Multi-omics studies of the molecular architecture of age, ageing and age-related disorders

Nuffield Department of Population Health

Sihao Xiao

BSc (Hons) MSc

Brasenose College, University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2025

Abstract

Background:

Ageing is a complex and heterogeneous process influenced by genetic, environmental, and lifestyle factors, with notable differences between sexes. High-throughput omics technologies, such as proteomics and metabolomics, provide new opportunities to quantify and understand the molecular underpinnings of ageing and the largest ageing contributor-smoking.

Methods:

This thesis applied machine learning approaches to large-scale omics datasets, including the UK Biobank and China Kadoorie Biobank. First, two sex-specific ageing clocks were developed: (1) sex-specific metabolomic ageing clocks and (2) sex-specific proteomic ageing clocks. Metabolic age gap (metAgeGap) and proteomic age gap (protAgeGap) were then derived as the residual between predicted age and chronological age. Next, the plasma proteome was used to study smoking, the largest contributor to ageing. Proteomic Smoking Index (pSIN) was developed to quantify pathologies induced by smoking and to capture its long-term health impacts. Their associations with incident health outcomes were further studied.

Results:

In females, the metabolomic clock explained 37% variance of age ($R^2 = 0.37$), whereas in males, the metabolomic clock achieved an R^2 of 0.29. A higher metAgeGap was correlated with incident cardiovascular diseases, neurodegenerative diseases and cancers in both sexes, with larger effects in males. Fertility-related factors, including later puberty, were correlated with a lower metAgeGap in both sexes. Genome-wide association study (GWAS) identified 147 loci in females, compared to 217 loci identified in males. Genes identified by the GWAS were enriched

in lipid metabolism (e.g., PPARG), transporters (e.g., SLC2A2), and inflammatory pathways. Sex-stratified proteomic clocks explained higher variance of age compared to the metabolic clock, reaching an R^2 of 0.88 in both males and females. In females, earlier menopause was correlated with higher protAgeGap, while in both sexes, later puberty was correlated with a lower protAgeGap. A higher protAgeGap was also associated with 14/15 non-cancer diseases tested, including neurodegenerative diseases, incident ischemic heart disease, and chronic kidney disease. In addition, protAgeGap was found to be associated with sex-specific cancer risks, including breast cancer in females and prostate cancer in males. Proteomics data can also be used to study smoking, one of the largest risk factors for ageing. The proteomic smoking model discriminated current versus never smokers with an area under the curve of 0.95 in the UK Biobank and 0.91 in CKB. Higher pSIN was strongly associated with incident lung cancer, chronic obstructive pulmonary disease (COPD), stroke and all-cause mortality independent of self-reported smoking history. pSIN can also be used to monitor the recovery of previous smokers. “Recovered” previous smokers defined by pSIN exhibited markedly lower risks of COPD, lung cancer, and mortality compared to “non-recovered” previous smokers. GWAS of pSIN identified 95 lead variants mapping to 129 genes (e.g., ALPP, CST5, IL12B) and revealed strong genetic correlations with body mass index, diabetes and smoking-related diseases. Exposome-wide analysis linked lower pSIN to favourable socioeconomic status, healthy diet, and higher levels of physical activity, whereas air pollution, high red meat intake, maternal smoking, and mood disorders were associated with higher pSIN.

Conclusion:

This thesis demonstrates the power of gradient boosting models in uncovering sex-specific and

multi-omics profiles of biological ageing and smoking exposure, revealing distinct proteomic and metabolomic signatures that differentially predict disease risk across males and females. By comparing proteomic and metabolic clocks, I highlight the superior robustness of proteomic age in capturing a broader spectrum of age-related outcomes, while metabolic age uniquely reflects adiposity-linked pathways. The development of pSIN further underscores how targeted protein panels can quantify the long-term effects of lifestyle risk factors that contribute to ageing and stratify chronic disease risk beyond traditional self-report measures.

Acknowledgements

I would like to extend my sincere gratitude to the participants, supporting staff, and the steering committee of the UK Biobank and China Kadoorie Biobank for their invaluable contributions to the creation of the UKB and CKB resources. These resources will undoubtedly benefit population health for generations to come. I also wish to acknowledge my funders, the Centre for Artificial Intelligence in Precision Medicine at Oxford, whose support made this research possible.

My deepest thanks go to my supervisor, Professor Cornelia van Duijn, for her generous scholarship and unwavering guidance throughout this research journey. I am particularly grateful for the autonomy she afforded me in designing the project and for placing her trust in me to undertake a completely novel initiative within the group. I would also like to express my appreciation to my other supervisors, Professor Najaf Amin and Professor Alejo Nevado-Holgado, for their continuous support and guidance during challenging moments. I also want to thank Zhengming Cheng for his guidance and suggestions during my DPhil.

I am also deeply grateful to my closest friends, Bowen Liu, Gaoyu Du, and Qishi Zhang, for their unwavering belief in me, and their help and inspiration during the development of new methods and concepts. Our long discussions, both in academic settings and social gatherings, were invaluable.

Finally, I would like to express my profound gratitude to my parents, Yu Xiao and Lan Bi, for their steadfast support throughout my DPhil journey, both financially and emotionally. Their encouragement has been a constant source of strength.

Statement of contribution to work

I was responsible for conceiving and developing the research questions addressed in this thesis. I designed and conducted all statistical analyses, and machine learning models and produced all the tables and figures presented apart from those cited from other studies in Chapters 1 and 2. The results of Chapter 2 were published in my third year of DPhil in Nature Medicine led by a postdoc of our group, Austin Argentieri, where I designed and coded the pipeline for the model development part. Prof Najaf Amin ran the Genome-Wide Association Study in Chapters 3 and 5, and I ran all the relevant follow-up analyses and generated the figures. I interpreted the findings and wrote up all aspects of the study, including the thesis and related publications.

Contents

List of figures

List of abbreviations

AGEs	Advanced Glycation End-Products
ALP	Alkaline Phosphatase
ALT	Alanine Transaminase
AMPK	AMP-Activated Protein Kinase
AP	Averaged Precision
APOB	Apolipoprotein B
APOC1	Apolipoprotein C1
APOE	Apolipoprotein E
AST	Aspartate Transaminase
AUC	Area Under the Curve
BA	Balanced Accuracy
BMI	Body Mass Index
BMP	Bone Morphogenetic Protein
bOHbutyrate	Beta-Hydroxybutyrate
Boruta	Boruta Feature Selection Algorithm
CGA	Chorionic Gonadotropin Alpha Subunit
CI	Confidence Interval
CKB	China Kadoorie Biobank
COPD	Chronic Obstructive Pulmonary Disease
Cox PH Model	Cox Proportional Hazards Model
CTCF	CCCTC-binding Factor
CV	Cross-Validation
CVD	Cardiovascular Disease
DALY	Disability-Adjusted Life Years
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EDA2R	Ectodysplasin A2 Receptor
EDTA	Ethylenediaminetetraacetic Acid
ELN	Elastin
F1	F1 Score (harmonic mean of precision and recall)
FDR	False Discovery Rate
FOXO3	Forkhead Box O3
FPR	False Positive Rate

FSHB	Follicle-Stimulating Hormone Beta Subunit
GDF11	Growth Differentiation Factor 11
GFAP	Glial Fibrillary Acidic Protein
GGT	Gamma-Glutamyl Transferase
GO	Gene Ontology
GRM	Genetic Relationship Matrix
GTEx	Genotype-Tissue Expression project
GWAS	Genome-Wide Association Study
HDL	High-Density Lipoprotein
HR	Hazard Ratio
HRT	Hormone Replacement Therapy
HSF1	Heat Shock Factor 1
IGF-1	Insulin-Like Growth Factor 1
IIS	Insulin/IGF-1 Signaling
IPAQ	International Physical Activity Questionnaire
KEGG	Kyoto Encyclopedia of Genes and Genomes
L1	L1 Regularization
L2	L2 Regularization
LASSO	Least Absolute Shrinkage and Selection Operator
LDL	Low-Density Lipoprotein
LDSC	Linkage Disequilibrium Score Regression
LightGBM	Light Gradient Boosting Machine
metAgeGap	Metabolic Age Gap
MLP	Multilayer Perceptron
mTOR	Mechanistic Target of Rapamycin
NAD^+	Nicotinamide Adenine Dinucleotide
NEFL	Neurofilament Light Chain
NIH	National Institutes of Health
NMR	Nuclear Magnetic Resonance
NPX	Normalized Protein eXpression
pQTL	Protein Quantitative Trait Locus
protAgeGap	Proteomic Age Gap
pSIN	Proteomic Smoking INdex
PUFAs	Polyunsaturated Fatty Acids
R^2	Coefficient of Determination
Reactome	Reactome Pathway Database

ResNet	Residual Neural Network
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SAIGE	Scalable and Accurate Implementation of Generalized mixed model
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SIRT1	Sirtuin 1
SMAD	SMAD Family (Sma and Mad proteins)
SNP	Single Nucleotide Polymorphism
TabR	Retrieval-augmented Neural Network for Tabular Data
TERT	Telomerase Reverse Transcriptase
TGF-β	Transforming Growth Factor-Beta
TOMM40	Translocase of Outer Mitochondrial Membrane 40
TPE	Tree-structured Parzen Estimator
UK	United Kingdom
UKB	United Kingdom Biobank
VLDL	Very Low-Density Lipoprotein

Chapter 1 Introduction

The concept of ageing

Ageing is characterised by the progressive decline of cellular, tissue, and systemic physiological functions, increasing the risk of diseases and leading to multi-morbidity and mortality[1]. Understanding the molecular changes underlying ageing is a medical and societal urgency, in light of the growing ageing population worldwide and its impact on healthcare systems that are globally facing their limits[2].

Although ageing is a universal phenomenon, it exhibits remarkable heterogeneity within and across species. In the vast majority of species, ageing and life span are strongly coupled with reproduction. Reproducing until the end of a lifespan is expected to result in greater evolutionary success, with more genetic variations being passed on to future generations[3]. In most mammals, the end of fertility heralds the end of life[4]. Humans are one of the few exceptions to the rule, as women spend on average 42.5% of their adult life post-reproductive, i.e., after menopause[5]. In our nearest relatives, i.e. chimpanzees, some females may have a prolonged period of post-reproductive life, but findings have been far from consistent, with some populations of female chimpanzees spending only 2% of their adult life post-reproductive[6]. In the natural world, the only other mammals with a well-established long post-reproductive life in females are toothed whales[7]. This phenomenon has puzzled evolution researchers and led to the grandmother hypothesis of aging[8]. The grandmother hypothesis suggests that, for humans, the cessation of direct reproduction (menopause) and a shift toward

investing resources in grandchildren can be a more effective way to propagate one's genes than continuing to bear children late in life. By ensuring that grandchildren survive to reproductive age, a grandmother improves her inclusive fitness—her genetic contribution to future generations—even after her fertility has ended. Below is a deeper exploration of how investing in offspring of offspring enhances evolutionary success under this framework.

Although both men and women are ageing, there is a clear difference between men and women due to their different roles in reproduction. While hormonal regulation of ovulation in women changes during a relatively short period (menopause), resulting in infertility, in men, testosterone levels decline throughout many years[9,10]. Differences between men and women in life span and ageing are well described[11]. The life expectancy in the UK of men in the period 2020 to 2022 was 78.6 years and of women is 82.6 years[11]. As the gap between life expectancy declined over time, the difference is largely attributed to risk behaviours and lifestyles[12]. However, from a mechanistic perspective, there is growing interest in the difference between men's and women's ageing.

In this chapter, I first review the general mechanism underlying ageing that has been summarised as the hallmarks of ageing. Next, I discuss the use biological clock to quantify biological aging. Further, I review the role of the genome and exposome, all exposures over time as drivers of the ageing process and the challenges to overcome in ageing research.

'Hallmarks of ageing' framework offers a systematic and comprehensive approach for deciphering the complex biology of ageing and identifying potential interventions to enhance health span and lifespan[13]. Since its introduction in 2013, the hallmarks framework has provided

a paradigm for ageing research, initially encompassing nine hallmarks: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication[14]. Each hallmark meets three criteria: it manifests depending on time, accelerates ageing when exacerbated, and is amenable to therapeutic intervention[1]. Recently, this framework expanded to include three additional hallmarks—disabled macroautophagy, chronic inflammation, and dysbiosis—highlighting advancements in understanding the complex interconnections between molecular, cellular, and systemic ageing processes[13] (***Fig 1***). These hallmarks form a dynamic network that collectively drives the ageing phenotype and provides a foundation for translational research aimed at promoting healthy ageing. Collectively, these hallmarks provide a comprehensive framework of interacting mechanisms that allow disentangling ageing as an integrative and dynamic process.

Genomic instability is a cornerstone of ageing biology, arising from cumulative DNA damage caused by intrinsic factors like replication errors and oxidative stress, as well as extrinsic factors such as radiation and toxins[15]. This DNA damage compromises cellular function, leading to mutations, chromosomal rearrangements, and telomere dysfunction, which impair tissue homeostasis and promote ageing. Interventions targeting PARP1[16], SIRT6[17], and OGG1[18] that enhance DNA repair mechanisms or mitigate DNA damage have been promising in extending lifespan in experimental models[15].

Telomere attrition reflects the gradual shortening of telomeres—the protective caps at chromosome ends—due to the end-replication problem[19]. Critically short telomeres trigger cellular senescence or apoptosis and in humans contribute to aging and age-related diseases

across organs such as cardiovascular disorders, metabolic diseases and cancer[20,21]. Experimental strategies to elongate telomeres, such as telomerase activation, have demonstrated potential in delaying ageing and restoring tissue function[19].

Epigenetic alterations encompass a diverse set of reversible regulatory changes—such as DNA methylation shifts, histone post-translational modifications, chromatin remodelling, and non-coding RNA activity—that accumulate with age and ultimately disrupt gene expression and cellular homeostasis. Collectively, these changes contribute to age-related pathologies (e.g., cancer, neurodegeneration, metabolic disorders, bone loss) by reshaping chromatin landscapes and affecting transcriptional programs[22]. Among these mechanisms, DNA methylation represents one of the most extensively characterized modifications. During ageing, the genome undergoes both global hypomethylation and locus-specific hypermethylation, often at tumour suppressor or Polycomb target sites[23]. Although most age-associated methylation changes occur in intronic or intergenic regions of uncertain function, they can be harnessed to build "epigenetic clocks" that predict biological age and disease risk[24].

Deregulated nutrient sensing, another hallmark of ageing, involves pathways such as insulin/IGF-1 signalling, mTOR, AMPK, and sirtuins, which regulate metabolism and cellular homeostasis [25]. Persistent activation of these pathways in adulthood drives ageing by prioritising growth over repair. Interventions such as caloric restriction and pharmacological modulation of these pathways have demonstrated significant lifespan and health span benefits in preclinical studies[25].

Altered intercellular communication disrupts signalling networks between cells, leading to chronic inflammation, impaired immune responses, and hormonal imbalances[26]. This dysregulation amplifies systemic ageing processes and underscores the importance of maintaining effective intercellular communication for healthy aging[27]. Ageing also causes numerous damages to the long-lived protein components of the extracellular matrix (ECM), including advanced glycation end-products (AGEs), carbonylation, carbamylation, elastin fragmentation, and collagen crosslinking[28]. These alterations contribute to tissue fibrosis, a phenomenon referred to as fibroaging. The excessive release of transforming growth factor-beta (TGF-β) and other growth factors, coupled with the nuclear translocation of transcription factors, further drives the expression of pro-fibrotic genes. On the other hand, ECM stiffness increasing with age also modulates the function of senescent cells by inducing matrix metalloprotease secretion[28]. This exacerbates ECM damage and promotes the activation of pro-inflammatory and pro-fibrotic signalling pathways, such as WNT, NOTCH, RAS, and TGF-β/SMAD[28].

Chronic inflammation, or "**inflammaging**" arises from immune system alterations, cellular senescence, and microbial dysbiosis[29,30] . It contributes to the pathogenesis of numerous age-related diseases, including cardiovascular disease, diabetes and neurodegeneration[31]. Multiple ageing hallmarks feed into inflammaging such as accumulation of cytosolic nuclear or mitochondrial DNA which activates DNA sensors and inflammatory cascades, epigenetic dysregulation which impairs proteostasis and reduces autophagy also upregulates pro-inflammatory genes[13]. Anti-inflammatory interventions, ranging from pharmacological agents to lifestyle modifications, are critical for improving health span[32].

Loss of proteostasis, the maintenance of protein homeostasis, deteriorates with age, leading to the accumulation of damaged or misfolded proteins[32]. Loss of proteostasis is intricately connected to other hallmarks of ageing spanning mitochondrial dysfunction, genomic instability, cellular senescence cellular communication, disabled macroautophagy and inflammation, forming a network of interdependent processes that collectively drive the ageing phenotype. For example, mitochondrial dysfunction increases oxidative stress, which exacerbates protein damage and overwhelms proteostasis systems[32]. Genomic instability introduces aberrant or misfolded proteins, further burdening the already declining proteostasis machinery[33]. These changes contribute significantly to the onset and progression of age-related diseases, including neurodegenerative disorders such as Alzheimer's, Parkinson's, and Huntington's[32], cardiovascular health[34], metabolic disorders[35], and immune senescence[36], highlighting its systemic influence on aging[37].

To date, the proteome is still the first target for developing new medication[38]. Despite these challenges, therapeutic strategies aimed at enhancing proteostasis—such as activating molecular chaperones, stimulating proteasome activity, or boosting autophagy—show great promise[13]. Dietary interventions like caloric restriction and the use of mimetics such as spermidine and rapamycin have demonstrated improved autophagic function and proteostasis in animal models, translating into greater cellular resilience and extended lifespan[39]. Advances in proteomics and systems biology are paving the way for precision medicine approaches to restore proteostasis, offering new hope for addressing ageing and age-related diseases[25].

Capturing Ageing with Biological Clocks

The advent of big data era in population-based studies have allowed the study of the molecular mechanisms of aging process through omics-based aging clocks. An ageing clock is a computational model that estimates an individual's biological age based on molecular, physical or clinical biomarkers, providing a more precise measure of ageing than chronological age. It reflects the functional state of an organism and the risk of age-related diseases based on ongoing disease-related processes[40]. Ageing clocks can be built from various biomarkers, ranging from visible traits such as greying hair to molecular indicators like leukocyte telomere length, epigenetic modifications, proteomic signatures, and metabolic shifts (***Fig2***)[40]. With the rise of high-throughput genomics, proteomics, and metabolomics, ageing clocks leverage large-scale biological data to map molecular changes that accumulate with age and determine the risk you developing an age-related disease. Although initially linear regression models were used to build such models, given the complexity of ageing biology, machine learning algorithms such as random forest and neural networks are increasingly used to distil omics data into composite ageing biomarkers. The predicted age serves as an indicator of an individual's biological age, with the difference between predicted and chronological age, known as Δage or age gap**,** representing differences in their past ageing rate[22]. This hypothesis is supported by findings showing that individuals with a positive age gap, referred to as age acceleration, have a larger association with the risk of mortality and age-related conditions such as heart disease, metabolic syndrome, and certain cancers[24,41]. It has been argued that these clocks can be seen as propensity scores[42]. This is due to the pitfalls of clock models where ΔAge estimates come from errors intrinsic to the model instead of from a true difference between biological and

chronological age. For instance, DNA methylation clocks built via penalized regression on cytosine-guanine (CpG) sites achieve high accuracy ($r \geq 0.8$), yet the residual (epigenetic age acceleration) only partially captures biological ageing, since age acceleration also includes intra-array measurement noise and shifts in blood cell composition[22] and provides limited information to risks of age-related diseases beyond chronological age. To address this, GrimAge clocks are trained on surrogate markers—such as plasma proteins and smoking pack-years—improving prediction of morbidity and mortality compared to chronological-age-trained clocks; nevertheless, they still rely on imperfect proxies, and ΔAge from GrimAge may reflect fluctuations in those surrogate biomarkers rather than fundamental ageing mechanisms[41]. Nonlinear models, such as convolutional neural networks (CNNs) trained on facial images, capture complex, non-linear features linked to perceived age and health status, but they risk overfitting to imaging artefacts and lack interpretability regarding which features drive biological age predictions[43].

Despite the shortcomings discussed above, ageing clocks is expected to have significant applications in health risk prediction, personalised medicine, and longevity research, aiding the identification of individuals at risk for age-related diseases, guiding interventions such as lifestyle modifications and pharmacological treatments, and evaluating the efficacy of anti-ageing therapies and rejuvenation strategies[44]. Ageing clocks provide a powerful tool for quantifying ageing at the molecular level, bridging the gap between fundamental ageing biology and clinical applications.

Ageing clocks can be built from different omics data, representing different layers of information. Epigenomics, for example, has been particularly successful in studying the

epigenetic changes related to ageing, which, in contrast to inherited genetic mutations, are reversible and do not change the sequence of DNA bases. These changes may or may not change the expression of the genome and its subsequent translation in proteins and metabolites. Aberrant DNA methylation patterns have been linked to the development of cancers, where hypermethylation of tumour suppressor genes silences their expression, contributing to tumorigenesis. In neurodegenerative diseases like Parkinson's, altered histone modifications and epigenetic dysregulation exacerbate disease progression[45].

Transcriptomics offers a dynamic view of gene expression changes underlying disease states. In cardiovascular diseases, transcriptomic studies have revealed the upregulation of inflammatory and fibrotic pathways, reflecting the cellular response to stress and damage[46]. Similarly, in cancer, differential expression of oncogenes and tumour suppressor genes provides valuable insights into tumour progression and treatment resistance[47]. However, transcriptomic signatures are often heterogeneous, complicating their direct association with specific diseases without tissue-specific analyses[48].

Proteomics has revealed disease-associated alterations in protein expression, modifications, and interactions. For example, in neurodegenerative diseases such as Alzheimer's, proteomic studies highlight the accumulation of misfolded proteins like amyloid-beta and tau, which form toxic aggregates[49]. Proteomic analyses of cardiovascular diseases uncover dysregulated structural proteins, inflammation and enzymes involved in oxidative stress[50]. In the next chapter, I will describe the establishment of a robust framework I designed to capture biological ageing using plasma proteomic data and its ability to generalize across populations which has been published in Nature Medicine[51].

Metabolomics captures changes in metabolic pathways linked to diseases. For instance, in diabetes and metabolic syndrome, alterations in glucose, lipid, and amino acid metabolism are well-documented, reflecting systemic metabolic reprogramming[52]. Metabolomics also uncovers biomarkers for cancer, such as altered levels of specific oncometabolites like 2-hydroxyglutarate in gliomas[53]. However, the sensitivity of metabolomics to diet, lifestyle, and environmental factors poses challenges in distinguishing disease-specific signals from background noise[54].

In summary, each omics approach offers unique insights into the molecular mechanisms underlying age-related diseases. While genomics and epigenomics highlight genetic and regulatory factors, transcriptomics, proteomics, and metabolomics provide dynamic and functional data. Together, these approaches enable a more comprehensive understanding of how ageing contributes to disease risk and progression, highlighting potential biomarkers and therapeutic targets. Integration of these omics layers through systems biology approaches is essential to fully unravel their complex associations with disease-specific processes. These may involve genetic and lifestyle determinants and biological responses (resilience).

Genetic Determinants of Ageing

From an evolutionary perspective, the ageing process is unlikely to be determined by the genome as those genes would escape natural selection[4,55]. However, there is increasing evidence suggesting that the ageing process, like other processes, is shaped by the interplay of inherited genetic variants, environmental, and lifestyle factors[56]. The *APOE* gene is one of the first linked to human longevity in a 1994 study of centenarians. *APOE* encodes three primary isoforms: ε2, ε3, and ε4, which arise from polymorphisms in two SNPs: rs7412 and rs429358[57].

The ε4 allele is associated with higher fertility and cognitive benefits in younger individuals in infectious environments, and increases the risk of cardiovascular diseases, Alzheimer's disease, and other age-related pathologies in modern post-industrial settings, highlighting the principle of antagonistic pleiotropy[58]. *APOE* ε4 carriers have higher low-density lipoprotein cholesterol and reduced apolipoprotein E levels compared to wide type, predisposing them to atherosclerosis, an association that was described before that of neurodegenerative disease, kidney function, and etc. [59]. The *APOE* gene also operates within a cluster including *TOMM40* and *APOC1*, where genetic variants influence promoter activity, impacting phenotypic expression and disease risk[49].

FOXO3, another major longevity gene, is a transcription factor integral to the insulin/IGF-1 signalling (IIS) pathway, regulating cellular metabolism, autophagy, and stress responses[60]. It enhances antioxidative defences and DNA repair, mitigating oxidative damage—a hallmark of aging[61]. *FOXO3* SNPs, such as rs2802292, have shown strong associations with longevity, particularly in male and Asian populations, as confirmed by multiple meta-analyses[60,62]. *FOXO3* interacts with transcription factors such as CTCF and heat shock factor 1 (HSF1), enhancing its role in stress resistance and lifespan extension[61,63]. Furthermore, *FOXO3* alleles associated with longevity are linked to reduced telomere attrition and lower cardiovascular mortality, underscoring its protective effects in aging[61,64].

In addition to *APOE* and *FOXO3*, other genetic factors play a role in ageing. *SIRT1*, for instance, is a gene encoding for a protein that regulates mitochondrial function and stress responses by deacetylating key proteins[65] and interacts with mTOR which regulates cellular growth and autophagy, with its inhibition extending lifespan in model organisms[66]. Further, genes such as

KLOTHO modulate calcium and phosphate metabolism, contributing to vascular and renal homeostasis[67], and TERT maintains telomere length, preventing cellular senescence[68]. Together, these genetic pathways form a complex protein network that governs cellular homeostasis, stress resistance, and ageing.

Lifestyles Determinants of Ageing

The near doubling of human lifespan over the past 200 years, despite genomic stability that can be captured by Hardy-Weinberg equilibrium[11], underscores the dominant role of environmental and behavioural factors[69]. Among these, lifestyle factors such as smoking, diet, physical activity, and stress management significantly shape health span and longevity[70]. Smoking, in particular, is a leading modifiable risk factor for ageing and premature death, with advanced machine learning models identifying behavioural and clinical parameters—including smoking status, socioeconomic factors, dietary habits, and physical activity—as critical predictors of life expectancy[70,71]. Additionally, study has also shown that smoking is the number one modifiable contributor to proteomic age acceleration[70].

Smoking accelerates the ageing process by inducing oxidative stress, systemic inflammation, and immune system impairments. It damages cellular components such as lipids, proteins, and DNA, thereby increasing the risk of chronic diseases including cardiovascular disease, cancer, and chronic obstructive pulmonary disease (COPD)[72]. A recent study ranked smoking as one of the most significant mortality risk factors among older populations, underscoring its profound impact on health outcomes[73]. Similarly, a large study in the UK Biobank aimed to rank the contribution of genes and environment on premature mortality and biological ageing also

found that the major determinants were 1) smoking 2) socioeconomic status, and 3) physical activity[70]. Moreover, smoking interacts with genetic predispositions, amplifying the effects of deleterious alleles and exacerbating age-related decline[74].

While recognised for a long as the major driver of ageing-related pathology, there is still lack of powerful tools to capture the pathological effects of smoking on ageing. Traditional biomarkers such as exhaled carbon monoxide and plasma cotinine are widely used to assess recent tobacco exposure[75,76]; however, their short half-life limits their utility in evaluating long-term smoking history or predicting future health risks. More recent approaches, such as DNA methylation-based smoking scores, have demonstrated promise in capturing cumulative smoking exposure[77,78]. However, these biomarkers have rarely provided additional predictive power for incident health outcomes beyond self-reported smoking status, limiting their broader applicability in epidemiological studies and clinical settings.

The constrained predictive ability of DNA methylation-based smoking scores may be attributed to several factors. First, most DNA methylation studies to date have been conducted on relatively small cohorts, often consisting of only a few thousand participants, which restricts their statistical power. Second, the follow-up duration in these studies has generally been short, hindering the ability to establish credible associations between epigenetic changes and long-term disease risks. Consequently, while DNA methylation markers have been extensively validated for distinguishing current and never smokers, their effectiveness in predicting incident health outcomes, such as cancer or cardiovascular diseases, remains suboptimal.

Moreover, most smoking-related biomarker studies have employed traditional linear modelling approaches, which are inherently limited in capturing the complex, non-linear relationships between smoking exposure, biological responses, and disease progression[77,78]. Given the dynamic and multi-faceted nature of smoking-induced physiological changes, advanced machine learning techniques such as gradient boosting could offer improved predictive accuracy by incorporating high-dimensional interactions among biomarkers.

Unlike genetic determinants, lifestyle factors like smoking are modifiable, offering opportunities for intervention. Cessation of smoking significantly reduces the risk of developing age-related diseases and slows the rate of biological ageing. A gap in our knowledge is how to monitor the residual risk over time in past smokers. Additionally, public health initiatives focusing on smoking cessation, promoting healthy diets, and encouraging regular physical activity have demonstrated substantial benefits in extending health span and improving overall quality of life[79].

Challenges to overcome in human ageing research

Current research on biological ageing and age-related risk factors predominantly relied on linear models, despite the well-established non-linearity in the associations between blood biomarkers and age[80]. While these models have provided foundational insights, their limited ability to capture complex interactions among biomarkers and non-linear relationships with ageing has constrained their predictive accuracy. More recent studies have adopted machine learning approaches, such as random forest models, which have demonstrated improved correlations between predicted and chronological age in omics-based ageing clocks[40,81].

However, random forest models, being ensemble-based and relatively opaque, do not fully exploit the hierarchical relationships within high-dimensional biomarker datasets.

Gradient boosting methods have emerged as powerful alternatives, particularly for highly correlated data[82]. These models iteratively refine predictions by minimising residual errors, enabling more precise age estimation compared to traditional machine learning techniques[82]. Notably, prediction models of ageing built with gradient boosting have exhibited superior generalisation when externally validated, while neural network-based models experienced heavy overfitting, reinforcing the gradient boosting model's robustness across diverse populations[51]. As part of my DPhil, I have developed a gradient-boosting-based method for assessing biological age based on metabolomics that will be discussed in Chapter 2 and has been published as part of a more general paper on proteomic ageing[70].

Beyond methodological limitations, current studies on ageing clocks are constrained by sample size and follow-up duration[40]. In the absence of a universally accepted golden standard for biological age, researchers frequently use associations between ageing clocks and age-related diseases or mortality as benchmarks. Larger cohort studies with extended follow-up periods are essential to enhance statistical power and enable more rigorous assessments of the relationships between biological clocks, multi-organ disease risk, and systemic ageing processes.

Another knowledge gap is the sex-specific differences further complicate ageing prediction[83]. Male and female populations exhibit distinct aging trajectories, influenced by genetic, hormonal, and environmental factors[83]. Yet, while prior studies have investigated the differential predictive capacity of ageing clocks in males and females with disease risks, none have

constructed sex-specific ageing clocks[84,85]. Given the known disparities in ageing between the sexes discussed above, developing separate predictive models may enhance the accuracy and uncover pathways with a differential importance between males and females.

In summary, while machine learning methods have substantially advanced ageing clock development and ageing-related risk factors modelling, several challenges remain. These include the need for more sophisticated models capable of capturing non-linear and hierarchical interactions, the necessity for larger and longer-term cohort studies to benchmark the model with incident health outcomes, and the imperative to develop sex-specific ageing models. Addressing these limitations will be pivotal in refining ageing clocks and translating their insights into clinical and epidemiological applications.

Thesis objectives

The primary aim of this thesis is to elucidate the relationship between multi-omics data, specifically metabolomics and proteomics, and the ageing process, as well as the primary risk factor associated with ageing, smoking. I seek to construct machine-learning-based scores that quantify these relationships while accommodating non-linear patterns and interactions inherent in omics data. I evaluate the associations between these scores and biochemical blood biomarkers, clinical risk factors, and their ability to differentiate incident disease risks, in addition to conventional risk factors. The specific objectives are as follows:

1. To develop a robust machine learning-based pipeline to capture molecular changes underly biological ageing and age-related risk factors based on proteomics and metabolomics measurement in plasma (chapter 2)

2. To study differences between males and females in ageing through the metabolic ageing clock, identifying key contributors, studying its underlying pathways and assessing their potential to differentiate disease risks across various organ systems (Chapter 3)
3. To study differences between males and females in ageing through the proteomic ageing clock, identifying key contributors, studying its underlying pathways and assessing their potential to differentiate disease risks across various organ systems (Chapter 4)
4. To study the pathological effects induced by smoking using proteomic data. To develop a score to quantify smoking-related biological damage, and assess the extent of recovery in previous smokers (Chapter 5)

Figures

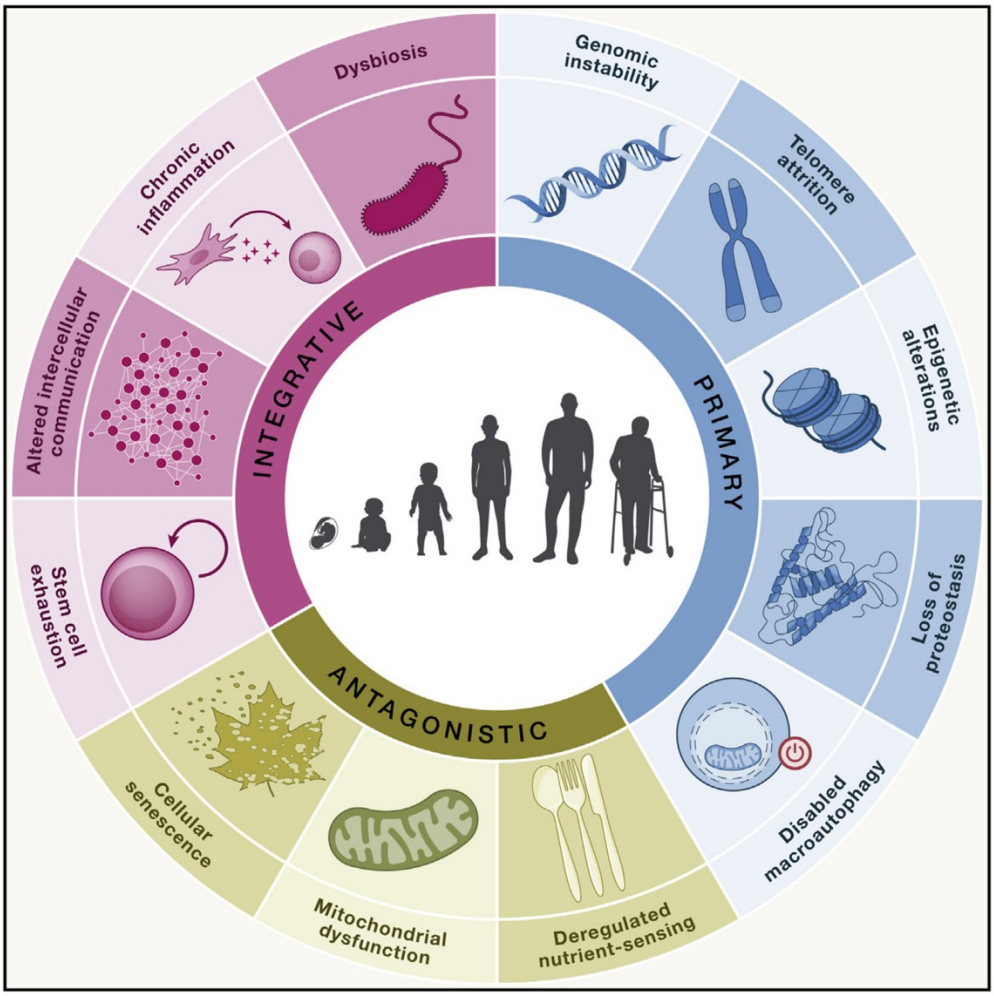


***Figure 1*: Hallmarks of aging.** Lopez-Otin proposed 12 hallmarks of aging and grouped them into 3 categories: integrative, antagonistic and primary. Figure cited from López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. Cell 186, 243–278 (2023).

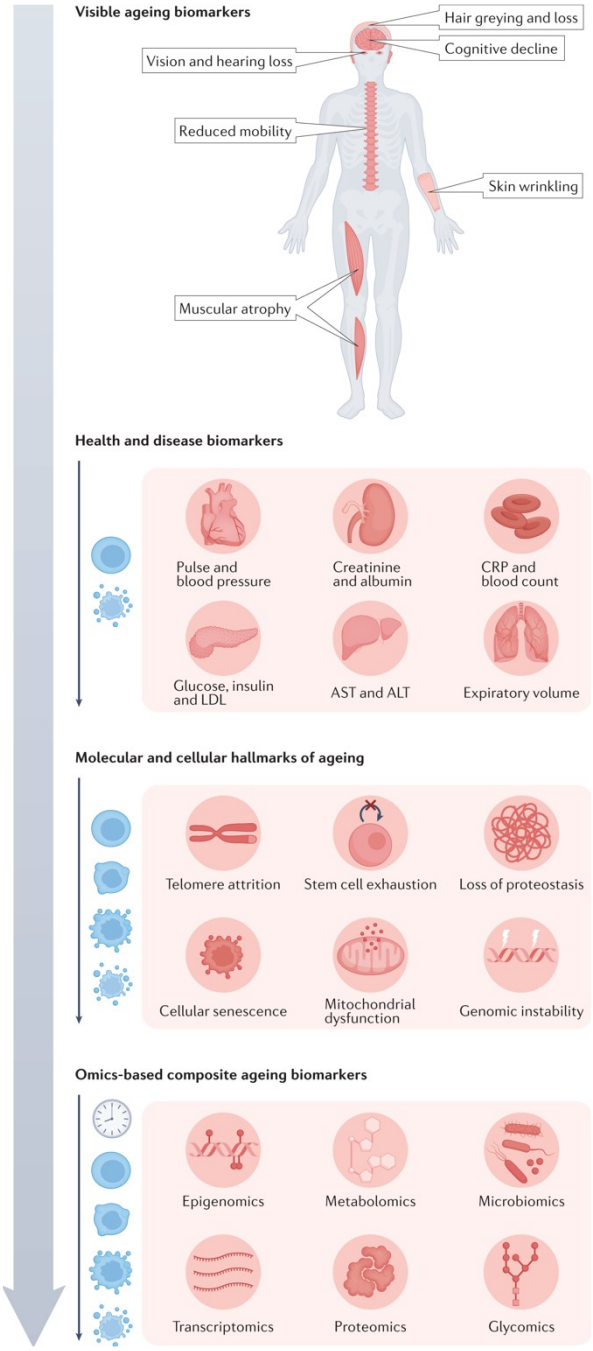


Figure 2: Aging clock can be built from various features. Obvious features of ageing (top), such as muscular frailty and greying hair. (second from top), such as blood pressure, inflammatory markers and metabolic markers, became the primary focus. Hallmarks of ageing (third from top), such as telomere shortening and cellular senescence, became the modern scientific framework for understanding ageing that has guided investigation of ageing at the molecular level. This has led, in part, to the development of omics-based ageing clock biomarkers of ageing (bottom), which attempt to integrate the entire breadth of molecular changes that occur with ageing into composite measures of biological age. Figure cited from Rutledge, J., Oh, H. & Wyss-Coray, T. Measuring biological age using omics data. Nat Rev Genet (2022) doi:10.1038/s41576-022-00511-7.

