0.X). The pair of related individuals with UKB ID 1 and 3 are assigned the unique pair identifier=1, while individuals with UKB ID 1 and 4 – are given pair identifier=456. The bottom table shows the estimated level of genetic relatedness between paired individuals. This example shows that UKB ID 1 is related to UKB ID 3 with a kinship coefficient of 0.15, which means that they are likely 2nd degree relatives. It is also shown that UKB ID 1 is also related to the UKB ID 4 with a kinship coefficient of 0.086, which means that they are close to being 3rd degree relatives. We cannot keep both UKB ID 1 and ID 3 – so we remove the sample with the most missing data relevant to the project or randomly exclude if the amount of missing data is equal. We would keep UKB ID 4 in study with UKB ID 1 as the kinship coefficient is less than 0.088.

1.6.2. UK Biobank phenotypic data

Now that the genetic data are described I will detail the UK Biobank phenotypic dataset. This includes health and lifestyle information for all ~500,000 individuals. The main phenotypic dataset is highly detailed and complex. The participants filled in a questionnaire that involved questions about individual background, lifestyle, physical and mental health, diet, and additional exposures.

For purposes of this project, the data collected on health from the questionnaire within the phenotypic dataset is referred to as self-reported data. The participants were asked to report non-cancer and cancer illnesses by selecting them from a list. In situations where there is uncertainty of the specific illness experienced a trained nurse would

attempt to find the appropriate illness code or a clinical doctor would examine a free-text description that was entered. For each illness participants were asked to enter the age or year when they were ill. This way most of self-reported data on health conditions was collected. It covered 385,933 participants and 1,127,398 items of data at the baseline visit.

In order to work with this data, it is important to understand its structure and how it was collected and stored. UKB provides a valuable online resource centre via the UKB data showcase. This provides researchers a summary for each available UKB data-field including statistical summaries of the variable and the total sample that it is available for. For example, data-field 3894 in **Figure 1.4** shows data for the age at which a heart attack was diagnosed. At the top of the page, summary tables show the number of participants that reported this variable, types of values and time range, when the data was collected. The first tab *Data* in the middle presents the graphical summary of the distribution of age at heart attack with summary statistics on the right. Other tabs provide additional information on instances, also known as assessments with different data collection times, notes describing this data-field, categories this data-field falls in, related fields with data on similar variables, and resources that were used to collect the data, which usually provides information on how exactly the question was asked in the questionnaire.

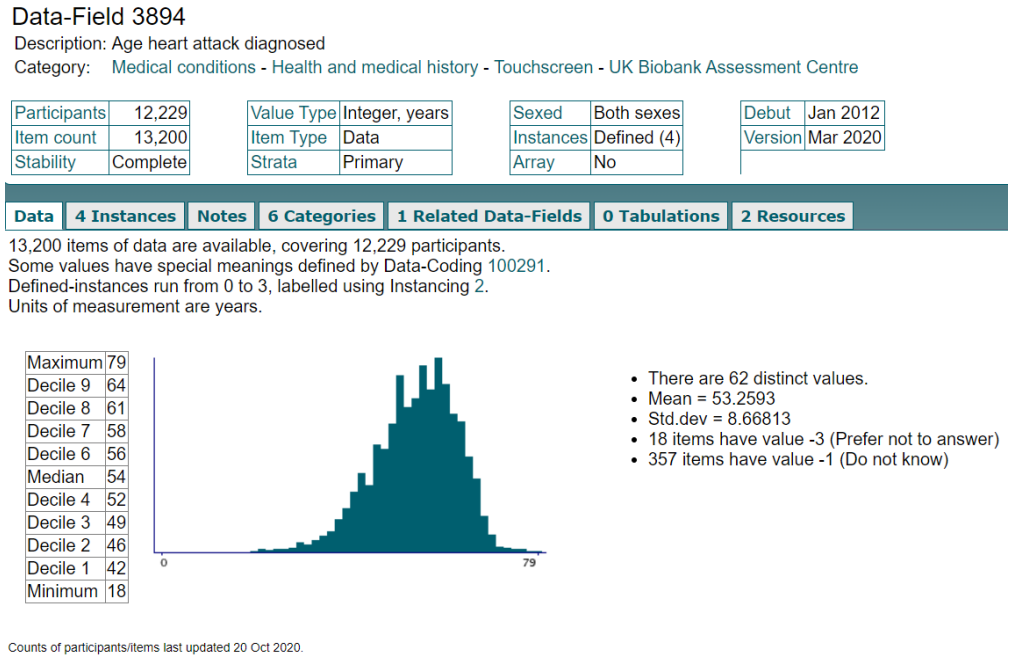


Figure 1.4. UK Biobank showcase of data-field 3894 (Sudlow *et al.*, 2015; UK Biobank, 2020).

The UKB phenotypic dataset allows for separate columns to represent a single variable, termed as a data-field in UKB, which is contained in a nested structure, incorporating repeated assessments and measurements. For example, variable 4080 that has data for the automated reading of systolic blood pressure (**Figure 1.5**).

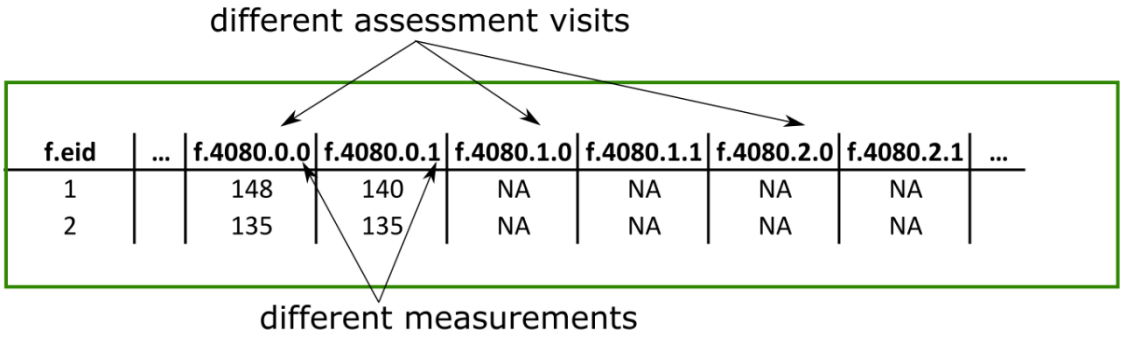


Figure 1.5. A nested structure of UK Biobank data-field 4080 for automated reading of systolic blood pressure.

Data-field 4080 has 6 columns of data:

1. 4080.0.0 – Initial assessment visit measurement 1.
2. 4080.0.1 – Initial assessment visit measurement 2.
3. 4080.1.0 – First repeat assessment visit measurement 1.
4. 4080.1.1 – First repeat assessment visit measurement 2.
5. 4080.2.0 – Imaging visit assessment visit measurement 1.
6. 4080.2.1 – Imaging visit assessment visit measurement 2.

There are three different assessment visits (0, 1, 2) that are encoded by the first digit after the data-field code 4080. It is followed by a final digit that shows the different measurements of blood pressure that were taken a few moments apart at the same visit. The initial assessment visit records have data recorded on all samples as this is the baseline visit, while other assessments are comparatively small data collections. For example, systolic blood pressure in data-field 4080 was measured for 472,374 individuals at initial assessment visit, and only for 20,287 at the first repeat assessment visit. It should also be noted that individuals at repeat visits may not be mutually exclusive, but measurements are independent.

The majority of UKB data are contained in the self-reported questionnaire and only collected at one specific time point for most data. All hospital events are recorded in the Hospital Episode Statistics (HES) data that were stored in separate tables at the time of this project:

1) *HESIN* is the main table containing hospital records. It contains all primary information on each hospital episode for each individual and stores information on primary diagnosis and primary operations. Each individual can appear multiple times and have multiple rows of data, but each new event has a unique record ID that appears only once.
2) *HESIN_diag10* contains secondary diagnoses coded using ICD10 (ICD - the international classification of diseases) for each hospital episode. There is also a complementary table, *HESIN_diag9*, containing older records that were encoded with ICD9 and this has the same structure as *HESIN_diag10*.
3) *HESIN_oper* contains all the secondary operations for each hospital episode.

The structure of these tables is illustrated in **Figure 1.6**.

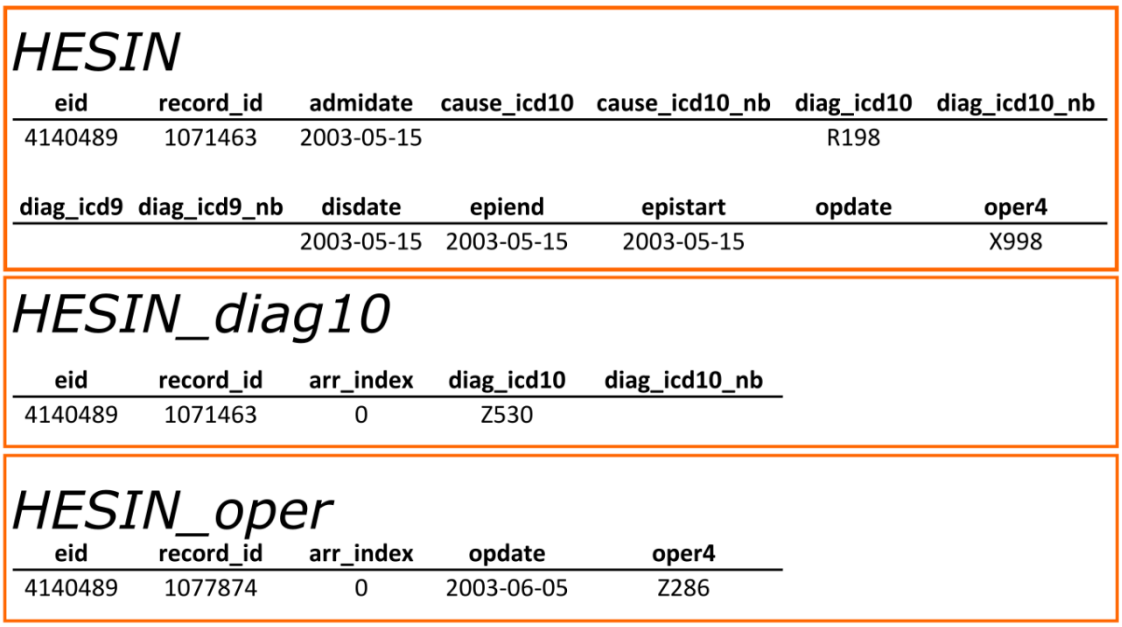


Figure 1.6. The structure of hospital episodes statistics data tables.

In the *HESIN* table we see an individual with UKB ID, *eid*, 4140489 who has a recorded event on the 15th of May 2003 with unique *record_id*=1071463. This event has a primary diagnosis code, *diag_icd10*, 'R198', which means 'Other specified symptoms and signs involving the digestive system and abdomen'. This same record *record_id*=1071463 is also in *HESIN_diag10* table and shows the secondary diagnosis with ICD10 code, *diag_icd10*, 'Z530', which is for 'Procedure and treatment not carried out because of contraindication'. In *HESIN_oper* this individual has another event on a different date with operation code, *oper4*, 'Z286', which means that a procedure or operation was performed on their sigmoid colon. These records for individual 4140489 are just a subset of hospital events for this individual for use as an example and there are more records for this individual in all three tables. It must also be noted that HES tables use both ICD9, ICD10, and have operation codes in OPCS-4 (Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures, 4th revision), while the phenotypic UKB data file has self-reported events utilising a different UKB specific encoding of events and operations.

To summarise, in this chapter I described the relevant background information that is required to research the stated aims and objectives of this project. These included basic genomics, telomere biology, telomere length and its association with ageing. Concepts of inherited, genetic, and modified TL were introduced, highlighting that this project focuses on investigating genetic TL and its effect on age-related conditions using data available in UKB, for which the data structure was detailed. In the following chapter I am going to provide an in-depth investigation of current uses of genetic TL using an up-to-date literature review of studies, and how these published data work within the current projects aims and objectives about the effects of genetic TL on age-related diseases.

Chapter 2. Observational association studies of telomere length and age-related diseases

2.1. Literature review

2.1.1. Known telomere epidemiology literature

Since the association was made between the 'end replication problem' of chromosomal ends and limited replicative capacity (Olovnikov, 1973) and between the subsequent association of telomere shortening and replicative capacity of the cell in vitro (Harley *et al.*, 1990; Allsopp *et al.*, 1992), telomeres, more specifically telomere length, have become the subject for a large array of studies examining the effects of telomere attrition on human health and mortality (Olovnikov, 1996; Turner *et al.*, 2019).

In this chapter I review the telomere length literature in the field of epidemiology. These data were collected via a literature search and are used to enforce the motivations behind the project hypotheses.

I start by describing the literature review strategy for telomere-related themes. Following this I present and clarify the current knowledge, identified in the search, relating the effects of TL on human health. I also highlight the implications and gaps in current knowledge that places this thesis within the field of telomere length epidemiological research.

2.1.2. Identifying telomere literature themes to review

To conduct the required literature searches I followed guidelines set out by Leite and colleagues (Leite *et al.*, 2019). The review for each telomere theme was performed using a specific strategy that consists of four steps:

1. **Identification**. I defined the main topics, identified keywords, and searched for publications via three literature search engines: PubMed, Google Scholar and the University of Leicester Library. The main topics covered: telomere biology, telomere length, TL association with ageing (including telomere syndromes), cardiovascular diseases, cancers and other disease phenotypes, telomere length

genome-wide association study, genetic risk score of TL, mendelian randomisation of TL.

2. **Screening**. I screened the titles and abstracts of all identified publications, selecting those that were relevant to the search terms and addressed the research question directly with telomeres.
3. **Eligibility**. I further read through and checked, whether the full text was available and the study provided a sufficient description of the methods related to the search terms used to identify it.
4. **Inclusion**. I included all relevant publications available up to the final search date of November 1st, 2020, that satisfied my eligibility criteria.

2.2. Literature review on telomere biology and telomere length

The literature review on telomere biology was performed to accumulate and summarise the biological understanding of telomeres, their structure and function. **Figure 2.1** shows an introductory example of the applied literature review procedure. Here I describe the entire strategy based on four steps in text and a figure. Other searches are going to be detailed in a figure.

The literature review of telomere biology was performed in four steps: identification, screening, eligibility, and inclusion as described above. The keyword list consisted of primary search terms 'telomere' or 'telomere components' and secondary terms such as 'biology', 'function', 'structure', 'maintenance', etc (**Figure 2.1**). I searched for combinations of these keywords in PubMed database, Google Scholar and University of Leicester Library. I screened the titles and abstracts of found publications, and selected ones that met the following criteria: 1) telomere was defined as a primary subject of a study, 2) telomere was defined, and its function and structure were described, 3) mechanisms related to telomeres were investigated. I further read through relevant publications and included those that met eligibility criteria such as: 1) availability in full text, 2) description of telomere and its function, 3) description of experimental design and analyses. Selected publications were used to summarise the current knowledge of telomere biology that was covered in the first chapter.

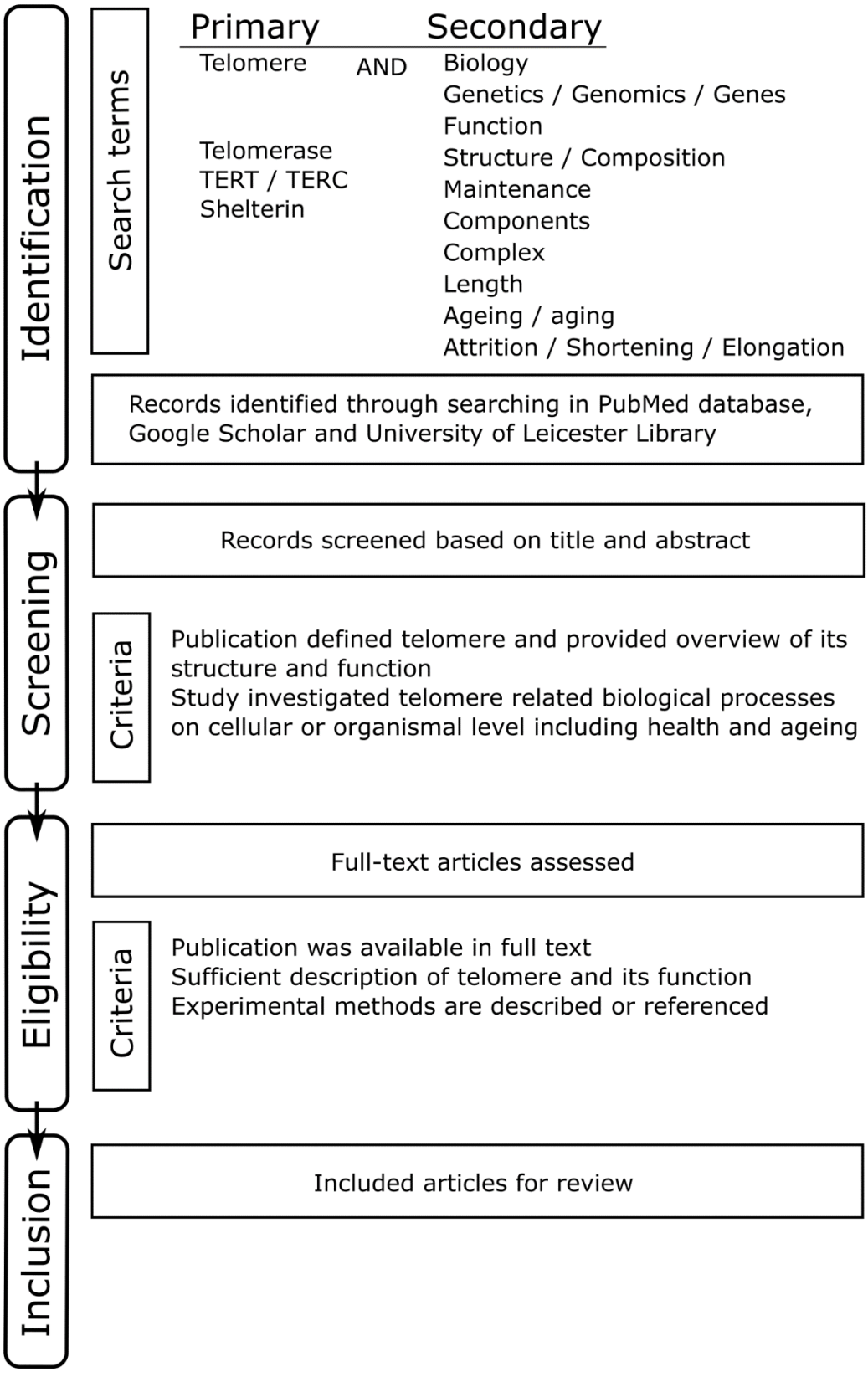


Figure 2.1. Flowchart of literature review on telomere biology.

Following literature review flowcharts can be read in a similar manner.

The identified publications that investigated telomere length, its heritability, genomics, and relationship with environmental factors were screened for eligibility and included if they met the required criteria using the described four steps as detailed in **Figure 2.2.**

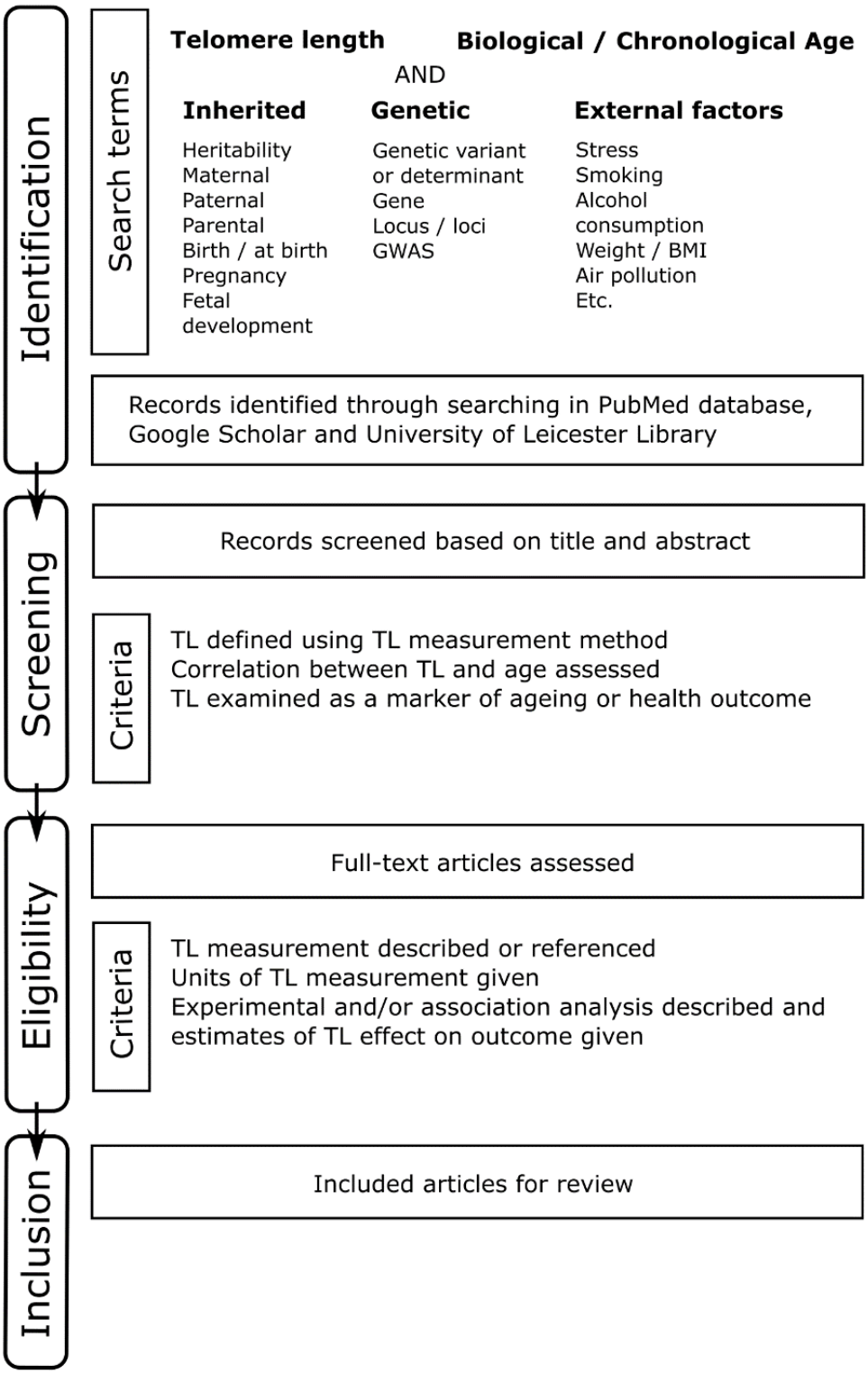


Figure 2.2. Flowchart of literature review on inherited, genetic, and environmentally modified telomere length.

Both literature reviews, on telomere biology and telomere length, were performed to give an introduction to the field, bring understanding to relevant telomere terms and concepts, and were detailed in the first chapter. For this reason, they will not be described further here. It should be noted that the number of publications identified have not been given due to the number of searches performed and the wide breadth of papers identified. The following chapters detail a full review of the literature that go beyond the background information already presented and as such are reviewed in more detail with respect to the project hypotheses.

2.3. Association between telomere length and age-related diseases

Since the association between telomere shortening and replicative capacity was observed on the cellular level (Harley *et al.*, 1990; Allsopp *et al.*, 1992), research interest was then drawn to the effects of telomere shortening on the organismal level. Mean and median TL measured in blood or other tissues became a useful metric to represent the overall organismal TL and has thus been used to investigate the effects of telomere attrition on human health and lifespan (Olovnikov, 1996; Turner *et al.*, 2019). Therefore, a literature review to investigate the current understanding of the association between measured TL and age-related diseases is necessary to introduce the current knowledge of the impact of TL on human health. This literature search is detailed using four step procedure with details of search terms and eligibility shown in **Figure 2.3**.

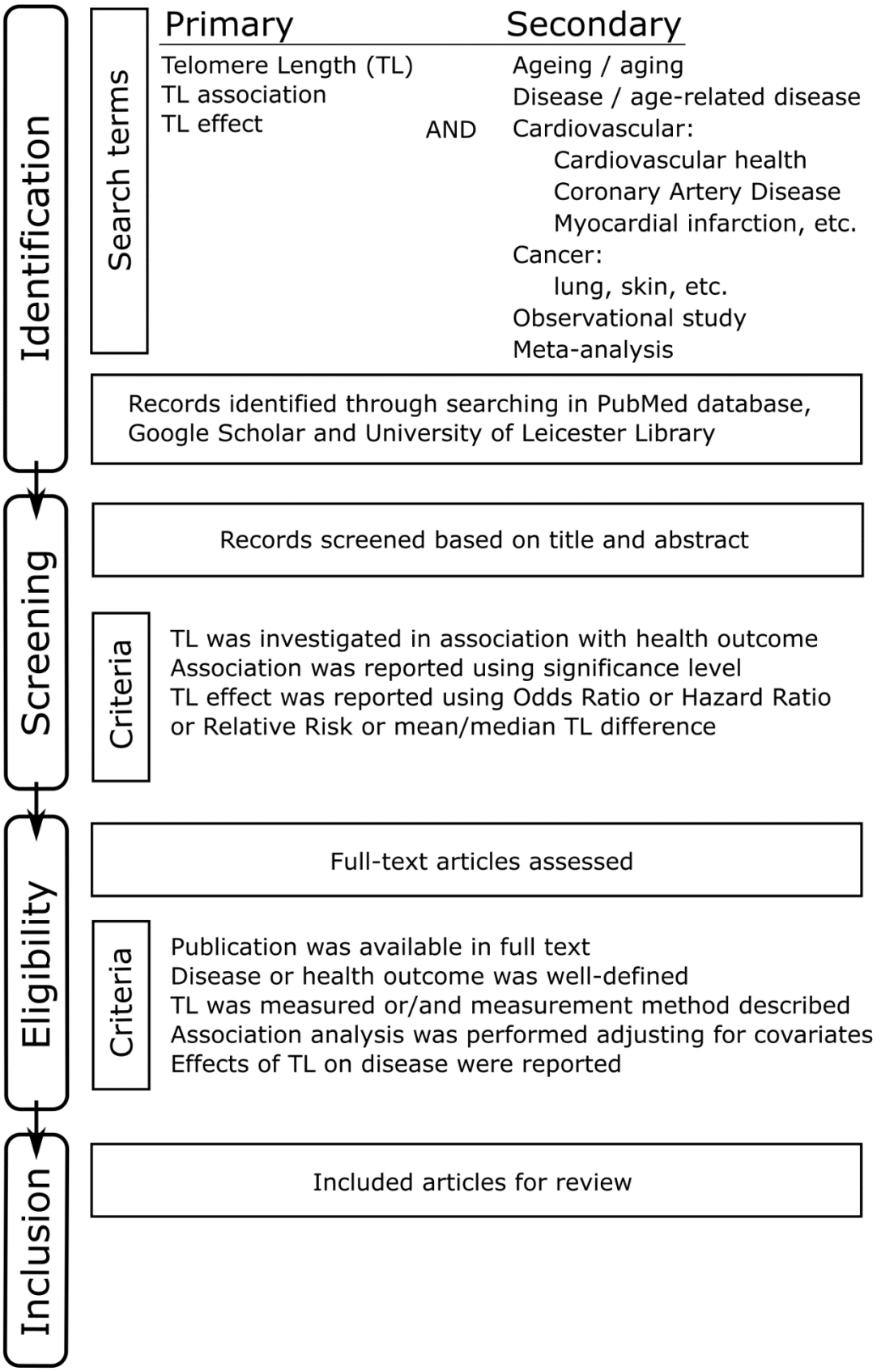


Figure 2.3. Flowchart of literature review on observational associations studies between telomere length and age-related diseases.

The range of age-related health outcomes is wide and while TL was studied in relation to many of them, I aimed to prioritise the search of publications on cardiovascular and cancer disease. Due to being leading causes of death worldwide these phenotypes were studied in more detail in previous research and, thus, had accumulated substantial literature and knowledge.

2.3.1. Telomere length and cardiovascular disease literature

Conditions that affect the heart or blood vessels are generally known as a cardiovascular disease, which is a leading cause of death worldwide and is related to ageing. This complex phenotype has many common contributing factors. It has high interindividual variability that poses difficulties in prevention and treatment. Telomere length, also related to ageing, is gaining more evidence as a link between telomere dysfunction and risk of cardiovascular outcomes (Tracy, 2003; Huzen *et al.*, 2010; Zhu *et al.*, 2013; Paneni *et al.*, 2017; Chiriacò *et al.*, 2019).

Ageing of the heart is characterised by physiological changes in aorta, arteries and epicardium. An increased turnover of cardiomyocytes, the cells of heart muscles, has been observed in the ageing heart as well as increased senescence of cardiac stem cells (Seki *et al.*, 2015; Nalobin *et al.*, 2020). This corresponds with the regenerative process, when senescent or dysfunctional cardiomyocytes need to be renewed, and corresponds with TL limiting the regenerative capacity of cells. One would expect to see shorter TL due to the accelerating loss of telomeric repeats via an exhaustion of somatic and stem cells that are stimulated to regenerate heart tissue.

Benetos *et al*., 2018, measured TL in muscle cells and leucocytes from patients with atherosclerotic cardiovascular disease (CVD) and controls. The authors propose that TL measured in muscle cells is more reflective of TL at birth, as these cells replicate less frequently than blood cells. Shorter LTL was observed and the difference between LTL and TL in muscles within an individual was reported to increase with age, as more cell divisions occur in blood cells. Authors reported a larger difference between LTL and TL in muscles in CVD patients, suggesting an accelerated attrition of LTL in cardiovascular disease (Benetos *et al.*, 2018).

Xu *et al*., 2020, analysed the association between measured LTL and cardiovascular health outcomes in 7,378 individuals. The study subjects were under 65 years old, with body mass index and blood cholesterol in the normal ranges. Individuals were engaged in physical activities and without health conditions such as hypertension or diabetes. LTL was measured during the initial assessment when participants reported any cardiovascular health related problems such as heart failure, coronary artery disease, angina, heart attack or stroke. In total, there were 821 subjects with at least one cardiovascular outcome. Longer LTL was associated with a protective effect on cardiovascular disease prevalence. Each 1 kilobase increase in LTL was associated with a 21% decrease in CVD prevalence (Odds Ratio (OR)=0.79, [95%CI:0.63-0.98]) (Xu *et al.*, 2020).

Haycock *et al*., 2014, systematically reviewed and meta-analysed 24 prospective and retrospective studies that reported the association between measured LTL and coronary heart disease. LTL distribution was divided into thirds and the shortest third was compared to the longest. Shorter LTL was associated with an increased relative risk (RR) for coronary heart disease (RR=1.54 [95%CI:1.30-1.83]). The direction of association was the same in both retrospective and prospective studies (Haycock *et al.*, 2014).

D'Mello *et al*., 2015, meta-analysed 27 observational studies and reported the association of a 1 standard deviation (SD) decrease in LTL with an increased risk of stroke (OR=1.21, [95%CI:1.06-1.37], based on 10 studies) and myocardial infarction (OR=1.24, [95%CI:1.04-1.47], based on 6 studies). Estimates were pooled from both cross-sectional and prospective studies. D'Mello and colleagues did not observe a significant association between LTL and CAD, when meta-analysing 7 studies (OR=1.03 [95%CI:0.98-1.08]) (D'Mello *et al.*, 2015).

Jin *et al*., 2018, studied the association between TL and stroke meta-analysing 11 studies. Authors reported that shortened TL was significantly associated with increased risk of stroke in combined analysis of retrospective and prospective studies (OR=1.50 [95%CI:1.13-2.0]), while this relationship was not statistically significant in only retrospective or only prospective studies (Jin *et al.*, 2018).

Fitzpatrick *et al*., 2007, measured LTL by mean terminal restriction fragments in 419 subjects and performed an association analysis of LTL and incident cases of myocardial infarction and stroke. They reported the association of shorter LTL and increased risk of

myocardial infarction (Hazard Ratio (HR)=3.08, [95%CI:1.22-7.73]) and stroke (HR=3.22, [95%CI:1.29-8.02]) in individuals of 73 years old and younger. The number of incident events was only 17 for myocardial infarction and 17 for stroke over 7 years of follow-up. Although the number of events was small, the result suggests that shorter LTL is associated with the incidence of cardiovascular disease and may be a causal factor involved in disease development (Fitzpatrick *et al.*, 2007).

Weischer *et al*., 2012, measured LTL in 19,838 individuals from 2 prospective studies, the participants of which were followed for up to 19 years. The study evaluated the association between LTL and incident cases of ischemic heart disease (n=2038) and myocardial infarction (n=929). A significant association was reported with an increased hazard for both ischemic heart disease (HR=1.06 [95%:1.00-1.11]) and myocardial infarction (HR=1.10 [95%:1.01-1.19]) for every 1000 base pair decrease in LTL. The results of the study suggest that shorter LTL may be a marker of degenerative processes, may be involved with cellular damage and stress and promoting tissue turnover with subsequent telomere shortening and the observed increase in risk of cardiovascular diseases (Weischer *et al.*, 2012).

In a study by Gebreab *et al*., 2017, cardiovascular health was assessed as a metric of seven cardiovascular disease related factors including blood pressure, total cholesterol, fasting blood sugar, smoking status, physical activity, diet, and body mass index. These metrics were divided into 3 categories to represent poor, intermediate and ideal cardiovascular health. The cardiovascular health metric and LTL were measured in 5194 adults (aged ≥20). The LTL was reported to be 3.4% shorter in individuals with poor cardiovascular health in comparison to the ideal category (Gebreab *et al.*, 2017).

There are many more studies telling a similar story and these are summarised in **Table 2.1**. Shortened TL is consistently associated with an increased risk of cardiovascular diseases, however, the strength of evidence and estimated effect sizes for specific diseases vary. The reliability of association depends on the selection of study subjects and the study type (retrospective or prospective), the number of participants and other cofactors. These are highlighted in the table along with the relevant result outcomes.

Pheno-type	Sample size	Results	TL effect	Reference
Cardio-vascular Health	5,194	Individuals with poor CVH had shorter LTL than individuals with ideal CVH	-3.4% [95%CI=-6.0,-0.8]	(Gebreab *et al.*, 2017)
CVD	7,378	LTL was associated with the risk of CVD	OR=0.79 [95%CI:0.63-0.98]	(Xu *et al.*, 2020)
CHD	589/653	LTL was shorter in CHD compared to the non-CHD subjects	CHD (overall 8.68 kb, SD 4.65) compared to the non-CHD (9.23 kb, SD 4.83; P=0.012)	(Maubaret *et al.*, 2010)
CHD	5,150/9,341*	TL in patients with CHD was shorter than in controls	Standard mean diference =-0.45 [95%CI:-0.65, -0.25)], P<0.0001	(Xu *et al.*, 2019)
CHD	8400*	Shorter LTL associated with increased risk of CHD	OR=1.54 [95%CI:1.30-1.83]	(Haycock *et al.*, 2014)
CAD	1,511/1,553	1 SD decrease in TL associated with increased risk of CAD	OR=1.17 [95%CI:1.09-1.26]	(Wang *et al.*, 2019)
CAD	566	Shorter TL associated with increased risk of cardiovascular outcomes	HR=1.8 [95%CI:1.1-2.0]	(Hammadah *et al.*, 2017)
CAD	366	Shorter TL associated with increased risk of CAD	HR=2.866 [95%CI:1.83-4.50]	(Sun *et al.*, 2020)
IHD	17,235	200-bp-shorter TL associated with increased risk of IHD	OR=1.02 [95%CI:1.01-1.03]	(Madrid *et al.*, 2016)
IHD	2,038	Short TL is associated with increased risk of IHD	HR=1.06 [95%CI:1.00-1.11] per 1kb decrease in TL	(Weischer *et al.*, 2012)
MI	203/180	MI cases had shorter mean TL than controls	Mean TRF length difference 299.7±69.3 bp, P<0.0001	(Brouilette *et al.*, 2003)
MI	419	Shortened kb of TL TRF corresponded with increased risk of MI	HR=3.08 [95%CI:1.22-7.73]	(Fitzpatrick *et al.*, 2007)
MI	929	Short TL associated with increased risk of MI	HR=1.10 [95%CI:1.01-1.19] per 1kb decrease in TL	(Weischer *et al.*, 2012)
MI	27 obs. studies*	A decrease in LTL associated with increased risk of MI	OR=1.24 [95%CI:1.04-1.47]	(Hunt *et al.*, 2015)
Stroke	27 obs. studies*	A decrease in LTL associated with increased risk of stroke	OR=1.21 [95%CI:1.06-1.37]	(Hunt *et al.*, 2015)
Stroke	187	Short LTL associated with an increased risk of cardioembolic stroke in AF patients	OR=2.93 [95%CI:1.24-6.94]	(Allende *et al.*, 2016)
Stroke	25,340 participants	Significant relationship between shortened TL and stroke	OR=1.50 [95%CI:1.13-2.0]	(Jin *et al.*, 2018)
Stroke	1,309/1,309	Mean TL significantly shorter in stroke patients	OR=2.12 [95%CI:1.62-2.77]	(Ding *et al.*, 2012)
Stroke	300/300	Shorter relative TL associated with an increased risk of stroke	OR=8.44 [95%CI:5.42-13.14]	(Gao *et al.*, 2018)
Stroke	543/616*	Shorter TL associated with increased risk of ischemic stroke	RR=1.12 [95%CI:1.05-1.19]	(Li *et al.*, 2018)
Stroke	419	Shortened TL (TRF) corresponded with increased risk of stroke	HR=3.22 [95%CI:1.29-8.02]	(Fitzpatrick *et al.*, 2007)
Stroke	504	No association found between relative TL and stroke in women	OR=0.82 [95%CI:0.52-1.32]	(Schürks *et al.*, 2013)
Stroke	486	Individuals with shorter TL had a higher presence of atherothrombotic stroke	OR=1.37 [95%CI:1.06-1.77]	(Zhang *et al.*, 2013)
Stroke	152	Longer TL associated with reduced risk of stroke	OR=0.748 [0.681-0.823]	(Xiao *et al.*, 2019)
Stroke	409	IS patients have shown longer RTL than controls, high-risk stroke populations have shorter RTL than controls	Median TL=1.52vs1.11 (P<0.001) and median TL=1.05vs1.11 (P=0.027), respectively	(Luo *et al.*, 2017)
Heart failure	620/183	TL shorter in patients with CHF compared with controls	Median TL ratio=0.64 (IQR:0.47-0.88) in CHF patients compared with 1.05 (IQR:0.86-1.29) in controls	(van der Harst *et al.*, 2007)
Heart failure	890	Shorter TL associated with increased risk of heart failure	HR=1.79 [95%CI:1.21-2.63]	(van der Harst *et al.*, 2010)

Atrial fibril-lation	14794*	No association found between TL and AF	Standard mean TL difference = -0.11±0.09, P=0.24	(Zhang *et al.*, 2018)
Atrial fibril-lation	379	Subjects with AF had shorter TL compared to non-AF	Mean T/S=0.87±0.29 compared to non-AF mean T/S=0.95±0.32	(Carlquist *et al.*, 2016)
Atrial fibril-lation	184	No significant association between LTL and incident AF	HR=1.01 [95%CI:0.86–1.19]	(Staerk *et al.*, 2017)
Atrial fibril-lation	367	TL associated with incident AF	HR=1.64 [95%CI:1.02-2.66], P=0.043	(Siland *et al.*, 2017)
Abdomi-nal Aortic Aneu-rysm	190/183	Shorter LTL in patients with AAA	Mean TL difference = 189 bp [95%CI:77-301], P=0.005	(Atturu *et al.*, 2010)
Hyper-tension	1,415/1,682*	Leucocyte telomers may be shorter in hypertensive than in normotensive individuals	Standardised mean difference of TL -0.288±0.039	(Tellechea *et al.*, 2017)
Hyper-tension	327	Hypertensive subjects exhibited shorter age-adjusted TL	Hypertensives=5.93±0.042kb, normotensives=6.07±0.040kb, P=0.025	(Demissie *et al.*, 2006)
Hyper-tension	388/379	The median TL ratio was shorter in hypertensive than in healthy normotensive subjects	The median telomere length ratio=0.57 (IQR:0.48-0.72)	(Yang *et al.*, 2009)
Hyper-tension	497	Hypertension risk higher in patients with shorter relative TL	OR=2.45 [95%CI:1.36-4.44]	(Zgheib *et al.*, 2018)
Hyper-tension	206	LTL was significantly shorter in hypertension patients than controls	0.96±0.52 vs 1.19±0.58, P = 0.001	(Cheng *et al.*, 2020)

Table 2.1. Summary of observational association studies of telomere length and risk of cardiovascular diseases. TL = telomere length, LTL = leucocyte telomere length, CVH = cardiovascular health, CHD = coronary artery disease, CAD = Coronary Artery Disease, IHD = ischemic heart disease, MI = Myocardial infarction, IS = Ischemic Stroke, HF = Heart failure, AF = Atrial fibrillation, OR=Odds Ratio, HR=Hazard Ratio, RR=Relative Risk, IQR=Interquartile Range. A star denotes meta-analysed studies. Sample size is given in number of cases (e.g., 206) or in number of cases/controls (e.g., 388/379).

2.3.2. Telomere length and cancer literature

Cancer along with cardiovascular disease is one of the leading causes of mortality worldwide and its incidence is increasing as the population continues to age. The hallmark of cancer is replicative immortality, the ability for cells to divide uncontrollably. Cells that exceed their replicative capacity enter senescence and undergo apoptosis, this ensures that the dysfunctional cells are cleared and do not accumulate further genomic abnormalities. Senescence can be triggered when a telomere loses its length and becomes too short to maintain its structure. An uncapped telomere is recognised as a DNA double-strand break and initiates the DNA damage response. However, the cell may become cancerous if it bypasses replicative senescence and regains the ability to replicate. A majority of human cancers exhibit the activation of telomerase, which enables cancerous cells to regain TL after its loss due to rapid replication (Robinson *et al.*, 2016; Turner *et al.*, 2019). This biological mechanism suggests that in observational studies we may observe cancer association with both short and long TL, which will depend on the cancer stage being studied.

Wentzensen *et al*., 2011, investigated 27 reports on 13 different cancers and performed meta-analyses to test the association between quartiles of TL, measured in blood or buccal cells, and cancer risk. The authors observed shorter TL to be associated with an increased risk of bladder, oesophageal, gastric, head and neck, renal and ovarian cancers. However, heterogeneity between studies was detected and the possibility of reverse causation was noted. A total of 25 studies, both retrospective and prospective, were included into a meta-analysis of TL quartiles and the overall risk of cancer. The authors reported a significant association between short TL and an increased risk of cancer (OR=1.96, [95%CI:1.37-2.81]). Separate meta-analyses were performed in retrospective and prospective studies and the shortest TL quartile was compared to the longest. Short TL was associated with increased risk of cancer in retrospective studies (OR=2.9 [95%CI:1.73-4.8]), but not in prospective studies (OR=1.16 [95%CI:0.87-1.54]) (Wentzensen *et al.*, 2011).

Ma *et al*., 2011, meta-analysed 21 studies of TL and cancer risk. The majority of the studies included were retrospective case-control studies that collected DNA and measured TL in cases after their cancer diagnosis. In total 11,255 cases and 13,101

controls were included in the meta-analysis and the association between relative TL and overall cancer risk was evaluated. The authors reported a significant association between shorter TL and increased risk of cancer (OR=1.35, [95%:CI:1.14-1.60]) and suggested that accelerating TL shortening may be a marker of cancer susceptibility (Ma *et al.*, 2011).

Zhang *et al*., 2015, analysed TL effects on cancer mortality using a meta-analysis of 45 studies. Authors included studies with newly diagnosed cancer patients into their meta-analysis and reported a significant association of shorter TL with increased cancer mortality risk (RR=1.30 [95%CI:1.06-1.59]) and poor cancer progression including cancer reoccurrence and treatment (RR=1.44 [95%CI:1.10-1.88]) (Zhang *et al.*, 2015).

Zhu *et al*., 2016, performed meta-analysis including 23,379 cancer cases and 68,792 controls from 51 independent studies to test the association of TL and cancer risk. TL was not significantly associated with the overall risk of cancer, but shorter TL was associated with an increased risk of gastrointestinal cancer (OR=1.62 [95%CI:1.33-1.97]) and head and neck cancer (OR=1.86 [95%CI:1.23-2.82]). Authors reported that shorter TL was associated with a decreased risk of lung cancer (OR=0.78 [95%CI:0.67-0.91]) in only prospective studies. In such studies TL is measured at the baseline in healthy individuals that may or may not develop cancer during followed-up. This allows to investigate the TL effect on cancer development (Zhu *et al.*, 2016).

Zhang *et al*., 2017, selected only prospective studies for their meta-analysis to investigate the effect of TL on the risk of cancer. They included 28 studies and a meta-analysis showed only a marginal association between longer TL and an increased risk of total cancers (OR=1.09 [95%CI:0.95-1.24]). The association was stronger in lung cancer (OR=1.69 [95%CI:1.25-2.28]). The authors suggested that longer TL, measured in blood, is a potential marker of lung cancer (Zhang *et al.*, 2017).

Barthel *et al*., 2017, used a different approach to investigate TL in tumours. In their study each patient contributed a tumour sample, normal blood sample and solid tissue control sample. TL was measured using TelSeq and compared between the patient's tumour and normal tissue samples. The authors reported shorter observational TL in tumours in comparison to normal tissue (Barthel *et al.*, 2017).

Adam *et al*., 2017, meta-analysed 61 studies with a total of 14,720 cancer patients and reported inconsistent results of TL association with the risk of, and overall survival in,

cancer. TL was not associated with overall survival in cancer (HR=0.88 [95%CI:0.69-1.11]), but longer TL was associated with decreased risk of chronic lymphatic leukaemia (HR=0.45 [95%CI:0.29-0.71]) and urothelial cancer (HR=0.68 [95%CI:0.46-1.00]). The authors noted the limitations of their meta-analysis due to significant heterogeneity between the included studies (Adam *et al.*, 2017).

In a prospective study by Luu *et al*., 2019, LTL was measured in 26,540 participants, 116 of which developed pancreatic cancer during the follow-up period. The authors reported the longest quartile of LTL to be associated with a higher risk of developing pancreatic cancer (HR=2.18 [95%CI:1.25-3.80]).

Kim *et al*., 2015, investigated LTL in relation to cancer while accounting for cancer stage. Their prospective study included 473 patients with early stage non-small-cell lung cancer and LTL was measured using qPCR. Patients that experienced recurrent cancer events during 61 months of follow-up were seen to have significantly longer LTL (1.13 versus 1.07, P=0.046) (Kim *et al.*, 2015).

The reported associations of TL and different cancers are very inconsistent. The majority of larger studies reported an association of shorter TL and a higher risk of cancers (**Table 2.2**) but many other studies report an increased risk with longer TL.

Sample size (case/control)	Results	TL effect	Ref
18,430 samples of 31 cancer types	Telomeres were shorter in tumours than in normal tissues (TL measured using TelSeq)	Relative mean TL shorter in tumours	(Barthel *et al.*, 2017)
11,255/13,101*	Shorter telomeres associated with increased cancer risk	OR=1.35 [95%CI=1.14-1.60]	(Ma *et al.*, 2011)
27 reports on 13 cancers*	Short surrogate tissue TL associated with increased risk of cancer	OR=1.96 [95%CI:1.37-2.81]	(Wentzensen *et al.*, 2011)
23,379/68,792*	Non-significant association between short telomeres and overall risk of cancer	OR=1.10 [95%CI:0.98-1.23]	(Zhu *et al.*, 2016)
11,429 (overall survival) and 4,293 (disease progression)	Short TL associated with increased cancer mortality risk and poor cancer progression	RR=1.30 [95%CI:1.06-1.59] and RR=1.44 [95%CI:1.10-1.88]	(Zhang *et al.*, 2015)
13,894 / 71,672*	Marginal association between longer TL and higher risk of total cancers	OR=1.09 [95%CI:0.95-1.24]	(Zhang *et al.*, 2017)
14,720	No significant association of TL and overall survival in cancer patients	HR=0.88 [95%CI=0.69-1.11]	(Adam *et al.*, 2017)
7,183 patients, 195 cancer events	Longer TL associated with decreased risk of cancer mortality	HR=0.37 [95%CI:0.25-0.54]	(Shen *et al.*, 2020)

Table 2.2. Summary of the largest observational studies of telomere length and risk of cancers. * Meta-analysed studies.

2.4. What previous research of telomere length tells us?

2.4.1. Age-related telomere shortening

Telomeres shorten with advancing age, and ageing is known as a primary risk factor for age-related diseases. Telomere shortening can be accelerated by inflammation and lead to premature senescence that exhausts the regenerative capacity of the tissue and may lead to its failure (Stone *et al.*, 2016; Zhang *et al.*, 2016). The process of ageing is characterised by a decreasing ability of the immune system to cope with the accumulating cell and tissue damage. Chronic inflammation and persistent oxidative stress may lead to tissue dysfunction.

Cellular ageing is an important source of inflammation that is recognised as a potential mechanism that accelerates vascular ageing and increases cardiovascular risk (Fitzpatrick *et al.*, 2007; Gebreab *et al.*, 2017; Chiriacò *et al.*, 2019). Telomere shortening may contribute to the ageing processes as well as play a role of a sensor that is exposed to cardiovascular risk factors (Weischer *et al.*, 2012). The associations between shorter TL and cardiovascular diseases add to the evidence that TL is a contributing factor driving the progression of cardiovascular outcomes. Multiple observation studies have reported the association of shorter TL with degenerative cardiovascular diseases (**Table 2.1**).

A somatic cell's capacity to divide is limited by the length of telomere that serves as a protective mechanism against cancer (Stone *et al.*, 2016; McNally *et al.*, 2019), while stem cells are able to maintain their TL due to the presence of telomerase activity. When the telomere length of at least one chromosome becomes critically short in the somatic cell, DNA damage signalling starts, and the cell's cycle stops. The cell becomes senescent, does not divide any further and is eventually cleared. However, some cells may have accumulated mutations that allow them to bypass the DNA damage checkpoint (Shay, 2016; McNally *et al.*, 2019). They continue to divide and lose telomeric repeats, extremely short telomeres are unable to maintain their structure and protect the chromosome ends, they become unstable and as such more mutations and rearrangements may occur. Such events would normally trigger cell crisis and the cell would undergo apoptosis. However, cell that bypass senescence may stabilise their telomeres by activating telomerase or the ALT pathway, which allows them to divide

indefinitely and become immortal (**Figure 2.4**) (Blasco, 2005; Hiyama, 2009; Teasley *et al.*, 2015; Shay, 2016). Up to 85-90% of cancers have telomerase activated, and 10-15% show telomere elongation through the ALT pathway.

Many early-stage cancers have shorter telomeres than in the surrounding normal tissue due to uncontrolled division and a loss of telomeric repeats. The disruption of a normal telomere structure within cancerous cells may lead to the activation of telomerase that would potentially explain the detection of elongated telomeres in cancers of later stages (Blasco, 2005; Teasley *et al.*, 2015; Thriveni *et al.*, 2019).

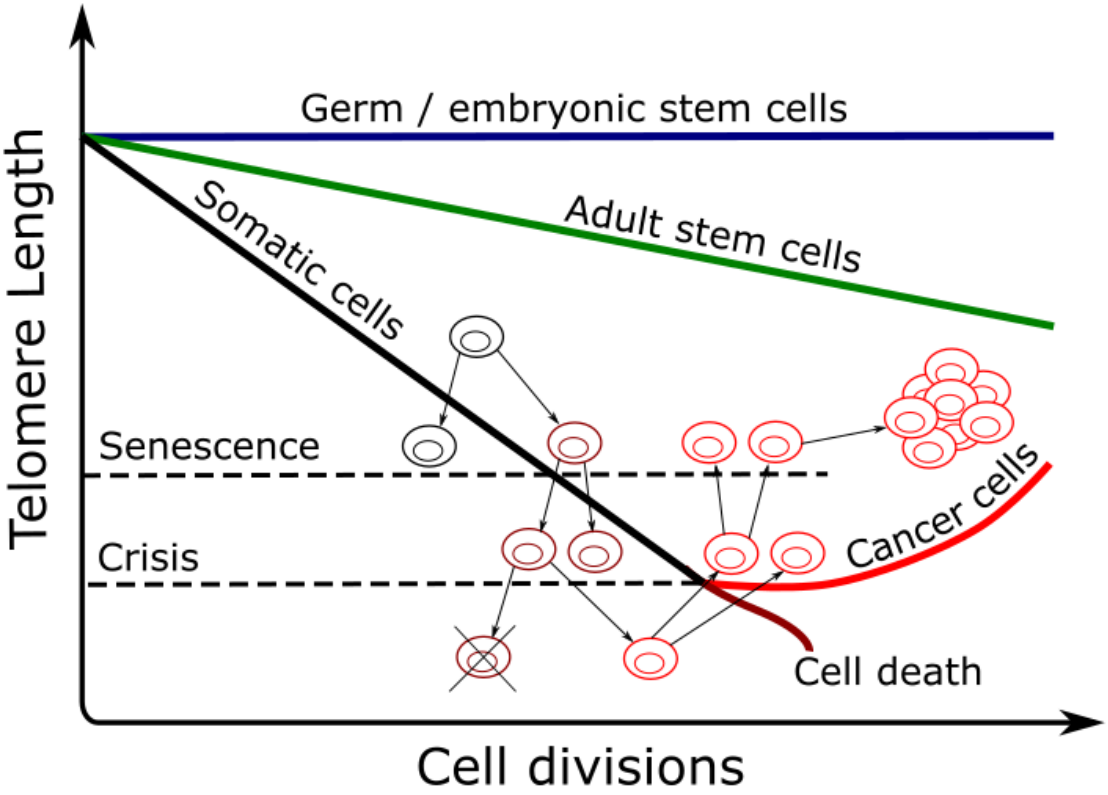


Figure 2.4. An illustration of telomere loss in different cell type and cancer development. Somatic cells have no telomerase activity, and their telomeres shorten with each cell division. Critically short telomeres initiate the DNA damage response and cell senescence. Some cells (in dark red) bypass senescence and the crisis stages and become cancerous. Adult stem cells have low regulated telomerase activity that enables them to maintain telomere length longer than in somatic cells. Germline and embryonic stem cells have high telomerase activity and their telomeres are maintained at the same length, considered to be immortalised (Blasco, 2005; Hiyama, 2009; Teasley *et al.*, 2015; Shay, 2016).

2.4.2. Inconsistent findings of associations between telomere length and age-related diseases

Telomere length is a marker of cellular ageing, and the average organismal TL has been proposed as a marker of biological age. The accelerated loss of TL was thought to increase the risk of age-related diseases, however, reports on the effects of TL on human health remain controversial. Observational studies cannot provide evidence of causality, due to unknown or unmeasured confounders, but only suggest possible links. Shorter observed TL in patients in comparison to controls may be due to reverse causality, where disease risk factors affect TL, causing accelerated TL shortening, or shorter observed TL in patients, rather than changes in TL driving disease risk. For example, increased body mass index (BMI) may cause accelerated telomere shortening and increase cardiovascular disease risk and as such linking TL to CAD causally would be confounded by BMI (Yeh *et al.*, 2016).

The identified literature investigating TL associations was detailed in the chapter above and highlighted several important points that pose difficulties for drawing conclusions about the effect of TL on age-related diseases.

The definition of the disease plays an important role, and while a disease is usually well-defined clinically in a single study, it might have significant differences when compared with other studies, which in turn might introduce imprecision into meta-analysis and increase heterogeneity.

The lack of statistical power due to low numbers of participants or events observed is another problem that may lead to spurious associations. Similar problems may occur due to sparse data at the end of the age distribution. Although many diseases manifest at later stages of human life, studies with a small number of cases suffer from relatively poor coverage of the age distribution in the general population. For example, LTL was a poor predictor of CVD-related survival in individuals older than >75 years due to insufficient data in the study by D'Mello *et al*., 2015 (D'Mello *et al.*, 2015).

A difference in study design may affect the drawing of reliable conclusions about the association. Meta-analyses on measured TL, presented in chapter 2.3. *Association between telomere length and age-related diseases*, combined retrospective, and prospective studies. In retrospective studies TL is assessed after the diagnosis and may

introduce reverse causation, where disease or disease associated risk factors affect TL. In prospective studies TL is assessed prior to diagnosis and potentially shows the effect of TL on the risk of disease development. This implication is especially prominent in studies of TL and cancers that appear to have some dependence upon cancer stages (Wentzensen *et al.*, 2011; Turner *et al.*, 2019).

It is not only the time point of the TL measurement, but the sample type and measurement technique that may also impose some difficulties when performing meta-analysis. TL can be measured in surrogate tissues such as blood leucocytes or buccal cells, or in other cells such as cardiomyocytes. The measurement methods define what is labelled as TL differently and the estimation of effect sizes may vary between studies included in meta-analysis (Ma *et al.*, 2011; Adam *et al.*, 2017). The transformation of risk estimates prior to meta-analysis may help to validate the combination of data and improve the power to detect significant associations (Zhang *et al.*, 2017).

Studies, depending on their initial design of data collection or available data, will often have study specific confounders and variables to consider and as such may use different adjustments for their association analyses and introduce heterogeneity between studies (Adam *et al.*, 2017). Some cofactors are crucial, as they may influence TL directly. For example, cancer studies with no information on administration of chemotherapy or its administration before or after sample collection for TL measurement may lead to misleading conclusions about the relationship between TL and cancer. In sex-specific cancers such as breast cancer or prostate cancer there is a significant influence of hormones, the levels of which should be adjusted for (Ma *et al.*, 2011; Wentzensen *et al.*, 2011).

To summarise, meta-analyses in the field of telomere epidemiology report conflicting evidence regarding TL effects on cardiovascular outcomes, cancers and longevity that is potentially due to heterogeneity between included studies (Hunt *et al.*, 2015). Methodological issues and differences make drawing robust conclusions difficult from the literature (Turner *et al.*, 2019).

Nonetheless, prospective studies on TL and age-related disease continue to emerge, where researchers aim to test the hypothesis that accelerated telomere shortening leads to disease rather than telomere shortening being a result of disease or that there is a common risk factor (Yeh *et al.*, 2016).

While observational studies may suffer from reverse causation, bias and unknown or unmeasured confounding, the association of genetic TL allows us to test for causality and avoid confounding.

2.5. Genetic telomere length as a driver of age-related disease

Once telomere length was associated with cellular ageing, more evidence was gathered in the literature investigating the role of telomere biology in the pathogenesis of age-related disease. Within observational studies on ageing in the general population telomeres were found to play an important role in premature ageing in familial studies. Such investigations showed that genetic mutations result in dysfunctional products of genes that encode proteins involved in telomere structure and regulation, DNA damage response and DNA damage repair. Such mutations cause extremely short TL in individuals with mutations in comparison to healthy individuals of the same age. Premature aging disorders that result due to telomere-related genetic mutations are referred as telomeropathies or telomere syndromes (Armanios *et al.*, 2012; Turner *et al.*, 2019).

The understanding of telomere genomics started to emerge with the identification of mutations in the dyskeratosis congenita 1 (*DKC1*) gene that leads to a rare disorder known as dyskeratosis congenita (Heiss *et al.*, 1998). The role of telomeres in human disease was explored further in studies of rare and common illnesses and many more telomere biology disorders were identified to be caused by rare pathogenic variants in genes encoding proteins required for maintaining telomere structure, replication and repair (Sarek, Marzec, *et al.*, 2015; Savage, 2018).

Familial studies of telomere syndromes allowed for identification of several genetic mutations crucial for telomere function (**Table 2.3**). These mutations have a significant effect on human health. Extremely short telomeres in carriers of these mutations are shown to accelerate telomere attrition and ageing. However, these mutations are very rare in the general population and are unlikely to be detected in cohorts, even in large samples such as 500,000 people as within the UKB due to the recruitment of individuals older than 40 years.

Process or complex	Defective gene	Disease	Reference
Telomerase core components	*TERC*	DC, IPF, aplastic anemia, liver disease	(Tsakiri *et al.*, 2007; Calado *et al.*, 2009)
	TERT	DC, HHS, IPF, aplastic anemia, liver disease	(Armanios *et al.*, 2005, 2007; Tsakiri *et al.*, 2007; Calado *et al.*, 2009)
Telomerase biogenesis	*DKC1*	DC, HHS, IPF, aplastic anemia	(Vulliamy *et al.*, 2001)
	NOP10	DC, IPF, aplastic anemia	(Walne *et al.*, 2007; Vulliamy *et al.*, 2008)
	NHP2	DC, IPF, aplastic anemia	(Vulliamy *et al.*, 2008)
Telomerase assembly and trafficking	*TCAB1*	DC	(Zhong *et al.*, 2011)
Shelterin components	*TINF2 (TIN2)*	DC, HHS and Revesz syndrome	(Savage *et al.*, 2008; Tsangaris *et al.*, 2008; Sasa *et al.*, 2012)
	TPP1	DC, HHS, aplastic anemia	(Guo *et al.*, 2014; Kocak *et al.*, 2014)
Telomeric DNA synthesis (t-loop dissociation)	*RTEL1*	DKC, HHS and IPF	(Ding *et al.*, 2004; Vannier *et al.*, 2012; Ballew, Joseph, *et al.*, 2013; Ballew, Yeager, *et al.*, 2013; Walne *et al.*, 2013; Sarek, Vannier, *et al.*, 2015)
CST complex	*CTC1*	DC, Coats plus syndrome	(Keller *et al.*, 2012; Martínez *et al.*, 2017)
RNA processing of *TERC, DKC1, RTEL1* and *TERF1*	*PARN*	DC, IPF	(Stuart *et al.*, 2015; Burris *et al.*, 2016)
RNA biogenesis	*NAF1*	IPF	(Stanley *et al.*, 2016)
TERT splicing, reduced *TERC* RNA	*LARP7*	Alazami disease	(Holohan *et al.*, 2016)
Overhang processing	*APOLLO (DCLRE1B)*	HHS	(Touzot *et al.*, 2010)
Signals uncapped telomeres and recruits telomerase	*ATM*	Ataxia telangiectasia	(Metcalfe *et al.*, 1996; Smilenov *et al.*, 1997; Wood *et al.*, 2001; Wong *et al.*, 2003)
Telomere replication, prevents telomere fragility and ALT	*BLM*	Bloom syndrome	(Du *et al.*, 2004; Zimmermann *et al.*, 2014)
Proper organization and association of telomeres with nuclear lamins	*LMNA*	Hutchinson– Gilford progeria	(Cao *et al.*, 2011; McCord *et al.*, 2013; Gordon *et al.*, 2014; Chojnowski *et al.*, 2015; Burla *et al.*, 2016; Dorado *et al.*, 2017; van Steensel *et al.*, 2017)
Telomere replication, prevents telomere fragility	*RECQL4*	*RECQL4* disorder / Rothmund-Thomson syndrome	(Kellermayer, 2006; Van Maldergem *et al.*, 2006; Holohan *et al.*, 2014)
Telomere replication, prevents chromatid telomere loss and ALT	*WRN*	Werner syndrome	(Chang *et al.*, 2004; Crabbe *et al.*, 2004, 2007; Du *et al.*, 2004; Opresko *et al.*, 2004; Edwards *et al.*, 2014)

Table 2.3. Telomere-related molecular processes affected by defective genes and resulting telomopathies. DC - dyskeratosis congenita, IPF - idiopathic pulmonary fibrosis, HHS - Hoyeraal–Hreidarsson syndrome.

In the general population telomere length is a polygenic trait, to which multiple genetic variants of modest effect sizes contribute. These genetic variants may be in telomere associated genes or in their regulation regions and are identified for their association with telomere length and may also relate to telomere attrition.

I have shown how genetic data is important for telomere length and how we can consider genetic information as generally free from confounding. The main aim of my study is to use genetically determined telomere length to investigate disease risk. Therefore, to analyse the effects of genetic TL this project required a set of genetic determinants for TL, which can be identified using a genome-wide association studies (GWASs) of telomere length. In the following chapter I am going to detail TL genetic determinants found to date using GWASs and the large-scale GWAS meta-analysis of TL, combining studies from the European Network for Genetic and Genomic Epidemiology (ENGAGE) and the European Prospective Investigation into Cancer and Nutrition (EPIC) consortiums, further referred as ENGAGE study, that was used to obtain most recent set of TL genetic variants.

Chapter 3. Genome-wide association studies of telomere length

A phenotype, a disease, or a trait such as telomere length, may be caused by genetic and environmental factors or by their interaction. The genetic background, although not causal in all instances, is an underlying set of rules that dictate how an organism develops, functions and responds to environmental stimuli (Dorak, 2017). Thus, it is crucial to investigate the genetic underpinnings of the phenotype to understand its biology.

The first genetic determinants of telomere length were detected through family-based studies, investigating affected individuals with telomeropathies (chapter 2.3. *Genetic telomere length as a driver of age-related disease*). These rare disorders are determined by highly penetrant mutations (**Table 2.3**) and exhibit extremely short TL due to their abolished activity of genes such as *TERC* and *TERT* that are required for telomere elongation. Mutations with low penetrance would not have such deleterious and immediate effects on health. However, the joint contribution of genetic variants with low penetrance and small effect sizes can be clinically significant and increase the risk of age-related complex diseases such as coronary artery disease.

This project focused on common genetic determinants of TL and their effects on human health. DNA sequence variation, identified to be associated with TL, may influence the activity of genes related to telomere biology, even if they are located outside the gene region. Common genetic variants, associated with a phenotype, are generally found in intergenic regions, between genes, and have a minimal effect on biological systems. They can affect the function of telomere related genes via regulatory elements, for example, by being a transcription factor binding site, a microRNA or microRNA-binding site, or a distal promoter (Bush *et al.*, 2012; Dorak, 2017). TL, determined by genetic variants, in turn, may affect or be involved in biological processes such as inflammation and lead to accelerated ageing and disease.

The identification of genetic determinants is a process of screening the whole genome and testing each genetic variation for an association with the phenotype. Such an approach is referred as a genome-wide association study.

3.1. Genome-wide association study basics

A genome-wide association study (GWAS) tests millions of SNPs against a phenotype to detect a statistical correlation between the presence of the genetic variant and a phenotype. The genome is screened with no prioritisation for specific genetic regions. Such analyses are free of hypotheses about the underlying biological causes of the disease or trait (Kitsios *et al.*, 2009; Clarke *et al.*, 2011). GWAS became a golden standard for association studies of common complex phenotypes, because it has the following advantages over candidate gene studies:

- No requirement for prior biological understanding of the disease aetiology (hypothesis free).
- Sample size tends to be large and yields better precision in estimation of effect.
- No limitation on gene number – whole genome is screened (Dorak, 2017).

In GWAS every test is performed under the assumption of no association present and that any observed difference is due to chance with a pre-specified level of statistical significance. Deviation from the null hypothesis and perceived statistical significance assumes that the probability of variation due to chance is small and the genetic association with the trait is likely (Balding, 2006; Donaldson *et al.*, 2016). The goal of GWAS is to identify significantly associated genetic variants that potentially have an impact on the disease or trait and can be used to make predictions about personalised risk in order to develop new preventative measures and treatments (Bush *et al.*, 2012).

The most common design of a GWAS involves a sample of unrelated individuals from a population with their genotypes arrayed, with whole genome imputed and a well-defined phenotype. The phenotype can be a continuous measure such as TL or blood pressure reading or binary such as with disease status for cases and healthy controls (Clarke *et al.*, 2011).

The case-control study design is often chosen to detect potentially causal genetic determinants of disease (Balding, 2006; Clarke *et al.*, 2011; Bush *et al.*, 2012; Reed *et al.*, 2015; Donaldson *et al.*, 2016; Dorak, 2017; Evangelou, 2018).

The genetic association is performed by regressing each SNP separately on a phenotype adjusting for individual and environmental factors (Bush *et al.*, 2012; Reed *et al.*, 2015). Different assumptions about the genetic effect may be made, and different models

(**Table 3.1**) selected to estimate the risk that the allele or genotype predisposes to (Lewis, 2002; Cordell *et al.*, 2005; Dorak, 2017). In dominant, over-dominant and recessive models three genotypes (AA, AB and BB) are collapsed into two (Risk and Reference) and encoded as 0 and 1. Dominant models (for the B allele) assume that having at least one B allele (having genotypes AB or BB) increases the risk compared to AA. Recessive models (for the B allele) assume that both copies of B are required to increase the risk (having genotype BB), while individuals with AA and AB are not affected. Additive models (for the B allele) assume that there is a linear increase in risk with each B allele and uses all three genotypes that are encoded as 0, 1, and 2. Co-dominant models (for the B allele) do not assume a linear change in risk, but instead assesses any change in any direction, treating SNP as categorical, and uses all three genotypes, also known as a 2 degrees of freedom model (Cordell *et al.*, 2005; Bush *et al.*, 2012; Donaldson *et al.*, 2016; Dorak, 2017; Evangelou, 2018).

Genetic models		
Model	**Risk**	**Reference**
Dominant	AB+BB	AA
Over-dominant	AB	AA+BB
Recessive	BB	AA+AB
Additive	BB > AB	AA
Co-dominant	BB > AB	AA

Table 3.1. Genetic models and encoding three genotypes with two variables, risk, and reference.

As is standard in statistical modelling, association testing for continuous, or quantitative, traits are performed using linear regression that assumes the residuals of the model fit are approximately normally distributed and that there is a linear relationship between the mean value of the trait and genotype (Balding, 2006). For binary, or dichotomous, traits association tests are generally performed using standard logistic regression. Linear regression is not an appropriate choice for binary outcomes, as it would predict values outside the range of 0 and 1. Logistic regression is limited to the values between 0 and 1 through the use of the logit link, and estimates the change in risk, as a set of odds ratios of being a case given a specific genotype (Balding, 2006; Dorak, 2017).

Using regression models allow for the adjustment of potential confounding variables and for the estimation of effect sizes using the equations shown below.

Linear regression can be written as:

$$\mathbf{y = \beta_0 + \beta_1 X}$$

where β_0 is an intercept, β_1 is the estimated slope, and X is the genotype.

Logistic regression can be written as:

$$\mathbf{logit}(p_i) \sim \beta_0 + \beta_1 X_i \quad \text{or} \quad p = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X)}}$$

Logistic regression is part of the generalised linear models family and uses logit transformation for the binomial distribution, where logit(p_i) = log (p_i / (1- p_i)), p_i is the expected value of the phenotype given the genotype: p_i = E (Y_i | X_i) of the *i*th individual, where Y_i is a phenotype of an individual *i*, and can take only two values (Y_i = 0 for controls, and Y_i = 1 for cases), and X_i is a genotype of an individual *i* at a particular SNP. Because each SNP consists of two alleles, the genotype can be of three states: AA (X_i = 0), AB (X_i = 1), and BB (X_i = 2) (Balding, 2006; Bush *et al.*, 2012). These are often collapsed into two, risk and reference (**Table 3.1**), for interpretability.

In the logistic regression, β_0 shows the movements of the curve to the left and right, and β_1 shows the steepness of the curve (slope). The model tests whether β_1, which is known as the log-odds ratio, differs from zero.

The genetic association is reported and interpreted using the estimated effect sizes that quantifies the strength of the association, while its confidence interval and P-value are used to assess its statistical significance (Dorak, 2017). In logistic regression the effect size, interpreted as the change in risk due to the genetic variation, is measured using an odds ratio (OR):

$$OR = \frac{P(event\ occurs)}{P(event\ does\ not\ occur)} = \frac{P(event\ occurs)}{1 - P(event\ occurs)}$$

If we are not adjusting for any confounding variables the OR can be simplified to a 2x2 frequency table with allele counts in cases and controls (**Table 3.2**).

	A	**B**
Cases	*a*	*b*
Controls	*c*	*d*
The odds that allele occurs in a case	*a/c*	*b/d*

Table 3.2. Allele frequency table for cases and controls under additive genetic model.

We calculate the ratio of two odds, that are calculated for allele carries and non-carriers, as follows:

$$OR = \frac{odds\ that\ A\ allele\ occurs\ in\ a\ case}{odds\ that\ B\ occurs\ in\ a\ case} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a\ d}{b\ c}$$

The OR shows the difference in odds of being a case, for example, for each additional A allele under an additive model of inheritance, where the A allele is the risk or effect allele. No significant association between genotype and disease would produce OR close to zero, while significant association with OR>1 would indicate that the A allele increases the risk of disease, and OR<1 decreases risk. In logistic regression we estimate the log-odds ratio, as this is on a linear scale, and we estimate the OR by exponentiating the β_1 coefficient.

The likeliness of resulting association to be true is assessed by evaluating the significance level that is expressed by a p-value. The p-value is defined as the probability of observing these data or more extreme under the null hypothesis, and interpreted as the probability of an error, when the null hypothesis is true but is rejected, which is also known as the Type I error. The most common p-value cut-off for GWAS is $5x10^{-8}$, which is assumed to be the Bonferroni-corrected genome-wide type I error rate of 0.05 for 1 million independent tests. Bonferroni-correction and False Discovery Rate can be used to account for multiple testing, as GWAS tests millions of SNPs, the p-value cut-off needs to be adjusted to control the Type I error rate due to multiplicity (Balding, 2006; Donaldson *et al.*, 2016; Dorak, 2017). Other common adjustments, when performing GWAS, include accounting for covariates and potential confounders that may bias the association.

3.2. Difficulties and limitations of genome-wide association studies

GWAS has many advantages over candidate gene studies, but it may also have issues in the study design and analysis, when we consider the existing association between genetic variants and the phenotype in a given population. The ideal result from a GWAS is the identification of causal genetic variants. Often the result is the association of non-causal variants that are likely in linkage with a causal variant that lies nearby, or a

spurious association due to population stratification or population admixture (Cordell *et al.*, 2005; Bush *et al.*, 2012; Jurj *et al.*, 2020).

The association of non-causal genetic variants that are correlated with the causal variant occurs due to Linkage Disequilibrium (LD), which means that allelic combinations in haplotypes, blocks of genetic variants that are inherited together, are not independent, and SNPs occur together in one haplotype more often or rarer than expected (Dorak, 2017). If the association is confounded by LD, a replication study might be able to tell whether the association was with a causal or a correlated genetic variant. Additionally, gene mapping, that identifies potentially causal variants in LD and proximity to the detected significant signal, may help to prioritise genes with biologically plausible functions that may explain the phenotype or disease aetiology.

Genetic association studies can also be confounded by population structure or population stratification. Structure in a population may show that subpopulations or ethnicities have different allele frequencies due to non-random mating between individuals. Population stratification issues may occur due to (1) the genetic risk of the phenotype depending on an individual's ethnicity (is only causal in a specific ethnicity), (2) the genetic variant has a different frequency within various ethnic subgroups (shows a difference in disease prevalence between ethnic groups rather than an effect of a genetic variant on disease risk), and (3) cases and controls are poorly matched and have different ancestry compositions in each group. Using Principal Component Analysis (PCA) on the genetic data and adjusting the GWAS for the estimated components helps to remove the confounding effect of population stratification and avoid spurious associations (Balding, 2006; Abegaz *et al.*, 2019). Also, mixed models that account for genetically correlated individuals can be used. For example, BOLT-LMM fits a linear mixed model to binary data that requires a transformation of the estimates to obtain an odds ratio (Loh *et al.*, 2015). REGENIE (Mbatchou *et al.*, 2020) and SAIGE (Zhou *et al.*, 2018) are alternative tools that can run mixed logistic models.

The success of GWAS depends on study specific and phenotype specific factors. For example, a study can be a prospective cohort that will generate a small number of disease cases unless the disease is very common or the study can be a case-control study that purposefully samples a large number of cases and controls to detect a difference. The definition of a case may limit the sample size with more stringent criteria or accept

the risk of misclassification bias while increasing the sample size to maintain statistical power to detect associations. The definition of a healthy control is sometimes cumbersome, as individuals may develop the disease later or have other conditions, or the age distribution may be different between cases and controls (Hattersley *et al.*, 2005; Clarke *et al.*, 2011; Bush *et al.*, 2012).

The reliability of GWAS results, or when to consider the association to be significant, is being debated. The decision lies in selecting an appropriate p-value threshold to account for the number of tests being performed. A P-value of $5x10^{-8}$ is a standard genome-wide significance level that avoids large numbers of false positives. When Bonferroni correction of a p-value is considered too stringent other methods such as the false discovery rate or permutations may be applicable to adjust this threshold (Hattersley *et al.*, 2005).

Small GWA studies are reported to overestimate the true effect sizes of associated genetic variants, which is also referred as winner's curse. This is intuitive in that the ability for a small study to detect a significant association is dependent on the size of the effect. More precise effects can be estimated in larger samples or by correcting estimates using statistical methods (Hattersley *et al.*, 2005; Bush *et al.*, 2012; Bigdeli *et al.*, 2016).

Other usual confounders such as gender, age, and socioeconomic status are less likely to influence the results of genetic association studies, because GWAS is testing genotypes that are inherited at birth and that are assumed to be randomly distributed across individuals in these groups. Nonetheless, these potential confounders may modify the effect of the association. For example, if gender modifies the effect of association, then categories of gender, females and males, will show different strength or direction in their association with the outcome. GWAS statistical tests need to be adjusted for cofactors that influence the trait (Bush *et al.*, 2012; Dorak, 2017).

The issues attributed to study design, methodology and interpretation in early GWAS sometimes made it difficult to obtain robust replication of reported results in independent samples and confirm the associations. Nonetheless, GWAS has proven to be an efficient technique in identifying genetic variants that elucidate the biological mechanism of the phenotype (Hattersley *et al.*, 2005; Clarke *et al.*, 2011).

3.3. Summary of genome-wide association studies of telomere length

The telomere length of an individual from the general population is a polygenic trait with multiple contributing factors with generally small effect sizes, and GWAS is a natural step to identify genetic regions that are associated with TL.

The first GWAS of mean LTL by Mangino *et al*., 2009, identified two variants in the region of gene *VPS34/PIKC3C* on chromosome 18q12.2, rs2162440 ($p=2.6x10^{-6}$), and rs7235755 ($p=5.5x10^{-6}$). Although associations did not reach genome-wide significance level (as $p>5x10^{-8}$), the results were thought to be biologically plausible, as *VPS34/PIKC3C* gene has been reported to control TL variation in yeast. The discovery sample consisted of 1,625 women with replication sample of 1,165 subjects from both genders (Mangino *et al.*, 2009).

The first genetic variants associated with TL to reach genome-wide significance were identified in the GWAS by Codd *et al*., 2010. The discovery sample was larger and consisted of 2,917 individuals with replication sample of 9,492 individuals. LTL was measured using a quantitative PCR-based technique. A locus on chromosome 3q26 that includes *TERC* which encodes the telomerase RNA component, was associated with TL, more specifically each copy of the minor allele (G) of rs12696304, which lies 1.5 kb downstream of *TERC*, was associated with shorter TL. The authors performed additional analyses that included sequencing the coding region of *TERC* together with flanking regions in individuals that were homozygous for the minor allele (n=16) and homozygous for the major allele (n=16). Unfortunately, no variants were identified within the coding sequence of *TERC* to be associated with TL. It was suggested that the identified variant, rs12696304, may possibly mediate the association with TL by affecting *TERC* expression or through one of the other genes in the 3q26 locus (Codd *et al.*, 2010).

The GWAS by Levy *et al*., 2010, merged four observational studies gaining a sample of 3,417 participants that identified and replicated SNPs in the region that contains *OBFC1* gene that encodes the human homolog of a protein involved in the replication and capping of telomeres in yeast (Levy *et al.*, 2010). GWASs by Mirabello *et al*., 2010 (Mirabello *et al.*, 2010), and Gu *et al*., 2011 (Gu *et al.*, 2011), did not identify any genetic variants at the genome-wide significance level. Prescott *et al*., 2013, performed GWAS

in 3,554 individuals, replicated the finding of *TERC* but identified no novel loci associated with TL (Prescott *et al.*, 2011).

The GWAS by Mangino *et al*., 2012, performed a meta-analysis of six independent GWASs that consisted of 9,190 individuals and further validated their results in 2,226 individuals from another four studies. The authors confirmed previous associations with *TERC* and *OBFC1* and identified two novel regions associated with LTL: one on chromosome 17p13.1 that lies near a conserved telomere maintenance complex component 1 (*CTC1*) and the another on chromosome 19p12 with zinc finger protein 676 (ZNF676). *CTC1* is involved in telomere biology and *ZNF676* function is unknown in relation to telomere biology (Mangino *et al.*, 2012).

The first large GWAS meta-analysis of TL by Codd *et al*., 2013, included 37,684 individuals in the discovery sample and 10,739 in the replication sample. The study replicated previous associations between TL and SNPs within *TERC* and *OBFC1*, identified three novel associations between TL and SNPs in the regions of genes *TERT*, *NAF1* and *RTEL1*, that are known to be involved in telomere biology, and reported two additional loci, 19p12 and 2p16.2 with no associated genes that are known to be involved in telomere biology (Codd *et al.*, 2013).

Pooley *et al*., 2013, conducted a meta-analysis of three GWASs with total of 2,240 individuals, and replicated their results in 15,065 healthy individuals, and in 11,024 cases of breast cancer. The authors confirmed the previous findings for three loci (*TERC*, *TERT* and *OBFC1*) at the genome-wide significance level and supported the evidence for association between TL and another three loci (*ACYP2*, *NAF1*, and *RTEL1*) at the nominal significance level. The novel association was detected between TL and rs6772228 at 3p14.1 that lies within intron 4 of the *PXK* gene that encodes for a serine/threonine kinase whose involvement with telomere biology is unknown (Pooley *et al.*, 2013), however, this finding was not replicated in subsequent TL GWASs.

Several other GWASs have been conducted and all TL GWAS published up to 2020 are summarised in **Table 3.3**.

SNP	CHR:BP	Associated Gene	EA	EAF	Beta	P-value	Discovery Set	Replication Set	Reference
rs2162440	18:35214006	*BRUNOL4, PIKC3C*	G	NR	-1.06	3.00E-06	1,625	1,165	(Mangino *et al.*, 2009)
rs12696304	3:170963963	*TERC*	G	0.26	-0.109	3.72E-14	2,917	9,492	(Codd *et al.*, 2010)
rs4452212	2:137015991	*CXCR4*	A	0.65	-0.08	2.00E-06	3,417	4,769	(Levy *et al.*, 2010)
rs2736428	6:31843924	*SLC44A4*	T	0.29	0.08	3.00E-06			
rs1975174	19:22515251	*ZNF676*	T	0.47	0.07	2.00E-06			
rs4387287	10:105677897	*OBFC1*	A	0.08	0.12	2.00E-11			
rs669976	11:64330165	*MEN1*	C	0.105	0.042	0.018	3,646	-	(Mirabello *et al.*, 2010)
rs820152	17:71127683	*RECQL5*	C	0.377	0.029	0.01			
rs13447720	11:93804974	*MRE11A*	G	0.227	0.037	0.012			
rs12549064	8:9479437	*TNKS*	C	0.177	0.041	0.014			
rs6028466	20:38129002	*DHX35*	A	NR	0.192	3.00E-07	459	1,160	(Gu *et al.*, 2011)
rs654128	6:117086378	*KPNA5*	T	NR	0.122	3.00E-06			
rs621559	1:43645411	*WDR65*	A	NR	0.16	2.00E-06			
rs398652	14:56525569	*PELI2*	A	NR	0.12	2.00E-06			
rs12696304	3:169481271	*TERC*	G	0.27	-0.03	2.00E-14	3,554	2,460	(Prescott *et al.*, 2011)
rs4452212	2:137015991	*CXCR4*	A	0.65	-0.08	2.00E-06	9,190	2,226	(Mangino *et al.*, 2012)
rs2736428	6:31843924	*SLC44A4*	T	0.29	0.08	3.00E-06			
rs1975174	19:22515251	*ZNF676*	T	0.47	0.07	2.00E-06			
rs4387287	10:105677897	*OBFC1*	A	0.08	0.12	2.00E-11			
rs11125529	2:54475866	*ACYP2*	C	0.86	-0.056	4.48E-08	37,684	10,739	(Codd *et al.*, 2013)
rs2736100	5:1286516	*TERT*	A	0.51	-0.078	4.38E-19			
rs7675998	4:164007820	*NAF1*	A	0.28	-0.074	4.35E-16			
rs8105767	19:22215441	*ZNF208*	A	0.71	-0.048	1.11E-09			
rs10936599	3:169492101	*TERC*	T	0.25	-0.097	2.54E-31			
rs9420907	10:105676465	*OBFC1*	A	0.87	-0.069	6.90E-11			
rs755017	20:62421622	*RTEL1*	A	0.87	-0.062	6.71E-09			
rs6772228	3:58376019	*PXK*	A	0.05	0.12	4.67E-17	2,240	26,089	(Pooley *et al.*, 2013)
rs10936601	3:169528449	*TERC*	C	0.27	0.00045	4.00E-15			
rs10466239	10:43849827	*FXYD4, RASGEF1A*	T	0.07	4.51d	7.00E-06	4,289	-	(Lee *et al.*, 2013)
rs34596385	6:141926004	*AK097143*	T	0.05	−4.53d	6.00E-06			
rs11787341	8:19102564	*LOC100128993*	A	0.06	4.91d	9.00E-07			
rs10904887	10:17188641	*TRDMT1*	T	0.47	4.61d	4.00E-06			
rs16859140	3:111792594	*TMPRSS7*	C	0.28	4.58d	5.00E-06			
rs73394838	22:30225973	*ASCC2*	G	0.06	4.44d	9.00E-06			
rs4902100	14:62549819	*SYT16*	G	0.28	4.64d	4.00E-06			
rs7680468	4:108304199	*DKK2, PAPSS1*	T	0.03	−5.47d	5.00E-08			
rs17653722	12:52587518	*KRT80*	T	NR	0.122	7.00E-06	2,632	3,917	(Liu *et al.*, 2014)
rs2098713	5:37144574	*C5orf42*	T	0.47	-0.25	3.00E-06	4,013	16,998	(Saxena *et al.*, 2014)
rs74019828	16:58209274	*CSNK2A2*	A	0.16	-0.38	5.00E-08			
rs2535913	14:73415233	*DCAF4*	A	0.306	-0.0493	6.38E-10	9,190	10,832	(Mangino *et al.*, 2015)
rs2297439	20:62289163	*RTEL1*	G	0.25	-0.12	2.82E-07	5,075	37,505	(Delgado *et al.*, 2017)*
rs1483898	14:42805905	*LRFN5*	A	0.236	0.148	7.86E-08	492	322	(Zeiger *et al.*, 2018)**
rs41293836	14:24721327	*TINF2*	C	NR	-0.233	2.47E-42	23,096	37,505	(Dorajoo *et al.*, 2019)*
rs3219104	1:226562621	*PARP1*	A	NR	-0.057	2.38E-18			
rs28365964	8:73920883	*TERF1*	T	NR	-0.27	6.96E-15			
rs227080	11:108247888	*ATM*	G	NR	-0.06	1.87E-10			
rs7776744	7:124599749	*POT1*	G	NR	-0.058	2.51E-10			
rs7095953	10:101274425	*NKX2-3*	C	NR	-0.042	9.59E-09			
rs2967374	16:82209861	*MPHOSPH6*	G	NR	-0.049	1.00E-11			
rs1001761	18:662103	*TYMS*	A	NR	-0.034	1.06E-08			

Table 3.3. Common genetic variants associated with telomere length. EA – effect allele, EAF – effect allele frequency, NR – not reported, d – results of t-test, * Asian and **African American samples.

One of the latest TL GWASs, was performed by Dorajoo *et al*., 2019. The discovery sample consisted of 16,759 individuals of Southern Han Chinese descent with replication sample of 6,337 individuals. These were further meta-analysed with summary data for an additional 37,505 individuals within the ENGAGE consortium of European studies (Codd *et al.*, 2013). In the Southern Han Chinese sample Dorajoo and colleagues identified ten genome-wide significant loci associated with TL, several of which contained candidate genes with biologically plausible functions in telomere biology and DNA repair. Such genes included *TINF2*, *PARP1*, *TERF1*, *ATM* and *POT1*. In the GWAS meta-analysis that combined both Chinese and European samples the authors identified further six loci that suggested the following candidate genes: *MPHOSPH6*, *NKX2-3* and *TYMS*. Most of their identified loci have biologically plausible roles in telomere biology, lying in or near genes that maintain or regulate TL or are involved in DNA repair pathways. The components of the shelterin complex, required for telomere structure, are encoded by *TINF2*, *POT1* and *TERF1*. *PARP1* and *ATM* are involved in DNA damage response, while *TYMS* is involved in the process of DNA replication and repair. The relevance to telomere biology of *MPHOSPH6* and *NKX2-3* has not been fully established yet (Dorajoo *et al.*, 2019).

A TopMED study by Taub *et al*., 2019 and 2020 (Taub *et al.*, 2019, 2020), estimated TL bioinformatically using whole genome sequencing in 75,176 individuals of multiple ethnicities. They used this TL estimate in a GWAS, unlike previous TL GWASs that predominantly used qPCR or Southern Blot derived TL measures. The discovery dataset consisted of 46,458 subjects using 28,718 subjects for replication. The authors identified 22 loci associated with TL at the genome-wide significance level. Twelve previously associated loci were confirmed and included *TERC, TERT, NAF1, RTEL1, OBFC1, DCAF4, ZNF676, ACYP2*, as well as *TERF1, TINF2, POT1* and *ATM* loci. Other known loci with variants near *PARP1, NKX2-3, MPHOSPH6, TYMS,* and *ZNF208* were confirmed with nominal significance. Ten identified loci were novel and included three genes (*TERF2*, *RFWD3*, and *SAMHD1*) with plausible telomere related functions such as telomere maintenance and DNA damage repair (Taub *et al.*, 2019, 2020).

3.4. A new genome-wide association study of telomere length

In our latest GWAS meta-analysis of LTL in 78,592 individuals of European descent from 21 cohorts we identified 49 genomic regions associated with LTL and prioritised genes in 31 of them (Li *et al.*, 2020). The analysis was performed on genotypes, imputed with a 1000 genomes reference panel, and mean LTL, measured using a qPCR-based method, with telomere length Z-standardised to be comparable across studies. Data were analysed in each study using SNPTEST (Marchini *et al.*, 2007) for fitting the regression model with additive effects. Data were then meta-analysed across 21 studies contributing to the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium and across the studies of the European Prospective Investigation into Cancer and Nutrition (EPIC) Cardiovascular Disease (CVD) and InterAct that included nine strata. The meta-analysis found 20 sentinel variants at 17 genomic loci associated with TL at genome-wide significance level. These included six novel loci: *SENP7, MOB1B, CARMIL1, PRRC2A, TERF2,* and *RFWD3*. Several of them were simultaneously reported in the study of Taub *et al*., 2019 (Taub *et al.*, 2019). Details are given in full within our paper, where we confirmed four recently reported loci identified in the Singaporean Chinese sample (*POT1, PARP1, ATM*, and *MPHOSPH6*) (Dorajoo *et al.*, 2019) and seven loci previously identified in European samples (*TERC, NAF1, TERT, STN1(OBFC1), DCAF4, ZNF208*, and *RTEL1*). To gain additional insights into telomere biology via genetic determinants of TL, FDR threshold (p-value≤0.05) that is less stringent than Bonferroni threshold was used on the data, which resulted in 52 variants that were significantly associated with LTL (**Appendix 1** *Genetic determinants of telomere length*). The study performed gene candidate prioritisation and newly identified loci contained genes (**Figure 3.1**) that: 1) have known genes in telomere regulation (*PARP1, POT1, ATM*, and *TERF2*), 2) that are involved in DNA damage repair and identified SNPs were linked to deleterious protein coding changes (*DCAF4* and *SENP7*) or associated with gene expression change (*RFWD3*), 3) that are involved in nucleotide metabolism (*TYMS, SAMHD1*, and *SMUG1*). While genetic variants associated with telomere structure components and DNA damage response and repair were previously reported, this study contributed a new finding, highlighting the nucleotide metabolism as a key pathway in regulating TL.

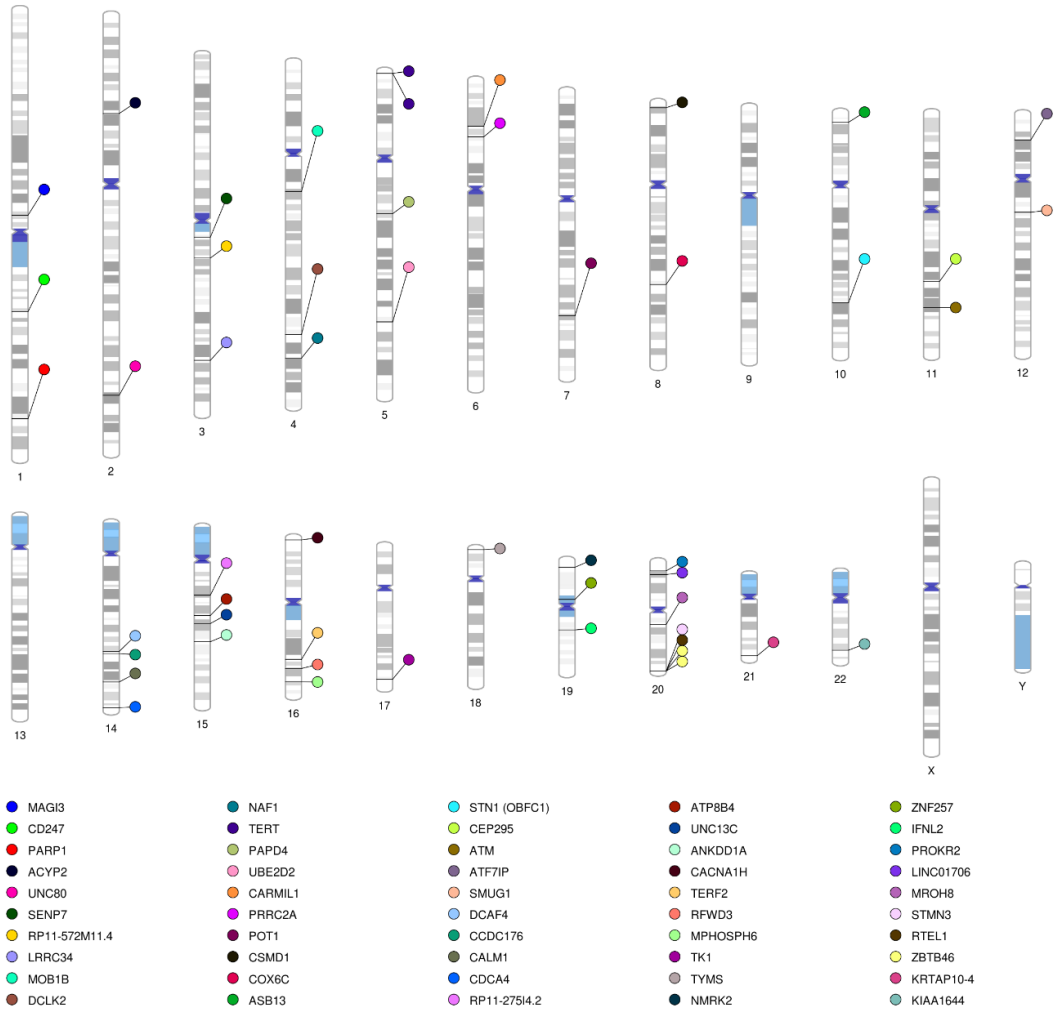


Figure 3.1. Distribution of identified telomere length genetic determinants across the genome and closest candidate genes.

3.5. How much heritability do the identified genetic variants explain?

A variation in the phenotypic trait such as TL can be attributed to both genetic and environmental factors. Heritability estimates how much variation can be attributed to genetic factors. Telomere length is a complex polygenic trait and its heritability was estimated to be between 44% and 80% (Vasa-Nicotera *et al.*, 2005; Njajou *et al.*, 2007; Mangino *et al.*, 2009; Codd *et al.*, 2010).

A single common genetic variant associated with a complex trait such as TL usually has a small effect and explains a very small amount of variance in TL. For example, locus 3q26 was estimated to explain from 0.32% to 1.0% variance in TL (Codd *et al.*, 2010). The study by Taub *et al*., 2019, estimated that 22 identified sentinel variants, associated

with TL, account for ~1.5% of the TL variance (Taub *et al.*, 2019). The 52 genetic determinants of TL identified in the study by Li *et al*., 2020, were estimated to account for ~2.93% of the variance in TL (Li *et al.*, 2020). This indicates that most of TL heritability remains unexplained and future GWASs may detect further variants related to TL.

3.6. Single variant links to health and disease

With the discovery of common genetic variants of TL, researchers have investigated how single variants influence human health. While mutations in telomerase or the shelterin complex have a great impact and result in serious conditions such as telomeropathies, the common variation identified by GWAS accounts for a small proportion of TL variance but may still contribute to the disease.

Directly measured TL has been associated with many age-related diseases in observational studies, suggesting that the genetic determinants of TL may also contribute to the development of adverse health conditions. In the study of Burnett-Hartman *et al*., 2012, SNPs in *OBFC1* (rs4387287 and rs9419958) and *TERC* (rs3772190) were genotyped and tested for association with cardiovascular death in 3,271 Caucasian women. There was a significant association between the minor allele of rs4387287 with CVD death (HR=0.7, 95%CI:[0.5-0.9] for CC vs. AC and HR=0.5, 95%CI:[0.2-1.4] for CC vs. AA genotype) (Burnett-Hartman *et al.*, 2012). Lu *et al*., 2018, investigated SNPs in the *RTEL1* gene in 596 CHD patients, and found that the G allele of rs6010620 and the C allele of rs4809324 in the *RTEL1* gene were associated with a decreased risk of CHD (Lu *et al.*, 2018).

The study by Gu *et al*., 2011, detected the association of rs398652 on 14q21 with a reduced risk of bladder cancer (OR=0.81, 95%CI:[0.67-0.97]) at a nominal significance level, which was consistent with the correlation of the variant with longer telomeres and the association of longer telomeres with reduced risk of bladder cancer. The authors also performed mediation analysis, which showed that TL is a significant mediator between rs398652 and bladder cancer and potentially explains 14% of the effect (Gu *et al.*, 2011).

LTL associated genetic variants were also associated with risk of oesophageal squamous cell carcinoma (Shi *et al.*, 2013), glioma (Walsh *et al.*, 2014), chronic lymphocytic

leukemia (Wysoczanska *et al.*, 2019), and myeloproliferative neoplasms (Giaccherini *et al.*, 2020). However, mentioned studies of single variant should be interpreted with caution due to several limitations.

3.7. Limitations of using single variants

Running a GWAS screens the whole genome, and the identified signals may not be causal genetic variants, but variants highly correlated with a causal one by being in high LD. Detection of the causal variant becomes more complex and requires additional analyses such as gene prioritisation, using bioinformatical techniques, and wet-lab experiments to validate the effect of each variant.

Common genetic variants that are identified using GWAS in the sample from a general population may be significantly associated with the phenotype, but usually have an effect too small because of the complexity of the phenotypic trait, such as TL. The estimated effects of single variants are also likely to be overestimated due to a small sample size, winner's curse, selection bias and unaccounted or unknown confounders or pleiotropy. In the previous chapter I covered several studies that had found quite a substantial effect of single variants on disease, while at the same time it is estimated that even the strongest, based on p-value, common genetic variant associated with TL accounts for less than 1% of TL variance.

The findings of recent TL GWASs with dozens of identified common genetic determinants associated with TL in the general population allow for a more systematic approach to analyse the links between the genetics of TL and human health by combining all TL associated variants into a single polygenic score (Barrett *et al.*, 2015). Having a large number of genetic determinants for TL allows us to build a more powerful genetic score that represents genetically determined telomere length, to explain more variance of TL, and to investigate TL association with age-related diseases with greater confidence and higher accuracy.

Chapter 4. Genetic risk score for shorter telomeres

In this chapter I cover an overview of the use of TL genetic determinants in the literature and in this project. I am going to introduce the concept and approaches to the generation of a genetic risk score (GRS), give details on how TL genetic determinants were used to build the GRS and how it was used to test the hypothesis that genetically determined telomere length is associated with age-related diseases. The results of GRS association analyses are going to be presented, discussed, and compared to previous findings.

4.1. Genetic risk score background

A genetic risk score (GRS), also known as a polygenic risk score (PRS), is a single estimate that aggregates the effects of multiple genetic variants associated with a trait. Most complex traits do not result from the effect of a single genetic variant, but instead are shaped by multiple SNPs with each having a small effect. Overall, a GRS represents a collective genetic predisposition of an individual for a phenotypic trait (Purcell *et al.*, 2009). The genetic information remains the same through life and using GRS to predict complex polygenic traits at the individual level is beneficial for biomedical research and medical practice, as it allows, already at birth, to calculate the risks and manage health via targeted personalised prevention, diagnosis and treatment (Spiliopoulou *et al.*, 2015).

For a single SNP, identified in a GWAS, we find that there is only a small contribution to the amount of variance explained in the trait (r^2). Using a GRS allows us to increase the amount of variance explained in the trait by the genomic information by combining genetic variants together into a single score. A single genetic risk score can be used to test associations with various outcomes such as diseases (Dudbridge, 2013; Goldstein *et al.*, 2014; Smith *et al.*, 2015; Vilhjálmsson *et al.*, 2015; Choi *et al.*, 2018; Evangelou, 2018).

GRS construction requires a selection of suitable SNPs. The selection of these genetic variants generally adhere to a set of criteria where each SNP should be statistically associated with the trait of interest at a pre-specified level of significance, often

genome-wide significance ($p<5x10^{-8}$), not in high linkage disequilibrium (LD) with other variants, and should be biologically relevant to the trait (Dorak, 2017; Evangelou, 2018). I will now describe these stages in more detail.

4.2. Data pre-processing for genetic risk score

4.2.1. Correction for multiple testing

A GWAS is run on millions of variants that produce thousands of results at the end. To determine variants that are truly associated we will most often consider p-values. A p-value is the probability of observing the result (or more extreme) and we test these at a threshold of the Type I error, the probability of rejecting the null hypothesis when it is true. In traditional epidemiology we consider a significance threshold of 5% to be acceptable, that there is a 5% chance that we will reject the null hypothesis even though it is true. It is therefore essential to try and account for the number of tests being performed when interpreting p-values from thousands of results. Usually SNPs that reach genome-wide significance ($p<5x10^{-8}$) are selected to generate the GRS (Smith *et al.*, 2015; Spiliopoulou *et al.*, 2015; Abraham *et al.*, 2016; Khera *et al.*, 2016; Ware *et al.*, 2017). This is based on the Bonferroni correction, that alters the significance level (α) to be more stringent by dividing the p-values by the number of tests, and is the most common method to correct for multiple testing (Chen *et al.*, 2009; Donaldson *et al.*, 2016). Let's consider an example to understand the Bonferroni correction. As stated, a p-value is the probability of observing these data or more extreme, thereby, performing many tests increases the probability of having a significant p-value just by chance. Assume that there is no true underlying association and that we perform 100 tests using the 5% level of significance ($P\leq0.05$). If we run these analyses under the null hypothesis then we would reject the null hypothesis 5% of the time. This is known as a false positive association. In our 100 tests this is likely to produce 5 significant results by chance, where the null hypothesis will be rejected even though it is true. To keep false positives under control one must correct the significance threshold to account for multiple testing. A Bonferroni correction for 100 tests would require a more significant result at $p<0.0005$. For a GWAS where we have LD between variants, we consider there to be

approximately one-million independent tests (where the MAF>1%), hence a threshold at $5x10^{-8}$=0.05/1,000,000. For genome-wide data this way of correction can be overly conservative. It is becoming increasingly common to use different techniques to correct for multiple testing that is less conservative and may include marginally associated SNPs. For example, the False Discovery Rate (FDR), that represents a proportion of false positives among significant results, allows more discoveries while keeping the type I error at a specified level (Benjamini *et al.*, 1995; Goeman *et al.*, 2014; Donaldson *et al.*, 2016).

The false discovery rate is defined as a ratio of false positives divided by the number of significant results (Benjamini *et al.*, 1995). The FDR method finds a cut-off p-value, denoted as a q-value, to keep the rate of false positives at a specific level, for example, at 5%. While there are many forms of the FDR, in principle it is a procedure that is performed via an algorithm that can be described as follows:

- Sort all p-values.
- Starting from the largest p-value we calculate the constraint as *0.05*i/N*, where ***i*** is the order sequence of the result and N is the total number of tests. We then repeat this for all p-values in descending order and test where the observed p-value ≤ constraint.
- We may find that in 100 tests it is the 4^{th} smallest p-value that satisfies the constraint p_4=0.001 ≤ 4*0.05/100=0.002
- We would then assume that p-values ≤ 0.002 are significant, so the FDR would result in 4 significant results.
- The FDR q-value threshold for each test may be calculated as the p-value that satisfies the constraint multiplied by the number of tests divided by the number of significant results. For p_4 in our example – the FDR q-value would be calculated as 0.001*100/4=2.5% where we would expect this number of false positives among our significant results.

4.2.2. Selection of independent genetic variants

The results from a GWAS will likely contain SNPs that are in high LD indicating that they more or less frequently occur together in the DNA sequence and cannot be treated as independent. One SNP with the lowest p-value is usually selected to represent the independent association within a locus, which often covers a DNA region of ~1Mb. This selected SNP may be causal, or it might be in LD with the causal variant(s). Selecting just one representative SNP from the locus allows us to eliminate over-representation of a specific genetic locus and prevent over-weighting in the GRS.

There are several methods to select independent SNPs. Each one of them defines independence criteria differently. Here we will discuss two of them.

The first method utilises Genome-wide Complex Trait Analysis - a conditional and joint (GCTA-COJO) multiple SNP analysis (Yang *et al.*, 2011, 2012) that selects the top independently associated SNPs. GCTA is a tool for genome-wide complex trait analysis that estimates the variance explained by multiple SNPs rather than testing the association of every single SNP to the trait (Yang *et al.*, 2011). GCTA-COJO is a specific method that performs conditional analysis using summary statistics from a GWAS meta-analysis and LD between SNPs. This method initiates a model containing the SNPs that reach the desired level of statistical significance, and then implements the association analysis conditioning on the other SNPs in the model. If the conditional p-value is lower than the cut-off p-value, then the other SNP is added to the model, and process is repeated to identify a third conditionally independent SNP and so on. When no more SNPs can be added to the model, all selected SNPs are fitted jointly in the model to obtain adjusted effect size estimate (Yang *et al.*, 2012).

Another way to remove correlation between SNPs is to perform a clumping procedure, commonly implemented in PLINK (Lewis, 2002; Purcell *et al.*, 2007). Here the process forms clumps out of variants that lie in physical proximity utilising the LD structure. First it selects the SNP with the lowest p-value and forms a clump based on a window with pre-specified size. The method removes all other SNPs that are in LD (r^2) with the top SNP from within the clump utilising a specified threshold. The process then selects the next smallest p-value from any remaining SNPs and completes the process of elimination by LD again, until we reach the end of the SNPs available in the clump. The next clump

is then selected, and the process is repeated. PLINK clumping requires genotype data for LD estimation and p-values of association.

4.2.3. Winner's curse correction

Summary statistics obtained from a GWAS will contain information on the estimated effect size of each SNP. These are expressed as a β that, for a binary trait, represents the risk increase in log-odds, or the estimated effect on a continuous phenotype. Under an additive model the β is the effect size that can be attributed to each copy of the effect allele (Marchini *et al.*, 2010; Goldstein *et al.*, 2014). We know that SNPs do not contribute equally, and each variant will explain a different amount of variation in the trait. For example, in Codd *et al*., 2013, the lead variant in the *TERC* locus is strongly associated with telomere length (β=-0.097, $p=2.54x10^{-31}$) and has twice the effect size of the genome-wide lead variant in the *ZNF208* locus (β=-0.048, $p=1.11x10^{-9}$) (Codd *et al.*, 2013).

A GRS is constructed by weighting each SNP by its effect size, so that SNPs with stronger effects are given more weight (Johnson, 2012; Smith *et al.*, 2015; Khera *et al.*, 2016). A GRS can predict the phenotype in a target dataset if the effects of the SNPs were estimated without error in the GWAS. However, that is rarely the case and effect sizes need to be adjusted, because estimated effect sizes in a single GWAS discovery set tend to suffer from winner's curse, a bias that overestimates the effect size of significant results. Failure to remove uncertainty and SNPs that do not influence the trait may result in a poor GRS and create spurious associations (Choi *et al.*, 2018). In most cases it is enough to replicate the GWAS hits in an independent dataset to confirm the significance of the association and, more importantly for a GRS, obtain an unbiased estimate of the true effect size that does not suffer from selection due to the winner's curse. A larger discovery set could also increase the chance of obtaining more precise estimates (Zöllner *et al.*, 2007; Faye and Bull, 2011; Faye, Sun, *et al.*, 2011; Bigdeli *et al.*, 2016; Grinde *et al.*, 2017; Palmer *et al.*, 2017).

In the absence of a dataset for replication we can adjust for the winner's curse using statistical methods to correct for this bias. Winner's curse generally affects the strongest

associations leading to estimated effect sizes being larger than the true effects. Winner's curse happens due to statistics having extreme values at one end of the distribution (Zöllner *et al.*, 2007; Bigdeli *et al.*, 2016). For example, the distribution of P-values from a GWAS should be uniform under the null hypothesis but have extreme values at the lower end in the presence of significant associations. The top hits from a GWAS sit in this extreme end and estimating effect sizes from extreme statistics may overestimate their magnitude (Bigdeli *et al.*, 2016).

Effect sizes, β, can be corrected for winner's curse using an FDR Inverse Quantile Transformation (FIQT) that was proposed by Bigdeli *et al*. (Bigdeli *et al.*, 2016). This method performs a multiple testing adjustment of P-values using the FDR and transforms the adjusted P-values to Z-scores, obtaining the non-centralities of the distribution. The effect sizes are then shrunk towards zero as will be detailed through application in the analysis later on (chapter 4.6.2. *Adjustment of effect sizes for telomere length associated genetic variants*).

4.3. Building a genetic risk score

A single variant that is common and has a low penetrance is not informative for assessing disease risk. An approach to combine the genetic information of variants across multiple loci that are associated with a phenotype may provide sufficient information necessary to measure the genetic risk of disease. Such an approach is known as a genetic risk score (GRS), which is also referred as a polygenic risk score (PRS), and is a calculated as a weighted sum of risk alleles for any individual (Lewis *et al.*, 2020):

$$GRS_i = \sum_{j=1}^{n} \hat{\beta}_j G_{ij}$$

Where *GRS* is an overall genetic score for shorter TL of individual *i*, *n* is the number of genetic variants to sum, $\hat{\beta}$ is the estimated effect size for j^{th} genetic variant and *G* is the genotype, coded using allele dosage, of the j^{th} genetic variant for the i^{th} individual (Dudbridge, 2013; Machiela *et al.*, 2015).

This is the most widely used method to construct GRS. The selection of SNPs plays an important role, and alternative models of construction were proposed to include whole

genome estimation of the genetic risk (Spiliopoulou *et al.*, 2015). Accuracy of the genetic risk score depends on the underlying polygenic architecture of the trait, and methods, such as LD score regression or Bayesian approaches, can be applied to incorporate the assessment of trait's polygenicity or shared genetic aetiology (Pasaniuc *et al.*, 2016).

4.4. Examples of genetic risk score use in the literature

Assessing the genetic risk is becoming an important part of clinical decision-making, and many studies are focusing on improving clinical assessments, based on non-genetic factors, by adding a genetic component.

For example, Horne *et al*., 2005, investigated the potential of a GRS to detect the risk of coronary artery disease (CAD) diagnosis. The authors selected a single SNP from three genes involved in the cholesterol metabolism pathway, which was linked to CAD, to build a GRS for 3,172 patients that were undergoing coronary angiography. They examined the risk of CAD diagnosis among patients, and showed a significant difference between GRS groups in the association analysis (Horne *et al.*, 2005).

Another study by Goldstein *et al*., 2014, illustrated how to improve the clinical risk score for coronary heart disease (CHD) by adding a GRS that was built from 50 SNPs, identified by the Coronary ARtery DIsease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease (CARDIoGRAMplusC4D) consortium to be associated with CHD (Consortium *et al.*, 2013). The authors calculated their GRS in a sample of 14,792 participants of Atherosclerosis in Communities Study (ARIC), combined it with the clinical risk score, and tested its association with incident CHD within 10 years of follow-up. They reported an improvement of overall risk discrimination when using both clinical and genetic risk scores (Goldstein *et al.*, 2014). In a similar study Ibrahim-Verbaas *et al*., 2014, showed a significant improvement in discrimination of incident stroke when adding a GRS, based on 324 SNPs implicated in stroke, to the clinical Framingham Stroke Risk Score (Ibrahim-Verbaas *et al.*, 2014).

Abraham *et al*., 2016, also showed the benefits of incorporating GRS together with a clinical risk score. In their study the authors generated a GRS based on 49,310 SNPs from CARDIoGRAMplusC4D consortium (Consortium *et al.*, 2013) and tested its predictive

capability on incident CHD in five prospective cohorts. The authors concluded that integration of the GRS with clinical risk score improves prediction (Abraham *et al.*, 2016). The GRS enables optimisation and improvement in discrimination and prediction of incident cases of complex diseases. Using GRS may help to enhance the screening for individuals at risk and manage preventive therapies.

4.5. Examples of telomere length genetic risk score use in literature

The GRS is built from genetic variants that are associated with a phenotype, where the phenotype can be the resulting disease or an intermediate phenotype associated with the disease, considered as a disease risk factor. Telomere length is such an intermediate phenotype. Measured TL has been associated with a number of age-related diseases, and several studies have investigated the potential influence of TL genetic variants on disease risk by combining TL SNPs and their effects to test relationship between a TL GRS and health conditions.

The first association study of a TL genetic risk score was performed by Codd *et al*., 2013. Seven loci associated with TL were identified in the GWAS meta-analysis and used to build a GRS for shorter TL. The GRS was incorporated into a mendelian randomisation approach (see chapter 5 *Mendelian randomisation study of telomere length*). Essentially, summary statistics from the CARDIoGRAM study (Schunkert *et al.*, 2011) that comprised 22,233 patients with CAD and 64,762 controls, were obtained based from a CAD GWAS. The authors reported that GRS for shorter LTL was associated with a 21% (95%CI:5-35%) increase in CAD risk per standard deviation decrease of LTL. The seven selected genetic determinants of TL estimated to account for less than 1% of TL variance, but even at this level it was possible to detect a significant association with age-related cardiovascular disease (Codd *et al.*, 2013).

Other association studies utilised genetic determinants of TL identified by Codd *et al*., 2013, and variants detected by others to investigate the associations between the GRS for TL and their phenotype of interest (**Table 4.1**). Primarily, a TL GRS was tested in association with age-related diseases to confirm previous associations between measured TL and risk of age-related conditions, assuming that the genetic component of TL would potentially elucidate causal links to disease via mendelian randomisation.

This was generally achieved by combining the effects of SNPs into a GRS, which is a sum of alleles, associated with TL in one direction, that are weighted by their effect sizes.

Phenotype	N (SNP)	Sample size	Results	Author, Year
CAD	7	22,233 / 64,762	1 SD decrease in LTL was associated with a 21% (95%CI:5-35%) higher risk of CAD	(Codd *et al.*, 2013)
Melanoma	7	11,108 / 13,933	Association between longer TL and increased melanoma risk	(Iles *et al.*, 2014)
Lung cancer	7	5,457 / 4,493	Longer TL suggested to increase lung cancer risk	(Machiela *et al.*, 2015)
Gastric cancer	8	1,136 / 1,012	U-shaped association between telomere length and gastric cancer risk	(Du *et al.*, 2015)
Non-Hodgkin lymphoma, chronic lymphocytic leukemia and small lymphocytic lymphoma	9	10,102 / 9,562	Longer TL may increase non-Hodgkin lymphoma risk, particularly risk of chronic lymphocytic leukemia or small lymphocytic lymphoma	(Machiela *et al.*, 2016)
Chronic lymphocytic leukemia	8	273 / 5,725	Association between longer LTL and increased CLL risk	(Ojha *et al.*, 2016)
Neuroblastoma, acute lymphoblastic leukemia and osteosarcoma	8	1,516 (NB), 958 (ALL), 660 (OS) / 6,892	Genetically longer LTL is a newly identified risk factor for neuroblastoma	(Walsh *et al.*, 2016)
Breast cancer	6	2,865 / 2,285	GRS for shorter telomeres was significantly associated with a decreased risk of breast cancer	(Luu *et al.*, 2016)
Renal Cell Carcinoma	9	10,784 / 20,406	Genetically longer TL is associated with increased risk of renal cell carcinoma	(Machiela *et al.*, 2017)
Pancreatic cancer	8	1,500 / 1,500	Genetically predicted TL is not associated with pancreatic cancer risk	(Antwi *et al.*, 2017)
Pancreatic cancer	10	2,374 / 4,326	Genetically shorter telomere associated with increased pancreatic cancer risk	(Campa *et al.*, 2018)
Thyroid cancer	9	118 / 5,206	TL GRS is not strongly associated with risk for thyroid subsequent malignant neoplasm in survivors of childhood cancer	(Gramatges *et al.*, 2019)
Depression and anxiety	9	17,693	No association between genetic predisposition to shorter TL and risk of depression and anxiety	(Chang *et al.*, 2018)

Table 4.1. Genetic risk score association studies of telomere length and diseases. Sample size is given in case/control number.

Genetically determined TL, estimated from up to 10 genetic variants identified through the initial GWASs of TL, explains only a small proportion of total TL variance (up to ~1%). Most studies detected promising significant associations between TL and disease, and some pointed out potential problems and complexity of establishing disease cause. For example, Iles *et al*., 2014, reported genetically longer TL to be associated with an increased risk of melanoma (Iles *et al.*, 2014), while Du *et al*., 2015, suggested a U-shaped associated of genetic TL with gastric cancer, where both the shortest and the longest genetic TL increase the risk (Du *et al.*, 2015), and Campa *et al*., 2019, found

genetically shorter TL to be associated with increased risk of pancreatic cancer (Campa *et al.*, 2018). Although a difference in TL direction association can sometimes be attributed to various types of cancer having different aetiology, the inconsistencies may also point to the limitation of a small set of TL genetic determinants being used to provide robust evidence of disease association.

To expand our previous knowledge and gain further insights into the relationship between genetic TL and age-related disease risk, I selected 52 genetic determinants of TL from the latest GWAS meta-analysis (Li *et al.*, 2020), described in chapter 3.4 *A new genome-wide association study of telomere length*, to build a genetic risk score for shorter TL. I test the TL GRS for association with a set of age-related diseases available from UK Biobank. I will describe and justify the approach used to build the TL GRS in more detail in the following chapters.

4.6. Genetic risk score construction and method justification

The GRS construction requires access to genetic information that is represented by genotypes or alleles for each individual in a study along with the estimated effects of genetic variants on TL from another study. This requirement needs pre-processing and data preparation: selection of genetic variants and adjustment of their effect sizes. When SNPs are chosen and their effects are estimated we can construct a GRS for any individual from their genetic information.

4.6.1. Selection of genetic variants for telomere length genetic risk score

In the ENGAGE study (Li *et al.*, 2020), there were in total 4,994 SNPs that reached significance according to FDR adjusted p-values. They were used for the primary selection of SNPs. I found 4,889 SNPs to be available as either genotyped or imputed in UKB and excluded five insertions or deletions. I extracted genotypic data for these 4,889 SNPs by chromosome number and position using *bgenix* (Band *et al.*, 2018) from the UKB genotype imputed dosage files (*bgen*) under project 9922, available to our research group.

Selected SNPs were identified using a p-value significance level of $1.03x10^{-5}$, equivalent to an FDR q-value≤0.05 (Li *et al.*, 2020), and were considered independent using two approaches, GCTA-COJO (Yang *et al.*, 2011, 2012) and PLINK clumping (Purcell *et al.*, 2007) as briefly described in chapter *4.2.2. Selection of independent genetic variants*.

In using GCTA conditional and joint analysis all p-values were adjusted by the number of loci that pass the selected p-value threshold. For this analysis this includes all 4,889 SNPs in the FDR list. A locus is defined as a cluster of SNPs that lie in a window of ~1Mb. When using this method GCTA identified 52 genetic determinants of telomere length (previously shown in **Figure 3.1** in Chapter 3.4. *A new genome-wide association study of telomere length* and listed in **Appendix 1** *Genetic determinants of telomere length*). These 52 SNPs were taken forward to represent GDTL in the GCTA constructed GRS for shorter TL.

I also applied the clumping procedure using PLINK (v1.90b3) to filter out SNPs in high LD where the algorithm keeps only independent signals from each locus (Purcell *et al.*, 2009; Spiliopoulou *et al.*, 2015). The clumping procedure requires a threshold for LD at which SNPs are to be considered independent, and to define this parameter I performed an analysis of testing what threshold is appropriate.

Independence between SNPs is defined by setting an appropriate LD r^2 threshold. By default, this is set to value of 0.5 indicating that SNPs with relatively low LD≤0.5 are treated as independent. A previous investigation of a CAD GRS reported that the selection of r^2 0.7 performed well in discriminating between cases and controls (Abraham *et al.*, 2016). To estimate the optimal choice of r^2 to use for clumping I performed an analysis using locally available data. This included individuals from the BIOlogy Study to TAilored Treatment in Chronic Heart Failure (BIOSTAT-CHF), that included both a Discovery and Replication cohort with 2,262 and 1,473 individuals respectively (Voors *et al*., 2016), the Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) with 1009 individuals (Tobin *et al*., 2008) and the Wellcome Trust Case Control Consortium Coronary Artery Disease (WTCCC-CAD) cohort with 2,909 (Burton *et al*., 2007). These data were used as all studies had both imputed genetic data and TL measurements. A Z-standardised TL was used in each study to allow them to be combined.

The BIOSTAT-CHF is a cohort study of heart failure patients from 11 European countries with the aim of creating a risk score for non-response to therapy using a systems biology approach (Voors *et al.*, 2016). BIOSTAT-CHF has two cohorts, both discovery and replication. GRAPHIC is a study designed to investigate genetic determinants of blood pressure and related cardiovascular traits in 520 nuclear families recruited from the general population (Tobin *et al.*, 2008; Codd *et al.*, 2010). WTCCC-CAD is the cardiovascular disease arm of the large Wellcome Trust Case Control Consortium, a validation study of GWASs for 7 major diseases with a pooled set of controls (Tobin *et al.*, 2008).

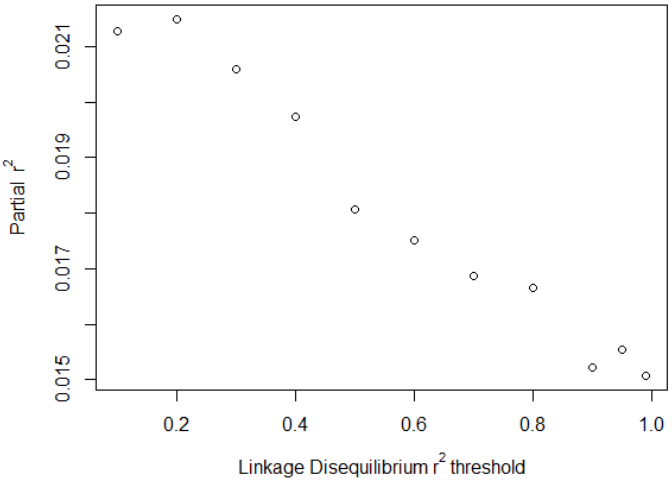
I then used these studies to test several GRSs where each was constructed using a set of SNPs obtained with a different threshold for LD when clumping. It was not possible to run this analysis using summary data, or to run this in UKB as TL data were not available at the time of study.

To find the optimal LD-independence parameter, r^2, I clumped our selected FDR SNP list from ENGAGE by varying r^2, ranging from 0.1 to 0.99, which gave a different number of SNPs for each study depending on their presence within a set of genotyped variants (**Figure 4.1**). I then built a GRS for each SNP set obtained with different thresholds and compared which GRS explained the most variance in TL when fitting a linear regression. This was adjusted for age, sex, and study. To estimate this, I calculated a partial r^2 as the difference in r^2 between the following two models:

$$Partial\, r^2 = r^2(zTL{\sim}GRS + sex + age + study) - r^2(zTL{\sim}age + sex + study)$$

where *zTL* is the Z-standardised TL measurement, and *study* is an indicator for the studies analysed (Biostat Discovery, Biostat Replication, GRAPHIC or WTCCC-CAD). **Figure 4.1** presents the results. The partial r^2, considered to be the explained variance in TL that is explained by the GRS, is plotted on Y-axis. The highest amount of TL variance explained is the highest point on the Y-axis, which corresponds to an LD threshold of r^2=0.2. Note that while this GRS includes over 120 variants the variance explained is still only ~2.1%. However, using r^2<0.2 is optimal for clumping to identify independent SNPs for use in a TL GRS. To provide additional support for this threshold a number of other

studies have also utilised r^2<0.2 to indicate independence (Consortium *et al.*, 2013; Goldstein *et al.*, 2014).



Study	Biostat Discovery	Biostat Replication	GRAPHIC	WTCCC
N(IDs)	2262	1473	1009	2909
r^2	N(SNPs)			
0.1	121	121	125	110
0.2	132	133	138	122
0.3	151	152	157	140
0.4	170	171	176	158
0.5	190	191	197	174
0.6	212	213	218	197
0.7	239	240	245	222
0.8	283	284	289	261
0.9	362	363	272	331
0.95	452	453	464	415
0.99	838	838	855	765

Figure 4.1. Estimated variance explained in telomere length by each constructed genetic risk score. On the X-axis – the r^2 values that represent the LD-independence threshold used for clumping are shown, while the Y-axis shows the estimated partial variance r^2 for TL.

As such, I selected r^2=0.2 as the optimal LD-independence threshold for my analysis and set other parameters to default. Within UK Biobank, utilising this threshold on the 4,889 variants found from the FDR list, the performed clumping procedure with set parameters resulted in 234 independent variants for inclusion in the GRS.

4.6.2. Adjustment of effect sizes for telomere length associated genetic variants

The variants from a GWAS are selected at the extreme of the p-value distribution. This leads to a potentially inflated estimate of the effect size β. This is known as winner's curse. In an effort to correct the effect sizes, β, for winner's curse I used FDR Inverse Quantile Transformation (FIQT), a method proposed by Bigdeli *et al*. (Bigdeli *et al.*, 2016). This method returns corrected Z-scores, but it has not been extended to return corrected standard errors, necessary to transform back to β. For this reason, I applied an approximation to this method to correct the effect sizes of selected SNPs. This shrinks the β coefficients by the level at which the Z-score is estimated to shrink after applying the correction. The original code and my modification can be found in **Appendix 2**

Correction for winner's curse. Analysis was performed in R (version 3.5.1) (R Core Team, 2013) and visualised using '*ggplots2*' (Wickham, 2016).

To investigate the modification for winner's curse it is possible to examine the change in β estimates before and after correction as shown in **Figure 4.2.** Here it is clear that the β estimates are shrunk towards the null and that the shift is not uniform due to being estimated for Z-scores. This approximation to the method proposed by Bigdeli *et al.*, 2016, aims to incorporate the precision of the estimate, where larger effect sizes are likely to be less precise. The distributions show that there are still a number of large effect sizes (red) but the majority are moved toward the null. This would suggest that the selection of many variants is likely to be due to winner's curse and the new β from the FIQT approximation provides a better, less biased, estimate.

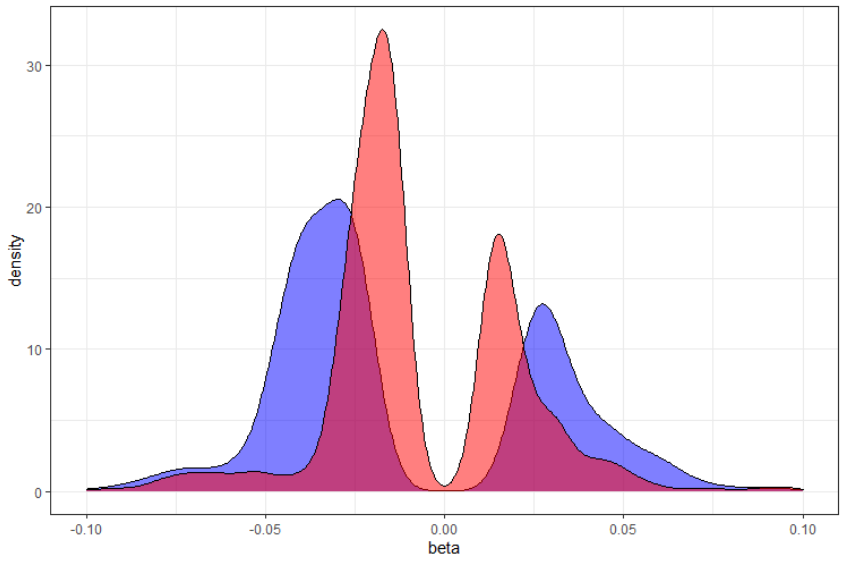


Figure 4.2. Correction of β estimates of nominally significant genetic variants associated with telomere length. Blue density plot shows frequencies of standardised β coefficients before FIQT correction, and red distribution – after the FIQT correction.

4.6.3. Generation of telomere length genetic risk score

I built a GRS for shorter telomeres within UK Biobank using genetic determinants of shorter TL identified in the latest ENGAGE study (Li *et al.*, 2020). The GRS acts as a surrogate of telomere length and can be thought of as the genetically determined telomere length (GDTL) of an individual. As measured TL is correlated with biological age

and is associated with a risk of various traits, I expected a GRS for shorter telomeres to have similar properties.

The generation of a GRS consisted of several steps as detailed in **Figure 4.3.** Summary statistics and their effect sizes were taken from the ENGAGE study and individual genotype and phenotype data from UK Biobank. After QC, described in detail in chapter *1.6.1.2. UK Biobank quality control*, I performed processing that involved:

1) Selection of telomere length associated genetic variants identified in ENGAGE with an FDR q-value≤5%.
2) Correction of effect size estimates, taken from ENGAGE, for winner's curse using an approximation of the FIQT method.
3) Selection of independent SNPs using two approaches, the first approach identified 52 conditionally independent using GCTA-COJO analysis, while the second selected 234 LD independent SNPs when using PLINK clumping.

To generate both GRSs I extracted genotypes for both SNP sets identified in step 3 from UKB genotype dosage files containing ~488k individuals with imputed genetic data. For each SNP I identified the direction of effect using the ENGAGE meta-analysis results. For each SNP the allele associated with a positive β was linked to longer TL, and the allele associated with a negative β – was linked to shorter TL. To aid interpretation of the generated scores and to avoid the problem of cancelling out the effects on TL when summing up the effects across multiple SNPs, I selected the allele that is associated with shorter TL as the effect allele. Note that an additive model was used for ENGAGE so the effect sizes for the alternate allele is estimated by changing the sign, for example, if for a A/G SNP the A allele estimate is 0.45 then I would use the G allele in my GRS with an estimate of -0.45.

The GRS was then calculated for each individual using the dosage of the effect allele, weighted by the estimated effect of the allele on TL (**Figure 4.3**). The dosage of the allele for imputed genetic data is given as a probability of a specific genotype at the genomic position, for example, two alleles A and G give three possible genotypes AA, AG and GG, the dosage of the allele A that is close to 2 shows the genotype is likely to be AA. It should be noted that the allele dosage will be a value between 0 and 2, rather than an integer, to account for the uncertainty in the imputation.

The allele dosage is weighted by SNP effect size when constructing the GRS. For example, if the G allele of a variant was the effect allele for shorter TL in ENGAGE, and in the UKB genotype allele G was the first allele, then the dosage of G is calculated from the probabilities as $2p_0+p_1$, i.e., twice the probability of GG plus the probability of AG as show in **Figure 4.3** for SNP1. If in the UKB imputed data allele G was the second allele, as in **Figure 4.3** for SNP2, then the dosage of G is calculated as p_1+2p_2. The allele dosage then was weighted by the effect size corrected for winner's curse. I then sum the weighted scores across all SNPs in the GRS separately for both the GCTA and clumped SNP sets to obtain two GRSs for shorter TL for each individual. The GRS construction was performed using Python (version 2.7.5) (van Rossum, 1995).

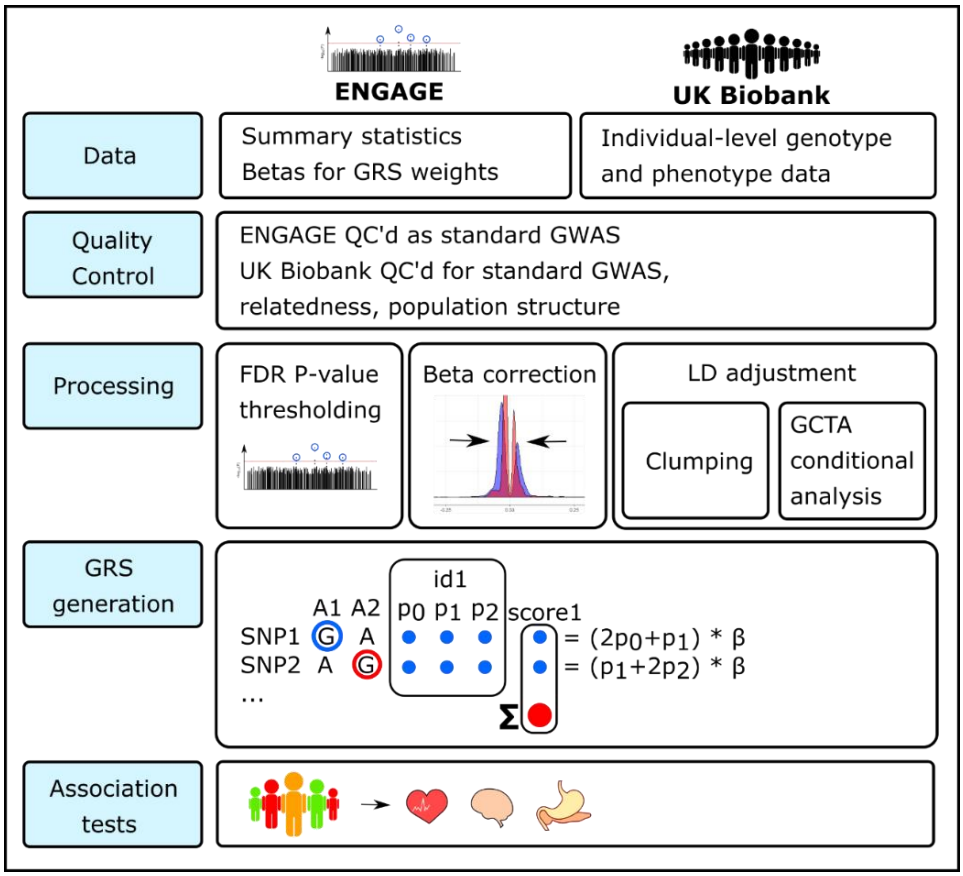


Figure 4.3. Genetic risk score generation workflow. The blue boxes on the left show steps that were taken to perform the association analyses using GRS: 1) Data shows two studies, ENGAGE and UK Biobank, from which SNPs and their associated effects on TL were taken and within which GRS was generated, respectively, 2) Quality Control shows important data quality checks applied to remove poor data, 3) Processing shows SNP selection by FDR q-value, correction for winner's curse and selection of independent SNPs, 4) GRS construction shows two formulas that were used on imputed individual genotype to generate the score, 5) Association tests show association analyses performed to test the relationship between TL GRS and age-related diseases.