Chapter 2 Method development

Declaration

Part of the method description and results shown in this chapter was either under review by Nature Communications (Sihao, X., et al. Proteomic signatures of smoking and their associations with risk of incident diseases and mortality in diverse populations) or published in Argentieri, M. A., Sihao, X., et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

Introduction

In the first year of my DPhil, I set out to build a biological-age clock from the metabolomics data in the UK Biobank, leveraging the unprecedented scale of nearly half a million participants. This same modelling framework—optimized in my metabolomics work—was then repurposed in a companion proteomics study in Nature Medicine, in which I contributed the methodology and pipeline (***Fig 1a, b***). Briefly, we first "discovered" the proteomic age clock in UK Biobank participants with Olink data, then replicated its performance in two independent cohorts (China Kadoorie Biobank and FinnGen) with distinct genetic and environmental backgrounds. The consistent variance explained and well-calibrated age predictions in these external samples confirmed the generalisability of our approach.

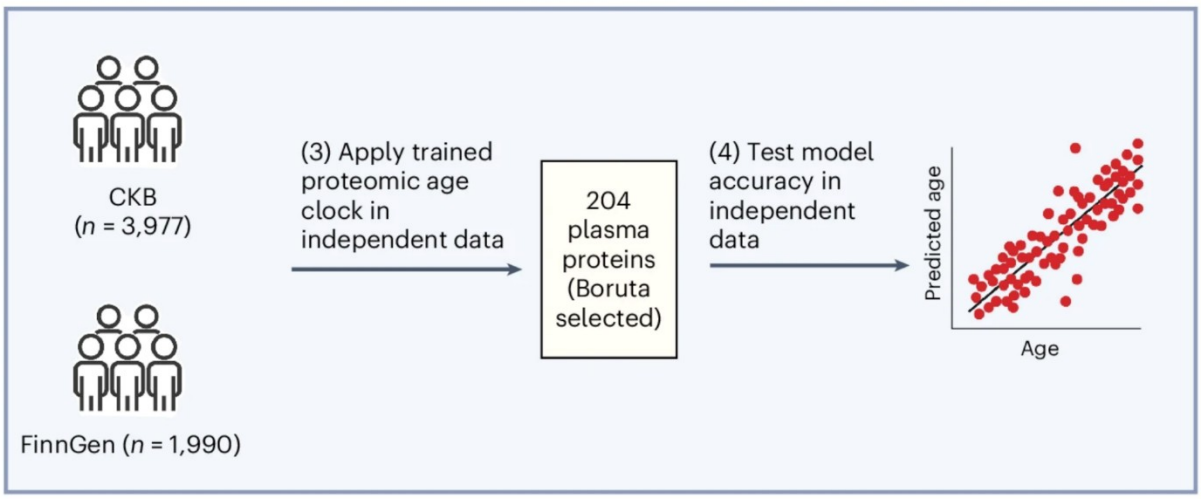
To capture the non-linear, high-dimensional relationships between hundreds of small-molecule biomarkers and chronological age, I compared three machine-learning approaches—Lasso, elastic net regression, gradient boosting, and three neural networks—and ultimately found that

tree-based models offered both the best predictive accuracy and the greatest robustness to sex-specific metabolic differences. Recognising that men and women follow distinct metabolic ageing trajectories, I trained separate clocks for each sex, trusting that gradient-boosted trees would automatically account for the major sex-driven shifts in metabolite distributions.

Finally, extending this pipeline beyond age prediction, I developed an analogous proteomic-smoking index, to model the molecular changes induced by the most important modifiable contributors of ageing: using the same gradient-boosting + Boruta feature-selection strategy to distinguish current from never smokers by their plasma protein profiles, then validating the index in independent populations. In the pages that follow, I will describe (1) the three cohorts used to discover and validate the unified proteomic clock, (2) the detailed modelling steps—data preprocessing, hyperparameter tuning, cross-validation, and feature-selection—used in both age and smoking-index applications, and (3) the downstream assessments of model performance, including accuracy metrics, external replication, and sensitivity.

Study populations

The **UK Biobank (UKB)** is a large, prospective cohort study that has collected extensive genetic and phenotypic data from 502,505 individuals residing in the United Kingdom, recruited between 2006



included 45,44

selected from t

selected by exc

be highly repre

ID 21022) is provided only as a whole integer, we derived a more precise estimate by using the month of birth (field ID 52) and year of birth (field ID 34) to construct an approximate date of birth, assuming the first day of the recorded birth month and year. Age at recruitment was then calculated as the number of days between the recruitment date (field ID 53) and this estimated birth date, divided by 365.25. Ages at the first and repeat imaging follow-ups were calculated by adding the elapsed time (in years) since recruitment to the decimal age at recruitment. During 11–16 years of follow-up, 4,828 participants (10.6%) died.

The **China Kadoorie Biobank (CKB**) is a prospective cohort of 512,724 adults aged 30–79 years, recruited between 2004 and 2008 from ten geographically diverse regions of China (five rural and five urban). The study design and methodology have been described in detail previously[89]. For this study, I included 3,977 participants from a nested case-cohort study of ischemic heart disease (IHD) with baseline Olink Explore proteomic data, ensuring all were genetically unrelated (54% female, age range 30–78 years). Age at recruitment in the CKB was already available as a decimal value. Proteomic profiling in the CKB was performed across two batches at Olink laboratories in Uppsala and Boston, with normalisation procedures applied using internal and inter-plate controls. Over 11–14 years of follow-up, 1,426 participants (36%) died.

FinnGen is a public-private research initiative integrating genomic and health registry data from approximately 500,000 participants across Finland to investigate the genetic basis of disease[90]. The project involves nine Finnish biobanks, research institutions, universities, university hospitals, 13 international pharmaceutical partners, and the Finnish Biobank Cooperative (FINBB). FinnGen utilises nationwide longitudinal health register data collected since 1969. For our analysis, 1,990 participants who were selected free from the disease were measured with

Olink Explore and passed proteomic quality control (52% female, age range 19–78 years). Proteomic profiling in FinnGen was conducted in three batches at Olink's Uppsala laboratory, with batch effects minimised by the inclusion of bridging samples and standardised normalisation protocols. Proteomic profiling in FinnGen primarily targeted healthy individuals, and only 1% (n = 22) of participants died during follow-up.

The age distribution of the three cohorts is shown in ***Fig 2a***. The distribution of age of death of UKB and CKB was shown in ***Fig 2b***. The number of prevalent cases for common diseases in UKB and the number of incident cases during follow-up are shown in ***Fig 2c***.

Proteomic Profiling

Across all cohorts, plasma proteomic profiling was performed using the Olink Explore 3072 platform, which comprises four panels (Cardiometabolic, Inflammation, Neurology, and Oncology). Data were reported in Normalised Protein eXpression (NPX) units on a log2 scale. Proteomic data were preprocessed and normalised within each cohort using internal controls and a predetermined correction factor. To ensure consistency, we restricted analyses to proteins measured across all three cohorts and excluded an additional three proteins with over 10% missingness in the UKB (CTSS, PCOLCE, and NPM1), resulting in a final set of 2,897 proteins. After the imputation of missing values, proteomic data were scaled between 0 and 1 using MinMaxScaler() from scikit-learn and centred on the median[91].

Missing data imputation

Although gradient-boosting algorithms can handle missing values natively, linear models require complete data for training. Missing protein expression values were imputed in Python using the

miceforest package with default settings. For each target protein, only proteins missing in ≤30% of participants were used as predictors in the imputation model. The imputed dataset was used only for the model-comparison experiments; all primary model development in the main analysis used the non-imputed data.

Methodology for Proteomic Age Prediction and Feature Selection

To estimate proteomic age, UK Biobank (UKB) participants (n = 45,441) were first partitioned into a training set (70%; n = 31,808) and a holdout test set (30%). Using the training data, we developed a gradient boosting model to predict participants' age at recruitment based on measurements from 2,897 proteins. Model optimisation was performed through hyperparameter tuning with fivefold cross-validation, leveraging a Tree-structured Parzen Estimator (TPE) based method provided by the Optuna module in Python[92]. A total of 200 trials were conducted to identify the optimal hyperparameters within a preset range of hyperparameters that maximised the average coefficient of determination (R^2) across all folds, ensuring robust model performance. Key hyperparameters include *num_leaves*, which controls the complexity of the tree structure; *subsample*, which determines the fraction of data sampled for training each tree; and *min_child_samples*, which specifies the minimum number of data points required to create a leaf node[82]. Additionally, the *learning_rate* was adjusted to balance the step size during optimization, while *min_child_weight* was used to regulate the sum of instance weights in child nodes[82]. Feature sampling was controlled via *colsample_bytree*, and regularisation was applied through *reg_alpha* (L1 regularisation) and *reg_lambda* (L2

regularisation) to prevent overfitting and improve generalisation[82]. The combination of L1 and L2 regularization (or weak L2 regularization) encourages the model to distribute the importance of correlated features more evenly, rather than forcing one's contribution to zero entirely[93]. This will reduce the chance of the model randomly choosing which feature to use when encountering a large cluster of correlated features and create bias when analyzing selected features.

To characterise the feature importance, SHapley Additive exPlanation (SHAP), a local tree explaining method based on game theory, was used[94]. SHAP calculates the contribution of each feature to the outcome in each participant and extends these local explanations to also capture interactions between features directly. Compared to traditionally used permutation feature importance, SHAP plots can display the magnitude, prevalence, and direction of a feature's effect. We then used a SHAP-based Boruta selection method provided by shap-hypetune package[95] to select all relevant features contributing to smoking status prediction. The Boruta method enhances feature selection by introducing shadow features—randomly permuted duplicates of the original features—which serve as a benchmark for relevance[96]. At each iteration, a gradient boosting model was trained on the full set of original and shadow features, and features that did not exceed the highest absolute SHAP value of the shadow features were eliminated. This process iterated until only features with statistically significant contributions to age prediction remained. 200 Boruta trials were conducted, applying a stringent 100% threshold, meaning that a feature was retained only if it outperformed all shadow features in

every trial. This approach ensured that only the most biologically relevant proteins were included in the final model, minimising the influence of noise.

Following feature selection, the gradient boosting model was trained using the identified subset of proteins and re-optimised its hyperparameters following the same fivefold cross-validation procedure. To ensure model generalisability and mitigate overfitting, we validated both the full-feature and Boruta-selected models through cross-validation within the training data and assessed their performance on the independent UKB test set. Throughout all modelling steps, we set gradient boosting to run with 5,000 estimators, 20 early stopping rounds, and R^2 as the custom evaluation metric to maximise the variance explained in age prediction. Models were then externally validated by using a model trained from the UKB training dataset and calculating predicted age in external cohorts.

The Boruta selected model achieved high accuracy in predicting chronological age with an R^2 of 0.88 and a person r of 0.94 (***Fig2d***). The model trained in UKB was also externally validated in two external cohorts with an R^2 of 0.82 in CKB and 0.87 in FinnGen (***Fig2e, f***).

In the paper, we also benchmarked a variety range of models and tested them in an external cohort to validate their performance and generalisability. We evaluated the performance of six machine learning models—LASSO, elastic net, LightGBM, and three neural network architectures (multilayer perceptron, residual feedforward network (ResNet), and retrieval-augmented neural network for tabular data (TabR))—for predicting chronological age from plasma proteomic data. In all models, we used 2,897 Olink protein expression variables as input

to train a regression model for age prediction. The models were trained using five-fold cross-validation on the UK Biobank (UKB) training dataset (n = 31,808) and evaluated on the UKB holdout test set (n = 13,633), along with independent validation sets from the China Kadoorie Biobank (CKB) and FinnGen cohorts.

LASSO and elastic net models were implemented using the scikit-learn package in Python[91]. For the LASSO model, we optimized the alpha parameter using the LassoCV function, testing a range of values: [1 × 10−15, 1 × 10−10, 1 × 10−8, 1 × 10−5, 1 × 10−4, 1 × 10−3, 1 × 10−2, 1, 5, 10, 50, 100]. Elastic net models were tuned for both alpha (using the same range) and the L1 ratio, which was evaluated across the values: [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1].

For the LightGBM model, hyperparameter tuning was performed using five-fold cross-validation with the Optuna module in Python, conducting 200 trials to optimise parameters for maximum average R2 across all folds[82].

The neural network architectures tested in this analysis were selected from a list of architectures that performed well on a variety of tabular datasets. The architectures tested included (1) multilayer perceptron, (2) ResNet, and (3) TabR. Hyperparameter tuning for all neural network models was carried out via five-fold cross-validation with Optuna, conducting 100 trials per architecture and optimizing for maximum average R2 across all folds.

In the UKB test set, both LightGBM and neural network–based methods demonstrated the best performance, with R^2 values exceeding 0.875; notably, ResNet achieved the highest R^2 of 0.880 (**Fig 3**)[97]. However, neural network–based models showed limited generalizability in external validation. In CKB, the performance of both MLP and TabR fell below that of Elastic Net,

whereas LightGBM emerged as the top performer with an R^2 of 0.847. A key factor contributing to the performance decline of neural network models was their inability to accurately predict participants younger than 40 or older than 75 years, as these age groups were not represented in the UKB cohort. This limitation became even more pronounced when validating the models in a cohort with a wider age range such as FinnGen. In FinnGen, neural network–based models exhibited clear cutoffs in predicted protein age around 40 and 75 years, leading to reduced performance that was even inferior to that of LASSO regression. By contrast, LightGBM consistently maintained the highest performance, achieving an R^2 of 0.867. The superior generalizability and robustness of LightGBM in predicting unseen data underpin our decision to adopt this model for subsequent analyses.

The sex specificity of the proteomic age model was also assessed by building a sex-stratified model and comparing the performances of age prediction between the sex-stratified model and sex unified model. Due to the nature of the tree-based model where branches can be developed when sex differences were notified by the gradient boosting model, as expected, performances between sex stratified model and sex unified model were similar and the correlation of age predictions was high (r=0.99 in female and r=0.98 in male) (***Fig4 a-d***). However, when looking at the SHAP values of the top 20, 9/20 in females and males are different (FSHB, PAEP, LECT2, GIP, AFP, CCDC80, IGDCC4, SUSD5, HAVCR1 in females and TSPAN1, KLK4, CXCL14, CDON, RET, AGRP, KLK3, ACTA2, ADAMTS16 in males) (***Fig5a, b***). This included FSHB (crucial to ovarian follicle development and estrogen production[98]) and PEAP (produced by the endometrium and involved in modulating the female reproductive tract environment[99]) which ranked at number 3 and 4 in females and KLK3 and KLK4 (predominantly produced in

prostate[100]) which ranked at number 4 and 17 in males. These differences indicated that although both models predicted age accurately, they relied on distinct proteins and captured different biological pathways. Therefore, to fully account for sex-specific ageing mechanisms, it is essential to build and analyze sex-stratified models when selecting proteomic biomarkers.

With the optimized model trained on Boruta-selected proteins, we calculated proteomic age (ProtAge) for all UKB participants (n = 45,441). This was achieved using a fivefold cross-validation approach, where a gradient boosting model—configured with the final set of hyperparameters—was trained on each fold's training subset and used to predict age in the corresponding test subset. The predicted values from all folds were then aggregated to generate a unified estimate of ProtAge for the full cohort.

Methodology for Proteomic Smoking INdex and Feature Selection

Similar methodology and pipeline were then transformed to deal with the classification problem imposed by the smoking project. Proteomic profiles of smoking were constructed by comparing current smokers with never smokers excluding passive smokers using gradient boosting. Samples were randomly split into 70% training (n=13,343 for never-smokers, n=3,312 for current smokers) and 30% testing dataset (n=5,719 for never-smokers, n=1,420 for current smokers). Similar to the model comparison step in the proteomic age calculation, here I have also compared the prediction ability of Lasso, Elastic net and LightGBM models. Withing the training dataset, LightGBM model obtained the highest 5-fold CV Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) with an AUC of 0.960 (SD=0.004), followed by the elastic net model with an AUC of 0.943 (SD=0.002). Unsurprisingly, the LASSO model had

the lowest AUC (AUC=0.933, SD=0.003). In the 30% UKB testing dataset, the elastic net model and the LightGBM model obtained a very similar AUC of 0.949 and 0.947 while the LASSO model still held the lowest AUC of all (AUC=0.940) (***Fig 5a***). However, when validated externally, the LightGBM showed the best generalisability, achieving an AUC of 0.914 (***Fig 5b***). This is followed by the elastic net model with an AUC of 0.908 and the LASSO model with an AUC of 0.869 (***Fig 5b***). Considering the great generalisability, the ability to capture non-linear relationship between proteins and the ability to consider missing value natively, LightGBM was chosen for the main analysis.

For the main analysis, within the training dataset, a gradient boosting machine learning model including all 2,911 proteins was trained to differentiate never smokers and current smokers. The model was the first hyperparameter tuned using a Tree-structured Parzen Estimator (TPE) based method provided by the Optuna package in Python[92]. Hyperparameters within a pre-set range were searched and optimised across 200 trials to maximise the 5-fold cross-validated ROC AUC score. After hyperparameter tunning, the performance of the best parameter in the training dataset with 5-fold cross-validation and in 30% left out testing dataset was assessed.

To characterise the feature importance, SHapley Additive exPlanation (SHAP), a local tree explaining method based on game theory, was used[94]. SHAP calculates the contribution of each feature to the outcome in each participant and extends these local explanations to also capture interactions between features directly. Compared to traditionally used permutation feature importance, SHAP plots can display the magnitude, prevalence, and direction of a feature's effect. We then used a SHAP-based Boruta selection method provided by shap-hypetune package[95] to select all relevant features contributing to smoking status prediction. The Boruta

algorithm enhances feature selection by creating randomly permuted shadow features that serve as a baseline for comparison[96]. Specifically, shadow features are duplicates of the original features with their values randomly shuffled to break any associations with the target variable. The algorithm compares the mean absolute SHAP values of the original features against those of the shadow features. Features are retained only if they demonstrate a higher mean SHAP value than the highest-scoring shadow feature, thereby ensuring their predictive importance is not due to random chance.

In our study, the algorithm was executed for 200 iterations to ensure robustness and stability in feature ranking. Features falling within the bottom 5% of importance scores (tail 5%) were systematically rejected as they were unlikely to contribute meaningfully to the model's performance. Following the feature selection process, the refined model—consisting of Boruta-selected features—underwent another round of hyperparameter tuning to optimise its performance before further analysis. All model tuning and feature selection steps were performed within the 70% training set in UKB.

The Proteomic Smoking INdex (pSIN) for the full UKB sample (n=43,914) was calculated using a robust methodology to mitigate the risk of overfitting. This process involved employing 5-fold cross-validation to ensure the reliability of the results. After identifying the best hyperparameters and selecting the proteins using the Boruta method, a gradient boosting model was trained within each fold. Subsequently, the predicted raw score for the corresponding test set was generated. For binary classification tasks, this raw score corresponds to the log odds of the positive class (in this case, being a current smoker). The LightGBM model typically outputs raw scores (logits) in the range of approximately -10 to 10, where a score of 0

indicates a neutral prediction, corresponding to a 50% probability of being in the positive class (sigmoid(0) = 0.5). Scores closer to 10 indicate a high confidence prediction for the positive class (sigmoid(10) ≈ 0.99995, or ~99.995% probability), while scores closer to -10 indicate a high confidence prediction for the negative class (sigmoid(-10) ≈ 4.5×10^{-5}, or ~0.0045% probability). In our analysis, we set the classification threshold at a raw score of -1.29, which corresponds to the point where the false positive rate (FPR) is 0.05. This threshold was chosen to balance sensitivity and specificity in our predictions. pSIN higher than the threshold indicates a higher likelihood of being in the positive class (smoker), with larger values reflecting greater confidence, while pSIN smaller than the threshold indicate a higher likelihood of being in the negative class (non-smoker), with more negative values reflecting greater confidence. These predicted raw scores from the test sets of each fold were then aggregated to create a comprehensive measure of smoking protein profiles for the entire population. This approach allowed for a more robust estimation of the impact of smoking on protein profiles across the UK Biobank cohort compared to using the model trained using 70% training data to calculate pSIN for the entire population. External validation in CKB was performed to further test the possibility of the overfitting problem. For external validation, the model with the optimized hyperparameter was trained in the UKB training dataset and was tested in the CKB. Performances of identifying current smokers from never smokers were compared.

Conclusion

In this chapter, I have detailed the development and validation of a robust, gradient-boosted framework for deriving biologically informative indices from high-dimensional plasma metabolomic and proteomic data. Beginning with large, population-scale cohorts (UK Biobank, China Kadoorie Biobank, and FinnGen), I described how rigorous data preprocessing—including quality control, imputation, and feature scaling—ensured a consistent proteomic matrix of 2,897 proteins across all studies. Leveraging LightGBM coupled with Optuna-guided hyperparameter tuning and SHAP-based Boruta feature selection, I constructed a proteomic age clock that explained approximately 88% of age variance in the UK Biobank holdout set and retained strong performance when applied to CKB and FinnGen. By comparing conventional and neural network architectures, I demonstrated that gradient-boosted trees not only outperformed alternative approaches in predictive accuracy but also exhibited superior generalisability, especially across cohorts with different age distributions and demographic characteristics. Furthermore, sex-stratified analyses underscored the importance of capturing sex-specific ageing signatures: although a combined model accurately predicted age in both males and females, sex-stratified models uncovered distinct protein drivers—such as hormonal markers (e.g., FSHB, PAEP) in females and prostate-related proteins (e.g., KLK3, KLK4) in males—thereby highlighting divergent biological pathways underpinning ageing in each sex.

Extending this modelling pipeline, I also developed a proteomic smoking index (pSIN) using the same LightGBM + Boruta procedures. By defining pSIN as the raw LightGBM log-odds score, I provided a continuous measure of smoking-related proteomic perturbation, rather than a

simple binary classification, trying to quantify and study both risk and recovery dynamics of the largest modifiable risk factor of ageing - smoking.

The approaches presented here lay a solid foundation for subsequent analyses of how plasma omic-biomarkers relate to health outcomes, disease risk, and mortality. By integrating advanced machine-learning techniques with rigorous external validation, this chapter has shown that a robust pipeline has been generated to build accurate proxies for quantitative biological ageing and qualitative measures of lifestyle exposures. Moving forward, scores generated using this pipeline will be applied in later chapters to investigate their associations with incident disease events, all-cause and cause-specific mortality, and the interplay with genetic and environmental modifiers.

Figures

Fig 2. Proteomic aging clock performance across cohorts. A) Density plot of age at recruitment in the UKB, CKB and FinnGen. **B)** Density plot of age at death in the UKB (4,784 deaths; 10.6%) and CKB (1,426 deaths; 36%). **C)** Counts of prevalent and incident cases of all common diseases studied in the UKB sample (*n* = 45,441). **D)** Performance of the trained proteomic aging model in the UKB holdout test set (*n* = 13,633). **E)** Performance of the trained proteomic aging model in the CKB (*n* = 3,977). **F)** Performance of the trained proteomic aging model in FinnGen (*n* = 1,990). Figure cited from Argentieri, M. A., Sihao, X., et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

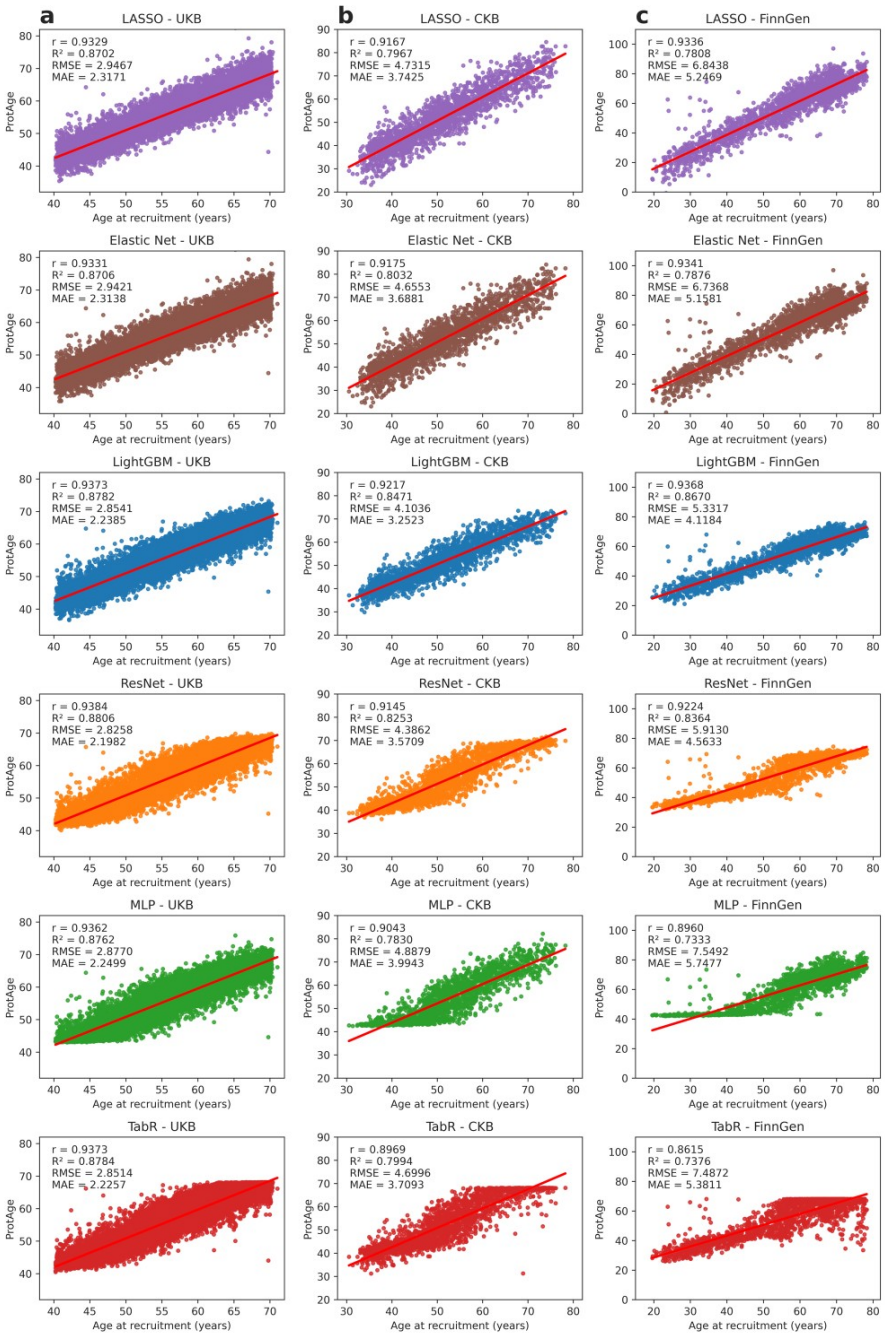


Fig 3. Model benchmarking for estimation of proteomic age in the UK Biobank and China Kadoorie Biobank. Scatterplots comparing actual chronological age (x-axis) versus protein predicted age (protAge; y-axis) in **a)** the UK Biobank test set (n=13,633); **b)** China Kadoorie Biobank (n=3,977); and **c)** FinnGen (n=1,990). Models compared included two penalized linear regression models (LASSO, elastic net), one gradient boosting machine learning model (LightGBM), and three neural network architectures (ResNet, MLP, TabR). LASSO: least absolute shrinkage and selection operator; MAE: mean absolute error; MLP: multilayer perceptron; RMSE: root mean square error. Figure cited from Argentieri, M. A., Sihao, X., et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

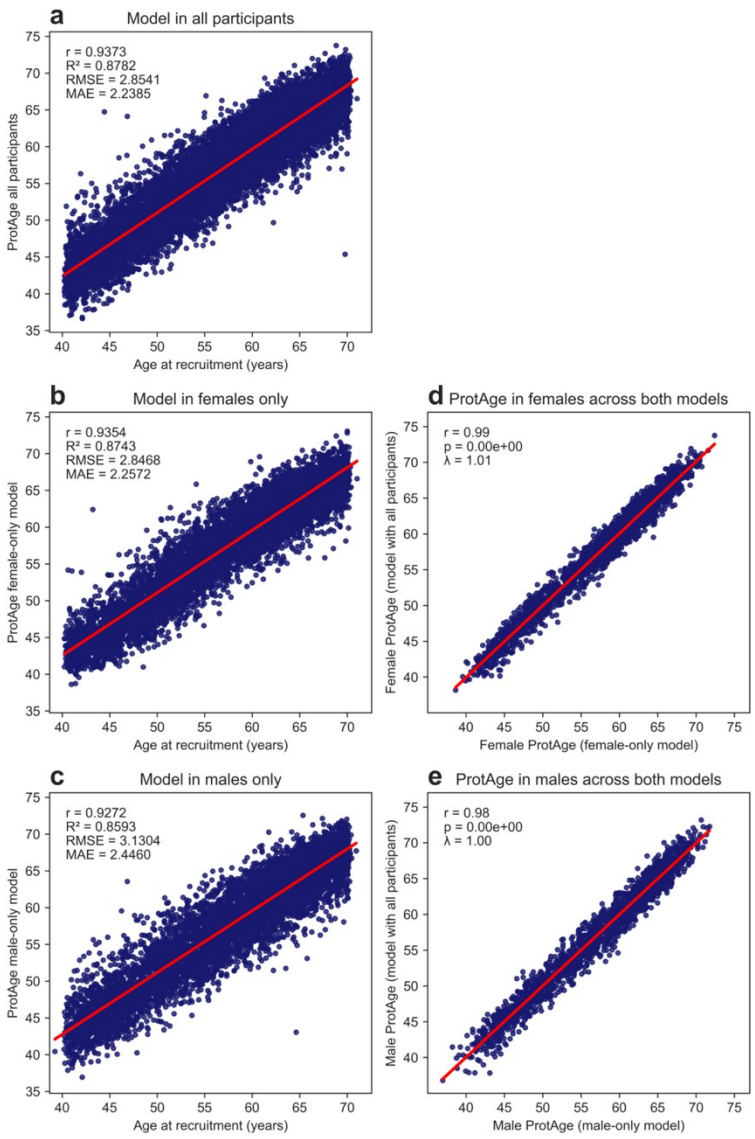


Fig 4. Proteomic age estimation accuracy by sex. Comparison of actual chronological age versus protein predicted age (protAge) for a model using: a) all participants (n=13,633), b) female participants only (n=7,374), c) male participants only (n=6,259). Model accuracy metrics comparing predicted versus actual age values are shown as Pearson r correlation coefficient, R^2 , root mean square error (RMSE) and mean absolute error (MAE). d) Comparison of protein predicted age (protAge) for the same female participants from the all-participant model (y-axis) and model with only female participants (x-axis). e) Comparison of protein predicted age (protAge) for the same male participants from the all-participant model (y-axis) and model with only male participants (x-axis). Figure cited from Argentieri, M. A., Sihao, X., et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

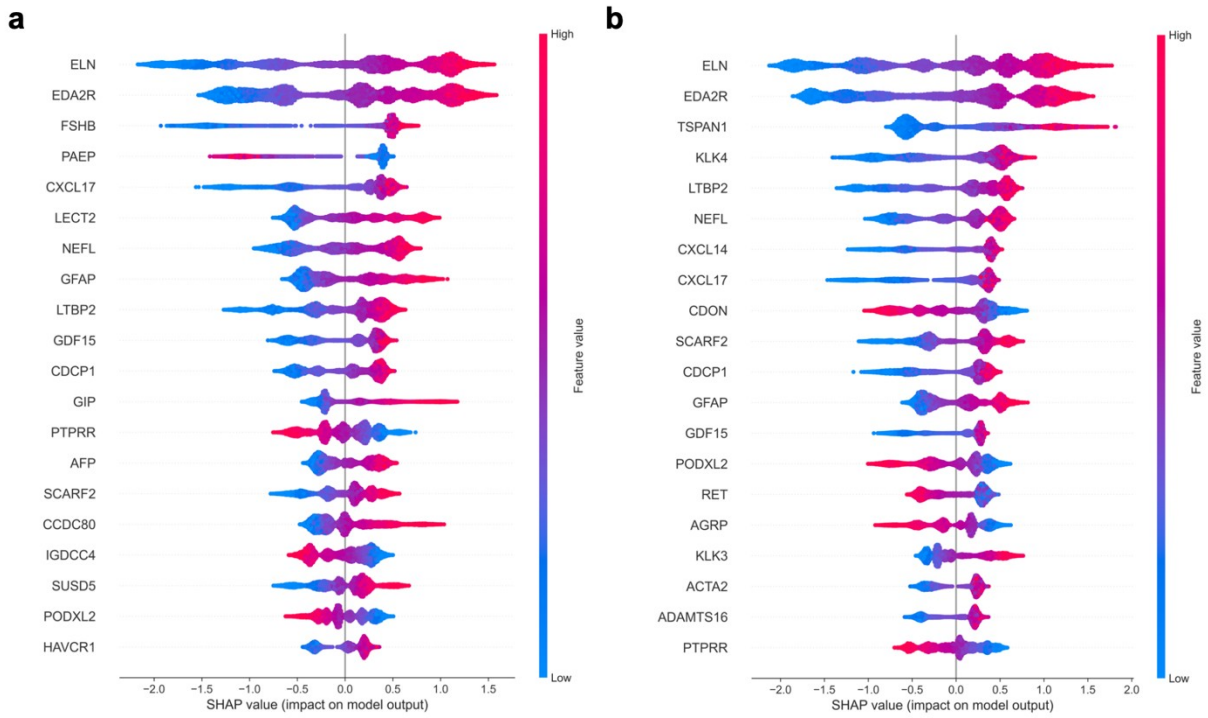


Fig 5. Top 20 most important proteins for predicting age in females and males. Comparison of the top 20 most important proteins for predicting chronological age for a) females (n=7,374) and b) males (n=6,259). Protein importance was calculated for each protein by taking the mean of the absolute SHAP values across all participants for that protein. The x-axis shows the SHAP values from the sex-specific LightGBM model for each protein, with the SHAP value for each individual participant as an individual dot. The sign of SHAP values represents the effect on the model outcome. Dots are color coded according to the value of the protein feature. Figure cited from Argentieri, M. A., Sihao, X., et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

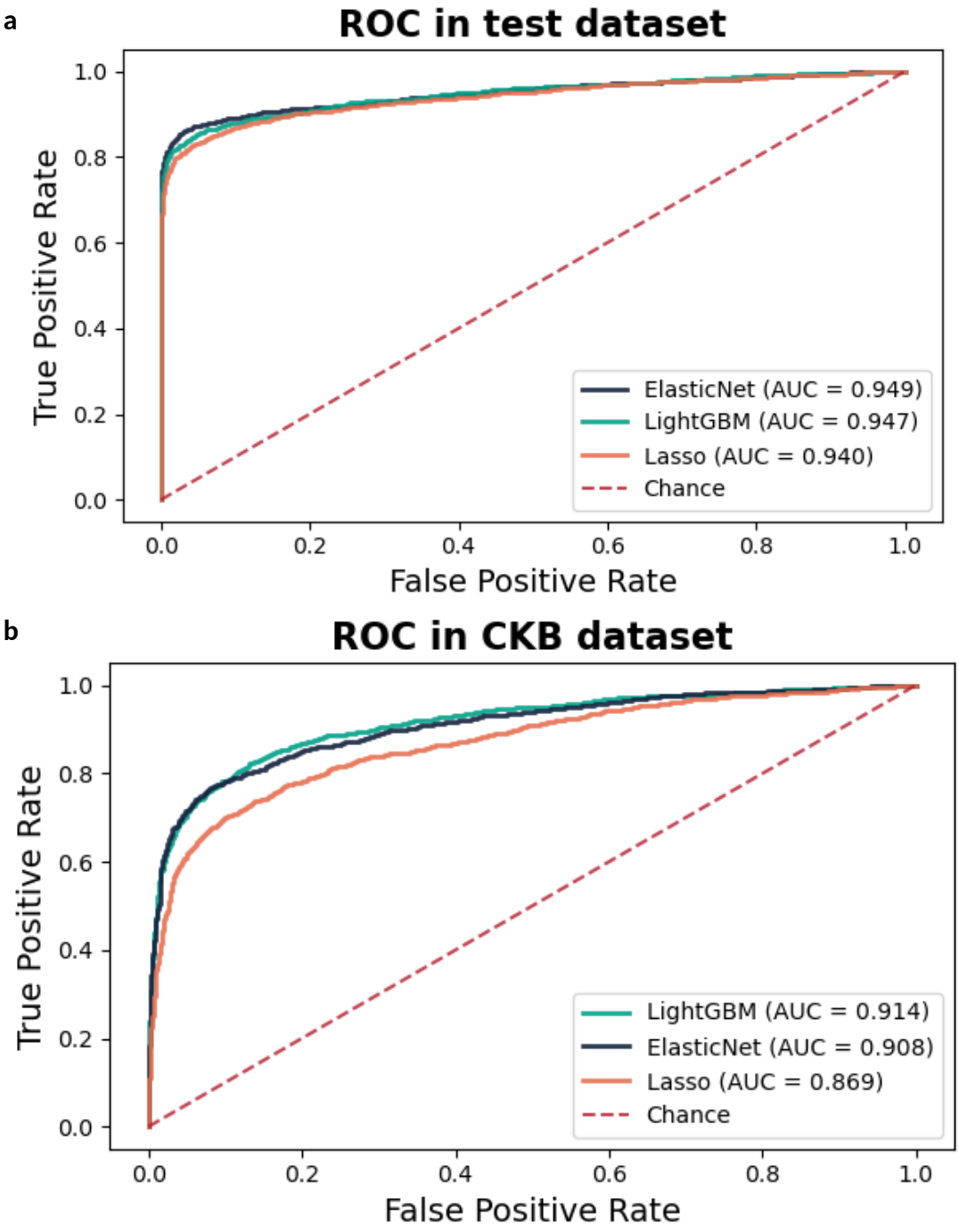


Fig 6. Model benchmarking for estimation of proteomic smoking index in the UK Biobank and China Kadoorie Biobank. ROC plots of performance of different in **a)** the UK Biobank test set and **b)** China Kadoorie Biobank were shown. Models compared included two penalized linear regression models (LASSO, elastic net) and one gradient boosting machine learning model (LightGBM).