4.7. Overview of constructed genetic risk score for shorter telomeres

I generated a GRS for shorter telomere for each individual within UKB. The score represents the individual genetic predisposition to shorter telomeres, which also may be referred as a risk for shorter telomeres. The weighted GRS with 52 SNPs (GRScojo52), has scores ranging from 1.015 to 2.25, while the weighted GRS of 234 SNPs (GRSclump234) has scores ranging from 8.348 to 9.876, where the shift between distributions is explained by the difference in the number of SNPs used. The lower risk value shows a genetic predisposition to longer telomeres and the higher risk value a genetic predisposition to shorter telomeres. When standardised, both GRS follow normal distributions with mean=0 and SD=1 and have low positive correlation (r=0.45) (**Figure 4.4**). The GRSclump234 is skewed slightly to the right, which is reflected in the difference between medians of these distributions, which are 0.0041 for GRScojo52 and 0.034 for GRSclump234.

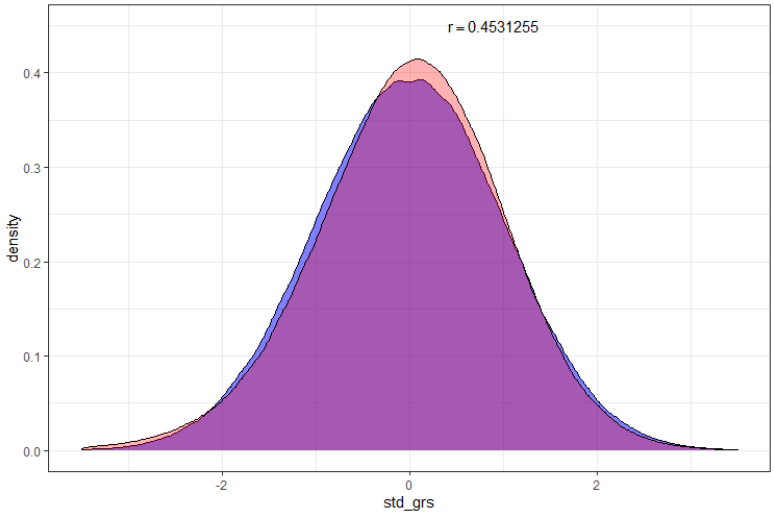


Figure 4.4. Distributions of standardised telomere length genetic risk scores. In blue – distribution of GRS of 52 SNPs, in orange – distribution of GRS of 234 SNPs, in purple – their overlap.

As I constructed GRS with an assumption that sex has no effect on GRS, because only genetic variants on the autosomal chromosomes were used, I expected the GRS distributions to be no different between sexes. There was no sex difference observed in either of distributions of GRScojo52 (P=0.918) nor of GRSclump234 (P=0.991).

The GRS represents a genetic component of TL and, thus, was also expected to be independent of the individual's age. I observed no significant correlation between either

GRS and age (r=0.015) (**Figure 4.5**) as well as no meaningful difference in means between age groups (**Figure 4.6).**

I concluded that the generated GRS for both SNP sets are, as expected, independent of both sex and age, two influential variables in observational studies of TL, and as such these GRS are suitable to use as a proxy for genetically determined TL in association studies.

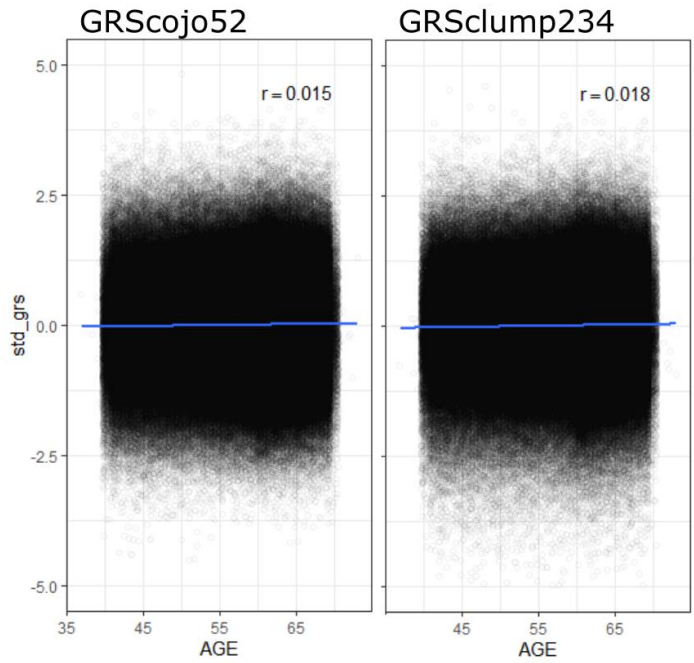


Figure 4.5. No relationship between telomere length genetic risk score and individual age. X-axis shows the age of UKB subjects and Y-axis corresponding generated GRScojo52 (left) and GRSclump234 (right). Blue line shows a fitted regression line, its trajectory through mean=0 of GRS for all ages indicates no relationship between estimated genetic TL and age.

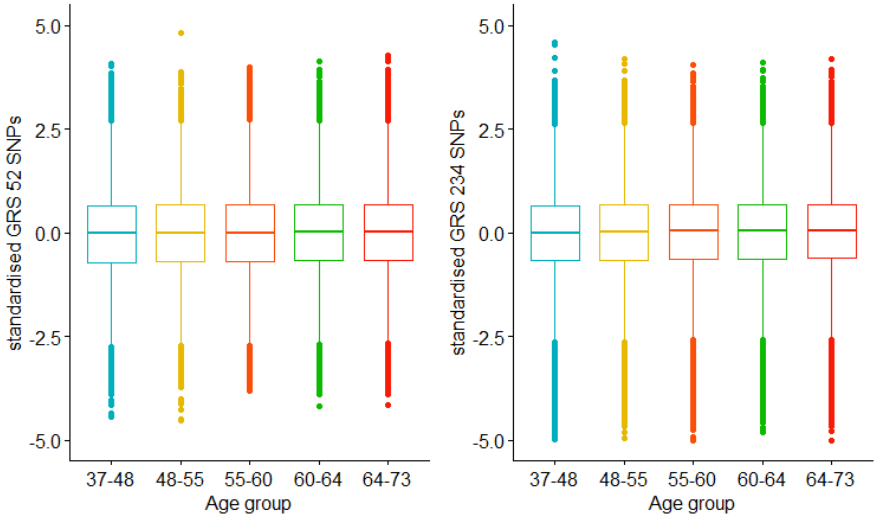


Figure 4.6. No relationship between telomere length genetic risk score and age groups.

4.8. Investigation of effect of genetically determined telomere length on health

4.8.1. Genetic risk score in association models

Association analysis when using a GRS is most commonly performed using standard generalised linear models such as:

Linear regression (Bagley *et al.*, 2001; Purcell *et al.*, 2007, 2009; Sebastiani *et al.*, 2012):

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{GRS}_i + \beta_2 \mathrm{C}_i + \beta_3 \mathrm{D}_i + \varepsilon$$

Logistic regression (Horne *et al.*, 2005):

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathrm{GRS}_i + \beta_2 \mathrm{C}_i + \beta_3 \mathrm{D}_i + \varepsilon$$

Where *i* is an individual, *Y* is a continuous phenotype, *GRS* – genetic risk score, *C* and *D* - covariates, and *ε* is the error term. β_1 shows the effect size of the GRS.

The GRS is usually Z-standardised to allow for easier interpretation when it can be shown to follow a normal distribution. Z-standardisation transforms the data to have a mean of zero and SD of one using the following equation:

$$Z_i = \frac{X_i - E[X]}{\sigma(X)},$$

where X_i is the observed value (GRS) for the i^{th} individual, *E[X]* the mean of *X*, and σ*(X)* is the standard deviation of *X*. This allows us to interpret the effect size estimates for the GRS as a 1 standard deviation (SD) increase in the GRS, in our case a SD reduction in genetically determined telomere length.

In practice a GRS may often be constructed from SNPs shown to be associated with disease to further evaluate the predictive ability of the genetic data for that disease. This is done to improve disease prediction in patient populations with a view to consider it for clinical use (Smith *et al.*, 2015; Abraham *et al.*, 2016). In this project, I construct two GRS using alleles associated with shorter telomeres, which I then use as a surrogate for genetically determined telomere length to investigate its relationship with age-related diseases.

4.8.2. Selecting phenotypes and assigning case-control status

Telomere length has previously been associated with various disease phenotypes ranging from cardiovascular diseases to cancers (chapters *2.3.1. Telomere length and cardiovascular disease literature* and *2.3.2. Telomere length and cancer literature*). In this project one of my aims is to investigate whether genetically determined shorter telomere length, defined in this chapter by a GRS, is associated with the risk of different age-related diseases in UK Biobank. I defined 112 general and 15 gender-specific phenotypes within UKB (**Appendix 3** *Disease definitions*) and assigned them to the following disease groups: cardiovascular diseases, endocrine disorders, mental illnesses, digestive diseases, genito-urinary diseases, musculoskeletal diseases, respiratory diseases, infections, eye problems, immune or inflammatory diseases, and cancers.

The set of diseases was formed depending on number of cases available for each condition within UKB that would provide a sufficient statistical power, a probability to detect true associations (Wassertheil-Smoller *et al.*, 2015). I calculated power using R package *powerMediation*, specifically function *powerLogisticCon()* that calculates power for logistic regression with continuous predictor (Qiu, 2020). The estimation required to specify an available number of cases and controls, significance level (α=0.05), and detectable effect size (10% or OR=1.1). The desired power of at least 80% was reached when using 870 cases. The diseases with number of cases close to 870 or larger were included in the set for further analyses.

I defined a case as an individual with any record of the disease, and a control as the remaining individuals in UKB without a record of the disease. To identify a record of an event and to assign a case-control status for an individual I defined each disease using either self-reported history of the disease, hospital admission for the disease, an operation or surgery for the disease, death due to the disease or death due to the operation used for disease treatment. For example, an individual was assigned as a case with CAD if at least one match for the disease is found in the following:

1) Individual reported history of heart attack or myocardial infarction (UKB variable 20002 with a code value of 1075) or reported that a heart attack was diagnosed by a doctor (UKB variable 6150 with a code value of 1).

2) There is a hospital record of admission due to CAD as the primary or secondary cause with ICD10 codes I21-I25 and ICD9 codes 410-412, 414. ICD10 codes are: I21 - acute myocardial infarction, I22 - subsequent myocardial infarction, I23 - certain current complications following acute myocardial infarction, I24 - other acute ischemic heart diseases, and I25 - chronic ischemic heart disease. ICD9 codes are: 410 - acute myocardial infarction, 411 - other acute and subacute forms of ischemic heart disease, 412 - old myocardial infarction, and 414 - other forms of chronic ischemic heart disease. This data is found in UKB tables *HESIN*, *HESIN_diag10* or *HESIN_diag9*.
3) The death registry holds a record of the individual's death, where the cause of death was attributed to CAD with ICD codes: ICD10 I21-I25 or ICD9 410-412, 414.
4) Hospital records indicated the individual had a record of an operation with following OPCS-4 codes: K40 - saphenous vein graft replacement of coronary artery, K41 - other autograft replacement of coronary artery, K42 - allograft replacement of coronary artery, K43 - prosthetic replacement of coronary artery, K44 - other replacement of coronary artery, K45 - connection of thoracic artery to coronary artery, K46 - other bypass of coronary artery, K49 - transluminal balloon angioplasty of coronary artery, K50.1 - percutaneous transluminal laser coronary angioplasty, K75 - percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery.
5) Individual reported to have had one of the following operations in their self-reported history: coronary angioplasty (PTCA) stent, coronary artery bypass grafts (CABG) or triple heart bypass (UKB variable 20004 with codes 1070, 1095, 1523, respectively).

All individuals that were not selected as a CAD case served as CAD controls and for this reason the number of controls differed between phenotypes. The definition of all diseases included are shown in **Appendix 3** *Disease definitions*, along with the number of cases.

4.8.3. Models of genetically shorter telomeres and the risk of diseases

I used two constructed GRSs to investigate the relationship between genetically determined telomere length and the risk of 112 general and 15 sex-specific disease outcomes. I Z-standardised the GRS to have a mean of zero and standard deviation of one to allow for easier interpretation and comparison of effect sizes between GRS. These were fit using logistic regression, where the GRS was modelled as an independent variable with the case-control status of a disease as the dependent binary variable. As the data are from an observational cohort study it is essential to consider confounding even though the predictor of interest is genetic as there are also genetic specific confounders:

1) LD is the most common confounding factor in a genetic association study. Because of LD multiple associations of SNPs with a trait are observed, where only one or several may or may not be causal (Dorak, 2017). I attempted to remove this confounding before the analysis by selecting independent SNPs using two different pruning procedures.
2) Ethnic background is another confounding factor in genetic association studies. Confounding by ethnicity is part of confounding by population substructure or population stratification. It becomes a problem when individuals from different ethnicities are included in a sample, but the disease is more common in one subpopulation or frequencies of genetic variants are different between subpopulations (Anderson *et al.*, 2011; Dorak, 2017). I adjusted for ethnicity using the first 5 principal components of the genetic data, calculated centrally at the UK Biobank. Principal components (PCs) calculated on the genomic data capture the substructure of a given population, and genetic diversity within a population that we may incorrectly assume to be homogenous (Reed *et al.*, 2015).
3) Sex is a common confounder of disease associations that is directly linked to telomere length. I have included this covariate, because differences in the association with the outcome were observed between sexes.

4) I included chronological age at the date of assessment as a covariate, because it is highly correlated with measured telomere length (Arbeev *et al.*, 2020), and is a strong risk factor for many diseases.
5) I also included BiLEVE status, which shows whether genotyping was performed in the UK Biobank centrally or in the UK BiLEVE project. This will adjust for possible differences between the genotyping arrays.

The final model was:

$$Disease \sim GRS + Sex + Age + PC1 + PC2 + PC3 + PC4 + PC5 + BiLEVE$$

Where *Disease* is a binary outcome (1 – case or 0 – control), *Sex* is a factor (female or male), *Age* is a continuous variable, *PC* is a principal component (ranging from first to fifth), and *BiLEVE* is a type of array used to genotype UKB data.

With this model I tested the association between both of the constructed GRSs for shorter telomeres and disease outcomes. Models were run in R (version 3.5.1) (R Core Team, 2013) and results visualised using R '*circlize*' package (Gu *et al.*, 2014).

4.8.4. Results and Discussion

To assess the association in the results a nominal p-value≤0.05 was considered. A more appropriate Bonferroni-corrected p-value 3.937×10^{-4} was used to consider statistical significance and to correct for multiple testing of 112 general and 15 gender-specific disease outcomes. I estimated the effect size using an odds ratio (OR) that estimates an increase or decrease in risk for a standard deviation increase in the GRS. I repeated analyses with both GRScojo52 and GRSclump234 and obtained significant results consistent between both GRSs (**Figures 4.7** and **4.8, and Table 4.2**).

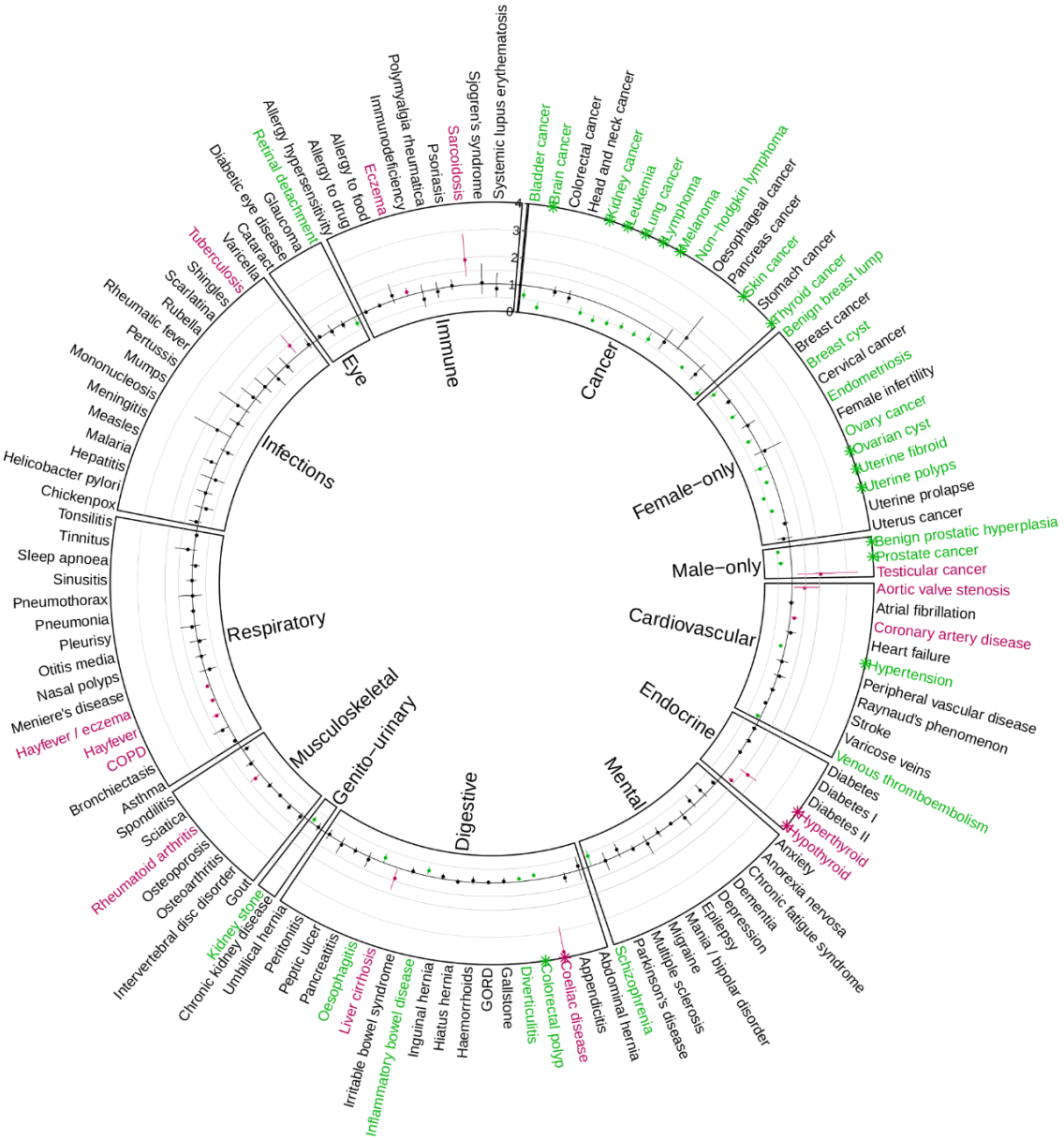


Figure 4.7. Association results of the genetic risk score based on 52 genetic variants with disease outcomes. Nominally significant results (p-value≤0.05) are coloured: red for associations, where the GRScojo52 increases the risk of disease (OR>1), green, where the risk is reduced (OR<1). Bonferroni significant results are marked with a star.

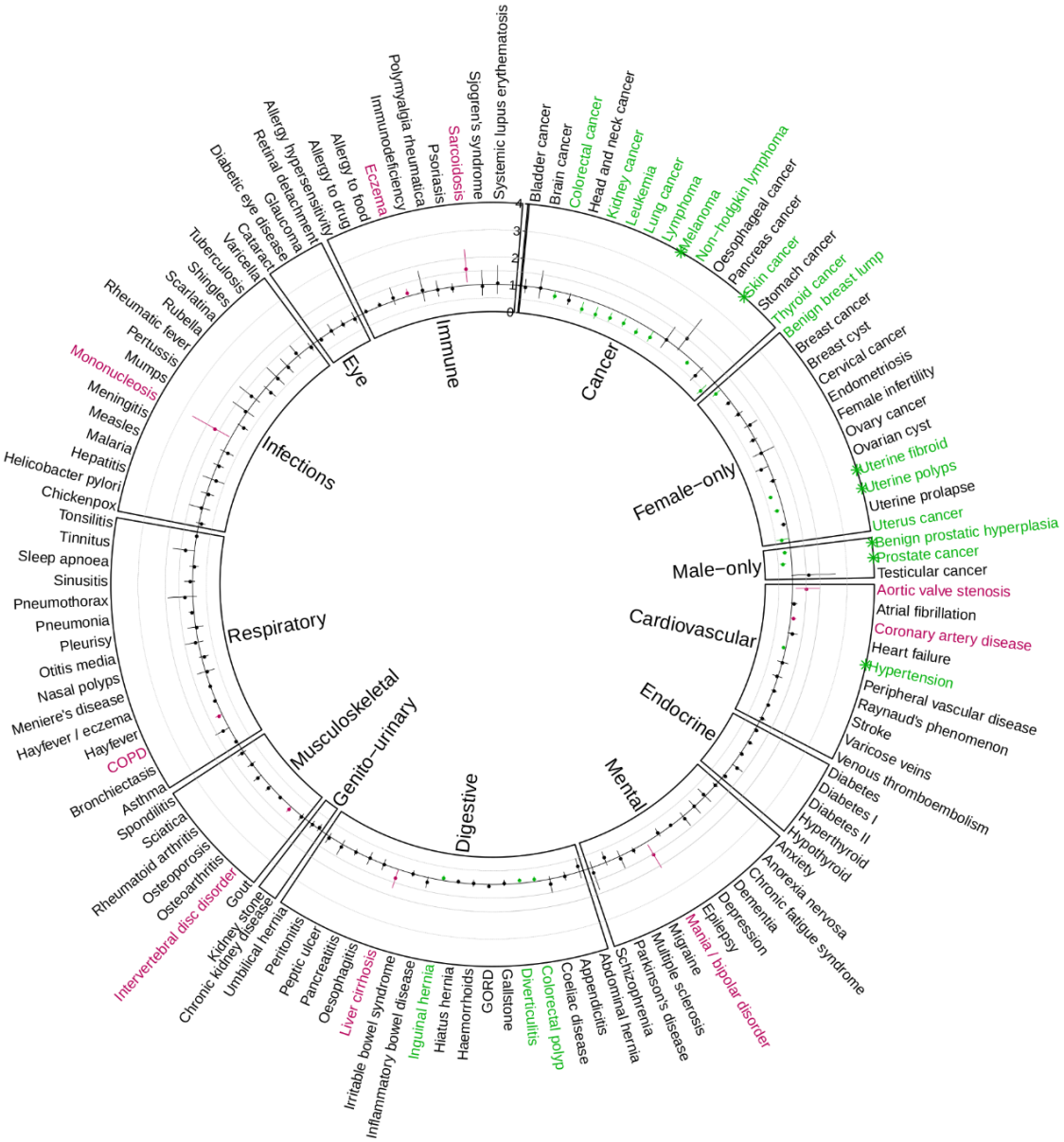


Figure 4.8. Association results of the genetic risk score based on 234 genetic variants with disease outcomes. Nominally significant results (p-value≤0.05) are coloured: red for associations where the GRSclump234 increases the risk of disease (OR>1), green where the risk is reduced (OR<1). Bonferroni significant results are marked with a star.

		GRScojo52		GRSclump234	
Disease phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Uterine fibroid	20556	1.49E-46	0.898(0.885,0.911)	3.03E-14	0.944(0.931,0.958)
Skin cancer	23320	1.61E-25	0.930(0.917,0.943)	4.10E-05	0.972(0.959,0.985)
Benign prostatic hyperplasia	19765	3.98E-25	0.922(0.908,0.936)	2.99E-06	0.964(0.949,0.979)
Coeliac disease	3005	1.16E-20	1.192(1.149,1.236)	2.37E-01	1.023(0.985,1.062)
Uterine polyps	14196	9.37E-15	0.933(0.917,0.950)	1.39E-05	0.962(0.945,0.979)
Hypertension	293487	1.47E-14	0.975(0.969,0.981)	7.89E-05	0.987(0.980,0.993)
Hypothyroid	29377	1.23E-10	1.041(1.029,1.054)	3.32E-01	1.006(0.994,1.019)
Prostate cancer	8967	2.04E-09	0.935(0.914,0.956)	1.96E-07	0.943(0.923,0.964)
Lung cancer	3298	2.50E-09	0.898(0.866,0.930)	2.00E-03	0.945(0.912,0.980)
Melanoma	5362	4.76E-09	0.921(0.895,0.946)	5.14E-06	0.938(0.912,0.964)
Leukemia	1624	1.30E-07	0.873(0.831,0.918)	5.41E-03	0.931(0.885,0.979)
Brain cancer	922	1.30E-07	0.836(0.782,0.894)	7.99E-01	0.991(0.927,1.060)
Hyperthyroid	5565	2.77E-07	1.074(1.045,1.103)	6.87E-02	1.026(0.998,1.054)
Colorectal polyp	34018	4.76E-07	0.971(0.960,0.982)	3.70E-03	0.983(0.972,0.994)
Ovarian cyst	14128	1.34E-06	0.958(0.941,0.975)	4.29E-01	0.993(0.976,1.011)
Lymphoma	3948	1.37E-06	0.923(0.894,0.954)	2.98E-03	0.952(0.922,0.983)
Thyroid cancer	756	4.45E-06	0.842(0.782,0.906)	9.43E-04	0.884(0.821,0.951)
Kidney cancer	1732	4.76E-05	0.904(0.861,0.949)	5.90E-03	0.934(0.890,0.981)
Hay fever / eczema	124509	5.33E-04	1.012(1.005,1.019)	7.77E-01	1.001(0.994,1.008)
Breast cyst	7370	6.41E-04	0.959(0.936,0.982)	6.55E-01	0.994(0.971,1.019)
Sarcoidosis	1536	9.43E-04	1.091(1.036,1.149)	1.02E-02	1.071(1.016,1.129)
Coronary artery disease	36974	9.95E-04	1.019(1.008,1.031)	1.92E-02	1.014(1.002,1.025)
Endometriosis	7753	1.02E-03	0.962(0.940,0.984)	3.09E-01	0.988(0.965,1.011)
Benign breast lump	4126	1.16E-03	0.948(0.919,0.979)	2.26E-02	0.963(0.933,0.995)
Diverticulitis	39848	1.22E-03	0.983(0.972,0.993)	7.95E-04	0.982(0.971,0.992)
COPD	19071	1.42E-03	1.025(1.010,1.041)	2.30E-02	1.018(1.002,1.034)
Bladder cancer	3017	3.71E-03	0.947(0.913,0.982)	8.06E-01	0.995(0.959,1.033)
Non-Hodgkin lymphoma	2641	4.89E-03	0.945(0.908,0.983)	3.51E-02	0.958(0.921,0.997)
Testicular cancer	885	6.75E-03	1.099(1.026,1.177)	5.88E-02	1.069(0.998,1.146)
Hay fever	49053	7.18E-03	1.013(1.004,1.023)	2.65E-01	1.006(0.996,1.015)
Tuberculosis	2952	7.21E-03	1.052(1.014,1.092)	3.28E-01	1.019(0.981,1.058)
Rheumatoid arthritis	8907	1.03E-02	1.029(1.007,1.051)	6.71E-02	1.021(0.999,1.043)
Aortic valve stenosis	2603	1.28E-02	1.052(1.011,1.095)	4.78E-03	1.060(1.018,1.104)
Inflammatory bowel disease	6761	1.45E-02	0.970(0.946,0.994)	9.75E-02	1.021(0.996,1.047)
Retinal detachment	5047	1.63E-02	0.966(0.938,0.994)	9.86E-01	1.000(0.971,1.029)
Liver cirrhosis	3076	2.03E-02	1.044(1.007,1.083)	3.07E-02	1.042(1.004,1.081)
Kidney stone	10299	2.07E-02	0.977(0.957,0.996)	8.71E-01	0.998(0.978,1.019)
Eczema	17142	2.14E-02	1.019(1.003,1.035)	1.60E-02	1.020(1.004,1.036)
Oesophagitis	4895	2.71E-02	0.968(0.940,0.996)	9.54E-01	0.999(0.970,1.029)
Venous thromboembolism	19170	3.50E-02	0.984(0.969,0.999)	5.05E-01	1.005(0.990,1.020)
Ovary cancer	1689	4.69E-02	0.951(0.905,0.999)	5.16E-02	0.952(0.906,1.000)
Inguinal hernia	21600	4.76E-02	0.986(0.971,1.000)	2.64E-02	0.984(0.970,0.998)

Table 4.2. Comparison of significant association results between two constructed genetic risk scores for shorter telomeres. The table is sorted by GRScojo52 P-value. Bonferroni significant results are highlighted with green, nominally significant results are given in black, and non-significant in light grey.

The GRScojo52 was associated with 42 disease phenotypes, 18 of these associations were Bonferroni significant. The GRSclump234 was associated with 27 disease phenotypes, 7 of these associations were Bonferroni significant. The two GRS results are similar especially with disease phenotypes such as cancers and diseases with high proliferative potential such as uterine fibroid and benign prostatic hyperplasia (**Table 4.2**). The main difference between two GRS results is that GRScojo52 has a greater number of significant associations, which also corresponds to lower p-values. The observed difference is due to the selection of SNPs that are included into the GRS, and while I performed additional investigation into the selection of the optimal LD cut-off for the clumping procedure in order to select independent SNPs, the SNP selection may have been affected by two factors:

1) Different lead SNPs from one generic region were selected for GRScojo52 and GRSclump234 with varying effects on TL.
2) GRSclump234 included more SNPs with potentially pleiotropic effects, more noise could have been added.

For this reason, I will mainly report GRScojo52 results. These 52 variants are the ones that we used for publication as our FDR list (Li *et al.*, 2020).

The lower values for the GRS show a predisposition to genetically longer TL, and higher values a predisposition to genetically shorter TL. The effect of GRS for shorter telomeres on the risk of disease was interpreted with Odds Ratios (OR) and interpreted as a 1 standard deviation (SD) increase in the GRS. For example, the GRS for shorter telomeres showed an increased risk of CAD (OR=1.02, 95% CI: 1.01-1.03), meaning that 1 SD increase in the GRS, increased the CAD risk by 2%. All results for 127 diseases are given in **Figures 4.7** above and in detail within **Appendix 4** *Genetic risk score association study results*. In the next chapters I am going to focus on specific results to place the findings in the context of the disease.

4.8.4.1. Shorter telomeres and cardiovascular diseases

Shorter telomeres in long-living mammals were suggested to be an evolutionary protective mechanism against cancer that uses telomere-driven replicative senescence to prevent cells from accumulating de novo somatic mutations when a critically short

telomere length is reached after a certain amount of cell divisions (Stone *et al.*, 2016; Haycock *et al.*, 2017). However, shorter telomeres limit regenerative capacity which may lead to tissue dysfunction. Cardiovascular diseases, such as CAD, is one such example, where shorter telomeres are associated with an increased risk of the disease.

In somatic cells telomeres shorten with every cell division and shorter telomeres reach a critically short length sooner. Critically short telomere length is recognised within the cell as DNA damage and the cell cycle is arrested. The damaged cell is fixed, or the cell becomes senescent and is removed and replaced by a new one. With no appropriate repair the DNA damage signalling may persist and the cell may begin to promote inflammation (O'Donovan *et al.*, 2011; Jose *et al.*, 2017; Anderson *et al.*, 2019). This will in turn promote clearance of senescent and dysfunctional cells, the process that is necessary to minimise an inflammatory load and keep the tissue functional.

Both shorter TL and chronic inflammation are associated with the promotion of biological ageing and an increased risk of cardiovascular outcomes (O'Donovan *et al.*, 2011). Shorter telomeres have been observed to promote inflammation with senescence, and inflammation and oxidative stress reported to promote telomere shortening (Prasad *et al.*, 2017). Moreover, senescent cells experience not only cell-cycle arrest but many other changes in gene expression, metabolism and secretome. The profile of their secretome, that includes molecules with which they communicate with the immune system and neighbouring cells, is known as the Senescence-Associated Secretory Phenotype (SASP). SASP normally signals and promotes the regeneration of tissue, but when SASP persists and becomes chronic, it may induce senescence in neighbouring young cells. For example, a senescent-like phenotype of rarely dividing cardiomyocytes, heart muscle cells, has been reported to lead to cardiac ageing, independent of cell division and telomere length, but accompanied by persistent DNA damage at telomere regions and mitochondrial dysfunction (Victorelli *et al.*, 2017; Anderson *et al.*, 2019; Wang *et al.*, 2019).

The association studies using measured TL, thus, have a challenge in being able to determine the true causal direction. This is especially true when taking into account that measured TL has vast variation at birth and in adulthood, is highly inherited (44-86%) (Njajou *et al.*, 2007; Broer *et al.*, 2013), and is also influenced by environmental factors (Stone *et al.*, 2016; Patel *et al.*, 2017).

Genetically determined telomere length, although free of reverse causation, is an even more complex trait. Genetic determinants, identified at the time of this project, explain less than 3% of the variance in measured telomere length. Nonetheless, even at this level, studies that have used a telomere length GRS were able to detect significant associations with cardiovascular outcomes.

In the GRS association results within this project a nominally significant association of GDTL was observed with an increased risk of CAD (OR=1.02, 95% CI: 1.01-1.03) and aortic valve stenosis (OR=1.05, 95% CI: 1.01-1.09), where a 1 SD increase in the GRS corresponded to a 2% and 5% increase in risk of CAD and aortic valve stenosis, respectively. This result suggests that genetic predisposition to shorter telomere length may contribute to accelerated ageing that leads to age-related disease like cardiovascular disease.

4.8.4.2. Shorter telomeres and hypertension

In this investigation a statistically significant association between the GRS and hypertension was found. This suggested that shorter telomere length is associated with a reduced risk of hypertension (OR=0.97, 95% CI: 0.97-0.98). This is inconsistent with previous studies. I investigated this phenotype in more detail for this reason.

The definition of the hypertension is a crucial step in study design, and hypertension in UKB was defined as:

1) Use of blood pressure lowering medications.
2) Hypertension diagnosis by a medical professional.
3) High systolic blood pressure (>140 mm Hg) and/or high diastolic blood pressure (>90 mm Hg).

One match to either of these three conditions would define a case, and case status was assigned to the majority, ~62%, of UKB individuals. As would be expected with such high prevalence, hypertension overlaps with many other diseases, almost every individual with hypertension had at least one other disease (90.8%). I observed my GRS for shorter telomeres to be significantly associated with stronger effects with cancer status, rather than with cardiovascular outcomes (**Table 4.3**).

		GRScojo52		GRSclump234	
Disease phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Hypertension	293487	1.47E-14	0.975(0.969,0.981)	7.89E-05	0.987(0.980,0.993)
Coronary artery disease	36974	9.95E-04	1.019(1.008,1.031)	1.92E-02	1.014(1.002,1.025)
Aortic valve stenosis	2603	1.28E-02	1.052(1.011,1.095)	4.78E-03	1.060(1.018,1.104)
Venous thromboembolism	19170	3.50E-02	0.984(0.969,0.999)	5.05E-01	1.005(0.990,1.020)
Skin cancer	23320	1.61E-25	0.930(0.917,0.943)	4.10E-05	0.972(0.959,0.985)
Prostate cancer	8967	2.04E-09	0.935(0.914,0.956)	1.96E-07	0.943(0.923,0.964)
Lung cancer	3298	2.50E-09	0.898(0.866,0.930)	2.00E-03	0.945(0.912,0.980)
Melanoma	5362	4.76E-09	0.921(0.895,0.946)	5.14E-06	0.938(0.912,0.964)
Leukemia	1624	1.30E-07	0.873(0.831,0.918)	5.41E-03	0.931(0.885,0.979)
Brain cancer	922	1.30E-07	0.836(0.782,0.894)	7.99E-01	0.991(0.927,1.060)
Lymphoma	3948	1.37E-06	0.923(0.894,0.954)	2.98E-03	0.952(0.922,0.983)
Thyroid cancer	756	4.45E-06	0.842(0.782,0.906)	9.43E-04	0.884(0.821,0.951)
Kidney cancer	1732	4.76E-05	0.904(0.861,0.949)	5.90E-03	0.934(0.890,0.981)
Bladder cancer	3017	3.71E-03	0.947(0.913,0.982)	8.06E-01	0.995(0.959,1.033)
Non-hodgkin lymphoma	2641	4.89E-03	0.945(0.908,0.983)	3.51E-02	0.958(0.921,0.997)
Testicular cancer	885	6.75E-03	1.099(1.026,1.177)	5.88E-02	1.069(0.998,1.146)
Ovary cancer	1689	4.69E-02	0.951(0.905,0.999)	5.16E-02	0.952(0.906,1.000)

Table 4.3. Nominally significant association results of two constructed genetic risk scores for shorter telomeres with cardiovascular and cancer phenotypes.

Therefore, it was considered that there is a possibility that the association in the observed direction might have happened due to an overlap of hypertension cases with cancer cases. I divided UKB subjects into hypertensive and non-hypertensive and estimated the prevalence of cases that had at least one cardiovascular disease (except hypertension) and at least one cancer (**Table 4.4**). There was no significant difference in proportions of cardiovascular and cancer cases between hypertensive and non-hypertensive individuals. This indicated a small chance of influence of cancer cases on the determined direction of effect for hypertension.

Disease phenotype	**Hypertensive**	**Non-hypertensive**
Cardiovascular	27.82%	15.56%
Cancer	11.39%	7.67%

Table 4.4. Cardiovascular disease and cancer prevalence in hypertensive and in non-hypertensive individuals within UK Biobank.

To investigate the association of each GRS with hypertension alone I reassigned case-control status in a new phenotype. A case was an individual that had only hypertension, and none of the other 126 diseases in the analysis, whereas a control was an individual that had none of the 127 diseases in the investigation. It should be mentioned that any

individual could have an additional disease or condition that was not included in the list of the 127 diseases investigated here. After reassigning cases and controls, I excluded all individuals with another disease. I repeated the analysis using 26,081 cases of hypertension and 22,626 healthy controls and found that there was still association between GRScojo52 for shorter telomeres and a decreased risk hypertension persisted (p-value=3.3×10^{-4}).

Whilst the probable lack of power cannot be ruled out in this new analysis, it is clear that consideration is needed in disease-wide scans. The hypertension phenotype highlights several problems when investigating the risk of common diseases with high prevalence: 1) it might not be well-defined, 2) it might be in the presence of other more severe phenotypes, and 3) a healthy control may develop the trait in the future. The non-hypertensive group is much younger as can be seen in **Figure 4.9**, where the distributions of ages in hypertensive and non-hypertensive individuals with no other disease are shown.

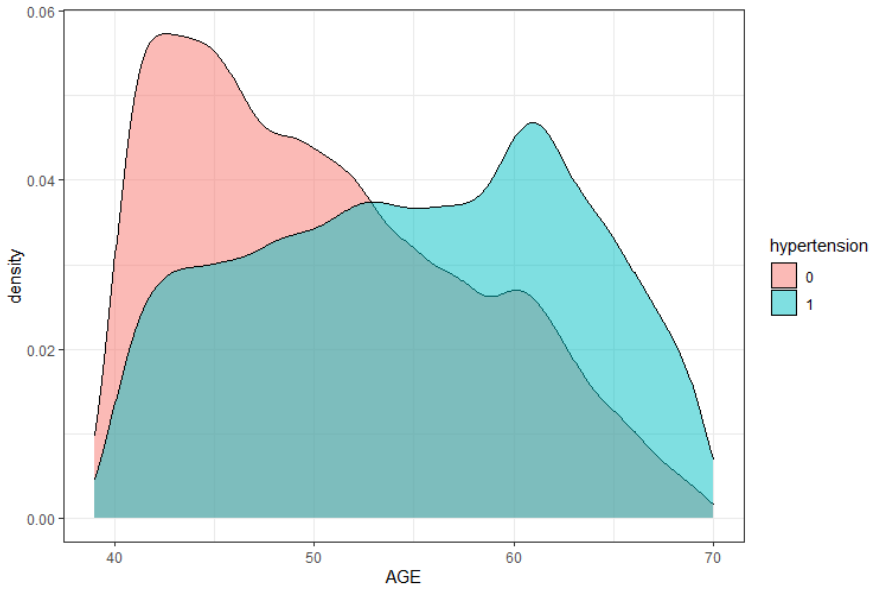


Figure 4.9. Distributions of age in hypertensive and non-hypertensive individuals within UK Biobank.

Of course, the relationship may be genuine in these data, even though it was not as expected. Using cases with only one phenotype and healthy or 'super'-controls that are free from any disease could be an option. However, in this analysis we cannot rule out the presence of unaccounted health conditions or missing data. We should also keep in mind the likelihood of an individual developing the trait. Using age-matching controls

could be another option, but the definition of hypertension seems to play the biggest role, and an association analysis of blood pressure as a continuous outcome might be more insightful rather that this analysis of case-control groups.
In my initial approach I allowed phenotypes to overlap, this works under the assumption that there is no pattern of disease linked to the disease of interest.

4.8.4.3. Shorter telomeres and cancer

Previous observational studies reported inconsistent associations with telomere length and cancer risk (**Table 2.2**). Some previous reports suggested a U-shape association with the extremes for both short and long telomeres increasing the risk of cancer (Barthel *et al.*, 2017; Thriveni *et al.*, 2019). Analysis in tumour cell lines have been shown to have shorter telomeres than surrounding normal tissue (Blasco, 2005; Barthel *et al.*, 2017). This corresponds to the idea that at an early-stage cancerous cells lose telomeric repeats with every cell division, which occurs more often due to uncontrolled division. However, around 90% of all cancer types have reactivated telomerase that elongates telomeres by adding new telomeric repeats, which provides certain stability for the genome and allows cancer cells to divide indefinitely (Blasco, 2005). This change in telomerase activity, from telomere shortening to telomere elongation, may be one of the reasons for inconsistent results seen in observational association studies between telomere length and cancer risk.
Shorter GDTL was reported to be associated with a decreased risk of cancer and my GRS results confirm the protective effect of genetically shorter telomeres against several cancers (**Table 4.5**). For example, in the association analysis of GRScojo52 shorter GDTL showed an 8% decreased risk of developing melanoma for every 1 SD increase in GRS (OR=0.92, 95% CI: 0.90-0.95). This result confirms the findings of the previous association study using a GRS built from 7 SNPs (Iles *et al.*, 2014).
I did not confirm the associations of disease phenotypes such as gastric cancer, pancreatic cancer, and breast cancer. However, the number of cases of gastric and pancreatic cancers within UKB was relatively small, which may have limited the possibility of detecting the association. The association of breast cancer, although sufficient sample size was available, did not reach the significance level.

Previous GRS studies					GRScojo52			
Disease Phenotype	**N (SNP)**	**Sample size (cases / controls)**	**Shorter TL association**	**Author, Year**	**Disease phenotype**	**N(Case)**	**P-value**	**OR(95%CI)**
Melanoma	7	11,108 / 13,933	Decreased risk	(Iles *et al.*, 2014)	Melanoma	5362	4.76E-09	0.92(0.90,0.95)
Lung cancer	7	5,457 / 4,493	Decreased risk	(Machiela *et al.*, 2015)	Lung cancer	3298	2.50E-09	0.90(0.87,0.93)
Gastric cancer	8	1,136 / 1,012	U-shaped association	(Du *et al.*, 2015)	Stomach cancer	895	1.89E-01	0.96(0.89,1.02)
Non-Hodgkin lymphoma	9	10,102 / 9,562	Decreased risk	(Machiela *et al.*, 2016)	Non-Hodgkin lymphoma	2641	4.89E-03	0.94(0.91,0.98)
Chronic lymphocytic leukemia					Leukemia	1624	1.30E-07	0.87(0.83,0.92)
Small lymphocytic lymphoma					Lymphoma	3948	1.37E-06	0.92(0.89,0.95)
Chronic lymphocytic leukemia	8	273 / 5,725	Decreased risk	(Ojha *et al.*, 2016)	Leukemia	1624	1.30E-07	0.87(0.83,0.92)
Breast cancer	6	2,865 / 2,285	Decreased risk	(Luu *et al.*, 2016)	Breast cancer	17691	1.55E-01	0.99(0.97,1.00)
Pancreatic cancer	8	1,500 / 1,500	No association	(Antwi *et al.*, 2017)	Pancreas cancer	949	1.01E-01	1.06(0.99,1.13)
Pancreatic cancer	10	2,374 / 4,326	Increased risk	(Campa *et al.*, 2018)	Pancreas cancer	949	1.01E-01	1.06(0.99,1.13)
Thyroid cancer	9	118 / 5,206	No strong association	(Gramatges *et al.*, 2019)	Thyroid cancer	756	4.45E-06	0.84(0.78,0.91)

Table 4.5. Association results of genetic risk score for shorter telomere length confirm previous associations of genetic telomere length and cancers. The results of GRScojo52 are coloured as follows: Bonferroni significant (green), nominally significant (black), non-significant (grey). OR<1 indicates the decreased associated risk with shorter GDTL.

4.9. Conclusions on the findings of effects of genetically determined telomere length on disease risk

Using TL GRS I analysed the relationship between the GDTL and age-related diseases in order to investigate the impact of the genetic component of TL on human health. Although the selected 52 SNPs that I used to build GRS explain only 2.93% of the TL variance (Li *et al.*, 2020), I was able to detect a number of significant results. Shorter GDTL was associated with an increased risk of two cardiovascular phenotypes, CAD and aortic valve stenosis, and a decreased risk of several cancers (melanoma, lung cancer, leukemia, lymphoma, and others) and proliferative diseases (uterine fibroid, benign prostatic hyperplasia, uterine polyps, and others). The effects of shorter GDTL were stronger in cancers than in cardiovascular group, when based on effect size and

statistical significance. For example, shorter GDTL was associated with a 10% decrease in the risk of lung cancer, and just a 2% increase in risk of CAD.

The findings suggest that shorter GDTL is protective against cancers potentially though limiting the cell's replicative capability as was demonstrated with inherited or measured TL, where TL serves as a measure of the replicative ability of the cell.

The TL GRS, in combination with measured TL and clinical risk scores, may be used to optimise and improve the discrimination and prediction of incident cases of age-related diseases, and enhance the screening for individuals at risk in order to manage individual preventive therapies.

The analyses performed had several limitations. This includes a reliance on the accuracy of defining the disease phenotype. While these disease phenotypes have been created with help of clinical colleagues, it cannot be ruled out other diseases not considered or any errors in the self-reported or hospital episode data. The selection of controls in this analysis works under the assumption of no underlying pattern of disease within the disease of interest. This is also true of the controls who are assumed to be at the same risk of other diseases as the case group. This will not be true of some comorbidities or common risk factors that may also be linked to TL. A small sample size can impact the ability to detect an association due to low statistical power. The selection of SNPs assumes independence that may hold true given the dual approach to GRS generation here. However, pleiotropic variants have not been investigated or removed. The analysis goes some way to reducing the impact of winner's curse through the application of a correction via the FIQT, a novel approximation of which was applied. This will not rule out the selection of SNPs that are not truly associated to TL but will reduce their impact. The results, thus, need to be interpreted with caution keeping in mind the study design described.

Moreover, GRS association does not imply causality, and cannot prove that the link from GDTL to outcome is genuine or that GDTL is causing the disease risk to change. One of the project aims is to investigate if GDTL may contribute or lead to a change in disease risk. To investigate potential causality of these reported associations I conducted a full mendelian randomisation study that is going to be described in the next chapter.

Chapter 5. Mendelian randomisation study of telomere length

In this chapter I am going to introduce the basic concepts of mendelian randomisation and its application in this project to test the causal association between telomere length and age-related outcomes using genetic instruments. I start with an introduction of the underlying ideas and assumptions of mendelian randomisation, and then describe the methods using an extract of the results to aid illustration before presenting and discussing all findings.

5.1. Mendelian randomisation background

5.1.1. The concept of mendelian randomisation

Classical epidemiology aims to study patterns of health and disease within populations and its main limitation is an inability to distinguish between correlation, or association, and causation. Observed correlations between measured factors and outcomes may be causal or not. This is mainly due to the problems that exist in observational research because of residual confounding, a bias that remains after adjusting for confounders. If we want to determine the cause of a disease, or the impact of a treatment, or to be able to develop medical interventions, or inform the public about lifestyle choices, we need to establish relationship as having a causal effect on the outcome. This process is known as causal inference.

Randomised controlled trials (RCTs) are a gold standard for drawing causal inferences. RCTs test the effect of an intervention of interest by randomly assigning individuals, for example, into two groups, one that receives the drug and one that receives placebo. However, RCTs have some limitations such as being expensive, time-consuming and cumbersome in cases, when risk factors cannot be allocated due to practical or ethical reasons (Burgess *et al.*, 2015).

Genetic epidemiology studies the role of genetic factors in health and disease, where genetic variants cannot be randomised in a controlled manner as, for example, a treatment in RCT. However, under Mendel's laws of inheritance we understand that genomic data are randomised by nature at conception and, thus, provide a natural

assignment into groups that have specific alleles and genotypes. This randomised assortment of genetic variants is known as Mendelian Randomisation (MR).

This naturally occurring phenomena can be considered as nature's randomisation through distribution of alleles in the population, and is described in Mendel's second law, which states that alleles are transmitted to the offspring at conception with equal probability (Mendel, 1865). External factors such as age, gender, socioeconomical status have no influence on which allele is transmitted at conception to the offspring and which genotype the offspring gets. This provides genetic association studies an advantage of being free of traditional epidemiological confounding and reverse causation, as genotype precedes phenotype. We use this property in MR, sometimes called nature's randomised trial, to draw causal inferences, while avoiding the expense and ethical issues of randomised trials (Hingorani *et al.*, 2005; Dorak, 2017; Hemani, Tilling, *et al.*, 2017; Evangelou, 2018; Hemani *et al.*, 2018; Burgess *et al.*, 2019; Morrison *et al.*, 2020). While RCTs require a high cost investment, and appropriate ethical approval due to their interventional nature, MR does not require interventions, as it utilises genetic information and available phenotypic information to provide evidence of potentially causal relationships (**Figure 5.1**) (Davies *et al.*, 2018). Thus, MR is a great solution to guide interventional research and provide guidelines for public health when RCTs are not possible (Smith *et al.*, 2003; Grover *et al.*, 2017; Morrison *et al.*, 2020).

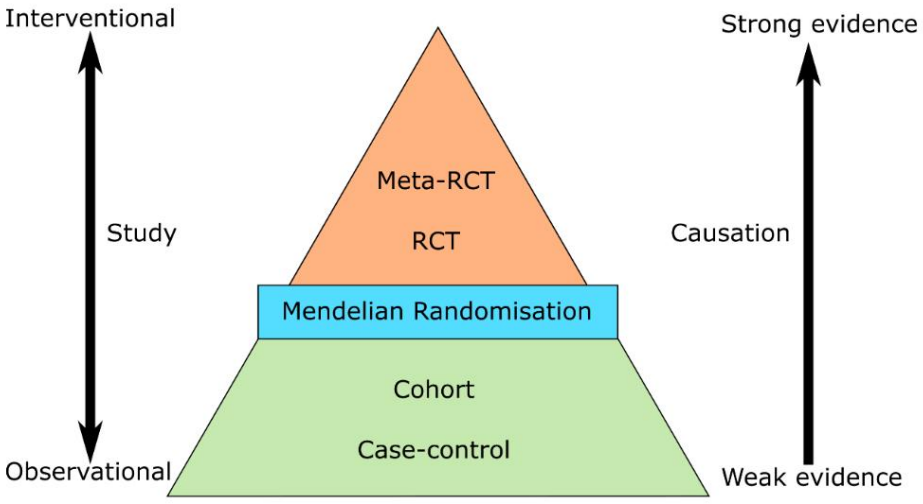


Figure 5.1. Mendelian randomisation's place in modern epidemiology.

5.1.2. Design and assumptions of a mendelian randomisation study

Genetic discoveries have added substantial knowledge into the understanding of biological pathways that lead to disease. Whilst genetic factors can be used to predict individual disease risk, they are not of direct interest from a clinical perspective, because the genome cannot be altered. However, MR fills this gap by using genetic information to detect and estimate causal effects of non-genetic modifiable risk factor or exposure that was associated with the disease in observational data (Smith *et al.*, 2003; Burgess *et al.*, 2015). It should be noted that identification of genetic variants associated with either disease risk or risk factors can lead to novel therapeutic interventions.

MR utilises the instrumental variable approach previously used more extensively in economics. In MR the genetic variants are known as the instruments or genetic instrument and defined by the genotype that affects the disease outcome indirectly (through the exposure or intermediate phenotype). The genetic instrument is assigned randomly at conception, and is independent of confounding factors (Didelez *et al.*, 2007). There are three key assumptions of MR, and only when all three are met can the genetic instrument be considered for MR. Such genetic instrument is referred as valid.

To illustrate the MR assumptions, we can consider MR via a directed acyclic graph as shown in **Figure 5.2**. Here we assume that a genetic variant *G* is a valid instrumental variable (IV) to infer the causal effect of the exposure *X* on the outcome *Y* (Burgess, 2011; Smith *et al.*, 2014; Boef *et al.*, 2015; Davies *et al.*, 2018; Burgess *et al.*, 2019; Kachuri *et al.*, 2019), if:

1) The relevance assumption: The genetic instrument is associated with the exposure *X* (assumption IV1).
2) The independence assumption: The genetic instrument is not associated with any confounders, *U*, of *X* and *Y* (IV2).
3) The exclusion restriction assumption: The genetic instrument is associated with the outcome *Y* only through its effect on the exposure *X* (IV3).

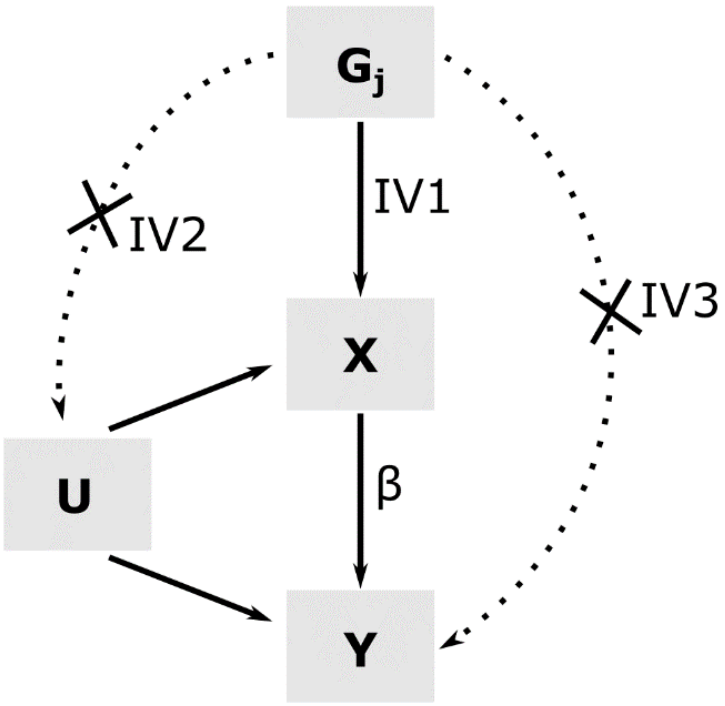


Figure 5.2. Directed acyclic graph representing the standard assumptions of mendelian randomisation. G_j is the j^{th} genetic instrument, *X* – exposure or risk factor, *Y* – outcome, *U* – unknown or unmeasured confounder. IV1 (solid line) is a SNP-exposure association, IV2 (dotted line) is a SNP-confounder association, and IV3 (dotted line) is a direct SNP-outcome association. Both IV2 and IV3 violate MR assumptions. β is an estimated causal effect of a unit increase in exposure X on the outcome *Y*.

To perform MR using TL as the exposure, or intermediate phenotype, we assume that *G* is a genetic determinant of the exposure *X*, telomere length, *Y* is the outcome, age-related disease, while *U* are environmental cofounders that may affect measured TL, disease, or both. We assume that genetic instruments are valid if all three MR assumptions are met, so we now look at these in more detail:

1) The genetic variants must be associated with TL. If the association between a genetic variant and TL is not true, it would violate assumption IV1, and, when there is no association found between the genetic variants and disease outcome, can be misinterpreted as evidence against a causal association between TL and disease. The selection of biologically plausible genetic variants, associated with TL, helps to minimise the probability of violating the IV1 assumption (Haycock *et al.*, 2016).

2) The genetic variants should not be associated with any confounders that affect measured TL or disease. For example, if the genetic determinants of TL were associated with smoking status, a known cofactor to influence measured TL and other biomedical parameters, it would violate the MR IV2 assumption. Such a violation would indicate that the genetic instrument is invalid due to the pleiotropic effect of being associated with another trait, which would in turn make it difficult to draw conclusions about true TL-disease association.
3) The genetic variants of TL should be associated with disease only through TL. The assumption IV3 is violated, if genetic variants of TL are directly associated with the disease outcome. A direct effect of a genetic variant on disease would indicates that the causal pathway does not involve TL, the hypothesised exposure (Haycock *et al.*, 2016).

Using valid genetic instruments, we can test their association with disease and draw causal inferences about the role of TL, the non-genetic intermediate phenotype, on age-related disease outcomes. TL is not measured directly, as this association is affected by many confounders. Instead TL is proxied by its genetic determinants (Dorak, 2017, p150).

5.1.3. Mendelian randomisation estimation of causal effect

Using a standard MR approach, we would obtain summary statistics from a genome-wide association study to identify genetic instruments G*j* that are associated with the exposure *X*, giving an estimated beta-coefficient $\hat{\beta}_{Xj}$ and standard error se($\hat{\beta}_{Xj}$). We would then estimate the association of each genetic variant *Gj* with the outcome *Y*, given by beta-coefficient $\hat{\beta}_{Yj}$ and standard error se($\hat{\beta}_{Yj}$). The causal effect $\hat{\beta}_j$ of the exposure *X* on the outcome *Y* is then estimated using the Wald ratio (Del Greco *et al.*, 2015; Hemani *et al.*, 2018; Burgess *et al.*, 2019):

$$\hat{\beta}_j = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}}$$

And its standard error se($\hat{\beta}_j$) is:

$$se(\hat{\beta}_j) = \frac{se(\hat{\beta}_{Yj})}{\hat{\beta}_{Xj}}$$

To estimate causal associations we synthesise evidence from multiple genetic variants using MR methods (Burgess *et al.*, 2016, 2018) such as the inverse-variance weighted (IVW) MR method that estimates the causal effect from *L* uncorrelated genetic instruments as:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{L} \omega_j \hat{\beta}_j}{\sum_{j=1}^{L} \omega_j}$$

where ω_j is the inverse-variance of $\hat{\beta}_{Xj}$.

Inverse-Variance Weighted MR is a common approach to estimate the causal effect. It utilises a standard meta-analysis approach, using either a fixed or random effects model that combines the evidence from multiple variants. Each genetic instrument contributes to the combined estimate and is weighted by the inverse of the variance $\omega_j = se(\hat{\beta}_j)^{-2}$ (Burgess *et al.*, 2013; Hemani *et al.*, 2018). This method helps to incorporate multiple genetic effects, thus increasing the statistical power and yielding a more precise estimate of the effect ($\hat{\beta}_{IVW}$).

Initially MR analyses were performed in a single study that would measure both exposure and outcome phenotypes, detect the genetic instruments of exposure, and draw a causal inference. This approach is known as a one-sample MR. Further developments in MR methods, and the increased availability of GWAS summary statistics, has allowed the expansion to two-sample MR. This is where one study provides a set of genetic instruments with estimated effects for the exposure, and a second study provides genetic data to extract the required instruments from an independent set of subjects with desired outcome (Zheng *et al.*, 2017). These advances have led to an explosion of MR analyses in the literature.

5.1.4. Sensitivity analysis – pleiotropy or mediation?

In two-sample MR the exposure and outcome are measured in independent samples. The exposure is proxied by the genetic instruments in the second sample, and causal inferences about the role of the exposure are made by investigating the association between the genetic instruments and the outcome (Dorak, 2017; Hemani *et al.*, 2018). If the outcome association is due to the genetic instruments, identified by their association with the exposure (TL), then MR is thought to estimate a reliable causal

association of exposure on outcome. However, each genetic instrument may influence exposure and outcome through independent pathways, which is known as horizontal pleiotropy as per the assumptions above. The instruments may also be associated with a covariate that is not a confounder but a mediator on the causal pathway from the exposure to the outcome, known as vertical pleiotropy (**Figure 5.3**) (Haycock *et al.*, 2016; Burgess, Bowden, *et al.*, 2017), though this does not violate the MR assumptions.

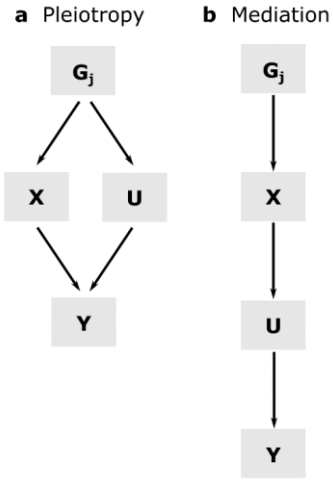


Figure 5.3. Horizontal and vertical pleiotropy. (a) Horizontal pleiotropy. The genetic variant is associated with the risk factor *X* and covariate *U* via different causal pathways. (b) Vertical pleiotropy or mediation. The genetic variant is associated with the risk factor *X* and consequently with the covariate *U*. G_j is the j^{th} genetic variant, *X* – exposure or risk factor, *U* – unknown or unmeasured confounder, *Y* – outcome.

Mediation is likely if several variants have the same direction of effect with the confounder. It may not be pleiotropy of the variants, but a representation of the downstream consequence of the exposure if the exposure is modified by intervention. In case of mediation, or vertical pleiotropy, the genetic variants are still valid MR instruments. This is because the causal pathway from the genetic instruments to the outcome is still only possible via the exposure, although there is a mediator on this pathway (Haycock *et al.*, 2016; Burgess *et al.*, 2017).

In MR, where multiple genetic variants are used, it is unlikely that all the genetic variants satisfy the assumptions of instrumental variables. Accounting for possible pleiotropic effects using sensitivity analysis is required to estimate correct causal effects. If multiple genetic variants for exposure are all concordantly associated with the confounder, it is likely due to mediation, not pleiotropy. If only a small number of variants are associated with the confounder, then it is likely due to horizontal pleiotropy. The sensitivity analyses aim to help to distinguish between these two (Burgess *et al.*, 2017; Slob *et al.*, 2019).

5.1.4.1. Egger's test for pleiotropy

One way to detect pleiotropy is to perform sensitivity analysis using Egger's test, originally proposed to detect small study bias in meta-analysis (Egger *et al.*, 1997; Bowden *et al.*, 2015). MR that uses multiple genetic instruments is analogous to meta-analysis, and Egger's regression was adapted for MR to estimate causal effect sizes and detect violations of a key MR assumption, requiring no horizontal pleiotropy. While MR-Egger gives a causal effect estimate, the intercept of MR-Egger provides a test for asymmetry or directional pleiotropy. It is possible to assess potential asymmetry in the causal estimates from MR-Egger visually in a funnel plot showing estimates of each genetic instrument (**Figure 5.4)**. In this plot the SNP-exposure estimates β_{Xj} are plotted against the SNP-outcome estimates β_{Yj}. If plot asymmetry occurs, then there is evidence of directional pleiotropy – that the genetic instruments have pleiotropic effects that are not balanced about the null (Bowden *et al.*, 2015).

MR-Egger regression performs weighted linear regression, similar to the IVW method, but with an unconstrained intercept. It can be written as:

$$\hat{\beta}_{Yj} = \beta_{0E} + \beta_{E}(\hat{\beta}_{Xj})$$

It regresses $\hat{\beta}_{Yj}$ coefficients on the $\hat{\beta}_{Xj}$ coefficients, where β_{0E} and β_{E} are the coefficients in the regression model (0 for the intercept, E – for Egger's) (Egger *et al.*, 1997; Bowden *et al.*, 2015). The intercept, β_{0E}, estimates the average effect across all genetic instruments. Testing the intercept allow us to identify, whether the intercept is different from zero. The intercept starting at the origin (0,0) suggests no pleiotropy, and the intercept significantly different from zero indicates evidence of directional pleiotropy (Bowden *et al.*, 2015; Haycock *et al.*, 2016; Burgess *et al.*, 2017; Hemani *et al.*, 2017; Slob *et al.*, 2019).

For example, MR-Egger using the 52 genetic TL determinants on CAD showed no evidence of pleiotropy, as the intercept passes through the origin (**Figure 5.4**). While the same test for kidney cancer showed significant evidence of pleiotropy, as the SNP effects were scattered, and the intercept does not pass through the origin (**Figure 5.5**).

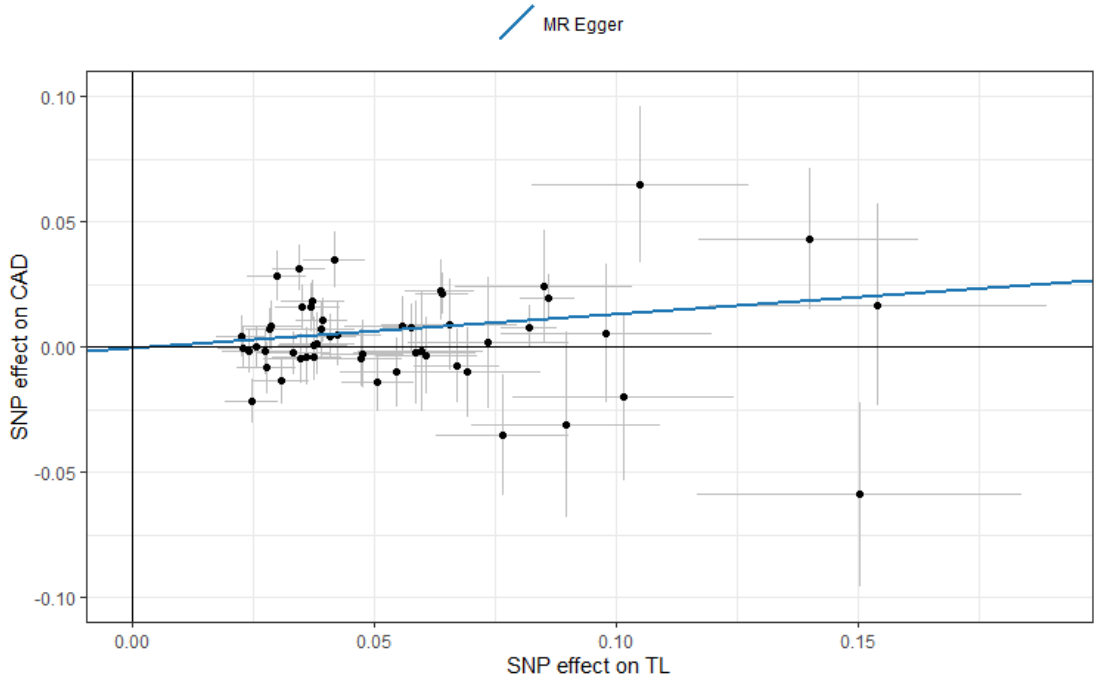


Figure 5.4. Egger's test detects no pleiotropy. The test was performed using 52 genetic determinants of TL and SNP-TL (X-axis) and SNP-CAD (Y-axis) effects were plotted. The MR-Egger did not detect any pleiotropic effects of 52 used SNPs, as the intercept passed through the origin (0,0).

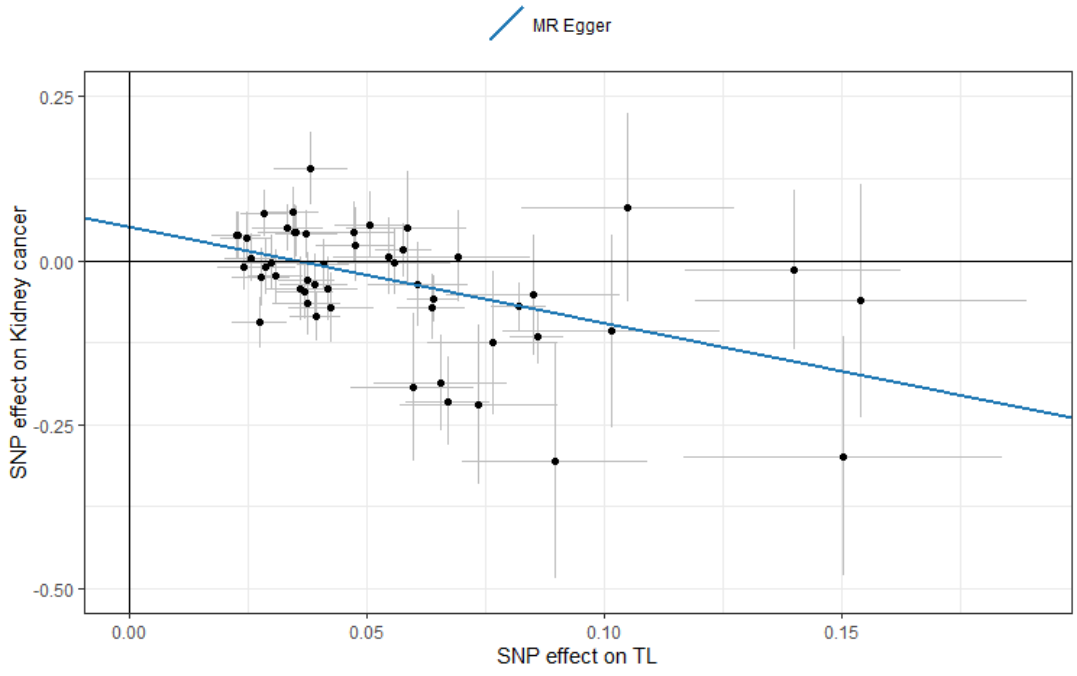


Figure 5.5. Egger's test detects pleiotropy. The test was performed using 52 genetic determinants of TL and SNP-TL (X-axis) and SNP-Kidney cancer (Y-axis) effects were plotted. The MR-Egger detected pleiotropic effects, as the intercept did not pass through the origin (0,0).

5.1.4.2. Mendelian randomisation Steiger test of directionality

The MR-Steiger test provides a different approach to sensitivity analysis with potential utility in causal inference. The principal idea is similar to a mediation-based analysis that orients the causal direction between the exposure and the outcome by following logical link. For example, if a genetic variant is associated with the exposure, and exposure – with the outcome, in mediation analysis we would assume that a direct influence of the genetic variant on the outcome is zero when conditioning on the exposure. In this case, the exposure would mediate the association between the genetic variant and the outcome, suggesting the direction of causal effect (Hemani, Tilling, *et al.*, 2017). Testing both directions of association, the exposure to the outcome and the outcome to the exposure using an MR is commonly known as bi-directional MR. In a mediation analysis the results of the exposure causing the outcome or the outcome causing the exposure will be different and the model with the strongest evidence would be chosen. Such an approach enables us to distinguish between a true causal association and an association that may be found due to reverse causation or residual confounding (Hemani, Tilling, *et al.*, 2017). However, mediation-based approaches require all variables to be measured in the same dataset, and they are not applicable for two-sample MR when the exposure and outcome are measured in different samples. Bi-directional MR can be used in such cases to orient causal directions. It tests whether the genetic instrument has a primary effect on *X* or *Y*. However, the biological knowledge of a valid instrument is required for both the exposure and outcome for the MR to be valid.

The MR-Steiger test was developed for application in scenarios where the biology behind the associated genetic instrument is not fully understood and only summary statistics for both the exposure and outcome are available. Thus, it can be applied to a two-sample MR study design. The MR-Steiger Z-test assesses the difference between two independent correlations, SNP-exposure and SNP-outcome, and distinguishes whether the exposure *X* causes the outcome *Y* or the outcome *Y* causes the exposure *X* (Steiger, 1980; Hemani, Tilling, *et al.*, 2017). Both correlations are identical under the null hypothesis and the probability *Z* of obtaining a difference between the two correlations, at least as large as that which is observed, is represented by the MR-Steiger p-value. The inference of causality and the most likely direction of effect is drawn from

a combination of bi-directional MR analyses and the MR-Steiger test. If MR-Steiger agrees with the direction of effect (denoted as sign(*Z*)), observed between exposure and outcome, then it supports the evidence of a potential causal association.

Assuming a significance threshold of α and probability of observing a difference between the two correlations of *Z*, the inference from the MR-Steiger test is as follows:

- MR-Steiger p-value < α and MR p-value < α and $Z > 0$, then a causal association of the exposure influencing the outcome is accepted (*X* -> *Y*).
- MR-Steiger p-value < α and MR p-value < α and $Z < 0$, then a causal association of the outcome influencing the exposure is accepted (*Y* -> *X*).
- MR-Steiger p-value > α or MR p-value > α, neither is accepted, as there is no evidence of a causal association.

The MR-Steiger test can be performed in the R package *TwoSampleMR* (version 0.4.26 used in this project) (Hemani *et al.*, 2018). It must be noted that the MR-Steiger test is prone to bias under 1) horizontal pleiotropy, 2) differential values of measurement error between the exposure and the outcome, or 3) unmeasured confounding between the exposure and the outcome (Hemani, Tilling, *et al.*, 2017). Other sensitivity analyses should be used to confirm MR-Steiger results.

5.1.4.3. Median-based and robust adjusted profile score mendelian randomisation – accounting for weak instruments

The conventional inverse-variance weighted method, commonly used for MR analysis, only provides consistent estimates if the genetic variants are valid instruments (5.1.2. *Design and assumptions of mendelian randomisation study*). When analysing complex traits with multiple genetic variants contributing small effects, it is unlikely that all genetic variants can be classified as valid MR instruments. Firstly, the association of genetic variants with the exposure is prone to measurement error, this highlights the potential for winner's curse to bias the selection of variants (Haycock *et al.*, 2016). Secondly, not all genetic variants are strongly associated with the exposure, which may make them weaker instruments and bias the estimation of the underlying true effects of the exposure on outcome, known as weak instrument bias (Zhao *et al.*, 2018).

The MR median-based approach accounts for bias, introduced by invalid or weak genetic instruments, and allows for up to 50% of variants to violate MR assumptions (Bowden *et al.*, 2016). This approach calculates causal estimates for each genetic variant individually, where genetic variants with more precise estimates receive more weight. Then causal estimates are sorted, and the weighted median is taken from the newly formed distribution of weighted causal estimates. The causal estimation is not affected by the number of outliers. With at least 50% of the genetic variants being valid instruments, the causal estimates from all valid instrumental variables will tend towards the same value as the sample size increases (Bowden *et al.*, 2016; Burgess *et al.*, 2017; Burgess, Thompson, *et al.*, 2017; Slob *et al.*, 2019). The method is therefore robust to weak instrument bias and winner's curse but may suffer from a reduction in statistical power.

Another approach that accounts for bias introduced by weak instruments is MR Robust Adjusted Profile Score (RAPS). MR-RAPS obtains causal estimates using a profile likelihood of the summary data and the variance of the pleiotropic effect distribution. The profile likelihood can be described as a linear regression of SNP-outcome on SNP-exposure using L_2-loss function, which stands for a special case of Least Square Errors and is used to minimise the error, where pleiotropic effects are thought to be normally distributed about zero with unknown variance (Hemani *et al.*, 2018; Zhao *et al.*, 2018; Slob *et al.*, 2019).

5.2. Investigation of genetic telomere length using mendelian randomisation

The previous chapter 5.1. *Mendelian randomisation background* covered the concepts of MR, the key assumptions and how to ensure we have a comprehensive assessment of the reliability of mendelian randomisation estimates. The various methods provide different approaches to account for any bias introduced by genetic instruments that violate MR assumptions.

In this chapter I describe the study design of my MR of TL on age-related disease. I am going to detail the methods used, and the sensitivity analyses performed to provide a robust inference from the MR findings. I then discuss the validity of the presented MR

results as well as the biological relevance and ways through which genetic TL may be causally associated to age-related diseases.

5.2.1. Mendelian randomisation study design of the project

In this project I utilised the IVW MR as my primary analysis method to test the hypothesis of a causal effect of TL (exposure) on over one-hundred age-related diseases (outcomes). All the techniques and methods that are used are described in chapter *5.1. Mendelian randomisation background*. To test the consistency of association, considering weak instruments and pleiotropic effects, I additionally employed several sensitivity analyses such as MR-Egger and MR-RAPS.

The workflow of data preparation and analysis are illustrated in **Figure 5.6** and consisted of the following steps:

1) Association of genetic variants with TL (SNP-TL)
2) Association of genetic variants with disease (SNP-outcome)
3) Data transformation
4) Primary MR analysis
5) Sensitivity MR analyses
6) Inference and interpretation

I used a two-sample MR study design that uses data from two independent study populations. This allows for higher statistical power to infer causal links with no requirement to measure both the exposure (TL) and outcome (disease) in the same study (Davies *et al.*, 2018). The ENGAGE study (Li *et al.*, 2020) performed a GWAS meta-analysis to estimate the effects of genetic variants on telomere length (TL) as previously described in chapter *3.4. A new genome-wide association study of telomere length*. The estimates are denoted as $\hat{\beta}_{Xj}$ in SNP-TL associations. The disease association data were estimated using the UK Biobank cohort and described further in this chapter. These estimates are denoted as $\hat{\beta}_{Yj}$ in SNP-outcome associations.

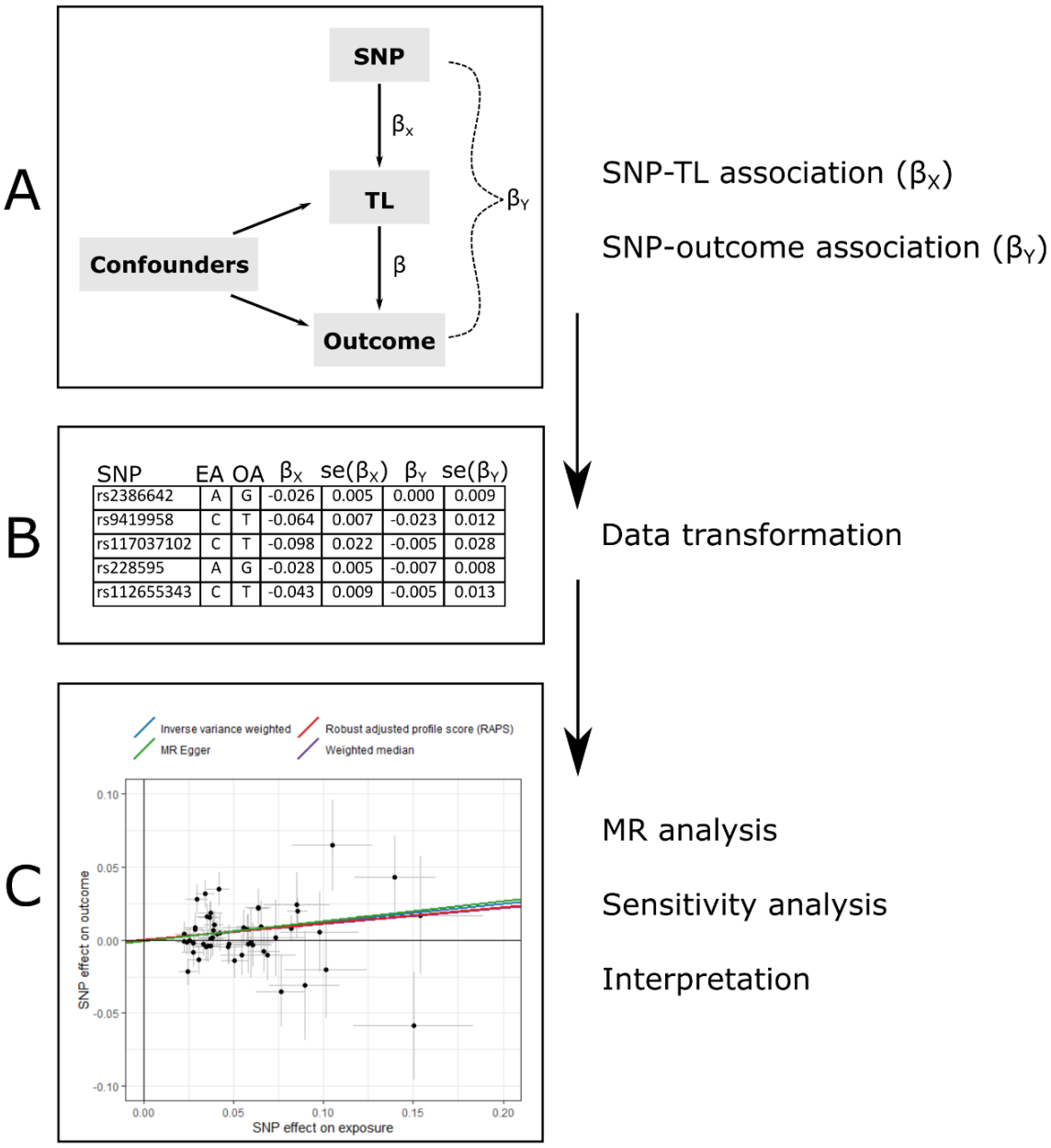


Figure 5.6. Mendelian randomisation study workflow. A) Data preparation consists of SNP-TL (β_X) and SNP-outcome (β_Y) associations that are used to estimate the TL-outcome causal association (β). B) Data transformation, where alleles are harmonised for consistency so both studies reference the same allele, here this represents the allele associated with shorter TL. C) Harmonised data are used in the MR analyses, using MR-IVW (blue line), where different MR methods are applied as sensitivity analyses, MR-Egger (green line) to detect pleiotropy and both MR-median (purple line) and MR-RAPS (red line) are used to assess bias introduced by weak instruments. Through assessment we consider the consistency of all MR estimates across the methods. The results are then reported and interpreted.

The six steps from above are described in further detail below. Given in sufficient detail so that the analysis could be replicated independently if required.

1) Association of genetic variants with TL (SNP-TL)

SNP-TL association data was generated in the ENGAGE study (Li *et al.*, 2020). In this large scale GWAS meta-analysis we identified 52 genetic variants that were significantly associated with TL and shown to be conditionally independent by GCTA joint analyses after applying a 5% false discovery rate. We defined the effect of each genetic instrument on TL as β_{Xj}. TL measurements were Z-standardised for consistency across the multiple cohorts in the meta-analysis and the estimates can be interpreted as the SD change in TL per copy of the effect allele. These data were made publicly available (Li *et al.*, 2020) and the summary statistics for the 52 variants used here are given in the **Appendix 1** *Genetic determinants of telomere length*.

2) Association of genetic variants with disease (SNP-outcome)

I defined 112 general and 15 gender-specific disease phenotypes as outcomes, the cases of which were extracted from the UK Biobank phenotypic and hospital data. These were described in *4.8.2. Selecting phenotypes and assigning case-control status*. I defined the effect of each genetic instrument on disease as β_{Yj}, obtained using logistic regression in SNPTEST (Marchini *et al.*, 2007) on genetically unrelated samples (kinship<0.088) adjusting for sex, age, the first 5 genetic principal components and the genotyping array. β_{Yj} can be interpreted as the change in disease log odds per copy of the effect allele.

3) Data transformation

Both β_{Xj} and β_{Yj} were harmonised to the allele associated with shorter TL to ensure that that disease and TL effects are consistent across datasets. Data preparation and harmonisation were performed using my own developed scripts in Python (version 2.7.5).

4) MR analysis

I estimated the potentially causal effect $\hat{\beta}$ using a Wald ratio. The effects of 52 SNPs, available from UK Biobank genetic data, were combined using the IVW method to calculate the MR estimate over all genetic variants. MR analysis was performed using *TwoSampleMR* (version 0.4.26), in R (version 3.5.1).

5) Sensitivity analysis

The assumption that each genetic instrument works on the outcome only through TL may be violated if there are any pleiotropic effects, which means that it affects the outcome through a different pathway from that which is hypothesised (horizontal pleiotropy). I performed sensitivity analysis using MR-Egger regression to determine the evidence of significant pleiotropy, and directionality test using MR-Steiger to determine the correct direction of association. Additionally, MR median-based and MR-RAPS were also used to account for different types of pleiotropy and to assess the consistency in causal estimates.

6) Inference and interpretation

Visualisation of results was performed using R libraries *TwoSampleMR* (Hemani *et al.*, 2017; Hemani *et al.*, 2018) and *Circlize* (Gu *et al.*, 2014).

5.2.2. Findings of mendelian randomisation study of telomere length

In this chapter I focus on the significant results from the performed MR analyses. The results are presented using MR IVW estimates, as MR IVW is considered as the standard method for reporting causal inference from MR. Estimates from the other sensitivity MR methods are presented alongside MR IVW in accompanying tables. Full detailed results for all 127 diseases using all MR methods can be found in **Appendix 5** *Mendelian randomisation study results*.

According to MR IVW estimates there were 33 nominally significant associations, 9 of which passed the Bonferroni corrected threshold ($P \leq 0.05/127 = 3.937 \times 10^{-4}$). The effect sizes presented all consider a 1 SD decrease in TL. Shorter TL was associated with an increased risk of cardiovascular, immune-related, and endocrine diseases, and decreased risk of several cancers and phenotypes with excessive growth (**Figure 5.7** and **Table 5.1**).

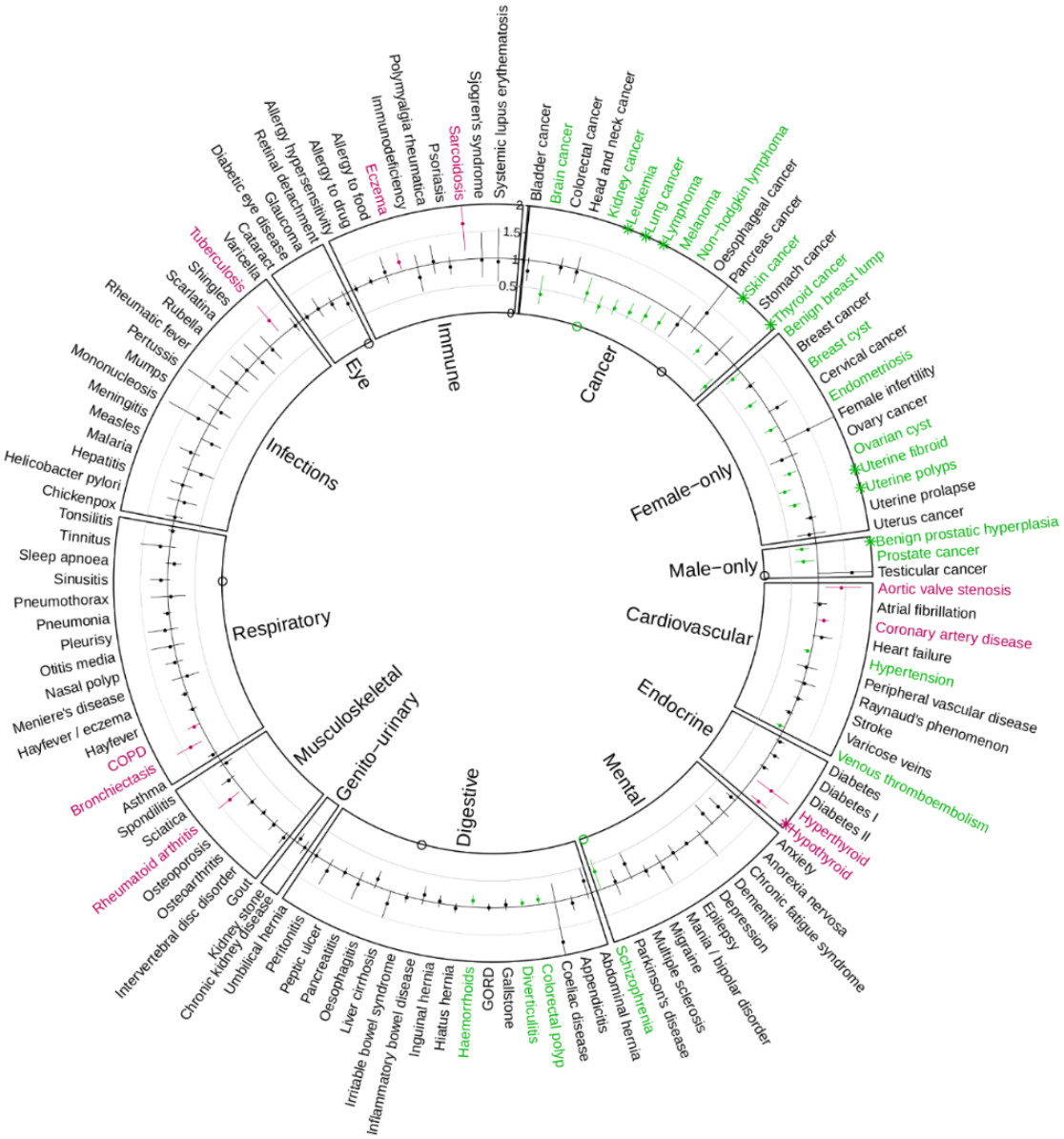


Figure 5.7. Causal association estimates for telomere length with 127 diseases. Shorter TL associated with an increased risk of disease with OR>1 (coloured red) and with a decreased risk of disease with OR<1 (coloured green) that show a nominal level of significance (P≤0.05). The estimated ORs are on the Y-axis. Presence of nominal pleiotropy, based on P-value of MR Egger's Intercept test, is marked with small circles on the inner circle with disease group names. The associations that survived Bonferroni correction for multiple testing are denoted with a star in front of a disease name.

Table 5.1 shows results of nominal significance according to MR IVW method. The results can be interpreted as follows: the first line of the table shows the association results for uterine fibroid. P-values lower than $3.937x10^{-4}$ show Bonferroni significant associations. P-values for all MR methods, IVW, MR-Egger, MR-median and MR-RAPS are significant for uterine fibroid, suggesting the strong evidence of a robust association. OR, representing the size and direction of effect, are 0.632, 0.455, 0.728, and 0.744 estimated by IVW, MR-Egger, MR-median and MR-RAPS, respectively, indicate agreement in the direction of the causal association. A good level of consistency in effect size estimation suggests that these values are close to the true estimate. MR-Egger's intercept P-value is non-significant, which indicates that there is no evidence of significant pleiotropic effects. MR-Steiger gives a significant P-value, suggestive of confidence in the direction of effect from TL to uterine fibroid.

Modest evidence of nominally significant pleiotropy, detected by MR-Egger's Intercept test, was observed for only two significantly associated diseases, kidney cancer and schizophrenia. Although neither would survive correction for multiple testing, the potential presence of pleiotropic effects suggests that there are other possible pathways between the genetic variants and the traits that do not go through TL. MR-Steiger, on the other hand, suggests that the significant result for kidney cancer is correct, supported by a significant P-value and suggested direction of effect from TL to disease. Non-significant P-values using MR-Steiger for the TL causal association with thyroid cancer, brain cancer and schizophrenia indicate that the test could not confirm the direction of association.

Phenotype	N(Case)	IVW		Egger's		Median		RAPS		Egger's Int		Steiger	
		P-value	OR(95%CI)	P-value	OR(95%CI)	P-value	OR(95%CI)	P-value	OR(95%CI)	P-value	Beta	P-value	Direction
Uterine fibroid	20551	8.37E-07	0.632(0.526,0.758)	7.95E-04	0.455(0.296,0.701)	2.11E-04	0.728(0.615,0.861)	3.64E-04	0.744(0.632,0.875)	1.09E-01	0.01691	4.26E-43	T
Thyroid cancer	756	4.09E-06	0.341(0.216,0.539)	6.73E-02	0.342(0.111,1.053)	1.72E-03	0.335(0.169,0.664)	4.65E-06	0.334(0.209,0.534)	9.95E-01	-0.00016	3.95E-01	T
Lung cancer	3298	6.17E-06	0.552(0.426,0.714)	8.13E-02	0.566(0.303,1.060)	1.27E-02	0.641(0.452,0.909)	1.25E-04	0.584(0.443,0.769)	9.29E-01	-0.00135	2.22E-07	T
Hypothyroid	29377	2.53E-05	1.339(1.169,1.534)	6.62E-02	1.372(0.986,1.908)	8.22E-08	1.399(1.237,1.582)	1.00E-03	1.235(1.089,1.401)	8.74E-01	-0.00125	3.29E-80	T
Leukemia	1624	4.85E-05	0.473(0.329,0.678)	1.09E-03	0.222(0.095,0.520)	1.49E-04	0.396(0.245,0.639)	3.06E-04	0.502(0.345,0.730)	6.14E-02	0.03875	9.61E-03	T
Uterine polyps	14186	6.87E-05	0.740(0.639,0.858)	9.41E-03	0.613(0.429,0.874)	8.78E-07	0.638(0.534,0.763)	2.76E-04	0.754(0.648,0.878)	2.57E-01	0.00973	1.04E-41	T
Skin cancer	23320	6.89E-05	0.718(0.610,0.845)	1.02E-03	0.504(0.343,0.741)	3.84E-05	0.727(0.624,0.846)	1.11E-11	0.740(0.678,0.807)	5.38E-02	0.01807	3.11E-58	T
Lymphoma	3948	8.45E-05	0.604(0.470,0.777)	2.26E-02	0.480(0.261,0.885)	4.99E-07	0.469(0.349,0.630)	1.13E-05	0.572(0.446,0.734)	4.24E-01	0.01173	3.26E-09	T
Benign prostatic hyperplasia	19759	3.03E-04	0.724(0.608,0.863)	1.65E-03	0.498(0.330,0.751)	5.05E-03	0.781(0.657,0.928)	1.40E-02	0.812(0.688,0.959)	5.50E-02	0.01923	5.82E-46	T
Coronary artery disease	36974	1.83E-03	1.133(1.047,1.226)	1.69E-01	1.145(0.946,1.386)	3.96E-02	1.116(1.005,1.239)	6.86E-03	1.118(1.031,1.212)	9.03E-01	-0.00056	1.28E-117	T
Bronchiectasis	3907	2.07E-03	1.343(1.113,1.620)	1.26E-02	1.810(1.155,2.837)	8.44E-02	1.277(0.967,1.686)	6.02E-03	1.320(1.083,1.609)	1.58E-01	-0.01539	1.02E-13	T
Aortic valve stenosis	2603	2.44E-03	1.427(1.134,1.797)	4.34E-01	1.249(0.719,2.169)	6.17E-01	1.092(0.773,1.543)	1.02E-02	1.373(1.078,1.750)	6.04E-01	0.00687	2.95E-08	T
Breast cyst	7199	2.81E-03	0.773(0.652,0.915)	8.50E-02	0.693(0.461,1.043)	7.52E-02	0.810(0.641,1.022)	1.60E-02	0.809(0.680,0.961)	5.70E-01	0.00559	2.86E-23	T
Tuberculosis	2952	3.17E-03	1.380(1.114,1.708)	3.15E-01	1.304(0.781,2.178)	6.69E-03	1.577(1.135,2.193)	5.62E-03	1.399(1.103,1.775)	8.14E-01	0.00291	1.01E-09	T
Brain cancer	922	3.90E-03	0.412(0.226,0.752)	3.72E-02	0.206(0.048,0.875)	8.33E-01	0.929(0.470,1.838)	1.65E-01	0.659(0.366,1.187)	3.06E-01	0.03587	9.48E-01	T
Melanoma	5362	4.53E-03	0.690(0.534,0.892)	5.54E-02	0.536(0.288,0.999)	1.70E-03	0.650(0.497,0.851)	4.34E-05	0.696(0.585,0.828)	3.88E-01	0.01289	3.87E-12	T
Sarcoidosis	1536	4.77E-03	1.666(1.169,2.375)	2.72E-02	2.649(1.145,6.132)	4.61E-02	1.591(1.008,2.513)	1.99E-03	1.680(1.209,2.334)	2.38E-01	-0.02422	3.07E-03	T
Ovarian cyst	14110	6.90E-03	0.827(0.721,0.949)	1.36E-01	0.772(0.553,1.079)	7.14E-02	0.855(0.721,1.014)	1.68E-01	0.908(0.792,1.041)	6.61E-01	0.00352	4.99E-46	T
Haemorrhoids	12795	7.59E-03	0.868(0.782,0.963)	7.75E-02	0.794(0.618,1.020)	4.92E-02	0.855(0.732,0.999)	7.03E-03	0.859(0.770,0.959)	4.50E-01	0.00456	9.42E-49	T
Rheumatoid arthritis	8907	7.81E-03	1.290(1.069,1.557)	2.02E-01	1.350(0.857,2.128)	1.21E-02	1.283(1.056,1.558)	2.28E-02	1.218(1.028,1.444)	8.30E-01	-0.00235	4.19E-26	T
Hyperthyroid	5565	9.09E-03	1.402(1.088,1.806)	2.15E-01	1.484(0.802,2.745)	3.25E-03	1.483(1.141,1.929)	4.02E-02	1.287(1.011,1.639)	8.43E-01	-0.00293	2.50E-14	T
Non-hodgkin lymphoma	2641	9.35E-03	0.705(0.541,0.918)	1.12E-01	0.588(0.309,1.119)	1.15E-03	0.546(0.379,0.786)	4.14E-03	0.673(0.514,0.882)	5.48E-01	0.00927	7.41E-07	T
Eczema	17142	1.01E-02	1.185(1.041,1.348)	5.79E-01	1.093(0.799,1.496)	5.03E-02	1.159(1.000,1.344)	3.96E-02	1.149(1.007,1.311)	5.84E-01	0.00412	3.08E-57	T
Prostate cancer	8962	1.28E-02	0.739(0.583,0.938)	1.00E-01	0.610(0.342,1.088)	7.47E-03	0.714(0.558,0.914)	3.75E-03	0.709(0.563,0.895)	4.77E-01	0.00983	6.81E-18	T
Kidney cancer	1732	1.92E-02	0.642(0.443,0.930)	1.49E-03	0.230(0.098,0.542)	5.28E-04	0.442(0.278,0.701)	8.45E-03	0.595(0.405,0.876)	1.30E-02	0.05222	4.66E-03	T
Endometriosis	7753	2.39E-02	0.845(0.730,0.978)	3.27E-02	0.676(0.476,0.959)	9.76E-03	0.749(0.601,0.932)	6.11E-02	0.863(0.739,1.007)	1.75E-01	0.01148	2.25E-27	T
Colorectal polyp	34018	3.22E-02	0.895(0.809,0.991)	8.82E-03	0.719(0.567,0.911)	1.03E-02	0.862(0.769,0.965)	4.56E-02	0.900(0.812,0.998)	5.21E-02	0.01127	1.06E-101	T
Hypertension	293487	3.72E-02	0.924(0.859,0.995)	1.14E-02	0.793(0.667,0.943)	3.41E-07	0.859(0.810,0.911)	1.37E-02	0.919(0.860,0.983)	6.15E-02	0.00792	2.64E-268	T
Benign breast lump	4076	3.75E-02	0.804(0.655,0.987)	2.60E-01	0.748(0.454,1.232)	4.89E-02	0.759(0.577,0.999)	4.59E-02	0.808(0.655,0.996)	7.58E-01	0.00370	1.15E-12	T
Schizophrenia	1509	3.98E-02	0.635(0.412,0.979)	3.07E-01	1.692(0.623,4.596)	3.76E-01	0.783(0.455,1.346)	6.92E-02	0.675(0.442,1.031)	3.93E-02	-0.05083	8.28E-02	T
Venous thromboembolism	19170	4.13E-02	0.911(0.832,0.996)	7.36E-01	1.038(0.837,1.286)	1.05E-01	0.890(0.773,1.024)	4.16E-02	0.905(0.823,0.996)	1.96E-01	-0.00673	9.69E-69	T
Diverticulitis	39848	4.24E-02	0.923(0.854,0.997)	1.30E-01	0.864(0.716,1.041)	1.22E-01	0.926(0.839,1.021)	8.89E-02	0.939(0.874,1.010)	4.46E-01	0.00343	5.72E-123	T
COPD	19071	4.55E-02	1.113(1.002,1.237)	6.94E-02	1.269(0.987,1.632)	3.76E-01	1.065(0.926,1.225)	9.25E-02	1.083(0.987,1.189)	2.67E-01	-0.00674	3.97E-68	T

Table 5.1. Significant findings of mendelian randomisation study of telomere length. The table presents MR estimates across five MR methods. Each line represents the association results for a single disease. All nominally significant results (P≤0.05) are highlighted in green.

My MR results confirmed some previous findings from association studies of TL and detected several new potentially causal associations that have not been reported previously.

I observed nominally significant increased risk with shorter TL for two cardiovascular phenotypes in the MR analysis: CAD (OR=1.13 [95%CI:1.05-1.23]) and aortic valve stenosis (OR=1.43 [95%CI:1.13-1.80]). Previous studies have reported shorter telomeres in both individuals with CAD (Weischer *et al.*, 2012; Codd *et al.*, 2013; Haycock *et al.*, 2014, 2017; Hunt *et al.*, 2015; Madrid *et al.*, 2016; Zhan *et al.*, 2017) and aortic valve stenosis (Kurz *et al.*, 2006). My results share the same direction, although with a smaller magnitude of risk increase in CAD per 1 SD of TL shortening. This reduced effect size may represent the lower disease prevalence in a cohort study when compared to case-control analyses, or a better estimate of the risk due to using an improved genetic instrument.

Perhaps surprisingly, decreased risk of hypertension (OR=0.92 [95%CI:0.86-0.99]) and venous thromboembolism (OR=0.91 [95%CI:0.83-0.99]) were nominally associated with shorter TL. These associations are not concordant with the results of previous observational studies and will be discussed further in the chapter *5.3.1. The causal effect of telomere length on cardiovascular diseases* together with the limitations of MR that might have contributed to this result.

I estimated an increase in risk of rheumatoid arthritis (OR=1.29 [95%CI:1.07-1.56]) and increase in risk of sarcoidosis (OR=1.67 [95%CI:1.17-2.38]) per 1 SD shorter TL. I also showed increased risk for COPD (OR=1.11 [95%CI:1.00-1.24]), bronchiectasis (OR=1.34 [95%CI:1.11-1.62]) and tuberculosis (OR=1.38 [95%CI:1.11-1.71]). All these phenotypes are associated with immune function.

I estimated a Bonferroni significant increase in hypothyroidism risk (OR=1.34 [95%CI:1.17-1.53]) and a nominally significant increase in hyperthyroidism (OR=1.40 [95%CI:1.09-1.81]). Thyroid related phenotypes were not studied in detail with a focus on TL previously, and this finding suggested a novel, potentially important, causal association between TL and endocrine diseases.

I showed a decreased risk of skin cancer, including melanoma, (OR=0.72 [95%CI:0.61-0.85]), lymphoma (OR=0.60 [95%CI:0.47-0.78]), lung cancer (OR=0.55 [95%CI:0.43-0.71]), leukemia (OR=0.47 [95%CI:0.33-0.68]), thyroid cancer (OR=0.34 [95%CI:0.22-

0.54]), uterine polyps (OR=0.74 [95%CI:0.64-0.86]), uterine fibroid (OR=0.63 [95%CI:0.53-0.76]), benign prostatic hyperplasia (OR=0.72 [95%CI:0.61-0.86]) and breast cyst (OR=0.77 [95%CI:0.65-0.92]), all of which survived Bonferroni correction.

I showed a decreased risk of melanoma (OR=0.69 [95%CI:0.53-0.90]), non-Hodgkin lymphoma (OR=0.71 [95%CI:0.54-0.92]), kidney cancer (OR=0.64 [95%CI:0.44-0.93]), brain cancer (OR=0.41 [95%CI:0.23-0.75]), ovarian cyst (OR=0.83 [95%CI:0.72-0.95]), prostate cancer (OR=0.74 [95%CI:0.58-0.94]), haemorrhoids (OR=0.87 [95%CI:0.78-0.96]), endometriosis (OR=0.85 [95%CI:0.73-0.98]), colorectal polyps (OR=0.90 [95%CI:0.81-0.99]) and benign breast lump (OR=0.80 [95%CI:0.66-0.99]), all of which were nominally significant.

I found that estimates for most associations showed consistency when using MR-IVW, MR-Egger, MR-median, and MR-RAPS methods (**Table 5.1**). Significant associations share similar size and direction of effects.

I noted that shorter TL in MR associations had a clear division between diseases that develop due to tissue degeneration and loss of function, and diseases with high proliferative potential that have excessive tissue growth. Potential causal roles of TL in associated disease groups will be discussed in more detail in the following chapters.

5.3. Causal role of telomere length in age-related diseases

Human aging is associated with a loss of immune functions that are crucial for protection against infections and malignancies. With age, the human immune system becomes senescent and gradually loses its ability to maximise protection against inflammation. With an inability to minimise the inflammatory damage and persistent stress signalling, the inflammation may become chronic and stress the immune system by affecting cellular turn-over through accelerating cell replication, which may exhaust the cell's replicative potential and increase the risk of developing age-related diseases. Telomere length as a marker of functional integrity of telomeres is tightly connected to cell longevity and is likely to represent both healthy and pathological aging (Hohensinner *et al.*, 2011; Jose *et al.*, 2017; Yeh *et al.*, 2019).

5.3.1. The causal effect of telomere length on cardiovascular diseases

One hallmark of pathological aging is atherosclerosis, a disease characterised by intense immunological activity that involves the formation of lesions in the arteries that are marked with inflammation, accumulation of lipids and cell death (Hansson *et al.*, 2006, 2011; Yeh *et al.*, 2019). Atherosclerosis is the most common underlying cause of cardiovascular diseases such as coronary artery disease and peripheral artery disease.

Telomere length was associated with the risk of cardiovascular diseases in previous studies that assessed the association using different approaches such as observational studies, meta-analyses, GRS and MR (previous **Tables 2.1** and **4.1**).

It is assumed that the most reliable and robust estimates are obtained from MR studies. MR circumvents residual confounding that plagues traditional observational studies when attempting to estimate a causal association. My MR findings were generally consistent with previous reports of shorter TL being a risk factor for cardiovascular and inflammatory age-related diseases.

I observed that a genetic predisposition to shorter TL increases the risk of CAD and aortic valve stenosis in my investigation of polygenic scores (chapter *4.8.4.1. Shorter telomeres and cardiovascular diseases*). Shorter TL was previously shown to be associated with CAD (Weischer *et al.*, 2012; Codd *et al.*, 2013; Haycock *et al.*, 2014, 2017; Hunt *et al.*, 2015; Madrid *et al.*, 2016; Zhan *et al.*, 2017). For example, shorter telomeres were observed in coronary endothelial cells of atherosclerosis patients when compared to healthy individuals (Minamino *et al.*, 2002; Yeh *et al.*, 2019). The observed effect of TL with aortic valve stenosis is also consistent with previous reports (Kurz *et al.*, 2006; Blunder *et al.*, 2018).

My MR results support the hypothesis that shorter telomeres may be a contributing factor to premature cellular senescence, tissue degeneration and loss of function. A one SD increase in shorter TL was associated with a 13% increase in risk of CAD (OR=1.13 [95%CI:1.05-1.23]) and with a 43% increase in risk of aortic valve stenosis (OR=1.43 [95%CI:1.13-1.80])) (**Table 5.2**).

		IVW		Egger's	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Coronary artery disease	36974	1.83E-03	1.133(1.047,1.226)	1.69E-01	1.145(0.946,1.386)
Aortic valve stenosis	2603	2.44E-03	1.427(1.134,1.797)	4.34E-01	1.249(0.719,2.169)
Hypertension	293487	3.72E-02	0.924(0.859,0.995)	1.14E-02	0.793(0.667,0.943)
Venous thromboembolism	19170	4.13E-02	0.911(0.832,0.996)	7.36E-01	1.038(0.837,1.286)
		Median		**RAPS**	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Coronary artery disease	36974	3.96E-02	1.116(1.005,1.239)	6.86E-03	1.118(1.031,1.212)
Aortic valve stenosis	2603	6.17E-01	1.092(0.773,1.543)	1.02E-02	1.373(1.078,1.750)
Hypertension	293487	3.41E-07	0.859(0.810,0.911)	1.37E-02	0.919(0.860,0.983)
Venous thromboembolism	19170	1.05E-01	0.890(0.773,1.024)	4.16E-02	0.905(0.823,0.996)
		Egger's Intercept		**Steiger**	
Phenotype	**N(Case)**	**P-value**	**Beta**	**P-value**	**Direction**
Coronary artery disease	36974	9.03E-01	-0.00056	1.28E-117	T
Aortic valve stenosis	2603	6.04E-01	0.00687	2.95E-08	T
Hypertension	293487	6.15E-02	0.00792	2.64E-268	T
Venous thromboembolism	19170	1.96E-01	-0.00673	9.69E-69	T

Table 5.2. Results of mendelian randomisation study of telomere length and cardiovascular diseases.

However, the MR estimate for the causal association is only nominally significant (**Figure 5.8**), suggesting a need to confirm results using either a stronger genetic instrument when larger GWAS of TL becomes available or a larger number of cases to increase the precision and statistical power. This would help not only to replicate the finding, but also increase the reliability of estimates.

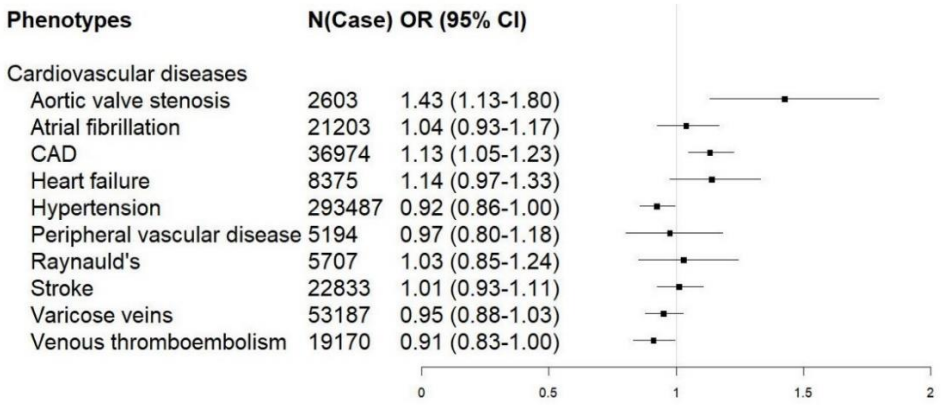


Figure 5.8. Inverse-variance weighted mendelian randomisation estimates of telomere length effect on cardiovascular diseases. The effect size is given as an odds ratio (OR) on X-axis, where OR=1 indicates no effect, OR<1 indicates a protective effect of shorter TL, and OR>1 indicates increase in risk with shorter genetic TL. The point estimate is denoted with black squares, and 95% confidence intervals with horizontal lines. A wide 95% CI suggests lack of precision in the estimate.

The GRS chapter, *4.8.4.1. Shorter telomeres and cardiovascular diseases,* covered the potential biological roles of telomere length in degenerative diseases, highlighting the idea of shorter telomeres limiting cell division to protect against cancer. However, a predisposition to shorter telomeres may also limit the regenerative capacity of a tissue, which may lead to its disfunction. Here, I tested the causal role of TL and identified nominal associations to support the idea that a genetic predisposition to shorter TL may contribute to accelerated telomere shortening and promote inflammation and ageing, increasing the risk of cardiovascular outcomes (O'Donovan *et al.*, 2011).

I also obtained results from the causal association analysis that were inconsistent with previous reports. Shorter TL was causally associated with a protective effect against hypertension and venous thromboembolism in MR analysis. Many previous observational studies reported shorter telomeres to be associated with an increased risk of hypertension (Demissie *et al.*, 2006; Lung *et al.*, 2008; Bhupatiraju *et al.*, 2012). Some authors suggested that hypertension may precede telomere shortening along with other factors such as inflammation and oxidative stress (Chiu *et al.*, 2016; Rietzschel *et al.*, 2016; Prasad *et al.*, 2017). Bi-directional effects may exist between TL and chronic inflammation, a common marker of cardiovascular disease (Yeh *et al.*, 2019).

Hypertension is a common health condition in an aging population, with high prevalence in UK Biobank (62.91%). Limitations of analysing hypertension were discussed in chapter *4.8.4.2. Shorter telomeres and hypertension* and included a look at the quality of the phenotype definition. I described how there is a high probability that the younger control group will develop hypertension later in life, and how the genetic variants for TL are explaining only a small of TL variance. The MR result for shorter TL having a protective effect against hypertension should also be taken with caution, because, although it is statistically significant, the upper limit of the confidence interval is close to one suggesting that these results could be due to chance. Further investigations are required to account for these concerns of disease definition, overlap with other diseases, distribution of age between cases and controls, medication against hypertension and other cofactors.

A decrease in risk of venous thromboembolism for every 1 SD decrease of genetic TL was found, suggesting that shorter TL is protective for this phenotype. Venous thromboembolism was nowhere near as common as hypertension in UKB, only 4.11%

of UKB participants were diagnosed with this condition. However, the same study limiting factors may influence the reported MR result: 1) Participants with venous thromboembolism may have other health conditions that could bias the association, 2) Some controls may develop this condition in future, 3) The detected causal associations did not survive Bonferroni correction, are of only nominal significance suggesting these results could be due to chance.

Nonetheless, the observed difference between observational and genetic estimates for TL highlights the need to investigate further within both observational and genetic studies under a high degree of scrutiny. The relationship between TL and inflammation, that contributes to development of cardiovascular disease, with the underlying immune functions should be explored in further work with consideration of traditional cardiovascular risk factors.

5.3.2. The causal effect of telomere length on immune-related diseases

Telomere shortening has previously been associated with the promotion of inflammation and immunosenescence (O'Donovan *et al.*, 2011; Jose *et al.*, 2017). Bi-directional influences may exist, according to observational studies. Telomere shortening may cause inflammation and inflammation may promote telomere shortening (Prasad *et al.*, 2017). The use of genetic variants as a surrogate for TL, on the other hand, can only have an effect in one direction, because genetic information cannot be altered by environmental factors. The genetic instrument can be considered as the ability to maintain telomere length. The shorter the physical TL the fewer divisions a cell can make. The shorter the genetic TL the poorer the capacity to maintain sufficient TL for the cell to function and divide. Both physical and genetic TL have an impact on a tissue's potential to renew.

MR investigation detected causal associations between shorter TL and an increased risk of several immune-related diseases. It must be noted that although reported diseases fall into different groups according to primary assignment by affected body part or type of disease, they are characterised by specific inflammatory processes and immune response as described below.

I estimated a 29% (OR=1.29 [95%CI:1.07-1.56]) increased risk of rheumatoid arthritis for a 1 SD decrease in TL (**Table 5.3**). Rheumatoid arthritis is characterised by a large accumulation of inflammatory cells. The arthritis estimate from MR is consistent with previous reports that showed shorter TL in peripheral T-cells in rheumatoid arthritis patients (Schönland *et al.*, 2003; Fujii *et al.*, 2009; Firestein, 2013). The previous causal association MR analysis of TL with arthritis reported a null result (Haycock *et al.*, 2017) and an increasing risk (Zeng *et al.*, 2020). The latter study used 7 SNPs and performed the analysis in a set of 911 patients and 2498 controls, estimating a 1 SD decrease in TL with a 47% (OR=1.47 [95%CI:1.14 -2.08] increase in risk of rheumatoid arthritis (Zeng *et al.*, 2020). Here sample size is much larger, and the estimated effect size is thought to be more precise, although confidence intervals are wide in both analyses.

		IVW		Egger's	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Bronchiectasis	3907	2.07E-03	1.343(1.113,1.620)	1.26E-02	1.810(1.155,2.837)
Tuberculosis	2952	3.17E-03	1.380(1.114,1.708)	3.15E-01	1.304(0.781,2.178)
Sarcoidosis	1536	4.77E-03	1.666(1.169,2.375)	2.72E-02	2.649(1.145,6.132)
Rheumatoid arthritis	8907	7.81E-03	1.290(1.069,1.557)	2.02E-01	1.350(0.857,2.128)
Eczema	17142	1.01E-02	1.185(1.041,1.348)	5.79E-01	1.093(0.799,1.496)
COPD	19071	4.55E-02	1.113(1.002,1.237)	6.94E-02	1.269(0.987,1.632)
		Median		**RAPS**	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Bronchiectasis	3907	8.44E-02	1.277(0.967,1.686)	6.02E-03	1.320(1.083,1.609)
Tuberculosis	2952	6.69E-03	1.577(1.135,2.193)	5.62E-03	1.399(1.103,1.775)
Sarcoidosis	1536	4.61E-02	1.591(1.008,2.513)	1.99E-03	1.680(1.209,2.334)
Rheumatoid arthritis	8907	1.21E-02	1.283(1.056,1.558)	2.28E-02	1.218(1.028,1.444)
Eczema	17142	5.03E-02	1.159(1.000,1.344)	3.96E-02	1.149(1.007,1.311)
COPD	19071	3.76E-01	1.065(0.926,1.225)	9.25E-02	1.083(0.987,1.189)
		Egger's Intercept		**Steiger**	
Phenotype	**N(Case)**	**P-value**	**Beta**	**P-value**	**Direction**
Bronchiectasis	3907	1.58E-01	-0.01539	1.02E-13	T
Tuberculosis	2952	8.14E-01	0.00291	1.01E-09	T
Sarcoidosis	1536	2.38E-01	-0.02422	3.07E-03	T
Rheumatoid arthritis	8907	8.30E-01	-0.00235	4.19E-26	T
Eczema	17142	5.84E-01	0.00412	3.08E-57	T
COPD	19071	2.67E-01	-0.00674	3.97E-68	T

Table 5.3. Results of mendelian randomisation study of telomere length and immune-related diseases.

I estimated a 67% (OR=1.67 [95%CI:1.17-2.38]) increased risk of sarcoidosis with shorter TL. This immune disorder has been previously associated with shorter TL in sarcoidosis patients when compared to controls (Maeda *et al.*, 2009; Georgin-Lavialle *et al.*, 2010).

However, this previous study was observational and could not consider a causal association.

Genetically shorter TL was causally associated with an increased risk of Chronic Obstructive Pulmonary Disease (COPD) – 11% increase in risk (OR=1.11 [95%CI:1.00-1.24]), bronchiectasis – 34% (OR=1.34 [95%CI:1.11-1.62]), and tuberculosis – 38% (OR=1.38 [95%CI:1.11-1.71]) all for a 1 SD decrease in TL (**Table 5.3**). Although none of them would survive Bonferroni correction, these associations are suggestive of a relationship between TL and immunity. COPD is a disease with persistent airflow limitation and is associated with chronic inflammation in the airways and the lungs (Vestbo *et al.*, 2013). Several TL related SNPs in the *TERT* region on chromosome 5 are associated with COPD (Van Moorsel *et al.*, 2018). The role of telomeres in cellular senescence that contributes to COPD has been reported previously (Albrecht *et al.*, 2014; Birch *et al.*, 2018).

I estimated an 18.5% (OR=1.19 [95%CI:1.04-1.35]) increased risk of eczema with 1 SD shorter TL. Although the confidence interval is relatively wide, this potentially causal relationship is a novel association. Previously, only Haycock *et al*., 2017, analysed genetic TL and atopic dermatitis and found no association (OR=1.05 [95%CI:0.88-1.24]) (Haycock *et al.*, 2017), likely due to the small number of genetic instruments used (13 SNPs) and the lower number of cases (10,788 subjects) in comparison to our 52 genetic instruments and 17,142 individuals. Atopic dermatitis, or eczema, is a chronic inflammatory skin condition that usually starts early in life. Immune dysregulation is suggested to be a causal factor for skin inflammation (Bang *et al.*, 2001). Increased telomerase activity and shorter measured TL was reported in T lymphocytes of atopic dermatitis patients when compared to healthy controls. This suggests that the immune system in eczema is constantly stimulated with increased cellular turnover (Wu *et al.*, 2000).

Overall, shorter TL showed a causal association with an increased risk of age-related diseases linked with a dysregulated immune response and chronic inflammation. Shorter TL may trigger the DNA damage response earlier (after fewer cell divisions) which may lead to persistent signalling of inflammation and the promotion of senescence.

5.3.3. The causal effect of telomere length on endocrine diseases

I identified a causal association with two endocrine diseases, hypo- and hyperthyroidism. I estimated a Bonferroni significant OR=1.34 [95%CI:1.17-1.53] increase in hypothyroidism risk and nominally significant OR=1.40 [95%CI:1.09-1.81] increase in hyperthyroidism risk per 1 SD shorter TL. The confidence intervals are wide, and the results should be considered with some caution until replicated in an independent dataset (**Table 5.4**). My findings represent a potentially new causal association of TL with thyroid diseases, the function of which was previously associated with senescence and longevity (Moreno-Navarrete *et al.*, 2018).

		IVW		Egger's	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Hypothyroid	29377	2.53E-05	1.339(1.169,1.534)	6.62E-02	1.372(0.986,1.908)
Hyperthyroid	5565	9.09E-03	1.402(1.088,1.806)	2.15E-01	1.484(0.802,2.745)
		Median		**RAPS**	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Hypothyroid	29377	8.22E-08	1.399(1.237,1.582)	1.00E-03	1.235(1.089,1.401)
Hyperthyroid	5565	3.25E-03	1.483(1.141,1.929)	4.02E-02	1.287(1.011,1.639)
		Egger's Intercept		**Steiger**	
Phenotype	**N(Case)**	**P-value**	**Beta**	**P-value**	**Direction**
Hypothyroid	29377	8.74E-01	-0.00125	3.29E-80	T
Hyperthyroid	5565	8.43E-01	-0.00293	2.50E-14	T

Table 5.4. Results of mendelian randomisation study of telomere length and endocrine diseases.

Both hyper- and hypothyroidism were observed to share the same direction of effect in MR estimates, despite being conflicting diseases of high and low thyroid response. However, this may be explained by a large overlap in cases between these two phenotypes. This is potentially due to individuals undergoing treatment for hyperthyroidism who then go on to develop hypothyroidism (Sheehan *et al.*, 2016). Previously, both diseases have been shown to associate with magnesium deficiency (Moncayo *et al.*, 2015), an important cellular ion involved in many processes, which itself has a potential role in telomere regulation via an effect on telomerase activity (Shah *et al.*, 2014). This suggests a biologically meaningful role of TL in the biological pathway of these diseases.

One can hypothesise that there is a potentially relevant biological relationship between the level of thyroid hormones, the immune response, inflammation and senescence in which TL plays a part. The thyroid stimulating hormone has previously been associated with senescence and longevity (Moreno-Navarrete *et al.*, 2018) as well as TL. MR results suggest that the direction of causal association is from TL to thyroid function, where shorter TL increases the risk of developing hypo- and hyperthyroidism. The exact pathway is yet to be established, but could be related to a deficiency of magnesium, an essential ion for nucleotide synthesis, the process that was linked to TL (Li *et al.*, 2020). Short TL, paired with a reduced ability of nucleotide synthesis and addition of telomere repeats, could result in a limited capability to keep thyroid tissue functional. Another possible pathway is from short telomeres through inflammation, early cell senescence and tissue exhaustion that may lead to disease. The identified causal associations are biologically plausible under these hypotheses, but the functional follow-up of this work is beyond the scope of this project.

The capability of shorter TL to limit cell division is not unique to endocrine disease and is more prominent in the relationship between TL and cancers.

5.3.4. The causal effect of telomere length on cancers and proliferative diseases

Telomeres allow a certain number of divisions before a cell becomes senescent and undergoes apoptosis. The length of the telomere is a limiting biological factor that was proposed as an evolutionary trait, developed to supress potential malignant formations (Young *et al.*, 2018). It was noted that the trade-off between short and long telomeres exist (Stone *et al.*, 2016) with shorter TL increasing the risk of cardiovascular diseases and longer TL increasing the risk of cancers. I saw this effect in the GRS analyses when considering just the genetically determined part of TL, and now confirmed this finding in the causal analyses using MR.

MR results demonstrated a strong protective effect of shorter TL on different types of cancer and highly proliferative phenotypes (**Table 5.1**). The protective effect of shorter TL against cancers is much stronger in comparison to the estimated effect on cardiovascular and immune-related diseases. For example, I estimated a 45% (OR=0.55 [95%CI:0.43-0.71]) decrease in risk of lung cancer for 1 SD shorter TL in comparison to

an estimated 13% (OR=1.13 [95%CI:1.05-1.23]) increase in risk of CAD for a 1 SD decrease in TL.

MR findings are consistent with previous studies that show an association with shorter TL and reduced cancer risk with both observed TL and a TL instrument used in MR (Haycock *et al.*, 2017; Zhang *et al.*, 2017). The significant MR results were generally consistent between the published MR from Haycock *et al*., 2017, and our published data (**Table 5.5**), with minor exceptions that could be attributed to an insufficient number of cases (testicular cancer) or possible pleiotropy in the genetic instruments (ovarian cancer). Haycock and colleagues reported longer TL, based on 10-13 SNPs, increasing the risk of cancers, and I have flipped the estimated ORs and 95% CIs for easier comparison of shorter TL effects. We reported shorter TL, based on 52 SNPs, to decrease the risk of cancer. Estimates were reported using an OR, for example, there was 28% decrease in risk of skin cancer for 1 SD decrease in TL. Significant associations from both studies show the consistency of effect direction that shorter TL is causally associated with a decreased risk of cancers. The estimates differ in magnitude between studies potentially due to different number of cases, and an increased number of genetic variants used in this study.

Study	**Haycock *et al*., 2017**				**Li *et al*., 2020**			
Cancer	**N(Case)**	**SNPs**	**OR(95%CI)**	**P-value**	**N(Case)**	**SNPs**	**OR(95%CI)**	**P-value**
Ovarian cancer	972	13	0.23 (0.13-0.42)	<0.001	1686	52	0.79 (0.57-1.11)	0.179
Lung cancer	3442	13	0.31 (0.24-0.42)	<0.001	3298	52	0.55 (0.43-0.71)	<0.001
Bladder cancer	1601	10	0.46 (0.27-0.76)	0.003	3017	52	0.81 (0.62-1.05)	0.107
Skin cancer	12814	13	0.53 (0.44-0.65)	<0.001	23320	52	0.72 (0.61-0.85)	<0.001
Testicular cancer	986	11	0.57 (0.33-0.98)	0.04	884	52	1.62 (0.98-2.67)	0.060
Kidney cancer	2461	12	0.65 (0.45-0.93)	0.02	1732	52	0.64 (0.44-0.93)	0.019

Table 5.5. Comparison of two mendelian randomisation studies of telomere length and cancer.

Previous observational studies that investigated TL in cancer tissues, or in blood, with the risk of various cancers have produced inconsistent results. Both shorter and longer TL have been significantly associated with an increased risk of cancers (Ma *et al.*, 2011; Jacobs *et al.*, 2013). The disagreement between the observational studies could be due to residual confounding, technical variation caused by measuring TL in different tissues,

using different measurement techniques or biological factors such as measuring TL at different cancer stages.

Some studies restricted TL measurements to LTL and quantitative PCR techniques for consistency, but the observed results remained inconsistent between different cancers. Then TL was demonstrated to differ in various cancer types depending on cell division rates within specific tissues, such as, different tumours (Zhu *et al.*, 2016). TL was observed to be shorter in tumours than in normal tissues (Barthel *et al.*, 2017), but was also found to be shorter at early cancer stages and longer during late stage cancer (Thriveni *et al.*, 2019). It might be the case that the rate of TL shortening is increased during the first stage of cancer due to unlimited cell replication. This may then be reversed to rapid lengthening during later cancer stages by reactivating telomerase or TL elongation by the ALT pathway (Zhdanova *et al.*, 2016; Apte *et al.*, 2017).

If genetic instrument is considered as a marker of TL maintenance, then we may consider that maintaining short TL might limit the capacity to replicate, whilst longer TL maintenance may improve the ability to maintain functional TL for longer, allowing more mutations to occur and promote cancer.

My results suggest a potentially causal association between shorter TL and lower risk of various cancers and proliferative diseases (**Table 5.6**). For example, shorter TL was significantly associated with decreased risk of skin cancer (including melanoma), consistent with previous meta-analysis (Caini *et al.*, 2015). TL was also associated with lymphoma and leukemia, which is consistent with previous GRS association analyses (Machiela *et al.*, 2016; Ojha *et al.*, 2016), and with lung (Cao *et al.*, 2019) and kidney cancers, which is consistent with a previous MR study (Haycock *et al.*, 2017). I also found a new potentially causal association between shorter TL and decreased risk of thyroid cancer that has not been reported before.

Table 5.6. Results of mendelian randomisation study of telomere length and proliferative diseases.

		IVW		Egger's		Median		RAPS		Egger's Int		Steiger	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**Beta**	**P-value**	**Direction**
Uterine fibroid	20551	8.37E-07	0.632(0.526,0.758)	7.95E-04	0.455(0.296,0.701)	2.11E-04	0.728(0.615,0.861)	3.64E-04	0.744(0.632,0.875)	1.09E-01	0.01691	4.26E-43	T
Thyroid cancer	756	4.09E-06	0.341(0.216,0.539)	6.73E-02	0.342(0.111,1.053)	1.72E-03	0.335(0.169,0.664)	4.65E-06	0.334(0.209,0.534)	9.95E-01	-0.00016	3.95E-01	T
Lung cancer	3298	6.17E-06	0.552(0.426,0.714)	8.13E-02	0.566(0.303,1.060)	1.27E-02	0.641(0.452,0.909)	1.25E-04	0.584(0.443,0.769)	9.29E-01	-0.00135	2.22E-07	T
Leukemia	1624	4.85E-05	0.473(0.329,0.678)	1.09E-03	0.222(0.095,0.520)	1.49E-04	0.396(0.245,0.639)	3.06E-04	0.502(0.345,0.730)	6.14E-02	0.03875	9.61E-03	T
Uterine polyps	14186	6.87E-05	0.740(0.639,0.858)	9.41E-03	0.613(0.429,0.874)	8.78E-07	0.638(0.534,0.763)	2.76E-04	0.754(0.648,0.878)	2.57E-01	0.00973	1.04E-41	T
Skin cancer	23320	6.89E-05	0.718(0.610,0.845)	1.02E-03	0.504(0.343,0.741)	3.84E-05	0.727(0.624,0.846)	1.11E-11	0.740(0.678,0.807)	5.38E-02	0.01807	3.11E-58	T
Lymphoma	3948	8.45E-05	0.604(0.470,0.777)	2.26E-02	0.480(0.261,0.885)	4.99E-07	0.469(0.349,0.630)	1.13E-05	0.572(0.446,0.734)	4.24E-01	0.01173	3.26E-09	T
Benign prostatic hyperplasia	19759	3.03E-04	0.724(0.608,0.863)	1.65E-03	0.498(0.330,0.751)	5.05E-03	0.781(0.657,0.928)	1.40E-02	0.812(0.688,0.959)	5.50E-02	0.01923	5.82E-46	T
Breast cyst	7199	2.81E-03	0.773(0.652,0.915)	8.50E-02	0.693(0.461,1.043)	7.52E-02	0.810(0.641,1.022)	1.60E-02	0.809(0.680,0.961)	5.70E-01	0.00559	2.86E-23	T
Brain cancer	922	3.90E-03	0.412(0.226,0.752)	3.72E-02	0.206(0.048,0.875)	8.33E-01	0.929(0.470,1.838)	1.65E-01	0.659(0.366,1.187)	3.06E-01	0.03587	9.48E-01	T
Melanoma	5362	4.53E-03	0.690(0.534,0.892)	5.54E-02	0.536(0.288,0.999)	1.70E-03	0.650(0.497,0.851)	4.34E-05	0.696(0.585,0.828)	3.88E-01	0.01289	3.87E-12	T
Ovarian cyst	14110	6.90E-03	0.827(0.721,0.949)	1.36E-01	0.772(0.553,1.079)	7.14E-02	0.855(0.721,1.014)	1.68E-01	0.908(0.792,1.041)	6.61E-01	0.00352	4.99E-46	T
Haemorrhoids	12795	7.59E-03	0.868(0.782,0.963)	7.75E-02	0.794(0.618,1.020)	4.92E-02	0.855(0.732,0.999)	7.03E-03	0.859(0.770,0.959)	4.50E-01	0.00456	9.42E-49	T
Non-hodgkin lymphoma	2641	9.35E-03	0.705(0.541,0.918)	1.12E-01	0.588(0.309,1.119)	1.15E-03	0.546(0.379,0.786)	4.14E-03	0.673(0.514,0.882)	5.48E-01	0.00927	7.41E-07	T
Prostate cancer	8962	1.28E-02	0.739(0.583,0.938)	1.00E-01	0.610(0.342,1.088)	7.47E-03	0.714(0.558,0.914)	3.75E-03	0.709(0.563,0.895)	4.77E-01	0.00983	6.81E-18	T
Kidney cancer	1732	1.92E-02	0.642(0.443,0.930)	1.49E-03	0.230(0.098,0.542)	5.28E-04	0.442(0.278,0.701)	8.45E-03	0.595(0.405,0.876)	1.30E-02	0.05222	4.66E-03	T
Endometriosis	7753	2.39E-02	0.845(0.730,0.978)	3.27E-02	0.676(0.476,0.959)	9.76E-03	0.749(0.601,0.932)	6.11E-02	0.863(0.739,1.007)	1.75E-01	0.01148	2.25E-27	T
Colorectal polyp	34018	3.22E-02	0.895(0.809,0.991)	8.82E-03	0.719(0.567,0.911)	1.03E-02	0.862(0.769,0.965)	4.56E-02	0.900(0.812,0.998)	5.21E-02	0.01127	1.06E-101	T
Benign breast lump	4076	3.75E-02	0.804(0.655,0.987)	2.60E-01	0.748(0.454,1.232)	4.89E-02	0.759(0.577,0.999)	4.59E-02	0.808(0.655,0.996)	7.58E-01	0.00370	1.15E-12	T
Diverticulitis	39848	4.24E-02	0.923(0.854,0.997)	1.30E-01	0.864(0.716,1.041)	1.22E-01	0.926(0.839,1.021)	8.89E-02	0.939(0.874,1.010)	4.46E-01	0.00343	5.72E-123	T

The association of TL with diseases that exhibit high proliferative potential confirms the hypothesis of TL being a limiting factor that restricts excessive cell replication. In my MR analysis, shorter TL is causally associated as protective against several phenotypes that exhibit tissue overgrowth: cancers, breast and ovarian cysts, uterine polyps and fibroids, and benign prostatic hyperplasia.