Chapter 3 Unveiling Sex Differences in Ageing through Metabolomics Clock

Declaration

Content of this chapter is planning to submit to Nature Medicine (Sihao, X., et al. Ticking Differently Over Time: Deciphering Sex Dimorphism in Human Aging Integrating Metabolomics, Proteomics and Genomics) in September 2025.

Introduction

Females outlive males globally[101] despite the fact that females in general are frailer[102], and have poorer physical functioning compared to males[103]. Further both historical and latest data suggest that females live longer even during severe famines and epidemics[104–107]. These differences have been attributed to various factors, including hormonal influences such as estrogen[108,109], genetics[83], telomere length[3,110,111], metabolic rate[83,112], and behavioural differences[113] where males tend to engage in risky behaviors[114] and often have poorer lifestyle habits, such as higher alcohol consumption and smoking[115,116]. Such environmental insults are captured by epigenetic processes like DNA methylation, which shows a higher entropy in males with advancing age[83,117,118].

Previous biological age predictors, often referred to as "clocks", have been developed to estimate biological ageing across different omics layers, demonstrating their substantial potential in ageing research[22,40,119–121]. Most of these clocks, however, did not investigate sex-

specific variations underlying ageing and the association of biological age with disease and mortality outcomes in each sex. A recent study[122] explored the differences in ageing between sexes but in a relatively small sample of 10,000 individuals using environmental exposures, physiological parameters, and a narrow platform of molecular markers mainly concentrating on lipids.

This study focused on developing metabolites-based ageing clocks leveraging the largest ever available data from 488,318 individuals in the UK Biobank (UKB). Metabolites reflect the outcomes of numerous physiological processes, including lipid metabolism, oxidative stress, and inflammation — all of which are regulated differently in males and females due to hormonal influences[119], genetic differences[123,124], body composition[125], etc. But as metabolites provide a more integrated and stable measure of these processes compared to hormones alone, they are a promising approach to add to the findings by Reicher et al.[122] to learn more about sex-specific ageing.

In this study, I also evaluated the respective biological age acceleration (metAgeGap) — the residual between predicted biological age and chronological age. This reflects the variation in individuals' past rate of ageing compared to others with the same chronological age group. We then tested its association with future risks of cancers and non-cancer common chronic and age-related diseases. We used metAgeGap to predict survival in both sexes. Finally, we performed a genome-wide association analysis of males and females specific metAgeGap to identify key genetic factors associated with metabolic age.

Methods

Study cohort

UK Biobank (UKB)

UKB population characteristics were described in Chapter 2.

Missing data was imputed using a random-forest-based algorithm provided by R package missRanger[126] when used as a covariate in linear association models (Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years). Imputation was performed with default hyperparameters using a maximum of 10 iterations and 200 trees. Linked hospital inpatient data, primary care data and cancer register data were accessed from the UKB data portal in August 2024, with a censoring date of Nov 30 2023, Dec 31 2023, Nov 30 2023 for participants recruited in England, Scotland, and Wales respectively. The follow-up time is between 8 and 16 years. Mortality data and cause of death information were accessed from the UKB data portal in August 2024, with a censoring date of Nov 30 2022. The follow-up time is between 12 and 16 years.

Assessment of Metabolites

NMR metabolic biomarkers were generated by Nightingale Health. Data is currently available for Phase 1 and Phase 2 of the study, measuring 249 metabolic biomarkers for EDTA plasma samples from 448,362 UK Biobank participants. Data from 280,000 individuals is available publicly and this study had pre-access to data from the rest of the individuals. Metabolic biomarkers were measured from randomly selected EDTA (Ethylenediaminetetraacetic acid)

plasma samples (aliquot 3) using a high-throughput NMR-based metabolic biomarker profiling platform developed by Nightingale Health Ltd. The measurements took place between June 2019 and April 2020 (Phase 1) and April 2020 and June 2022 (Phase 2) using eight spectrometers at Nightingale Health, based in Finland. The biomarkers span multiple metabolic pathways, including lipoprotein lipids in 14 subclasses, fatty acids and fatty acid compositions, as well as various low-molecular-weight metabolites, such as amino acids, ketone bodies, and glycolysis metabolites quantified in molar concentration units.

Statistical analysis

A descriptive analysis of population characteristics was performed using the r package CBCgrps[127]. The study design and analysis pipeline are illustrated in ***Fig 1***. The method of predicting biological age, feature importance calculation and feature selection were described in Chapter 2.

Metabolic Age Gap (metAgeGap)

MetAgeGap for the full UKB sample (n=448,362) was predicted using a robust methodology to mitigate the risk of overfitting. This process involved employing 5-fold cross-validation to ensure the reliability of the results. After identifying the best hyperparameters and selecting the metabolites using the Boruta method, a gradient boosting model was trained within each fold. Subsequently, the predicted biological age for the corresponding test set was generated. MetAgeGap was evaluated as the residual between predicted age and chronological age to estimate the variation in individuals' past rate of ageing compared to others with the same

chronological age. This approach allowed for a robust estimation of the biological age across the UK Biobank cohort.

Association of lifestyle, clinical biomarkers and risk factors with metAgeGap

To test the association of self-reported lifestyle factors, blood biochemistry biomarkers, and clinical risk factors with metAgeGap, generalised linear models from statsmodel v.0.14.0 package[128] were used. For continuous exposure variables, standardisation was applied before inclusion in the models. Associations were adjusted for recruitment centre, ethnicity, education years, and Townsend deprivation index. P-values resulting from these analyses were corrected for FDR multiple testing.

Association metAgeGap with future health-related outcomes

To test the association between metAgeGap and incident health outcomes, all prevalent cases were removed beforehand. Multi-variate Cox proportional hazard model provided by lifeline v.0.27.8 package[129] was used with a pre-set step size of 0.1. Survival outcomes were defined using follow-up time to the event and the binary incident event indicator. For all incident outcomes in the whole UKB population, three successive models were tested with an increasing number of covariates: model 1 adjusted for chronological age; model 2 was adjusted for recruitment centre, Townsend deprivation index, and ethnicity; model 3 was further adjusted for physical activity and BMI; model 4 was further adjusted for smoking status and alcohol frequency. P-values of the hazard ratio were corrected for FDR multiple testing. Forest plots were generated with a minimum sample size threshold of 80 to ensure adequate statistical power and reliable interpretation.

Cumulative incidence plots were generated utilising the KaplanMeierFitter function from the lifelines package[129]. Due to limitations in case numbers or at-risk numbers at both ends, the x-axis of the plot was constrained to the age range of 45 to 75. This adjustment ensured a more focused visualisation of the cumulative incidence curve within a clinically relevant age range. P-values between cumulative incidence curves were calculated using a log-rank test with adjustment for FDR multiple testing.

GWAS

Sex-specific GWAS were conducted using SAIGE[130] software version 1.09. For constructing a genetic relationship matrix (GRM) in step 1, we used the pruned genotype dataset. Genotype pruning was conducted in PLINK[131] software using the 'indep-pairwise' option with an r^2 of 0.5, a window size of 1000 markers and a step size of 100 markers. We further used the 'LOCO= TRUE' option to construct the GRM. The GWAS analyses were adjusted for age, ethnicity, batch effects, and 40 genetic principal components identified within UKB genotyping data[132].

Pathway enrichment study

Ensemble variant effect predictor webtool[133] was used to annotate the top-hit variants ($p<5*10^{-8}$). Mapped gene symbols were extracted and R package clusterProfiler was then used to perform molecular function enrichment analysis[134]. The cutoff p-value for molecular function and biological pathway enrichment analysis was set to 0.5 and the top 15 enriched molecular functions were shown on the enrichment network plot.

Results

Descriptive Statistics

The study population comprised 488,318 participants, with 223,396 males and 264,922 females. Participants taking lipids-lowering drugs were removed from this study resulting in 168,460 males (mean age=55.86, sd=8.23) and 222,481 females (mean age=56.01, sd=8.00). Over the follow-up period, which averaged 12.1 years for males and 12.4 years for females, there were 2,343 recorded deaths in males and 1,602 in females. An overview of the participants' selection is shown in ***Fig 1A***.

Males had a higher prevalence of cancers than females except for lung cancer and non-Hodgkin lymphoma. Females had a higher prevalence of most common diseases except cardiovascular (ischemic heart disease, ischemic stroke, all stroke) and neurodegenerative diseases (all-cause dementia and vascular dementia) **(*Table s1*)**. Furthermore, males had higher median systolic/diastolic blood pressure, and more often used alcohol and smoked cigarettes, while females reported higher frequency of suffering from sleep difficulties and tiredness **(*Table s2*)**.

Metabolomic ageing clock for males and females

To build biological age, we built a gradient boosting tree-based model using NMR-metabolomics data from the UK Biobank (UKB) for males and females separately. For both males and females, the dataset was split into 70% training and 30% test sets with participants randomly assigned to each set. We first trained sex-specific metabolic estimators of chronological age with 249 metabolic features in the training data. ***Fig 1B*** summarizes the outline of the study. After

hyperparameter tuning, the mean R^2 for 5-fold cross-validation (CV) within the training dataset was 0.29 (sd=0.002) in males and 0.37 (sd=0.001) in females. The R^2 in the held-out test set was 0.29 in males and 0.37 in females **(*Fig 1C*)**. The distribution of metAgeGap was similar in males and females **(*Fig s1*)**. Subsequently, the Boruta algorithm was used to select all relevant metabolic features in males and females separately. A total of 93 metabolic features were selected in females and 76 in males (***Fig s2***, ***Fig 1C*)**. Among the selected features, 61 were shared between sexes, indicating shared metabolic pathways. 32 metabolomics were uniquely selected for females including cholines, and phosphatidylcholines**,** and 15 metabolites were uniquely selected for males, including HDL size and triglyceride percentages in large HDL **(*Fig s2*)**.

To assess the importance of each feature selected by Boruta, Shapley Additive exPlanations (SHAP) were used. Some shared features contributed differently across sexes - for instance, glutamine (Gln) ranked 5th in females but only 12th in males (***Fig s3,s4***).

Correlation of disease-related risk factors with metAgeGap

Next, we tested the association of metAgeGap with 22 common disease-related risk factors. Obesity (females: beta=0.94, $p=2.25*10^{-308}$, males: beta=0.30, $p=1.58*10^{-40}$), BMI (females: beta=0.54, $p=2.25*10^{-308}$, males: beta=0.12, $p=1.16*10^{-37}$), and blood pressure (systolic blood pressure - females: beta=0.26, $p=1.72*10^{-196}$, males: beta=0.10, $p=1.08*10^{-24}$; diastolic blood pressure - females: beta=0.42, $p=2.25*10^{-308}$, males: beta=0.07, $p=9.91*10^{-12}$) showed stronger association to metabolic ageing in females, whereas type-II diabetes (females: beta=0.31, $p=2.05*10^{-03}$, males: beta=1.05, $p=3.02*10^{-35}$) had the highest effect estimate in males (***Fig 2***).

Lifestyle factors (poor self-rated health, sleep hours, alcohol frequency, and smoking packyears) were positively associated with similar effect sizes in both sexes.

Late puberty is associated with younger metabolic age

Next, we tested the association of metAgeGap with sex-specific factors. Late puberty (beta=-0.19, $p=3.77*10^{-110}$), the number of live births (beta=-0.16, $p=1.39*10^{-82}$), higher age at the first (beta=-0.13, $p=1.63*10^{-32}$) and last live births (beta=-0.18, $p=1.81*10^{-64}$) were associated with lower metabolic age in females. Interestingly age at menopause was not associated with metabolic aging (***Fig 3A***).

Balding was associated with higher metabolic age in males (beta=0.05, $p=3.05*10^{-02}$) while factors related to puberty like higher age at first facial hair appearance (beta=-0.22, $p=5.34*10^{-16}$) and voice break (beta= -0.23, $p=2.99*10^{-09}$), and number of children (beta=-0.10, $p=3.04*10^{-25}$) were associated with lower metabolic age (***Fig 3B***).

Genome-wide association analysis (GWAS) of metAgeGap

To gain genetic insights into the aetiology of biological aging I conducted a GWAS of metAgeGap (N: females=219,115, males=165,933). In females, I identified 23,756 genome-wide significant (p-value$<5*10^{-08}$) polymorphisms associated with metAgeGap mapped to 147 independent genomic loci harbouring 615 genes. In males, I identified 15,110 significantly associated polymorphisms with metAgeGap mapped to 120 independent genomic loci harbouring 510 genes. The corresponding Manhattan and QQ plots and associations are provided. 217 genes including the APOE gene that has been consistently associated with longevity in GWAS[135], overlapped between males and females **(*Fig s5-s8, Table s3, s4*)**. There were equally a large

number of genes that appeared to be specific to either sex. For instance, in females we identified several loci that were absent in males, e.g., locus on chromosome 1p13 including gene CELSR2; several independent loci on chromosome 2 harbouring genes ABCG5, ABCG8, LCT, MCM6, GLS, USP37 and STK25; several loci on chromosome 3 harbouring PPARG, DNAJC13, SLC2A2 and AHSG; locus 5q33 harbouring virus receptors HAVCR1 & HAVCR2; locus 6p21-22 harbouring CDKAL2 and VEGFA; loci on chromosome 7p harbouring DNAH11 and TMED4 and 7q33 harbouring STRA8; several independent loci on chromosome 8 harbouring genes MFHAS1, XKR6, SLC7A2, NAT2, TRIB1, PTK2 and VPS28; locus 9q34 harbouring SURF4 and SLC2A6; locus 10p15 harbouring AKR1CL1 and 10q25 harbouring TCF7L2 gene; locus 11q24 harbouring ST3GAL4; several loci on chromosome 12 harbouring genes TULP3, PHC1, SLC38A4 and TIMELESS; loci on chromosome 13 harbouring N4BP2L2 and GAS6; chromosome 14q harbouring MAP4K5 and ELMSAN1; loci on chromosome 17 harbouring GLTPD2 and SLC25A10; loci on chromosome 19p13 harbouring LDLR and SUPG1 genes; several loci on chromosome 20 harbouring RIN2, FAM182B, GDF5, PLCG1, HNF4A, and PLTP; and locus 22q13 that harbours CELSR1 and PPARA.

Similarly, there were several genomic regions that appeared associated with males only. These include several loci on chromosome 1 harbouring MIER1, IL6R, LGR6 and SARG genes; loci on chromosome 2 harbouring SH3YL1, GCKR and SPC25; chromosome 3p14 harbouring FRMD4B; chromosome 5 including FBXO4 and SNCAIP; locus 7q21 harbouring BRI3; loci on chromosome 8 harbouring RBPMS and TMEM70; loci on chromosome 10 harbouring FAM107B, MRC1L1, HKDC1, CYP26A1, PKD2L1 and NDUFB8; locus 11p15 harbouring SBF2; locus 14q32harbouringg SLC25A47; locus 15q15harbouringg MAP1A; loci on chromosome 16harbouringg BCAR1 and

GCSH; chromosome 17harbouringg SERPINF2, RAB5C and APOH; locus 18q21harbouringg ATP8B1; locus 19q13.11harbouringg HPN and locus 22q12harbouringg APOL3. As a sensitivity analysis, I used the Cochran's Q test to investigate the heterogeneity by sex. For 737 lead SNP in males and females, 722 showed significant heterogeneity between sexes after FDR correction. Top significant SNPs were mapped to BUD13, ZNF259, APOA, APOC, APOE, SLC22A2 and etc. Volcano plots showing the relationship between GWAS beta values and heterogeneity p values were shown in ***Fig s9a*** for females and ***Fig s9b*** for males.

Genes associated with metAgeGap in both males and females were expressed in the liver, digestive and endocrine glands and were enriched in cholesterol metabolism and signalling pathways among several others **(*Fig s10, s11*)**. Genes associated with metAgeGap in females only did not identify any specific pathway. Comparatively, genes associated with metAgeGap in males were expressed in diverse tissues including reproductive system, muscle and skeletal system (***Fig s12***) and were enriched in alcohol binding in addition to several others that were observed in the common pathways.

LD Score Regression (LDSC) analysis shows high correlation between male metAgeGap and female metAgeGap but with differences (r=0.71, $p=5.30*10^{-28}$). When comparing with publicly available GWAS, LDSC suggests significant genetic correlations of metAgeGap with Cystatin C (males: r=0.22, $p=3.23*10^{-04}$, females: r=0.17, $p=3.10*10^{-06}$), Hemoglobin A1c (HbA1c) (males: r=0.18, p=0.02, females: r=0.23, $p=6.01*10^{-08}$), estimated Glomerular Filtration Rate (eGFR) (males: r=-0.24, $p=9.44*10^{-04}$, females: r=-0.20, $p=1.08*10^{-07}$), and albumin (males: r=-0.49, $p=9.54*10^{-11}$, females: r=-0.24, $p=3.48*10^{-08}$) among others in both sexes. Mouth ulcers (r=-0.18, p=0.02), and atrial fibrillation (r=0.14, p=0.01) were found significant only in males while iron

deficiency-related anaemias (females: r=0.17, p=0.016), BMI (females: r=0.16, p=8.30*10^{-05}), type 2 diabetes (females: r=0.20, p=4.77*10^{-07}), ischemic heart disease (females: r=0.18, p=1.59*10^{-05}), and high cholesterol (females: r=0.56, p=1.00*10^{-28}) were found significant only in females (***Table s5, s6, Fig s15, s16***).

MetAgeGap and the risks of chronic diseases

When studying the association of metabolic ageing with 15 incident morbidities and mortality (***Fig 4A, B***), except for neurodegenerative disorders (Alzheimer's and Parkinson's), metAgeGap was significantly associated with all diseases in males. MetAgeGap was significantly associated with All-Cause Dementia (ACD) (HR=1.07, CI=1.02, 1.12), ischemic stroke (HR=1.14, CI=1.10, 1.18), macular degeneration (HR=1.05, CI=1.01, 1.08), all strokes (HR=1.11, CI=1.08, 1.15), rheumatoid arthritis (HR=1.08, CI=1.04, 1.12), and osteoporosis (HR=1.11, CI=1.07, 1.16) in males but not in females. The impact of metAgeGap on mortality and the risk of heart disease was far more pronounced in males. The HR for all-cause mortality was 8 times higher in males, and 2 times higher for ischemic heart**.** Effect estimates were overall stronger in males compared to females. Although remaining significant adjusting for BMI led to a higher dilution of effects in females but not in males especially for kidney, liver, and heart diseases, suggesting that BMI can have a different impact on ageing in males and females

Stark differences were observed between males and females for cancers. MetAgeGap was significantly associated with liver (HR=1.74, CI=1.52, 1.99), oesophageal (HR=1.30, CI=1.19, 1.42) and lung cancers (HR=1.10, CI=1.04, 1.17) in males but not in females (***Fig 4C, D***). The hazard ratios for liver and oesophageal cancers in males were particularly high, indicating a

strong link with metabolic ageing. Further, adjusting for alcohol use and smoking did not attenuate these associations (***Fig s17-s20***). Colorectal cancer was significantly associated with metAgeGap in both sexes. In females after adjusting for BMI and physical activity, metAgeGap's association with kidney (HR=1.11, CI=0.99,1.24) and breast (HR=1.02, CI=1.01,1.04) cancers became non-significant, suggesting that these factors are potentially confounding the association between metAgeGap and these cancers in females (***Fig 4C, D, Table s7-s10***).

When comparing the disease risks of the biologically youngest group (bottom 25% metAgeGap) with that of the biologically oldest group (top 25% metAgeGap), youngest males had a higher cumulative risk for Ischemic Heart Disease (IHD) than the oldest females, reflecting a higher baseline risk in males **(*Fig 5*)**. Further, the differences in cumulative risks of COPD and all-cause mortality (ACM) between the top and bottom 25% were more pronounced in males than in females. The increase in cumulative risk for osteoarthritis was more gradual and linear in both sexes, with females being more at risk than males after the age of 50, suggesting a different risk trajectory in both sexes. For chronic kidney and liver diseases, no differences were observed between males and females where biologically younger individuals showed significantly lower cumulative risks compared to biologically older individuals (***Fig 5*).** Detailed cumulative risks were shown in ***supplementary tables s11-s14***.

Discussion

In this study, I developed and validated sex-specific metabolomic ageing clocks using NMR-based metabolomic data from the UK Biobank, enabling me to investigate biological ageing through a metabolic lens in a population of over 380,000 individuals. Our findings demonstrate

that metabolic ageing, as captured by the metAgeGap metric, is a complex and sex-differentiated process with distinct biochemical, phenotypic, and genetic correlates.

First, the fact that our models could predict chronological age with moderate accuracy (R^2 = 0.29 for males, 0.37 for females) suggests that the metabolome holds reliable, albeit incomplete, clues about the biological clock. Interestingly, females consistently showed stronger prediction performance—a possible reflection of the tighter regulation or narrower variability of metabolic processes in women. While 61 features were common to both sexes—suggesting conserved core pathways—the presence of 32 female-specific and 15 male-specific features points to divergent metabolic processes driving ageing. In females, unique contributions from cholines and phosphatidylcholines align with known roles in membrane biology and lipid signaling[136,137], whereas males exhibited distinct associations with HDL particle size and triglyceride composition[138,139], implicating differences in lipid transport and cardiovascular risk.

The metAgeGap metric showed strong and consistent associations with cardiometabolic risk factors, but the magnitude and pattern of these associations were sex-dependent. Obesity, BMI, and blood pressure exhibited stronger associations in females, while type 2 diabetes showed the largest effect in males. These differences underscore that while the roads to ageing may be parallel, they aren't identical and that sex should be considered when examining the metabolic determinants of biological age and their relevance to disease risk.

Perhaps the most intriguing finding of the current study is the association of puberty and hormonal exposures with metabolic ageing. In women, later puberty, higher parity, later age at first and last live births and hormone use painted a portrait of youth preserved, with

significantly younger metabolic ages. Meanwhile, in men, delayed signs of puberty—like voice breaking or facial hair— and the number of children fathers were similarly linked to a slower metabolic clock. These findings suggest that earlier exposure to sex hormones may accelerate biological ageing in both sexes, consistent with hypotheses proposing a trade-off between early reproductive maturation and long-term somatic maintenance. These findings align with previous evidence suggesting that hormonal and physiological factors related to reproduction can modulate long-term metabolic health and ageing trajectories[140,141]. However, our findings of protective effects of parity in both females and males contradict the evolutionary theories of ageing which hypothesize a trade-off between reproduction and lifespan[55,142–144]. Literature shows conflicting results about parity and longevity in women[55,144] but somewhat consistent in males[145]. Our findings appear more plausible with the hypothesized effects of fetal microchimerism on maternal health and longevity[146]. Further later age at birth has also been found to be associated with longevity in females[145].

Crucially, metAgeGap was significantly associated with incident morbidity and mortality outcomes, again revealing pronounced sex differences. Males with higher metabolic age showed elevated risks across nearly all disease categories, including cardiovascular, renal, hepatic, and oncological outcomes. Notably, liver, oesophagal, and lung cancers exhibited strong associations in males, independent of lifestyle factors such as alcohol use and smoking. In contrast, females displayed fewer and generally weaker associations, although osteoarthritis risk increased significantly after midlife, and some disease risks (e.g., breast and kidney cancer) were attenuated after adjustment for BMI and physical activity. The attenuating effect of BMI was consistently more pronounced in females, suggesting that the interplay between adiposity

and metabolic ageing may differ by sex. These findings are consistent with earlier studies which showed a stronger association between BMI and mortality in females than in males[147]. The sex difference can be explained by the fact that females tend to accumulate more subcutaneous fat, while men store more visceral fat[148–150], which is more closely linked to metabolic risks, highlighting the importance of developing sex-specific clocks. The female body seemed to carry a certain metabolic resilience, especially when it came to cancers and neurodegenerative diseases.

All-cause mortality and IHD were markedly associated with higher metabolic age in males, with effect estimates significantly exceeding those observed in females. This finding is consistent with the observation that IHD more often affects males than females, and its incidence in males typically starts in the 40s and increases with age[151]. Males in the bottom decile of metAgeGap had a higher risk of mortality than all females up to the age of 72 years. I hypothesize that this might be the result of the metabolic panel we use which is focused on lipid metabolism. Past research has indicated that males with hyperlipidemia are diagnosed earlier than females, suggesting a heightened vulnerability in males to lipid-related disorders, particularly before menopause[152]. This earlier diagnosis in males could explain the stronger association between metAgeGap and mortality risk in this group, while females' metabolic profiles may be influenced by hormonal factors that mitigate these risks until later in life. These differences may be due to lifestyle differences observed in our dataset, with higher rates of alcohol consumption and smoking among males than females. These habits are well-established risk factors for liver, lung, and oesophageal cancers, and they also contribute to cognitive decline through mechanisms like neuroinflammation, vascular damage, and oxidative stress.

In the GWAS, 217 genes overlapped between males and females indicating a shared genetic basis of metabolic aging bringing out cholesterol metabolism and PPAR signalling as common pathways of metabolic aging. These included the APOE gene, which is one of the two genes consistently associated with longevity in large-scale GWAS[153]. Of note is that a large number of genomic loci (147 vs 120) were identified in females suggesting potentially higher polygenic complexity. Despite a larger gene set, no unique pathway was observed for female-specific genes beyond those shared. There were, however, some interesting genes including those involved in glucose transport, e.g., SLC2A2 and SLC2A6, lactose intolerance and gut microbiome composition LCT, circadian rhythm TIMELESS, virus receptors HAVCR1 and HAVCR2, and genes involved in metabolic and brain health CELSR2 and CELSR1 that showed no association with metabolic ageing in males. In contrast to females, male-specific genes were enriched in alcohol binding pathways, suggesting that lipid-related proteins in males may interact more strongly with alcohol or alcohol-derived compounds, potentially due to higher alcohol consumption patterns in males. Further, male-specific loci included genes such as IL6R, GCKR, SNCAIP, HKDC1 and APOH some of which are implicated in inflammation and metabolism. Sex-specific genetic correlations were found with iron-deficiency anaemia, BMI, type 2 diabetes, ischemic heart disease, and high cholesterol, suggesting that metabolic ageing in females may be more closely tied to cardiometabolic health, while genetic correlations with mouth ulcers and atrial fibrillation were observed only in males, suggesting a distinct health profile associated with male metabolic ageing.

This study has several strengths, including the use of the largest available metabolites dataset and its contributions of new insights into sex-specific differences in metAgeGap. However, there

are some limitations. Although the model was validated well in the random split testing dataset, the findings were not externally validated. The accuracy of the model and the impact of metAgeGap on disease risks in other ethnic groups are in plan to be validated in future projects. Additionally, the dataset predominantly captures lipid-related metabolites, which may limit the scope of targeting clocks. While lipids are critical in metAgeGap, other metabolic pathways and molecules not represented in this dataset may also play important roles in ageing. Another limitation was that the GWAS results showed mild inflation, with LDSC intercepts of 1.22 in females and 1.17 in males. However, the analysis was performed using SAIGE, applying a linear mixed model with a genetic relationship matrix (GRM) and principal components to account for population stratification and relatedness. Comparable levels of inflation have been reported in large-scale GWAS, such as those of height (~5 million participants, LDSC intercept = 1.48)[154] and educational attainment (~3 million participants, LDSC intercept = 1.66)[155]. Previous studies have attributed such inflation primarily to large sample sizes and high trait heritability[156]. Importantly, the significant loci identified in our GWAS align with biologically plausible pathways, supporting the robustness and validity of our findings.

In conclusion, although ageing affects everyone, its pace and impact differ between males and females due to genetic, hormonal, or environmental factors. The need to understand these differences extends beyond scientific curiosity, impacting real-world healthcare practices and outcomes. For example, cardiovascular disease is often underdiagnosed or misdiagnosed in women because they frequently experience atypical symptoms, such as nausea or fatigue, rather than the more well-known chest pain seen in males[157,158]. Our study demonstrated that sex-specific metabolomics aging clocks are powerful tools for measuring biological age and

capturing aging signatures that are linked to common age-related diseases in both males and females. Our findings highlighted the potential of these clocks to identify the biological mechanisms underlying common diseases and emphasized the importance of sex differences in ageing processes. These clocks can shape our knowledge of sex-specific disease susceptibility, rates of physiological decline, and overall longevity, paving the path for more personalised prevention, treatment strategies and refined public health policies.

Figures

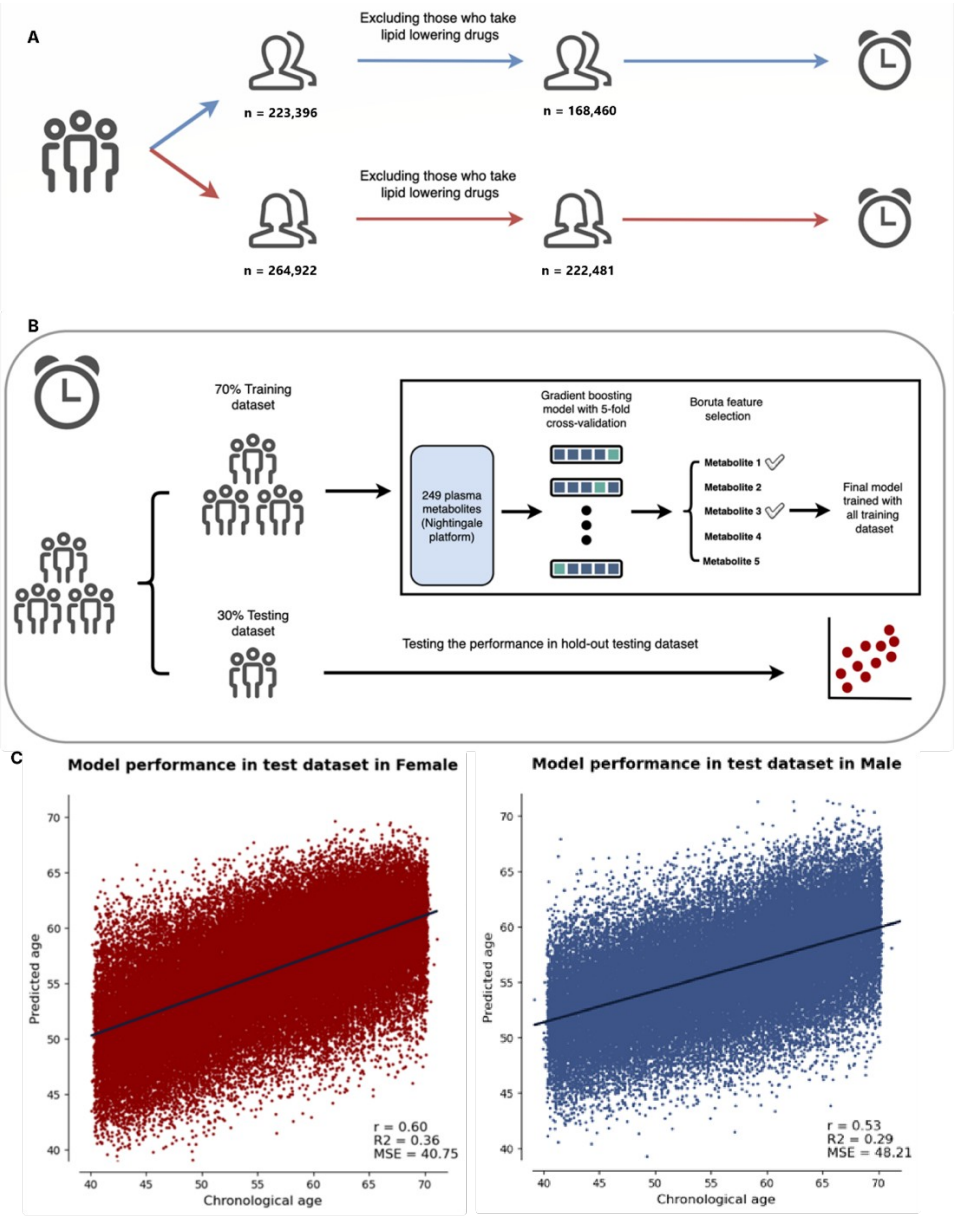


Figure 1: Overview of the study and model performance. (A) A metabolome-based age clock was built for each sex using gradient boosting. Criteria used to select participants for the study. **(B)** The classification model was trained on 70% randomly selected men/females and tested on the remaining data. Boruta feature selection algorithm was then used to select only relevant features for downstream analysis. **(C)** Model performance in predicting the age of participants in the test set in the model trained using 93, and 76 metabolites selected by Boruta in females and males respectively.

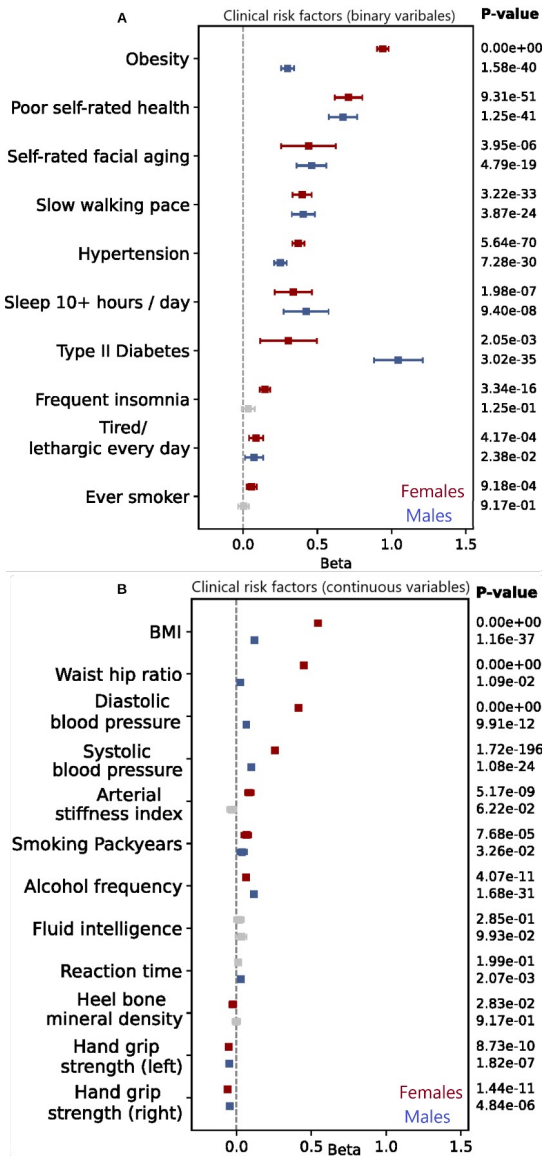


Figure 2: Association of age-related clinical risk factors with metAgeGap in females (red) and males (blue). Linear regressions were performed between each exposure and metAgeGap adjusting for recruitment centre, ethnicity, education years, and Townsend deprivation index. Values were standardized if quantitative. The grey colour showed if the association was not significant after FDR correction. Males and females have different associations in particular obesity, BMI and waist-to-hip ratio are more associated with metAgeGap in females than males. (A) Binary variables. (B) Continuous variables. The plots display per standard deviation change for easy comparison.

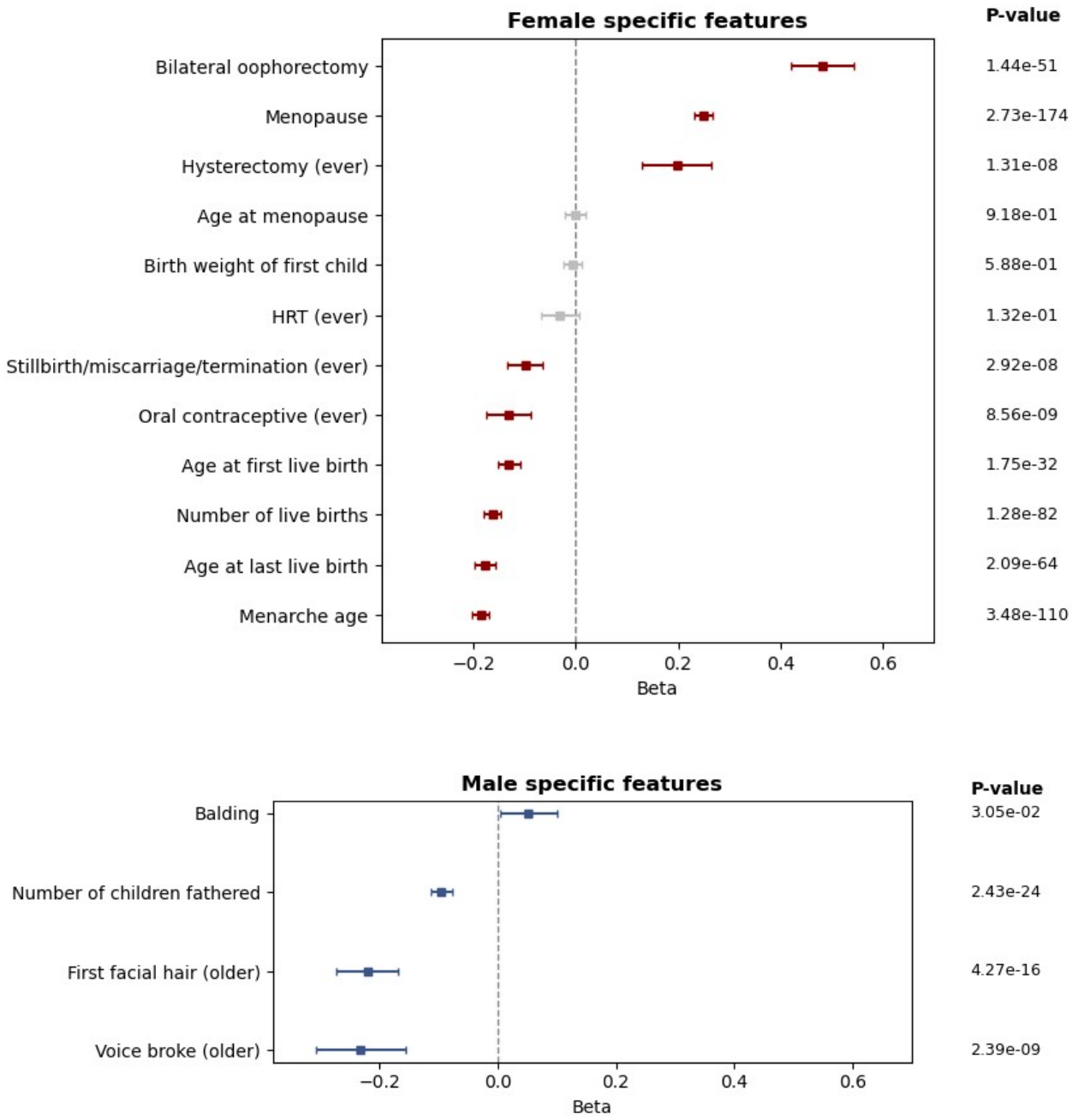


Figure 3: Sex-specific factors' association with metAgeGap. Linear regressions were performed between each exposure and metAgeGap adjusting for recruitment centre, ethnicity, education years, and Townsend deprivation index. Values were standardized if quantitative. The grey colour showed if the association was not significant after FDR correction. Balding in males and menopause in females is associated with higher metAgeGap.

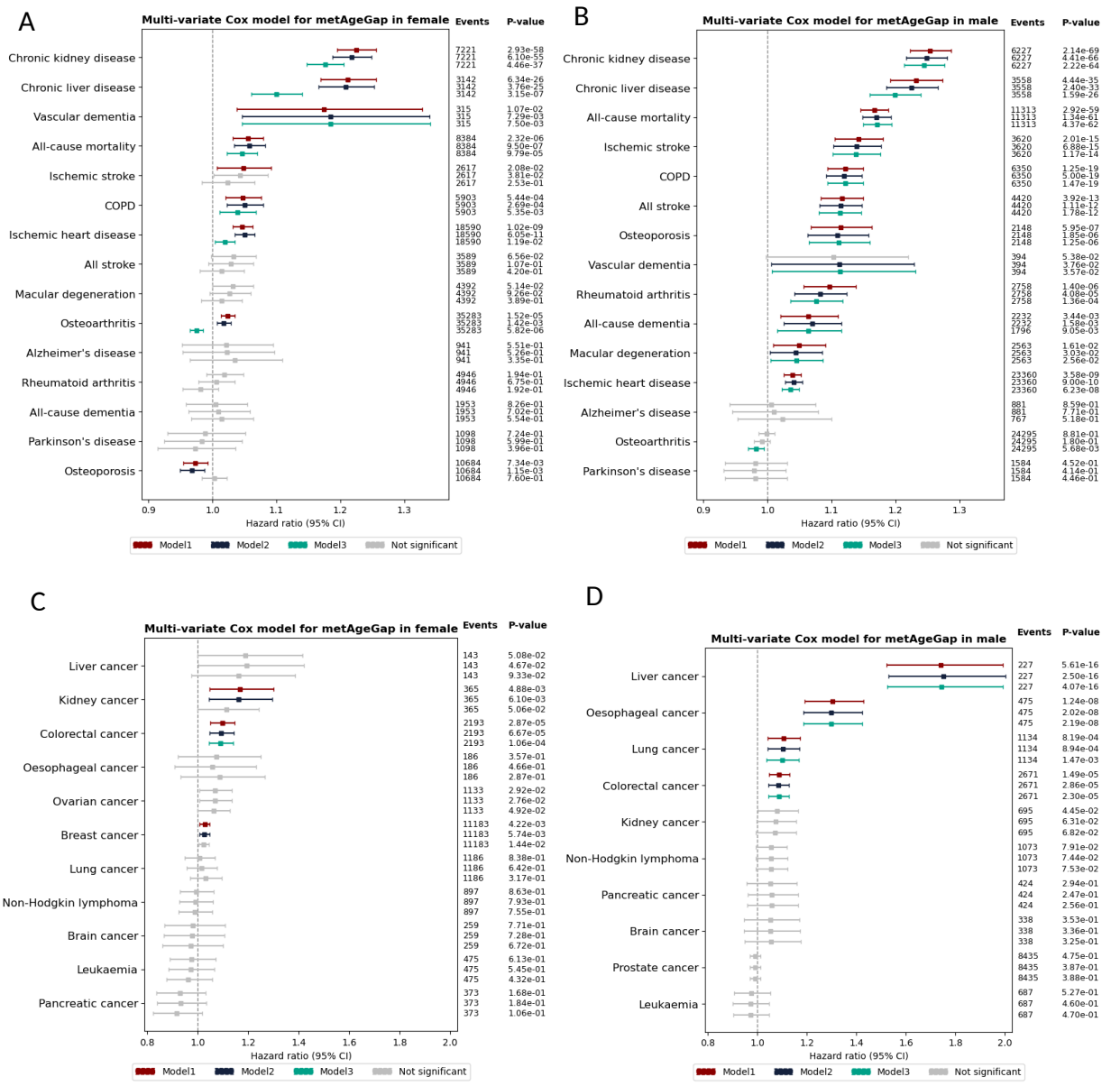


Figure 4: Association of metAgeGap to risk of morbidities (including sex-specific cancers) and mortality. (A) shows the association between metAgeGap and common diseases in females using a multivariate Cox proportional hazard model. In model 1, the exposure was metAgeGap with adjustment of chronological age. Model 2 was further adjusted for recruitment centre, Townsend deprivation index, and ethnicity. Model 3 was further adjusted for physical activity and BMI. P-values (p-val) were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour. **(B)** shows the association between metAgeGap and common diseases in males using a multivariate Cox proportional hazard model. **(C)** shows the association between metAgeGap and cancers in females using a multivariate Cox proportional hazard model. **(D)** shows the association between metAgeGap and cancers in males using a multivariate Cox proportional hazard model. P-values and number of cases for each model are shown in **Supplementary Tables 3-6**.

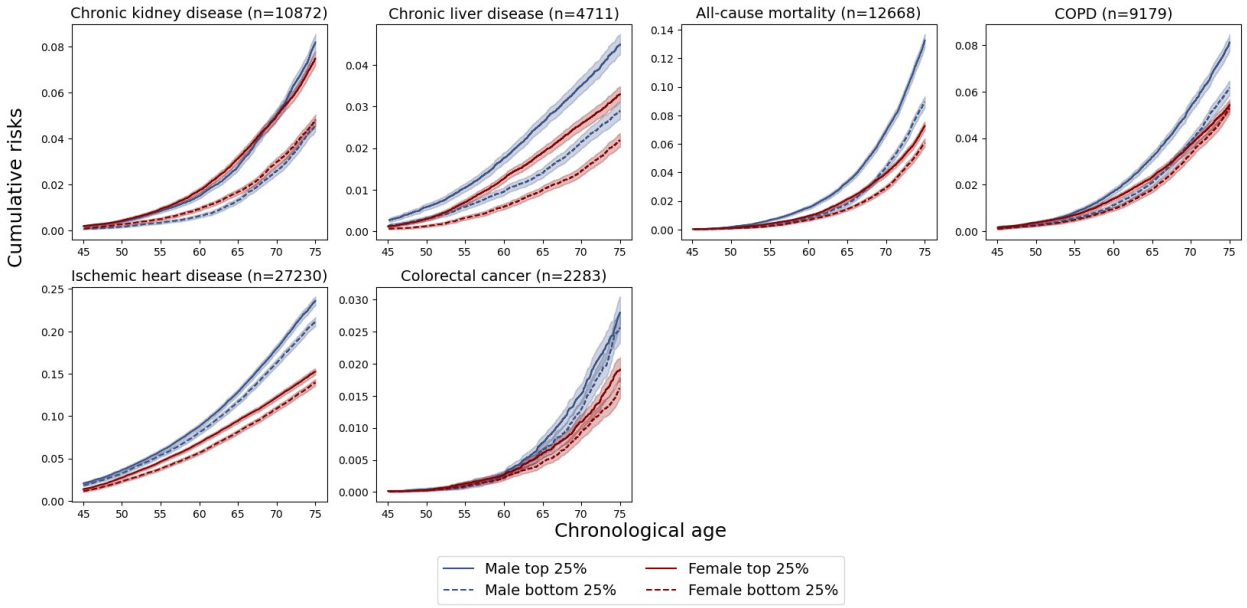
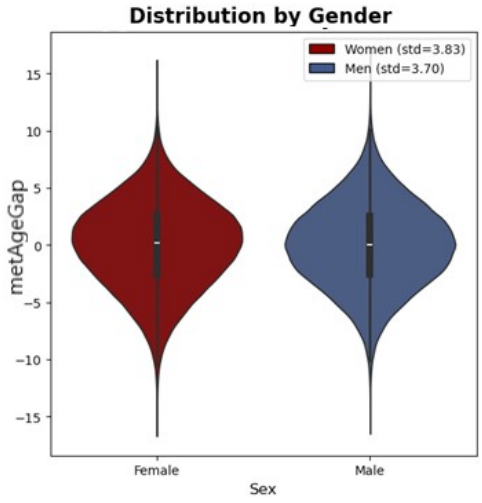
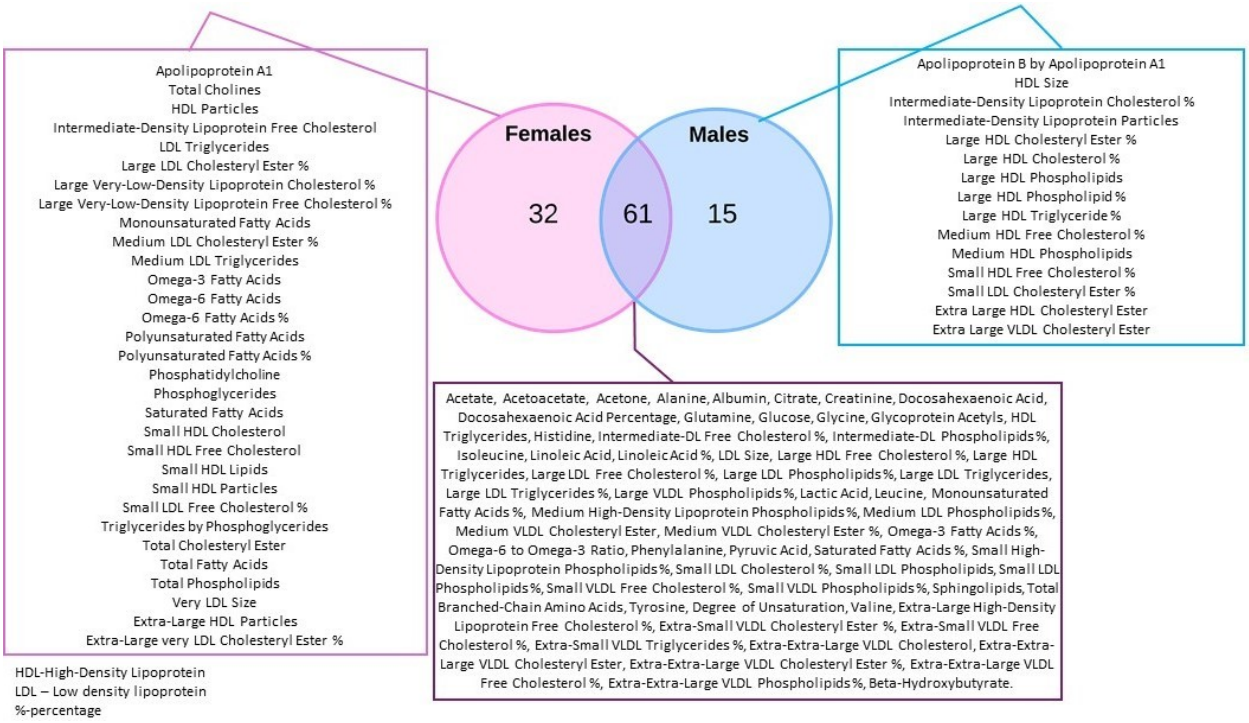


Figure 5: Deciles of metAgeGap lead to strongly diverging cumulative incidence of major incident diseases and mortality. Cumulative incidence plot of top and bottom 25% of the metAgeGap in males and females with 95% confidence interval shown as lighter shading. The X-axis denotes the chronological age and the Y-axis denotes cumulative risk. Cumulative incidence and number at risk at each age point are shown.

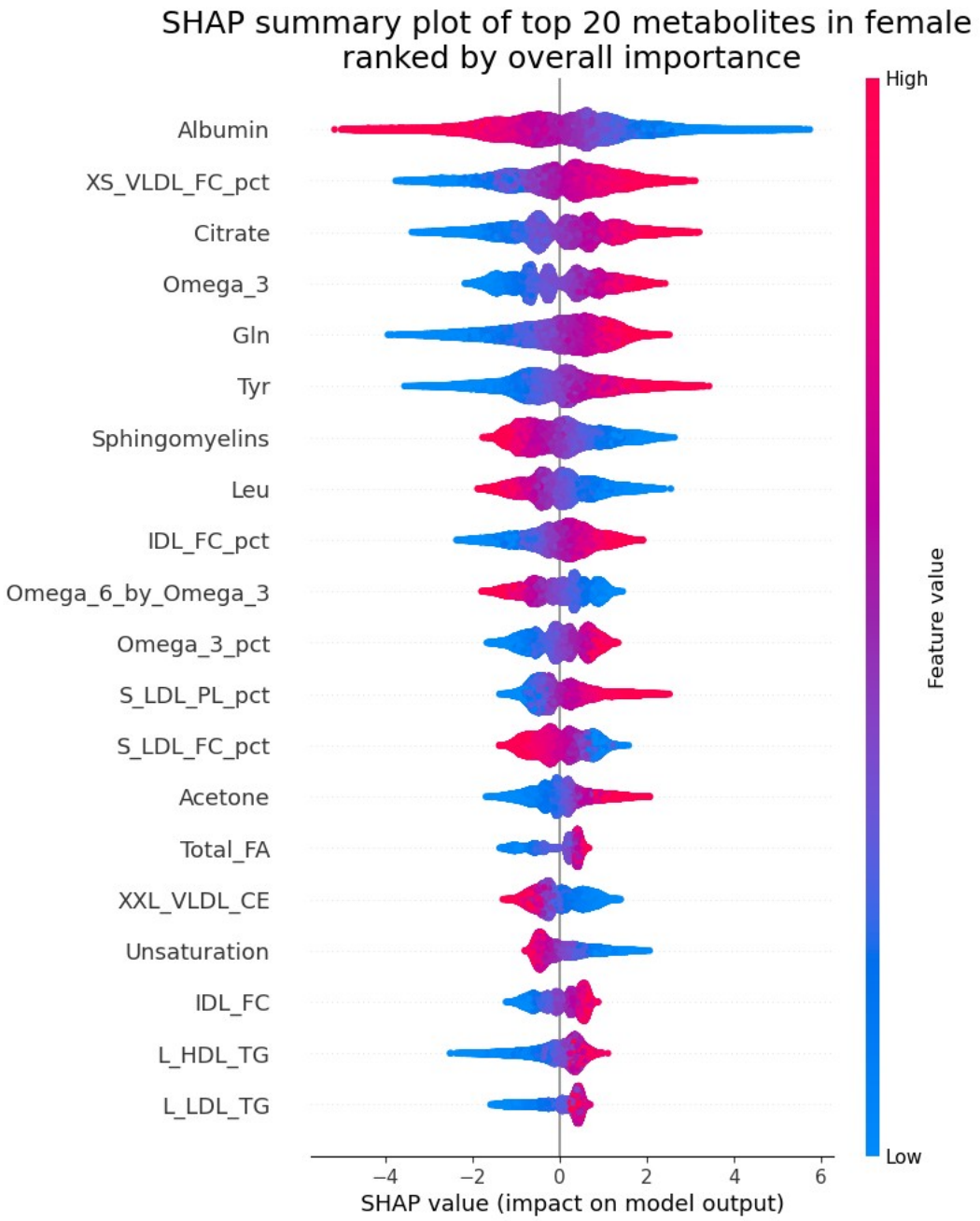
Supplementary figures



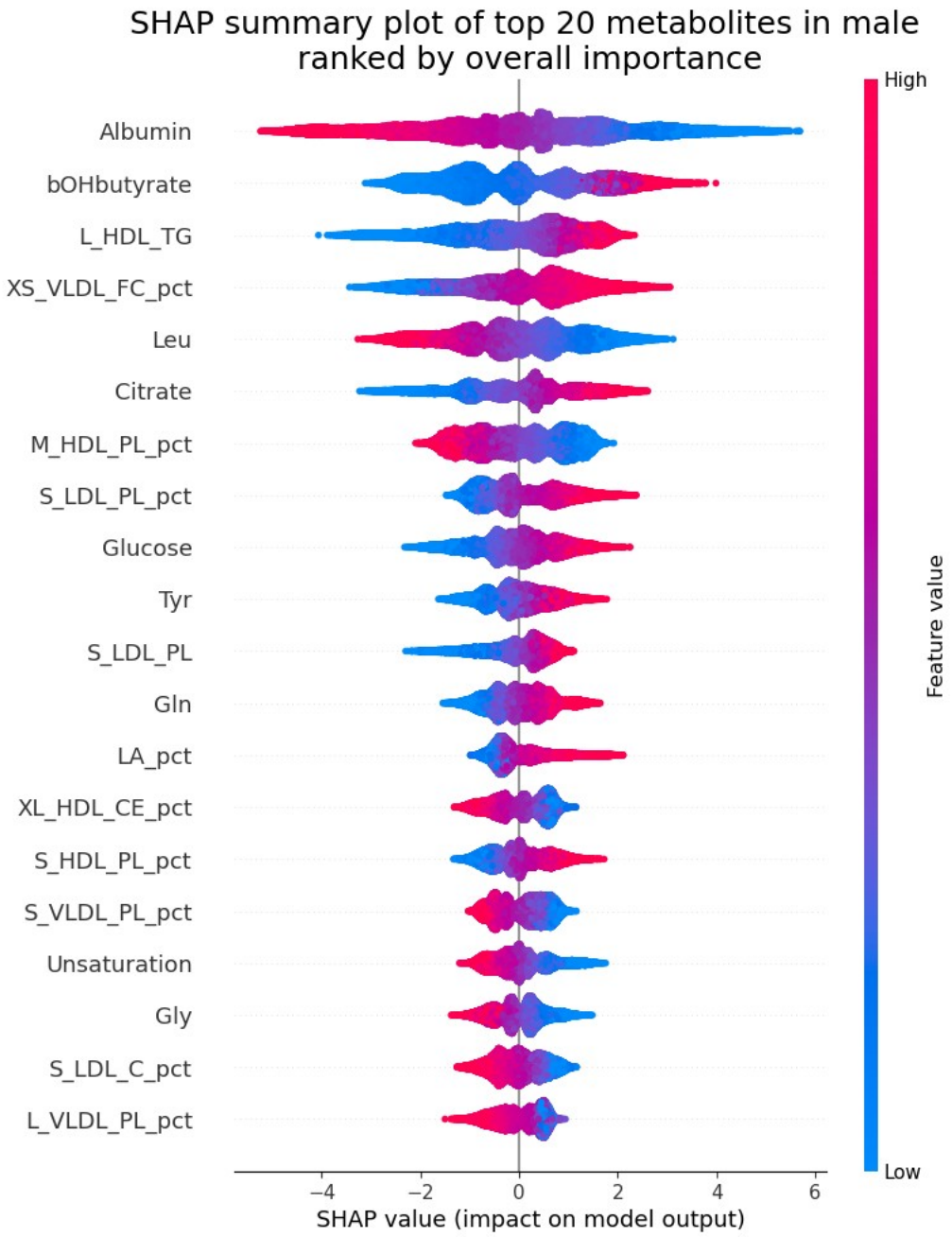
Supplementary Figure 1: Distribution of residual between chronological age and predicted metabolic age - metabolic age gap (metAgeGap) in males and females.



Supplementary Figure 2: An overview of the metabolites selected by the Boruta algorithm. 61 metabolites are common between males and females, while 32 and 15 are uniquely selected in females and males respectively.

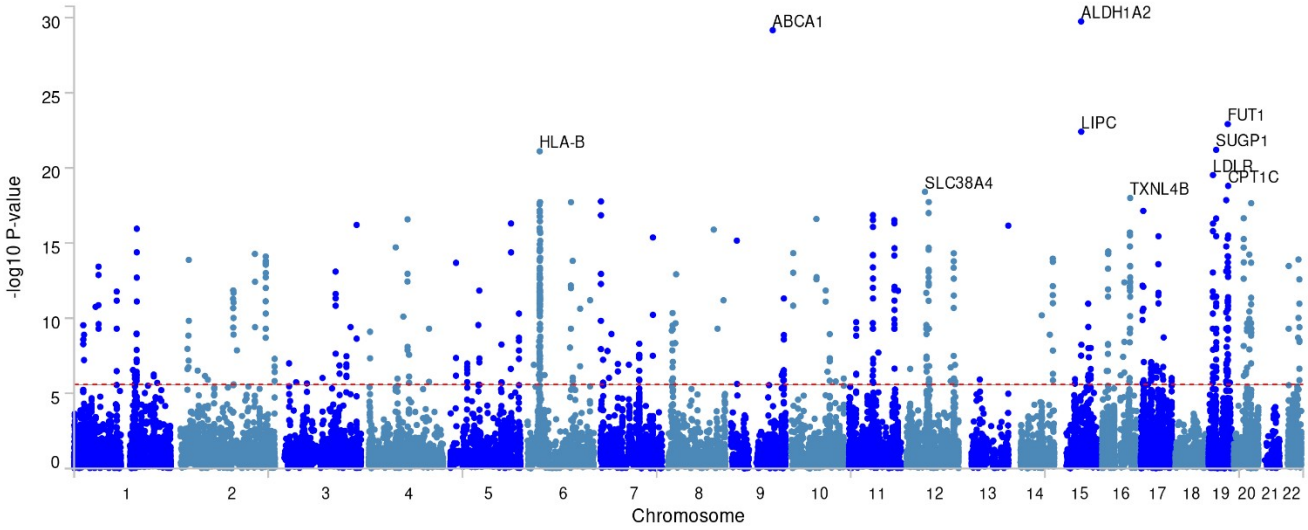


Supplementary Figure 3: Different metabolites are deemed important by the model in males and females. SHAP values of the top 20 selected metabolites in females. Each dot denotes a participant, the colour of the dots denotes the metabolome expression level and the X-axis denotes its contribution to the model decision.



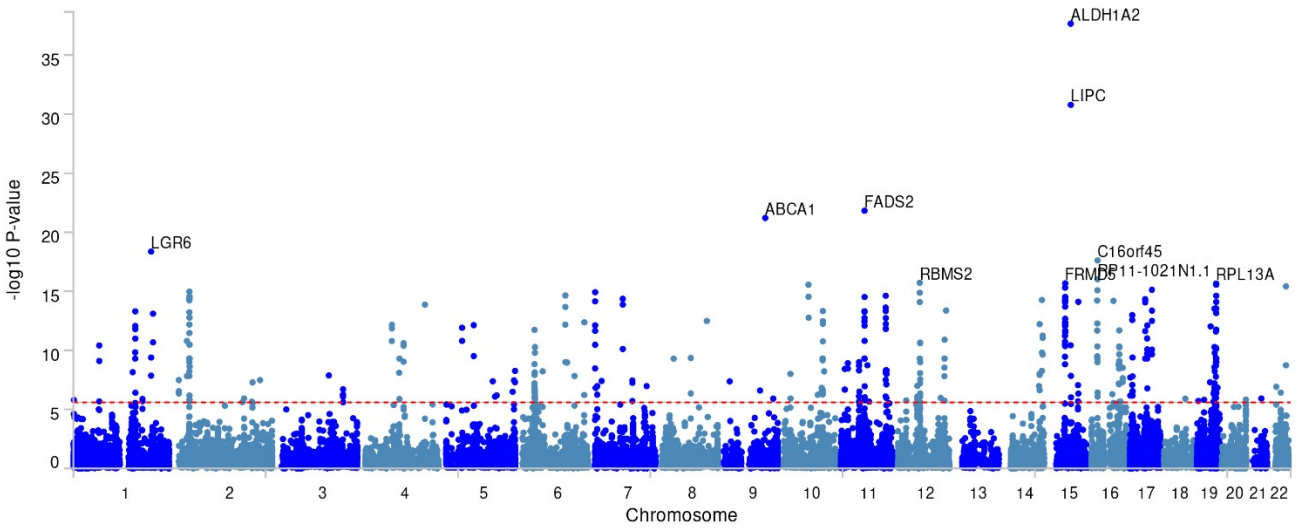
Supplementary Figure 4: Different metabolites are deemed important by the model in males and females. SHAP values of the top 20 selected metabolites in males. Each dot denotes a participant, the colour of the dots denotes the metabolome expression level and the X-axis denotes its contribution to the model decision.

Supplementary Figure 5: QQ plot for females. LDSC analysis showed an observed h2 of 0.13 (std=0.021), Lambda GC of 1.46, LDSC intercept of 1.22 (std=0.014) and an attenuation ratio of 0.27 (std=0.017), suggesting mild but acceptable inflation.

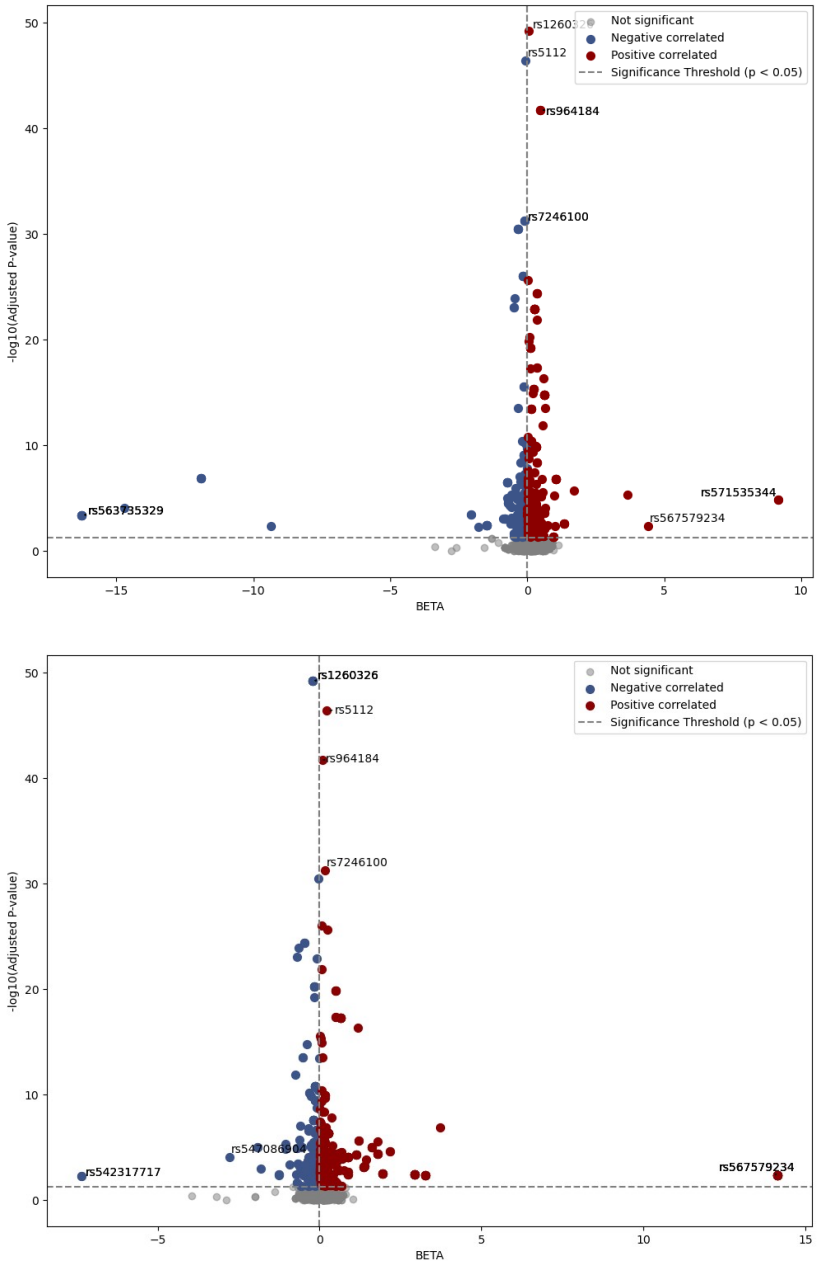


Supplementary Figure 6: The Manhattan plot showing top genes associated with metAgeGap for females.

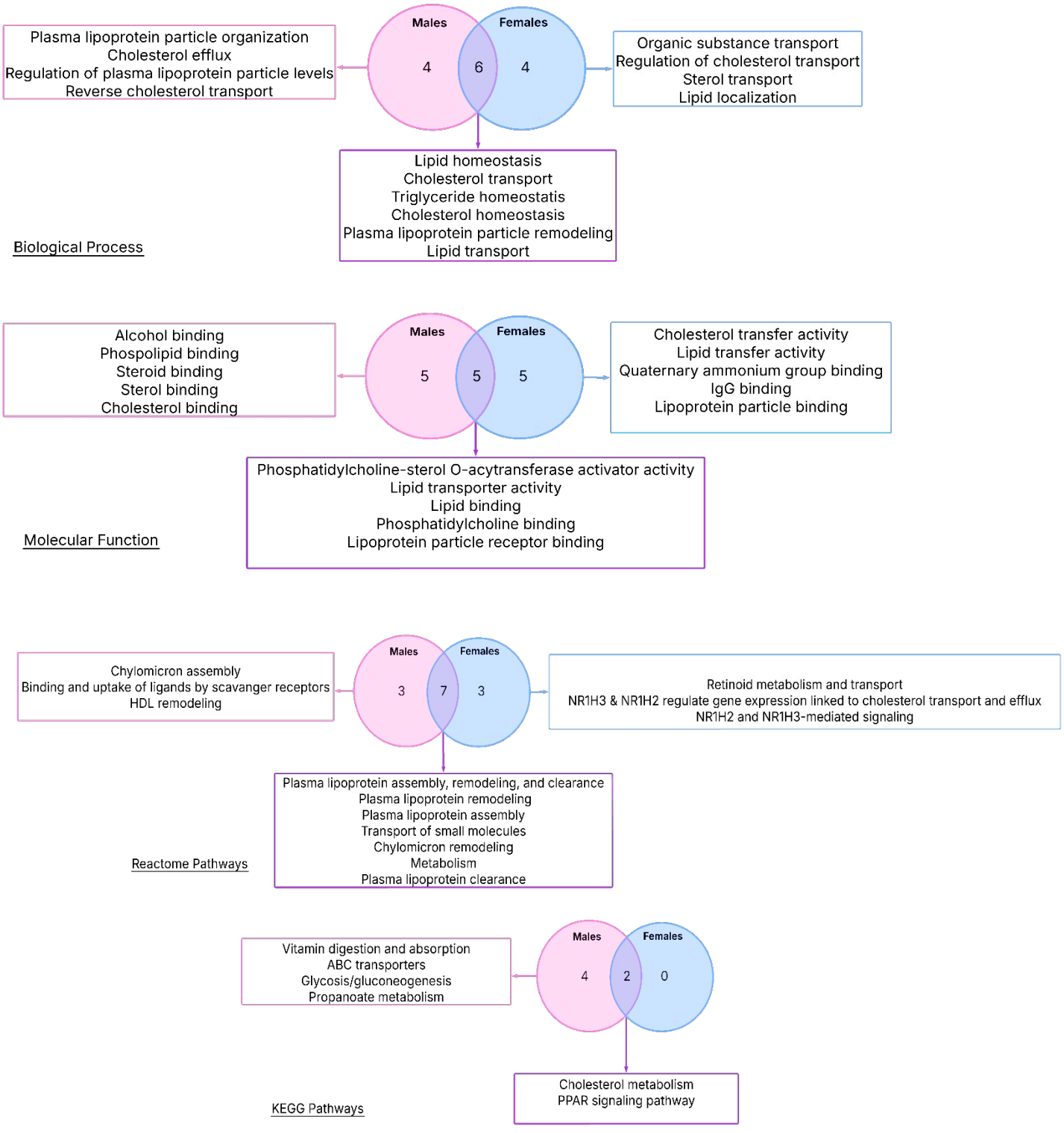
Supplementary Figure 7: QQ plot for males. LDSC analysis showed an observed h2 of 0.14 (std=0.027), Lambda GC of 1.36, LDSC intercept of 1.17 (std=0.013) and an attenuation ratio of 0.28 (std=0.021), suggesting mild but acceptable inflation.



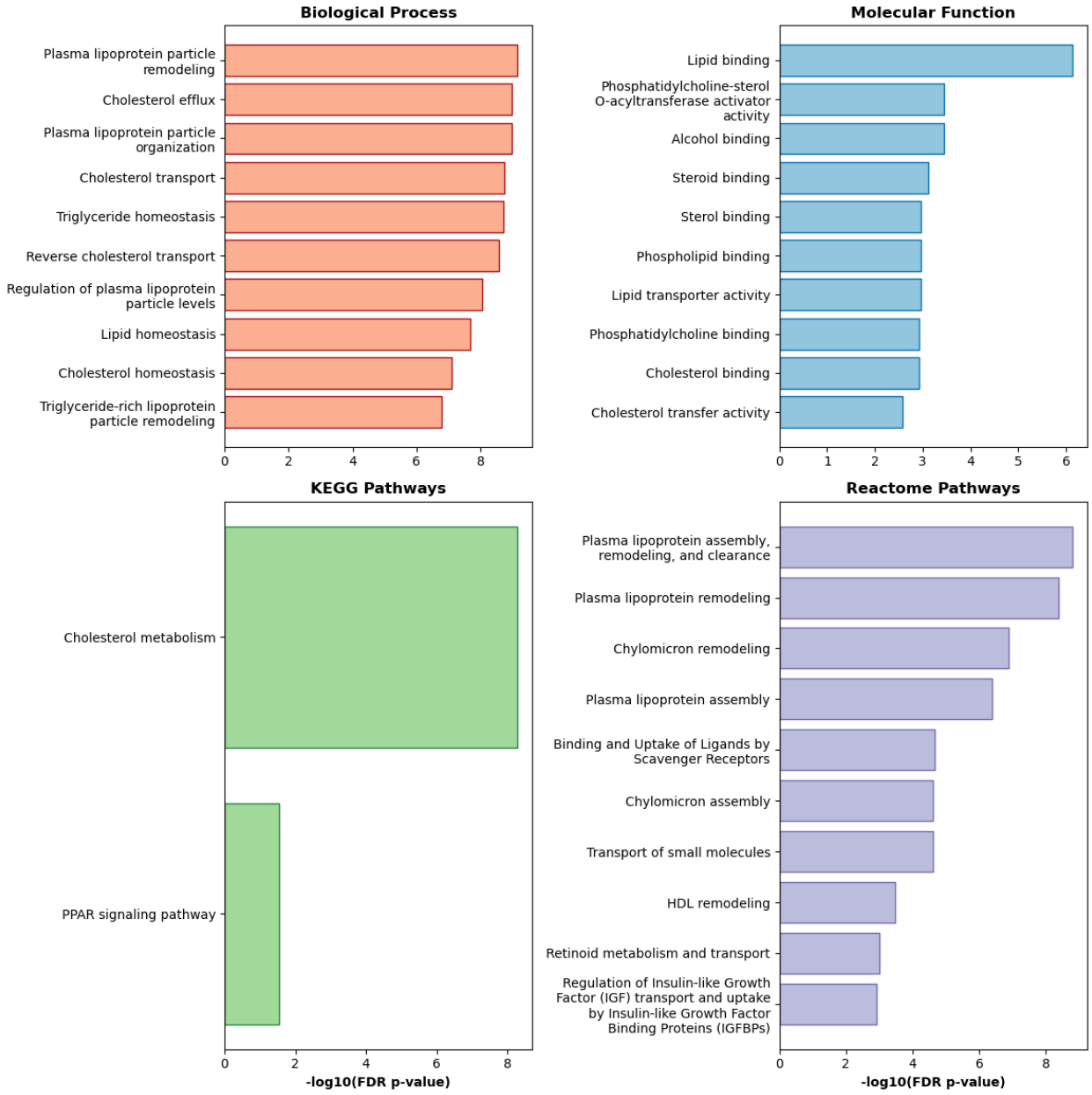
Supplementary Figure 8: The Manhattan plot showing top genes associated with metAgeGap for males.