It must be noted that longer TL is unlikely to be the only factor that may lead to proliferative disease. One of the MR causal estimates, between TL and kidney cancer, showed the potential presence of nominally significant pleiotropy (MR-Egger's Intercept P=0.013). This indicates that another pathway may exist between the genetic instrument and the disease other than through tested intermediary, TL. Moreover, the association of TL with cancers may not be linear and both the extremes of short and long telomeres may increase the risk of malignancies (Cheng *et al.*, 2017; Haycock *et al.*, 2017). It was previously shown that patients with telomerase syndromes, with extremely short telomeres, have an increased risk for certain malignancies (Armanios *et al.*, 2012).

To sum up, I observed that shorter TL is causally associated with a decreased risk of cancers and diseases with high proliferative potential. Both genetic and physical TL serve as factors limiting cell replication and protecting against malignancies.

5.4. Potential biological mechanisms of telomere length in age-related diseases

I observed a clear division between causal associations of TL with diseases, exhibiting tissue degeneration and loss of function, with diseases of excessive cell replication and tissue overgrowth. Both disease types show accelerated cell division, loss of TL repeats and signalling of DNA damage, all of which initiate processes that replace dysfunctional and senescent cells and renew the damaged tissue. However, in cardiovascular, immune, and endocrine diseases the inflammation often persists and results in cell exhaustion and loss of tissue function. In proliferative diseases the resulting outcome is not only cell exhaustion, but uncontrolled cell division that increases the probability of malignant transformation by accumulation of mutations.

Shorter telomeres may drive cells to reach the Hayflick limit faster, where the telomere becomes critically short, and enter premature senescence that promotes pro-

inflammation. If the clearance of senescent cells is not sufficient, they may induce senescence in neighbouring cells. Inflammation may progress and this could trigger cellular proliferation (Kong *et al.*, 2013; Victorelli *et al.*, 2017). Stem cells may be exhausted by a high need in tissue repair, namely high proliferative turn-over of cells, the disability of accumulated senescent cells to keep the tissue functional, and tissue degeneration and loss of function (Ruzankina *et al.*, 2007; Haycock *et al.*, 2017; Victorelli *et al.*, 2017).

Shorter telomeres, inducing senescence, limit the proliferative capacity of the cells and supress tumours (Ruzankina *et al.*, 2007; Hanahan *et al.*, 2011; Kong *et al.*, 2013). By limiting cell replication, shorter TL decreases the risk of cancers and other highly proliferative phenotypes. However, it should be noted that malignancies may also originate from cells with critically short telomeres, which may be explained by non-linear associations of measured TL and cancer risk (Cheng *et al.*, 2017). Cells with critically short telomeres may overcome senescence and divide further, which may introduce genomic instability such as chromosome fusions and further DNA alterations leading to cancer initiation (Kong *et al.*, 2013). Individuals with telomerase syndromes, where extremely short telomeres are reached early in life, show an increased risk of malignancies, but incidence of cancer in such patients is low, and degenerative diseases account for the majority of mortality (Armanios *et al.*, 2012). TL has been shown to be shorter in tumours compared to normal tissues (Barthel *et al.*, 2017), which may indicate that some cancers initiate at shorter TL when genomic instability is easier to acquire. Then telomerase is reactivated, and malignant cells can elongate the telomeres indefinitely.

Longer telomeres are known to accumulate more DNA damage, where a packed telomere structure prevents efficient repairs, and signalling of DNA damage may persist for a long time. With no repair this may contribute to genomic instability and cancer formation (Victorelli *et al.*, 2017).

It is not clear how exactly genetic TL affects these biological processes, but my results suggest that shorter TL along with inherited or measured TL has the potential to limit cell replication and protect against genomic instability.

The presented MR analyses suggest that TL has a potentially causal role in the prevalence of many age-related diseases. Shorter telomeres were causally associated

with an increased risk of degenerative diseases including cardiovascular, immune-related, and endocrine diseases. Where there is a decreased risk of diseases with high proliferative capacity such as cancers. The effect size is stronger for proliferative diseases and points to the important potential for TL to limit excessive cell replication.

If we consider genetic TL as the cell's ability to maintain telomere length, we may hypothesise, given the same inherited and environmentally affected physical TL, that genetic TL may set the initial capacity for maintaining the cell's lifespan. In such a case, the genetic predisposition to shorter TL would have stricter genetic rules for mechanisms that control and limit cell replication. The genetic predisposition to longer TL, on the other hand, may employ mechanisms to maintain telomeres at the longest possible length in order to keep the cell viable and allow longer lifespan. The exact underpinnings of the biological pathways, that 52 genetic determinants of TL are involved in, need to be investigated in the future.

5.5. Mendelian randomisation study limitations

The MR analysis performed has several limitations that should be taken into account when interpreting the results and comparing to other studies.

The first limitation is in the selection of genetic variants used as the MR instruments. This was restricted to results from a single large-scale meta-analysis, the ENGAGE study (Li *et al.*, 2020), which is the largest published GWAS of TL to date. These were restricted further due to some SNPs not being available in UK Biobank because of imputation differences and quality control of the genetic variants. However, unlike previous studies I investigated a much larger instrument with 52 genetic variants, making this the most up-to-date and largest investigation of causal associations between TL and age-related diseases.

Here disease definition treats all diseases as independent. In reality, a single individual is highly likely to have several disease phenotypes in their lifetime and would therefore contribute to multiple analyses as a case. I did not select controls to contain individuals with no known diseases, sometimes considered as super-controls. As such, individuals with other diseases would act as a control for any disease they did not have. In case of cancers and degenerative diseases, which have the observed opposite effects this may

reduce the estimated effect size towards the null. Moreover, diseases analysed here are just a subset and an individual may have had a health condition that was not recorded or not analysed due to an insufficient number of cases. I estimated that there were approximately only 5% of UK Biobank individuals with none of the diseases being investigated here.

A third limitation would be the lack of adjustment for confounding. Whilst confounding in MR is not possible in the traditional sense, as the genetic instrument cannot be altered, it is accepted as appropriate to remove the effect of some risk factors from the estimation of the β for selected diseases. However, due to the nature of investigating a broad spectrum of over 120 diseases it was not possible to adjust each disease for all possible covariates that have been previously reported as disease's specific risk factors. To counter this, I performed an extensive investigation of pleiotropy, where confounding would be detected.

The instrument consisted of 52 selected genetic determinants of TL. These only account for a small proportion of the variance explained in TL with r^2<3% (Li *et al.*, 2020). As such the instrument is not overly powerful. Additional factors such as an individual's inherited TL at birth, along with environmental impacts, including stress, may also lead to accelerated telomere shortening and affect health. I was unable to account for many of these factors.

The final limitation is due to the resource being used. The risks of various diseases were estimated within UK Biobank (Marchini, 2015; UK Biobank, 2015), and, although large and incredibly data rich, this dataset is known to be not ideally representative of the general population due to the tendency of participants within UK Biobank to be more conscious of their health, and thus relatively healthier. UK Biobank participants could be considered as the worried healthy and as such generalisations are more challenging to other populations.

5.6. Potential use of genetic telomere length

The difference in genetic predisposition to specific TL on the organismal level is useful to estimate not only the potential risks of developing disease, but also the time when disease is likely to develop. For example, a genetic predisposition to longer TL was

associated with an increased risk of cancers, but the age of onset may be different depending on individual genotypes, sex, age, ethnicity, and additional disease specific cofactors. For this reason, I took this investigation further to perform time-to-event analyses. Using methods of survival analysis, I planned to estimate the impact of genetic TL on the time to disease development, utilising the genetic risk score for shorter telomeres. These methods will be explored in the next chapter.

Chapter 6. Survival analysis and genetically determined telomere length

In this chapter I introduce the basic concepts of time-to-event analyses, more commonly known as survival analysis, and its application in this project to test the effects of genetically determined telomere length (GDTL) on the age of disease onset. I am going to present significant associations with GDTL and discuss the potential for GDTL to predict time to disease.

6.1. Introduction to survival analysis

Here I cover the questions, goals and problems addressed by survival analysis, define the outcome variables, survival times and censoring strategy. I am going to present survival and hazard functions, the most commonly used methods, study designs and how to interpret the estimated effects.

6.1.1. Survival analysis background

Survival analysis or time-to-event analysis is a statistical approach to analyse the length of time to occurrence of an event of interest. It is widely used in medical and epidemiological research, where it can explore the time to an event between groups of patients or, as I will apply, the effects of a continuous trait on the time to event.

The main aim of time-to-event analysis is to estimate the likelihood of an event in a group of patients. Calculating the proportion of patients that experienced the event after a certain period of time is a simple solution. But it would require all patients to be in the study for the same length of time, which hardly corresponds to the realities of clinical research, where patients may enter the study at different time points, may be lost to follow-up or will die from other causes. Survival analysis uses censoring to take different follow-up times into account, which distinguishes it from other types of analyses (Flynn, 2012).

To investigate survival time, we need to define what we classify as an event. In epidemiological research the event is usually a binary outcome and has two states, the event has or has not occurred. An event could be defined as an occurrence of death,

onset or diagnosis of disease or other change in health state. The event needs to be well-defined and easily observable. Patient death is an unambiguous end point and is a primary choice for survival analysis. CAD, via myocardial infarction or surgical intervention, is an event that requires hospitalisation and can also be classified as a clinically relevant and unambiguous event suitable for time-to-event analysis. Cancer events may be more poorly defined because it does not occur instantaneously. The incidence of cancer may not be diagnosed for some time and the exact event time of occurrence may be unknown (Kleinbaum *et al.*, 2005, 2012; Moore, 2016; Schober *et al.*, 2018).

With a well-defined event we also need to define what we mean by time, the **time** until the event occurs, or the study ends. We could use years, months or days from the beginning of the study or, for a genetic analysis, we may define time as the age of an individual, modelling from age zero (Flynn, 2012; Kleinbaum *et al.*, 2012).

In survival analysis, we usually refer to an event as a failure, because we are usually interested in events such as death, hospitalisation, disease diagnosis or other negative experience. We refer to the time variable as survival time, as it provides the length of time that individual 'survived' before experiencing the event or until the study ends.

The length of follow-up should be sensibly chosen to ensure that a sufficient number of events are observed as it is the number of events that provide statistical power to these analyses. Some individuals will not experience the event until after the completion of the study and their survival times will be censored at study end. This situation is referred to as right censoring and is the most common censoring type. Survival times of censored patients are included in the estimation of survival probabilities before they are censored and excluded from analyses after the censoring time. Each individual enters the study event free and is considered at risk until they have the event or until censoring. The actual survival times are not observed for censored individuals (even if the event is experienced after the study end point), but censoring provides the knowledge that they survived up to the censoring time point and their survival time is greater that the observation time (Kleinbaum *et al.*, 2005, 2012; Flynn, 2012; Moore, 2016; Schober *et al.*, 2018).

Time-to-event analysis aims to answer questions related to time, such as:

- What is the median time to an event, or how long will an individual survive?
- How does the probability of an event change over time?
- What is the probability of surviving to a certain time point?

To answer these questions survival analysis estimates the effect sizes of variables that influence the time to an event of interest (Schober *et al.*, 2018). The following chapter describes the formal definitions used, the estimation of effects and their interpretation.

6.1.2. Terminology and notation for survival analysis

We will consider genetically determined telomere length, estimated via a GRS, as the primary predictor, and aim to estimate the effect of the TL GRS on time to an event of interest or disease-free survival.

We encode disease as a binary event with 1 indicating that event occurred during the study time, and 0 otherwise. Diseases were defined previously in chapter 4.8.2. *Selecting phenotypes and assigning case-control status*. In addition, I analyse age at death, age at menopause and parental longevity. Given that each individual inherits their genetic variation at birth, this study timeline is age from birth until the end of follow-up in UK Biobank health records. Most individuals do not experience the disease and their participation time was censored due to any one of the following reasons:

- Study ended before the individual experiences the disease.
- The individual died from an event that is not our event of interest.

Censoring allows us to keep the censored individual's time at risk, because we know that up to the censoring point the individual was disease-free (Schober *et al.*, 2018).

For example, we may consider CAD, where disease status equals to one if CAD was diagnosed during the study time, and zero if the survival time was censored. Each individual's survival time we denoted by T that represents the individual's age. Any chosen value of time variable T we denote with t. If we were interested in evaluating whether a person with the genetic predisposition to shorter TL remains disease-free for 75 years, we would choose t=75, and then we estimate how probable it is that T >75?

To answer this question, we would need to consider two terms of survival analysis: the survival function, denoted $S(t)$, and the hazard function, denoted $h(t)$.

We define the survival function as:

$$\boldsymbol{S(t) = P(T > t), 0 < t < \infty}$$

Where *P* is the probability that the random variable *T* is greater than some specified time *t*. In other words, the survival function estimates the probability of surviving past a specific time point, *t*, or as the probability that the event has not occurred by time *T* (Kleinbaum *et al.*, 2005; Stevenson, 2007; Cole *et al.*, 2010; Moore, 2016; Schober *et al.*, 2018). When plotting the survival function, we might expect to see a smooth curve with *t* ranging from 0 to infinity. However, in practice we usually observe a step-stair shaped slope, where stairs down indicate that one or more events have occurred (**Figure 6.1**).

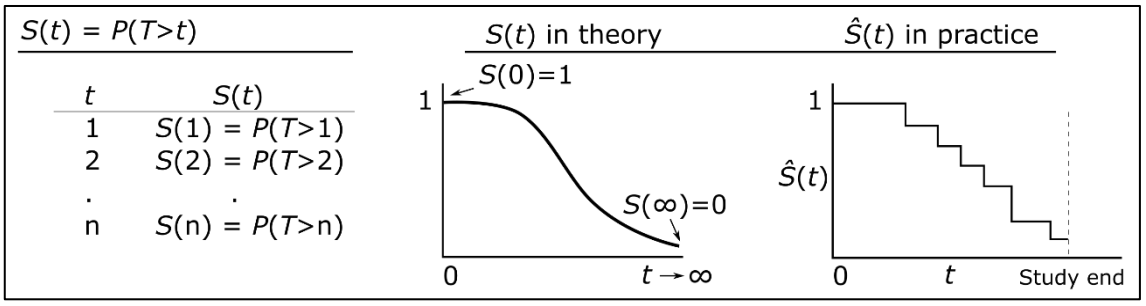


Figure 6.1. Survival function estimation and plotting. The survival function, *S(t),* is estimated at each time *t* as the probability that the event has not occurred by time *T*. The survival curve is smooth in an ideal analysis, while real examples will often show step-stair lines as events continue to occur.

The survival function has the following characteristics:

- The survival probability decreases as *t* increases – more events occur over time.
- The survival probability at time point zero equals 1, as no one has experienced the event: $t = 0$, $S(t) = S(0) = 1$.
- The survival curve will eventually fall to zero, if we allow the study period to increases without limit, as we assume that eventually everyone would experience the event: $t = \infty$, $S(t) = S(\infty) = 0$.

We define the hazard function as the instantaneous rate of events over time. The hazard at time t is a ratio of the probability of an event at time t, given that an individual survived to that time (Cole & Hudgens, 2010; Kleinbaum & Klein, 2005; Kleinbaum 2012):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Where Δt is the time interval, and P is the conditional probability of an individual's survival time, T, lying between t and $t+\Delta t$, given that that individual's survival time, T, is equal to or greater than t.

The hazard function is not a probability but a rate, the values of which range between 0 and infinity ($h(t) \geq 0$ and has no upper bound), depending on whether time is measured in days, months, or years. The hazard function can vary over different values of t. For example, a constant hazard shows that the instantaneous potential to become ill is the same throughout the entire study duration, and in other cases hazard can be increasing or decreasing with increasing time.

The survival and hazard functions are closely related and can be converted to each other. The survival function is commonly used to directly describe the survival of the study cohort. The hazard function is used as a measure of instantaneous potential and to identify a specific model that is suitable for analysis (Kleinbaum *et al.*, 2012; Schober *et al.*, 2018).

Using time-to-event analyses I aim to estimate the effect that describes the relationship between our exposure of interest and the outcome of interest adjusting for relevant confounders. The outcome is binary and the measure of effect is similar to an odds ratio, but called a Hazard Ratio (HR), which is expressed as the exponential of the regression coefficient from the model (Kleinbaum *et al.*, 2012). The HR has a similar interpretation to an OR: HR=1 means there is no effect, HR=2 means the hazard is double when compared to the reference group, and HR=0.5 means the hazard is half that of the reference group (Kleinbaum *et al.*, 2012; Schober *et al.*, 2018).

6.1.3. Survival analysis methods

Two methods are commonly used in survival analysis, the Kaplan-Meier (KM) Estimator (Kaplan *et al.*, 1958) and a Cox Proportional Hazards Model (Cox, 1972). Both were used in this project as descriptive or analytical methods to assess the time to disease onset.

The Kaplan-Meier Estimator is a simple technique developed for randomised controlled trials that considers the number of patients in the study at different time points and the number of events that have occurred by a specific time point (Kaplan *et al.*, 1958; Flynn, 2012).

The Kaplan-Meier Estimator is nonparametric, which does not make assumptions about the distribution of survival time, nor does it assume a specific relationship between covariates and the survival time. It estimates the unadjusted probability of survival past a certain time point:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{(r_j - d_j)}{r_j}, \qquad \text{for } 0 \leq \mathrm{t} \leq \mathrm{t}^{+}$$

Where time t_j, is estimated at j = 1, …, n event times and t^+ is the maximum event time, d_j is the number of events at time t_j and r_j is the number of individuals at risk at time t_j (Stevenson, 2007; Kleinbaum *et al.*, 2012; Schober *et al.*, 2018). An example of how the survival probability is calculated is provided in **Table 6.1**. This data shows 30 patients at t_0, the start of the study (n_0=30). Before t_1 2 events were observed (d_0=2) and 3 individuals were censored (w_0=3), making 27 individuals at risk between time point t_1 and t_0 with survival probability equal to 0.93.

Time	Start n_j	Event d_j	Censored w_j	At risk r_j	P(Survival) $P_j = (r_j - d_j) / r_j$
0	30	2	3	30-3=27	(27-2)/27=0.93
1	25	1	2	25-2=23	(23-1)/23=0.96
…	…	…	…	…	…
n					

Table 6.1. Example of calculating Kaplan-Meier survival estimates.

As we are using discrete time intervals KM survival curve uses a step function. In the case of this project, the baseline for all individuals was birth (**Figure 6.2 A**), while in

clinical studies participants may enter the study at different time points and aligning entry time points is required before ordering the data (**Figure 6.2 B**). However, while the calendar date of entry is unique for each individual in both cases, we set entry to be t_0 (**Figure 6.2 C**).

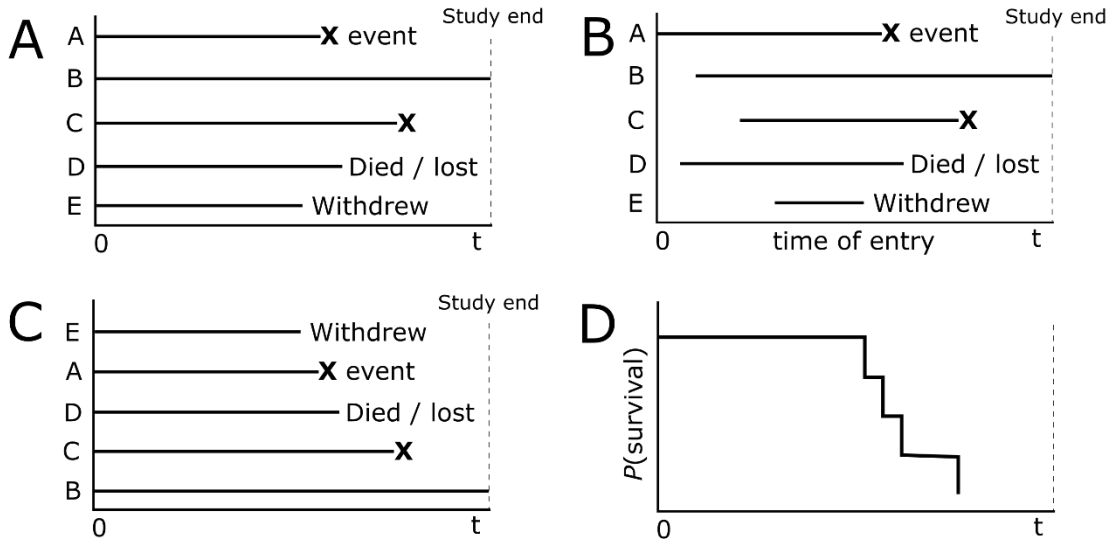


Figure 6.2. Visual representation of survival data. (A) shows timelines originating at the same baseline and (B) shows timelines with different study entry points, (C) shows data ordered by duration of follow-up and (D) the corresponding Kaplan-Meier survival curve.

Cox Proportional Hazards (PH) regression (Cox, 1972) is the most commonly used approach in survival analysis. It is a semiparametric method, with no assumptions about the distribution of survival time. Cox PH models assume a linear relationship between the covariates and the hazard function and models the hazard function, while it does not directly model survival probabilities or survival times (Flynn, 2012; Schober *et al.*, 2018). The hazard function for a Cox PH model can be written as:

$$h(t) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$$

Where $h(t)$ is the hazard at time t, and $h_0(t)$ is the baseline hazard that the Cox model does not estimate, i.e., when all variables X_1, X_2, X_p are equal to zero. The PH model assumes that all study participants have a common baseline hazard function that depends on time. The individual hazard function has a common baseline hazard multiplied by a time-independent function of an individual's covariates. As can be seen in the hazard function, the baseline hazard depends on time, but none of $X_1, X_2, \ldots, X_p$

interact with time *t*, meaning the model assumes that underlying effects of the predictor variables (or the ratio of the hazards) for any two individuals are proportional or constant over time (Gogtay *et al*., 2017; Schober 2018). The PH assumption can be checked using Schoenfeld residuals (Schoenfeld, 1980, 1982). The variable that violates PH can still be used in a Cox PH model using stratification by that variable or by fitting an interaction between the independent variable and time (Kleinbaum *et al.*, 2012).

Cox models with time-varying covariates are more difficult to interpret and potential inference from them is prone to errors (Schober *et al.*, 2018). For this reason, and because violations of PH are common, there are other models that can be applied to time-to-event data that model the baseline hazard.

6.2. Time-to-event study design of the project

I aimed to estimate the effects of genetically determined TL on time to event, which were disease onset (self-reported or from health records), death, menopause, and parental longevity, defined by age at death of the parent. Using TL GRS, the study start was birth, as genetic data is not altered through time. The end of the study, used for censoring, was the latest available date in health records, at the time of analysis, in UK Biobank. These depend on country of residence: 2016-02-29 for individuals living in England, 2015-02-14 – in Scotland, and 2016-03-01 - in Wales.

I used KM and Cox PH models to estimate the effect sizes and make predictions about survival. The concepts and methods used are described above in more detail, and the study design is summarised in **Figure 6.3.**

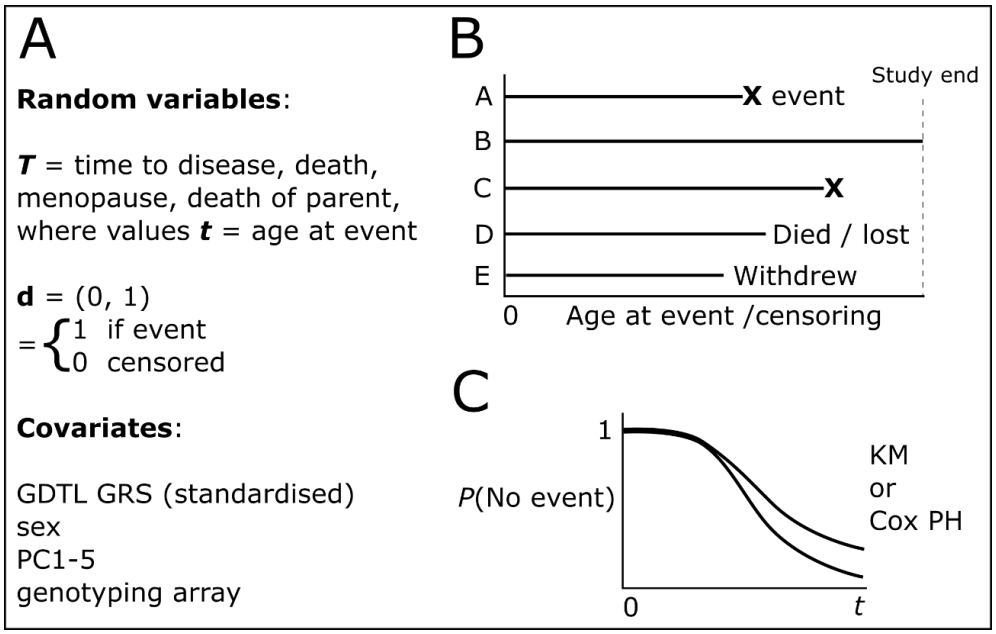


Figure 6.3. Time-to-event study design for analysing effect of genetically determined telomere length. (A) variables and covariates, (B) visual representation of events and censoring reasons, where X denotes the event, (C) graph of survival probability (not experiencing the event).

As with the previous analyses in chapters 4. *Genetic risk score for shorter telomeres* and 5. *Mendelian randomisation study of telomere length* I investigated 127 curated diseases. I also included death, menopause, and parental longevity as outcomes for these time-to-event analyses. TL has been previously associated with the risk of age-related diseases and longevity, and in this chapter I aimed to estimate the GDTL effect on time to disease. For example, GDTL for shorter telomeres was causally associated with higher risk of CAD, and time-to-event analysis may show that individuals with shorter GDTL develop CAD earlier in life, compared to individuals with longer GDTL.

6.2.1. Event definition

I defined disease onset as the first recorded occurrence of disease. For example, I selected individuals with a myocardial infarction (MI) using the corresponding hospital record date of admission in hospital episode data, or the self-reported in the UKB phenotypic file. Individuals that had a MI, but for whom the date of diagnosis was missing, were excluded. MI is a well-defined outcome for time-to-event analysis, as a specific event resulting in hospitalisation is generally recorded. Other diseases such as

tuberculosis, uterine fibroid and cancers have potentially inaccurate event dates, because these types of diseases are not instantaneous and may develop long before the time of diagnosis. However, within the scope of this project, I used time of diagnosis as the event time for all diseases selected.

6.2.2. Model strategy

I used the Kaplan-Meier Estimator to illustrate data trends and Cox PH to estimate the effect of GDTL and other covariates on time to event. The general Cox PH model, used for time to disease onset analyses, included the following predictors and covariates:

$$Event(time, status) \sim GRS + Sex + PC1 + PC2 + PC3 + PC4 + PC5 + BiLEVE$$

Where *time* is the age when the individual experienced the event or was censored, *status* is a binary outcome (1 – event or 0 – censored), *GRS* is the standardised score that I built from 52 genetic variants and that represents GDTL, *Sex* is a factor (female or male), *PC* are principal components (ranging from the first to fifth), and *BiLEVE* is the array used to genotype UKB data.

As I observed strong gender differences for several diseases, the analyses are stratified for sex, and a Cox PH model for time to disease onset was modified to:

$$Event(time, status) \sim GRS + Strata(Sex) + PC1 + PC2 + PC3 + PC4 + PC5 + BiLEVE$$

The estimated coefficient β can be interpreted as a change in the expected log of the hazard ratio relative to a 1 SD change in GDTL with all other predictors held constant. To additionally investigate the dose response relationship, the GRS distribution was divided into five quintiles, where quintile 1 corresponded to the group with the longest TL, and quintile 5 - the shortest TL. These are fit as a factor in another Cox PH model:

$$Event(time, status) \sim qGRS + Strata(Sex) + PC1 + PC2 + PC3 + PC4 + PC5 + BiLEVE$$

For all models I performed tests of the proportional hazard assumption. Data preparation were performed using Python programming language (version 2.7.5) and statistical analyses using R statistical package (version 3.5.1). Survival analysis was performed using *survival* (version 2.42.3) and *survminer* (version 0.4.7) R packages.

6.3. Genetically determined telomere length predicts time to disease

In this chapter I am going to present and discuss the results of association between GDTL and time to disease.

6.3.1. Overview of survival analysis results

To investigate whether shorter GDTL associates with early disease onset I performed time-to-event analyses using the TL GRS, described in chapter 4 *Genetic risk score for shorter telomeres*.

Using age as the time scale I found that GDTL is a statistically significant predictor for 37 diseases, 18 of which passed the Bonferroni threshold, $P \leq 0.05/127 = 3.937 \times 10^{-4}$ (**Figure 6.4**).

Shorter GDTL was associated with increased hazard, the probability of experiencing the disease per given time unit, of cardiovascular, immune, and inflammatory diseases. Longer GDTL was associated with decreased hazard of cancers and diseases with high proliferative potential. Full detailed time-to-event results showing all results can be found in **Appendix 6** *Time to disease onset results*. This chapter only focuses on the statistically significant associations between GDTL and the time to disease onset.

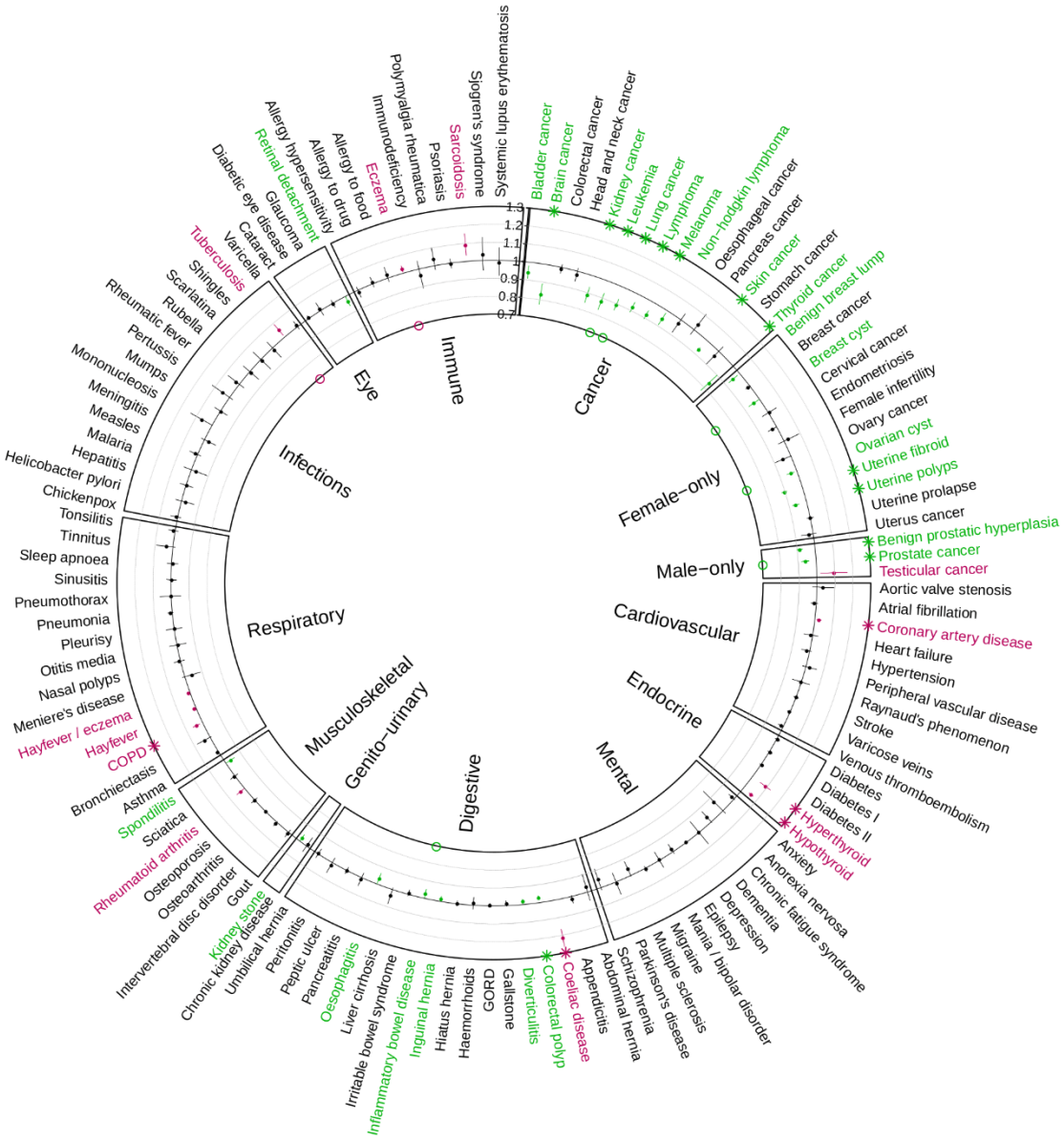


Figure 6.4. Shorter genetically determined telomere length association results for 127 diseases using time-to-event analysis. The outer circle shows disease phenotypes. All coloured phenotype names reach a nominal significance, where a star denotes that it passes Bonferroni correction. The forest circular plot shows estimated hazard ratios (HR). HR>1 (coloured red) indicates increased risk to experience disease earlier in life. HR<1 (coloured green) indicates decreased risk to experience disease earlier in life.

I observed shorter GDTL significantly associated with an increasing hazard for CAD, COPD, thyroid dysfunction and coeliac disease, whilst longer GDTL was associated with an increased hazard for cancers and diseases with high proliferative potential, which include female- and male-only diseases (**Table 6.2**).

Disease group	Phenotype	N(Event)	P-value	HR(95%)
Cardiovascular	Coronary artery disease	28272	1.95E-04	1.023(1.011,1.035)
Endocrine	Hypothyroid	22155	1.89E-15	1.055(1.042,1.070)
	Hyperthyroid	3958	1.23E-07	1.089(1.055,1.124)
Digestive	Coeliac disease	2402	4.54E-21	1.215(1.167,1.265)
	Colorectal polyp	28601	1.59E-06	0.972(0.960,0.983)
	Diverticulitis	22547	8.21E-03	0.982(0.969,0.995)
	Oesophagitis	3902	3.47E-02	0.966(0.936,0.998)
	Inflammatory bowel disease	5805	4.37E-02	0.974(0.949,0.999)
Genito-urinary	Kidney stone	8403	3.70E-02	0.977(0.956,0.999)
Musculoskeletal	Spondylitis	5486	1.93E-02	0.969(0.943,0.995)
	Rheumatoid arthritis	5978	2.95E-02	1.029(1.003,1.056)
	Inguinal hernia	20554	3.29E-02	0.985(0.972,0.999)
Respiratory	COPD	9866	2.51E-04	1.038(1.017,1.059)
	Hay fever / eczema	108546	4.33E-04	1.011(1.005,1.017)
	Hay fever	45780	1.37E-02	1.012(1.002,1.021)
Infections	Tuberculosis	2744	8.71E-03	1.052(1.013,1.092)
Immune	Sarcoidosis	1285	1.73E-03	1.092(1.034,1.155)
	Eczema	13178	7.75E-03	1.024(1.006,1.042)
Eye	Retinal detachment	4723	3.68E-02	0.970(0.942,0.998)
Cancer	Skin cancer	22402	1.34E-25	0.932(0.919,0.944)
	Melanoma	5124	3.03E-09	0.920(0.894,0.945)
	Lung cancer	2845	3.66E-09	0.894(0.861,0.928)
	Prostate cancer	7976	3.39E-08	0.939(0.919,0.961)
	Brain cancer	824	1.26E-07	0.830(0.775,0.889)
	Lymphoma	3635	9.02E-07	0.921(0.891,0.952)
	Thyroid cancer	708	2.60E-06	0.836(0.776,0.901)
	Leukemia	1333	3.10E-06	0.879(0.832,0.928)
	Kidney cancer	1538	4.40E-06	0.888(0.845,0.934)
	Bladder cancer	2857	2.25E-03	0.944(0.909,0.979)
	Non-Hodgkin lymphoma	2381	6.84E-03	0.945(0.908,0.985)
	Testicular cancer	845	9.47E-03	1.094(1.022,1.172)
Female-only	Uterine fibroid	14453	6.59E-39	0.896(0.881,0.911)
	Uterine polyps	10500	9.73E-11	0.938(0.920,0.956)
	Ovarian cyst	7853	9.63E-04	0.963(0.942,0.985)
	Benign breast lump	3641	3.26E-03	0.952(0.921,0.984)
	Breast cyst	3203	2.99E-02	0.962(0.929,0.996)
Male-only	Benign prostatic hyperplasia	13146	2.22E-24	0.914(0.899,0.930)

Table 6.2. Significant associations of genetically determined telomere length and 37 disease phenotypes in time-to-event analysis. All nominally significant associations are marked green (P<0.5). HR>1 indicates increased risk and HR<1 indicates decreased risk of developing disease earlier in life.

The estimated Hazard Ratio (HR) represents the change in hazard for every 1 SD shorter GDTL. My association results showed that a 1 SD shorter GDTL is associated with a 2.3% increase in hazard for CAD. There was no significant association found for the other cardiovascular phenotypes in time-to-event analyses.

There is a protective effect for uterine fibroid with an estimated 10.4% decrease in hazard for every 1 SD of GDTL shortening. This suggests that individuals with a 1 SD shorter GDTL have, on average, 10.4% fewer events of uterine fibroid at a particular time point.

The detected associations and estimated effect sizes from the time to disease onset analyses allow us to estimate the probability of disease occurring at a specific age for an individual based on their TL GRS having considered and adjusted for other covariates. Two examples, time to CAD and skin cancer onset, are going to be described in more detail in the following chapters.

6.3.2. Genetically shorter telomeres and earlier onset of cardiovascular diseases

The distinctive feature of this project is the use of GDTL to estimate the risk and hazard for age-related disease. Previously researchers have investigated measured TL, or a small number of genetic determinants of TL to assess the influence of TL on health outcomes. The design of many observational studies has focused on the difference between cases and controls, i.e., participants who already have cardiovascular disease compared to those who do not. For example, 890 patients with heart failure had shorter TL and were found to be at an increased risk for reaching adverse events of interest (van der Harst *et al.*, 2010).

It is possible to estimate the risk of an event from baseline over specific time intervals, when investigating survival in a case-control setting (van der Harst *et al.*, 2010; Zhang *et al.*, 2013; Haver *et al.*, 2015; Pusceddu *et al.*, 2018; Vecoli *et al.*, 2019). However, a single measurement of TL within patients cannot provide evidence for the causal direction. Disease may itself cause TL shortening and independently increase the risk of mortality making any identified TL associations potentially confounded by disease.

Other observational studies designed their analyses of TL on the time to incident disease such as atrial fibrillation and myocardial infarction in cohorts of elderly healthy

individuals, where only some experience the event of interest during follow-up (Østhus *et al.*, 2017; Siland *et al.*, 2017; Staerk *et al.*, 2017; Stefler *et al.*, 2018). These studies had inconsistent results and shorter TL was associated with both increased and decreased risk or was not shown to be significantly associated with time to disease outcome.

There are examples of other studies that used TL SNPs in time-to-event analyses, but no effect of GDTL on cardiovascular disease onset was found (Roberts *et al.*, 2014) or associations were detected only in women (Burnett-Hartman *et al.*, 2012).

In this project I investigated time to disease onset using a GRS of 52 genetic determinants of TL as the main predictor of interest. GDTL and CAD were found to be significantly associated indicating a 2.3% increase in CAD hazard over time for every 1 SD increase in TL GRS (HR=1.023 [95%CI:1.011-1.035]). This suggest that in addition to being at increased risk of CAD, individuals with shorter GDTL are predisposed to develop CAD earlier in life.

I also observed that males had ~3.5-fold higher risk of experiencing CAD in comparison to women. It is known that CAD occurs more prevalently in men, 74.83% of CAD cases in UKB. The distributions of age at onset of CAD between men and women, although of similar shape, show that men develop CAD on average two years earlier than women (**Figure 6.5**).

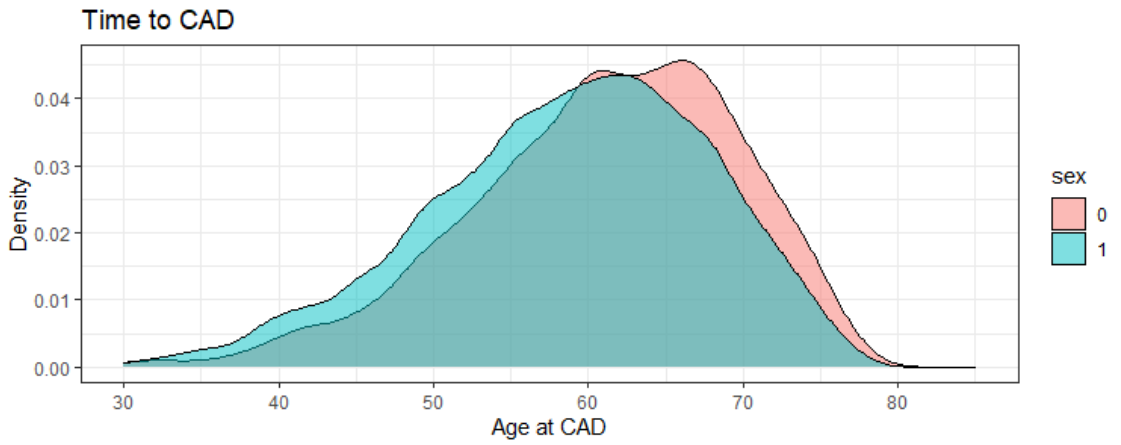


Figure 6.5. Distribution of age at diagnosis of coronary artery disease. Where 0 is a female, and 1 is a male.

Additionally, I performed a time-to-event analysis using GDTL quintiles. The results showed a dosage response relationship with the effect of shorter GDTL increasing the hazard of CAD (**Figure 6.6**). The significant effect is observable when comparing the extremes of the GDTL data, i.e., quintile 5 compared to the reference quintile 1. Comparison of extreme quintiles suggests that individuals predisposed to have the shortest GDTL according to TL GRS have a 7% increase in CAD hazard in comparison to individuals predisposed to have the longest GDTL.

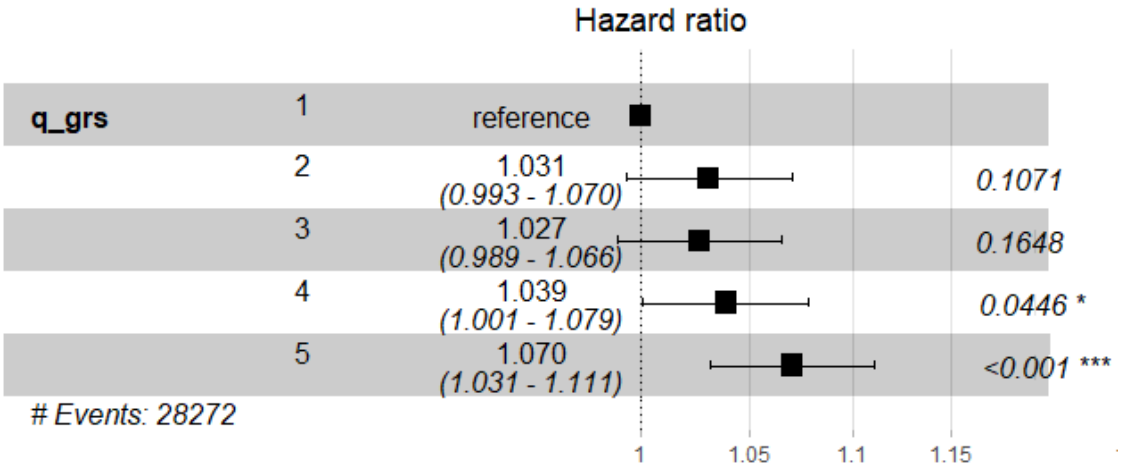


Figure 6.6. Quintiles of telomere length genetic risk score and dosage effect in association with time to coronary artery disease diagnosis. Forest plot shows the estimates for each quintile in the Cox PH model. Reference quintile 1 of TL GRS stands for predisposition to the longest GDTL and quintile 5 – to the shortest.

Time-to-event analysis provides us with the ability to estimate the probability of being CAD-free at a particular time point. It should be noted, however, that the effect of the TL GRS is rather small and the difference in probability of being disease-free when comparing TL GRS quintiles is only visible when examining older ages (**Figure 6.7**).

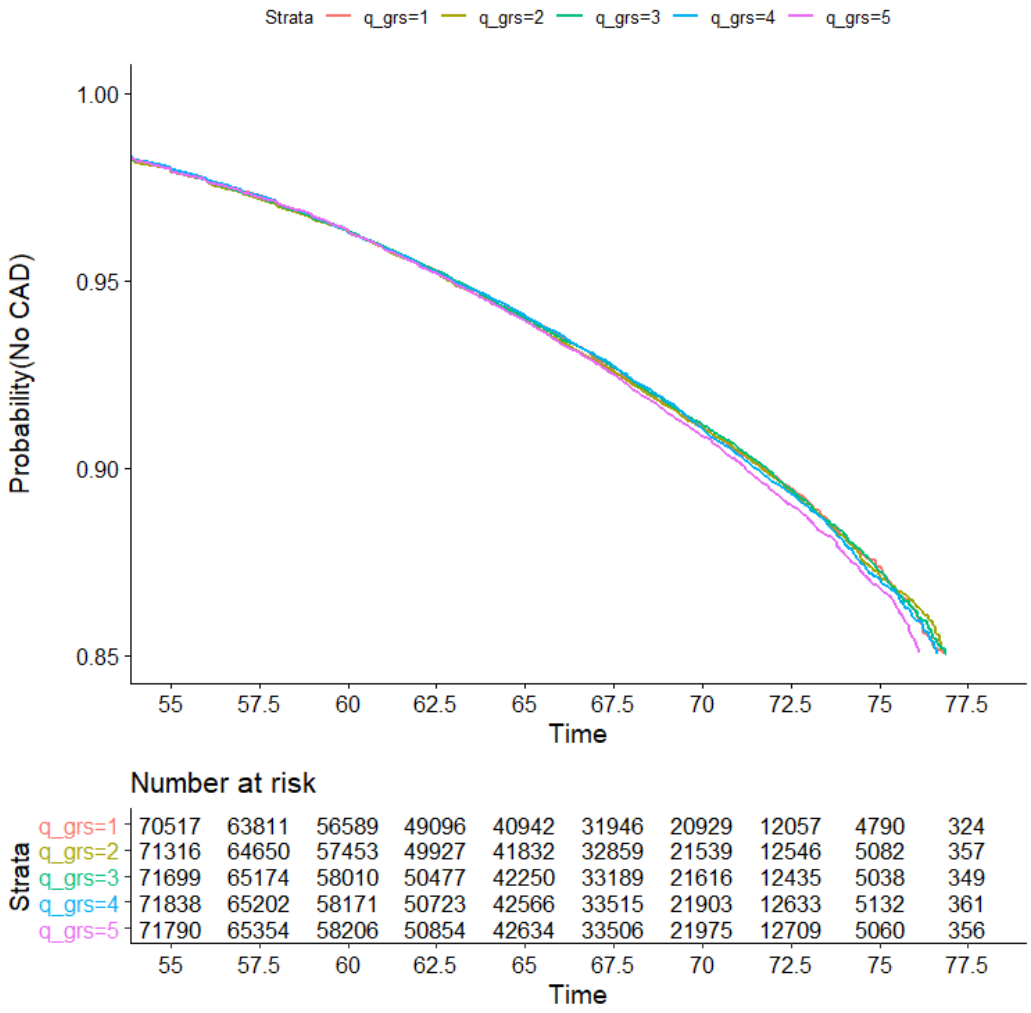


Figure 6.7. Change in the probability of being free of coronary artery disease over time depending on genetically determined telomere length quintile. The shortest GDTL (quintile 5) shows that probability of being CAD free decreases slightly faster and events occur earlier in life in comparison to the other quintiles.

One of the main interests of time-to-event analysis is to estimate the probability of staying disease-free for a specific individual, which can be modelled for any important variables and covariates. My Cox PH model suggests that at the population level the hazard of CAD is higher for individuals with shorter GDTL and higher for men. To make individual predictions we can use the model to determine risk. For example, using the data for four individuals in **Table 6.3**, we observe two men and two women with extreme TL GRS quintiles. They enrolled in the study at a certain age (column *age*) and at the end of the study (indicated by their age in column *time*) they had not experienced our event of interest, CAD, where *status*=0.

id	age	sex	time	std_grs	q_grs	S(55)	S(60)	S(65)	S(70)
1	41	1	48.54	-1.211	1	0.964	0.937	0.901	0.858
2	48	1	55.46	0.981	5	0.963	0.936	0.899	0.856
3	56	0	63.54	-1.347	1	0.992	0.985	0.974	0.958
4	46	0	53.54	1.002	5	0.992	0.985	0.974	0.957

Table 6.3. Example data for the prediction of time to coronary artery disease. id – unique identifier, age – age at recruitment, sex – gender (0 - female, 1 - male), time – age at event or censoring, std_grs – Z-standardised TL GRS, q_grs – quintile of std_grs (ranging from the longest, quintile 1, to the shortest, quintile 5), S(55), S(60), S(65), S(70) – probability of being CAD free at age 55, 60, 65 and 70, respectively.

With the built Cox PH model, we can estimate the probability that these individuals will be CAD-free when they are 55, 60, 65, and 70 years old. Visually, their CAD-free survival looks like a smooth curve (**Figure 6.8**), where we observe the probability of CAD-free survival decreasing more rapidly for men compared to women and decreasing faster for shorter GDTL in both women and men. The figure highlights how the effect of sex is much greater than the effect of GDTL over time. The probability of CAD-free survival for a man with the shortest GDTL is around 96.4% at age 55, 93.7% at age 60, 90.1% at age 65, and 85.8% at 70. The extrapolation to older ages is possible within the bounds of the model, but it should be noted that model precision is lower with data near the limits, as the sample size is much smaller.

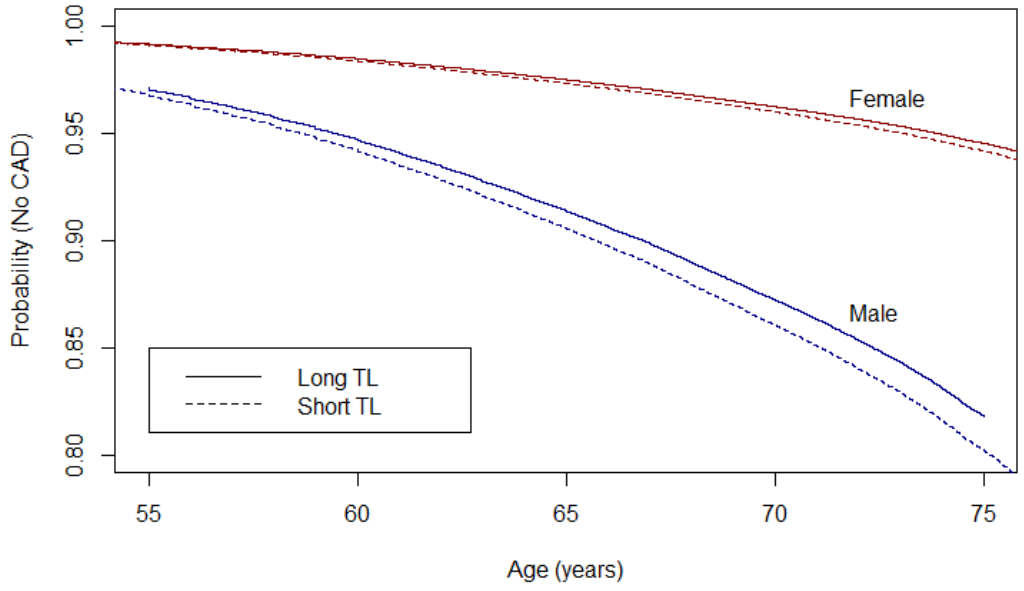


Figure 6.8. Prediction of probability of being free of coronary artery disease over time. The survival curves were estimated for the individuals shown in **Table 6.3**. The Cox PH model was stratified by sex.

Although GDTL is significantly associated with CAD risk, its effect on time to disease onset is relatively small. The GDTL used in this analysis explains only a small amount of TL variance. The TL GRS can be potentially expanded with more genetic determinants of TL. However, it is likely that the effect of genetic TL will not be greater than the effect of other known CAD risk factors, such as gender or modifiable exposures like smoking. It is also important to note that the effect of GDTL is not that prominent in cardiovascular and immune-related phenotypes in comparison to cancers and similar diseases, where TL serves as a factor limiting cell proliferation. To continue with this discussion we will consider cancer, with a focus on the association of GDTL with time to skin cancer onset.

6.3.3. Genetically longer telomeres and earlier onset of cancers

In the previous analyses using GRS I observed that shorter GDTL was strongly associated with a protective effect against most cancer types that I studied, some of which were potentially causal in the MR. The Cox PH models also found similar results. The effect of GDTL was found to be greater in cancer related disease phenotypes in comparison to cardiovascular and immune-related phenotypes. The diagnosis of cancer, used in this analysis, is not an acute event, rather, disease onset is the first occurrence in the health records, likely the date of first diagnosis.

Using time-to-event analyses I detected significant change in hazards between GDTL and 12 cancer types and between GDTL and 6 phenotypes that are linked with tissue overgrowth (**Table 6.4**).

Disease group	Phenotype	N(Event)	P-value	HR(95%CI)
Cancer	Skin cancer	22402	1.34E-25	0.932(0.919,0.944)
	Melanoma	5124	3.03E-09	0.920(0.894,0.945)
	Lung cancer	2845	3.66E-09	0.894(0.861,0.928)
	Brain cancer	824	1.26E-07	0.830(0.775,0.889)
	Lymphoma	3635	9.02E-07	0.921(0.891,0.952)
	Thyroid cancer	708	2.60E-06	0.836(0.776,0.901)
	Leukemia	1333	3.10E-06	0.879(0.832,0.928)
	Kidney cancer	1538	4.40E-06	0.888(0.845,0.934)
	Bladder cancer	2857	2.25E-03	0.944(0.909,0.979)
	Non-Hodgkin lymphoma	2381	6.84E-03	0.945(0.908,0.985)
	Head and neck cancer	2816	2.70E-01	0.979(0.943,1.016)
	Stomach cancer	817	3.08E-01	0.965(0.900,1.034)
	Pancreas cancer	843	3.14E-01	1.036(0.967,1.109)
	Oesophageal cancer	957	4.76E-01	0.977(0.916,1.042)
	Colorectal cancer	6280	4.95E-01	0.991(0.967,1.016)
Female-only	Uterine fibroid	14453	6.59E-39	0.896(0.881,0.911)
	Uterine polyps	10500	9.73E-11	0.938(0.920,0.956)
	Ovarian cyst	7853	9.63E-04	0.963(0.942,0.985)
	Benign breast lump	3641	3.26E-03	0.952(0.921,0.984)
	Breast cyst	3203	2.99E-02	0.962(0.929,0.996)
	Ovary cancer	1549	5.34E-02	0.952(0.905,1.001)
	Endometriosis	5199	5.51E-02	0.973(0.947,1.001)
	Breast cancer	16503	1.31E-01	0.988(0.973,1.004)
	Uterus cancer	2152	2.34E-01	0.974(0.934,1.017)
	Female infertility	1019	4.20E-01	1.026(0.964,1.092)
	Cervical cancer	1916	8.25E-01	1.005(0.961,1.052)
	Uterine prolapse	11709	9.48E-01	0.999(0.981,1.018)
Male-only	Benign prostatic hyperplasia	13146	2.22E-24	0.914(0.899,0.930)
	Prostate cancer	7976	3.39E-08	0.939(0.919,0.961)
	Testicular cancer	845	9.47E-03	1.094(1.022,1.172)
Digestive	Colorectal polyp	28601	1.59E-06	0.972(0.960,0.983)

Table 6.4. Associations of genetically determined telomere length and disease phenotypes with high proliferative potential using time-to-event analysis. All nominally significant associations are marked green. HR>1 indicates increased risk and HR<1 indicates decreased risk of developing disease earlier in life.

Focusing on skin cancer, I estimated a 6.8% decrease in hazard over time for every 1 SD increase in shorter GDTL. The model suggests that shorter GDTL is protective against early development of skin cancer in comparison to longer GDTL.

Additional time-to-event analyses using quintiles of GDTL showed a strong dosage effect where shorter GDTL is associated with a decreased hazard of skin cancer (**Figure 6.9**). This significant effect was observable for all four quintiles when compared to the reference quintile. A comparison of extreme quintiles suggests that individuals predisposed to the shortest GDTL, according to the built TL GRS, have a decrease in skin cancer hazard of ~18.4% when compared to individuals predisposed to the longest GDTL.

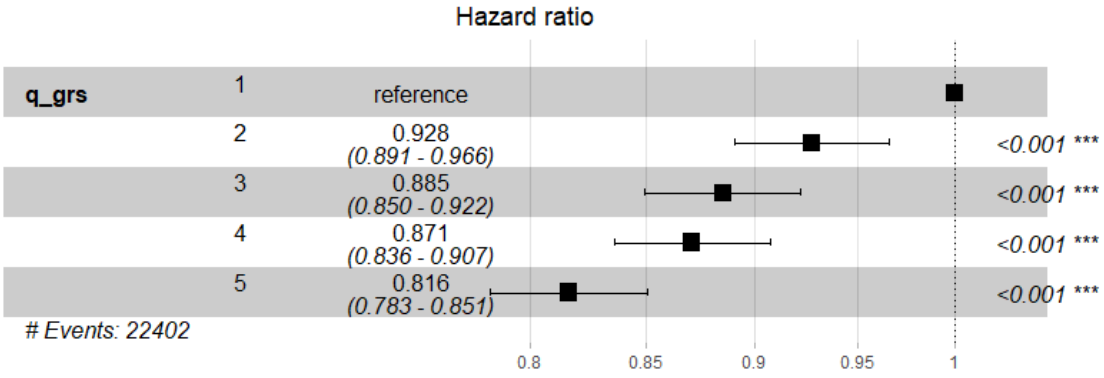


Figure 6.9. Quintiles of telomere length genetic risk score and the dosage effect on time to skin cancer diagnosis. Forest plot shows estimates of GRS quintiles from additional Cox PH model. Reference quintile 1 is a predisposition to the longest GDTL and quintile 5 the shortest.

Time-to-event analysis provides us with ability to estimate the probability of being skin cancer free at a specific time point (**Figures 6.10**).

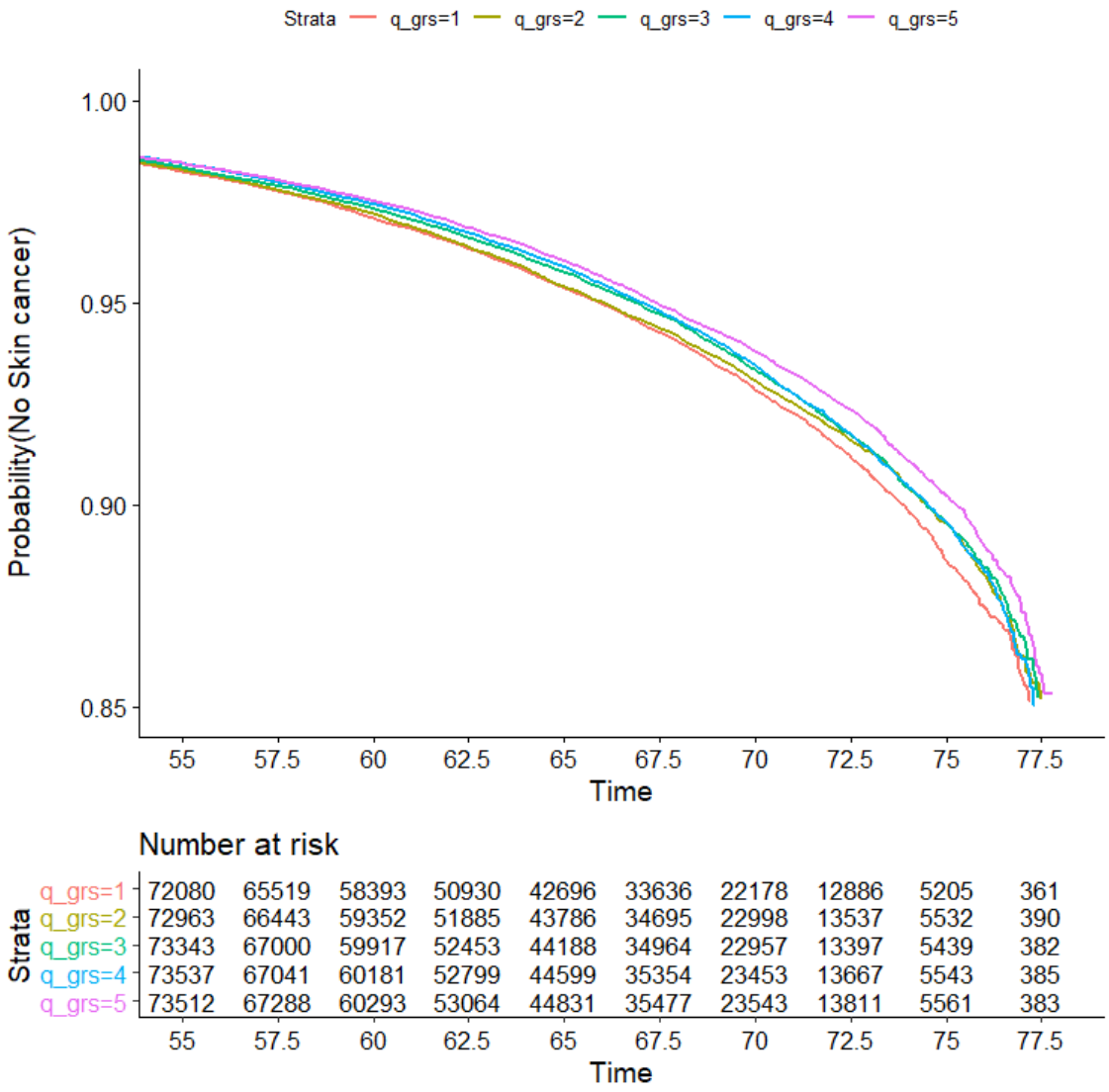


Figure 6.10. Change in the probability of skin cancer free survival depending on genetically determined telomere length quintile. The longer GDTL (quintile 1) shows that the probability of being skin cancer free decreases and events occur earlier in comparison to the shorter GDTL (quintile 5).

To estimate the probability of staying skin cancer free I used the results from the Cox PH model, which suggests that the hazard of skin cancer is higher for individuals with longer GDTL and, as for CAD, also higher for men. To make individual predictions I will again use an example for four individuals (**Table 6.5**), two men and two women with extreme TL GRS quintiles. They enrolled in the study at a certain age (column *age*) and by the time of study end (column *time*) they had not experienced skin cancer (column *status*).

id	age	sex	time	std_grs	q_grs	S(55)	S(60)	S(65)	S(70)
1	41	1	48.54	-1.211	1	0.985	0.974	0.955	0.925
2	48	1	55.46	0.981	5	0.986	0.976	0.959	0.933
3	56	0	63.54	-1.347	1	0.980	0.969	0.953	0.932
4	46	0	53.54	1.002	5	0.982	0.973	0.958	0.939

Table 6.5. Example data for the prediction of time to skin cancer. id – unique identifier, age – age at recruitment, sex – gender (0 - female, 1 - male), time – age at event or censoring, std_grs – Z-standardised TL GRS, q_grs – quintile of std_grs (ranging from the longest, quintile 1, to the shortest, quintile 5), S(55), S(60), S(65), S(70) – probability of being skin cancer free at age 55, 60, 65 and 70, respectively.

Based on the results from the Cox PH model we can estimate the probability of being skin cancer free for these individuals when they are 55, 60, 65 and 70 years old. As shown for CAD, visually, their probability of remaining skin cancer free looks like a smooth curve (**Figure 6.11**). I observed the probability of being skin cancer free decreasing more rapidly after an individual reaches 55 years old and decreasing slightly faster for longer GDTL and for men. The probability of survival to 55 years old without experiencing skin cancer for a man with the shortest GDTL is around 98.6%, to age 60 – 97.6%, to age 65 – 95.9%, and can be extended out to 70 – 93.3%.

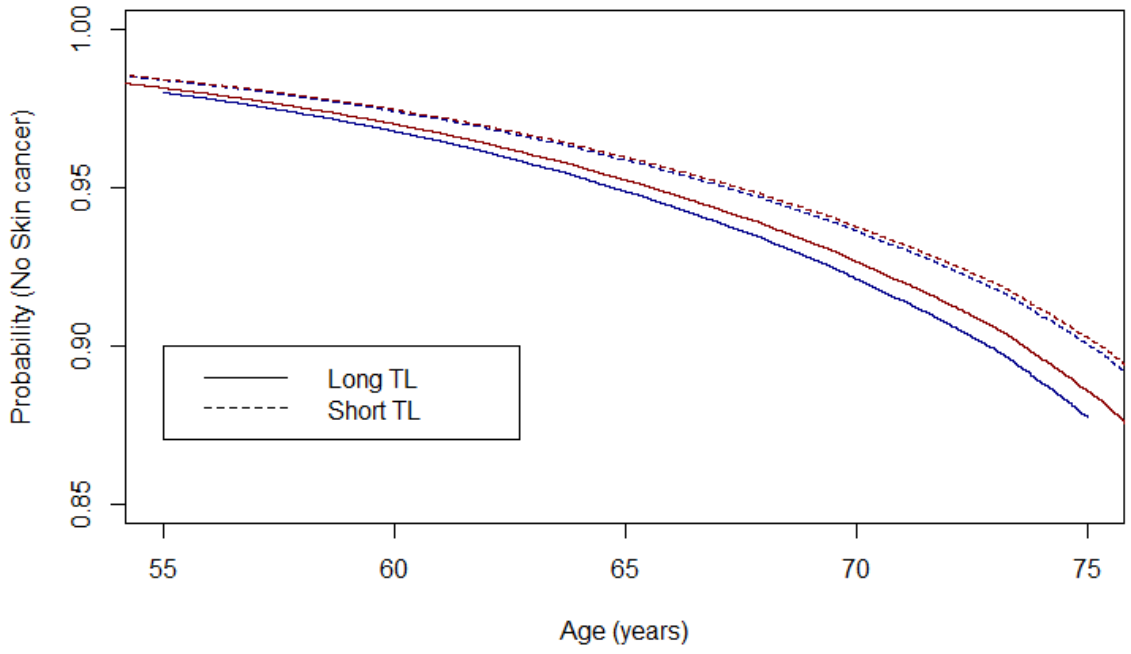


Figure 6.11. Prediction of probability of being skin cancer free over time. The survival curves were estimated for the participants shown in **Table 6.5**, where men are coloured blue and women red.

In the time-to-event Cox PH model for skin cancer I observed that GDTL has a greater effect relative to CAD. For skin cancer, the effect of sex is not overly strong, despite skin cancer being more frequent in males (Rubin *et al.*, 2020).

It must again be noted that cancer is not an acute event, and cancer onset comes with some assumptions. I used the date of cancer diagnosis as the outcome, which is only an estimation of the true cancer onset time. Thus, although I observed a significant result the estimates may be prone to bias due to this.

My approach, using GDTL, cannot be fully compared to previous observational studies that have used measured TL to report TL associations with time to cancer onset, as many studies have investigated survival following a cancer diagnosis.

When modelling the effect over time I have only considered the linear effect of GDTL. However, observed evidence of association fits pre-conceived ideas of GDTL effect direction, as it is consistent across multiple cancer phenotypes.

Overall, my results suggest that GDTL may be used as a predictor to estimate not only the risk of developing age-related diseases as was shown with GRS and MR, but also to estimate the probability of disease at a specific age. This can be considered as a representation of human healthspan that is no less important than longevity.

6.4. Telomeres and longevity

Average TL, measured in one tissue or cell line, can be considered to be an indicator of the biological age of an organism and, importantly, a predictor of longevity and mortality (Pusceddu *et al.*, 2018). At the organismal level, TL is suggested as a marker that reflects the current health status and the capacity to maintain genome integrity and functions (Lidzbarsky *et al.*, 2018).

Studies of TL effects on mortality were suggestive of shorter TL being linked to greater mortality (Epel *et al.*, 2009; Pusceddu *et al.*, 2018; Stefler *et al.*, 2018). This corresponds with the biological telomere attrition with age. The contribution of genetic predisposition to telomere length has not previously been studied with lifespan. Using the TL GRS I investigated the effects of GDTL on longevity within UK Biobank, and I will present the analysis and results in the following chapter.

6.4.1. Genetically determined telomere length and individual longevity

As several studies reported short TL to be associated with an increased risk of mortality in small observational studies, I hypothesised that short GDTL may have a similar effect. In other words, a genetic predisposition to shorter TL was thought to potentially contribute to acceleration of the decline in tissue function, which would lead to earlier death. Conversely, being predisposed to longer TL may increase longevity.

There were 27,009 death events available from the death registry in UK Biobank. I excluded any deaths that were due to external reasons, encoded with ICD10 codes starting from O to Z. For example, ICD10 codes starting with “O” relate to pregnancy or delivery and are not likely to be influenced greatly by TL genetics or be due to biological reasons. To minimise the noise in the model these deaths were excluded.

I used a Cox PH model included standardised TL GRS and found that the TL GRS was not significantly associated with longevity (P-value=0.0807).

Fitting TL GRS quintiles also showed no significant association. In **Figure 6.12** the survival curves for TL GRS quintiles intersect and may have different effects at various ages indicating that the TL GRS violates the PH assumption.

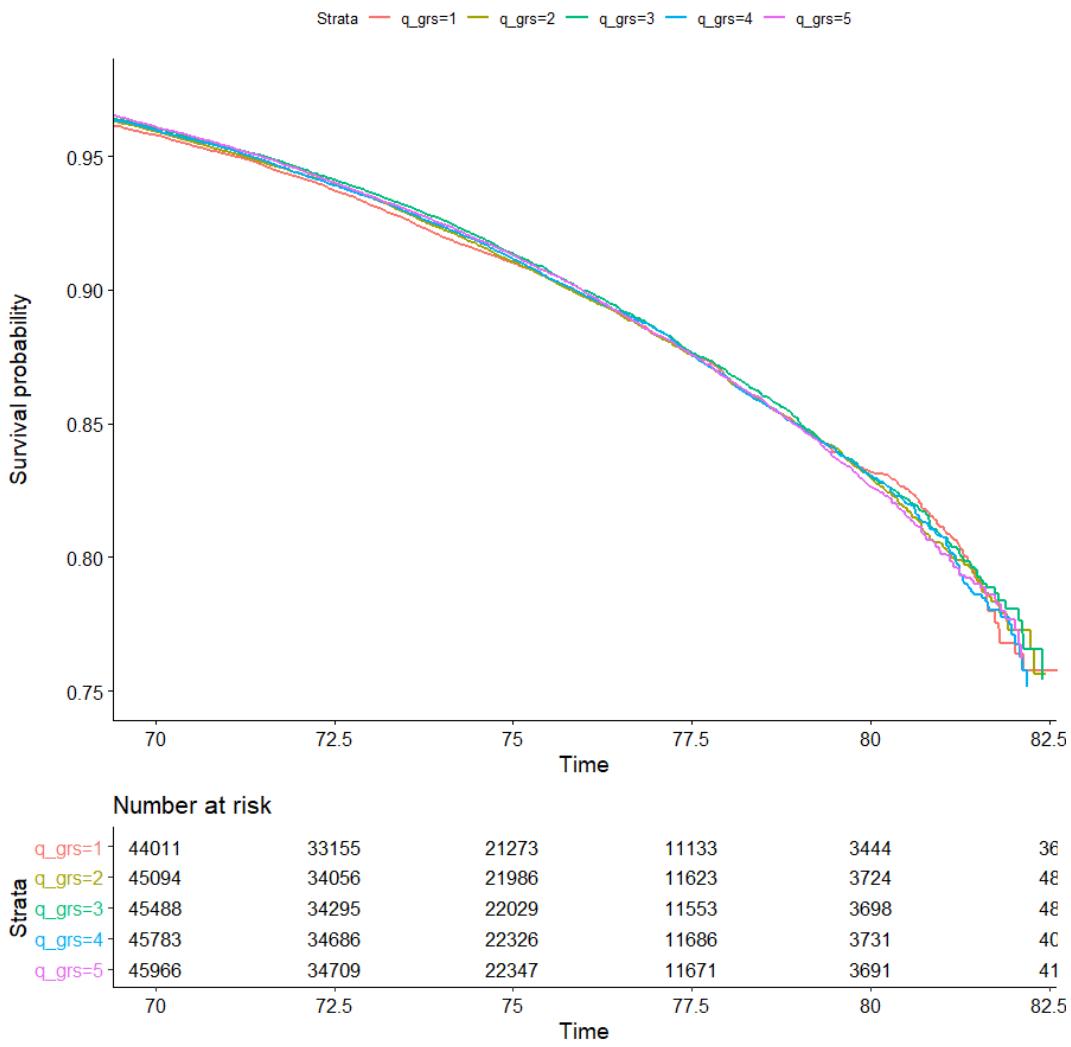


Figure 6.12. Kaplan-Meier Estimate of survival probability for UK Biobank Death data using telomere length genetic risk score quintiles.

To account for non-proportional hazards, I grouped individuals into three age groups, according to their time of event or censorship: [0-75]), [75-80) and [80-100) and used a stratified Cox PH model. However, no evidence of an association was found between longevity and the TL GRS in any age group.

The possibility of TL GRS interaction with time was explored. The significance of both TL GRS and its interaction with time is difficult to interpret. It is likely that TL GRS has non-linear or non-proportional relationship with age at death. The limitations of this analysis will be discussed in chapter 6.5.3. *Limitations of time to death or parent's death analysis*.

6.4.2. Genetically determined telomere length and case-specific survival

GDTL has an opposite direction of effect on the risk of degenerative diseases and cancers. I hypothesised that GDTL effect may also be disease-specific for mortality.

To further investigate GDTL and longevity I analysed different groups of outcomes: all-cause mortality, cardiovascular-specific mortality, and cancer-specific mortality both combined and stratified by sex. Cardiovascular-specific mortality was defined as death caused by diseases of the circulatory system with ICD10 codes I00-I99, and cancer-specific mortality as death caused by neoplasms with ICD10 codes C00-D49.

The TL GRS was not associated with all-cause mortality or cardiovascular death but was found significant for cancer mortality in all three models (**Table 6.6**).

The sex of an individual is an important factor, with men at higher risk of death over time than women. Investigation of mortality in women and men separately showed that the TL GRS is nominally associated with all-cause and cancer mortality in females, and only cancer mortality in men. However, the direction of the association between TL GRS and all-cause mortality in women is opposite from the one hypothesised, suggesting that shorter TL in women is associated with improved longevity. This result is possibly driven by the majority of deaths caused by cancer in women group. Moreover, PH assumption is violated by TL GRS in this model.

I observed a significant association between TL GRS and cancer-specific mortality in women, where 1 SD increase in TL GRS was associated with a 6.4% decrease in hazard of cancer death. This association is consistent with previous findings and confirms shorter TL protective effect against cancers. I observed similar significant association between TL GRS and cancer-specific mortality in men, where 1 SD increase in TL GRS is associated with a 3.4% decrease in hazard of cancer death. TL GRS did not violate PH assumptions in analyses stratified by sex.

Model	Mortality	N(Event)	P-value	HR(95%CI)
Default	All-cause	27009	8.56E-02	0.989(0.978,1.001)
	Cardiovascular	6168	9.12E-01	0.999(0.974,1.024)
	Cancer	12829	3.51E-08	0.952(0.935,0.969)
Female-only	All-cause	10942	8.38E-03	0.975(0.956,0.993)
	Cardiovascular	1837	6.66E-01	0.990(0.945,1.037)
	Cancer	6050	3.81E-07	0.936(0.912,0.960)
Male-only	All-cause	16067	9.39E-01	0.999(0.984,1.015)
	Cardiovascular	4331	8.94E-01	1.002(0.972,1.033)
	Cancer	6779	5.15E-03	0.966(0.943,0.990)

Table 6.6. Results of time-to-death analysis. The default Cox PH model included TL GRS, sex, PC1-5 and BILEVE variables. Sex-stratified models did not need to adjust for *sex*.

To reiterate the limitation of this analysis, it should be noted that my predictor of interest, GDTL, is created using 52 SNPs that explain only a little amount of the variation in TL and may not be sufficient to predict age at death even in a large cohort.

Another key consideration is that the available UK Biobank death data is not representative of the general population in many ways, and this includes the distribution of age at death (**Figure 6.13)**. UKB recruited participants that were over 40 years old, and the latest recorded UKB age at death was 82 years old. In the general population, the median age of death is ~82-90 years old (Canudas-Romo, 2010; de Beer *et al.*, 2016; Butt, 2017; Basellini *et al.*, 2020). UKB sample may not currently be old enough to investigate longevity as the distribution of deaths will not have reached its peak for several years. Accumulation of death data may be used in the future to test GDTL effects on time to death.

Moreover, UKB participants are thought to be the worried healthy, those who are more cautious about their health and are therefore more likely to live healthier and longer lives in comparison to the general population, which will also skew the distribution of age at death, perhaps offsetting the effect of TL in this population to an even later time.

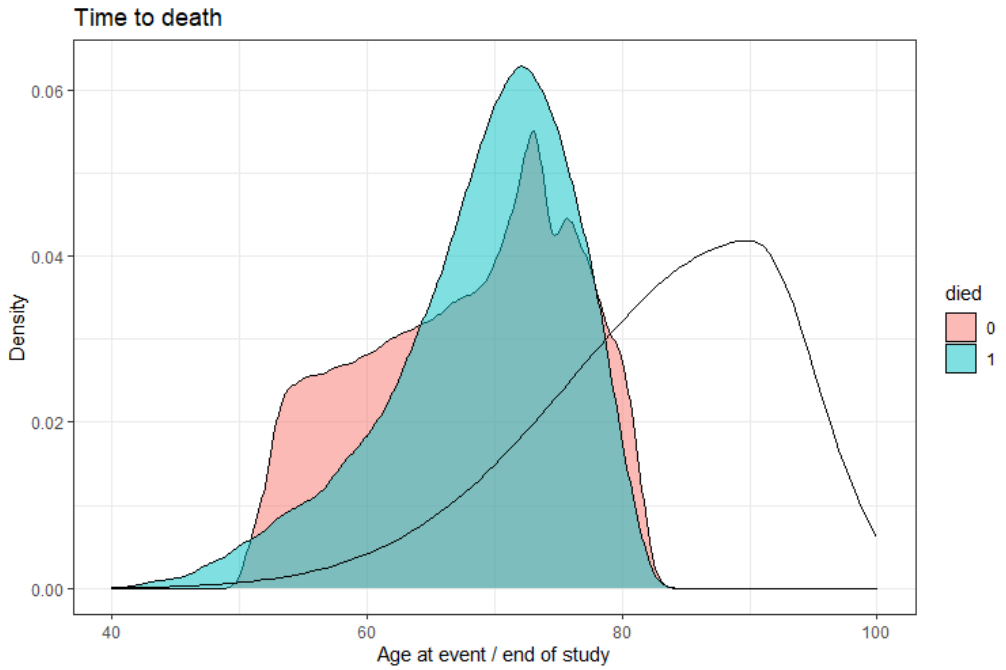


Figure 6.13. Distribution of the age at death or censorship in UK Biobank differs from the general population. Age at death (turquoise), age at censorship (orange) for participants that were still alive by the end of the study and the approximate distribution of age at death (black line) (Butt, 2017; Basellini *et al.*, 2020).

Mortality is often disease-driven, where there is often a starting point from which death is the end point. I did not observe GDTL associations with all-cause or cardiovascular mortality possibly due to death data limitations within UKB, prevalence of cancer-related death, and unaccounted effects of external factors such as smoking and body mass index that are likely to have a greater impact on survival than GDTL.

Nonetheless, I observed a significant GDTL associations with cancer-specific mortality that is consistent with statistically significant associations between GDTL and time to cancer onset. The TL GRS, thus, can be a predictor of healthspan and disease-specific lifespan. The type of disease is important, as I previously reported, shorter TL is protective against cancers and in the case of cancer prognosis shorter GDTL might be beneficial for longer survival.

6.4.3. Genetically determined telomere length and menopause onset

I continued the investigation of GDTL effects on healthspan by analysing age at menopause. Menopause is recorded for more than 150,000 women in the UKB.

Early age at menopause was previously associated with an increased risk of cardiovascular disease and mortality (Muka *et al.*, 2016; D. Zhu *et al.*, 2019; Y. Zhu *et al.*, 2019). This suggests that age at menopause may reflect the healthspan of a woman. Moreover, previous studies have reported longer measured TL to be associated with a later onset of menopause (Gray *et al.*, 2014; Shenassaa *et al.*, 2015). I hypothesised that GDTL may have an effect on time to menopause rather than mortality.

Although death as an outcome is a finite instantaneous event, theoretically more suitable for survival analysis than menopause, investigation into the age of menopause has its advantages. The distribution of age at menopause is long established, while the distribution of age at death has changed throughout the centuries due to an increased understanding of disease, medical advances in treatment and prevention and lifestyle improvements. Menopausal age is more affected by an internal hormonal system due to aging, although environmental factors as smoking may also have a role (Ertunc *et al.*, 2015).

The distribution of age at menopause for UKB is shown in **Figure 6.14**. Initially there were 154,697 events of menopause events out of 264,231 women. However, there was a noticeable discrepancy in the data. Approximately one quarter of women over ~60 years old do not report menopause, which is physiologically unlikely in such a large sample. As I was unable to estimate when the menopause may have started in these individuals, I removed women over 60 years old that did not report menopause from the analysis.

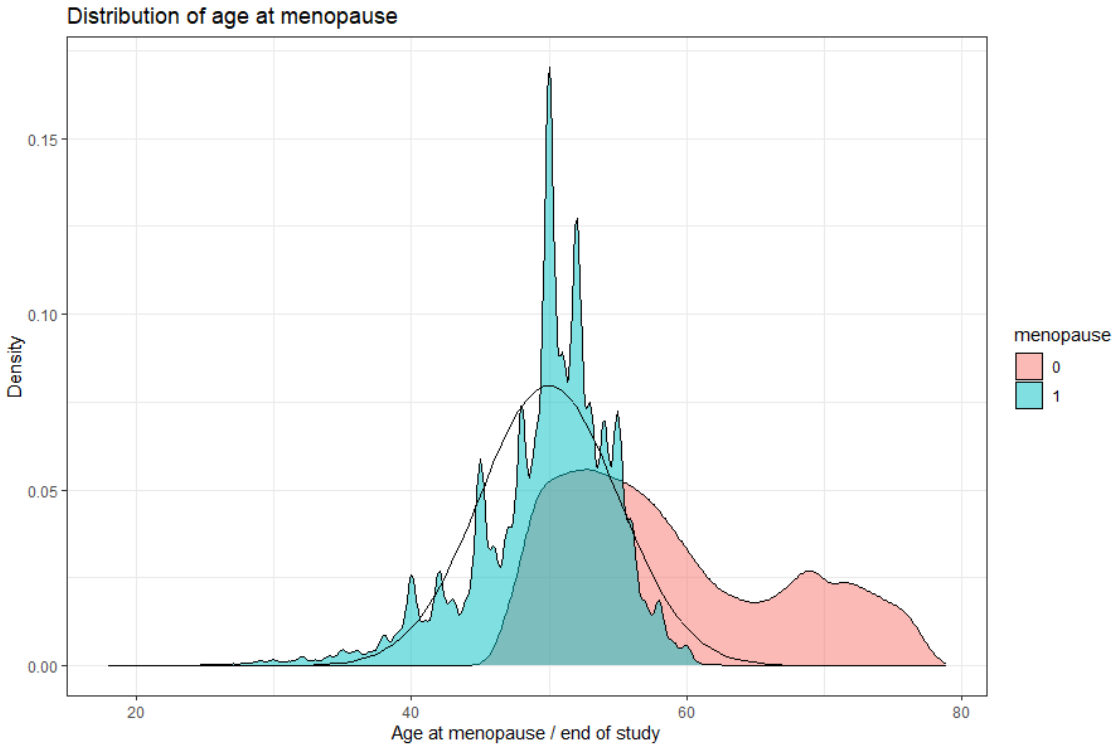


Figure 6.14. Distribution of age at menopause in UK Biobank data. Two coloured distributions are shown: age at menopause (turquoise) and age at censoring for women that did not experience menopause (orange). The approximated distribution for age at menopause in the general population is given by a bell-shaped black curve (mean=50, SD=5).

To estimate the effect of GDTL on time to menopause I used a Cox PH model adjusting for the first 5 genetic principal components and the BILEVE genotyping array. There were 154,697 events of menopause events out of 239,717 women after exclusion of women aged over 60 years old without menopause. The TL GRS was significantly associated with time to menopause. A 1 SD increase in shorter GDTL was associated with a 1.2% increase in hazard for menopause, suggesting that shorter GDTL predisposes women to undergo menopause earlier in life. However, I detected that the TL GRS violates the PH assumption (P-value=0.0007). It is difficult to see, but in the Cox PH model using TL GRS quintiles (**Figure 6.15**), the curves intersect at earlier ages.

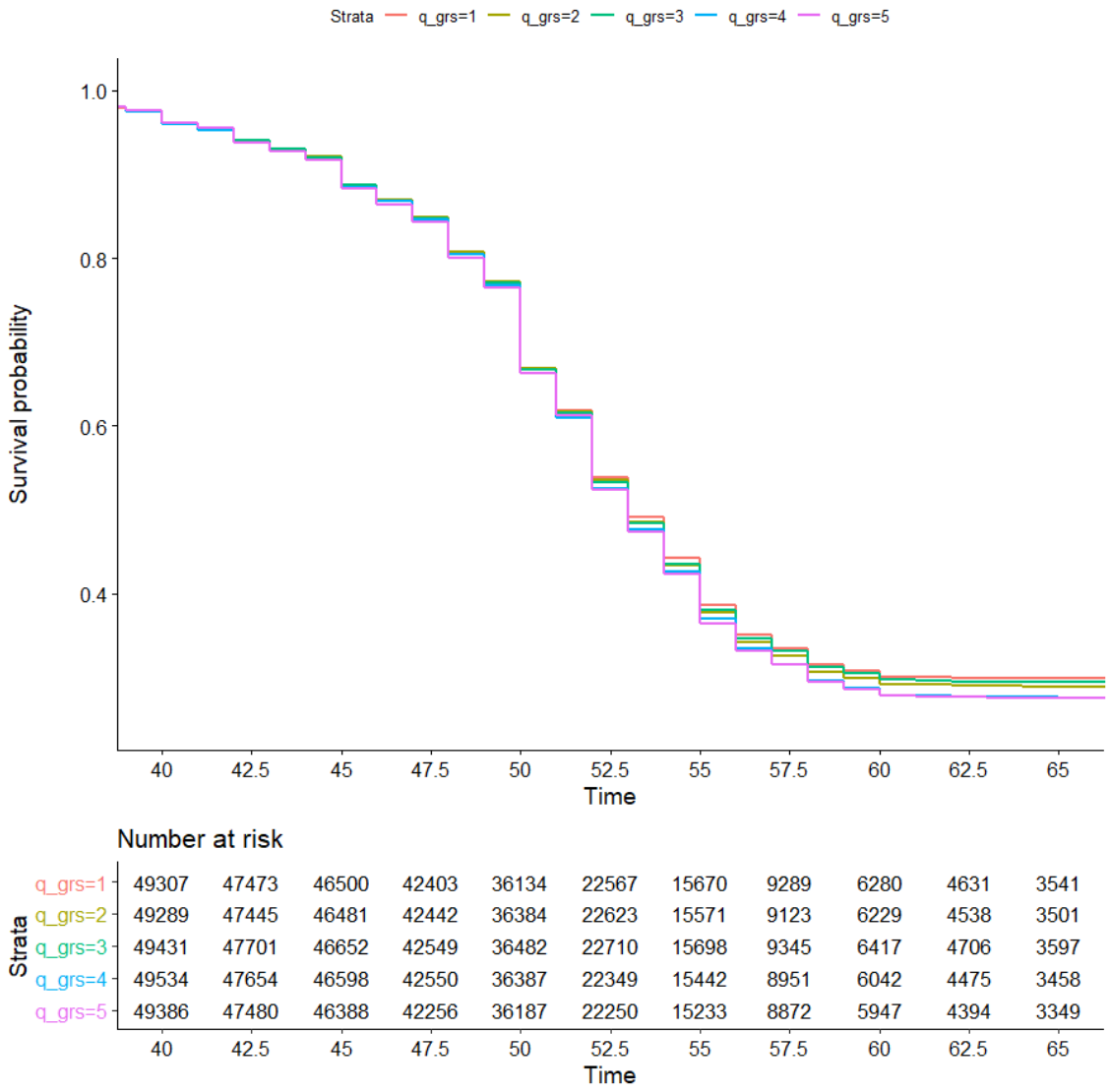


Figure 6.15. Shorter genetically determined telomere length is associated with earlier age at menopause. The graph demonstrates the curves of survival probability for each of 5 TL GRS quintiles (ranging from the longest 1 to the shortest 5). X-axis – age at event or censoring, Y-axis – survival probability (without menopause).

To account for non-PH I investigated whether the effect of TL GRS may be different in groups of women with early, average and late menopause. I divided the subjects into three groups depending on time of event or censorship: [0-45), [45-55), and [55+]. The new Cox PH model stratified by menopause age group and detected no significant effects within each group, possibly due to the reduced power to detect a difference.

Additionally, I checked whether TL GRS is time-dependent by fitting an interaction term between TL GRS and time of menopause or censorship, no significant interaction was found.

A significant association between the GDTL and time to menopause with the expected direction of effect was seen, however, violation of PH was detected. The effect of GDTL might not be independent or it may be masked by stronger external factors such as smoking, stress, chronic conditions, etc. It must also be noted that the distribution of time to menopause does not match the general population perfectly. The main reason for this is that many women did not report an age at menopause due to various reasons, all of which may affect the analysis and coefficient estimation. Nonetheless, this association is expected to become more significant with a more powerful TL GRS and more data in the future.

6.4.4. Genetically determined telomere length and parental longevity

GDTL was not significantly associated with survival using ~27,000 available death records in UK Biobank. Nonetheless, it is well-known that the survival of parents is a determinant for offspring survival (Shadyab *et al.*, 2018) and that parent-offspring TL is correlated (Delgado *et al.*, 2019; Nguyen *et al.*, 2019). As almost all UKB participants reported parental age at death or at the time of assessment, I hypothesised that my TL GRS in the offspring may be a predictor of parental longevity, as each parent contributes roughly half to their offspring GDTL.

I used survival analysis to estimate the effects of individual's GDTL on parental survival. To prepare data for analysis I excluded all individuals, who reported themselves to be adopted, and those who reported the parent to have died before the age of 40. This was done so that death at an early age was removed, to exclude events unlikely to be due to biological or ageing reasons related to common TL genetic determinants. The numbers of reported ages of mothers and fathers as well as the mean age are summarised in **Table 6.7**. This shows that parents who have died were on average younger than the parents still alive at the baseline UK Biobank visit.

Parent	Status	Number	Mean age (SD)
Mother	Alive	181618	78.34 (8.10)
	Dead	260409	75.09 (11.97)
Father	Alive	106556	77.91 (7.29)
	Dead	335471	71.56 (11.65)

Table 6.7. Parental age in UK Biobank. Parents of adopted children and parents who died early were excluded.

To estimate the effect of GDTL on maternal and paternal survival I used Cox PH models stratified by parental sex that included standardised TL GRS as a main predictor. The primary results showed no association of GDTL with parental age at death for either mothers or fathers (**Table 6.8**). Additionally, I analysed only those whose parents had died. This case-only analysis showed a significant association of GDTL on time to mother's death (**Table 6.8**), however, the PH assumption was violated.

Dataset	Parent	N(events)	OR	SE	P-value
Parents alive or who died after age 40	Mothers	262049	1.000	0.002	9.41E-01
	Fathers	337443	1.002	0.002	2.14E-01
Parents who died after age 40	Mothers	234879	0.994	0.002	5.33E-03
	Fathers	234879	1.002	0.002	2.57E-01
Only mothers who died after age 40	Mothers	262049	0.993	0.002	7.82E-04
Only fathers who died after age 40	Fathers	337443	1.001	0.002	5.61E-01

Table 6.8. Estimated effects of genetically determined telomere length on parental survival.

It is estimated that TL heritability is between 40-80% (Broer *et al.*, 2013; Hjelmborg *et al.*, 2015). The individual's GDTL is composed of genetic variants that are inherited from their parents. However, parental GDTL is roughly half of the offspring GDTL, while the other half is unknown.

I analysed the effects of GDTL on time to death of an individual and their parents and detected no statistically significant effects in primary analyses. In secondary analyses GDTL was found to be associated with time to cancer-specific death. I conclude that the TL GRS, may not be a sufficient predictor of human longevity in UK Biobank, unless it is cause-specific and has a sufficient sample size. The potential limitations of the study will be listed in the following chapter to highlight the complexity and missingness of the data, study design limitations and possible solutions for future investigations.

6.5. Limitations of survival analysis study

One of the main advantages of survival analysis is an efficient use of available data, where durations of follow-up may vary and the time that individuals contribute to the study is included even if they do not experience the event or are lost to follow up. Techniques such as Kaplan-Meier survival curves and the Cox PH model are widely used for their ease of interpretation and simplification of underlying principles and assumptions (Flynn, 2012).

Limitations of survival analysis often involve the study design, methodology and a need to satisfy strict assumptions. In this chapter I analysed the effects of GDTL on time to age-related diseases, death, menopause, and the age of parental death. The limitations can be divided into 1) general limitations that affect all performed analyses, and 2) outcome-specific limitations that involve quality of the data and the approaches used to analyse it.

6.5.1. General limitations of time-to-event analysis

The first general limitation of all analyses was with the predictor of interest. Here I use the GRS, built using 52 genetic determinants of TL, which explains <3% of TL variance. Further research and detection of additional TL genetic determinants might help to improve the predictive ability of a TL GRS and detect stronger associations and greater effects on time to health and age-related outcomes. However, it should be noted that this GRS was a good predictor of disease risk, and its components are a valid tool to use as instrumental variables in causal analyses for disease. It was also the most up to date TL GRS at the time of study, using our latest TL meta-GWAS and as such the most appropriate tool to use for analysis before the UK Biobank release of TL measurements and subsequent GWAS.

A second general limitation is the use of the UK Biobank cohort, although phenotypically rich, it is a subset of the general population and does not reflect the true prevalence of diseases in the general population. This is mainly because the cohort participants tend to be healthier than the general population and may have joined UKB as they are more

concerned about their health. Taking this into account comparing estimates to other studies must be done with caution. It should be noted, however, that the UK Biobank resource is very well-powered and analyses within the cohort hold validity within these limitations.

The final general limitation is that gender often violates the PH assumptions and stratified analyses were required. This is mostly driven by the difference in disease prevalence between genders for most diseases. UK Biobank is, however, a cohort study and as such the data are approximately equally split between men and women and whilst women tend to experience disease at an older age a simple analysis on stratified data is sufficient to counter the lack of proportional hazards.

6.5.2. Study-specific limitations of time-to-event analysis

Time-to-event analyses of age-related diseases have specific limitations. This includes statistical power being dependent on the number of events. To maximise the case number in previous chapters, cases were assigned using self-reported data. Unfortunately, for many of these events the date, or age of onset, of the first occurrence was missing due to various reasons. In order to analyse time to an event it is not possible to include participants with a known disease but a missing time of onset.

This reduced the number of events and the power to detect associations. The power of the study is improved by the use of a survival model as opposed to a logistic model as we are able to include all individuals at risk through censoring. This is the strength of survival analyses.

Not all disease events are instantaneous, which is the desired outcome type to have for survival analysis, to obtain a precise estimate of the hazard. For example, myocardial infarction is a well-defined acute event, as for most cases it requires hospitalisation. Cancer, on the other hand, cannot be defined with precision as it could have started developing long before it was diagnosed. However, for these diseases I model the time to diagnosis, which has clinical utility as a useful metric based on practice. Furthermore, health records may be unreliable and contain errors. Nonetheless, for these analyses the disease definitions are considered robust, as they have been reviewed by clinical members of staff. Dates have been reviewed so that any obvious issues with the

reporting of dates in health records are eliminated, including the removal of events that occur after the censorship date.

The TL GRS violated PH for a small number of time-to-event analyses. However, this indicates that investigations into the variability of GDTL effects during various ages or via an interaction with disease specific covariates are required. Nonetheless, my time-to-event analyses were mostly consistent with the previous GRS and MR associations, which confirms the validity of results and potential use of the TL GRS in estimating age specific hazards for a set of significant diseases, for example, CAD and skin cancer.

6.5.3. Limitations of time to death or parent's death analysis

Time to death analysis is limited due to the distribution of age at death within UK Biobank, which was not representative of the general population. UK Biobank recruited participants that were 40-69 years old between 2006 and 2010 and collected death records up to the year 2020 (Marchini, 2015; UK Biobank, 2015). During the study period ~27,000 death events occurred within the cohort. However, the maximum age at death was 82 years old, which is approximately the median age of death in the general population within the UK. This indicates that longer follow-up time is required to accumulate enough data. The data are rich and when the cohort ages over the next 5-10 years we will be able to run similar analyses again to determine the effect of GDTL on longevity.

Another limitation was the influence of cause of death. Shorter GDTL was associated with an increased risk of cardiovascular diseases and reduced risk of cancers. These types of diseases have different distributions of age at onset and may themselves have a direct contribution to mortality. The effect of all-cause mortality may have been cancelled-out by the effects of GDTL pulled in both directions. To account for this, I analysed cardiovascular- and cancer-specific mortality.

In case of parental longevity, I was limited by the use of GDTL that potentially only explains half of the variation in TL compared to the offspring. It might be insufficient to detect a significant association or to estimate the true effect underlying effect in the parent. Nonetheless, I utilised all data on individual and parental longevity available to

us via UK Biobank. With further accumulation of data on parental death events, or by improving TL GRS, it might be possible to detect significant associations.

6.5.4. Limitations of time-to-event analysis of menopause

The distribution of age at menopause was consistent with the distribution of age at menopause in the general population for women, who reported the event. However, the distribution of age at menopause for some premenopausal women was strongly skewed towards older ages. This indicates that some women did not report the event, though many likely had menopause, or there were data entry or reporting issues. In these analyses this resulted in the exclusion of a large number of women from the study. It is also unclear whether the data can be considered missing at random, as was assumed, because it could be missing due to the sensitivity of the question. It was decided that women over 60 who do not report menopause should be excluded rather than included as premenopausal.

The TL GRS was again found to violate the PH assumption in the menopause analysis but was not found to be time-dependent. Modelling the GDTL effect on time to menopause may require consideration of additional phenotype specific covariates such as hormone replacement therapy use, cancer or other disease influences, environmental factors such as smoking, as well as any possible interaction with the TL GRS. Study design improvements as well as choosing an alternative model to Cox PH might be beneficial in tackling the problem. Nonetheless, this analysis had a large number of events to detect associations and highlighted important data limitations crucial for future investigations.

6.6. Conclusions on the findings of effects of genetically determined telomere length on time to event

In this chapter I investigated the effects of GDTL on time to event, where many events were considered, including age-related disease onset, longevity, menopause, and parental longevity. Age-related diseases and menopause are characterised to represent elements of an individual's healthspan, the period of change in life when the effects of ageing become evident to the individual. Death and parental longevity represented the analysis of lifespan, or longevity, the number of years that an individual survives.

I found GDTL to be a significant predictor for the early onset of several disease phenotypes including CAD, cancers, menopause, and individual cancer-specific mortality, but not a good predictor of individual or parental all-cause mortality. I estimated the hazard of developing age-related diseases at specific time points in life and showed that shorter GDTL increases the hazard of CAD and menopause and decreases the hazard of cancers.

The analyses highlight a difference between an individual's healthspan and longevity. Healthspan is represented by age-related diseases seemed to be better defined in UK Biobank for a genetically determined measure. GDTL was able to predict, for example, the time to CAD onset, but not to all-cause mortality. Information on age at death is still being accumulated by UK Biobank and future work may provide a more powerful analysis of longevity and estimate the potential effect of GDTL on survival. Investigations can be extended into the exploration of not only single outcomes, but multiple, as in multi-state models that look into the transition between states, i.e., time to CAD and time from CAD to death. The interaction between genetic TL and environmental factors and their combined influence on the healthspan or other outcomes may bring a better understanding of early disease onset and healthy ageing.

Chapter 7. Discussion and conclusion

In this project I aimed to investigate the influence of genetically determined TL on age-related diseases, and in this chapter, I summarise the key findings, study limitations and highlight potential future work.

7.1. Summary of key findings

7.1.1. The use of telomere length

Telomere length has been proposed as a marker of biological age, where a biological count down is represented by the loss of telomeric repeats that happens with each cell division due to the 3' replication problem. This mechanism allows only a certain number of cell divisions before the cell becomes senescent, or dysfunctional, and needs to be replaced.

Telomere length, measured in blood or tissue, can be considered as an average representation of individual's true TL. Loss of telomeric repeats is similar across tissues, and thus TL may reflect the biological age of an individual. In the literature TL was associated with a number of age-related diseases. Shorter TL was repeatedly reported to increase the risk of cardiovascular diseases such as CAD, and longer TL to increase the risk of cancers. However, association studies of measured TL are not free of residual confounding or of reverse causation, where disease or related inflammation may affect TL.

7.1.2. The genetic determinants of telomere length

Based on the limitations of observational research utilising measured TL I aimed to investigate genetic TL, represented through the combination of 52 genetic determinants of TL, in relation to age-related diseases. It is known that genetic information cannot be altered and is therefore free of reverse causation and residual confounding in the traditional sense. Whilst measured TL is thought to reflect the current biological age and health status of an individual, genetically determined TL may set the initial rules within an individual for TL maintenance. Genetically determined TL may have varying effects

on TL attrition trajectories through time in individuals with the same amount of directly inherited TL.

A genetic predisposition to shorter telomeres may limit the cell's potential to provide the required maintenance for telomeres, which would accelerate the loss of repeats during cell division. In turn, this may contribute to premature cellular senescence, tissue degeneration and loss of function. A genetic predisposition to longer telomeres may allow for prolonged telomere maintenance, which may compensate or minimise the loss of TL and delay cell senescence.

To test these hypotheses and gain an understanding into genetic TL effects I employed three statistical methods that each answer a different question relating to disease risk: Genetic Risk Score, Mendelian Randomisation and Survival Analysis.

I first had to identify genetic variants that were associated with TL. In our latest genome-wide meta-analysis we identified 52 genetic determinants of TL using a 5% false discovery rate following GCTA conditional analysis (Li *et al.*, 2020), that could be used to represent genetically determined TL. I calculated a TL GRS for each UKB participant using 52 SNPs and a second TL GRS using 234 SNPs that were defined independent via PLINK clumping. I defined 127 diseases, most of which represent the ageing processes, and assigned disease status to each individual utilising the wealth of data available in UKB, including self-reported data and electronic health records.

7.1.3. Three analyses to answer different questions

I performed GRS association analyses to investigate the change in disease risk due to a genetic predisposition to shorter telomere length using two genetic risk scores for TL. I then ran causal inference to estimate the effect size of a causal association between TL and disease risk using MR. As part of the MR strategy, I performed sensitivity analyses to determine consistency and to detect pleiotropic effects. Finally, I ran time-to-event analyses to determine the impact of GDTL on disease onset. I extended analyses in that chapter to investigate the ability of the TL GRS to predict time to menopause and death. These three analyses can be interpreted independently due to their ability to ask different questions, as described within each chapter. Here I consider the impact of the three approaches when taken together.

7.1.4. The role of genetically determined telomere length in disease

Across the three analyses, shorter genetically determined TL was associated with an increased risk of cardiovascular, endocrine, and immune-related diseases, and with a decreased risk of cancers and diseases with high proliferative potential. This confirmed the findings of the observational studies of measured TL.

My results suggest that genetically determined TL is causally associated with degenerative diseases that are characterised by increased cellular turnover and chronic inflammation. GDTL, along with measured TL, may contribute to accelerated ageing of the tissue by limiting its regenerative capacity and promoting inflammation and senescence, that may lead to loss of tissue function. Genetically determined TL may be linked to immune and inflammatory pathways through DNA damage repair or other TL maintenance-related mechanisms, including synthesis of nucleotides, as was reported in our latest TL GWAS (Li *et al.*, 2020). I discuss the genetic TL association and its potential biological relevance to cardiovascular, endocrine, and immune-related diseases in detail in chapter 5.3. *Causal role of telomere length in age-related diseases*.

Given that my detected associations are consistent across methods the results are given more weight as the association is independent of the method. However, many associations in the group of degenerative diseases were only of nominal significance with wide confidence intervals. It must also be noted that the degenerative disease phenotypes investigated here may be nested, and further investigations into possible mediators and TL interactions with external cofactors are needed.

For shorter telomere length, I observed there to be an opposite effect for diseases with a high capacity to proliferate. Shorter TL was causally associated with a decreased risk of many cancers and proliferative diseases. The association results within this group are consistent in the level of statistical significance, direction, and effect size. In contrast to degenerative diseases, the results for cancers appear to be more robust, and strongly point to TL primary function of limiting cell division and preventing unnecessary proliferation.

To summarise, I identified several potentially causal links between genetic TL and age-related diseases that are biologically plausible. A TL GRS can be used to estimate disease risk and determine the hazard at specified age. In combination with measured TL and

clinical risk scores, the genetic determinants of TL, when combined, could help to improve the screening for individuals at risk in order to manage individual preventive therapies.

7.2. Study limitations

This project had several data and method limitations that should be considered when interpreting these results.

The selection of genetic determinants was restricted to results from a single large-scale TL meta-analysis, the ENGAGE study (Li *et al.*, 2020). The selected 52 variants explain only ~2.9% of TL variance. Thus, it should be highlighted that the genetic instrument is not overly powerful. Nonetheless, ENGAGE was the largest published TL GWAS at the time of this project, making this the largest and most up-to-date investigation of associations between genetic TL and age-related diseases. It is also clear that despite the lack of power, which may result in several associations being missed, the findings presented were consistent across the approaches used.

The power and reliability of the genetic instrument relied on an assessment of SNP independence to avoid potential biases due to LD that may have caused a single underlying effect to be counted multiple times. I performed primary analyses using the 52 genetic variants based on the FDR list. Whilst I also assessed SNP selection approaches using PLINK clumping, which resulted in the selection of 234 SNPs. It should be noted that other alternative approaches to select genetic variants could have been used. Indeed, it is possible that a GRS based on more than significant results could have been used, or a score that takes biological relevance of each SNP, with TL, into account. These scores provide potential for future work.

UK Biobank, the resource used, comprises of genetic and phenotypic information of individuals that are not ideally representative of the general UK population. UKB participants are thought to be more conscious about their health, and thus relatively healthier. For this reason, generalisations are more challenging to other populations. Nonetheless, UKB is the largest dataset accessible to researchers with extensive phenotypic information. Through the use if UKB I was able to explore over 120 age-related diseases.

In this project I relied on the accuracy of defining over 120 disease phenotypes in UKB. Although disease definitions were curated with the help of clinical colleagues, we cannot rule out that an individual who we consider as disease case may have been missed due to errors in data collection or coding. It is also not possible to know if there are any errors in the self-reported data or health records. It was evident that within UKB we have individuals that are cases identified through electronic health records but who did not report disease at baseline and vice versa, we have self-reported events that are not in the electronic records. I treated the participants as a case if either of these records were reported. Of course, it may be a cleaner phenotype, if I had only included cases with evidence that occurred in multiple data sources. However, this strategy is more restrictive and has its own limitations. To select controls for analyses I assumed that there is no underlying pattern of disease. As such, individuals with other diseases would act as a control for any disease they did not have. This assumption may not be true in the case of comorbidities or where there is a strong common causal risk factor.

Due to the nature of investigating a broad spectrum of 127 diseases it was not feasible to adjust each disease for all possible confounders, which could bias the associations identified. However, this was not true of the MR, where pleiotropy was assessed, and results show consistency helping alleviate this concern. During all analyses I adjusted for well-known covariates such as age, sex, and ethnicity throughout methods, and, if appropriate, performed sensitivity analyses to detect confounding.

The results of this work should be interpreted alongside these limitations in data, study design and methods, as well as taken forward in future analyses.

7.3. Future work

The investigation of genetically determined TL described in this thesis promotes a range of ideas and possibilities for future TL research.

Within our downstream analysis of the 52 genetic determinants of TL and their relation to genes and their products we highlighted pathways relevant to TL (Li *et al.*, 2020). However, this could be taken further through using bioinformatic approaches as well as wet-lab experiments to determine the exact mechanisms of TL maintenance to support the proposed hypothesis of TL-disease associations.

A TL GWAS in a larger sample size, perhaps with different ethnicities, may identify novel genetic determinants of TL that could be used to build a more robust genetic instrument. Such a study is possible and has been lead and performed within UK Biobank by Codd *et al*., 2021, at the Cardiovascular Research Centre at the University of Leicester. A project that measured TL in all UKB participants, which was only completed towards the end of my PhD, at the time of writing this thesis. A consequent TL GWAS, which I am also involved with, to identify novel genetic variants associated with TL is available as a pre-print (Codd, Wang, *et al.*, 2021). This study is going to provide state-of-the-art findings in the epidemiological research of telomeres, where both genetic and measured TL have been explored in relation to multiple traits including age-related diseases. Moreover, TL measurements are going to be released by UKB, and available to researchers internationally, greatly enhancing the progress towards understanding telomeres and their effects on disease risk.

TL association analyses can be expanded to mediation analyses, where different additional modifiable exposures, associated with TL, may potentially lie on the causal pathway to disease. Another avenue for research could be the analysis of interactions between TL and disease specific or external factors. Bi-directional MR can also be employed to investigate measured TL mediators. We know that physical TL can be affected by external factors, such as inflammation or disease, therefore an investigation into the role of disease on TL would inform our understanding of the potential complexities of telomere maintenance. For example, there is evidence that insulin is a mediator on the potentially causal pathway between TL and CAD (Zhan *et al.*, 2017).

The effects of TL on time to event can be explored using models that account for an interaction with time or assume non-linear relationships. Indeed, this project has shown that TL can predict time to disease but is a poor predictor of time to death. Extensions beyond the Cox PH model could also provide better modelling, allowing for the estimation of the baseline hazard. Future event data accumulation may be useful to re-investigate the TL effects on time to disease and to death using a multi-state modelling framework to model the transition between diseases, ultimately leading to death.

With new TL data sources becoming available multiple areas of telomere research may be expanded to explore TL effects on human health and provide individual and public advice to improve the healthspan.

7.4. Conclusion

I have shown through multiple analyses that genetically determined telomere length is consistently linked to several diseases. Many of these relationships are also estimated to be causal. Shorter telomere length increases the risk of cardiovascular, endocrine, and immune-related diseases, while also being associated with a decreased risk of cancers and diseases with high proliferative potential. These consistent results suggest that telomeres are an important biomarker for disease risk, are associated with the time to disease onset and are involved in the causal pathways of disease.

Appendix

Appendix 1. Genetic determinants of telomere length

Table S1. Independent variants associated with leucocyte telomere length at false discovery rate ≤0.05. **refA** - reference allele, **Freq** – allele frequency in the original data, **Beta** - effect size, **SE** - standard error, **P-value** – significance level from the original GWAS meta-analysis, **N** - estimated effective sample size.

SNP	CHR:BP	Closest gene	refA	Freq	Beta	SE	P-value	N
rs10936600	3:169514585	*LRRC34 (TERC)*	T	0.243	-0.0858	0.0057	6.42E-51	80402
rs7705526	5:1285974	*TERT*	A	0.328	0.0820	0.0058	4.82E-45	64656
rs2853677	5:1287194	*TERT*	A	0.592	-0.0638	0.0055	3.12E-31	66348
rs4691895	4:164048199	*NAF1*	C	0.783	0.0577	0.0061	1.47E-21	77751
rs9419958	10:105675946	*STN1 (OBFC1)*	C	0.862	-0.0636	0.0071	4.77E-19	79674
rs75691080	20:62269750	*STMN3*	T	0.091	-0.0671	0.0089	5.75E-14	73300
rs59294613	7:124554267	*POT1*	A	0.293	-0.0407	0.0055	1.12E-13	77807
rs8105767	19:22215441	*ZNF208*	G	0.289	0.0392	0.0054	5.21E-13	80103
rs73624724	20:62436398	*ZBTB46*	C	0.129	0.0507	0.0074	6.08E-12	79451
rs3219104	1:226562621	*PARP1*	C	0.830	0.0417	0.0064	9.31E-11	82702
rs932827	20:62380527	*ZBTB46*	T	0.238	-0.0374	0.0060	3.28E-10	75271
rs2736176	6:31587561	*PRRC2A*	C	0.313	0.0345	0.0055	3.41E-10	74733
rs3785074	16:69406986	*TERF2*	G	0.263	0.0351	0.0056	4.50E-10	78947
rs7194734	16:82199980	*MPHOSPH6*	T	0.782	-0.0369	0.0060	6.72E-10	79221
rs34978822	20:62291599	*RTEL1*	G	0.015	-0.1397	0.0227	7.04E-10	64579
rs34991172	6:25480328	*CARMIL1*	G	0.068	-0.0608	0.0105	6.03E-09	69563
rs228595	11:108105593	*ATM*	A	0.417	-0.0285	0.0050	1.39E-08	79131
rs2302588	14:73404752	*DCAF4*	C	0.100	0.0476	0.0084	1.64E-08	75515
rs13137667	4:71774347	*MOB1B*	C	0.959	0.0765	0.0137	2.37E-08	65744
rs55749605	3:101232093	*SENP7*	A	0.579	-0.0373	0.0067	2.38E-08	44478
rs62053580	16:74680074	*RFWD3*	G	0.169	-0.0389	0.0071	3.96E-08	68785
rs754017156	2:54482703	*ACYP2*	D	0.165	0.0471	0.0088	7.52E-08	45835
rs12909131	15:50387678	*ATP8B4*	T	0.231	-0.0308	0.0058	1.15E-07	80707
rs1744757	20:35734863	*MROH8*	T	0.851	0.0359	0.0068	1.38E-07	82223
rs2124616	18:661917	*TYMS*	A	0.140	-0.0374	0.0072	1.72E-07	78571
rs2613954	3:112847045	*RP11-572M11.4*	T	0.886	-0.0381	0.0078	1.10E-06	78133
rs12065882	1:114078755	*MAGI3*	G	0.208	0.0298	0.0062	1.36E-06	77171
rs2386642	10:5702259	*ASB13*	A	0.673	-0.0256	0.0053	1.44E-06	78325
rs56810761	2:210663697	*UNC80*	T	0.270	0.0275	0.0057	1.45E-06	75730
rs62365174	5:78925743	*TENT2*	G	0.088	-0.0544	0.0113	1.50E-06	47138
rs112655343	12:14430807	*ATF7IP*	T	0.102	0.0425	0.0090	2.22E-06	65703
rs55710439	15:65229816	*ANKDD1A*	T	0.014	0.1050	0.0223	2.65E-06	69380
rs11640926	16:1249877	*CACNA1H*	G	0.139	0.0557	0.0119	2.93E-06	28513
rs60160057	4:151000830	*DCLK2*	A	0.211	-0.0287	0.0062	3.15E-06	76459
rs117536281	14:105494403	*CDCA4*	G	0.034	0.0850	0.0183	3.31E-06	43901
rs7510583	22:44698803	*KIAA1644*	G	0.290	0.0347	0.0075	3.38E-06	42137
rs59192843	14:74514120	*BBOF1 (CCDC176)*	G	0.059	0.0655	0.0141	3.52E-06	43632
rs57415150	8:2882469	*CSMD1*	A	0.042	-0.0584	0.0126	3.68E-06	76210
rs6038821	20:7402809	*LINC01706*	T	0.038	0.0596	0.0129	3.98E-06	78795
rs144204502	17:76183233	*TK1*	T	0.014	-0.0896	0.0196	4.92E-06	90239
rs6107615	20:5310273	*PROKR2*	C	0.422	-0.0228	0.0050	5.30E-06	79236
rs9972513	15:38930961	*RP11-275I4.2*	T	0.281	0.0247	0.0055	5.75E-06	80585
rs117037102	11:93404608	*CEP295*	T	0.018	0.0979	0.0218	6.81E-06	58251
rs7276273	21:45994841	*KRTAP10-4*	C	0.007	-0.1502	0.0334	6.90E-06	58816
rs11665818	19:39768216	*IFNL2*	A	0.195	0.0278	0.0062	7.04E-06	80995
rs3213718	14:90869913	*CALM1*	T	0.583	0.0224	0.0050	7.22E-06	79728
rs112347796	5:138964816	*UBE2D2*	D	0.049	0.0691	0.0154	7.29E-06	43936
rs143276018	19:3939249	*NMRK2*	C	0.018	-0.1015	0.0229	9.02E-06	51875
rs201375979	8:100917632	*COX6C*	D	0.317	0.0332	0.0075	9.11E-06	39878
rs7311314	12:54592103	*SMUG1*	A	0.317	0.0240	0.0054	9.50E-06	75916
rs35675808	1:167399643	*CD247*	G	0.028	0.0736	0.0166	9.54E-06	64172
rs117610974	15:55105443	*UNC13C*	G	0.009	-0.1540	0.0350	1.05E-05	42499

Appendix 2. Correction for winner's curse

R script for winner's curse correction:

```
FIQT <- function(z=z, min.p=10^-300)
  pvals<-2*pnorm(abs(z),low=F)
  pvals[pvals<min.p]<- min.p
  adj.pvals<-p.adjust(pvals,method="fdr") # multiple testing adjustment (MTA) using FDR
  mu.z<-sign(z)*qnorm(adj.pvals/2,low=F) # back-transforming MTA P-values on Z-score scale
  mu.z[abs(z)>qnorm(min.p/2,low=F)] <- z[abs(z)>qnorm(min.p/2,low=F)]
  mu.z

z <- beta / se # obtaining z-scores

fiqt_z <- FIQT(z) # correcting z-scores

z_shrink <- z / fiqt_z # obtaining shrinkage level

fiqt_beta <- beta / z_shrink # correcting betas by shrinkage level
```

Appendix 3. Disease definitions

Table S2. Diseases defined by self-reported and hospital episode data.

Phenotype	Definition		
	Self-reported (variable:value)	ICD-10	ICD-9
Cardiovascular diseases			
Coronary artery diseases (CAD)	Heart attack / myocardial infarction (20002:1075), heart attack diagnosed by a doctor (6150:1), age heart attack diagnosed (3894:age), date of myocardial infarction (42000:date), date of STEMI (42002:date), date of NSTEMI (42004:date)	I21-I25	410-412, 414
Atrial fibrillation (AF)	Atrial fibrillation (20002:1471) or atrial flutter (20002:1483)	I48	427.3
Heart failure (HF)	Heart failure / pulmonary odema (20002:1076)	I50, I11.0, I13.0, I13.2	428
Peripheral vascular disease (PVD)	Peripheral vascular disease (PVD) (20002:1067) or leg claudication / intermittent claudication (20002:1087)	I73.9, I74	443.9, 444
Venous thromboembolism	Venous thromboembolic disease (20002:1068), pulmonary embolism (20002:1093) or deep venous thrombosis (DVT) (20002:1094)	I26, I80, I81, I82, I74, I63.6, I67.6	415.1, 451-453
Aortic valve stenosis	Aortic stenosis (20002:1490)	I35.0	424.1
Hypertension	Blood pressure medications (6177:2, 6153:2), vascular / heart problems diagnosed by doctor - high blood pressure (6150:4), high systolic blood pressure (BP), automated reading (4080:>140), high systolic BP, manual reading (93:>140), high diastolic BP, automated reading (4079:>90), high diastolic BP, manual reading (94:>90), age high blood pressure diagnosed (2966:age)	I10-I13, I15	401-405
Stroke	Stroke (20002:1081), subarachnoid haemorrhage (20002:1086), ischaemic stroke (20002:1583), vascular/heart problems diagnosed by doctor - stroke (6150:3), age stroke diagnosed (4056:age), date of stroke (42006:date), age of ischaemic stroke (42008:date), age of intracerebral haemorrhage (42010:date), date of subarachnoid haemorrhage (42012:date)	I60-I64, G46.3, G46.4, G46.5, G46.6, G46.7, G46.8	430-432
Varicose veins	Varicose veins (20002:1494), varicose ulcer (20002:1593)	I83, I84	454
Raynaud's phenomenon / disease	Raynaud's phenomenon / disease (20002:1561)	I73.0	443
Endocrine disorders			
Diabetes	Diabetes (20002:1220), diabetes type I (20002:1222), diabetes type II (20002:1223), diabetes diagnosed by doctor (2443:1), medication for cholesterol, blood pressure, diabetes, or take exogenous hormones - insulin (6153:3), medication for cholesterol, blood pressure or diabetes - insulin (6177:3), age diabetes diagnosed by doctor (2976:age), started insulin within one year diagnosis of diabetes	E10-E14	250
Diabetes type I	Diabetes type I (juvenile type) (20002:1222)	E10	250.01, 250.03, 250.11, 250.13, 250.21, 250.23, 250.31, 250.33, 250.41, 250.43, 250.51, 250.53, 250.61, 250.63, 250.71, 250.73, 250.81, 250.83, 250.91, 250.93

Diabetes type II	Diabetes type II (20002:1223), generic diabetes (20002:1220) and age of onset 35+ years old (2976:>35)	E11, E13-E14	250.00,250.02,250.10,250.12,250.20,250.22,250.30,250.32,250.40,250.42,250.50,250.52,250.60,250.62,250.70,250.72,250.80,250.82,250.90,250.92
Hyperthyroid	Hyperthyroidism / thyrotoxicosis (20002:1225)	E05	242.9
Hypothyroid	Hypothyroidism / myxoedema (20002:1226)	E03.9	244.9
Mental illnesses			
Anxiety	Anxiety / panic attacks (20002:1287)	F40-F41	300
Depression	Depression (20002:1286), Probable Recurrent major depression (severe) (20126:3), Probable Recurrent major depression (moderate) (20126:4), Single Probable major depression episode (20126:5), age at first depression (20433:age)	F32-F33	296.2-296.3
Multiple sclerosis	Multiple sclerosis (20002:1261)	G35	340
Epilepsy	Epilepsy (20002:1264)	G40-G41	345
Dementia	Dementia / Alzheimer's / cognitive impairment (20002:1263)	F00-F03, G30-G31	290, 330-331
Parkinsons' disease	Parkinson's disease (20002:1262)	G20-G21, F02.3	332
Migraine	Migraine (20002:1265)	G43	346
Mania / bipolar disorder / manic depression	Mania / bipolar disorder / manic depression (20002:1291), Bipolar I Disorder (20126:1), Bipolar II Disorder (20126:2)	F30-F31	296.0-296.1, 296.4-296.8
Anorexia nervosa	Anorexia / bulimia / other eating disorder (20002:1470)	F500, F502, F508, R630	3071, 7830, 30751
Schizophrenia	Schizophrenia (20002:1289)	F20, F21, F22, F25, F28, F29	295
Chronic fatigue syndrome	Chronic fatigue syndrome (20002:1482)	R5382, F48.0, G93.3, M79.7	78071
Digestive diseases			
Gastro-oesophageal reflux disease (GORD)	Gastro-oesophageal reflux or gastric reflux (20002:1138)	K21	530.11, 530.81
Irritable bowel syndrome (IBS)	Irritable bowel syndrome (20002:1154)	K58	564.1
Inflammatory bowel disease (IBD)	Inflammatory bowel disease (20002:1461), Crohn's disease (20002:1462) or ulcerative colitis (20002:1463)	K50-K51	555-556
Gallstone	Cholelithiasis / gall stones (20002:1183)	K80	574
Peptic ulcer	Peptic ulcer (20002:1400), duodenal ulcer (20002:1457) or gastric / stomach ulcers (20002:1142)	K25-K28	531-533
Liver cirrhosis	Liver failure / cirrhosis (20002:1158), primary biliary cirrhosis (20002:1136), alcoholic liver disease or alcoholic cirrhosis (20002:1604)	K70, K74	571
Appendicitis	Appendicitis (20002:1502)	K35-K37	540-543
Oesophagitis	Oesophagitis / Barrett's oesophagus (20002:1139)	K20, K22.7	530.10, 530.85
Hiatus hernia	Hiatus hernia (20002:1474)	K44.0, K44.1, K44.9	552.3, 553.3, 551.3
Abdominal hernia	Abdominal hernia (20002:1511)	K45-K46	-
Umbilical hernia	Umbilical hernia (20002:1512)	K42	5511, 5521, 5531
Inguinal hernia	Inguinal hernia (20002:1513)	K40	5500, 5501, 5509
Malabsorption / coeliac disease	Malabsorption / coeliac disease (20002:1456), coeliac disease or gluten sensitivity (21068:1)	K90.0	579

Diverticular disease / diverticulitis	Diverticular disease / diverticulitis (20002:1458)	K57, K38.2	562
Rectal or colon adenoma / polyps	Rectal or colon adenoma / polyps (20002:1460) or benign neoplasms	K63.5, K62.0, K62.1, D12	5690, 211.3, 211.4, 2095
Haemorrhoids	Haemorroids / piles (20002:1505)	K64	455
Pancreatitis	Pancreatitis (20002:1165)	K85, K86.0–K86.1, K86.3, B25.2, B26.3, K87.1	577.0–577.1, 0723
Peritonitis	Peritonitis (20002:1190)	K65, K67, N733, N734, N735, A1831, A5485, A7481	567, 56889, 0140, 03283, 0952, 09886, 6145, 6147
Musculoskeletal diseases			
Gout	Gout (20002:1466)	M10	274
Rheumatoid arthritis	Rheumatoid arthritis (20002:1464)	M05-M06	714
Osteoarthritis	Osteoarthritis (20002:1465)	M15-M19	715
Osteoporosis	Osteoporosis (20002:1309)	M80-M82	733.0
Sciatica	Sciatica (20002:1476)	M54.3-M54.4	724.3
Intervertebral disc disorder	Prolapsed disc / slipped disc (20002:1312) or disc degeneration (20002:1533)	M50-M51	722
Spine arthritis / spondylitis	Spine arthritis / spondylitis (20002:1311) or ankylosing spondylitis (20002:1313)	M46.0, M46.1, M46.5-M46.9, M08.1, M45	721.90, 721.91, 720.0
Respiratory diseases			
Chronic obstructive pulmonary disease (COPD)	COPD (20002:1112) or emphysema / chronic bronchitis (20002:1113), age COPD diagnosed (22150:age)	J40-J44	490-492, 495-496
Asthma	Asthma (20002:1111), age asthma diagnosed by doctor (22147:age), age asthma diagnosed (3786:age)	J45-J46	493
Lower respiratory infection / pneumonia	Pneumonia (20002:1398)	J10-J18, J20-J22	466, 480-487
Otitis media	Not self-reported	H65-H66	381-382
Hay fever / allergic rhinitis / eczema	Hay fever, allergic rhinitis (20002:1387), eczema / dermatitis (20002:1452) or contact dermatitis (20002:1669), Age hay fever, rhinitis or eczema diagnosed (3761:age), Hay fever, allergic rhinitis or eczema (6152:9)	J30-J31, L20-L30	477, 692, 6918
Hay fever	Hay fever or allergic rhinitis (20002:1387)	J30, J31	477
Bronchiectasis	Bronchiectasis (20002:1114)	J47, Q33.4	494, 748.61
Sleep apnoea	Sleep apnoea (20002:1123)	G47.3	327.2, 780.57
Pleurisy	Pleurisy (20002:1125)	R09.1	511.0, 511.1
Pneumothorax	Spontaneous pneumothorax / recurrent pneumothorax (20002:1126)	J93.0, J93.1, J9381	512.0, 512.81, 512.82
Chronic sinusitis	Chronic sinusitis (20002:1416)	J01, J32	461, 473
Nasal polyps	Nasal polyps (20002:1417)	J33	471
Tonsillitis	Tonsillitis (20002:1598)	J03, J35.0	463, 474.0
Meniere's disease	Meniere's disease (20002:1421)	H81.0	386.0
Tinnitus	Tinnitus (20002:1597)	H93.1	388.3
Infections			
Rheumatic fever	Rheumatic fever (20002:1479)	I00-I02	390, 391, 392
Meningitis	Meningitis (20002:1247)	G00-G02, A39	320–321
Measles	Measles / morbillivirus (20002:1568)	B05	055
Rubella	Rubella / German measles (20002:1570)	B06	056

Chickenpox	Chickenpox (20002:1571) or varicella (20002:1674)	B01	052
Shingles	Shingles (20002:1573)	B02	053
Varicella	Chickenpox (20002:1571), varicella (20002:1674) or shingles (20002:1573)	B01-B02	052, 053
Infectious mononucleosis	Infectious mononucleosis / glandular fever / Epstein-Barr virus (1567)	B27	075
Mumps	Mumps / epidemic parotitis (20002:1569)	B26	072
Helicobacter Pylori	Helicobacter pylori (20002:1442)	B9681	041.86
Tuberculosis	Tuberculosis (20002:1440), Age tuberculosis diagnosed by doctor (22157:age)	A15-A19	010-018
Pertussis	Whooping cough / pertussis (20002:1572)	A37	033
Scarlatina	Scarlet fever / scarlatina (20002:1677)	A38	034.1
Malaria	Malaria (20002:1441)	B50-B54	084
Eye Problems			
Retinal detachment	Retinal detachment (20002:1281)	H330, H332, H333, H334	3610, 3612, 3618
Diabetic eye disease	Diabetic eye disease (20002:1276), Age when diabetes-related eye disease diagnosed (5901:age)	H36, E10.3, E11.3, E12.3, E13.3, E14.3	250.5, 3620, 36641
Glaucoma	Glaucoma (20002:1277), Age glaucoma diagnosed (4689:age)	H40-H42	365
Cataract	Cataract (20002:1278), Age cataract diagnosed (4700:age)	H25-H26, H28, Q12.0	366
Genito-urinary diseases			
Chronic kidney diseases	Renal / kidney failure (20002:1192) requiring dialysis (20002:1193) or not requiring dialysis (20002:1194)	N18	585
Kidney stone	Kidney stone / ureter stone / bladder stone (20002:1197)	N20.0-N20.9, N21, N22, N13.2	592.0, 592.1,592.9,594
Immune			
Sarcoidosis	Sarcoidosis (20002:1371)	D86	135
Psoriasis	Psoriasis (20002:1453)	L40	696
Allergy / hypersensitivity / anaphylaxis	Allergy / hypersensitivity / anaphylaxis (20002:1374)	T78.2, T78.4, T780-T781, Z9101, Z9102, T886, Z88, Z91103-Z9109, K0855	995.0, V1381, 9956, 997, 99527, 52566, 9953, V150
Allergy or anaphylactic reaction to food	Allergy or anaphylactic reaction to food (20002:1385)	T780-T781, Z9101, Z9102	9956, 997
Allergy or anaphylactic reaction to drug	Allergy or anaphylactic reaction to drug (20002:1386)	T886, T887, Z88	99527
Polymyalgia rheumatica	Polymyalgia rheumatica (20002:1377)	M31.5, M35.3	725
Systemic lupus erythematosus	Systemic lupus erythematosus (20002:1381)	M32, H0112, L93	710.0, 37334, 6954
Sjogren's syndrome/sicca syndrome	Sjogren's syndrome / sicca syndrome (20002:1382)	M35.0	710.2
Eczema	Eczema (20002:1452) or contact dermatitis (20002:1669)	L20-L30	692, 6918
Cancer			
Lung cancer	Lung cancer, small cell or non-small cell lung cancer or trachea cancer (20001:[1001,1027,1028,1080], Age lung cancer (not mesothelioma) diagnosed by doctor (22160:age)	C33-C34	162
Colorectal cancer	Large bowel / colorectal cancer, colon cancer / sigmoid cancer, rectal cancer or anal cancer (20001:[1020, 1022, 1023, 1021])	C18-C21	153, 154.0-154.1
Thyroid cancer	Thyroid cancer (20001:1065)	C73	193

Oesophageal cancer	Oesophageal cancer (20001:1017)	C15	150
Stomach cancer	Stomach cancer (20001:1018)	C16	151
Liver cancer	Liver / hepatocellular cancer (20001:1024)	C22	155
Pancreas cancer	Pancreas cancer (20001:1026)	C25	157
Melanoma	Malignant melanoma (20001:1059)	C43	172
Skin cancer (including melanoma)	Skin cancer (20001:1003), malignant melanoma (20001:1059), non-melanoma skin cancer (20001:1060), basal cell carcinoma or squamous cell carcinoma (20001:1062)	C43-C44	172-173
Kidney cancer	Kidney / renal cell cancer (20001:1034)	C64, C65, C66	189
Bladder cancer	Bladder cancer (20001:1035)	C67	188
Non-Hodgkin lymphoma	Non-Hodgkin lymphoma (20001:1053)	C82-C86	200, 202
Lymphomas and multiple myeloma	Lymphoma (20001:1047), Hodgkin (20001:1052) or non-Hodgkin lymphoma (20001:1053), multiple myeloma (20001:1050)	C81-C88, C90	200-203
Leukaemia	Leukaemia (20001:1048), acute myeloid leukaemia (20001:1074), chronic lymphocytic (20001:1055) or chronic myeloid (20001:1056)	C91-C95	204-208
Brain cancer / primary malignant brain tumour	Brain cancer / primary malignant brain tumour (20001:1032)	C70, C71, C72, C75.1, C75.2, C75.3, C75.4, C75.5	191
Head and neck cancer	Cancer of larynx / throat (20001:1006), parotid gland (20001:1015), other salivary gland (20001:1016), lip (20001:1010), tongue (20001:1011), gum (20001:1012), mouth (20001:1077), tonsil (20001:1078), oropharynx/oropharyngeal (20001:1079), nasal cavity (20001:1007), sinus (20001:1009), lip/mouth/pharynx/oral cavity (20001:1004)	C32, C30, C00–C14	161, 140–149
Female-only			
Breast cysts	Breast cysts (20002:1367)	N60.0-N60.4	610.0-610.4
Benign breast lump	Breast lump (20002:1666)	D24, N608, N609	217, 6108, 6109
Breast cancer	Breast cancer (20001:1002)	C50, D05	174
Endometriosis	Endometriosis (20002:1402)	N80	617
Female infertility	Female infertility (20002:1403)	N97.0	628
Ovarian cyst	Ovarian cyst or cysts (20002:1349), polycystic ovaries / polycystic ovarian syndrome (20002:1350)	N83.0-N83.2, E282, D27	620.0-620.2, 2564
Uterine prolapse	Vaginal prolapse / uterine prolapse (20004:1353)	N81	618.0-618.4, 618.6-618.9
Uterine fibroid	Uterine fibroids (20002:1351)	D25	218
Uterine polyps	Uterine polyps (20002:1352)	N84.0, N84.1, D26	6210, 2190, 2191
Cervical cancer	Cervical cancer (20001:1041) or cin/pre-cancer cells cervix (20001:1072)	C53	180
Uterus cancer	Uterine / endometrial cancer (20001:1040)	C54-C55	179, 182
Ovary cancer	Ovarian cancer (20001:1039) or fallopian tube cancer (20001:1087)	C56	183
Male-only			
Benign prostatic hyperplasia	Enlarged prostate (20002:1396) or benign prostatic hypertrophy (20002:1516)	N40	600
Prostate cancer	Prostate cancer (20001:1044)	C61	185
Testicular cancer	Testicular cancer (20001:1045)	C62	186

Table S3. Diseases defined by operation codes.