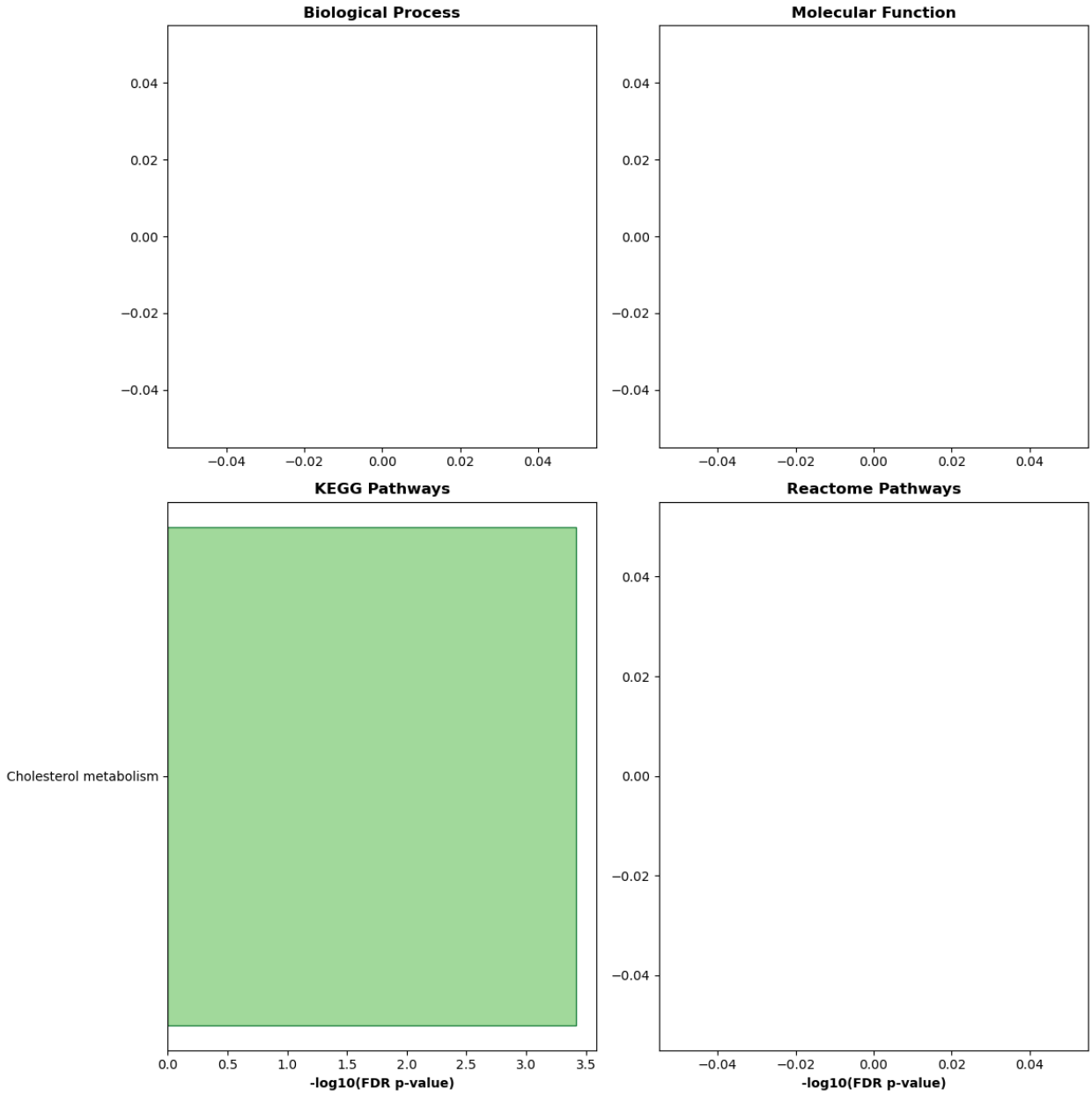
Supplementary Figure 9: Volcano plot of GWAS beta values and heterogeneity test p values. Cochran's Q test was performed for all lead SNPs in males and females. (a) shows the relationship between beta value and p value in females. (b) shows the relationship between beta value and p value in males. P values were FDR corrected.



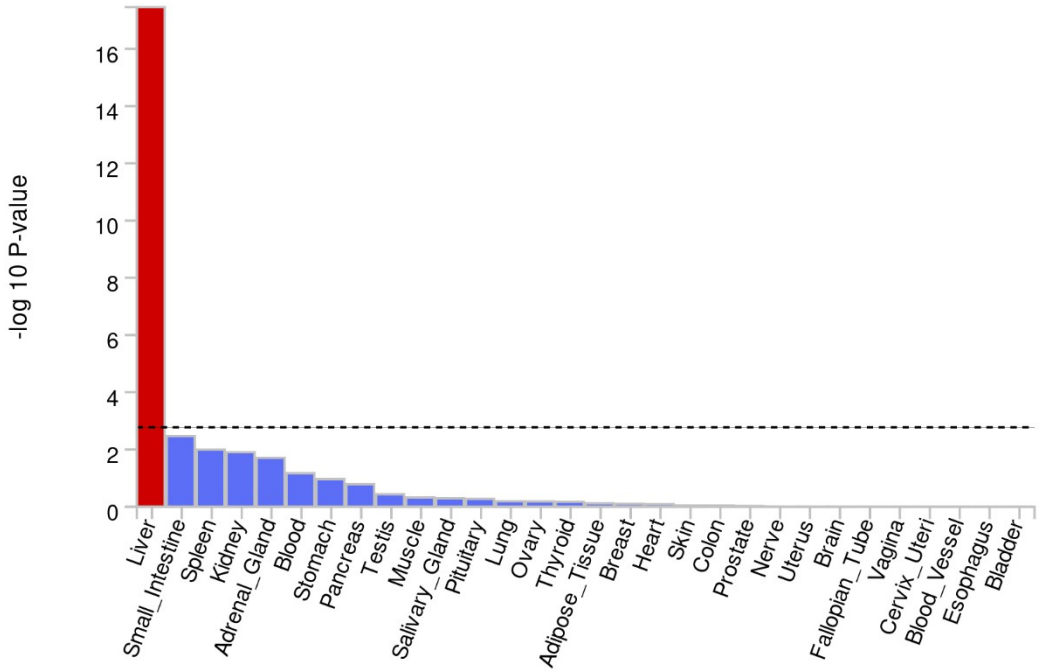
Supplementary Figure 10: Pathway enrichment results in both sexes.



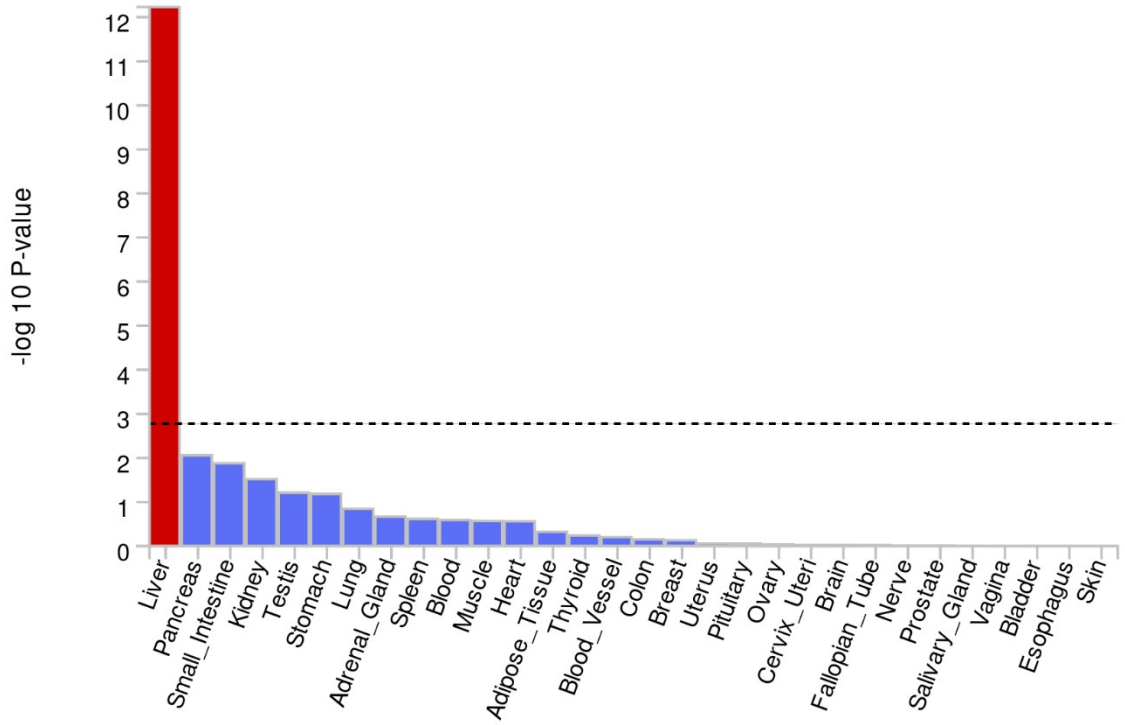
Supplementary Figure 11: Pathway enrichment results in males.



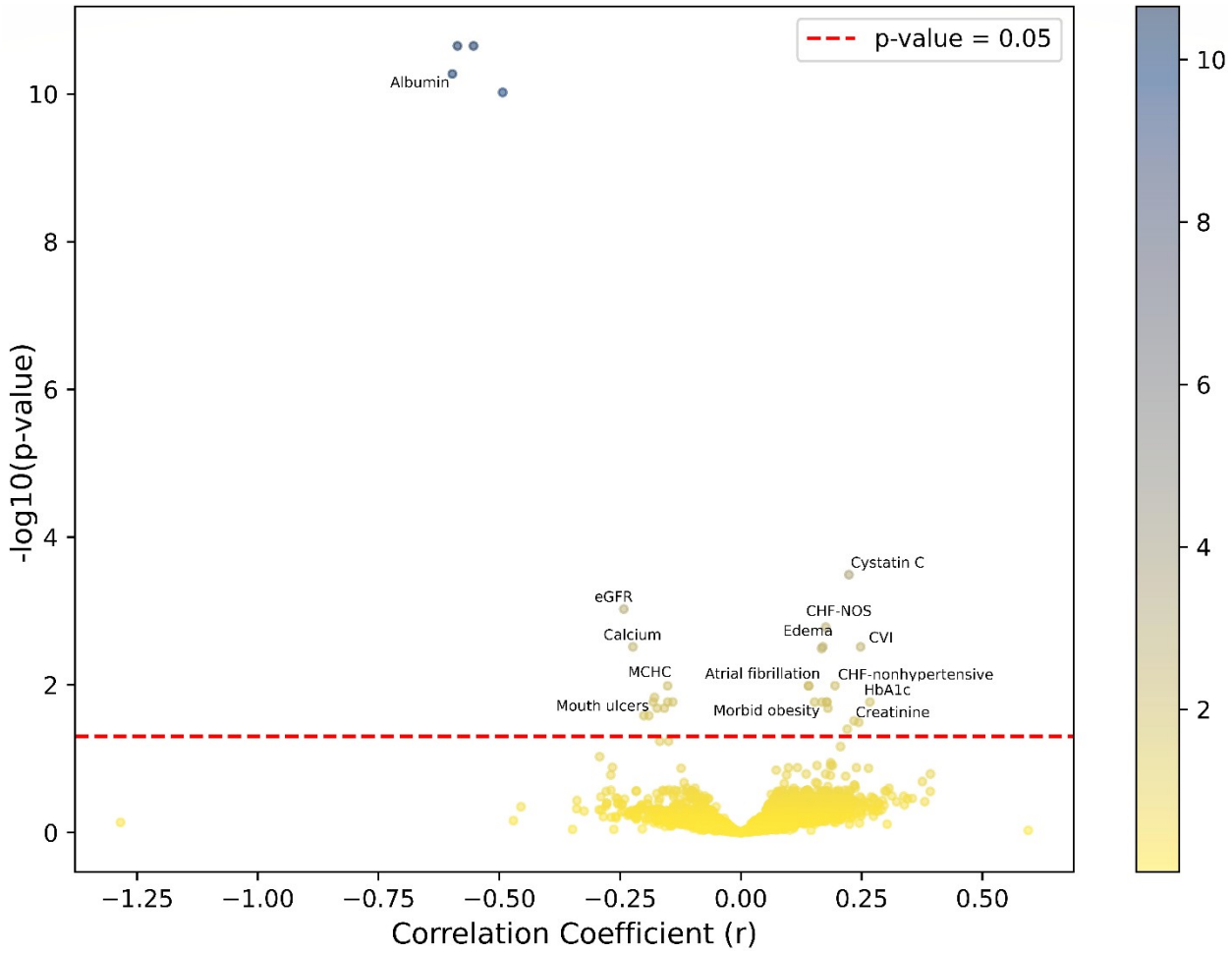
Supplementary Figure 12: Pathway enrichment results in females.



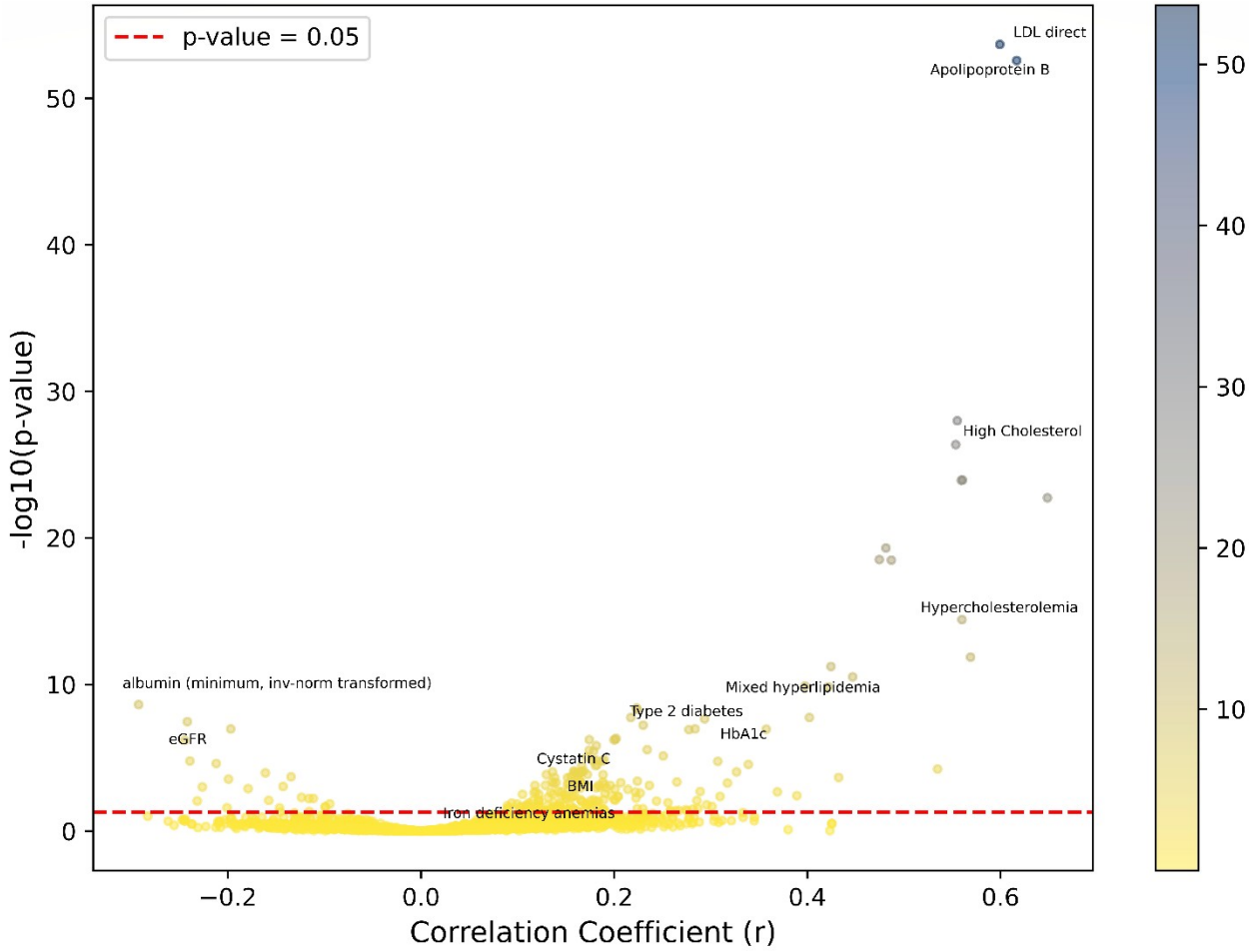
Supplementary Figure 13: Differentially Expressed Genes (DEGs) in females.



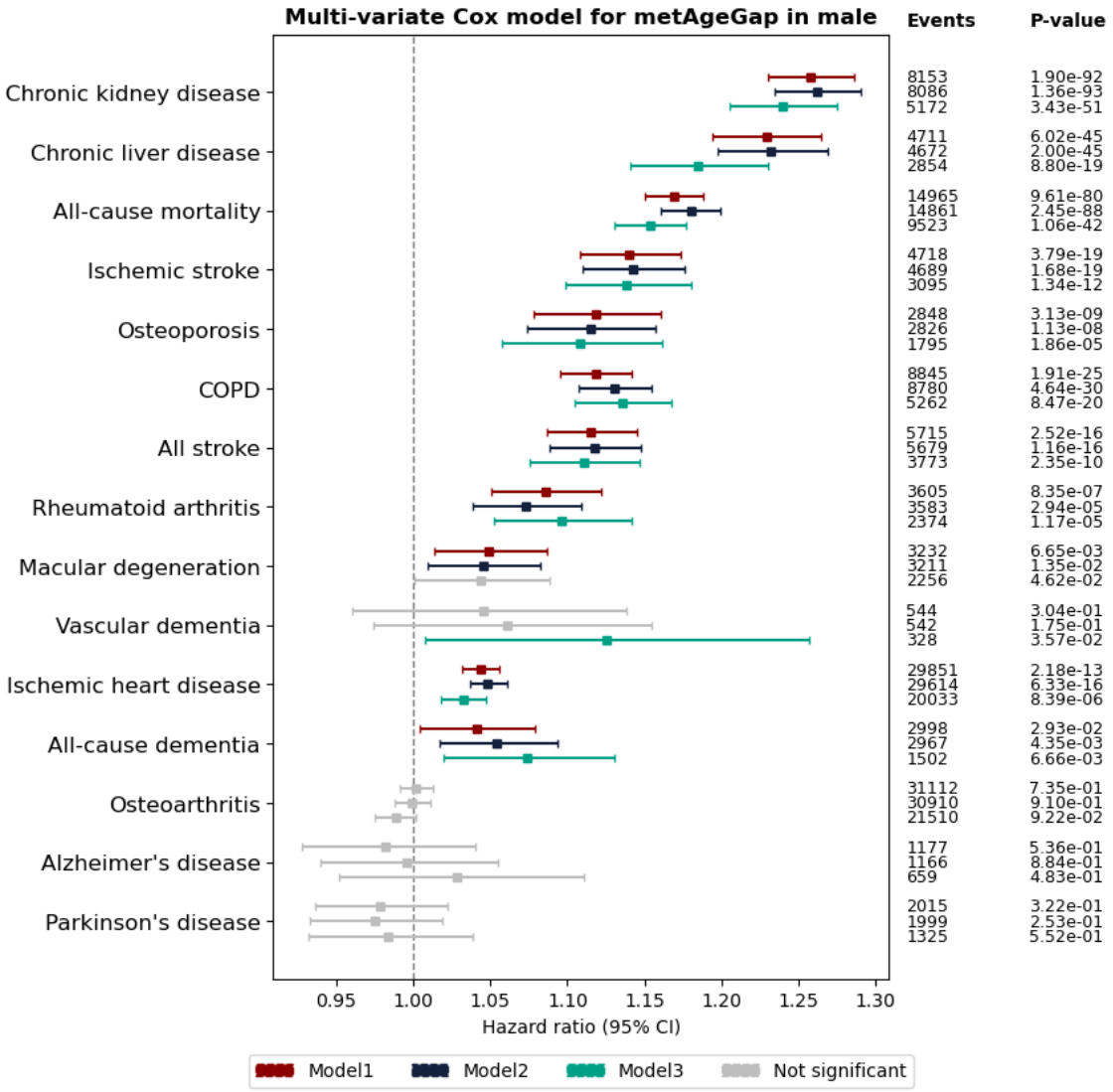
Supplementary Figure 14: Differentially Expressed Genes (DEGs) in males.



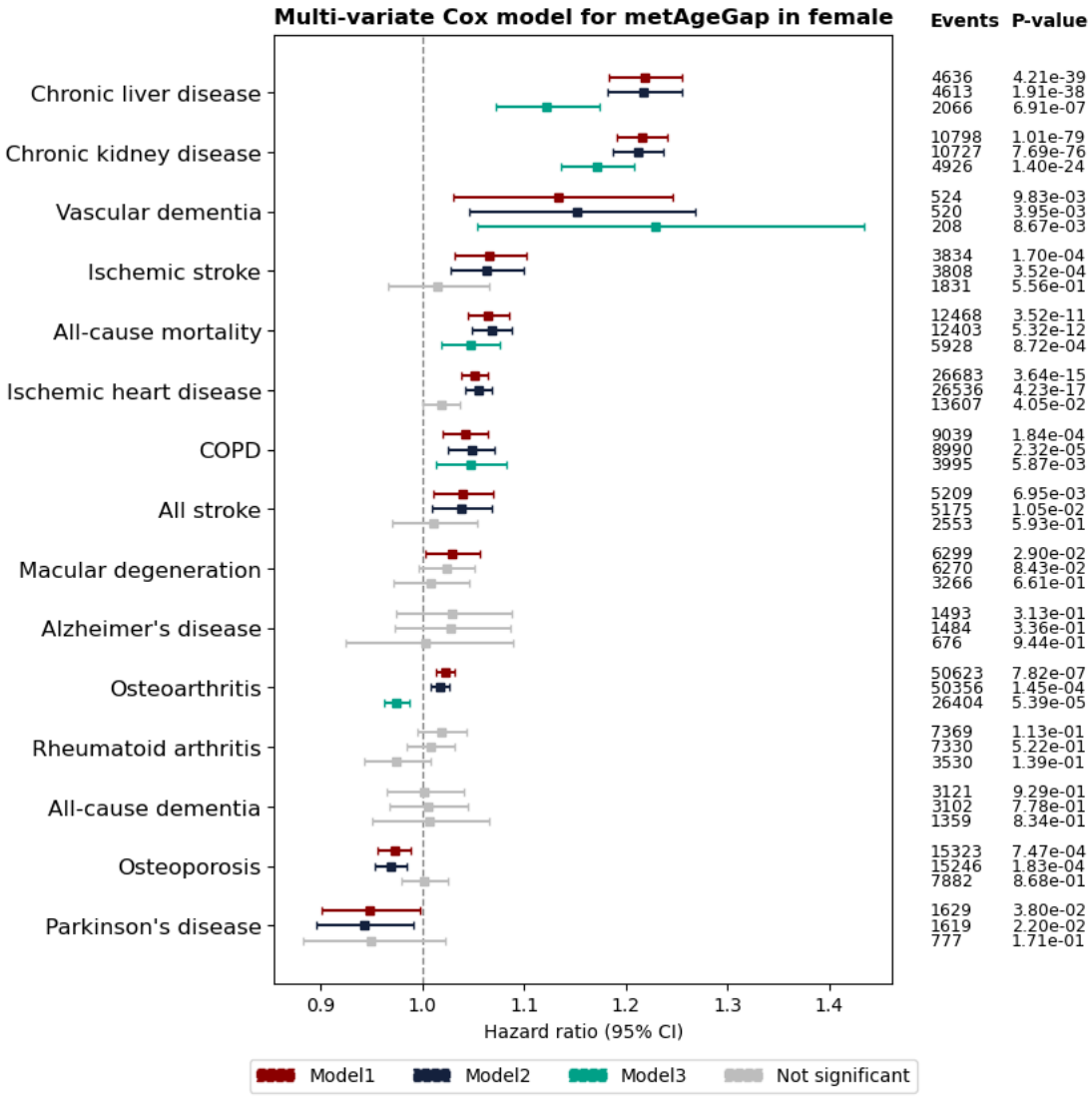
Supplementary Figure 15: Genetic correlations using LD Score Regression (LDSC) analysis highlighting strong genetic correlations of metAgeGap with various phenotypes in males. This plot contains the top correlations and the phenotypes discussed in Section Genetic association with metAgeGap of the manuscript. (CVI: Chronic venous insufficiency, CHF-NOS: Congestive heart failure not otherwise specified, MCHC: mean corpuscular haemoglobin concentration)



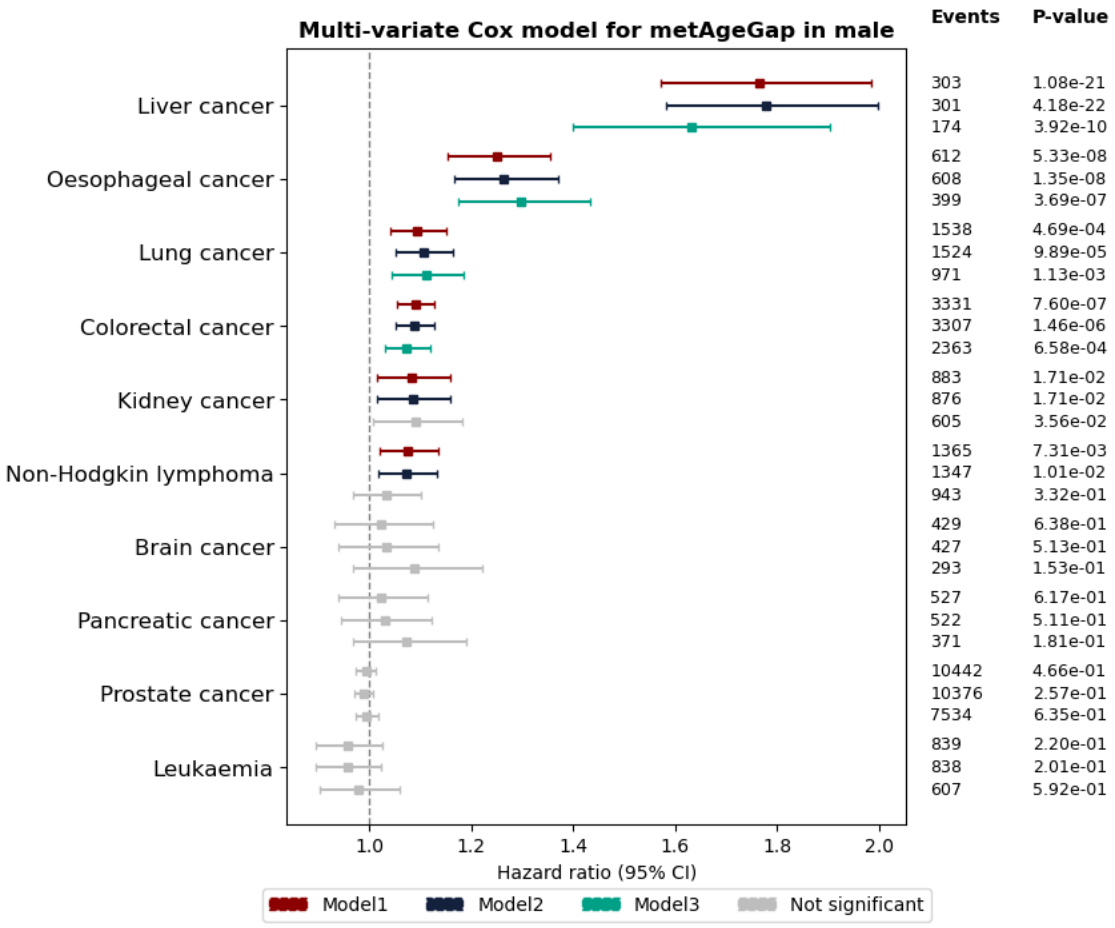
Supplementary Figure 16: Genetic correlations using LD Score Regression (LDSC) analysis highlighting strong genetic correlations of metAgeGap with various phenotypes in females. This plot contains the top correlations and the phenotypes discussed in Section Genetic association with metAgeGap of the manuscript. (LDL direct: direct measurement of low-density lipoprotein)



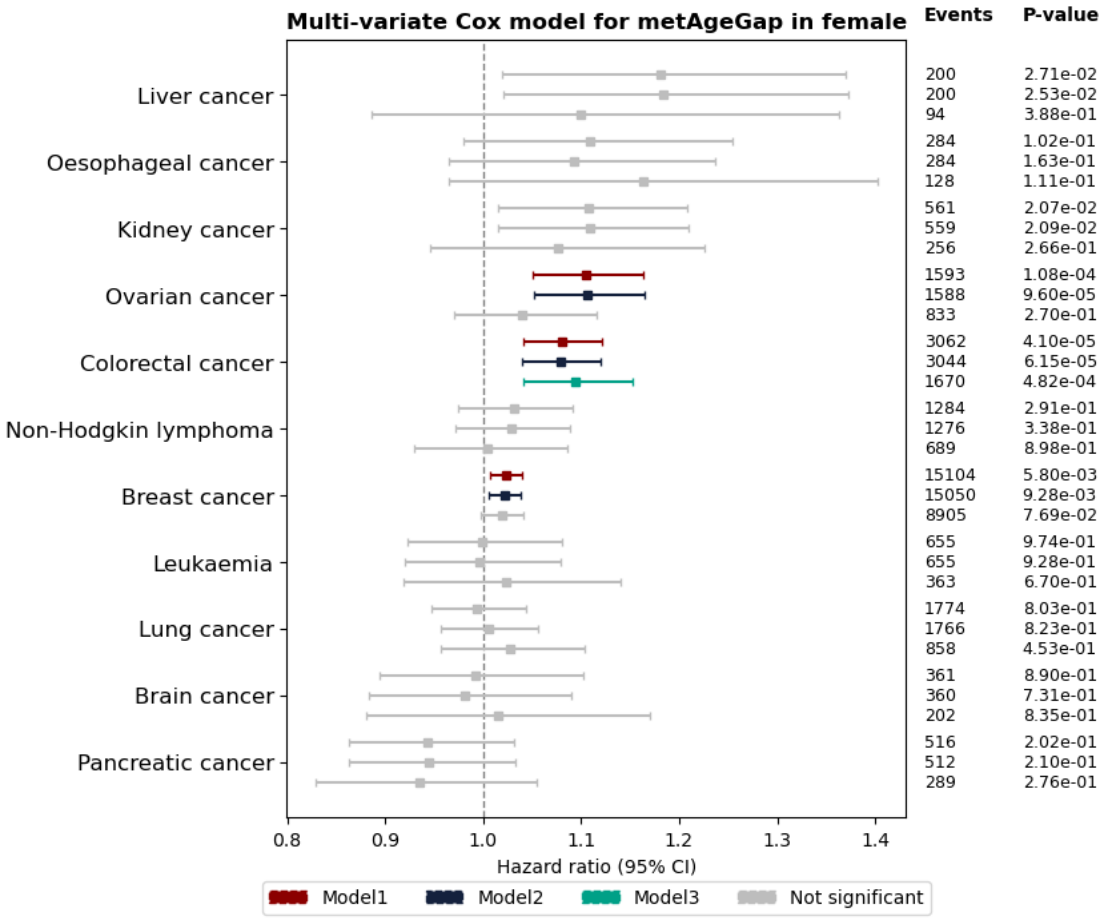
Supplementary Figure 17: Association of metAgeGap to risk of common diseases in males. Model 1 was adjusted for chronological age; model 2 was adjusted for recruitment centre, Townsend deprivation index, and ethnicity; model 3 was further adjusted for physical activity, BMI, smoking status, and alcohol frequency.



Supplementary Figure 18: Association of metAgeGap to the risk of common diseases in females. Mode 1 was adjusted for chronological age; model 2 was adjusted for recruitment centre, Townsend deprivation index, and ethnicity; model 3 was further adjusted for physical activity, BMI, smoking status, and alcohol frequency.

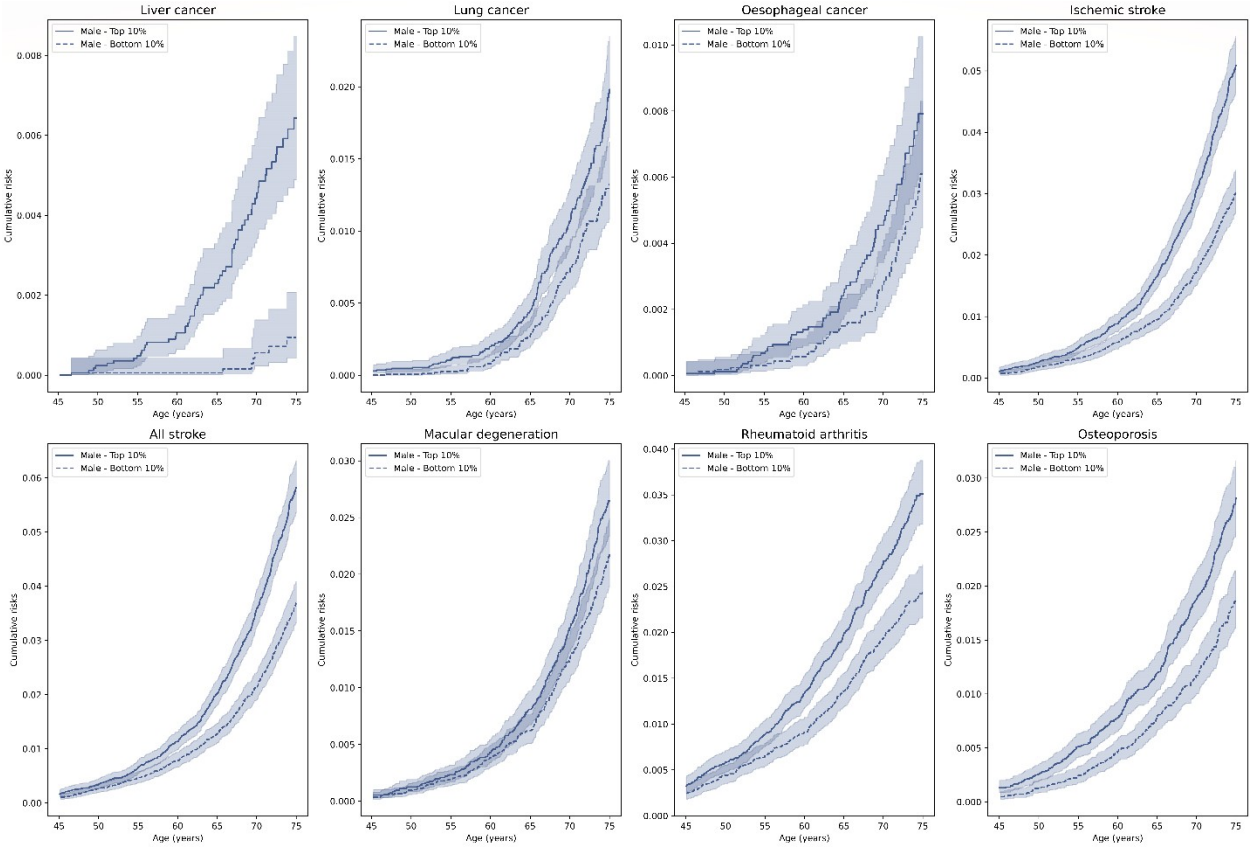


Supplementary Figure 19: Association of metAgeGap to risk of cancers in males. Mode 1 was adjusted for chronological age; model 2 was adjusted for recruitment centre, Townsend deprivation index, and ethnicity; model 3 was further adjusted for physical activity, BMI, smoking status, and alcohol frequency.

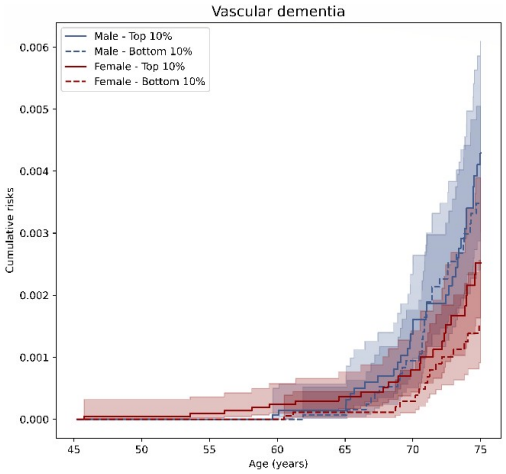


Supplementary Figure 20: Association of metAgeGap to risk of cancers in females. Mode 1 was adjusted for chronological age; model 2 was adjusted for recruitment centre, Townsend deprivation index, and ethnicity; model 3 was further adjusted for physical activity, BMI, smoking status, and alcohol frequency.

A



B



Supplementary Figure 21: **(A)** Cumulative incidence plot of top, and bottom 10% of the metAgeGap in diseases significant only in males (Cox model the 3). The X-axis denotes the chronological age and the Y-axis denotes the cumulative incidence and number at risk at each

age point is shown in Supplementary Tables 10-13. **(B)** Cumulative incidence plot of top, and bottom 10% of the metAgeGap in vascular dementia.

Chapter 4 Unveiling Sex Differences in Ageing through Proteomic Ageing Clock

Introduction

While our MetAgeGap analyses in Chapter 3 illuminated sex-specific ageing trajectories, several limitations temper its interpretability. First, the Nightingale panel's 249 measurements provide only a snapshot of lipid- and small-molecule metabolism, leaving many pathways unmeasured. Second, in females, the MetAgeGap–disease/mortality associations were largely mediated by BMI, suggesting that metabolic age signals may partly recapitulate adiposity rather than capture distinct ageing biology. Finally, the exclusion of statin users during model training introduced selection bias when studying ageing mechanisms and MetAgeGap's association with diseases.