Phenotype	OPCS-4	OPER (variable:value)
Cardiovascular diseases		
Coronary artery diseases (CAD)	K40-K46, K49, K50.1, K75	Coronary angioplasty (PTCA) stent (20004:1070), coronary artery bypass grafts (CABG) (20004:1095), triple heart bypass (20004:1523)
Varicose veins		Varicose vein surgery (20004:1479)
Digestive diseases		
Gallstone		Gallstones removed (20004:1528)
Peptic ulcer	G35, G52	Peptic ulcer surgery (20004:1566), Gastric ulcer surgery (20004:1567)
Haemorrhoids / piles		Haemorrhoidectomy / piles surgery / banding of piles (20004:1483)
Musculoskeletal diseases		
Intervertebral disc disorder	V29, V29.1-6, V29.8-9, V30, V30.1-6, V30.8-9, V31, V31.1-.4, V31.8-9, V32, V32.1-4, V32.8-9, V33, V33.1-9, V34, V34.1-9, V35, V35.1-2, V35.8-9, V36, V36.1-3, V36.8-9, V51, V51.1, V51.8-9, V52, V52.1-5, V52.8-9, V58, V58.1-3, V58.8-9, V59, V59.1-3, V59.8-9, V60, V60.1-3, V60.8-9, V61, V61.1-3, V61.8-9, V62, V62.1-3, V62.8-9, V63, V63.1-3, V63.8-9	
Respiratory diseases		
Nasal polyps		Nasal polyp surgery / nasal polypectomy (20004:1559)
Tonsillitis		Tonsillectomy +/- adenoids (20004:1478)
Glaucoma	C60.1-6, C60.8-9, C61.1-5, C61.8-9, C62.1-4, C62.8-9	Glaucoma surgery / trabeculectomy (20004:1436)
Eye Problems		
Cataract	C71.1-3, C71.8-9, C72.1-3, C72.8-9, C73.1-4, C73.8-9, C74.1-3, C74.8-9, C75.1-4, C75.8-9, C77.1-2, C77.6, C77.8-9	Cataract extraction / lens implant (20004:1435)
Genito-urinary diseases		
Kidney stone	M06.1, M09, M09.1-4, M09.8-9, M14, M14.1, M14.8-9, M16.4, M23.1, M26.1-3, M27.1-3, M28, M28.1-5, M28.8-9, M31, M31.1, M31.8-9, M39.1, M44.2, M75.4, M86.1	Percutaneous / open kidney stone surgery / lithotripsy (20004:1197)
Cancer		
Melanoma		Removal of malignant melanoma (20004:1593)
Skin cancer (including melanoma)		Removals of squamous cell carcinoma (20004:1595), malignant melanoma (20004:1593), rodent ulcer / basal cell carcinoma (20004:1596)
Female-only		
Breast cysts		Breast cyst / abscess removal (20004:1513)
Ovarian cyst	Q474, Q493	Ovarian cyst removal / surgery (20004:1506)
Uterine fibroid		Myomectomy / fibroids removed (20004:1509)
Uterine polyps		Uterine polypectomy/uterine polyps removed (20004:1539)

Appendix 4. Genetic risk score association study results

Table S4. Results of association analysis between telomere length genetic risk score and 127 diseases.

		GRScojo52		GRSclump234	
Disease phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**
Abdominal hernia	999	2.98E-01	0.967(0.907,1.030)	3.32E-01	0.969(0.908,1.033)
Allergy hypersensitivity	39873	9.78E-01	1.000(0.990,1.011)	3.33E-01	1.005(0.995,1.016)
Allergy to drug	39522	8.34E-01	1.001(0.991,1.012)	8.23E-02	1.010(0.999,1.021)
Allergy to food	2555	2.32E-01	1.025(0.985,1.066)	3.57E-01	1.019(0.979,1.061)
Anorexia nervosa	1602	1.50E-01	0.964(0.916,1.013)	5.62E-01	0.985(0.936,1.036)
Anxiety	16709	1.11E-01	0.987(0.972,1.003)	6.65E-01	0.996(0.981,1.013)
Aortic valve stenosis	2603	1.28E-02	1.052(1.011,1.095)	4.78E-03	1.060(1.018,1.104)
Appendicitis	8899	1.73E-01	1.015(0.993,1.037)	9.48E-01	1.001(0.979,1.023)
Asthma	65325	5.31E-02	1.008(1.000,1.017)	3.28E-01	1.004(0.996,1.013)
Atrial fibrillation	21203	9.71E-01	1.000(0.985,1.014)	1.44E-01	1.011(0.996,1.026)
Bladder cancer	3017	3.71E-03	0.947(0.913,0.982)	8.06E-01	0.995(0.959,1.033)
Brain cancer	922	1.30E-07	0.836(0.782,0.894)	7.99E-01	0.991(0.927,1.060)
Bronchiectasis	3907	6.54E-02	1.031(0.998,1.065)	6.72E-02	1.031(0.998,1.066)
Coronary artery disease	36974	9.95E-04	1.019(1.008,1.031)	1.92E-02	1.014(1.002,1.025)
Cataract	39713	2.85E-01	1.006(0.995,1.017)	2.25E-01	0.993(0.982,1.004)
Chickenpox	5755	3.77E-01	1.012(0.985,1.039)	3.99E-01	0.989(0.962,1.015)
Chronic fatigue syndrome	3903	5.78E-01	0.991(0.959,1.024)	2.26E-01	1.020(0.988,1.055)
Chronic kidney disease	8843	7.31E-01	0.996(0.975,1.018)	4.95E-01	0.992(0.971,1.014)
Coeliac disease	3005	1.16E-20	1.192(1.149,1.236)	2.37E-01	1.023(0.985,1.062)
Colorectal cancer	6736	6.02E-01	0.993(0.969,1.018)	2.57E-02	0.972(0.948,0.997)
Colorectal polyp	34018	4.76E-07	0.971(0.960,0.982)	3.70E-03	0.983(0.972,0.994)
COPD	19071	1.42E-03	1.025(1.010,1.041)	2.30E-02	1.018(1.002,1.034)
Dementia	2892	4.25E-01	1.016(0.978,1.055)	5.09E-01	1.013(0.975,1.053)
Depression	105158	3.94E-01	0.997(0.990,1.004)	2.26E-01	0.996(0.988,1.003)
Diabetes I	24498	4.56E-01	1.005(0.992,1.019)	1.32E-01	1.010(0.997,1.024)
Diabetes II	35139	7.44E-01	0.998(0.987,1.010)	3.30E-01	1.006(0.994,1.017)
Diabetes	35441	7.39E-01	0.998(0.987,1.009)	3.34E-01	1.006(0.994,1.017)
Diabetic eye disease	3858	8.08E-01	0.996(0.964,1.029)	7.78E-01	1.005(0.972,1.039)
Diverticulitis	39848	1.22E-03	0.983(0.972,0.993)	7.95E-04	0.982(0.971,0.992)
Eczema	17142	2.14E-02	1.019(1.003,1.035)	1.60E-02	1.020(1.004,1.036)
Epilepsy	6153	7.12E-01	0.995(0.970,1.021)	3.13E-01	0.987(0.961,1.013)
Gallstone	18902	2.66E-01	0.991(0.977,1.007)	2.96E-01	0.992(0.977,1.007)
Glaucoma	10443	3.97E-01	0.991(0.972,1.011)	9.41E-01	1.001(0.981,1.021)
GORD	48219	4.13E-01	0.996(0.986,1.006)	2.57E-01	1.006(0.996,1.016)
Gout	9614	3.73E-01	1.010(0.989,1.031)	6.84E-01	1.004(0.983,1.026)
Haemorrhoids	12795	2.10E-01	0.989(0.971,1.007)	5.97E-01	0.995(0.977,1.013)
Hay fever	49053	7.18E-03	1.013(1.004,1.023)	2.65E-01	1.006(0.996,1.015)
Hay fever / eczema	124509	5.33E-04	1.012(1.005,1.019)	7.77E-01	1.001(0.994,1.008)
Head and neck cancer	2953	2.69E-01	0.979(0.943,1.016)	1.02E-01	0.969(0.934,1.006)
Heart failure	8375	3.69E-01	1.010(0.988,1.033)	1.16E-01	1.018(0.995,1.042)
Helicobacter pylori	1407	4.45E-01	0.979(0.928,1.033)	8.43E-01	1.005(0.953,1.061)
Hepatitis	3186	8.71E-01	1.003(0.968,1.040)	6.30E-01	0.991(0.956,1.028)
Hiatus hernia	40173	9.07E-01	1.001(0.990,1.011)	1.97E-01	0.993(0.982,1.004)
Hypertension	293487	1.47E-14	0.975(0.969,0.981)	7.89E-05	0.987(0.980,0.993)
Hyperthyroid	5565	2.77E-07	1.074(1.045,1.103)	6.87E-02	1.026(0.998,1.054)
Hypothyroid	29377	1.23E-10	1.041(1.029,1.054)	3.32E-01	1.006(0.994,1.019)
Immunodeficiency	813	2.92E-01	0.962(0.896,1.034)	7.13E-01	1.014(0.944,1.089)
Inflammatory bowel disease	6761	1.45E-02	0.970(0.946,0.994)	9.75E-02	1.021(0.996,1.047)

Irritable bowel syndrome	16916	2.77E-01	0.991(0.976,1.007)	8.13E-01	0.998(0.982,1.014)
Inguinal hernia	21600	4.76E-02	0.986(0.971,1.000)	2.64E-02	0.984(0.970,0.998)
Intervertebral disc disorder	19689	2.43E-01	1.009(0.994,1.024)	7.07E-02	1.014(0.999,1.029)
Kidney cancer	1732	4.76E-05	0.904(0.861,0.949)	5.90E-03	0.934(0.890,0.981)
Kidney stone	10299	2.07E-02	0.977(0.957,0.996)	8.71E-01	0.998(0.978,1.019)
Leukemia	1624	1.30E-07	0.873(0.831,0.918)	5.41E-03	0.931(0.885,0.979)
Liver cirrhosis	3076	2.03E-02	1.044(1.007,1.083)	3.07E-02	1.042(1.004,1.081)
Lung cancer	3298	2.50E-09	0.898(0.866,0.930)	2.00E-03	0.945(0.912,0.980)
Systemic lupus erythematosus	897	4.90E-01	0.976(0.912,1.045)	8.16E-01	1.008(0.942,1.079)
Lymphoma	3948	1.37E-06	0.923(0.894,0.954)	2.98E-03	0.952(0.922,0.983)
Malaria	831	4.77E-01	0.975(0.909,1.045)	4.45E-01	0.973(0.908,1.043)
Mania / bipolar disorder	2053	5.10E-01	1.015(0.971,1.062)	9.16E-03	1.062(1.015,1.111)
Measles	4574	7.80E-02	1.027(0.997,1.058)	5.22E-01	0.990(0.961,1.020)
Melanoma	5362	4.76E-09	0.921(0.895,0.946)	5.14E-06	0.938(0.912,0.964)
Meniere's disease	1702	3.32E-01	0.976(0.930,1.025)	7.07E-01	0.991(0.943,1.040)
Meningitis	2095	2.31E-01	0.973(0.931,1.017)	3.17E-01	0.978(0.935,1.022)
Migraine	17207	8.09E-01	1.002(0.986,1.018)	5.91E-01	1.004(0.989,1.020)
Mononucleosis	1004	7.36E-02	1.059(0.994,1.129)	2.34E-02	1.077(1.010,1.149)
Multiple sclerosis	2077	1.58E-01	0.968(0.926,1.013)	6.47E-01	1.011(0.966,1.057)
Mumps	2814	8.78E-01	1.003(0.966,1.042)	9.64E-02	0.968(0.932,1.006)
Nasal polyps	5987	9.32E-01	0.999(0.973,1.025)	2.37E-01	0.984(0.959,1.011)
Non-Hodgkin lymphoma	2641	4.89E-03	0.945(0.908,0.983)	3.51E-02	0.958(0.921,0.997)
Oesophageal cancer	1034	6.51E-01	0.986(0.926,1.049)	9.73E-01	1.001(0.940,1.066)
Oesophagitis	4895	2.71E-02	0.968(0.940,0.996)	9.54E-01	0.999(0.970,1.029)
Osteoarthritis	81976	9.36E-01	1.000(0.992,1.008)	4.22E-01	1.003(0.995,1.011)
Osteoporosis	14950	9.16E-01	1.001(0.984,1.018)	1.26E-01	1.014(0.996,1.031)
Otitis media	2221	6.69E-01	0.991(0.949,1.034)	5.22E-01	1.014(0.971,1.059)
Pancreas cancer	949	1.01E-01	1.056(0.989,1.128)	8.37E-02	1.060(0.992,1.133)
Pancreatitis	3361	4.93E-01	1.012(0.978,1.048)	9.27E-01	0.998(0.964,1.034)
Parkinson's disease	2024	9.15E-01	0.998(0.953,1.044)	1.83E-01	0.970(0.926,1.015)
Peptic ulcer	14497	8.38E-01	1.002(0.985,1.019)	8.65E-01	1.001(0.984,1.019)
Peripheral vascular disease	5194	9.63E-01	0.999(0.971,1.028)	9.99E-01	1.000(0.972,1.029)
Peritonitis	2698	6.83E-01	1.008(0.970,1.048)	5.14E-01	1.013(0.974,1.054)
Pertussis	1093	1.56E-01	1.045(0.983,1.111)	7.68E-01	1.009(0.949,1.073)
Pleurisy	2303	4.95E-01	1.015(0.973,1.058)	8.19E-01	0.995(0.954,1.038)
Pneumonia	28981	7.84E-01	1.002(0.989,1.014)	6.91E-02	1.012(0.999,1.024)
Pneumothorax	1287	9.60E-01	0.999(0.944,1.056)	2.12E-01	1.037(0.980,1.097)
Polymyalgia rheumatica	2455	5.78E-01	0.988(0.949,1.030)	9.88E-01	1.000(0.959,1.042)
Psoriasis	7402	6.26E-01	1.006(0.982,1.030)	5.66E-01	0.993(0.970,1.017)
Raynaud's phenomenon	5707	5.46E-01	1.008(0.981,1.036)	6.38E-01	0.993(0.967,1.021)
Retinal detachment	5047	1.63E-02	0.966(0.938,0.994)	9.86E-01	1.000(0.971,1.029)
Rheumatic fever	1412	5.98E-01	1.014(0.962,1.070)	2.92E-01	1.029(0.975,1.087)
Rheumatoid arthritis	8907	1.03E-02	1.029(1.007,1.051)	6.71E-02	1.021(0.999,1.043)
Rubella	1581	8.55E-01	1.005(0.955,1.057)	1.35E-01	0.962(0.914,1.012)
Sarcoidosis	1536	9.43E-04	1.091(1.036,1.149)	1.02E-02	1.071(1.016,1.129)
Scarlatina	925	5.70E-01	0.981(0.918,1.048)	7.41E-01	0.989(0.925,1.057)
Schizophrenia	1509	2.09E-01	0.967(0.917,1.019)	7.67E-02	1.049(0.995,1.107)
Sciatica	8003	6.34E-01	1.006(0.983,1.029)	8.32E-01	1.002(0.980,1.026)
Shingles	1570	7.49E-01	0.992(0.942,1.044)	7.53E-01	0.992(0.942,1.044)
Sinusitis	6797	9.20E-01	1.001(0.977,1.026)	3.13E-01	1.013(0.988,1.038)
Sjogren's syndrome	939	8.35E-01	1.007(0.943,1.076)	8.59E-01	0.994(0.930,1.062)
Skin cancer	23320	1.61E-25	0.930(0.917,0.943)	4.10E-05	0.972(0.959,0.985)
Sleep apnoea	6951	2.30E-01	0.985(0.961,1.009)	6.00E-01	0.993(0.969,1.018)
Spondylitis	7250	6.75E-02	0.978(0.955,1.002)	4.24E-01	1.010(0.986,1.034)
Stomach cancer	895	1.89E-01	0.956(0.894,1.023)	9.80E-02	0.945(0.883,1.011)
Stroke	22833	8.97E-01	0.999(0.985,1.013)	5.58E-01	1.004(0.990,1.018)

Thyroid cancer	756	4.45E-06	0.842(0.782,0.906)	9.43E-04	0.884(0.821,0.951)
Tinnitus	2274	2.02E-01	1.028(0.985,1.072)	5.97E-02	1.042(0.998,1.088)
Tonsilitis	74060	5.35E-01	0.997(0.989,1.006)	7.16E-01	0.998(0.990,1.007)
Tuberculosis	2952	7.21E-03	1.052(1.014,1.092)	3.28E-01	1.019(0.981,1.058)
Umbilical hernia	5492	7.34E-01	0.995(0.968,1.023)	5.95E-01	1.008(0.980,1.036)
Varicella	6970	3.98E-01	1.011(0.986,1.035)	5.72E-01	0.993(0.969,1.018)
Varicose veins	53187	5.99E-02	0.991(0.982,1.000)	5.22E-01	1.003(0.994,1.013)
Venous thromboembolism	19170	3.50E-02	0.984(0.969,0.999)	5.05E-01	1.005(0.990,1.020)
Benign breast lump	4126	1.16E-03	0.948(0.919,0.979)	2.26E-02	0.963(0.933,0.995)
Breast cancer	17691	1.55E-01	0.988(0.973,1.004)	5.10E-02	0.984(0.968,1.000)
Breast cyst	7370	6.41E-04	0.959(0.936,0.982)	6.55E-01	0.994(0.971,1.019)
Cervical cancer	5046	8.41E-01	1.004(0.970,1.039)	3.84E-01	1.016(0.981,1.051)
Endometriosis	7753	1.02E-03	0.962(0.940,0.984)	3.09E-01	0.988(0.965,1.011)
Female infertility	1084	6.16E-01	1.016(0.956,1.080)	8.99E-01	0.996(0.936,1.060)
Ovarian cyst	14128	1.34E-06	0.958(0.941,0.975)	4.29E-01	0.993(0.976,1.011)
Ovary cancer	1689	4.69E-02	0.951(0.905,0.999)	5.16E-02	0.952(0.906,1.000)
Uterine fibroid	20556	1.49E-46	0.898(0.885,0.911)	3.03E-14	0.944(0.931,0.958)
Uterine polyps	14196	9.37E-15	0.933(0.917,0.950)	1.39E-05	0.962(0.945,0.979)
Uterine prolapse	15457	5.68E-01	0.995(0.978,1.012)	7.87E-02	0.985(0.968,1.002)
Uterus cancer	2286	2.16E-01	0.973(0.933,1.016)	3.93E-02	0.956(0.916,0.998)
Benign prostatic hyperplasia	19765	3.98E-25	0.922(0.908,0.936)	2.99E-06	0.964(0.949,0.979)
Prostate cancer	8967	2.04E-09	0.935(0.914,0.956)	1.96E-07	0.943(0.923,0.964)
Testicular cancer	885	6.75E-03	1.099(1.026,1.177)	5.88E-02	1.069(0.998,1.146)

Appendix 5. Mendelian randomisation study results

Table S5. Results of causal inference between telomere length and 127 diseases.

		IVW		Egger's		Median		RAPS		Egger's Int		Steiger	
Phenotype	**N(Case)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**OR(95%CI)**	**P-value**	**Beta**	**P-value**	**Direction**
Abdominal hernia	999	1.30E-01	0.744(0.508,1.091)	2.58E-01	0.585(0.233,1.466)	2.02E-01	0.690(0.390,1.220)	1.23E-01	0.728(0.487,1.089)	5.75E-01	0.01238	2.93E-02	T
Allergy hypersensitivity	39873	9.17E-01	0.996(0.925,1.072)	4.79E-01	0.937(0.784,1.120)	3.88E-01	1.042(0.949,1.144)	9.52E-01	1.002(0.926,1.085)	4.65E-01	0.00314	1.79E-127	T
Allergy to drug	39522	6.30E-01	1.017(0.948,1.092)	3.26E-01	0.918(0.776,1.086)	7.21E-01	1.017(0.926,1.117)	4.66E-01	1.028(0.955,1.107)	1.95E-01	0.00527	8.77E-129	T
Allergy to food	2555	4.62E-01	1.090(0.866,1.372)	1.07E-01	1.590(0.915,2.765)	6.60E-01	1.084(0.756,1.555)	7.91E-01	1.033(0.812,1.315)	1.48E-01	-0.01943	7.65E-09	T
Anorexia nervosa	1602	2.45E-01	0.841(0.629,1.126)	7.20E-01	0.879(0.436,1.773)	3.59E-01	0.820(0.536,1.254)	1.75E-01	0.809(0.596,1.099)	8.94E-01	-0.00225	1.79E-04	T
Anxiety	16709	7.85E-02	0.898(0.796,1.012)	6.97E-01	1.059(0.794,1.413)	9.91E-01	1.001(0.862,1.162)	6.42E-02	0.884(0.776,1.007)	2.21E-01	-0.00851	6.38E-55	T
Aortic valve stenosis	2603	2.44E-03	1.427(1.134,1.797)	4.34E-01	1.249(0.719,2.169)	6.17E-01	1.092(0.773,1.543)	1.02E-02	1.373(1.078,1.750)	6.04E-01	0.00687	2.95E-08	T
Appendicitis	8899	2.33E-01	1.080(0.952,1.225)	5.35E-02	1.357(1.003,1.836)	6.22E-01	1.048(0.870,1.263)	1.92E-01	1.095(0.955,1.254)	1.10E-01	-0.01178	4.48E-34	T
Asthma	65325	1.86E-01	1.062(0.972,1.160)	3.59E-01	1.107(0.893,1.371)	6.01E-01	1.022(0.941,1.111)	1.77E-01	1.037(0.984,1.094)	6.79E-01	-0.00214	5.75E-166	T
Atrial fibrillation	21203	5.17E-01	1.039(0.925,1.167)	7.66E-01	0.958(0.724,1.268)	5.91E-01	1.037(0.909,1.183)	1.65E-01	1.080(0.969,1.204)	5.36E-01	0.00418	1.32E-70	T
Bladder cancer	3017	1.07E-01	0.807(0.622,1.048)	6.59E-02	0.550(0.295,1.026)	2.65E-02	0.655(0.451,0.952)	1.65E-01	0.832(0.642,1.079)	1.91E-01	0.01968	3.19E-08	T
Brain cancer	922	3.90E-03	0.412(0.226,0.752)	3.72E-02	0.206(0.048,0.875)	8.33E-01	0.929(0.470,1.838)	1.65E-01	0.659(0.366,1.187)	3.06E-01	0.03587	9.48E-01	T
Bronchiectasis	3907	2.07E-03	1.343(1.113,1.620)	1.26E-02	1.810(1.155,2.837)	8.44E-02	1.277(0.967,1.686)	6.02E-03	1.320(1.083,1.609)	1.58E-01	-0.01539	1.02E-13	T
Coronary artery disease	36974	1.83E-03	1.133(1.047,1.226)	1.69E-01	1.145(0.946,1.386)	3.96E-02	1.116(1.005,1.239)	6.86E-03	1.118(1.031,1.212)	9.03E-01	-0.00056	1.28E-117	T
Cataract	39713	9.71E-01	1.001(0.932,1.075)	3.78E-01	1.081(0.911,1.283)	6.92E-01	1.020(0.925,1.124)	9.89E-01	1.001(0.926,1.081)	3.41E-01	-0.00394	1.17E-127	T
Chickenpox	5755	9.99E-01	1.000(0.857,1.167)	5.88E-01	1.109(0.765,1.607)	7.69E-01	1.036(0.820,1.307)	9.38E-01	1.007(0.843,1.203)	5.51E-01	-0.00532	1.43E-21	T
Chronic fatigue syndrome	3903	6.62E-02	0.746(0.546,1.020)	4.92E-02	0.463(0.219,0.979)	7.22E-02	0.685(0.454,1.035)	8.71E-02	0.755(0.547,1.042)	1.76E-01	0.02447	7.12E-05	T
Chronic kidney disease	8843	8.97E-01	0.991(0.866,1.134)	5.83E-01	0.912(0.659,1.263)	2.50E-01	0.888(0.726,1.087)	9.08E-01	1.009(0.872,1.166)	5.85E-01	0.00427	3.33E-32	T
Coeliac disease	3005	2.66E-01	1.744(0.655,4.644)	7.06E-02	8.950(0.875,91.546)	7.34E-03	1.635(1.141,2.342)	1.19E-01	1.443(0.910,2.287)	1.36E-01	-0.08455	4.34E-03	F
Colorectal cancer	6736	7.26E-01	0.971(0.825,1.144)	9.07E-01	1.024(0.689,1.523)	6.60E-01	1.054(0.834,1.333)	6.40E-01	0.959(0.805,1.143)	7.75E-01	-0.00272	3.11E-23	T
Colorectal polyp	34018	3.22E-02	0.895(0.809,0.991)	8.82E-03	0.719(0.567,0.911)	1.03E-02	0.862(0.769,0.965)	4.56E-02	0.900(0.812,0.998)	5.21E-02	0.01127	1.06E-101	T
COPD	19071	4.55E-02	1.113(1.002,1.237)	6.94E-02	1.269(0.987,1.632)	3.76E-01	1.065(0.926,1.225)	9.25E-02	1.083(0.987,1.189)	2.67E-01	-0.00674	3.97E-68	T
Dementia	2892	1.74E-01	1.187(0.927,1.519)	8.61E-01	1.055(0.582,1.913)	7.85E-01	1.049(0.745,1.476)	2.49E-01	1.169(0.897,1.523)	6.72E-01	0.00610	2.22E-08	T
Depression	105158	4.11E-01	0.979(0.932,1.029)	6.76E-01	1.026(0.910,1.157)	4.23E-01	0.974(0.912,1.039)	5.28E-01	0.984(0.934,1.036)	4.07E-01	-0.00239	7.46E-226	T
Diabetes I	24498	6.25E-02	1.088(0.996,1.188)	7.26E-01	1.039(0.840,1.286)	4.77E-01	1.043(0.928,1.173)	9.20E-02	1.087(0.986,1.198)	6.46E-01	0.00236	4.31E-86	T
Diabetes II	35139	3.04E-01	1.043(0.962,1.131)	8.05E-01	1.025(0.843,1.246)	5.36E-01	1.035(0.928,1.154)	2.01E-01	1.059(0.970,1.158)	8.46E-01	0.00091	2.33E-114	T
Diabetes	35441	2.96E-01	1.044(0.963,1.131)	8.81E-01	1.015(0.836,1.232)	4.94E-01	1.038(0.933,1.156)	2.12E-01	1.058(0.968,1.156)	7.57E-01	0.00145	3.47E-115	T
Diabetic eye disease	3858	8.94E-01	1.013(0.835,1.230)	5.66E-01	0.872(0.549,1.387)	5.42E-01	0.913(0.681,1.224)	8.64E-01	1.019(0.824,1.259)	4.89E-01	0.00781	6.15E-13	T
Diverticulitis	39848	4.24E-02	0.923(0.854,0.997)	1.30E-01	0.864(0.716,1.041)	1.22E-01	0.926(0.839,1.021)	8.89E-02	0.939(0.874,1.010)	4.46E-01	0.00343	5.72E-123	T
Eczema	17142	1.01E-02	1.185(1.041,1.348)	5.79E-01	1.093(0.799,1.496)	5.03E-02	1.159(1.000,1.344)	3.96E-02	1.149(1.007,1.311)	5.84E-01	0.00412	3.08E-57	T
Epilepsy	6153	7.69E-01	0.976(0.829,1.149)	8.76E-01	0.969(0.652,1.440)	4.37E-01	0.915(0.732,1.144)	9.45E-01	0.994(0.845,1.170)	9.69E-01	0.00037	1.99E-21	T

Migraine	17207	8.18E-01	1.014(0.901,1.141)	5.73E-01	1.087(0.816,1.448)	5.13E-01	1.048(0.910,1.207)	5.27E-01	1.038(0.926,1.163)	6.05E-01	-0.00356	2.38E-60	T
Mononucleosis	1004	3.12E-01	1.233(0.822,1.852)	4.55E-01	1.461(0.544,3.924)	5.17E-02	1.745(0.996,3.059)	3.77E-01	1.218(0.787,1.885)	7.13E-01	-0.00871	3.62E-02	T
Multiple sclerosis	2077	7.84E-02	0.722(0.502,1.038)	9.57E-02	0.467(0.194,1.124)	9.29E-03	0.588(0.394,0.877)	6.01E-02	0.716(0.505,1.014)	2.92E-01	0.02223	9.07E-04	T
Mumps	2814	5.76E-01	0.930(0.722,1.199)	7.53E-01	1.105(0.596,2.045)	4.08E-01	1.158(0.818,1.638)	5.18E-01	0.912(0.690,1.206)	5.51E-01	-0.00883	3.72E-08	T
Nasal polyp	5987	5.47E-01	1.048(0.899,1.222)	8.27E-01	1.043(0.717,1.516)	3.06E-01	1.127(0.896,1.417)	5.84E-01	1.046(0.889,1.231)	9.77E-01	0.00026	3.91E-22	T
Non-Hodgkin lymphoma	2641	9.35E-03	0.705(0.541,0.918)	1.12E-01	0.588(0.309,1.119)	1.15E-03	0.546(0.379,0.786)	4.14E-03	0.673(0.514,0.882)	5.48E-01	0.00927	7.41E-07	T
Oesophageal cancer	1034	4.67E-01	0.863(0.579,1.285)	1.55E-01	1.971(0.784,4.954)	4.33E-01	0.795(0.448,1.410)	7.53E-01	0.931(0.595,1.456)	5.83E-02	-0.04273	2.62E-02	T
Oesophagitis	4895	1.35E-01	0.863(0.712,1.047)	5.47E-01	0.866(0.543,1.380)	3.37E-02	0.749(0.574,0.978)	6.90E-02	0.847(0.708,1.013)	9.89E-01	-0.00015	6.64E-16	T
Osteoarthritis	81976	5.39E-01	0.978(0.912,1.049)	9.04E-01	0.990(0.835,1.173)	6.97E-02	0.929(0.858,1.006)	3.29E-01	0.965(0.899,1.036)	8.86E-01	-0.00059	1.87E-192	T
Osteoporosis	14950	8.93E-01	1.007(0.912,1.111)	5.50E-01	0.929(0.732,1.179)	6.07E-01	0.961(0.827,1.117)	9.85E-01	0.999(0.901,1.108)	4.73E-01	0.00410	1.46E-58	T
Otitis media	2221	6.48E-01	1.069(0.803,1.423)	8.01E-01	0.913(0.454,1.839)	8.85E-01	0.971(0.649,1.453)	5.60E-01	1.094(0.809,1.479)	6.31E-01	0.00802	7.60E-06	T
Pancreas cancer	949	1.53E-01	1.370(0.889,2.109)	8.11E-03	3.997(1.493,10.701)	2.81E-01	1.372(0.773,2.436)	1.21E-01	1.433(0.910,2.259)	2.28E-02	-0.05555	5.85E-02	T
Pancreatitis	3361	4.72E-01	1.093(0.858,1.392)	8.48E-01	0.944(0.526,1.694)	7.80E-01	1.045(0.768,1.420)	6.54E-01	1.053(0.841,1.319)	5.92E-01	0.00752	1.40E-09	T
Parkinson's disease	2024	6.52E-01	0.942(0.725,1.223)	4.28E-01	1.292(0.689,2.421)	7.90E-01	0.949(0.644,1.398)	4.93E-01	0.909(0.691,1.195)	2.83E-01	-0.01629	1.72E-06	T
Peptic ulcer	14497	9.52E-01	1.003(0.909,1.107)	4.50E-01	1.097(0.865,1.391)	6.05E-01	1.038(0.901,1.197)	8.72E-01	1.009(0.909,1.119)	4.22E-01	-0.00460	5.88E-56	T
Peripheral vascular disease	5194	7.97E-01	0.975(0.803,1.183)	9.29E-01	0.979(0.613,1.564)	1.97E-01	0.838(0.641,1.096)	6.58E-01	0.955(0.778,1.172)	9.86E-01	-0.00020	5.82E-17	T
Peritonitis	2698	3.11E-01	1.127(0.894,1.422)	4.35E-01	0.801(0.461,1.392)	8.87E-01	0.976(0.703,1.357)	4.02E-01	1.107(0.873,1.404)	1.89E-01	0.01760	2.78E-08	T
Pertussis	1093	1.28E-01	1.314(0.924,1.869)	7.43E-02	2.208(0.942,5.173)	1.06E-01	1.588(0.906,2.781)	2.16E-01	1.264(0.872,1.834)	1.96E-01	-0.02657	7.98E-03	T
Pleurisy	2303	3.90E-01	1.115(0.870,1.429)	2.67E-01	1.408(0.775,2.560)	6.65E-01	1.085(0.751,1.566)	2.74E-01	1.160(0.889,1.515)	4.03E-01	-0.01201	4.32E-07	T
Pneumonia	28981	4.39E-01	1.030(0.956,1.110)	3.72E-01	1.087(0.907,1.302)	4.58E-01	0.960(0.862,1.069)	6.16E-01	1.021(0.941,1.107)	5.27E-01	-0.00276	3.87E-103	T
Pneumothorax	1287	7.37E-01	0.946(0.682,1.311)	8.64E-02	0.495(0.225,1.088)	9.38E-01	1.019(0.629,1.653)	8.34E-01	0.964(0.686,1.356)	8.30E-02	0.03318	4.25E-04	T
Polymyalgia rheumatica	2455	7.15E-01	1.048(0.815,1.347)	8.61E-01	1.057(0.572,1.953)	5.93E-01	1.099(0.777,1.554)	9.86E-01	1.002(0.780,1.287)	9.77E-01	-0.00042	3.70E-07	T
Psoriasis	7402	2.49E-01	0.901(0.754,1.076)	5.30E-01	1.148(0.749,1.758)	4.02E-01	1.098(0.882,1.367)	3.13E-01	0.919(0.780,1.083)	2.27E-01	-0.01241	7.32E-24	T
Raynaud's phenomenon	5707	7.64E-01	1.029(0.852,1.243)	9.23E-01	0.978(0.619,1.544)	4.23E-01	0.896(0.685,1.172)	6.34E-01	1.049(0.862,1.277)	8.09E-01	0.00266	1.40E-18	T
Retinal detachment	5047	1.39E-01	0.824(0.637,1.065)	9.26E-03	0.439(0.242,0.797)	3.06E-02	0.729(0.548,0.971)	9.83E-02	0.798(0.610,1.043)	2.73E-02	0.03233	3.65E-10	T
Rheumatic fever	1412	9.15E-01	1.017(0.746,1.386)	7.86E-01	1.110(0.525,2.347)	6.61E-01	1.111(0.693,1.781)	9.03E-01	1.020(0.738,1.411)	8.02E-01	-0.00448	1.37E-04	T
Rheumatoid arthritis	8907	7.81E-03	1.290(1.069,1.557)	2.02E-01	1.350(0.857,2.128)	1.21E-02	1.283(1.056,1.558)	2.28E-02	1.218(1.028,1.444)	8.30E-01	-0.00235	4.19E-26	T
Rubella	1581	9.15E-01	1.018(0.736,1.407)	1.72E-01	1.724(0.797,3.726)	1.18E-01	1.453(0.909,2.323)	9.19E-01	1.019(0.708,1.465)	1.47E-01	-0.02702	4.93E-04	T
Sarcoidosis	1536	4.77E-03	1.666(1.169,2.375)	2.72E-02	2.649(1.145,6.132)	4.61E-02	1.591(1.008,2.513)	1.99E-03	1.680(1.209,2.334)	2.38E-01	-0.02422	3.07E-03	T
Scarlatina	925	6.08E-01	0.894(0.583,1.372)	8.41E-01	0.899(0.317,2.549)	9.59E-01	1.016(0.543,1.902)	6.76E-01	0.907(0.573,1.435)	9.92E-01	-0.00026	6.72E-02	T
Schizophrenia	1509	3.98E-02	0.635(0.412,0.979)	3.07E-01	1.692(0.623,4.596)	3.76E-01	0.783(0.455,1.346)	6.92E-02	0.675(0.442,1.031)	3.93E-02	-0.05083	8.28E-02	T
Sciatica	8003	2.09E-01	1.088(0.954,1.240)	7.25E-01	1.059(0.772,1.451)	1.70E-01	1.151(0.941,1.407)	2.10E-01	1.092(0.952,1.252)	8.55E-01	0.00138	2.82E-31	T
Shingles	1570	6.69E-01	0.937(0.697,1.261)	6.92E-01	1.158(0.563,2.384)	6.22E-01	0.901(0.595,1.364)	5.16E-01	0.901(0.659,1.233)	5.31E-01	-0.01087	2.19E-04	T

Gallstone	18902	2.80E-01	0.942(0.844,1.050)	8.63E-01	0.977(0.750,1.273)	1.29E-01	0.895(0.776,1.033)	2.33E-01	0.938(0.844,1.042)	7.66E-01	-0.00189	4.81E-67	T
Glaucoma	10443	4.40E-01	0.947(0.824,1.088)	1.47E-01	0.780(0.561,1.086)	7.57E-01	1.032(0.846,1.259)	2.53E-01	0.918(0.792,1.064)	2.12E-01	0.01000	1.20E-36	T
GORD	48219	7.14E-01	0.988(0.925,1.055)	5.85E-01	0.956(0.815,1.122)	2.47E-01	0.949(0.868,1.037)	6.48E-01	0.984(0.920,1.053)	6.64E-01	0.00167	1.92E-147	T
Gout	9614	2.11E-01	1.080(0.957,1.219)	8.32E-01	1.032(0.772,1.380)	3.40E-01	0.916(0.764,1.097)	4.45E-01	1.051(0.926,1.192)	7.38E-01	0.00234	3.37E-38	T
Haemorrhoids	12795	7.59E-03	0.868(0.782,0.963)	7.75E-02	0.794(0.618,1.020)	4.92E-02	0.855(0.732,0.999)	7.03E-03	0.859(0.770,0.959)	4.50E-01	0.00456	9.42E-49	T
Hay fever	49053	8.09E-01	1.009(0.935,1.090)	4.06E-01	1.082(0.899,1.302)	8.06E-02	1.085(0.990,1.188)	2.05E-01	1.039(0.979,1.104)	4.22E-01	-0.00358	1.73E-143	T
Hay fever / eczema	124509	2.29E-01	1.039(0.976,1.106)	2.39E-01	1.096(0.942,1.276)	2.49E-01	1.044(0.971,1.122)	4.34E-01	1.026(0.963,1.093)	4.50E-01	-0.00276	3.73E-233	T
Head and neck cancer	2953	7.37E-01	0.957(0.742,1.235)	7.67E-01	1.099(0.590,2.048)	5.51E-01	0.901(0.638,1.271)	3.55E-01	0.897(0.714,1.128)	6.35E-01	-0.00709	3.63E-08	T
Heart failure	8375	1.02E-01	1.139(0.975,1.331)	9.92E-01	1.002(0.689,1.457)	1.14E-01	1.173(0.963,1.429)	1.42E-01	1.119(0.963,1.300)	4.64E-01	0.00663	1.67E-28	T
Helicobacter pylori	1407	4.56E-01	0.889(0.653,1.211)	1.42E-01	0.570(0.273,1.192)	3.72E-01	0.816(0.522,1.276)	4.20E-01	0.875(0.633,1.210)	1.99E-01	0.02297	3.31E-04	T
Hepatitis	3186	9.20E-01	1.011(0.821,1.244)	6.75E-01	1.113(0.676,1.834)	4.13E-01	1.144(0.828,1.581)	9.55E-01	1.006(0.809,1.253)	6.78E-01	-0.00499	1.65E-10	T
Hiatus hernia	40173	3.00E-01	1.042(0.964,1.125)	1.27E-01	1.157(0.963,1.391)	3.15E-01	1.049(0.956,1.150)	1.89E-01	1.053(0.975,1.137)	2.24E-01	-0.00542	6.15E-126	T
Hypertension	293487	3.72E-02	0.924(0.859,0.995)	1.14E-02	0.793(0.667,0.943)	3.41E-07	0.859(0.810,0.911)	1.37E-02	0.919(0.860,0.983)	6.15E-02	0.00792	2.64E-268	T
Hyperthyroid	5565	9.09E-03	1.402(1.088,1.806)	2.15E-01	1.484(0.802,2.745)	3.25E-03	1.483(1.141,1.929)	4.02E-02	1.287(1.011,1.639)	8.43E-01	-0.00293	2.50E-14	T
Hypothyroid	29377	2.53E-05	1.339(1.169,1.534)	6.62E-02	1.372(0.986,1.908)	8.22E-08	1.399(1.237,1.582)	1.00E-03	1.235(1.089,1.401)	8.74E-01	-0.00125	3.29E-80	T
Immunodeficiency	813	3.21E-01	0.811(0.537,1.226)	2.47E-01	0.554(0.207,1.488)	3.93E-01	0.768(0.420,1.406)	3.65E-01	0.815(0.524,1.268)	4.09E-01	0.01988	7.48E-02	T
Inflammatory bowel disease	6761	2.33E-01	0.894(0.743,1.075)	1.02E-01	0.686(0.440,1.068)	8.67E-01	0.979(0.769,1.247)	1.53E-01	0.872(0.723,1.052)	2.04E-01	0.01361	1.55E-21	T
Irritable bowel syndrome	16916	6.85E-01	1.020(0.928,1.120)	1.02E-01	0.829(0.664,1.034)	7.70E-01	0.980(0.853,1.125)	8.98E-01	1.006(0.914,1.108)	4.87E-02	0.01064	4.49E-64	T
Inguinal hernia	21600	4.40E-01	0.957(0.855,1.070)	5.15E-01	0.913(0.695,1.199)	5.57E-01	0.963(0.849,1.092)	4.64E-01	0.963(0.869,1.066)	7.12E-01	0.00242	1.74E-71	T
Intervertebral disc disorder	19689	6.89E-01	1.025(0.909,1.155)	3.29E-01	1.156(0.867,1.542)	7.43E-01	0.978(0.856,1.117)	8.66E-01	0.992(0.905,1.087)	3.71E-01	-0.00620	2.70E-66	T
Kidney cancer	1732	1.92E-02	0.642(0.443,0.930)	1.49E-03	0.230(0.098,0.542)	5.28E-04	0.442(0.278,0.701)	8.45E-03	0.595(0.405,0.876)	1.30E-02	0.05222	4.66E-03	T
Kidney stone	10299	1.09E-01	0.903(0.798,1.023)	8.16E-01	1.036(0.769,1.398)	3.78E-01	0.921(0.766,1.106)	1.00E-01	0.902(0.798,1.020)	3.26E-01	-0.00707	4.52E-38	T
Leukemia	1624	4.85E-05	0.473(0.329,0.678)	1.09E-03	0.222(0.095,0.520)	1.49E-04	0.396(0.245,0.639)	3.06E-04	0.502(0.345,0.730)	6.14E-02	0.03875	9.61E-03	T
Liver cirrhosis	3076	6.08E-02	1.224(0.991,1.511)	2.28E-01	1.370(0.827,2.271)	4.08E-01	1.141(0.835,1.559)	7.43E-02	1.225(0.980,1.532)	6.32E-01	-0.00583	3.16E-10	T
Lung cancer	3298	6.17E-06	0.552(0.426,0.714)	8.13E-02	0.566(0.303,1.060)	1.27E-02	0.641(0.452,0.909)	1.25E-04	0.584(0.443,0.769)	9.29E-01	-0.00135	2.22E-07	T
Systemic lupus erythematosus	897	8.29E-01	0.947(0.576,1.556)	9.58E-01	0.968(0.289,3.239)	1.15E-01	0.621(0.343,1.124)	2.37E-01	0.779(0.514,1.179)	9.69E-01	-0.00114	1.39E-01	T
Lymphoma	3948	8.45E-05	0.604(0.470,0.777)	2.26E-02	0.480(0.261,0.885)	4.99E-07	0.469(0.349,0.630)	1.13E-05	0.572(0.446,0.734)	4.24E-01	0.01173	3.26E-09	T
Malaria	831	1.49E-01	0.748(0.504,1.109)	2.80E-01	1.672(0.665,4.202)	1.07E-01	0.600(0.322,1.116)	1.25E-01	0.722(0.477,1.094)	6.40E-02	-0.04300	2.52E-02	T
Mania / bipolar disorder	2053	4.57E-01	1.110(0.843,1.462)	3.05E-01	1.418(0.732,2.746)	4.43E-01	1.162(0.792,1.705)	3.59E-01	1.152(0.851,1.559)	4.28E-01	-0.01266	7.97E-06	T
Measles	4574	3.95E-01	1.082(0.903,1.296)	4.57E-02	1.559(1.019,2.384)	2.73E-01	1.163(0.888,1.522)	2.57E-01	1.122(0.920,1.368)	6.95E-02	-0.01880	6.06E-16	T
Melanoma	5362	4.53E-03	0.690(0.534,0.892)	5.54E-02	0.536(0.288,0.999)	1.70E-03	0.650(0.497,0.851)	4.34E-05	0.696(0.585,0.828)	3.88E-01	0.01289	3.87E-12	T
Meniere's disease	1702	4.38E-01	0.894(0.675,1.186)	8.12E-01	0.920(0.466,1.817)	5.60E-01	0.880(0.572,1.352)	3.51E-01	0.868(0.644,1.169)	9.28E-01	-0.00148	3.07E-05	T
Meningitis	2095	6.07E-02	0.787(0.613,1.011)	8.81E-01	1.048(0.571,1.921)	6.87E-01	0.926(0.637,1.346)	4.76E-02	0.766(0.588,0.997)	3.16E-01	-0.01463	1.59E-06	T

Sinusitis	6797	1.23E-01	1.130(0.967,1.321)	1.75E-01	0.776(0.540,1.114)	6.64E-01	0.952(0.763,1.188)	1.47E-01	1.132(0.957,1.339)	2.95E-02	0.01935	3.54E-24	T
Sjogren's syndrome	939	9.30E-01	0.982(0.654,1.475)	8.31E-01	0.898(0.334,2.409)	8.33E-01	0.943(0.545,1.630)	6.55E-01	0.913(0.613,1.361)	8.45E-01	0.00464	2.98E-02	T
Skin cancer	23320	6.89E-05	0.718(0.610,0.845)	1.02E-03	0.504(0.343,0.741)	3.84E-05	0.727(0.624,0.846)	1.11E-11	0.740(0.678,0.807)	5.38E-02	0.01807	3.11E-58	T
Sleep apnoea	6951	1.44E-01	0.889(0.759,1.041)	9.02E-01	0.976(0.666,1.430)	5.47E-01	0.934(0.747,1.167)	8.94E-02	0.860(0.722,1.024)	5.99E-01	-0.00483	4.60E-24	T
Spondylitis	7250	3.27E-01	0.929(0.801,1.077)	8.88E-01	0.974(0.681,1.395)	1.22E-01	0.848(0.689,1.045)	1.27E-01	0.892(0.770,1.033)	7.74E-01	-0.00247	1.15E-26	T
Stomach cancer	895	6.38E-01	0.910(0.613,1.349)	4.80E-01	1.407(0.549,3.608)	9.20E-01	0.970(0.533,1.765)	9.75E-01	0.993(0.629,1.566)	3.22E-01	-0.02254	3.27E-02	T
Stroke	22833	7.99E-01	1.012(0.926,1.105)	8.19E-01	1.025(0.827,1.271)	9.87E-01	1.001(0.887,1.129)	9.38E-01	1.003(0.920,1.095)	8.92E-01	-0.00070	5.71E-82	T
Thyroid cancer	756	4.09E-06	0.341(0.216,0.539)	6.73E-02	0.342(0.111,1.053)	1.72E-03	0.335(0.169,0.664)	4.65E-06	0.334(0.209,0.534)	9.95E-01	-0.00016	3.95E-01	T
Tinnitus	2274	2.47E-01	1.173(0.896,1.535)	1.10E-01	1.704(0.897,3.236)	5.39E-02	1.458(0.994,2.138)	1.35E-01	1.239(0.935,1.642)	2.15E-01	-0.01925	2.68E-06	T
Tonsilitis	74060	8.95E-01	0.996(0.933,1.062)	6.67E-01	1.035(0.885,1.211)	9.67E-01	0.998(0.927,1.075)	7.49E-01	0.990(0.930,1.054)	5.95E-01	-0.00200	1.29E-187	T
Tuberculosis	2952	3.17E-03	1.380(1.114,1.708)	3.15E-01	1.304(0.781,2.178)	6.69E-03	1.577(1.135,2.193)	5.62E-03	1.399(1.103,1.775)	8.14E-01	0.00291	1.01E-09	T
Umbilical hernia	5492	9.05E-01	0.990(0.844,1.162)	5.49E-01	0.889(0.606,1.304)	5.50E-01	1.079(0.840,1.386)	9.31E-01	0.993(0.839,1.174)	5.46E-01	0.00558	2.40E-20	T
Varicella	6970	9.79E-01	1.002(0.862,1.164)	4.39E-01	1.155(0.804,1.660)	5.24E-01	1.069(0.870,1.314)	8.69E-01	1.014(0.863,1.190)	4.02E-01	-0.00732	8.95E-26	T
Varicose veins	53187	1.98E-01	0.951(0.880,1.027)	5.36E-01	0.942(0.782,1.136)	2.53E-01	0.950(0.870,1.037)	1.25E-01	0.956(0.902,1.013)	9.20E-01	0.00045	5.69E-148	T
Venous thromboembolism	19170	4.13E-02	0.911(0.832,0.996)	7.36E-01	1.038(0.837,1.286)	1.05E-01	0.890(0.773,1.024)	4.16E-02	0.905(0.823,0.996)	1.96E-01	-0.00673	9.69E-69	T
Benign breast lump	4076	3.75E-02	0.804(0.655,0.987)	2.60E-01	0.748(0.454,1.232)	4.89E-02	0.759(0.577,0.999)	4.59E-02	0.808(0.655,0.996)	7.58E-01	0.00370	1.15E-12	T
Breast cancer	17571	6.86E-01	0.979(0.884,1.085)	2.09E-01	0.852(0.666,1.090)	9.41E-01	1.006(0.864,1.171)	9.11E-01	0.994(0.893,1.106)	2.30E-01	0.00714	3.07E-60	T
Breast cyst	7199	2.81E-03	0.773(0.652,0.915)	8.50E-02	0.693(0.461,1.043)	7.52E-02	0.810(0.641,1.022)	1.60E-02	0.809(0.680,0.961)	5.70E-01	0.00559	2.86E-23	T
Cervical cancer	3490	8.56E-02	1.165(0.979,1.385)	3.35E-01	1.232(0.809,1.876)	5.21E-01	1.086(0.844,1.398)	1.16E-01	1.167(0.962,1.416)	7.73E-01	-0.00292	1.69E-17	T
Endometriosis	7753	2.39E-02	0.845(0.730,0.978)	3.27E-02	0.676(0.476,0.959)	9.76E-03	0.749(0.601,0.932)	6.11E-02	0.863(0.739,1.007)	1.75E-01	0.01148	2.25E-27	T
Female infertility	1084	9.75E-02	1.469(0.932,2.314)	6.35E-01	0.766(0.257,2.285)	6.22E-01	1.173(0.622,2.212)	1.52E-01	1.433(0.876,2.342)	2.06E-01	0.03346	8.99E-02	T
Ovarian cyst	14110	6.90E-03	0.827(0.721,0.949)	1.36E-01	0.772(0.553,1.079)	7.14E-02	0.855(0.721,1.014)	1.68E-01	0.908(0.792,1.041)	6.61E-01	0.00352	4.99E-46	T
Ovary cancer	1686	1.79E-01	0.794(0.568,1.111)	1.74E-01	0.566(0.252,1.271)	1.44E-01	0.711(0.450,1.123)	4.81E-01	0.882(0.623,1.250)	3.71E-01	0.01749	7.04E-04	T
Uterine fibroid	20551	8.37E-07	0.632(0.526,0.758)	7.95E-04	0.455(0.296,0.701)	2.11E-04	0.728(0.615,0.861)	3.64E-04	0.744(0.632,0.875)	1.09E-01	0.01691	4.26E-43	T
Uterine polyps	14186	6.87E-05	0.740(0.639,0.858)	9.41E-03	0.613(0.429,0.874)	8.78E-07	0.638(0.534,0.763)	2.76E-04	0.754(0.648,0.878)	2.57E-01	0.00973	1.04E-41	T
Uterine prolapse	15457	6.44E-01	0.978(0.889,1.076)	5.78E-01	1.068(0.849,1.343)	8.98E-01	0.991(0.858,1.144)	8.02E-01	0.987(0.893,1.092)	4.12E-01	-0.00453	1.25E-60	T
Uterus cancer	2286	4.87E-01	0.903(0.677,1.204)	7.02E-01	0.872(0.433,1.754)	3.84E-01	0.856(0.604,1.214)	2.65E-01	0.856(0.652,1.125)	9.14E-01	0.00181	4.04E-06	T
Benign prostatic hyperplasia	19759	3.03E-04	0.724(0.608,0.863)	1.65E-03	0.498(0.330,0.751)	5.05E-03	0.781(0.657,0.928)	1.40E-02	0.812(0.688,0.959)	5.50E-02	0.01923	5.82E-46	T
Prostate cancer	8962	1.28E-02	0.739(0.583,0.938)	1.00E-01	0.610(0.342,1.088)	7.47E-03	0.714(0.558,0.914)	3.75E-03	0.709(0.563,0.895)	4.77E-01	0.00983	6.81E-18	T
Testicular cancer	884	5.96E-02	1.617(0.981,2.665)	7.47E-03	5.172(1.630,16.417)	4.41E-01	1.294(0.672,2.493)	1.21E-01	1.507(0.897,2.529)	3.46E-02	-0.05985	2.11E-01	T

Appendix 6. Time to disease onset results

Table S6. Results of time-to-event analyses between genetically determined telomere length and 127 diseases.

Phenotype	N(Event)	P-value	HR(95%CI)	PH P-value
Abdominal hernia	638	4.38E-01	0.969(0.896,1.049)	7.04E-01
Allergy hypersensitivity	3898	4.55E-01	0.988(0.957,1.020)	3.06E-01
Allergy to drug	4685	4.94E-01	1.010(0.981,1.040)	3.12E-02
Allergy to food	2303	3.41E-01	1.020(0.979,1.063)	4.27E-01
Anorexia nervosa	452	1.24E-01	0.929(0.847,1.020)	1.01E-01
Anxiety	7537	1.45E-01	0.983(0.961,1.006)	1.90E-01
Aortic valve stenosis	1138	2.66E-01	1.034(0.975,1.097)	9.47E-01
Appendicitis	8403	1.54E-01	1.016(0.994,1.038)	5.01E-01
Asthma	53445	8.28E-02	1.008(0.999,1.016)	2.87E-01
Atrial fibrillation	11223	2.44E-01	0.989(0.971,1.008)	5.92E-01
Bladder cancer	2857	2.25E-03	0.944(0.909,0.979)	8.90E-02
Brain cancer	824	1.26E-07	0.830(0.775,0.889)	1.74E-01
Bronchiectasis	2228	9.99E-02	1.036(0.993,1.080)	1.11E-01
Coronary artery disease	28272	1.95E-04	1.023(1.011,1.035)	1.72E-01
Cataract	36819	2.29E-01	1.006(0.996,1.017)	5.18E-01
Chickenpox	4643	4.84E-01	1.010(0.981,1.040)	3.47E-01
Chronic fatigue syndrome	2129	7.89E-02	0.962(0.922,1.004)	9.01E-01
Chronic kidney disease	1729	6.55E-01	0.989(0.943,1.037)	3.09E-01
Coeliac disease	2402	4.54E-21	1.215(1.167,1.265)	8.20E-01
Colorectal cancer	6280	4.95E-01	0.991(0.967,1.016)	7.64E-01
Colorectal polyp	28601	1.59E-06	0.972(0.960,0.983)	3.54E-01
COPD	9866	2.51E-04	1.038(1.017,1.059)	1.29E-01
Dementia	619	9.70E-01	0.998(0.922,1.081)	9.96E-01
Depression	94116	3.00E-01	0.997(0.990,1.003)	3.70E-01
Diabetes I	23237	4.98E-01	1.004(0.992,1.018)	7.09E-04
Diabetes II	24090	4.28E-01	1.005(0.992,1.018)	2.49E-03
Diabetes	24222	4.45E-01	1.005(0.992,1.018)	2.02E-03
Diabetic eye disease	2479	6.55E-01	1.009(0.970,1.050)	9.10E-02
Diverticulitis	22547	8.21E-03	0.982(0.969,0.995)	6.75E-01
Eczema	13178	7.75E-03	1.024(1.006,1.042)	3.55E-02
Epilepsy	4289	2.19E-01	0.981(0.952,1.011)	8.43E-01
Gallstone	15612	1.24E-01	0.988(0.972,1.003)	5.51E-01
Glaucoma	7476	7.71E-01	0.997(0.974,1.020)	7.33E-01
GORD	33436	4.97E-01	0.996(0.986,1.007)	1.15E-01
Gout	6959	5.08E-01	1.008(0.984,1.032)	4.87E-01
Haemorrhoids	3409	5.02E-01	0.988(0.955,1.023)	3.40E-01
Hay fever	45780	1.37E-02	1.012(1.002,1.021)	1.30E-01
Hay fever / eczema	108546	4.33E-04	1.011(1.005,1.017)	9.96E-01
Head and neck cancer	2816	2.70E-01	0.979(0.943,1.016)	2.21E-01
Heart failure	2505	6.20E-01	0.990(0.952,1.030)	5.39E-03
Helicobacter pylori	1359	4.31E-01	0.979(0.927,1.033)	6.93E-01
Hepatitis	2716	8.69E-01	0.997(0.960,1.035)	8.40E-01
Hiatus hernia	19864	6.97E-01	0.997(0.983,1.011)	5.87E-01
Hypertension	4733	2.06E-01	0.982(0.954,1.010)	2.36E-02
Hyperthyroid	3958	1.23E-07	1.089(1.055,1.124)	1.34E-01
Hypothyroid	22155	1.89E-15	1.055(1.042,1.070)	7.19E-01
Immunodeficiency	612	3.81E-01	0.965(0.891,1.045)	8.25E-01
Inflammatory bowel disease	5805	4.37E-02	0.974(0.949,0.999)	1.38E-02

Irritable bowel syndrome	12228	3.53E-01	0.992(0.974,1.009)	2.13E-01
Inguinal hernia	20554	3.29E-02	0.985(0.972,0.999)	8.44E-01
Intervertebral disc disorder	16571	1.46E-01	1.011(0.996,1.027)	1.04E-01
Kidney cancer	1538	4.40E-06	0.888(0.845,0.934)	3.99E-01
Kidney stone	8403	3.70E-02	0.977(0.956,0.999)	2.05E-01
Leukemia	1333	3.10E-06	0.879(0.832,0.928)	5.59E-03
Liver cirrhosis	1945	1.31E-01	1.035(0.990,1.083)	3.14E-01
Lung cancer	2845	3.66E-09	0.894(0.861,0.928)	3.10E-02
Systemic lupus erythematosus	648	7.77E-01	0.989(0.915,1.069)	6.86E-01
Lymphoma	3635	9.02E-07	0.921(0.891,0.952)	7.66E-01
Malaria	795	4.62E-01	0.974(0.908,1.045)	1.67E-01
Mania / bipolar disorder	1600	8.54E-01	0.995(0.947,1.046)	6.45E-02
Measles	3636	3.77E-01	1.015(0.982,1.049)	8.08E-01
Melanoma	5124	3.03E-09	0.920(0.894,0.945)	7.99E-01
Meniere's disease	1369	1.34E-01	0.960(0.910,1.013)	5.94E-01
Meningitis	1899	2.57E-01	0.974(0.931,1.019)	4.13E-01
Migraine	15096	6.05E-01	1.004(0.988,1.021)	5.08E-01
Mononucleosis	946	5.23E-02	1.066(0.999,1.137)	5.18E-01
Multiple sclerosis	1778	8.61E-02	0.960(0.916,1.006)	9.09E-01
Mumps	2248	8.67E-01	0.996(0.956,1.039)	7.71E-01
Nasal polyps	4474	6.72E-01	1.006(0.977,1.037)	1.08E-01
Non-Hodgkin lymphoma	2381	6.84E-03	0.945(0.908,0.985)	3.66E-01
Oesophageal cancer	957	4.76E-01	0.977(0.916,1.042)	8.08E-01
Oesophagitis	3902	3.47E-02	0.966(0.936,0.998)	6.75E-01
Osteoarthritis	62846	9.14E-01	1.000(0.993,1.008)	5.69E-02
Osteoporosis	8631	7.89E-01	1.003(0.982,1.025)	2.52E-01
Otitis media	1772	7.78E-01	0.993(0.948,1.041)	7.20E-01
Pancreas cancer	843	3.14E-01	1.036(0.967,1.109)	3.03E-01
Pancreatitis	2737	5.98E-01	1.010(0.973,1.049)	9.84E-01
Parkinson's disease	984	8.03E-01	1.008(0.946,1.074)	2.14E-01
Peptic ulcer	10036	9.91E-01	1.000(0.980,1.020)	8.60E-02
Peripheral vascular disease	2515	3.16E-01	1.020(0.981,1.062)	2.99E-01
Peritonitis	1952	4.87E-01	1.016(0.971,1.063)	3.18E-02
Pertussis	882	1.86E-01	1.046(0.979,1.118)	4.67E-02
Pleurisy	2008	9.92E-01	1.000(0.957,1.045)	5.43E-01
Pneumonia	19845	1.76E-01	1.010(0.996,1.024)	4.07E-02
Pneumothorax	1191	9.52E-01	1.002(0.946,1.061)	7.34E-01
Polymyalgia rheumatica	1125	2.19E-01	1.038(0.978,1.101)	2.87E-01
Psoriasis	5698	8.60E-01	0.998(0.972,1.024)	8.93E-01
Raynaud's phenomenon	1591	9.16E-01	0.997(0.949,1.048)	3.87E-01
Retinal detachment	4723	3.68E-02	0.970(0.942,0.998)	8.63E-02
Rheumatic fever	1340	4.24E-01	1.022(0.968,1.079)	4.07E-02
Rheumatoid arthritis	5978	2.95E-02	1.029(1.003,1.056)	3.77E-01
Rubella	1214	7.15E-01	0.989(0.935,1.047)	1.51E-01
Sarcoidosis	1285	1.73E-03	1.092(1.034,1.155)	1.72E-01
Scarlatina	835	7.69E-01	0.990(0.924,1.060)	2.04E-01
Schizophrenia	948	1.13E-01	0.949(0.890,1.012)	6.93E-01
Sciatica	6379	7.11E-01	1.005(0.980,1.030)	9.35E-01
Shingles	1168	6.21E-01	0.985(0.930,1.044)	5.67E-01
Sinusitis	5151	7.80E-01	0.996(0.969,1.024)	1.51E-02
Sjogren's syndrome	500	4.68E-01	1.033(0.946,1.129)	7.24E-01
Skin cancer	22402	1.34E-25	0.932(0.919,0.944)	2.92E-01
Sleep apnoea	4171	6.71E-01	0.993(0.963,1.024)	5.73E-02
Spondylitis	5486	1.93E-02	0.969(0.943,0.995)	2.44E-01
Stomach cancer	817	3.08E-01	0.965(0.900,1.034)	1.57E-01
Stroke	20678	9.58E-01	1.000(0.987,1.014)	6.84E-01

Thyroid cancer	708	2.60E-06	0.836(0.776,0.901)	3.81E-01
Tinnitus	1548	1.46E-01	1.038(0.987,1.092)	4.77E-02
Tonsilitis	5315	4.96E-01	1.009(0.982,1.037)	2.48E-02
Tuberculosis	2744	8.71E-03	1.052(1.013,1.092)	7.87E-04
Umbilical hernia	4821	7.69E-01	0.996(0.968,1.025)	9.22E-01
Varicella	5513	5.00E-01	1.009(0.983,1.037)	5.36E-01
Varicose veins	40048	2.24E-01	0.994(0.984,1.004)	1.90E-02
Venous thromboembolism	15938	5.66E-02	0.985(0.970,1.000)	5.57E-01
Benign breast lump	3641	3.26E-03	0.952(0.921,0.984)	2.38E-01
Breast cancer	16503	1.31E-01	0.988(0.973,1.004)	9.83E-01
Breast cyst	3203	2.99E-02	0.962(0.929,0.996)	1.81E-02
Cervical cancer	1916	8.25E-01	1.005(0.961,1.052)	7.74E-01
Endometriosis	5199	5.51E-02	0.973(0.947,1.001)	9.38E-01
Female infertility	1019	4.20E-01	1.026(0.964,1.092)	5.52E-01
Ovarian cyst	7853	9.63E-04	0.963(0.942,0.985)	5.31E-05
Ovary cancer	1549	5.34E-02	0.952(0.905,1.001)	4.01E-01
Uterine fibroid	14453	6.59E-39	0.896(0.881,0.911)	2.79E-01
Uterine polyps	10500	9.73E-11	0.938(0.920,0.956)	8.27E-02
Uterine prolapse	11709	9.48E-01	0.999(0.981,1.018)	8.56E-02
Uterus cancer	2152	2.34E-01	0.974(0.934,1.017)	9.18E-01
Benign prostatic hyperplasia	13146	2.22E-24	0.914(0.899,0.930)	7.04E-01
Prostate cancer	7976	3.39E-08	0.939(0.919,0.961)	1.12E-03
Testicular cancer	845	9.47E-03	1.094(1.022,1.172)	2.03E-01

References

Abegaz, F. *et al.* (2019) 'Principals about principal components in statistical genetics', *Briefings in Bioinformatics*, 20(6), pp. 2200–2216.

Abraham, G. *et al.* (2016) 'Genomic prediction of coronary heart disease', *European Heart Journal*, 37(43), pp. 3267–3278.

Adam, R. *et al.* (2017) 'Prognostic role of telomere length in malignancies: A meta-analysis and meta-regression', *Experimental and Molecular Pathology*, 102(3), pp. 455–474.

Albrecht, E. *et al.* (2014) 'Telomere length in circulating leukocytes is associated with lung function and disease', *European Respiratory Journal*, 43(4), pp. 983–992.