Proteomics, however, as a more direct product of DNA and a functional representation of each of the pathways, provide a more comprehensive insight into ageing biology[97]. Further, the OLINK explore panel which included almost 3000 proteins covers all major pathways[159], giving us an opportunity to capture biological changes more comprehensively. Indeed, my previous work on the proteomic ageing clock has shown its capability to differentiate risks for 18 common diseases and 4 cancers independent of chronological age[97]. The study has also indicated the

association between ProtAgeGap and health outcomes were free of confounding factors including BMI, smoking, and other lifestyle and socioeconomic factors[97]. Although my paper has shown that the correlation between sex-specific ageing clock and unified ageing clock is high, proteins used to predict proteomic age between sexes were very different. How future health is correlated to ageing in males and females differently needs to be addressed. Moreover, associations between sex-specific factors and the proteomic ageing clock also remain unknown.

To fill this gap of knowledge, I built separate proteomic age clock models for males and females using the OLINK plasma proteomics panel from the UK Biobank (n=43,914) and studied their differences in proteomic architecture and functional pathways involved. I further tested how the proteomic Age Gap(protAgeGap) was correlated with blood biomarkers including biomarkers of ageing, inflammation, organ functions and hormones. I then studied how sex-specific factors including fertility and puberty timing were correlated with protAgeGap. Finally, I investigated the association between protAgeGap and 14 non-cancer common diseases, mortality and 8 cancers in each sex.

Method

Study cohort

UK Biobank (UKB)

UKB population was described in Chapter 2. Cohort population characteristics were summarised in **Table s1**. The mean recruitment age for females in this cohort was 58.42 (Q1=50.83, Q3=63.67), which is significantly younger than males recruited in the cohort ($p=2.48*10^{-08}$) (mean=59.08, Q1=50.83, Q3=64.50). Townsend deprivation index was not significantly different

between the two sexes ($p=8.68*10^{-2}$), while ethnicity ($p=2.35*10^{-07}$), International Physical Activity Questionnaire (IPAQ) physical activity group ($p=5.39*10^{-20}$), smoking status ($p=1.38*10^{-129}$) and education years ($p=3.83*10^{-138}$) were statistically different between male and female. Missing data was imputed using a random-forest-based algorithm provided by R package missRanger[126] when used as a covariate in linear association models (Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years). Imputation was performed with default hyperparameters using a maximum of 10 iterations and 200 trees.

Statistical analysis

A descriptive analysis of population characteristics was performed using the r package CBCgrps[127]. The study design and analysis pipeline are illustrated in **Fig 1**. The method of predicting biological age, feature importance calculation and feature selection were described in Chapter 2.

Proteomic Age Gap(protAgeGap)

Once the final model with Boruta-selected proteins was trained in the UKB training dataset, I calculated protein predicted age for the entire UKB cohort for each sex using fivefold cross-validation. Within each fold in male and female, a gradient boosting model was trained using the final hyperparameters and predicted age was generated for the test set of that fold. I then combined the predicted age values from each of the folds to create a measure of protein-predicted age for the entire population of males and females separately. Finally, the Proteomic

Age Gap (protAgeGap) was calculated as the residual between protein-predicted age and chronological separately in males and females.

Functional enrichment analysis

Functional enrichment analysis (GO: biological pathway, GO: molecular functions, KEGG pathways and Reactome pathways) was performed and visualised using custom Python scripts based on results extracted from String database[160].

Association of blood biochemistry biomarkers, clinical risk factors and sex-specific factors with protAgeGap

To test the association of blood biochemistry biomarkers, clinical risk factors, and sex-specific factors with protAgeGap, generalised linear models from statsmodel v.0.14.0 package[128] were used. For continuous exposure variables, standardisation was applied before inclusion in the models. Associations were adjusted for recruitment centre, ethnicity, education years, and Townsend deprivation index, for all exposures. P-values resulting from these analyses were corrected for FDR multiple testing. Beta values of the linear associations for males and females were plotted together in order for comparisons. The blue colour showed the beta value of male participants, red showed the female participants, and grey showed those associations which were not significant.

Associating pSIN with future health-related outcomes

To test the association between protAgeGap and incident health outcomes, all prevalent cases were removed before running the association model for each outcome. Multi-variate Cox proportional hazard model provided by lifeline v.0.27.8 package[129] was used with a pre-set step

size of 0.1. Survival outcomes were defined using follow-up time to the event and the binary incident event indicator. For all incident outcomes in both males and females, three successive models were tested with an increasing number of covariates: model1 adjusted for chronological age; model 2 was adjusted additionally for recruitment centre, Townsend deprivation index, ethnicity and education years; model3 was adjusted additionally for IPAQ physical activity group, BMI and smoking status. P-values of the hazard ratio were corrected for FDR multiple testing. Forest plots were generated with a minimum sample size threshold of 80 to ensure adequate statistical power and reliable interpretation.

For outcomes where the association was significant in all 3 Cox models, cumulative incidence plots were generated utilising the KaplanMeierFitter function from the lifelines package[129]. Due to limitations in case numbers or at-risk numbers at both ends, the x-axis of the plot was constrained to the age range of 45 to 75. This adjustment ensured a more focused visualisation of the cumulative incidence curve within a clinically relevant age range. P-values between cumulative incidence curves were calculated using a log-rank test with adjustment for FDR multiple testing.

Results

Proteomic data predicts age in both sexes

In the random split 70% training dataset of UKB participants where plasma proteomic features were measured using the Olink explore panel, I developed separate models using the gradient boosting tree method with proteomics data regressed to chronological age in males and females (**Fig 1**). After hyperparameter tunning, models achieved high accuracy in predicting

chronological age in both males and females with R^2 of 0.88 (std=0.003) and 0.88 (std=0.005) respectively for 5-fold cross-validation (5CV) in the training dataset (***Fig S1a, b***). The performance of the model in the 30% hold-out testing dataset showed no decrease in performance with R^2 of 0.88 in both sexes, suggesting no overfitting of the model. Protein-predicted age in the testing dataset of both sexes showed a high Pearson correlation with chronological age with r = 0.94 in both sexes. We then used the Boruta algorithm to select all relevant features contributing to chronological age prediction which yielded 174 proteins for females and 136 proteins in males respectively compared to the 2,917 proteins in the full Olink panel. Of the Boruta-selected proteins, 41 proteins were only selected in males, 79 were only selected in females and 95 were selected by both sexes (***Fig S2***). Details of selected proteins are shown in supplementary table ***Table S2***. The chronological prediction model with selected proteins showed a minimal decrease in performance with a 5CV R^2 in the training dataset of 0.88 (std=0.004) and 0.87 (std=0.003) in females and males respectively. Model performance in the testing dataset showed no change, maintaining an R^2 of 0.88 for females and 0.87 for males (**Fig2 a, b**).

Pathway enrichment analysis was then performed based on GO, KEGG and Reactome databases to investigate pathways involved in ageing models in both sexes and also the differences (***FigS3,4***). Proteins selected only in females were enriched in biological processes such as regulation of multicellular organism development, multicellular organism process, development process, cell population proliferation and multiple signalling pathways (PI3K/AKT, RAS, MAPK) which was shown to be involved in regulating the cell cycle and hence directly related to cellular proliferation, cancer development and longevity[161] (***FigS3***). Other signalling pathways enriched

included MAPK signalling, and RAS signalling pathways were shown to be related to cancer developments. Male-specific proteins, on the other hand, were only enriched for cell adhesion migration which recent studies have also shown to be related to aging and longevity[162] (***FigS4***).

Subsequently, I investigated the top proteins that contributed to the age prediction models in both males and females (**Fig2 c, d**). ELN and EDA2R proteins were ranked as the top three in both sexes and exhibited a positive correlation with older age. ELN is a crucial component of elastic fibres that constitute the extracellular matrix (ECM) and confers elasticity to organs. It has been demonstrated to contribute to cellular homeostasis, which is a hallmark of ageing. EDA2R is a type III transmembrane protein belonging to the tumour necrosis factor receptor superfamily and is associated with cytokine signalling pathways and TNFR activities. Notably, FSHB, which ranked second in the female model, was not selected in the male model. It is known as the follicle-stimulating hormone beta subunit, which contributes to promoting fertility and follicular oocyte development[98]. Furthermore, both NEFL and GFAP were selected as the top 20 proteins in both males and females, which are known as markers for neurodegenerative activity and brain ageing.

Using the model after selection, I calculated the proteomic Age Gap (protAgeGap) as the residual between protein-predicted age and chronological age in females and males separately (**Fig2 e**). This is to infer if an individual's protein predicted age, and hence if biological age is younger or older than the people in the same chronological age. Differences between protAgeGap of 25 percentile and 75 percentiles in females and males were 3.5 years and 3.6 years respectively with standard deviation of 2.6 and 2.7 respectively (**Fig2 f**).

Association to biochemical markers, clinical risk factors and sex-specific variables

To validate, I explore how biochemistry biomarkers and clinical risk factors were correlated with protAgeGap in women and men (**Fig3 a, b**) (***Table S3-6***). After adjusting for social economic confounders including recruitment centre, Townsend deprivation index, ethnicity and education years, protAgeGap was associated with all biochemistry biomarkers apart from albumin, vitamin D, and lipoprotein A (**Fig3 a**) (***Table S3***). This includes markers of inflammation such as C-reactive protein (CRP), indicators of glucose metabolism such as higher HbA1c levels, ageing-related biomarkers like telomere length and insulin-like growth factor 1 (IGF1), and biomarkers related to kidney and liver function. In addition, significant associations were observed with sex hormones such as lower levels of SHBG and testosterone, although associations with SHBG became nonsignificant when adjusting out the effect of menopause (**Fig S5A**).

On the other hand, in males, 11 out of 28 markers were not significantly associated with protAgeGap such as vitamin D, biomarkers for liver function (ALP, GGT, AST), biomarkers for lipid metabolisms etc. (**Fig3 a**) (***Table S4***). However, protAgeGap in males was strongly associated with biomarkers for ageing, inflammation, glucose metabolism, and kidney function. Interestingly, protAgeGap's association with sex hormones such as SHBG and testosterone were inversed compared to which in females. Higher testosterone in older males may be explained by even higher levels of SHBG blocking the bioactivity of testosterone in males. We observed that for cholesterol, LDL cholesterol, APOB, and calcium, the associations differed between males and females, displaying opposing directions. I hypothesize that this divergence may be due to the use of antihypertensive and lipid-lowering medications. To test, I conducted a sensitivity analysis excluding participants who reported using such medications. This sensitivity analysis

revealed that the negative associations observed in males became statistically nonsignificant (***Fig S5B***).

Additional analysis on clinical risk factors showed general markers of ageing including poor self-rated health, slow walking pace, longer reaction time, and less hand grip strength were all correlated with protAgeGap in females (**Fig3 b**) (***Table S5***). Prevalent disease risk factors such as type II diabetes, hypertension, and obesity were also associated with older protAgeGap. Furthermore, lifestyle risk factors such as ever smoking, smoking amount (packyears) and sleeping longer than 10 hours a day were also correlated with older protAgeGap in females. Interestingly, self-rated facial ageing ($p=6.50*10^{-02}$), alcohol frequency ($p=4.76*10^{-01}$), and fluid intelligence ($p=1.42*10^{-01}$) were not significantly correlated with protAgeGap in females.

On the other hand, in males, both older self-rated facial ageing ($p=6.35*10^{-11}$) and lower fluid intelligence ($2.32*10^{-02}$) were significantly correlated with protAgeGap (***Table S6***). Similar to the associations in females, protAgeGap in males correlated most strongly with poor self-rated health ($p=3.17*10^{-26}$) followed by type II diabetes ($p=2.79*10^{-18}$), and slow walking pace ($p=3.25*10^{-23}$). Noticeably, while associations stayed significant in females, tired every day ($p=7.72*10^{-02}$), frequent insomnia ($p=7.55*10^{-02}$), blood pressure (systolic:$p=7.65*10^{-02}$, diastolic:$p=1.36*10^{-01}$), and heel bone mineral density (4.52*10-01) the associations were not significant in males.

Noticing the difference between males and females, we next explored how sex-specific variables correlated with protAgeGap in females (**Fig4 a**) (***Table S7***) or males (**Fig4 b**) (***Table S8***). Reproductive factors such as years since last live birth, menopause, younger menopause age

and bilateral oophorectomy were correlated with older protAgeGap and number of live births, birth weight of first child, age at last live birth, and age at first live birth were correlated with younger protAgeGap. Additionally, ever using HRT or oral contraceptives was associated with older protAgeGap as well. In both males and females, delayed puberty was correlated with younger protAgeGap such as older menarche age in females and older than average when first growing facial hair and voice break in males. Additionally, in males, balding was positively associated with older protAgeGap while the number of children fathered was correlated with younger protAgeGap.

protAgeGap is associated with morbidities and mortality

To explore how protAgeGap is associated with health outcomes, we tested the association of protAgeGap with 14 common diseases, all-cause mortality, and 8 cancers including breast cancer and ovarian cancer for females and prostate cancer for males in both sexes (**Fig5**). In females, protAgeGap was significantly associated with 14/15 common diseases and mortality with the exception of macular degeneration ($p=7.74*10^{-02}$) when adjusted for chronological age (**Fig5 a**) (***Table s9***). Further adjustment of socio-demographic factors such as recruitment centre, Townsend deprivation index, and ethnicity in model 2 and lifestyle factors such as IPAQ activity group, BMI, and smoking status in model 3 did not change the significance. For the fully adjusted model, the association was led by neurodegenerative diseases including vascular dementia (HR=1.54, CI:1.23,1.93), Alzheimer's disease (1.51, 1.32, 1.71), and all-cause dementia (1.43, 1.29, 1.58) and then followed by chronic kidney disease (1.31, 1.23, 1.38) and ischemic stroke (1.28, 1.18, 1.38). Incident all-cause mortality was also significantly associated with protAgeGap independent of chronological age and other confounding factors ($P=1.12*10^{-15}$)

with HR of 1.24 (CI:1.17, 1.30). In males, the top associations were also led by neurodegenerative diseases such as vascular dementia and Alzheimer's disease (**Fig5 b**) (***Table s10***). The association was followed by all-cause mortality, all-cause dementia, chronic kidney disease and COPD in males. Noticeably, Ischemic stroke, which was ranked at No.5 in females, ranked at No.8 in males. Moreover, Parkinson's disease which ranked at No.6 in females was not significant in males.

In cancers, protAgeGap only showed a significant association in breast cancer in females after adjusting for confounders (p=9.85*10-5) with an HR of 1.10 (CI:1.05,1.16) (**Fig5 c**) (***Table s11***). In males, four cancers were significantly associated with protAgeGap after full adjustment of social-demographic and lifestyle factors (**Fig5 d**) (***Table s12***). Non-Hodgkin lymphoma ranked top with an HR of 1.44 (CI:1.23, 1.70) followed by lung cancer (HR=1.39, CI:1.21, 1.59), oesophageal cancer (1.33, 1.09, 1.62), and prostate cancer (1.15, 1.08, 1.22).

For mortality and 13 common diseases that were significantly associated with protAgeGap in all models, I further tested if participants in the top and bottom quartile of protAgeGap exhibited divergent cumulative incidence by chronological age in each health outcome (***Fig 6***) (***Table s13-16***). In females, I noticed that the largest divergence in cumulative incidence at the age of 75 between the top and bottom quartile of protAgeGap was seen in Alzheimer's disease (***Fig 6***). Participants within the top quartile of protAgeGap showed a 3.56-time higher cumulative incidence of Alzheimer's disease compared to those within the bottom quartile. This was followed by all-cause dementia and vascular dementia which all showed more than doubled cumulative incidence when comparing participants within the top and bottom quartile of protAgeGap. As for all-cause mortality, participants with the top quartile of protAgeGap showed

a 1.57 times higher cumulative incidence compared to the bottom quartile. In males, only two health outcomes showed a more than two times higher cumulative incidence in the top quartile of protAgeGap at the age of 75, i.e. vascular dementia (2.49-time) and all-cause mortality (2.02-time) (***Fig 6***). We also noticed that in males the biologically youngest group (lowest 25% protAgeGap) had higher cumulative disease risks in ischemic stroke, all-cause mortality, all stroke, chronic liver disease, and ischemic heart diseases compared to the biologically oldest group of females (highest 25% protAgeGap) denoting the disease risk difference in both baseline and ageing process. Cumulative risks and number at risk at each time point were shown in ***supplementary Table s13-16.***

For cancers significant in all three models, I further investigated how participants within the top and bottom quartile of protAgeGap showed divergent cumulative incidence rates associated with chronological age. In breast cancers of females, while the difference is small, we still observed a difference in cumulative incidence of the top and bottom quartile of protAgeGap (***Fig s6a***). In fact, at the age of 75, participants within the top quartile of protAgeGap showed a 1.11-time higher cumulative incidence compared to those within the bottom quartile of protAgeGap. In males, the differences between cumulative incidence were more pronounced (***Fig s6b***). For non-Hodgkin lymphoma, participants within the top quartile of protAgeGap exhibited 2.19 times cumulative incidence compared to those within the bottom quartile at the age of 75. This was followed by oesophageal cancer, lung cancer, and prostate cancer with a 1.92-time, 1.72-time, and 1.39-time higher cumulative incidence respectively. Cumulative risks and number at risk at each time point were shown in ***supplementary Table s17-20.*** Associations

of top20 proteins selected in males and females with incident health outcomes were shown in **Fig S7a,b.**

Discussion

Using the machine learning model, I built separate models predicting chronological age in females and males based on plasma proteomic data in UK Biobank and showed although different proteins were used in males and females, both models predicted chronological age accurately. I determined the residual between protein-predicted age and chronological age as a measure of an individual's ageing rate compared to those with the same age and termed it as protAgeGap. I found protAgeGap in both sexes correlated with various biomarkers of ageing and organ function decline while in males protAgeGap was less correlated with liver function biomarkers. I also found that protAgeGap correlated with clinical risk factors and reproductive factors independent of social-demographic factors in both sexes. This study also demonstrated the ability of protAgeGap to differentiate risks of 13 common morbidities and mortality in both sexes. Of note, protAgeGap was more prevalent in associating with cancer incidence in males compared to females with significant associations with 4 types of cancers in males while only 1 cancer in females was significant.

The findings from the present study corroborate and expand upon previous research studies, highlighting the proteomic differences of ageing in males and females. Compared to previous proteomic ageing models which combine male and female, sex-specific models achieved similar accuracy in predicting chronological age in both sexes but with only 174 proteins selected for females and 136 proteins selected for males[97]. My sex-specific models selected 51 distinct

proteins compared to the combined model including 35 proteins selected in females, 15 proteins selected in males and 1 protein (LEFTY2) selected by both sexes (**Fig S9**). Although none of the 15 proteins selected in males was presented as the top 20 most important proteins in my study, CGA which was ranked at 4th place in the female model was not picked up by the combined model.

From pathway enrichment analysis, my noticed that most proteins common to both sexes were enriched in multiple developmental processes including anatomical structure development, system development, multicellular organism development etc. which is one of the most fundamental biological processes that's related to multiple hallmarks of aging[13]. Common proteins selected by both sexes are also enriched in pathways such as PI3k-AKT, a signalling pathway that regulates cell growth, survival and cell cycle[161]. Interestingly, previous studies have also shown that full activation of AKT leads to a phosphorylation event which results in the inhibition of pro-apoptotic FOXO proteins[163], a well-established protein for aging and longevity[164]. The PI3k-Akt signalling pathway is also a major signalling pathway in various cancer types as it acts as a regulator for hallmarks of cancers including cell survival, metastasis, and metabolism and also tumour environment including angiogenesis and inflammatory factor recruitment[165,166]. Proteins selected in the female model also enriched in RAS and MAPK pathway, also a major signalling pathway for cells to control survival, differentiation, proliferation, metabolism and motility similar to PI3K-Akt[167]. Previous evidence has shown the cross-inhibition[168] and cross-activation[169] activities between two signalling pathways. RAS/MAPK pathway transduces signals from the extracellular milieu to the cell nucleus in which it regulates gene expressions responsible for cell growth, cell cycle, tissue repair and etc[170]. It is also one of

the most frequently mutated oncogenes in cancers as it is also able to activate angiogenesis processes[171]. Proteins selected in males were especially enriched in biological processes involving cell adhesion although ECM in general was enriched in both sexes. As a scaffold to cell and tissue structure, it also modulates cell processes, repairing and conducting intracellular signals. Ageing leads to functional decay of ECM differently in different tissues i.e. ECM degradation leads to osteoarthritis, ECM deposition leads to tissue fibrosis, and change in ECM composition leads to arterial stiffening and ultimately chronic liver diseases and cardiovascular diseases[162]. ECM is also crucial for maintaining stem cell function, a hallmark of ageing, by transducing integrin signals[172].

Of note, among top20 proteins ranked by mean absolute SHAP value, half of them were different in males and females (FSHB, CGA, PAEP, LECT2, ROBO1, AFP, CCDC80, SUSD5, FBLN2, C0L6A3 in females; LTBP2, TSPAN1, KLK4, CDON, KLK3, GDF15, CXCL14, AGRP, ENG, ACRV1 in males). FSHB and CGA which ranked second and fourth in the female model were all involved in female reproductive functions. FSHB is the beta subunit of follicle-stimulating hormone (FSH) which promotes follicular oocyte production, maturation and also menstrual cycle[173]. Research has also shown that FSH is associated with bone mass loss, obesity and altered energy metabolism in particular during the last menstrual period[173]. CGA, similar to FSH, is a protein produced in the placenta which belongs to the glycoprotein hormones. Apart from its role in controlling fertility and pregnancy, it was also shown to be a receptor of estrogen and was elevated in breast cancers[174] which leads to cell proliferation. CGA was also found to be correlated with multiple solid cancer development and progression besides breast cancer including lung[175] and neuroendocrine cancer[176]. In addition, CGA was also used as a biomarker

for premenstrual psychoemotional symptoms[177]. AFP, on the other hand, is a major plasma protein initially produced by the yolk sac, liver and gastrointestinal tract whose expression level remains high in fetal states until born. The expression level of AFP will, however, increase in adults when liver cells are experiencing proliferation which usually indicates hepatocellular carcinoma[178]. The presence of AFP only in the top 20 of the female model but not the male model may also be the explanation why liver damage biomarkers such as GGT, AST, and ALT were only significantly associated with protAgeGap in females. GDF15, which ranked in the top 20 only in male models, was another well-known protein for human ageing. It is one of the crucial protein stress responses and is also involved in tissue tolerance during inflammation and tissue injury[179]. It has also been found to be associated with multiple chronic diseases where the risks increase as age grows[180]. Indeed, many studies have used GDF15 as a biomarker for cardiovascular diseases[181], mitochondrial diseases[182], diabetes[183] and cognitive degenerative diseases[184]. In males, we also discovered that ENG and ACRV1 were ranked at the top 20op20 in only males but not females. These two proteins all belong to the bone morphogenetic protein (BMP) pathway. The BMP pathway is found in many tissues and plays a crucial role in vascular signalling and development and homeostasis[185]. BMP pathway is fine-tuned and sophisticatedly controlled, variations in different pathways lead to morphological, functional and molecular heterogeneity of endothelial cells in arteries, veins, lymphatic vessels and capillaries[186]. Apart from its association with cardiovascular diseases[187], the BMP pathway is also a key regulator of age-related cognitive decline[185]. Studies have shown that age-related BMP signalling increase will lead to inhibition of hippocampal neurogenesis and hence contribute to the age-related cognitive function decline[188]. SiA similar function was also observed in mice models[189].

Interestingly, recent studies have also shown that the BMP pathway promotes tumorigenesis and progression by activation of cancer cell proliferation and regulation of angiogenesis[190]. This includes increased risks for leukemia[191], lung cancer[192], colorectal cancer[193] etc.

In this study, I found that, in both sexes, older puberty age (Menarche age in females; first facial hair and age when voice breaks in males) was correlated with younger protAgeGap. Limited studies have directly shown the relation between ageing and puberty timing but evidence has proved a significant correlation between later puberty timing and shorter lifespan[194], higher disease risks[195] and telomere shortening[196]. Menopause is another important sex-specific variable for females. We found a strong correlation with not only menopause but also age at menopause was associated with protAgeGap. During menopause, many biological changes will occur including hormonal changes (essentially estrogen and progesterone), bone loss, cardiovascular changes, and weight gain among others. Those changes will ultimately lead to several health conditions including cardiovascular diseases[197], Alzheimer's disease[198], Osteoporosis[199] and others. Furthermore, menopause was also found to be associated with epigenetic age[200] similar to the findings of this study. Other female reproductive variables such as the number of live births were also correlated with younger protAgeGap after adjusting for socio-demographic confounding factors. Interestingly, this study has shown an even stronger association between age at last birth and younger protAgeGap. I then showed that it is the time between the last birth and now that is correlated most strongly with protAgeGap (Beta:0.32, 95%CI:0.24,0.41). Indeed previous studies have shown that fetal stem cells were helpful for repairing cardiac injury and that the mechanism was potentially useful for developing cardiovascular regenerative therapy[201]. Interestingly, previous studies suggested that pregnancy

led to faster ageing due to a rapid increase in diverse forms of stress although this process will soon be recovered after live birth[202,203].

In general, I found protAgeGap differentiated risks of most non-cancer common diseases in both sexes independent of chronological age, social-demographic and lifestyle confounding factors. These associations were led by vascular dementia, Alzheimer's disease and all-cause dementia in both sexes. This could be explained by biomarkers for neurodegenerations such as NEFL, GFAP and GDF15 were selected as the top 20 most important proteins in the protAgeGap model. However, I did see that protAgeGap in females was correlated strongly with Parkinson's disease while it was not significant in males. I also noticed that protAgeGap was associated with all-cause mortality stronger in males (ranked top 3) compared to females (ranked top 8). As for cancers, protAgeGap was only associated with breast cancer in females while it was associated with non-Hodgkin lymphoma, lung cancer, oesophageal cancer, and prostate cancer in males. While we did find proteins related to those cancers or pathways involved in the development of those cancers in females as well, the absence of the association may be because the subset of the UK Biobank used in this study was limited in number and did not have enough cases for testing.

Despite the Olink Explore panel used in this study is at the time one of the largest targeted protein panels available, only around 3000 proteins were tested in contrast to more than 20,000 proteins estimated in the human proteome. Other platforms such as SOMAscan (>10,000 proteins) consisted of more proteins, although consistency between different platforms needs to be tested. Secondly, Olink measurements are relative quantifications suitable for large cohort studies due to high throughput and cost considerations, any translation of such assays into

clinical practice would require replication of these findings using absolute quantification. In addition, although the model was validated well in the random split testing dataset, the findings were not externally validated. The accuracy of the model and the impact of protAgeGap on disease risks in other ethnic groups are in plan to be validated in future projects.

Overall, this study demonstrates that plasma proteome is a powerful tool to distinguish between the biological differences of ageing in males and females and can be used to quantify biological ageing and the relative risks of morbidities and mortality. From this, we can easily identify different proteins and potential pathways involved in the ageing of males and females. It also provides a means of validating how sex-specific variables including puberty timing and reproductive factors are related to biological age. In addition, my work has shown the ability of protAgeGap as a reliable tool to discover the underlying mechanisms of overall ageing and age-related diseases and how they are different between sexes.

Figures

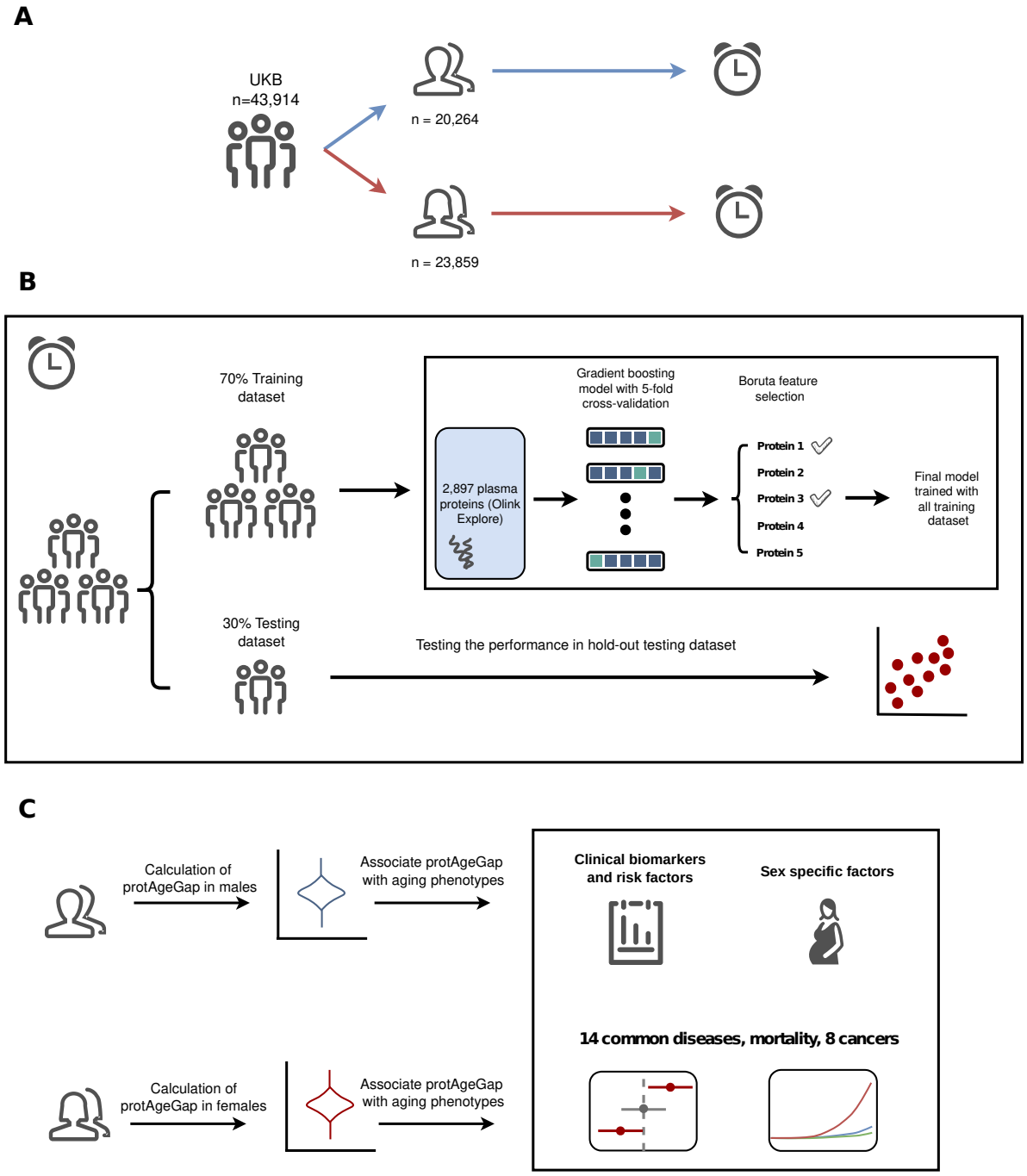


Figure 1: Overview of the study design and analytic approach. (a) gradient boosting model was built stratified by sex in UKB population. **(b)** For males and females, model training and feature selection process was carried out in 70% training dataset. Boruta feature selection algorithm was used to select only relevant features for downstream analysis. Performance of age prediction was tested in the 30% testing dataset **(c)** proteomic Age Gap (protAgeGap) was calculated for each sex and association against aging-related phenotypes were tested.

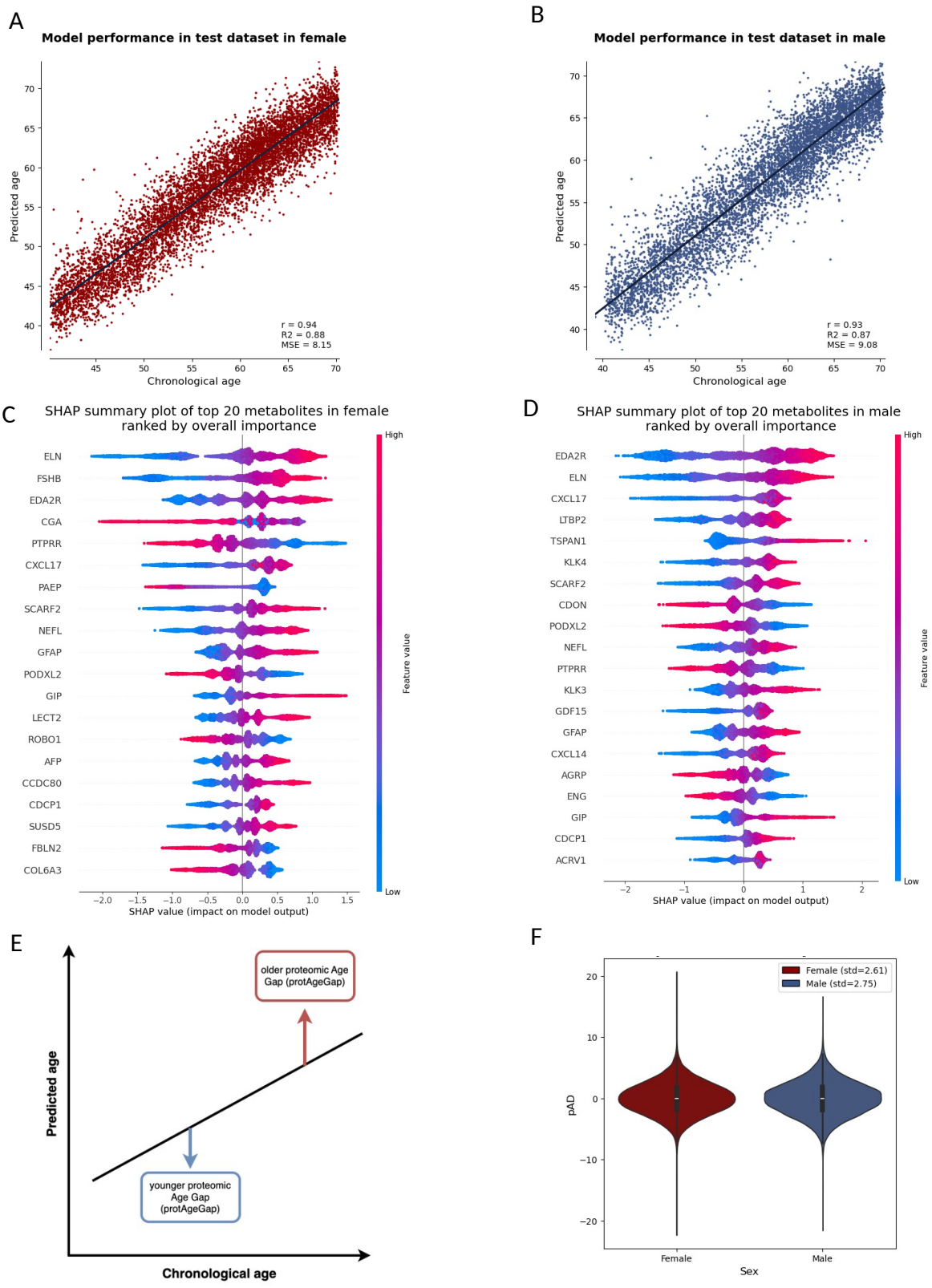


Figure2: Proteomic age model performance. (a) Scatter plot of model performance of female model in the testing dataset (R2=0.88). **(b)** Scatter plot of model performance of male model in the testing dataset (R2=0.87). **(c)** SHAP summary plot of top 20 protein in female model ranked by mean absolute importance. **(d)** SHAP summary plot of top 20 protein in male model ranked by mean absolute importance. **(e)** protAgeGap was calculated by the residual between protein predicted age and chronological age. **(f)** comparison of protAgeGap distribution of females and males.

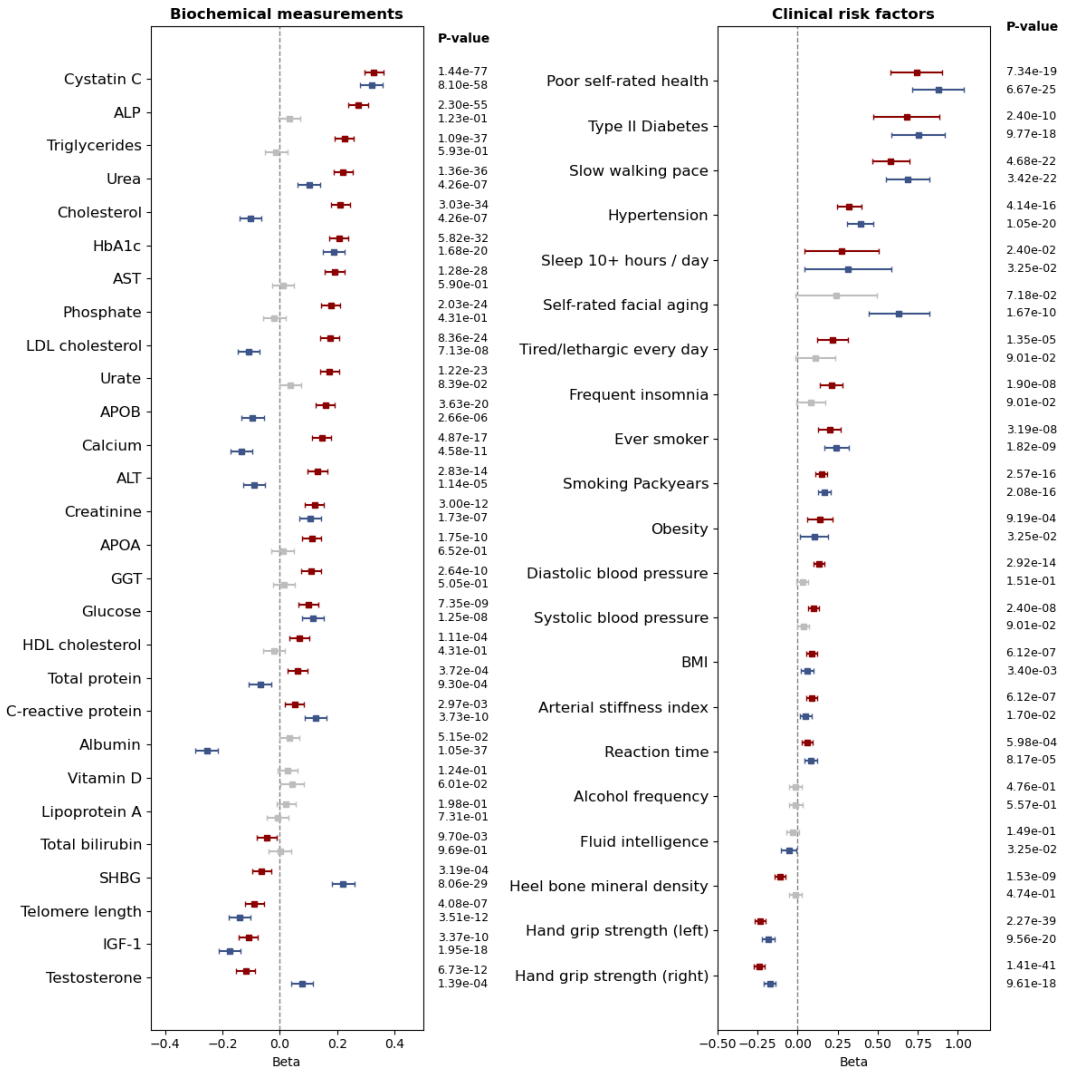


Figure 3: Associations between biochemical measurement and clinical risk factors and protAgeGap. Linear regressions were performed between each exposure and protAgeGap adjusting for recruitment centre, ethnicity, education years, and Townsend deprivation index. Values were standardized if quantitative. Red shows the associations in females and blue shows the associations in males. Grey colour showed if the association was not significant after FDR correction. **(a)** showed associations for biochemical measurements. **(b)** showed associations for clinical risk factors.

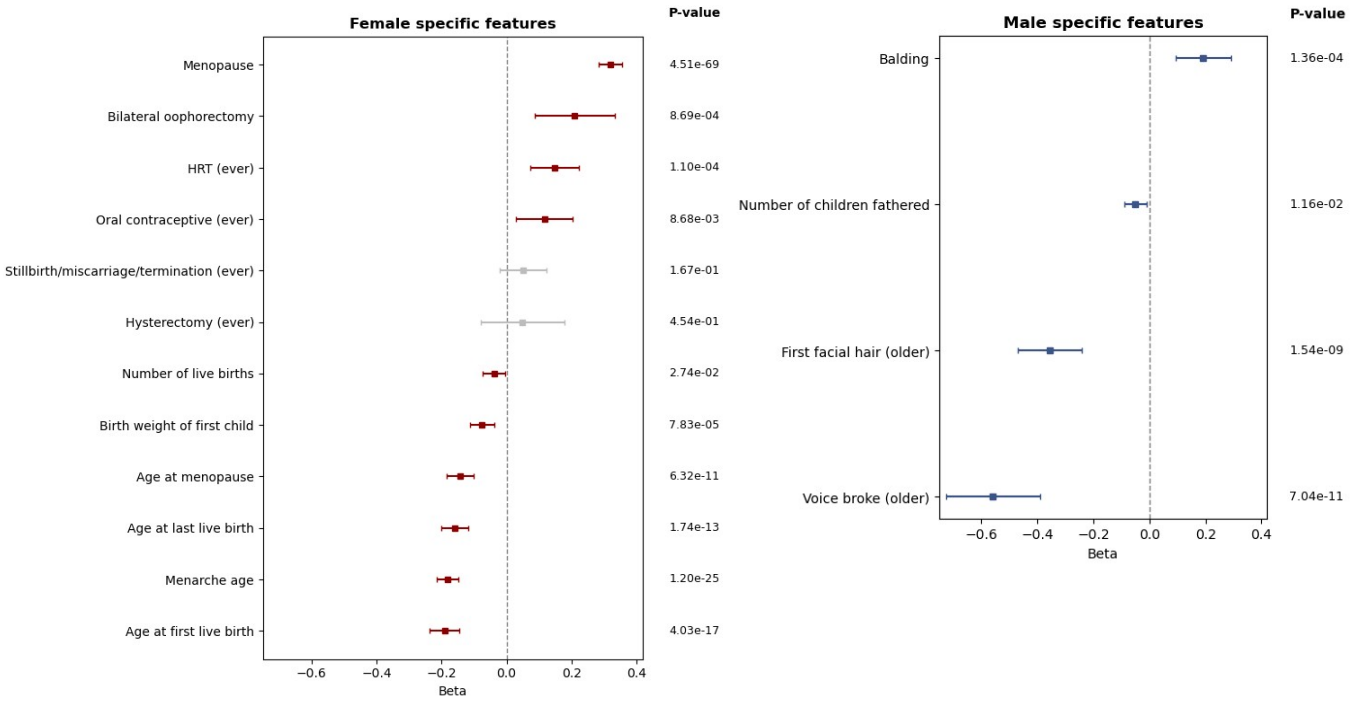


Figure 4 Associations between sex specific factors and protAgeGap. Linear regressions were performed between each exposure and protAgeGap adjusting for recruitment centre, ethnicity, education years, and Townsend deprivation index. Values were standardized if quantitative. Grey color showed if the association was not significant after FDR correction. **(a)** showed associations for female specific factors. **(b)** showed associations for male specific factors.

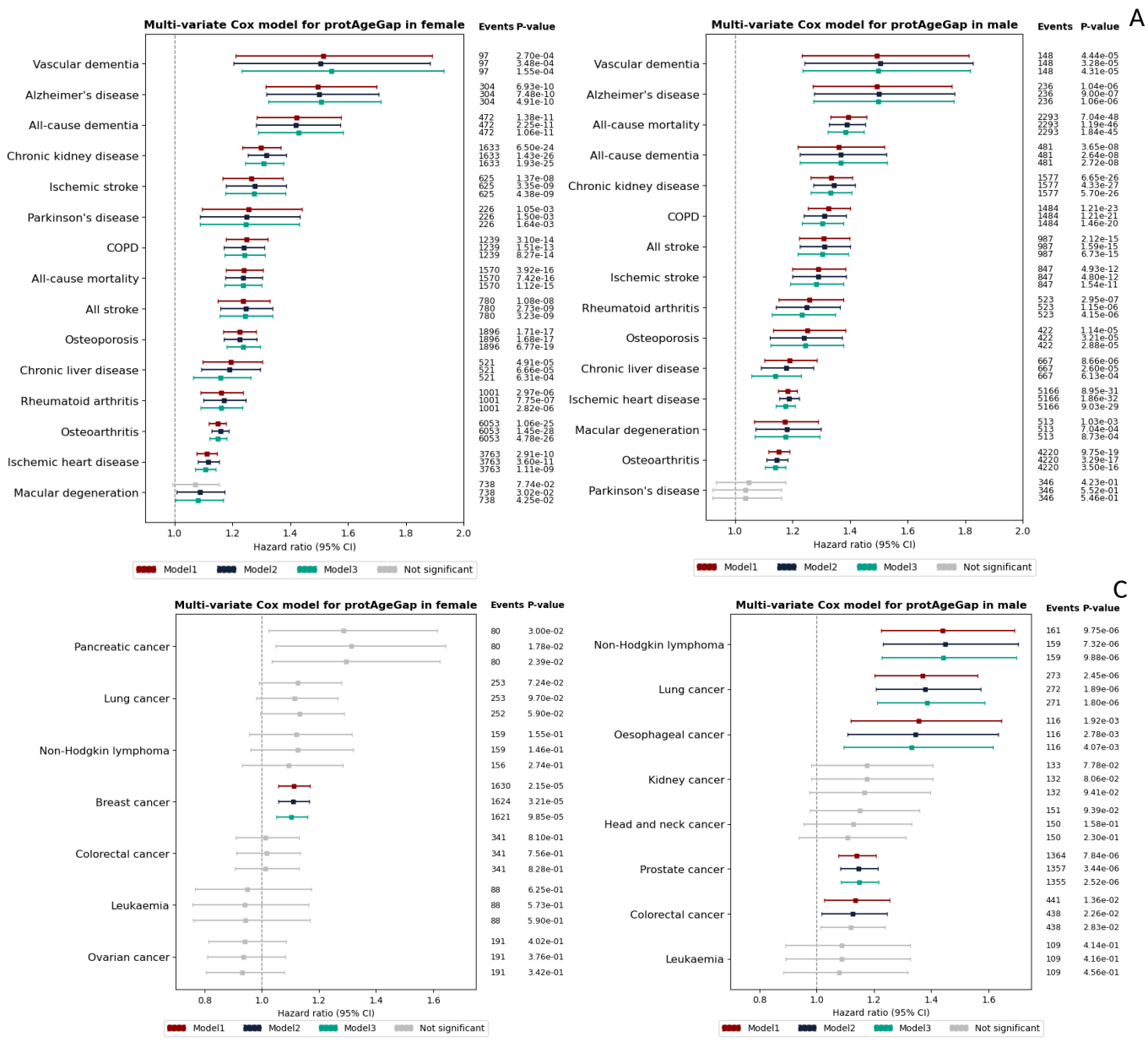


Figure 5 protAgeGap differentiates incident morbidity and mortality risks in both females and males. Cox proportional hazard models are used to investigate associations between protAgeGap and incidence of morbidity and mortality. Model1 (red) adjusted for chronological age, model2 (blue) additionally adjusted for recruitment center, Townsend deprivation index, and ethnicity, model3 (green) additionally adjusted for IPAQ activity groups, BMI and smoking status. Grey colour denotes if the association is not significant after FDR correction. **(a)** Associations of non-cancer morbidities and mortality in females. **(b)** Associations of non-cancer morbidities and mortality in males. **(c)** Associations of cancers in females. **(d)** Associations of cancers in males.

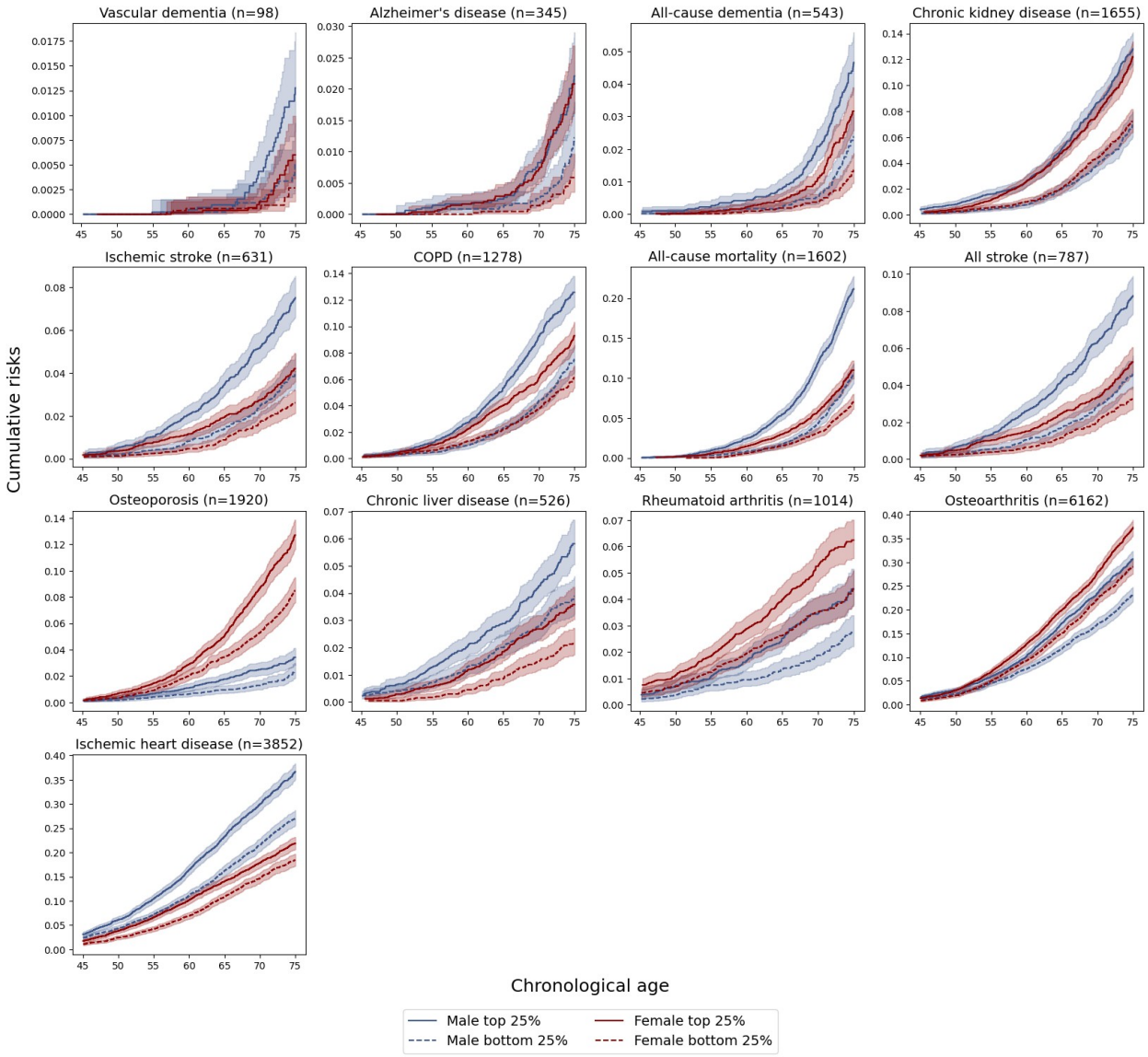
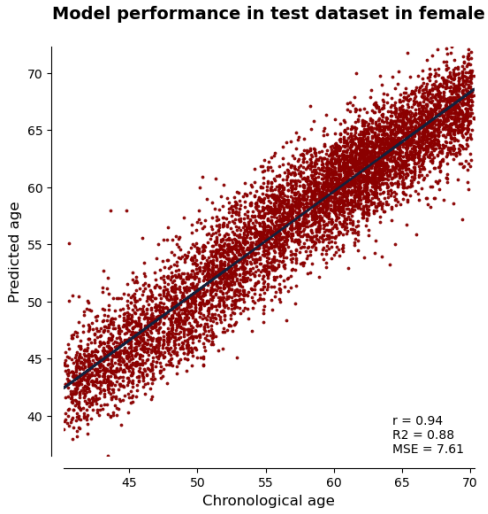


Figure 6* protAgeGap differentiates future risks of none-cancer morbidities and mortality.** Cumulative incidence plot of top and bottom 25% of the protAgeGap was shown in none-cancer health outcome significant in the cox model in female with 95% confidence interval shown as lighter shading. X-axis denotes the chronological age and Y-axis denotes the cumulative incidence. Cumulative incidence and number at risk at each age point for female is shown in ***Table s11 and ***Table s12***. Cumulative incidence and number at risk at each age point for male is shown in ***Table s13*** and ***Table s14***.

Supplementary figures

A



B

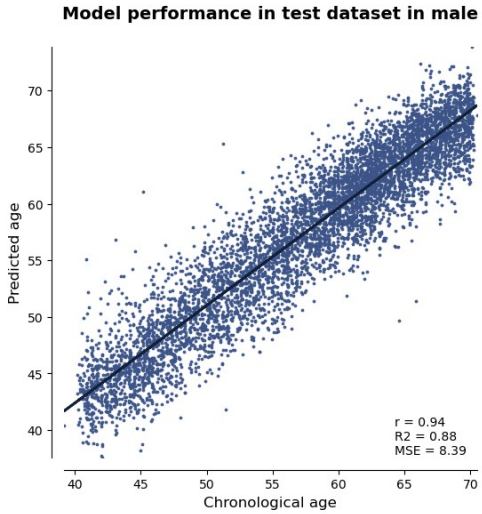


Figure S1: Proteomic age model performance before feature selection. (a) Scatter plot of model performance of female model in the testing dataset (R2=0.88). **(b)** Scatter plot of model performance of male model in the testing dataset (R2=0.88).

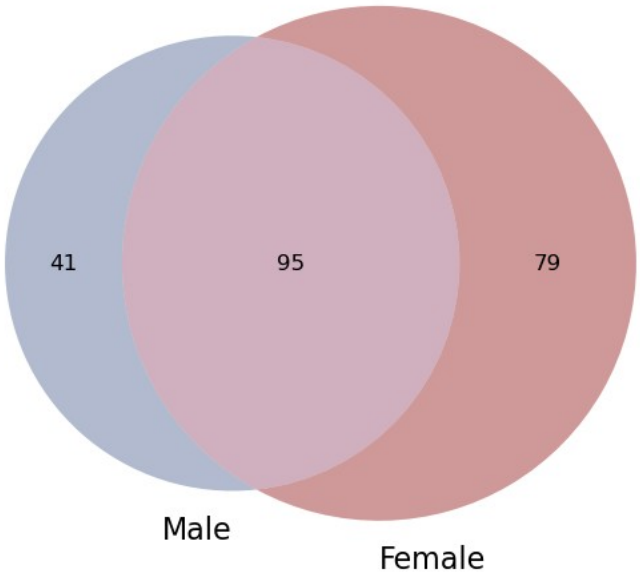


Figure S2: Venn plot of protein selected by male and female after Boruta selection. 41 proteins were only selected by male, 79 proteins were only selected by female and 95 proteins were selected by both sexes. Details of proteins selected were shown in ***Table S2.***

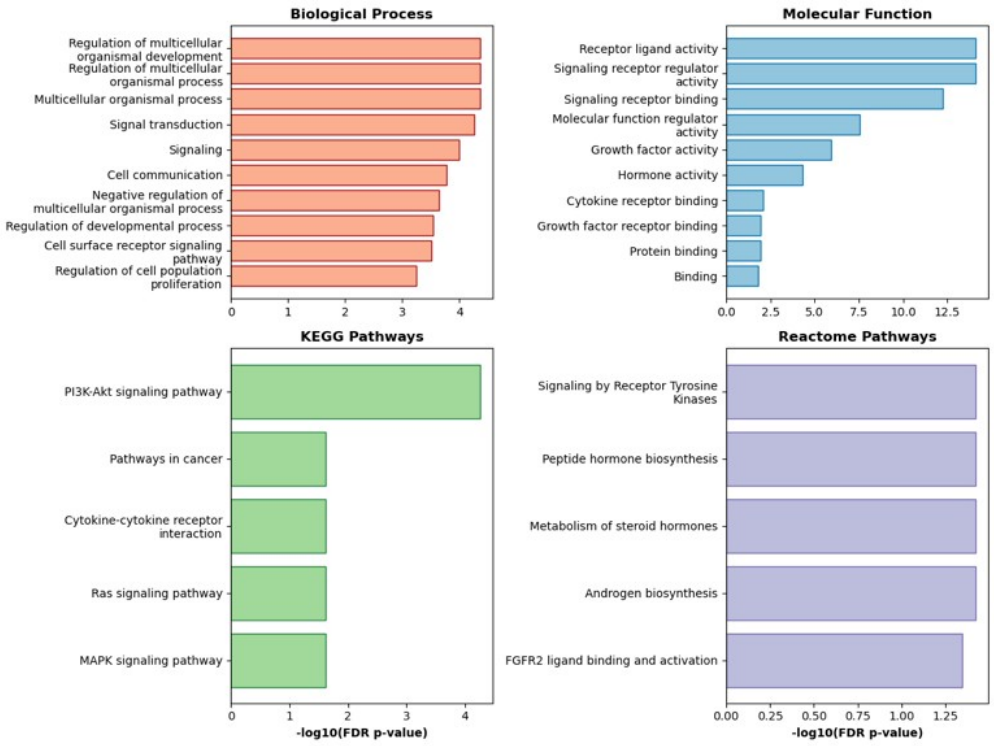


Figure S3: GO, KEGG, and Reactome pathway enrichment analysis in females. Enrichment pathways were only shown if FDR adjusted p-value is smaller than 0.05. If there were more than 10 pathways enriched, only top 10 pathways ranked by FDR adjusted p-value were shown.

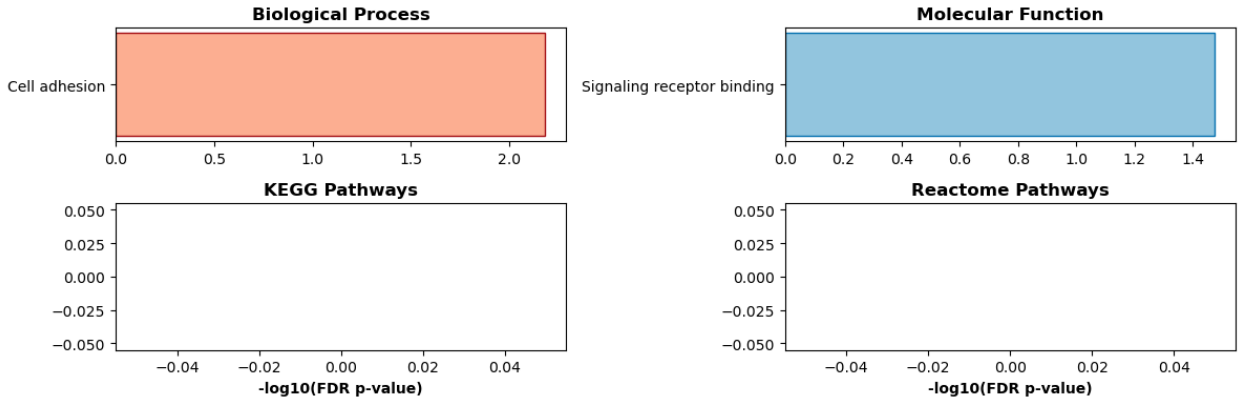


Figure S4: GO, KEGG, and Reactome pathway enrichment analysis in males. Enrichment pathways were only shown if FDR adjusted p-value is smaller than 0.05. If there were more than 10 pathways enriched, only top 10 pathways ranked by FDR adjusted p-value were shown.

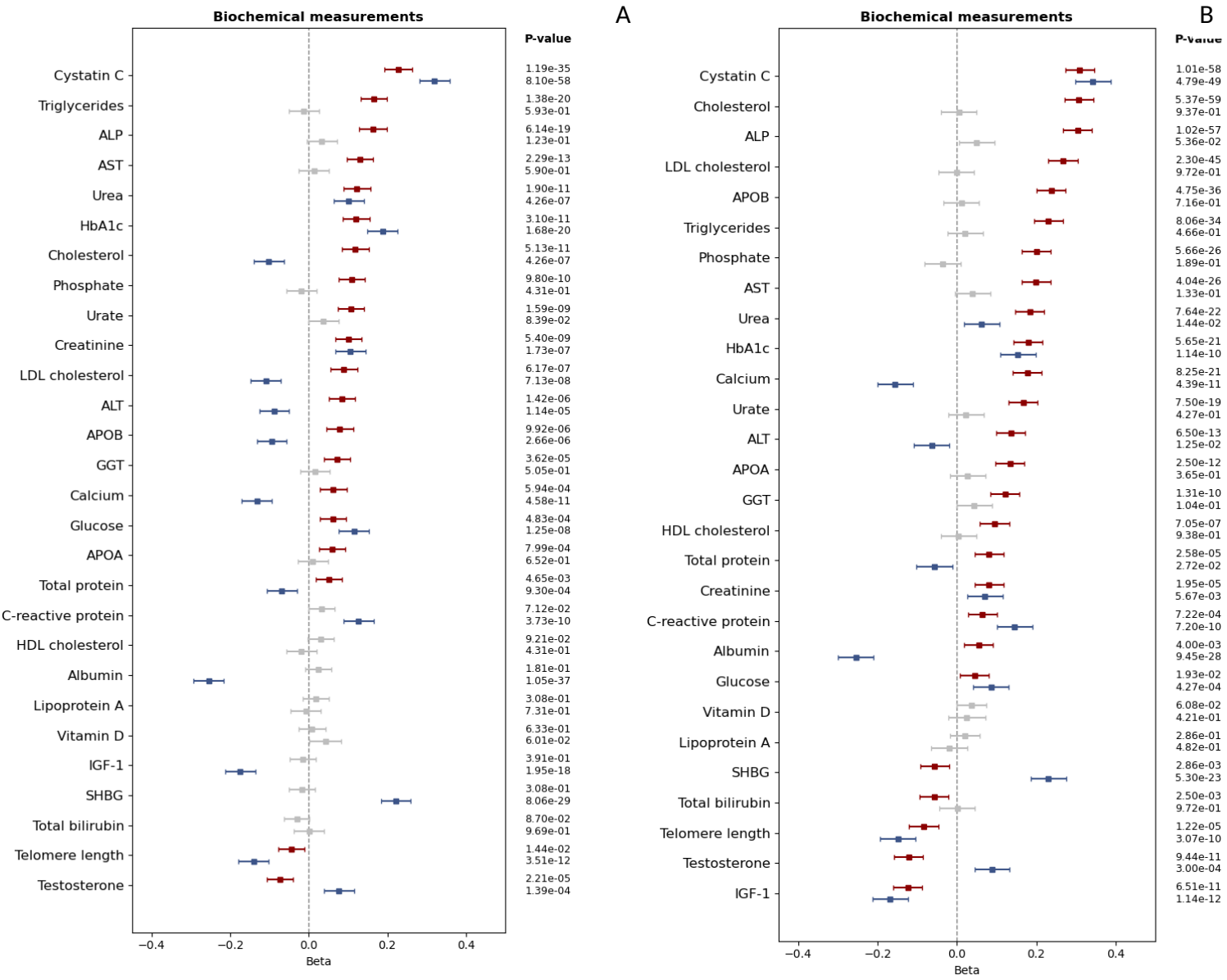


Figure s5: Sensitivity analysis on associations between biochemical measurement and protAgeGap. All linear regressions were performed between each exposure and protAgeGap adjusting for recruitment centre, ethnicity, education years, and Townsend deprivation index. Values were standardized if quantitative. Red shows the associations in females and blue shows the associations in males. Grey colour showed if the association was not significant after FDR correction. a) Sensitivity analysis on effect of menopause by adjusting additionally for menopause status. b) Sensitivity analysis on effect of drugs by removing participants who took anti-hypertensive and lipid lowering medications.

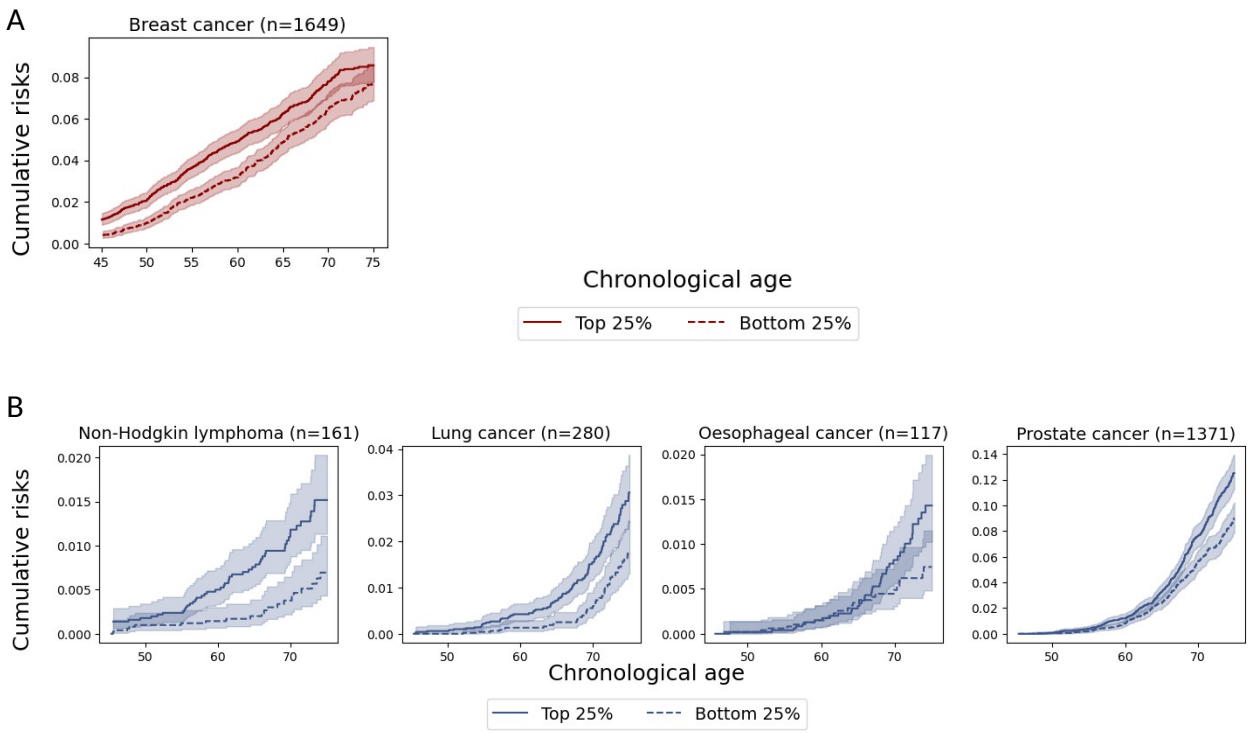
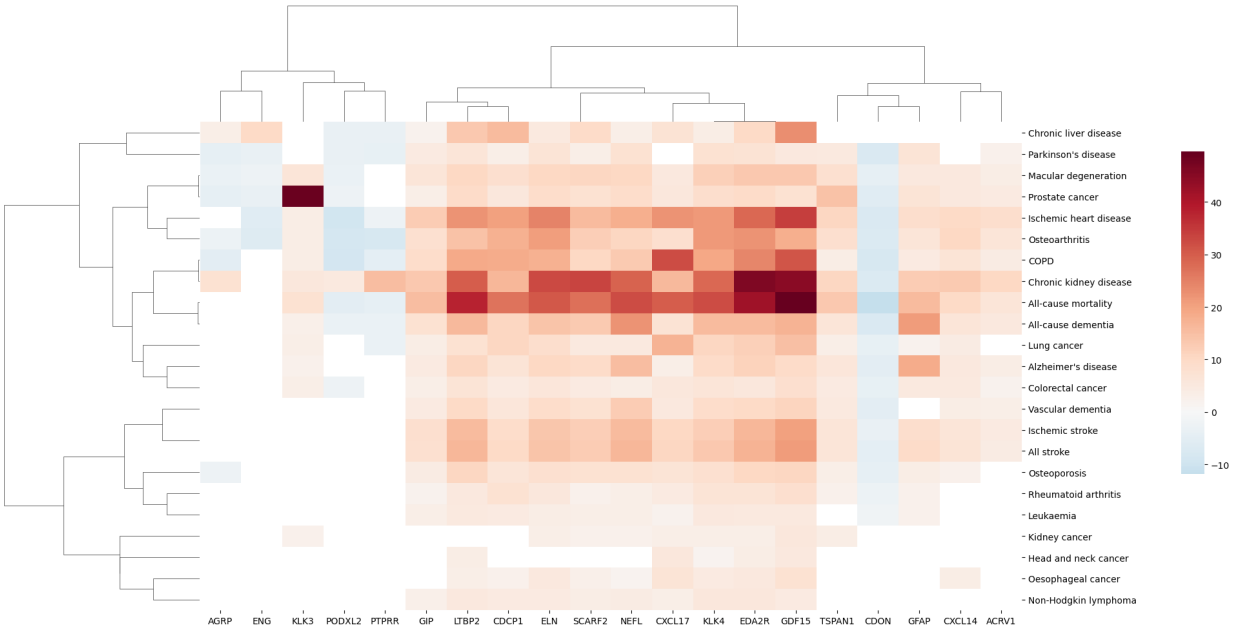
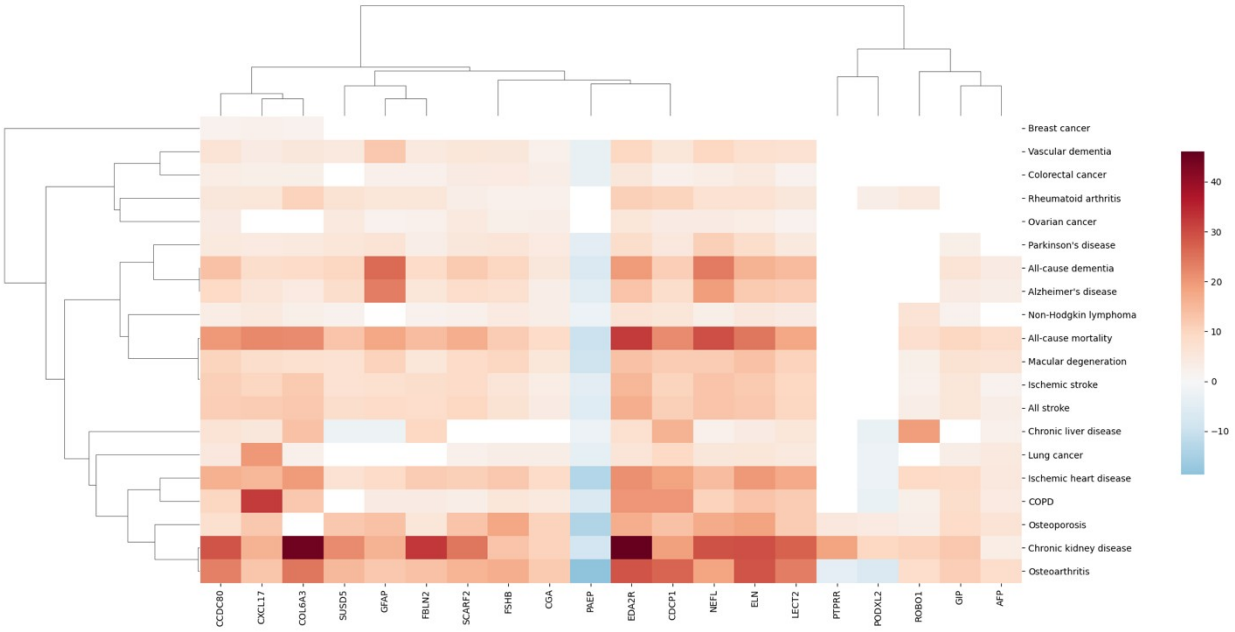


Figure s6* protAgeGap differentiates future risks of cancer. a)** shows cumulative incidence plot of top and bottom 25% of the protAgeGap in cancers significant in the cox model in female with 95% confidence interval shown as lighter shading. X-axis denotes the chronological age and Y-axis denotes the cumulative incidence. Cumulative incidence and number at risk at each age point is shown in ***Table s17 and ***Table s18***. (b) shows the cumulative incidence plot for cancers significant in the cox model in male. Cumulative incidence and number at risk at each age point is shown in ***Table s19*** and ***Table s20***.


Chronic liver disease
Parkinson's disease
Macular degeneration
Prostate cancer
Ischemic heart disease
Osteoarthritis
COPD
Chronic kidney disease
All-cause mortality
All-cause dementia
Lung cancer
Alzheimer's disease
Colorectal cancer
Vascular dementia
Ischemic stroke
All stroke
Osteoporosis
Rheumatoid arthritis
Leukaemia
Kidney cancer
Head and neck cancer
Oesophageal cancer
Non-Hodgkin lymphoma
AGRP ENG KLK3 PODXL2 PTPRR GIP LTBP2 CDCP1 ELN SCARF2 NEFL CXCL17 KLK4 EDA2R GDF15 TSPAN1 CDON GFAP CXCL14 ACRV1
40
30
20
10
0
−10


Breast cancer
Vascular dementia
Colorectal cancer
Rheumatoid arthritis
Ovarian cancer
Parkinson's disease
All-cause dementia
Alzheimer's disease
Non-Hodgkin lymphoma
All-cause mortality
Macular degeneration
Ischemic stroke
All stroke
Chronic liver disease
Lung cancer
Ischemic heart disease
COPD
Osteoporosis
Chronic kidney disease
Osteoarthritis
CCDC80 CXCL17 COL6A3 SUSD5 GFAP FBLN2 SCARF2 FSHB CGA PAEP EDA2R CDCP1 NEFL ELN LECT2 PTPRR PODXL2 ROBO1 GIP AFP
40
30
20
10
0
−10

A

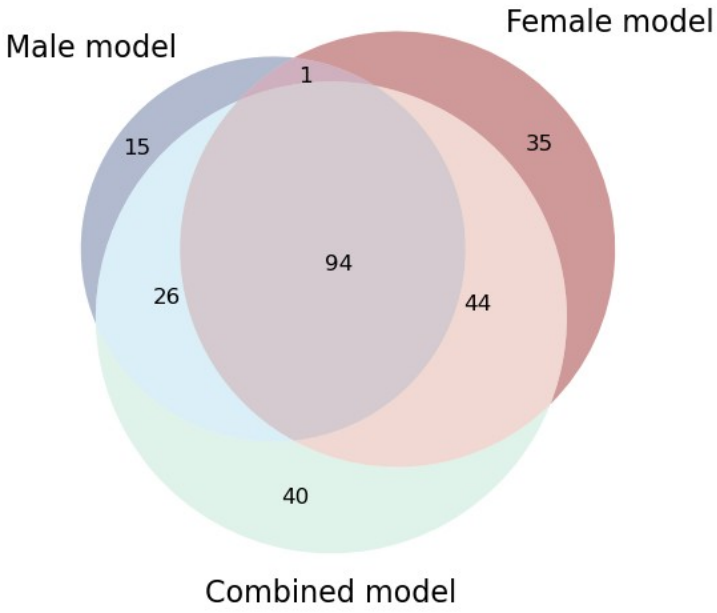


Figure S8: Venn plot of proteins selected by sex specific models and combined sex models.