Allende, M. *et al.* (2016) 'Short Leukocyte Telomere Length Is Associated With Cardioembolic Stroke Risk in Patients', *Stroke*, (Mar), pp. 863–865.

Allsopp, R. C. *et al.* (1992) 'Telomere length predicts replicative capacity of human fibroblasts', *Proc.Natl.Acad.Sci.U.S.A*, 89(Nov), pp. 10114–10118.

Anderson, C. a *et al.* (2011) 'Data quality control in genetic case-control association studies', *Nature protocols*, 5(9), pp. 1564–1573.

Anderson, R. *et al.* (2019) 'Length-independent telomere damage drives post-mitotic cardiomyocyte senescence', *The EMBO Journal*, 38(5), pp. 1–21.

Antwi, S. O. *et al.* (2017) 'Genetically Predicted Telomere Length is not Associated with Pancreatic Cancer Risk', *Cancer Epidemiology Biomarkers & Prevention*, 26(6), pp. 971–974.

Apte, M. S. *et al.* (2017) 'Life and cancer without telomerase: ALT and other strategies for making sure ends (don't) meet', *Critical Reviews in Biochemistry and Molecular Biology*, 52(1), pp. 57–73.

Arbeev, K. G. *et al.* (2020) 'Association of Leukocyte Telomere Length With Mortality Among Adult Participants in 3 Longitudinal Studies', *JAMA network open*, 3(2), pp. 1–16.

Armanios, M. *et al.* (2005) 'Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita', *Proceedings of the National Academy of Sciences of the United States of America*, 102(44), pp. 15960–15964.

Armanios, M. (2013) 'Telomeres and age-related disease: How telomere biology informs clinical paradigms', *Journal of Clinical Investigation*, 123(3), pp. 996–1002.

Armanios, M. *et al.* (2012) 'The telomere syndromes', *Nature Reviews Genetics*, 13(10), pp. 693–704.

Armanios, M. Y. *et al.* (2007) 'Telomerase Mutations in Families with Idiopathic Pulmonary Fibrosis', *New England Journal of Medicine*, 356(13), pp. 1317–1326.

Atturu, G. *et al.* (2010) 'Short Leukocyte Telomere Length is Associated with Abdominal Aortic Aneurysm (AAA)', *European Journal of Vascular and Endovascular Surgery*, 39(5), pp. 559–564.

Aubert, G. *et al.* (2012) 'Collapse of Telomere homeostasis in hematopoietic cells caused by heterozygous mutations in Telomerase genes', *PLoS Genetics*, 8(5), pp. 1–11.

Aubert, G. (2014) 'Telomere dynamics and aging', *Progress in Molecular Biology and Translational Science*, 125, pp. 89–111.

Baerlocher, G. M. *et al.* (2006) 'Flow cytometry and FISH to measure the average length of telomeres (flow FISH)', *Nature Protocols*, 1(5), pp. 2365–2376.

Bagley, S. C. *et al.* (2001) 'Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain', *Journal of Clinical Epidemiology*, 54(10), pp. 979–985.

Baird, D. M. *et al.* (2003) 'Extensive allelic variation and ultrashort telomeres in senescent human cells', *Nature Genetics*, 33(2), pp. 203–207.

Balding, D. J. (2006) 'A tutorial on statistical methods for population association studies', *Nature Reviews Genetics*, 7(10), pp. 781–791.

Ballew, B. J., Joseph, V., *et al.* (2013) 'A Recessive Founder Mutation in Regulator of Telomere Elongation Helicase 1, RTEL1, Underlies Severe Immunodeficiency and Features of Hoyeraal Hreidarsson Syndrome', *PLoS Genetics*, 9(8), pp. 1–10.

Ballew, B. J., Yeager, M., *et al.* (2013) 'Germline mutations of regulator of telomere elongation helicase 1, RTEL1, in Dyskeratosis congenita', *Human Genetics*, 132(4), pp. 473–480.

Band, G. *et al.* (2018) 'BGEN: a binary file format for imputed genotype and haplotype data', *bioRxiv*, pp. 1–6.

Bang, K. *et al.* (2001) 'CD4+ CD8+ (thymocyte-like) T lymphocytes present in blood and skin from patients with atopic dermatitis suggest immune dysregulation', *British Journal of Dermatology*, 144(6), pp. 1140–1147.

Barrett, J. H. *et al.* (2015) 'Telomere length and common disease: study design and analytical challenges', *Human Genetics*, 134, pp. 679–689.

Barthel, F. P. *et al.* (2017) 'Systematic analysis of telomere length and somatic alterations in 31 cancer types', *Nature Genetics*, 49(3), pp. 349–357.

Basellini, U. *et al.* (2020) 'An age-at-death distribution approach to forecast cohort mortality', *Insurance: Mathematics and Economics*, 91, pp. 129–143.

de Beer, J. *et al.* (2016) 'A new parametric model to assess delay and compression of mortality', *Population Health Metrics*, 14(1), pp. 1–21.

Benetos, A. *et al.* (2018) 'Short leukocyte telomere length precedes clinical expression of atherosclerosis the blood-And-muscle model', *Circulation Research*, 122(4), pp. 616–623.

Benjamini, Y. *et al.* (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society*, 57(1), pp. 289–300.

Bhupatiraju, C. *et al.* (2012) 'Short Report Association of Shorter Telomere Length with

Essential Hypertension in Indian Population', *American Journal of Human Biology*, 24, pp. 573–578.

Bigdeli, T. B. *et al.* (2016) 'A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans', *Bioinformatics*, 32(17), pp. 2598–2603.

Birch, J. *et al.* (2018) 'Mitochondria, telomeres and cell senescence: Implications for lung ageing and disease', *Pharmacology and Therapeutics*, 183(Mar), pp. 34–49.

Blackburn, E. *et al.* (2017) 'The Telomere Effect: A Revolutionary Approach to Living Younger, Healthier, Longer', in, pp. 10–40.

Blackburn, E. H. (1992) 'Telomerases', *Annu. Rev. Biochem.*, 61, pp. 113–129.

Blackburn, E. H. (2005) 'Telomeres and telomerase: Their mechanisms of action and the effects of altering their functions', *FEBS Letters*, 579(4), pp. 859–862.

Blackburn, E. H. *et al.* (2015) 'Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection', *Science*, 350(6265), pp. 1193–1198.

Blasco, M. A. (2005) 'Telomeres and human disease: Ageing, cancer and beyond', *Nature Reviews Genetics*, 6(8), pp. 611–622.

Blauwkamp, M. N. *et al.* (2017) 'Analytical Validation of Relative Average Telomere Length Measurement in a Clinical Laboratory Environment', *The Journal of Applied Laboratory Medicine: An AACC Publication*, 2(1), pp. 4–16.

Blunder, S. *et al.* (2018) 'Targeted gene expression analyses and immunohistology suggest a pro-proliferative state in tricuspid aortic valve-, and senescence and viral infections in bicuspid aortic valve-associated thoracic aortic aneurysms', *Atherosclerosis*, 271, pp. 111–119.

Boef, A. G. C. *et al.* (2015) 'Mendelian randomization studies: A review of the approaches used and the quality of reporting', *International Journal of Epidemiology*, 44(2), pp. 496–511.

Bowden, J. *et al.* (2016) 'Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator', *Genetic Epidemiology*, 40(4), pp. 304–314.

Bowden, J. *et al.* (2015) 'Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression', *International Journal of Epidemiology*, 44(2), pp. 512–525.

Broer, L. *et al.* (2013) 'Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect', *European Journal of Human Genetics*, 21(10), pp. 1163–1168.

Brouilette, S. *et al.* (2003) 'White cell telomere length and risk of premature myocardial infarction', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 23(5), pp. 842–846.

Burgess, S. (2011) *Statistical issues in Mendelian randomization: use of genetic instrumental variables for assessing causal associations*. University of Cambridge.

Burgess, S., Thompson, D. J., *et al.* (2017) 'Dissecting Causal Pathways Using Mendelian

Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer', *Genetics*, 207(2), pp. 481–487.

Burgess, S., Bowden, J., *et al.* (2017) 'Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants', *Epidemiology*, 28(1), pp. 30–42.

Burgess, S. *et al.* (2019) 'A robust and efficient method for Mendelian randomization with hundreds of genetic variants: unravelling mechanisms linking HDL-cholesterol and coronary heart disease', *bioRxiv*, pp. 1–22.

Burgess, S. *et al.* (2016) 'Beyond Mendelian randomization: How to interpret evidence of shared genetic predictors', *Journal of Clinical Epidemiology*, 69, pp. 208–216.

Burgess, S. *et al.* (2013) 'Mendelian randomization analysis with multiple genetic variants using summarized data', *Genetic Epidemiology*, 37(7), pp. 658–665.

Burgess, S. *et al.* (2018) 'Inferring Causal Relationships Between Risk Factors and Outcomes from Genome-Wide Association Study Data', *Annu. Rev. Genom. Hum. Genet.*, 19, pp. 303–327.

Burgess, S. *et al.* (2015) *Mendelian Randomization: Methods for using Genetic Variants in Causal Estimation*.

Burla, R. *et al.* (2016) 'Mammalian telomeres and their partnership with lamins', *Nucleus*, 7(2), pp. 187–202.

Burnett-Hartman, A. N. *et al.* (2012) 'Telomere-associated polymorphisms correlate with cardiovascular disease mortality in Caucasian women: The Cardiovascular Health Study', *Mech Ageing Dev.*, 133(5), pp. 275–281.

Burris, A. M. *et al.* (2016) 'Hoyeraal-Hreidarsson Syndrome Due to PARN Mutations: Fourteen Years of Follow-up', *Pediatr Neurol.*, 56(Mar), pp. 62–68.

Burton, P. R. *et al.* (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447, pp. 661–678.

Bush, W. S. *et al.* (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*, 8(12), pp. 1–11.

Butt, A. (2017) *Most common age at death, by socio-economic position in England and Wales: a 30 years comparison*.

Caini, S. *et al.* (2015) 'Telomere length and the risk of cutaneous melanoma and non-melanoma skin cancer: A review of the literature and meta-analysis', *Journal of Dermatological Science*, 80(3), pp. 168–174.

Calado, R. T. *et al.* (2009) 'A spectrum of severe familial liver disorders associate with telomerase mutations', *PLoS ONE*, 4(11), pp. 1–9.

Campa, D. *et al.* (2018) 'Genetic determinants of telomere length and risk of pancreatic cancer: A PANDoRA study', *International Journal of Cancer*, 144, pp. 1275–1283.

Canudas-Romo, V. (2010) 'Three measures of longevity: Time trends and record values', *Demography*, 47(2), pp. 299–312.

Cao, K. *et al.* (2011) 'Progerin and telomere dysfunction collaborate to trigger cellular senescence in normal human fibroblasts', *The Journal of Clinical Investigation*, 121(7), pp. 2833–2844.

Cao, X. *et al.* (2019) 'Mendelian randomization study of telomere length and lung cancer risk in East Asian population', *Cancer Medicine*, 8(17), pp. 7469–7476.

Carlquist, J. F. *et al.* (2016) 'Shortened telomere length is associated with paroxysmal atrial fibrillation among cardiovascular patients enrolled in the Intermountain Heart Collaborative Study', *Heart Rhythm*, 13(1), pp. 21–27.

Cawthon, R. M. (2009) 'Telomere length measurement by a novel monochrome multiplex quantitative PCR method', *Nucleic Acids Research*, 37(3), pp. 1–7.

Chang, S. *et al.* (2004) 'Essential role of limiting telomeres in the pathogenesis of Werner syndrome', *Nature Genetics*, 36(8), pp. 877–882.

Chang, S. C. *et al.* (2018) 'Polygenic risk score of shorter telomere length and risk of depression and anxiety in women', *Journal of Psychiatric Research*, 103, pp. 182–188.

Chen, X. *et al.* (2009) 'The null distributions of test statistics in genomewide association studies', *Stat Biosci*, 1(2), pp. 214–227.

Cheng, G. *et al.* (2020) 'Shorter leukocyte telomere length coupled with lower expression of telomerase genes in patients with essential hypertension', *International Journal of Medical Sciences*, 17(14), pp. 2180–2186.

Cheng, Y. *et al.* (2017) 'Genetic association of telomere length with hepatocellular carcinoma risk: A Mendelian randomization analysis', *Cancer Epidemiology*, 50, pp. 39–45.

Chiriacò, M. *et al.* (2019) 'Inflammation and Vascular Ageing: From Telomeres to Novel Emerging Mechanisms', *High Blood Pressure and Cardiovascular Prevention*, 26(4), pp. 321–329.

Chiu, C. L. *et al.* (2016) 'Does telomere shortening precede the onset of hypertension in spontaneously hypertensive mice?', *Twin Research and Human Genetics*, 19(5), pp. 422–429.

Choi, S. W. *et al.* (2018) 'A guide to performing Polygenic Risk Score analyses', *bioRxiv*, 2, pp. 1–22.

Chojnowski, A. *et al.* (2015) 'Progerin reduces LAP2α-telomere association in hutchinson-gilford progeria', *eLife*, 4(Aug), pp. 1–21.

Clarke, G. M. *et al.* (2011) 'Basic statistical analysis in genetic case-control studies', *Nature protocols*, 6(2), pp. 121–133.

Codd, V. *et al.* (2010) 'Common variants near TERC are associated with mean telomere length', *Nature Genetics*, 42(3), pp. 197–199.

Codd, V. *et al.* (2013) 'Identification of seven loci affecting mean telomere length and their association with disease', *Nature Genetics*, 45(4), pp. 422–427.

Codd, V., Denniff, M., *et al.* (2021) 'A major population resource of 474,074 participants

in UK Biobank to investigate determinants and biomedical consequences of leukocyte telomere length', *medRxiv*, pp. 1–26.

Codd, V., Wang, Q., *et al.* (2021) 'Polygenic basis and biomedical consequences of telomere length variation', *medRxiv*, pp. 1–18.

Cole, S. R. *et al.* (2010) 'Survival analysis in infectious disease research: Describing events in time', *Aids*, 24(16), pp. 2423–2431.

Consortium, C. A. D. *et al.* (2013) 'Large-scale association analysis identifies new risk loci for coronary artery disease', *Nature Genetics*, 45(1), pp. 25–33.

Cooper, J. N. *et al.* (2017) 'Telomere Biology and Disease', in *Congenital and Acquired Bone Marrow Failure*, pp. 181–194.

Cordell, H. J. *et al.* (2005) 'Genetic association studies', *The Lancet*, 366, pp. 1121–1131.

Cox, D. R. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society*, 26(2), pp. 211–252.

Crabbe, L. *et al.* (2004) 'Defective telomere lagging strand synthesis in cells lacking WRN helicase activity', *Science*, 306(5703), pp. 1951–1953.

Crabbe, L. *et al.* (2007) 'Telomere dysfunction as a cause of genomic instability in Werner syndrome', *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp. 2205–2210.

D'Mello, M. J. J. *et al.* (2015) 'Association between shortened leukocyte telomere length and cardiometabolic outcomes: Systematic review and meta-analysis', *Circulation: Cardiovascular Genetics*, 8, pp. 82–90.

Davies, N. M. *et al.* (2018) 'Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians', *BMJ*, 362, pp. 1–11.

Dekker, P. *et al.* (2011) 'Relation between maximum replicative capacity and oxidative stress-induced responses in human skin fibroblasts in vitro', *J Gerontol A Biol Sci Med Sci*, 66A(1), pp. 45–50.

Delgado, D. A. *et al.* (2017) 'Genome-wide association study of telomere length among South Asians identifies a second RTEL1 association signal', *Journal of Medical Genetics*, 55(1), pp. 64–71.

Delgado, D. A. *et al.* (2019) 'The contribution of parent-to-offspring transmission of telomeres to the heritability of telomere length in humans', *Hum Genet*, 138(1), pp. 49–60.

Demanelis, K. *et al.* (2020) 'Determinants of telomere length across human tissues', *Science*, 369, pp. 1–12.

Demissie, S. *et al.* (2006) 'Insulin resistance, oxidative stress, hypertension, and leukocyte telomere length in men from the Framingham Heart Study', *Aging Cell*, 5(May), pp. 325–330.

Didelez, V. *et al.* (2007) 'Mendelian randomization as an instrumental variable approach to causal inference', *Statistical Methods in Medical Research*, 16(4), pp. 309–330.

Ding, H. *et al.* (2004) 'Regulation of Murine Telomere Length by Rtel: An Essential Gene Encoding a Helicase-like Protein', *Cell*, 117, pp. 873–886.

Ding, H. *et al.* (2012) 'Telomere length and risk of stroke in Chinese', *Stroke*, 43(3), pp. 658–663.

Ding, Z. *et al.* (2014) 'Estimating telomere length from whole genome sequence data', *Nucleic Acids Research*, 42(9), pp. 7–10.

Donaldson, P. *et al.* (2016) *Genetics of Complex Disease*. Garland Science.

Dorado, B. *et al.* (2017) 'A-type lamins and cardiovascular disease in premature aging syndromes', *Current Opinion in Cell Biology*, 46(2), pp. 17–25.

Dorajoo, R. *et al.* (2019) 'Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies', *Nature Communications*, 10(2491), pp. 1–12.

Dorak, M. T. (2017) *Genetic Association Studies: Background, Conduct, Analysis, Interpretation*. Garland Science.

Du, J. *et al.* (2015) 'Telomere length, genetic variants and gastric cancer risk in a Chinese population', *Carcinogenesis*, 36(9), pp. 963–970.

Du, X. *et al.* (2004) 'Telomere Shortening Exposes Functions for the Mouse Werner and Bloom Syndrome Genes', *Molecular and Cellular Biology*, 24(19), pp. 8437–8446.

Dudbridge, F. (2013) 'Power and Predictive Accuracy of Polygenic Risk Scores', *PLoS Genetics*, 9(3), pp. 1–17.

Dugdale, H. L. *et al.* (2018) 'Heritability of telomere variation: It is all about the environment!', *Phil. Trans. R. Soc. B*, 373, pp. 1–14.

Edwards, D. N. *et al.* (2014) 'Strand exchange of telomeric DNA catalyzed by the Werner syndrome protein (WRN) is specifically stimulated by TRF2', *Nucleic Acids Research*, 42(12), pp. 7748–7761.

Egger, M. *et al.* (1997) 'Bias in meta-analysis detected by a simple, graphical test', *BMJ*, 315(629), pp. 1–11.

Entringer, S. *et al.* (2013) 'Maternal psychosocial stress during pregnancy is associated with newborn leukocyte telomere length', *American Journal of Obstetrics and Gynecology*, 208(134), pp. 1–7.

Entringer, S. *et al.* (2015) 'Maternal estriol concentrations in early gestation predict infant telomere length', *Journal of Clinical Endocrinology and Metabolism*, 100(1), pp. 267–273.

Entringer, S. *et al.* (2018) 'The fetal programming of telomere biology hypothesis: an update', *Phil. Trans. R. Soc. B*, 373, pp. 1–15.

Epel, E. *et al.* (2017) *Background Information and Our Views on Telomere Dynamics and Measurement*. Available at: https://amecenter.ucsf.edu/telomere-testing.

Epel, E. S. *et al.* (2009) 'The rate of leukocyte telomere shortening predicts mortality

from cardiovascular disease in elderly men.', *Aging*, 1(1), pp. 81–88.

Erdel, F. *et al.* (2017) 'Telomere Recognition and Assembly Mechanism of Mammalian Shelterin', *Cell Reports*, 18(1), pp. 41–53.

Ertunc, D. *et al.* (2015) 'Passive smoking is associated with lower age at menopause', *Climacteric*, 18(1), pp. 47–52.

Evangelou, E. (2018) *Genetic Epidemiology*. Springer Science.

Factor-Litvak, P. *et al.* (2016) 'Leukocyte telomere length in newborns: Implications for the role of telomeres in human disease', *Pediatrics*, 137(4), pp. 1–9.

Faye, L. L., Sun, L., *et al.* (2011) 'A flexible genome-wide bootstrap method that accounts for rankingand threshold-selection bias in GWAS interpretation and replication study design', *Statistics in Medicine*, 30(15), pp. 1898–1912.

Faye, L. L. and Bull, S. B. (2011) 'Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: Accuracy of genetic effect estimates', *BMC Proceedings*, 5(564), pp. 1–7.

Fice, H. E. *et al.* (2019) 'Telomere Dynamics Throughout Spermatogenesis', *Genes*, 10(525), pp. 1–12.

Firestein, G. S. (2013) 'Etiology and Pathogenesis of Rheumatoid Arthritis', in *Kelley's Textbook of Rheumatology*. Tenth Edit. Elsevier Inc., pp. 1115–1166.

Fitzpatrick, A. L. *et al.* (2007) 'Leukocyte telomere length and cardiovascular disease in the cardiovascular health study', *American Journal of Epidemiology*, 165(1), pp. 14–21.

Flynn, R. (2012) 'Survival analysis', *Journal of Clinical Nursing*, 21(Oct), pp. 2789–2797.

Frej, F. *et al.* (2015) 'Telomere Biology and Vascular Aging', in *Early Vascular Aging*, pp. 201–211.

De Frutos, C. *et al.* (2016) 'Spermatozoa telomeres determine telomere length in early embryos and offspring', *Reproduction*, 151(1), pp. 1–7.

Fujii, H. *et al.* (2009) 'Telomerase insufficiency in rheumatoid arthritis', *Proceedings of the National Academy of Sciences*, 106(11), pp. 4360–4365.

Gao, D. *et al.* (2018) 'Relative Telomere Length and Stroke Risk in a Chinese Han Population', *Journal of Molecular Neuroscience*, 66(4), pp. 475–481.

Gebreab, S. Y. *et al.* (2017) 'Less than ideal cardiovascular health is associated with shorter leukocyte telomere length: The national health and nutrition examination surveys, 1999-2002', *Journal of the American Heart Association*, 6(2), pp. 1999–2002.

Georgin-Lavialle, S. *et al.* (2010) 'The telomere/telomerase system in autoimmune and systemic immune-mediated diseases', *Autoimmunity Reviews*, 9(10), pp. 646–651.

Giaccherini, M. *et al.* (2020) 'Genetic polymorphisms associated with telomere length and risk of developing myeloproliferative neoplasms', *Blood Cancer Journal*, 10(89), pp. 1–7.

Goeman, J. J. *et al.* (2014) 'Multiple hypothesis testing in genomics', *Statistics in Medicine*, 33(11), pp. 1946–1978.

Gogtay, N. J. *et al.* (2017) 'Key Concepts in Survival', *Journal of The Association of Physicians of India*, 65(May), pp. 80–84.

Goldstein, B. A. *et al.* (2014) 'Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: Coronary heart disease as an example', *Frontiers in Genetics*, 5(Aug), pp. 1–9.

Gordon, L. B. *et al.* (2014) 'Progeria: A paradigm for translational medicine', *Cell*, 156(3), pp. 400–407.

Gorenjak, V. *et al.* (2018) 'The future of telomere length in personalized medicine', *Frontiers in Bioscience*, 23(9), pp. 1628–1654.

Gorenjak, V. *et al.* (2019) 'Telomere length determinants in childhood', *Clinical Chemistry and Laboratory Medicine*, 58(2), pp. 162–177.

Gramatges, M. M. *et al.* (2019) 'Telomere length-associated genetic variants and the risk of thyroid cancer in survivors of childhood cancer: A report from the Childhood Cancer Survivor Study (CCSS)', *Cancer Epidemiology Biomarkers and Prevention*, 28(2), pp. 417–419.

Gray, K. E. *et al.* (2014) 'Leukocyte telomere length and age at menopause', *Epidemiology*, 25(1), pp. 139–146.

Del Greco, F. *et al.* (2015) 'Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome', *Statistics in Medicine*, 34(21), pp. 2926–2940.

Grinde, K. E. *et al.* (2017) 'Illustrating, quantifying, and correcting for bias in post-hoc analysis of gene-based rare variant tests of association', *Frontiers in Genetics*, 8(117), pp. 1–11.

Grover, S. *et al.* (2017) 'Mendelian Randomization', in *Statistical Human Genetics: Methods and Protocols*, pp. 581–628.

Gu, J. *et al.* (2011) 'A Genome-wide Association Study Identifies a Locus on Chromosome 14q21 as a Predictor of Leukocyte Telomere Length and as a Marker of Susceptibility for Bladder Cancer Jian', *Cancer Prev Res*, 4(4), pp. 514–521.

Gu, Z. *et al.* (2014) 'Circlize implements and enhances circular visualization in R', *Bioinformatics*, 30(19), pp. 2811–2812.

Guo, Y. *et al.* (2014) 'Inherited bone marrow failure associated with germline mutation of ACD, the gene encoding telomere protein TPP1', *Blood*, 124(18), pp. 2767–2774.

Hammadah, M. *et al.* (2017) 'Telomere Shortening, Regenerative Capacity, and Cardiovascular Outcomes', *Circ Res.*, 120(7), pp. 1130–1138.

Hanahan, D. *et al.* (2011) 'Hallmarks of cancer: The next generation', *Cell*, 144(4), pp. 646–674.

Hansson, G. K. *et al.* (2011) 'The immune system in atherosclerosis', *Nature Immunology*,

12(3), pp. 204–212.

Hansson, G. K. *et al.* (2006) 'The immune response in atherosclerosis: A double-edged sword', *Nature Reviews Immunology*, 6(7), pp. 508–519.

Harley, C. B. *et al.* (1990) 'Telomeres shorten during ageing of human fibroblasts', *Nature*, 345, pp. 458–460.

van der Harst, P. *et al.* (2007) 'Telomere Length of Circulating Leukocytes Is Decreased in Patients With Chronic Heart Failure', *Journal of the American College of Cardiology*, 49(13), pp. 1459–1464.

van der Harst, P. *et al.* (2010) 'Telomere length and outcome in heart failure', *Annals of Medicine*, 42(1), pp. 36–44.

Hattersley, A. T. *et al.* (2005) 'What makes a good genetic association study?', *Lancet*, 366(9493), pp. 1315–1323.

Haver, V. G. *et al.* (2015) 'Telomere length and outcomes in ischaemic heart failure: Data from the COntrolled ROsuvastatin multiNAtional Trial in Heart Failure (CORONA)', *European Journal of Heart Failure*, 17(3), pp. 313–319.

Haycock, P. C. *et al.* (2014) 'Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis', *BMJ*, 349(Jul), pp. 1–11.

Haycock, P. C. *et al.* (2016) 'Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies', *The American Journal of Clinical Nutrition*, 103, pp. 965–978.

Haycock, P. C. *et al.* (2017) 'Association between telomere length and risk of cancer and non-neoplastic diseases a mendelian randomization study', *JAMA Oncology*, 3(5), pp. 636–651.

Hayflick, L. *et al.* (1961) 'The serial cultivation of human diploid cell strains', *Experimental Cell Research*, 25, pp. 585–621.

Heiss, N. S. *et al.* (1998) 'X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions', *Nature Genetics*, 19(May), pp. 32–38.

Hemani, G., Bowden, J., *et al.* (2017) 'Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome', *bioRxiv*, pp. 1–22.

Hemani, G. *et al.* (2018) 'The MR-Base platform supports systematic causal inference across the human phenome', *eLife*, 7(May), pp. 1–29.

Hemani, G., Tilling, K., *et al.* (2017) 'Orienting the causal relationship between imprecisely measured traits using GWAS summary data', *Plos Genetics*, 13(11), pp. 1–22.

Hingorani, A. *et al.* (2005) 'Nature's randomised trials', *The Lancet*, 366, pp. 1906–1908.

Hiyama, K. (2009) *Telomeres and Telomerase in Cancer*. Edited by K. Hiyama. Humana Press.

Hjelmborg, J. B. *et al.* (2015) 'The heritability of leucocyte telomere length dynamics', *Journal of Medical Genetics*, 52(5), pp. 297–302.

Hohensinner, P. J. *et al.* (2011) 'Telomere dysfunction, autoimmunity and aging', *Aging and Disease*, 2(6), pp. 524–537.

Holohan, B. *et al.* (2015) 'Decreasing initial telomere length in humans intergenerationally understates age-associated telomere shortening', *Aging Cell*, 14, pp. 669–677.

Holohan, B. *et al.* (2016) 'Impaired telomere maintenance in Alazami syndrome patients with LARP7 deficiency', *BMC Genomics*, 17(749), pp. 1–9.

Holohan, B. *et al.* (2014) 'Telomeropathies : An emerging spectrum disorder', 205(3), pp. 289–299.

Horne, B. D. *et al.* (2005) 'Generating Genetic Risk Scores from Intermediate Phenotypes for Use in Association Studies of Clinically Significant Endpoints', *Ann Hum Genet*, 69(Mar), pp. 176–186.

Huang, J. *et al.* (2015) 'Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel', *Nature Communications*, 6, pp. 1–9.

Hunt, S. C. *et al.* (2015) 'Association Between Shortened Leukocyte Telomere Length and Cardiometabolic Outcomes', *Circ Cardiovasc Genet*, 8, pp. 4–7.

Huzen, J. *et al.* (2010) 'The emerging role of telomere biology in cardiovascular disease', *Frontiers in Bioscience*, 15, pp. 35–45.

Ibrahim-Verbaas, C. A. *et al.* (2014) 'Predicting stroke through genetic risk functions: The CHARGE risk score project', *Stroke*, 45(2), pp. 403–412.

Iles, M. M. *et al.* (2014) 'The effect on melanoma risk of genes previously associated with telomere length', *Journal of the National Cancer Institute*, 106(10), pp. 1–5.

Jacobs, E. G. *et al.* (2013) 'Accelerated Cell Aging in Female APOE-ε4 Carriers: Implications for Hormone Therapy Use', *PLoS ONE*, 8(2), pp. 1–7.

Jafri, M. A. *et al.* (2016) 'Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies', *Genome Medicine*, 8(69), pp. 1–18.

Jin, X. *et al.* (2018) 'Relationship between short telomere length and stroke', *Medicine*, 97(39), pp. 1–7.

Johnson, T. (2012) *Efficient Calculation for Multi-SNP Genetic Risk Scores*.

Jose, S. S. *et al.* (2017) 'Chronic inflammation in immune aging: Role of pattern recognition receptor crosstalk with the telomere complex?', *Frontiers in Immunology*, 8(Sep), pp. 1–10.

Jurj, M. A. *et al.* (2020) 'Critical analysis of genome-wide association studies: Triple negative breast cancer quae exempli causa', *International Journal of Molecular Sciences*, 21(16), pp. 1–22.

Kachuri, L. *et al.* (2019) 'Mendelian Randomization and mediation analysis of leukocyte

telomere length and risk of lung and head and neck cancers', *International Journal of Epidemiology*, 48(3), pp. 751–766.

Kaplan, E. L. *et al.* (1958) 'Nonparametric estimation from incomplete samples', *Journal of the american statistical association*, 73(282), pp. 457–481.

Keller, R. B. *et al.* (2012) 'CTC1 Mutations in a patient with dyskeratosis congenita', *Pediatric Blood and Cancer*, 59(2), pp. 311–314.

Kellermayer, R. (2006) 'The versatile RECQL4', *Genetics in Medicine*, 8(4), pp. 213–216.

Khera, A. V. *et al.* (2016) 'Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease', *New England Journal of Medicine*, 375(24), pp. 2349–2358.

Kido, T. *et al.* (2018) 'Are minor alleles more likely to be risk alleles?', *BMC Medical Genomics*, 11(1), pp. 1–11.

Kim, E. S. *et al.* (2015) 'Telomere length and recurrence risk after curative resection in patients with early-stage non-small-cell lung cancer: a prospective cohort study', *J Thorac Oncol*, 10(2), pp. 302–308.

Kimura, M. *et al.* (2008) 'Offspring's leukocyte telomere length, paternal age, and telomere elongation in sperm', *PLoS Genetics*, 4(2), pp. 1–9.

Kimura, M. *et al.* (2010) 'Measurement of telomere length by the southern blot analysis of terminal restriction fragment lengths', *Nature Protocols*, 5(9), pp. 1596–1607.

Kitsios, G. *et al.* (2009) 'Genome-Wide Association Studies: hypothesis-"free" or "engaged"?', *Transl Res.*, 154(4), pp. 161–164.

Kleinbaum, D. G. *et al.* (2005) *Survival Analysis A Self-Learning Text*. 2nd edn. Springer.

Kleinbaum, D. G. *et al.* (2012) *Survival Analysis: A Self-Learning Text*. 3rd edn. Springer.

Kocak, H. *et al.* (2014) 'Hoyeraal-Hreidarsson syndrome caused by a germline mutation in the TEL patch of the telomere protein TPP1', *Genes and Development*, 28(19), pp. 2090–2102.

Kong, C. M. *et al.* (2013) 'Telomere shortening in human diseases', *FEBS Journal*, 280(14), pp. 3180–3193.

Kurz, D. J. *et al.* (2006) 'Degenerative aortic valve stenosis, but not coronary disease, is associated with shorter telomere length in the elderly', *Arterioscler Thromb Vasc Biol.*, 26(6), pp. 114–117.

Lazarides, C. *et al.* (2019) 'Maternal pro-inflammatory state during pregnancy and newborn leukocyte telomere length: A prospective investigation', *Brain, Behavior, and Immunity*, 80(Mar), pp. 419–426.

Lazzerini-Denchi, E. *et al.* (2016) 'Stop pulling my strings-what telomeres taught us about the DNA damage response', *Nature Reviews Molecular Cell Biology*, 17(6), pp. 364–378.

Lee, J. H. *et al.* (2013) 'Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: The long life family study', *Frontiers in Genetics*, 4(Jan), pp. 1–13.

Leite, D. F. B. *et al.* (2019) 'Approaching literature review for academic purposes: The literature review checklist', *Clinics*, 74, pp. 1–8.

Levy, D. *et al.* (2010) 'Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology', *Proceedings of the National Academy of Sciences*, 107(20), pp. 9293–9298.

Lewis, C. M. (2002) 'Genetic association studies: design, analysis and interpretation', *Brief Bioinform*, 3(2), pp. 146–153.

Lewis, C. M. *et al.* (2020) 'Polygenic risk scores: From research tools to clinical instruments', *Genome Medicine*, 12(1), pp. 1–11.

Li, C. *et al.* (2020) 'Genome-Wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length', *The American Journal of Human Genetics*, 106(Mar), pp. 1–16.

Li, J. *et al.* (2018) 'The association of telomere attrition with first-onset stroke in Southern Chinese: A case-control study and meta-analysis', *Scientific Reports*, 8(1), pp. 1–11.

Lidzbarsky, G. *et al.* (2018) 'Genomic Instabilities, Cellular Senescence, and Aging: In Vitro, In Vivo and Aging-Like Human Syndromes', *Frontiers in medicine*, 5(Apr), pp. 1–16.

Liu, Y. *et al.* (2014) 'A Genome-Wide Association Study Identifies a Locus on TERT for Mean Telomere Length in Han Chinese', *PLoS ONE*, 9(1), pp. 1–6.

Loh, P. R. *et al.* (2015) 'Efficient Bayesian mixed-model analysis increases association power in large cohorts', *Nature Genetics*, 47(3), pp. 284–290.

Lu, S. *et al.* (2018) 'Genetic analysis of the relation of telomere length-related gene (RTEL1) and coronary heart disease risk', *Molecular Genetics and Genomic Medicine*, 7(e550), pp. 1–7.

Lung, F. W. *et al.* (2008) 'Telomere length may be associated with hypertension', *Journal of Human Hypertension*, 22, pp. 230–232.

Luo, D. *et al.* (2017) 'Telomere length associated with the risks of high-risk and ischemic stroke in southern Chinese Han population', *Oncotarget*, 8(62), pp. 105915–105922.

Luu, H. N. *et al.* (2016) 'Association between genetic risk score for telomere length and risk of breast cancer', *Cancer Causes and Control*, 27(10), pp. 1219–1228.

Ma, H. *et al.* (2011) 'Shortened Telomere length is associated with increased risk of cancer: A meta-analysis', *PLoS ONE*, 6(6), pp. 1–9.

Machiela, M. J. *et al.* (2015) 'Genetic variants associated with longer telomere length are associated with increased lung cancer risk among never-smoking women in Asia: A report from the female lung cancer consortium in Asia', *International Journal of Cancer*, 137(2), pp. 311–319.

Machiela, M. J. *et al.* (2016) 'Genetically predicted longer telomere length is associated with increased risk of B-cell lymphoma subtypes', *Human Molecular Genetics*, 25(8), pp. 1663–1676.

Machiela, M. J. *et al.* (2017) 'Genetic Variants Related to Longer Telomere Length are Associated with Increased Risk of Renal Cell Carcinoma', *European Urology*, 72, pp. 747–754.

Maciejowski, J. *et al.* (2017) 'Telomeres in cancer: Tumour suppression and genome instability', *Nature Reviews Molecular Cell Biology*, 18(3), pp. 175–186.

Madrid, A. S. *et al.* (2016) 'Short telomere length and ischemic heart disease: Observational and genetic studies in 290 022 individuals', *Clinical Chemistry*, 62(8), pp. 1140–1149.

Maeda, T. *et al.* (2009) 'Aging-related alterations of subtelomeric methylation in sarcoidosis patients', *J Gerontol A Biol Sci Med Sci*, 64(7), pp. 752–760.

Van Maldergem, L. *et al.* (2006) 'Revisiting the craniosynostosis-radial ray hypoplasia association: Baller-Gerold syndrome caused by mutations in the RECQL4 gene', *Journal of Medical Genetics*, 43(2), pp. 148–152.

Mangino, M. *et al.* (2009) 'Short report A genome-wide association study identifies a novel locus on chromosome 18q12.2 influencing white cell telomere length', *J Med Genet*, 46, pp. 451–454.

Mangino, M. *et al.* (2012) 'Genome-wide meta-analysis points to CTC1 and ZNf676 as genes regulating telomere homeostasis in humans', *Human Molecular Genetics*, 21(24), pp. 5385–5394.

Mangino, M. *et al.* (2015) 'DCAF4, a novel gene associated with leucocyte telomere length', *Journal of Medical Genetics*, 52(3), pp. 157–162.

Manichaikul, A. *et al.* (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26(22), pp. 2867–2873.

Marchini, J. *et al.* (2007) 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nature Genetics*, 39(7), pp. 906–913.

Marchini, J. (2015) *UK Biobank Phasing and Imputation Documentation*.

Marchini, J. *et al.* (2010) 'Genotype imputation for genome-wide association studies', *Nature Reviews Genetics*, 11(7), pp. 499–511.

Martens, D. S. *et al.* (2016) 'Maternal pre-pregnancy body mass index and newborn telomere length', *BMC Medicine*, 14(1), pp. 1–10.

Martínez, P. *et al.* (2017) 'Telomere-driven diseases and telomere-targeting therapies', *The Journal of Cell Biology*, 216(4), pp. 875–887.

Mathur, M. B. *et al.* (2016) 'Perceived stress and telomere length: A systematic review, meta-analysis, and methodologic considerations for advancing the field', *Brain, Behavior, and Immunity*, 54, pp. 158–169.

Maubaret, C. G. *et al.* (2010) 'Telomeres are shorter in myocardial infarction patients compared to healthy subjects: Correlation with environmental risk factors', *Journal of Molecular Medicine*, 88(8), pp. 785–794.

Mbatchou, J. *et al.* (2020) 'Computationally efficient whole genome regression for

quantitative and binary traits', *bioRxiv*, pp. 1–88.

McClintock, B. (1941) 'The Stability of Broken Ends of Chromosomes in Zea Mays', *Genetics*, 26(2), pp. 234–282.

McCord, R. P. *et al.* (2013) 'Correlated alterations in genome organization, histone methylation, and DNA–lamin A/C interactions in Hutchinson-Gilford progeria syndrome', *Genome Research*, 23, pp. 260–269.

McNally, E. J. *et al.* (2019) 'Long telomeres and cancer risk: the price of cellular immortality', *J Clin Invest*, pp. 1–8.

Mendel, G. (1865) *Versuche über Plflanzenhybriden*.

Metcalfe, J. A. *et al.* (1996) 'Accelerated telomere shortening in ataxia telangiectasia', *Nature Genetics*, 13(Jul), pp. 350–353.

De Meyer, T. *et al.* (2007) 'Paternal age at birth is an important determinant of offspring telomere length', *Human Molecular Genetics*, 16(24), pp. 3097–3102.

Minamino, T. *et al.* (2002) 'Endothelial cell senescence in human atherosclerosis: Role of telomere in endothelial dysfunction', *Circulation*, 105(13), pp. 1541–1544.

Mirabello, L. *et al.* (2010) 'The association of telomere length and genetic variation in telomere biology genes', *Human Mutation*, 31(9), pp. 1050–1058.

Moncayo, R. *et al.* (2015) 'The WOMED model of benign thyroid disease : Acquired magnesium de fi ciency due to physical and psychological stressors relates to dysfunction of oxidative phosphorylation', *BBA Clinical*, 3, pp. 44–64.

Moore, D. F. (2016) *Applied Survival Analysis Using R*. Springer.

Van Moorsel, C. H. M. (2018) 'Trade-offs in aging lung diseases: A review on shared but opposite genetic risk variants in idiopathic pulmonary fibrosis, lung cancer and chronic obstructive pulmonary disease', *Current Opinion in Pulmonary Medicine*, 24(3), pp. 309–317.

Moreno-Navarrete, J. M. *et al.* (2018) 'Adipose TSHB in humans and serum TSH in hypothyroid rats inform about cellular senescence', *Cellular Physiology and Biochemistry*, 51(1), pp. 142–153.

Morrison, J. *et al.* (2020) 'Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics', *Nature Genetics*, 52(7), pp. 740–747.

Muka, T. *et al.* (2016) 'Association of age at onset of menopause and time since onset of menopause with cardiovascular outcomes, intermediate vascular traits, and all-cause mortality: A systematic review and meta-analysis', *JAMA Cardiology*, 1(7), pp. 767–776.

Muller, H. J. (1938) 'The remaking of chromosomes', *Collect. Net*, 13, pp. 1181–1198.

Nalobin, D. *et al.* (2020) 'Telomeres and Telomerase in Heart Ontogenesis, Aging and Regeneration', *Cells*, 9(503), pp. 1–12.

Nandakumar, J. *et al.* (2013) 'Finding the end: Recruitment of telomerase to telomeres',

Nature Reviews Molecular Cell Biology, 14(2), pp. 69–82.

Neumann, A. A. *et al.* (2013) 'Alternative lengthening of telomeres in normal mammalian somatic cells', *Genes and Development*, 27, pp. 18–23.

Nguyen, M. T. *et al.* (2019) 'Telomere length: population epidemiology and concordance in Australian children aged 11 – 12 years and their parents', *BMJ Open*, 9, pp. 118–126.

Njajou, O. T. *et al.* (2007) 'Telomere length is paternally inherited and is associated with parental lifespan', *PNAS*, 104(29), pp. 12135–12139.

O'Donovan, A. *et al.* (2011) 'Cumulative inflammatory load is associated with short leukocyte telomere length in the health, aging and body composition study', *PLoS ONE*, 6(5), pp. 1–7.

Ojha, J. *et al.* (2016) 'Genetic variation associated with longer telomere length increases risk of chronic lymphocytic leukemia', *Cancer Epidemiology Biomarkers and Prevention*, 25(7), pp. 1043–1049.

Olovnikov, A. M. (1973) 'A theory of marginotomy: The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon', *J. Theor. Biol.*, 41, pp. 181–190.

Olovnikov, A. M. (1996) 'Telomeres, telomerase, and aging: Origin of the theory', *Experimental Gerontology*, 31(4), pp. 443–448.

Opresko, P. L. *et al.* (2004) 'The werner syndrome helicase and exonuclease cooperate to resolve telomeric D loops in a manner regulated by TRF1 and TRF2', *Molecular Cell*, 14(Jun), pp. 763–774.

Osorio-Yáñez, C. *et al.* (2020) 'Early life tobacco exposure and children's telomere length: The HELIX project', *Science of the Total Environment*, 711(Apr), pp. 1–11.

Østhus, I. B. Ø. *et al.* (2017) 'Association of Telomere Length With Myocardial Infarction: A Prospective Cohort From the Population Based HUNT 2 Study', *Progress in Cardiovascular Diseases*, 59(6), pp. 649–655.

Ozturk, S. (2015) 'Telomerase Activity and Telomere Length in Male Germ Cells', *Biology of Reproduction*, 92(Jan), pp. 1–11.

Palm, W. *et al.* (2008) 'How Shelterin Protects Mammalian Telomeres', *Annual Review of Genetics*, 42(1), pp. 301–334.

Palmer, C. *et al.* (2017) 'Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies', *PLoS Genetics*, 13(7), pp. 1–14.

Paneni, F. *et al.* (2017) 'The Aging Cardiovascular System: Understanding It at the Cellular and Clinical Levels', *Journal of the American College of Cardiology*, 69(15), pp. 1952–1967.

Pasaniuc, B. *et al.* (2016) 'Dissecting the genetics of complex traits using summary association statistics', *Nature Reviews Genetics*, pp. 1–12.

Patel, C. J. *et al.* (2017) 'Systematic correlation of environmental exposure and

physiological and self-reported behaviour factors with leukocyte telomere length', *International Journal of Epidemiology*, 46(1), pp. 44–56.

Pooley, K. A. *et al.* (2013) 'A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk', *Human Molecular Genetics*, 22(24), pp. 5056–5064.

Prasad, K. N. *et al.* (2017) 'Telomere shortening during aging: Attenuation by antioxidants and anti-inflammatory agents', *Mechanisms of Ageing and Development*, 164(Apr), pp. 61–66.

Prescott, J. *et al.* (2011) 'Genome-wide association study of relative telomere length', *PLoS ONE*, 6(5), pp. 1–9.

Purcell, S. *et al.* (2007) 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *The American Journal of Human Genetics*, 81(3), pp. 559–575.

Purcell, S. M. *et al.* (2009) 'Common polygenic variation contributes to risk of schizophrenia and bipolar disorder', *Nature*, 460(7256), pp. 748–752.

Pusceddu, I. *et al.* (2018) 'Telomere length and mortality in the ludwigshafen risk and cardiovascular health study', *PLoS ONE*, 13(6), pp. 1–13.

Qiu, W. (2020) *Package 'powerMediation'*.

R Core Team (2013) 'R: A language and environment for statistical computing'.

Ravlić, S. *et al.* (2018) 'Mechanisms of fetal epigenetics that determine telomere dynamics and health span in adulthood', *Mechanisms of Ageing and Development*, 174(Aug), pp. 55–62.

Reed, E. *et al.* (2015) 'A guide to genome-wide association analysis and post-analytic interrogation', *Statistics in Medicine*, 34(28), pp. 3769–3792.

Rietzschel, E. R. *et al.* (2016) 'Telomeres and Atherosclerosis: The Attrition of an Attractive Hypothesis', *Journal of the American College of Cardiology*, 67(21), pp. 2477–2479.

Rivera, T. *et al.* (2017) 'A balance between elongation and trimming regulates telomere stability in stem cells', *Nature Structural and Molecular Biology*, 24(1), pp. 30–39.

Roberts, J. D. *et al.* (2014) 'Telomere Length and the Risk of Atrial Fibrillation: Insights into the Role of Biological versus Chronological Aging', *Circ Arrhythm Electrophysiol*, 7(6), pp. 1026–1032.

Robinson, N. J. *et al.* (2016) 'Means to the ends: The role of telomeres and telomere processing machinery in metastasis', *Biochimica et Biophysica Acta*, pp. 320–329.

van Rossum, G. (1995) *Python Tutorial*, *Technical Report CS-R9526*.

Rubin, J. B. *et al.* (2020) 'Sex differences in cancer mechanisms', *Biology of Sex Differences*, 11(1), pp. 1–29.

Ruzankina, Y. *et al.* (2007) 'Relationships between stem cell exhaustion, tumour suppression and ageing', *British Journal of Cancer*, 97(9), pp. 1189–1193.

Samani, N. J. *et al.* (2008) 'Biological ageing and cardiovascular disease', *Heart*, 94(5), pp. 537–539.

Sanders, J. L. *et al.* (2013) 'Telomere Length in Epidemiology: A Biomarker of Aging, Age-Related Disease, Both, or Neither?', *Epidemiologic Reviews*, 35, pp. 112–131.

Sarek, G., Marzec, P., *et al.* (2015) 'Molecular basis of telomere dysfunction in human genetic diseases', *Nature Structural and Molecular Biology*, 22(11), pp. 867–874.

Sarek, G., Vannier, J. B., *et al.* (2015) 'TRF2 recruits RTEL1 to telomeres in s phase to promote t-loop unwinding', *Molecular Cell*, 57(4), pp. 622–635.

Sasa, G. S. *et al.* (2012) 'Three novel truncating TINF2 mutations causing severe dyskeratosis congenita in early childhood', *Clinical Genetics*, 81(5), pp. 470–478.

Savage, S. A. *et al.* (2008) 'TINF2, a Component of the Shelterin Telomere Protection Complex, Is Mutated in Dyskeratosis Congenita', *American Journal of Human Genetics*, 82(2), pp. 501–509.

Savage, S. A. (2018) 'Beginning at the ends: telomeres and human disease', *F1000Research*, 7(May), pp. 1–16.

Saxena, R. *et al.* (2014) 'A Genome-Wide Association Study Identifies Variants in Casein Kinase II (CSNK2A2) to be Associated with Leukocyte Telomere Length in a Punjabi Sikh Diabetic Cohort', *Circ Cardiovasc Genet.*, 7(3), pp. 287–295.

Schober, P. *et al.* (2018) 'Survival analysis and interpretation of time-to-event data: The tortoise and the hare', *Anesthesia and Analgesia*, 127(3), pp. 792–798.

Schoenfeld, D. (1980) 'Chi-squared goodness-of-fit tests for the proportional hazards regression model', *Biometrika*, 67, pp. 145–153.

Schoenfeld, D. (1982) 'Partial residuals for the proportional hazards regression model', *Biometrika*, 69, pp. 239–241.

Schönland, S. O. *et al.* (2003) 'Premature telomeric loss in rheumatoid arthritis is genetically determined and involves both myeloid and lymphoid cell lineages.', *PNAS*, 100(23), pp. 13471–13476.

Schunkert, H. *et al.* (2011) 'Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease', *Nature Genetics*, 43(4), pp. 333–340.

Schürks, M. *et al.* (2013) 'Telomere length and ischaemic stroke in women: A nested case-control study', *European Journal of Neurology*, 20(7), pp. 1068–1074.

Sebastiani, P. *et al.* (2012) 'Naive Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: Not so different after all!', *Frontiers in Genetics*, 3(Feb), pp. 1–9.

Seki, A. *et al.* (2015) 'Age-related Cardiovascular Changes and Diseases', *Cardiovascular Pathology: Fourth Edition*, pp. 57–83.

Send, T. S. *et al.* (2017) 'Telomere length in newborns is related to maternal stress during pregnancy', *Neuropsychopharmacology*, 42(12), pp. 2407–2413.

Shadyab, A. H. *et al.* (2018) 'Parental longevity predicts healthy ageing among women', *Age and Ageing*, 47(6), pp. 853–860.

Shah, N. C. *et al.* (2014) 'Short-term magnesium deficiency downregulates telomerase, upregulates neutral sphingomyelinase and induces oxidative DNA damage in cardiovascular tissues: Relevance to atherogenesis, cardiovascular diseases and aging', *International Journal of Clinical and Experimental Medicine*, 7(3), pp. 497–514.

Shay, J. W. (2016) 'Role of telomeres and telomerase in aging and cancer', *Cancer Discovery*, 6(6), pp. 584–593.

Sheehan, M. T. *et al.* (2016) 'Transient hypothyroidism after radioiodine for graves' disease: Challenges in interpreting thyroid function tests', *Clinical Medicine and Research*, 14(1), pp. 40–45.

Shen, G. *et al.* (2020) 'The Relationship between Telomere Length and Cancer Mortality: Data from the 1999–2002 National Healthy and Nutrition Examination Survey (NHANES)', *Journal of Nutrition, Health and Aging*, 24(1), pp. 9–15.

Shenassaa, E. D. *et al.* (2015) 'Telomere length and age-at-menopause in the US', *Maturitas*, 82(2), pp. 215–221.

Shi, J. *et al.* (2013) 'Leukocyte telomere length-related genetic variants in 1p34.2 and 14q21 loci contribute to the risk of esophageal squamous cell carcinoma', *International Journal of Cancer*, 132(12), pp. 2799–2807.

Siland, J. E. *et al.* (2017) 'Telomere length and incident atrial fibrillation – data of the PREVEND cohort', *PLoS ONE*, 12(2), pp. 1–12.

Slob, E. A. W. *et al.* (2019) 'A Comparison Of Robust Mendelian Randomization Methods Using Summary Data', *bioRxiv*, pp. 1–31.

Smilenov, L. B. *et al.* (1997) 'Influence of ATM function on telomere metabolism', *Oncogene*, 15(22), pp. 2659–2665.

Smith, G. D. *et al.* (2003) '"Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease?', *International Journal of Epidemiology*, 32(1), pp. 1–22.

Smith, G. D. *et al.* (2014) 'Mendelian randomization: Genetic anchors for causal inference in epidemiological studies', *Human Molecular Genetics*, 23(Jun), pp. 1–10.

Smith, J. A. *et al.* (2015) 'Current Applications of Genetic Risk Scores to Cardiovascular Outcomes and Subclinical Phenotypes', *Current Epidemiology Reports*, 2(3), pp. 180–190.

Song, L. *et al.* (2019) 'Effects of maternal exposure to ambient air pollution on newborn telomere length', *Environment International*, 128(May), pp. 254–260.

Spiliopoulou, A. *et al.* (2015) 'Genomic prediction of complex human traits: Relatedness, trait architecture and predictive meta-models', *Human Molecular Genetics*, 24(14), pp. 4167–4182.

Staerk, L. *et al.* (2017) 'Association Between Leukocyte Telomere Length and the Risk of

Incident Atrial Fibrillation: The Framingham Heart Study', *J Am Heart Assoc*, 6, pp. 1–9.

Stanley, S. E. *et al.* (2016) 'Telomerase and the genetics of emphysema susceptibility: Implications for pathogenesis paradigms and patient care', *Annals of the American Thoracic Society*, 13(8), pp. S447–S451.

van Steensel, B. *et al.* (2017) 'Lamina-associated domains: links with chromosome architecture, heterochromatin and gene repression', *Cell*, 169(5), pp. 780–791.

Stefler, D. *et al.* (2018) 'Leukocyte telomere length and risk of coronary heart disease and stroke mortality: prospective evidence from a Russian cohort', *Scientific Reports*, 8(1), pp. 1–6.

Steiger, J. H. (1980) 'Tests for Comparing Elements of a Correlation Matrix', *Psychological Bulletin*, 87(2), pp. 245–251.

Stevenson, M. (2007) 'An Introduction to Survival Analysis', pp. 1–31.

Stindl, R. (2016) 'The paradox of longer sperm telomeres in older men's testes: A birth-cohort effect caused by transgenerational telomere erosion in the female germline', *Molecular Cytogenetics*, 9(1), pp. 10–13.

Stone, R. C. *et al.* (2016) 'Telomere Length and the Cancer-Atherosclerosis Trade-Off', *PLoS Genetics*, 12(7), pp. 1–10.

Stuart, B. D. *et al.* (2015) 'Exome Sequencing Links Mutations in PARN and RTEL1 with Familial Pulmonary Fibrosis and Telomere Shortening', *Nat Genet.*, 47(5), pp. 512–517.

Sudlow, C. *et al.* (2015) 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLoS Medicine*, 12(3), pp. 1–10.

Sun, Y. *et al.* (2020) 'Leukocyte telomere length: A potential biomarker for the prognosis of coronary artery disease', *Biomarkers in Medicine*, 14(11), pp. 933–941.

Supplement, S. *et al.* (2010) 'Supplements: Common variants near TERC are associated with mean telomere length', *Nature genetics*, 42(3), pp. 197–9.

Taub, M. A. *et al.* (2019) 'Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed', *bioRxiv*, (Sep), pp. 1–31.

Taub, M. A. *et al.* (2020) 'Novel genetic determinants of telomere length from a trans-ethnic analysis of 109, 122 whole genome sequences in TOPMed', *bioRxiv*, pp. 1–42.

Teasley, D. C. *et al.* (2015) *Telomere Biology*, *Encyclopedia of Cell Biology*.

Tellechea, M. *et al.* (2017) 'The impact of hypertension on leukocyte telomere length: a systematic review and meta-analysis of human studies', *Journal of Human Hypertension*, 31, pp. 99–105.

Thriveni, K. *et al.* (2019) 'Patterns of Relative Telomere Length is Associated With hTERT Gene Expression in the Tissue of Patients With Breast Cancer', *Clinical Breast Cancer*, 19(1), pp. 27–34.

Tobin, M. D. *et al.* (2008) 'Common variants in genes underlying monogenic hypertension and hypotension and blood pressure in the general population', *Hypertension*, 51(6), pp. 1658–1664.

Touzot, F. *et al.* (2010) 'Function of Apollo (SNM1B) at telomere highlighted by a splice variant identified in a patient with Hoyeraal-Hreidarsson syndrome', *PNAS*, 107(22), pp. 10097–10102.

Tracy, R. P. (2003) 'Emerging relationships of inflammation, cardiovascular disease and chronic diseases of aging', *International Journal of Obesity*, 27, pp. 29–34.

Tsakiri, K. D. *et al.* (2007) 'Adult-onset pulmonary fibrosis caused by mutations in telomerase', *Proceedings of the National Academy of Sciences of the United States of America*, 104(18), pp. 7552–7557.

Tsangaris, E. *et al.* (2008) 'Ataxia and pancytopenia caused by a mutation in TINF2', *Human Genetics*, 124(5), pp. 507–513.

Tsoukalas, D. *et al.* (2019) 'Discovery of potent telomerase activators: Unfolding new therapeutic and anti-aging perspectives', *Molecular Medicine Reports*, 20(4), pp. 3701–3708.

Turner, J. K. *et al.* (2019) 'Telomere Biology and Human Phenotype', *Cells*, 8(73), pp. 1–19.

UK Biobank (2015) *Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers (Interim Data Release, 2015)*.

UK Biobank (2020) *UK Biobank showcase of Data-Field 3894*. Available at: https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=3894.

Vannier, J. B. *et al.* (2012) 'RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity', *Cell*, 149(4), pp. 795–806.

Vasa-Nicotera, M. *et al.* (2005) 'Mapping of a major locus that determines telomere length in humans', *American Journal of Human Genetics*, 76(1), pp. 147–151.

Vecoli, C. *et al.* (2019) 'Independent and combined effects of telomere shortening and mtDNA4977 deletion on long-term outcomes of patients with coronary artery disease', *International Journal of Molecular Sciences*, 20(5508), pp. 1–11.

Vestbo, J. *et al.* (2013) 'Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary', *American Journal of Respiratory and Critical Care Medicine*, 187(4), pp. 347–365.

Victorelli, S. *et al.* (2017) 'Telomeres and Cell Senescence - Size Matters Not', *EBioMedicine*, 21, pp. 14–20.

Vilhjálmsson, B. J. *et al.* (2015) 'Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores', *American Journal of Human Genetics*, 97(4), pp. 576–592.

Voors, A. A. *et al.* (2016) 'A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure: rationale, design, and baseline characteristics of BIOSTAT-CHF', *European*

Journal of Heart Failure, 18(6), pp. 716–726.

Vulliamy, T. *et al.* (2001) 'The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita', *Nature*, 413(6854), pp. 432–435.

Vulliamy, T. *et al.* (2008) 'Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita', *PNAS*, 105(23), pp. 8073–8078.

Wain, L. V. *et al.* (2015) 'Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank', *The Lancet Respiratory Medicine*, 3(10), pp. 769–781.

Walne, A. J. *et al.* (2007) 'Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase- associated protein NOP10', *Hum Mol Genet.*, 16(13), pp. 1619–1629.

Walne, A. J. *et al.* (2013) 'Constitutional mutations in RTEL1 cause severe dyskeratosis congenita', *American Journal of Human Genetics*, 92(3), pp. 448–453.

Walsh, K. M. *et al.* (2014) 'Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk', *Nat Genet.*, 46(7), pp. 731–735.

Walsh, K. M. *et al.* (2016) 'Common genetic variants associated with telomere length confer risk for neuroblastoma and other childhood cancers', *Carcinogenesis*, 37(6), pp. 576–582.

Wang, X. bin *et al.* (2019) 'Leukocyte telomere length, mitochondrial DNA copy number, and coronary artery disease risk and severity: A two-stage case-control study of 3064 Chinese subjects', *Atherosclerosis*, 284(Nov), pp. 165–172.

Ware, E. B. *et al.* (2017) 'Heterogeneity in polygenic scores for common human traits', *bioRxiv*, (5), pp. 1–13.

Wassertheil-Smoller, S. *et al.* (2015) *Biostatistics and Epidemiology*.

Watson, J. D. (1972) 'Origin of Concatemeric T7 DNA', *Nature New Biology*, 239, pp. 197–201.

Weischer, M. *et al.* (2012) 'Short telomere length, myocardial infarction, ischemic heart disease, and early death', *Arteriosclerosis, Thrombosis, and Vascular Biology*, 32(3), pp. 822–829.

Wentzensen, I. M. *et al.* (2011) 'The association of telomere length and cancer: a meta-analysis.', *Cancer Epidemiol Biomarkers Prev.*, 20(6), pp. 1238–1250.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wong, K. K. *et al.* (2003) 'Telomere dysfunction and Atm deficiency compromises organ homeostasis and accelerates ageing', *Nature*, 421(6923), pp. 643–648.

Wood, L. D. *et al.* (2001) 'Characterization of ataxia telangiectasia fibroblasts with extended life-span through telomerase expression', *Oncogene*, 20(3), pp. 278–288.

Wu, K. *et al.* (2000) 'Telomerase Activity Is Increased and Telomere Length Shortened in

T Cells from Blood of Patients with Atopic Dermatitis and Psoriasis', *The Journal of Immunology*, 165(8), pp. 4742–4747.

Wysoczanska, B. *et al.* (2019) 'Variability within the human TERT gene, telomere length and predisposition to chronic lymphocytic leukemia', *OncoTargets and Therapy*, 12, pp. 4309–4320.

Xiao, J. *et al.* (2019) 'The telomere length of peripheral blood cells is associated with the risk of ischemic stroke in Han population of northern China', *Medicine*, 98(7), pp. 1–6.

Xu, C. *et al.* (2020) 'Association between leucocyte telomere length and cardiovascular disease in a large general population in the United States', *Scientific Reports*, 10(1), pp. 1–10.

Xu, X. *et al.* (2019) 'Differences in Leukocyte Telomere Length between Coronary Heart Disease and Normal Population: A Multipopulation Meta-Analysis', *BioMed Research International*, pp. 1–9.

Yang, J. *et al.* (2011) 'GCTA: A tool for genome-wide complex trait analysis', *American Journal of Human Genetics*, 88(1), pp. 76–82.

Yang, J. *et al.* (2012) 'Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits', *Nature Genetics*, 44(4), pp. 369–375.

Yang, Z. *et al.* (2009) 'Short telomeres and prognosis of hypertension in a Chinese population', *Hypertension*, 53(4), pp. 639–645.

Yeh, J. K. *et al.* (2019) 'Telomeres as Therapeutic Targets in Heart Disease', *JACC: Basic to Translational Science*, 4(7), pp. 855–865.

Yeh, J. K. *et al.* (2016) 'Telomeres and Telomerase in Cardiovascular Diseases', *Genes*, 7(58), pp. 1–18.

Young, A. J. *et al.* (2018) 'The role of telomeres in the mechanisms and evolution of life-history trade-offs and ageing', *Phil. Trans. R. Soc. B*, 373.

Zeiger, A. M. *et al.* (2018) 'Genetic Determinants of Telomere Length in African American Youth', *Nature Scientific Reports*, 8(May), pp. 1–9.

Zeng, Z. *et al.* (2020) 'Association of telomere length with risk of rheumatoid arthritis: A meta-analysis and Mendelian randomization', *Rheumatology*, 59(5), pp. 940–947.

Zgheib, N. K. *et al.* (2018) 'Short telomere length is associated with aging, central obesity, poor sleep and hypertension in lebanese individuals', *Aging and Disease*, 9(1), pp. 77–89.

Von Zglinicki, T. (2002) 'Oxidative stress shortens telomeres', *Trends in Biochemical Sciences*, 27(7), pp. 339–344.

Zhan, Y. *et al.* (2017) 'Exploring the Causal Pathway from Telomere Length to Coronary Heart Disease: A Network Mendelian Randomization Study', *Circulation Research*, 121(3), pp. 214–219.

Zhang, C. *et al.* (2015) 'The association between telomere length and cancer prognosis:

Evidence from a meta-analysis', *PLoS ONE*, 10(7), pp. 1–17.

Zhang, J. *et al.* (2016) 'Ageing and the telomere connection: An intimate relationship with inflammation', *Ageing Research Reviews*, 25, pp. 55–69.

Zhang, N. *et al.* (2018) 'Leucocyte telomere length and paroxysmal atrial fibrillation: A prospective cohort study and systematic review with meta-analysis', *Journal of Clinical Laboratory Analysis*, 32(9), pp. 1–8.

Zhang, W. *et al.* (2013) 'Short telomere length in blood leucocytes contributes to the presence of atherothrombotic stroke and haemorrhagic stroke and risk of post-stroke death', *Clinical Science*, 125(1), pp. 27–36.

Zhang, X. *et al.* (2017) 'The association of telomere length in peripheral blood cells with cancer risk:ASystematic review and meta-Analysis of prospective studies', *Cancer Epidemiology Biomarkers and Prevention*, 26(9), pp. 1381–1390.

Zhao, Q. *et al.* (2018) 'Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score', *arXiv*, pp. 1–55.

Zhdanova, N. S. *et al.* (2016) 'Telomere Recombination in Normal Mammalian Cells', *Russian Journal of Genetics*, 52(1), pp. 14–23.

Zheng, J. *et al.* (2017) 'Recent Developments in Mendelian Randomization Studies', *Current Epidemiology Reports*, 4(4), pp. 330–345.

Zhong, F. *et al.* (2011) 'Disruption of telomerase trafficking by TCAB1 mutation causes dyskeratosis congenita', *Genes and Development*, 25(1), pp. 11–16.

Zhou, W. *et al.* (2018) 'Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies', *Nature Genetics*, 50(9), pp. 1335–1341.

Zhu, D. *et al.* (2019) 'Age at natural menopause and risk of incident cardiovascular disease: a pooled analysis of individual patient data', *The Lancet Public Health*, 4(11), pp. e553–e564.

Zhu, H. *et al.* (2013) 'Telomere Biology in Senescence and Aging: Focus on Cardiovascular Traits', in *Inflammation, Advancing Age and Nutrition*, pp. 71–84.

Zhu, X. *et al.* (2016) 'The association between telomere length and cancer risk in population studies', *Scientific Reports*, 6, pp. 1–10.

Zhu, Y. *et al.* (2019) 'Telomere and its role in the aging pathways: telomere shortening, cell senescence and mitochondria dysfunction', *Biogerontology*, 20(1), pp. 1–16.

Zimmermann, M. *et al.* (2014) 'TRF1 negotiates TTAGGG repeatassociated replication problems by recruiting the BLM helicase and the TPP1/POT1 repressor of ATR signaling', *Genes and Development*, 28(22), pp. 2477–2491.

Zöllner, S. *et al.* (2007) 'Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data', *The American Journal of Human Genetics*, 80(4), pp. 605–615.