Comparison of proteins selected by our previous combined sex model from Nature Medicine paper (204 proteins) and our sex specific models (male:136, female:174)

Chapter 5 Proteomic signatures of smoking and their associations with risk of incident diseases and mortality in diverse populations

Declaration

Content of this chapter is under reviewed by Nature Communications (Sihao, X., et al. Proteomic signatures of smoking and their associations with risk of incident diseases and mortality in diverse populations).

Introduction

Seventy years after the British Doctors Study first demonstrated an increased risk of death from lung cancer[204], myocardial infarction and chronic obstructive pulmonary disease[205] among smokers, smoking still accounts for about 14% of total deaths and 8% of disability-adjusted life years (DALY) globally[206]. Indeed, smoking is associated with higher risks of numerous diseases across different organ systems, including most cancers, respiratory diseases, cardiovascular diseases, and diseases of the liver, brain, kidney, and bladder, among others[207]. While smoking is often initiated in youth, smoking-related diseases typically manifest in middle age or later. Throughout the life course, smoking behaviour, including the type and quantity of tobacco consumed, may fluctuate, and many individuals attempt to quit smoking multiple times without success due to its addictive nature. Additionally, as smoking is becoming socially unacceptable in an increasing number of societies, the validity of smoking history is further

compromised[208,209]. Consequently, there is an urgent need to develop objective measures that assess smoking history and possible differences in smoking behaviour dynamics, including the recovery status of previous smokers. Such biomarkers will provide more accurate measurements of smoking exposure and will help to identify molecular mechanisms linking smoking with disease risks.

Exhaled carbon monoxide and plasma cotinine levels are widely used to validate smoking status,[210,211] but both capture smoking status within the last 24 hours, with limited ability to characterise long-term smoking habits and health effects and to reveal mechanisms of action associated with smoking. There have been major advances in our understanding of the epigenomic signatures of cigarette smoking during the last decade, particularly using smoking-related DNA methylation[212]. A meta-analysis of 16 epigenome-wide studies demonstrated smoking-induced DNA methylation of 2623 CpGs annotated to 1405 genes among current smokers[213]. Of these, 185 CpGs showed persistent alterations in previous smokers, and 36 CpGs showed persistent alterations for up to 30 years after smoking cessation. While DNA methylation markers reflect smoking history in smokers and previous smokers, DNA methylation-based smoking profiles have only been associated with the risks of a limited number of disorders, including chronic obstructive pulmonary disease (COPD), lung cancer, stroke and all-cause mortality, independent of smoking history[214–216].

Circulating proteins, however, may reflect not only smoking exposure as captured by DNA methylation but also the cumulative biological effects of responses to cigarette smoking (including oxidative stress or other mechanisms) or the biological responses to limit the hazards of smoking[77,212,217]. Thus, plasma proteomic profiles may enhance the ability to quantify the

direct effects and the physiological responses to smoking and ultimately predict the risk of subsequent morbidity and mortality among current and previous smokers.

Building on the chapters that established metabolomic (MetAgeGap) and proteomic (ProtAgeGap) clocks to quantify biological ageing and its sex-specific implications, this chapter aims to further dissect ageing-related pathology by focusing on smoking—a major modifiable risk factor with systemic impacts on aging[218]. While MetAgeGap and ProtAgeGap highlighted intrinsic biological processes underpinning differential ageing trajectories and disease susceptibilities in males and females, here we explore how an exogenous lifestyle factor like smoking is captured through the blood proteome and whether it reflects cumulative biological damage across organs. By developing a proteomic smoking index (pSIN), we assess not only the molecular imprint of current smoking but also the extent of recovery in former smokers, offering a complementary dimension to the intrinsic ageing clocks. Together, these chapters present an integrative framework linking internal ageing biology with external exposures to delineate individualised disease risk.

Methods

Study cohorts

UK Biobank (UKB)

The UKB population was described in Chapter 2. Smoking status stratified cohort population characteristics were summarized in **Table s1**. Smoking behaviours including smoking status, number of cigarettes per day, smoking start age, age stopped smoking, type of tobacco smoked, exposure to tobacco smoke at home and exposure to tobacco smoke outside were collected at

baseline using touchscreen-based questionnaires. Participants were classified as current smokers, previous smokers and never smokers. We classified "occasional smokers" as never smokers to ensure consistency in our comparison groups and to align with prior epidemiological studies. Occasional smokers comprised a very small proportion of the population and did not report a sustained smoking history. Given that our study focuses on the long-term molecular impact of regular smoking, we categorised occasional smokers with never smokers to avoid potential misclassification bias and to maintain a clear distinction between individuals with significant smoking exposure and those without. However, while building the model, an extremely pure never smokers set was used, excluding all occasional smokers and those with passive smoking exposures. Smoking packyears were calculated as the Number of cigarettes per day / 20 * (Age stopped smoking - Age start smoking). Passive exposure was defined as a binary variable, indicating whether an individual had been exposed to tobacco smoking either at home or outside. UKB blood biomarkers were measured using the non-fasting blood serum samples collected at baseline.

Missing data was imputed using a random-forest-based algorithm provided by R package missRanger[126] when used as a covariate in linear association models (Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years). Recruitment centre has a missingness of 0.00%, ethnicity has a missingness of 0.55%, alcohol frequency has a missingness of 0.30%, BMI has a missingness of 0.62%, IPQA activity group has a missingness of 19.93%, Townsend deprivation index has a missingness of 0.12%, education years has missingness of 2.02%. Imputation was performed with default hyperparameters using a maximum of 10 iterations and 200 trees. Linked hospital inpatient data, primary care data and

cancer register data were accessed from the UKB data portal on May 23 2023, with a censoring date of Oct 31 2022, Jul 31 2021, Feb 28 2018 for participants recruited in England, Scotland, and Wales respectively. The follow-up time is between 8 and 16 years. Mortality data and cause of death information were accessed from the UKB data portal on May 23 2023, with a censoring date of November 30 2022. The follow-up time is between 12 and 16 years. Methods and ICD diagnosis codes used to identify prevalent and incident chronic disease in UKB are shown in **Table s30**.

Proteomics assessment

Generation of the proteomic data and quality control steps were described in Chapter 2. To address the concern regarding the potential residual signal of age and sex in the proteomic expression data after linear regression, we performed additional analyses to rigorously evaluate the effectiveness of the regression process. Specifically, we trained gradient boosting models to assess whether any predictive signal for age or sex remained in the regressed data. For age, we trained regression models to predict age using each protein in the regressed dataset. The results showed a mean R^2 of -0.0042 (SD = 0.0089), indicating that the models performed no better than random chance. This suggests that the linear regression successfully removed age-related signals from the proteomic data. For sex, we trained classification models to predict sex using each protein in the regressed dataset. The models achieved a mean AUC of 0.51 (SD = 0.017), which is equivalent to random guessing. This confirms that sex-related signals were effectively removed by the regression process.

Genomics assessment

UK Biobank genotyping was conducted by Affymetrix using a bespoke BiLEVE Axiom array for ~50K participants and the remaining ~ 450K on the Affymetrix UK Biobank Axiom array. As the two arrays are broadly comparable with over 95% overlap in assessed gene variants, they were combined. All genetic data were quality controlled and imputed by UK Biobank. Detailed information on the genotyping process and technical methods are available online[219] Genetic data was phased before imputation with SHAPEIT3 followed by imputations using IMPUTE2. Details on genetic imputations are provided elsewhere[220].

China Kadoorie Biobank

The CKB population was described in Chapter 2. Smoking variables were collected by laptop-based questionnaires from baseline. Smoking status was collected as never smokers, occasional smokers, ex-regular smokers and smokers. We then re-define never-smokers and occasional smokers as never-smokers, ex-regular smokers as previous smokers and smokers as current smokers. This study restricted analysis on participants where the proteomic profile was measured with the Olink platform (n=3,977). Details of study design and methods were previously described[221].

Proteomics assessment

Proteomic assessments in CKB were described in Chapter 2. After regressing out age and sex in the CKB cohort, proteomic data was further processed by first rescaling to the value between 0 and 1 and then centring on the median.

Statistical analysis

A descriptive analysis of population characteristics was performed using the r package CBCgrps[127]. The study design and analysis pipeline are illustrated in **Fig 1**. Generation of the smoking status prediction model, feature importance assessment and feature selection steps were described in Chapter 2.

Proteomic Smoking INdex (pSIN)

The pSIN for the full UKB sample (n=43,914) was calculated using a robust methodology to mitigate the risk of overfitting. This process involved employing 5-fold cross-validation to ensure the reliability of the results. After identifying the best hyperparameters and selecting the proteins using the Boruta method, a gradient-boosting model was trained within each fold. Subsequently, the predicted raw score for the corresponding test set was generated. For binary classification tasks, this raw score corresponds to the log odds of the positive class (in this case, being a current smoker). The LightGBM model typically outputs raw scores (logits) in the range of approximately -10 to 10, where a score of 0 indicates a neutral prediction, corresponding to a 50% probability of being in the positive class (sigmoid(0) = 0.5). Scores closer to 10 indicate a high confidence prediction for the positive class (sigmoid(10) ≈ 0.99995, or ~99.995% probability), while scores closer to -10 indicate a high confidence prediction for the negative class (sigmoid(-10) ≈ 4.5×10^{-5}, or ~0.0045% probability). In our analysis, we set the classification threshold at a raw score of -1.29, which corresponds to the point where the false positive rate (FPR) is 0.05. This threshold was chosen to balance sensitivity and specificity in our predictions. pSIN higher than the threshold indicates a higher likelihood of being in the positive

class (smoker), with larger values reflecting greater confidence, while pSIN smaller than the threshold indicate a higher likelihood of being in the negative class (non-smoker), with more negative values reflecting greater confidence. These predicted raw scores from the test sets of each fold were then aggregated to create a comprehensive measure of smoking protein profiles for the entire population. This approach allowed for a more robust estimation of the impact of smoking on protein profiles across the UK Biobank cohort compared to using the model trained using 70% training data to calculate pSIN for the entire population. External validation in CKB was performed to further test the possibility of the overfitting problem. For external validation, the model with the optimised hyperparameter was trained in the UKB training dataset and was tested in the CKB. Performances of identifying current smokers from never smokers were compared.

Function annotation for proteins

Individual protein function annotation (GO: molecular functions) was extracted from the GO database using clusterProfiler v.4.2.2 in R[134]. Tissue-specific protein expression data was extracted from the Genotype-Tissue Expression (GTEx) project[222] database v.8 and the heatmap was plotted using the FUMAGWAS webtool. Values showed on the heatmap were the average of normalised expression per gene. Differential expression genes (DEG) were identified by a 2-sided t-test per tissue type versus all other tissue types. Genes with a Bonferroni corrected p-value < 0.05 and absolute log fold change > 0.58 were selected as DEG and were shown as red colour.

Association of smoking history, clinical biomarkers and risk factors, haematological measurements, exposome-wide association analysis with pSIN

To test the association of self-reported smoking habits, social-demographic factors, lifestyle factors, blood biochemistry biomarkers, and clinical risk factors with pSIN, generalised linear models from statsmodel v.0.14.0 package[128] were used.

For associations between smoking history and pSIN, models were adjusted for basic socioeconomic factors including recruitment centre, ethnicity, education years, and Townsend deprivation index.

For associations between blood biomarkers/clinical risk factors/haematological measurements to pSIN, continuous exposure variables, standardisation was applied before inclusion in the models. Associations were adjusted additionally for lifestyle factors including IPAQ activity group and alcohol intake frequency as they are known to cofound physiology status and liver damage. To explore the added value of pSIN compared to self-reported smoking status, sensitivity analysis was performed adjusting additionally for self-reported smoking status.

For exposome-wide association analysis (all available social-economic and lifestyle variables available in UKB), models were only adjusted for the most basic recruitment centre, ethnicity and smoking status to allow exploration of a wide range of potential associations between the exposome and pSIN without masking potential signals by controlling for too many variables.

P-values resulting from these analyses were corrected for FDR multiple testing.

Identification of genes influencing pSIN

Genome-wide association study was conducted using SAIGE software V1.09. For constructing a genetic relationship matrix (GRM) in step 1, we used the pruned genotype dataset. Genotype pruning was conducted in PLINK software using the 'indep-pairwise' option with an r^2 of 0.5, a window size of 1000 markers and a step size of 100 markers. We further used the 'LOCO= TRUE' option to construct the GRM. The GWAS analyses were adjusted for age, sex, batch effects, and 40 genetic principle components identified within UKB genotyping data.[220] Gene mapping was performed using the FUMAGWAS web tool where the maximum p-value for lead SNP was set as $5*10^{-8}$, and the maximum distance to genes was set as 10 kb. LDSC analysis was performed using online tools Complex-Traits Genetics Visual Labs (CTG-VL)[223] where summary statistics of GWAS against pSIN were correlated to publicly available GWAS results of 1,461 traits in the database. Significant correlations were identified if FDR adjusted p-value was smaller than 0.05.

Calculating the contribution of each category to pSIN

To investigate how much of the additional variance explained does each category of genetics, smoking history, social-demographic and lifestyle, and clinical biomarkers and risk factors has on pSIN, we constructed four gradient boosting models. These models were sequentially augmented by adding each category, starting with genetics, followed by smoking history, then social-demographic and lifestyle factors, and finally clinical biomarkers and risk factors. Each model was initially trained on the 70% training dataset, and subsequently, the variance explained was assessed using the 30% testing dataset using scikit-learn package[91]. The additional variance explained by each category was determined by subtracting the variance explained by the model that contains the information on this category from that of the previous

model that did not include this category. The same procedure was carried out for the whole population, but also in current smokers, previous smokers and never smokers to investigate how each category contributes to pSIN differently among smoking status stratified population. The additional variance explained was then plotted in a stacked bar chart for comparison.

Associating pSIN with future health-related outcomes

To test the association between pSIN and incident health outcomes, all prevalent cases were removed beforehand. Multi-variate Cox proportional hazard model provided by lifeline v.0.27.8 package[129] was used with a pre-set step size of 0.1. Survival outcomes were defined using follow-up time to the event and the binary incident event indicator. For all incident outcomes in the whole UKB population, two successive models were tested with an increasing number of covariates: model 1 did not adjust for any additional covariate as age and sex had already been regressed out in protein level; model 2 was adjusted for recruitment centre, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years. In current smokers, a third model was applied, which further adjusted for smoking pack years and the number of years smoked. For previous smokers, an additional third model was tested that included adjustments for smoking pack years and the number of years since cessation. P-values of the hazard ratio were corrected for FDR multiple testing. Forest plots were generated with a minimum sample size threshold of 80 to ensure adequate statistical power and reliable interpretation.

Cumulative incidence plots were generated utilising the KaplanMeierFitter function from the lifelines package[129]. Due to limitations in case numbers or at-risk numbers at both ends, the x-axis of the plot was constrained to the age range of 45 to 75. This adjustment ensured a more

focused visualisation of the cumulative incidence curve within a clinically relevant age range. P-values between cumulative incidence curves were calculated using a log-rank test with adjustment for FDR multiple testing.

Results

To model the relationships between the plasma proteome and cigarette smoking, we used a subset of the UK Biobank (UKB) with available measurements of the plasma proteome. The study population consisted of 43,914 participants, with 4,732 self-reported as current regular cigarette smokers, and 23,778 as never smokers. Baseline characteristics of study participants in the UKB are provided in **Table S1**. We observed significant age ($p=3.44*10^{-26}$) and sex ($p=1.64*10^{-71}$) differences between current and never smokers. Therefore, we regressed age and sex from the expression of each protein to eliminate their effects for downstream analyses. For external validation, we used a subset of China Kadoorie Biobank (CKB), in which plasma proteome was measured using an identical OLINK assay panel (**Table S2** for baseline characteristics).

Proteomic signatures of smoking

In a training dataset comprised of 70% of UKB participants, we developed a model using a gradient boosting tree model with plasma protein data to discriminate current smokers from never-smokers who reported no passive exposure to cigarette smoking. The model achieved a high area under receiver operating characteristic curve (AUC) after 5-fold cross-validation (CV) within the training dataset (mean AUC=0.96, SD=0.004; F1=0.98, SD=0.006; Averaged precision

(AP)=0.93, SD=0.006; Balanced accuracy(BA)=0.90, SD=0.006) (**Fig S1a**) and showed similar performance in the 30% UKB holdout test dataset (AUC=0.95, SD=0.004, F1=0.98, SD=0.004; AP=0.91, SD=0.006; BA=0.88, SD=0.006). After using the Boruta algorithm to identify all proteins that contribute to predicting smoking status, we found that 51 of 2,917 proteins discriminated current smokers and never smokers (98.3% reduction in protein size) with a mean AUC of 0.95 in 5-fold CV in the training dataset (SD=0.005)(F1=0.97, SD=0.005; AP=0.93, SD=0.004, BA=0.90, SD=0.004) (**Fig S1b, Table S3**) and an AUC of 0.95 (SD=0.004) in the test dataset (F1=0.97, SD=0.000; AP=0.91, SD=0.000; BA=0.89, SD=0.000) (**Fig 2a**). This model yielded a sensitivity of 84.6% and a specificity of 95%. To assess the added value of pSIN compared to using a single biomarker, we evaluated the predictive performance of the top three proteins ranked by SHAP value within the UKB test dataset (**Fig S1c**). ALPP and CXCL17 demonstrated relatively strong discrimination with AUCs of 0.88 and 0.87, respectively, while ACVRL1 showed lower performance with an AUC of 0.76. However, none of these individual proteins matched the predictive power of pSIN, which achieved an AUC of 0.95. These results highlight the advantage of using a multi-protein composite score over a single biomarker for more accurate stratification of current smokers and never smokers. The UKB model derived in the training set was subsequently externally validated in CKB, achieving an AUC of 0.91 (SD=0.000; **Fig 2a**) (F1=0.93, SD=0.000; AP=0.87, SD=0.000; BA=0.79, SD=0.000), and a sensitivity of 70.8% and specificity of 95%. We next calculated a predicted score for the overall UKB population including previous smokers defined as a Proteomic Smoking INdex (pSIN) (**Fig 2c)**. Although the curves largely overlapped, the mean pSIN was significantly higher in previous smokers (mean=2.8; SD=1.8) than in never smokers (mean=-3.7; SD=1.4, $p<2.2*10^{-308}$).

Overall, the proteins selected were predominantly involved in biological processes including epithelial cell proliferation (IL12B, IGFBP4, TNFSF12, ACVRL1, MMP12, PRL, KIT), regulation of immune system (IL12B, MMP12, SCGB1A1, MERTK, PDCD1, IL7R), cell growth (IGFBP4, TNR, ACVRL1, SCGB3A1, ISLR2), and T cell activation (IL12B, CD1C, SCGB1A1, KIT, IL7R). Tissue enrichment analysis was performed among selected proteins using RNA expression data from Genotype-Tissue Expression (GTEx) project[222]. Results indicated that many of these proteins were differentially expressed in tissues either directly exposed to or affected by smoking including lung, salivary glands, colon, esophagus and adipose tissues (**Fig s2a, b**). The top proteins contributing to the model prediction (**Fig 2b**) included ALPP, CXCL17 and ACVRL1, all of which showed higher levels in smokers. All three are predominantly expressed in the lungs, oesophagus and minor salivary glands. ALPP is synthesised in the liver by a metalloenzyme that catalyses the hydrolysis of phosphoric acid monoesters and was previously found to be associated with COPD and cancers[224]. CXCL17 is a mucosal chemokine that attracts immature dendritic cells and blood monocytes to the lungs and was previously associated with lung cancer[225]. ACVRL1 is an activin receptor-like kinase which has recently been shown to separately mediate transcytosis of low-density lipoprotein (LDL) into arterial endothelium[226]. Individuals with loss of function *ACVRL1* DNA variants have lower rates of progression of atherosclerotic vascular diseases[227].

Determinants and correlates of pSIN

pSIN is predominantly determined by smoking history, as the number of years smoked, cigarettes smoked per day, and pack years that were smoked were each positively correlated with pSIN scores in both current and previous smokers (**Table s4**). Among previous smokers, a longer duration since cessation was strongly associated with lower pSIN scores. Analysis of the

data in 5-year intervals, indicated a continuous decline in pSIN scores up to and exceeding 30 years of smoking cessation, suggesting ongoing recovery from effects of regular smoking-related effects (**Table s4**). In individuals who had never smoked, passive exposure to smoking was correlated with higher pSIN scores, demonstrating its high sensitivity. Similar trends were observed in CKB current smokers, where smoking duration, smoking exposure, and smoking intensity were all significantly and positively associated with pSIN. However, among former smokers, significant associations were observed only for years since smoking cessation and smoking duration, whereas associations with smoking intensity and smoking exposure were not statistically significant. This may be because there are more modifiers in previous smokers than in current smokers. (**Table s5**).

To understand the genetic architecture underlying the biological consequences of smoking pathology, as indexed by pSIN, I performed a genome-wide association study (GWAS). The GWAS was conducted in the UKB, using pSIN as the outcome variable. We applied a linear mixed model, adjusting for age, sex, genotyping batch effects, and the first 40 principal components to account for population structure. I identified 95 lead-independent genome-wide significant variants mapped to 129 genes, of which 8 *(ALPP*, *CST5*, *IL12B*, *ACVRL1*, *IL7R*, *SCGB1A1*, *NCAM1*, *and ICAM5)* encoded one of the selected proteins in the pSIN score. Of the 95 significant lead-independent variants, 32 were cis-pQTLs mapping to 16 genes, with 8 of these genes encoding proteins included in the pSIN model. Additionally, 7 variants were identified as trans-pQTLs mapping to 366 genes, of which 21 encoded proteins were selected in the pSIN model. Of the 129 genes associated with pSIN, 10 genes were previously identified as GWAS smoking loci[212,228], and 54 genes were previously found in epigenetic studies[77,78,213], 75 (58%) were novel *and have previously been reported to be associated with* body mass index (BMI), diabetes, cancer development, and immunological/haematological traits including lymphocyte counts, eosinophil counts, and white blood cell counts (**Fig S3**, **Table s6-s9**). Importantly, haematological

measurements also showed strong associations with pSIN in the UKB before and after adjusting for smoking status(**Fig S4A, B**). Genetic correlations using LD score regression (LDSC) analysis highlighted strong genetic correlations of pSIN with current smoking (r=0.78, $p=1.08*10^{-98}$), never smoking (r=-0.65, $p=1.05*10^{-59}$), maternal smoking around birth (r=0.66, $p=2.08*10^{-42}$), cannabis use (r=0.51,$p=1.50*10^{-27}$), smoking-related lung disorders (lung cancer (r=0.71, $p=1.14*10^{-06}$), COPD (r=0.54, $p=1.04*^{10-31}$), depression (r=0.28, $p=9.46*10^{-16}$), ADHD (r=0.47, $p=8.26*10^{-16}$), multi-site chronic pain (r=0.31, $p=7.31*10^{-16}$), obesity and fat distribution (BMI r=0.33, $p=1.832*10^{-22}$), % of leg body fat (r=0.33, $p=1.05*10^{-23}$) and diabetes (r=0.27, $p=4.08*10^{-10}$) among others (**Table S10**).

I then studied the relation of pSIN with all available environmental exposures in the UKB (i.e., the exposome). The exposome-wide analysis was conducted using generalised linear models, where each factor was analyzed separately for its association with pSIN. Adjustments were made for the recruitment centre, ethnicity, and smoking status. Favourable socio-economic indicators and healthier lifestyle choices, including better housing conditions, lower Townsend deprivation index, higher household income, higher levels of education, high consumption of fruit and fibre intake, higher levels of physical activity and social interactions were associated with lower levels of pSIN (**Table S11**). Conversely, unhealthy lifestyle choices, poor environment and mood disorders including high salt intake, high consumption of red/processed meat, and coffee, alcohol consumption, exposure to air pollutants (PM10, PM2.5, NO2, and NO), maternal smoking during pregnancy, evening chronotype, sleeping greater than 9 hours or less than 7 hours, or feeling of tiredness or low mood were associated with higher pSIN scores. These findings suggest that beyond smoking behaviour itself, a range of socio-economic, dietary,

environmental and behavioural factors each contributed independently to smoking-related damage as reflected by effects on pSIN.

I next explored how different clinical biomarkers and risk factors that were read-outs of the exposome (e.g., obesity as a read-out of diet and physical activity) and those indicative of overall health status at baseline (e.g., lipid levels, creatinine, and blood pressure) influenced the pSIN. Firstly, blood biochemistry profiles revealed significant correlations with pSIN. These included markers of inflammation such as glycA (glycoprotein acetyl, antichymotrypsin) and C-reactive protein (CRP), indicators of glucose metabolism such as higher HbA1c levels, and biomarkers related to kidney and liver function. In addition, significant associations were observed with sex hormones, lower vitamin D levels, and aging-related biomarkers like telomere length and insulin-like growth factor 1 (IGF1) (**Fig S5a, Table S12**). Additional analyses of association with lower clinical function indicated that self-rated health status, older facial ageing, but also to clinically assessed ones including systolic blood pressure, heel bone density, fluid intelligence, and lung function, were positively correlated with pSIN (**Fig S5b, Table S12**). Moreover, baseline disease including type 2 diabetes and arterial stiffness were strongly correlated with higher pSIN scores (**Table S12**). The findings of this study highlight the substantial and diverse impact of smoking physiological systems, as reflected by the associations of pSIN with a wide array of clinical biomarkers and health indicators at baseline. I conducted additional sensitivity analyses where the associations between pSIN and clinical biomarkers were further adjusted for smoking status (**Fig S6a, b**). I observed that while the effect sizes of biomarkers such as HbA1c, GlycA, and Triglycerides were attenuated, they remained statistically significant. However, biomarkers such as APOA, HDL cholesterol, and Creatinine became non-

significant after this adjustment. Similarly, the association between pSIN and clinical risk factors, such as poor self-rated health, was weakened when adjusting for smoking status while the associations with other risk factors were maintained at a similar level.

To explore the relative contributions of the genome, smoking history and exposome to pSIN, I used a gradient-boosting model. I first regressed GWAS significant SNPs against pSIN and calculated the variance explained (**Fig 2d**). In our analysis, genetic factors explained a larger proportion of variance in pSIN never smokers (2.7%) and previous smokers (1.6%) compared with current smokers (0.3%). However, overall, genetic factors had a minimal contribution to pSIN. I next estimated the contribution of the exposome, starting with the contribution of smoking history, then added sociodemographic and lifestyle factors, and finally added clinical biomarkers and risk factors to the models, calculating the additional proportion of variance explained by each factor (**Fig 2d**). Smoking history accounted for the largest proportion of pSIN variance in both the overall population (65.8%), current (52.1%) and previous smokers (16.4%); passive smoking explained 0.2% of the variance in never-smokers. Sociodemographic and lifestyle factors contributed most of the variance in never-smokers (4.8%), with factors such as air pollution, diet, and alcohol consumption playing a substantial contribution to levels of pSIN in this subgroup. Lastly, clinical biomarkers and general health indicators provided a comparable amount of additional information across current (2.0%), previous (1.9%) and never (2.0%) smokers. These findings underscore the multifactorial nature of pSIN, it is predominantly determined by smoking history in the population overall and in current and previous smokers but also genome and exposome are related to pSIN, even in non-smokers.

Associations of pSIN with risk of morbidity and mortality

Over a mean follow-up time of 13.3 years (SD=2.2) in the UKB, 4,615 deaths were observed. I tested the association of pSIN with 27 major disease outcomes and all-cause mortality. In the overall UKB population, pSIN was significantly associated (FDR<0.05) with 19/28 outcomes independent of the confounding factors (**Fig S7**, **Table S13**). These included lung cancer (HR=1.97, CI:1.83, 2.11), COPD (1.72, 1.67, 1.78), and head and neck cancer (1.64, 1.44, 1.86) ranking as top 3 based on their hazard ratios per SD increase in pSIN. pSIN was able to capture different hazard ratio of subtypes of incident inflammatory bowel disease with higher risks in Crohn's disease (1.23, 1.04, 1) comparing to ulcerative colitis (451.14, 1.00, 1.29) (***Fig s8***). Interestingly, pSIN displayed a protective effect for Parkinson's disease (0.84, 0.75, 0.94), one of the few common disorders with a lower risk in smokers[229]. pSIN was also significantly associated with mortality (1.32, 1.28, 1.36). Associations of individual proteins with incident health outcomes are shown in **Fig S9.** I further validated the association with lung cancer, COPD, any vascular disease, any respiratory disease, ischemic stroke, all stroke, ischemic heart disease and mortality in the CKB, all of which stayed significant after adjusting for confounding factors (***Table s14***).

For mortality and the 18 diseases that were significantly associated with pSIN, I further tested if participants in the top, median and bottom quartiles of the pSIN exhibited divergent cumulative incidence in each outcome (**Fig 3**). We found that 11 diseases showed more than 2-fold higher cumulative incidence when comparing the top and bottom quartiles of pSIN at age 75 years including lung cancer, peripheral artery disease, COPD, head and neck cancer, oesophagal

cancer, congestive cardiac failure, all-cause mortality, chronic liver disease, ischemic stroke, all stroke, and bladder cancer. The cumulative incidence and number of risks in each age group are shown in **Table S15** and **Table S16**.

To evaluate the additional information provided by the pSIN compared to self-reported smoking habits, I analysed their associations with current smokers and previous smokers. In current smokers, pSIN was significantly associated with 6/15 major smoking-related health outcomes (with a more than 80 incident cases cutoff) including lung cancer, peripheral artery disease, COPD, all-cause mortality, osteoporosis, and ischemic heart disease after adjusting for lifestyle factors and smoking packyears (**Fig 4a, Table S17**). Further, participants in the highest quartile of pSIN had greater than 2-fold higher cumulative incidence at age 75 years compared to those in the lowest quartile for all evaluated outcomes, except for ischemic heart disease, which showed a 1.59-fold higher risk (**Fig S10a**). Cumulative incidence and number at risk at each age are provided in **Tables S18** and **S19.**

Among previous smokers, pSIN was significantly associated with higher risks of 12 disease outcomes including COPD, lung cancer, peripheral artery disease, and mortality among others after adjusting for pack-years smoked and number of years since cessation (**Fig 4b, Table S20**). Participants in the highest quartile of pSIN had greater than 2-fold higher cumulative incidence at age 75 years compared to those in the lowest quartile in 2 of the morbidities including lung cancer (3.43-fold higher risk) and COPD (2.06-fold higher risk). The cumulative incidence associated with pSIN for each disease is shown in **Fig S10b**, **Table S21** and **Table S22**.

As a sensitivity analysis, I examined whether conventional smoking history variables—including smoking status, pack-years (in both current and previous smokers), duration of smoking (in current and previous smokers), number of cigarettes smoked (in current and previous smokers), years since cessation (in previous smokers), and passive smoking exposure (in never smokers)—retained predictive value after adjustment for pSIN (***Fig s11***). Among current smokers, adjusting for pSIN substantially attenuated the HRs of incident diseases, with the most pronounced reductions observed for smoking status and smoking duration, where nearly all associations lost statistical significance. Among previous smokers, HRs also decreased after adjustment for pSIN, but significant associations persisted, particularly for lung cancer, COPD, and PAD. These findings indicate that while pSIN captures molecular insights into smoking pathology and predicts incident smoking-related diseases, conventional questionnaire-based smoking history still contains residual information not fully captured by proteomic markers.

Use of pSIN to differentiate the recovery status of previous smokers

Analyses of changes in mean pSIN levels by years since smoking cessation among previous smokers indicated that it took only 2 years for the pSIN to decline to a threshold that differentiated current from never smokers at FPR of 0.05 (**Fig s12**). However, the variation of pSIN in each individual with the same cessation years is large and for a substantial number of people, the pSIN remains high despite quitting smoking. Indeed, at least 10 years were needed for more than 80% of the previous smokers to have a pSIN below the threshold, consistent with findings in the British Doctors study[230] (**Fig 5a**). Additionally, the mean levels of pSIN of previous smokers plateaued, approaching the levels seen in never smokers after approximately 25 years

after cessation of smoking (**Fig s12**). Using this cut-off, we identified two clinically relevant subgroups among previous smokers: (1) those whose proteomic profile was similar to that of the current smokers and (2) those whose proteomic profiles were similar to the never smokers – henceforth referred to as the 'recovered' group (**Fig 5b**; n=2,576/15,404).

To determine whether this differentiation between these two groups of previous smokers was clinically relevant, I evaluated the disease and mortality risk differences between the two groups, adjusting for lifestyle factors, smoking packyears and smoking cessation years. Within the group of previous smokers, the recovered group had significantly lower risks for 10 of the disease outcomes, including COPD (HR=0.33; CI:0.29, 0.38), lung cancer (0.37, 0.27, 0.52) and all-cause mortality (0.52, 0.46, 0.59). (**Fig S13, Table s23**). Importantly, for most diseases, cumulative lifetime risks for recovered groups were comparable to those for never smokers (**Fig 5c**).

Conversely, among those not recovered from smoking damage based on pSIN, their lifetime cumulative incidence curve showed no significant differences compared with current smokers for diseases such as asthma (p=0.72), chronic kidney disease (p=0.62), chronic liver disease (p=0.13), and congestive cardiac failure (p=0.12). However, for most respiratory and cardiovascular diseases, their risks were significantly lower among previous smokers compared to current smokers. Notably, conditions such as lung cancer and peripheral artery disease showed more than a 50% reduction in cumulative incidence by age 75 (**Tables S24, S25**).

In addition, I also identified current smokers whose pSIN scores were similar to those of never smokers (**Fig S14a**, n=737/4,732). After adjusting for lifestyle factors and smoking packyears,

this population had significantly lower risks for peripheral artery disease (HR=0.13, CI:0.03,0.53), lung cancer (0.15, 0.05, 0.46) and COPD (0.25, 0.16, 0.39) in addition to all-cause mortality (0.53, 0.38, 0.73) (**Fig S14b, Table S26**). **Fig S14c** shows that when current smokers had similar pSIN to never-smokers, they exhibited cumulative disease risks comparable to those of never-smokers for all tested health-related outcomes except for COPD (risk at 70 years: 2.5%, $p=9.5*10^{-5}$) and all-cause mortality (risk at 70:6.4%, $p=4.59*10^{-2}$) which were associated with a lower absolute risk compared to high-risk current smokers (**Table S27, S28**).

Discussion

Using a machine learning analysis of plasma proteomic data in UKB, I generated the pSIN score and showed that it differentiated current smokers from never smokers with high accuracy and validated this score in CKB. I found that pSIN levels in the population and cigarette smokers were determined mainly by smoking history but also by genetic, environmental, and clinical risk factors and blood-based biomarkers of morbidity and mortality. Higher levels of pSIN were strongly associated with mortality, and 18 out of 27 major diseases, and the associations were independent of self-reported smoking exposure in smokers. Among previous smokers, the risks of morbidity and mortality were attenuated compared with current cigarette smokers but pSIN levels predicted the residual risk of mortality and morbidity.

The findings from the present study corroborate and expand upon previous research studies, particularly highlighting associations of smoking with COPD, peripheral artery disease and lung cancer[231,232]. Comparisons with the available worldwide evidence on this topic indicated that 36 of 51 smoking-associated proteins identified in the present study were independently

associated with smoking in genome[212,228,233], epigenome[77,78,213], transcriptome[217,234,235] (**Fig S15**) or individual protein-based studies (**Table s29**). Most of the top 20 proteins contributing to pSIN align with these findings, with novel associations identified for the first time such as DKK4, SLITRK1, KLK13, and CA4. Assessing the predictive performance of our pSIN model against previous epigenomic studies as the gold standard, I achieved comparable discrimination power for smoking status. For instance, Sugden et al[212] achieved an AUC of 0.93 and 0.81 based on 2,623 CpGs; Bollepalli et al[77] achieved an average sensitivity of 0.81 with a specificity of 0.85 when using 121 CpGs; and Maas et al[78] achieved a 5-fold CV AUC of 0.897 using 13 CpGs. In contrast, our pSIN model demonstrated a robust performance with a 5-fold cross-validation AUC of 0.96 within the training set and an AUC of 0.95 in the test dataset. Furthermore, for identifying current smokers with high specificity, pSIN outperformed existing models with a sensitivity of 0.85 and specificity of 0.95. We find that pSIN is genetically correlated with maternal and passive smoking. However, in never smokers their contribution to the pSIN levels is small (0.2%). Interestingly, beyond matched performance, I also observed a similar distribution of the DNA methylation smoking score built by Elliott et al[236] in current, previous and never smokers (**Fig s16**; unpublished) compared with pSIN in UKB.

Beyond the new proteins associated with smoking and the high levels of accuracy of the models, the major novelty of the present study was the link between the composite effects of these proteins with individual risk of morbidity and mortality. Previous omics studies, which have been limited by sample size or short duration of follow-up and primarily focused on genetic correlations with diseases, were unable to fully assess associations of selected omics markers with incident disease outcomes. In contrast, the present study provided a more

comprehensive analysis which included an investigation of associations of pSIN with 27 major incident diseases and with all-cause mortality. I demonstrated that pSIN effectively differentiated risks of disease outcomes in the general population and provided additional predictive value for assessing risks of morbidity and mortality among both current and previous smokers, independent of confounding factors and smoking history. The present study highlights the clinical relevance of pSIN for predicting disease risks, providing an objective, personalized, measure of smoking history that can be translated into a risk estimate of various diseases and death. For example, previous smokers who have similar pSIN to never-smokers, are found to have significantly reduced incident morbidity and mortality risks, potentially comparable with those of never-smokers. This suggests that pSIN can serve as an objective test to assess the magnitude of recovery from smoking-related damage among previous smokers. I also demonstrated that for diseases like lung cancer, and peripheral artery disease, the lifetime risks declined by greater than 50% immediately after smoking cessation, even for the highest-risk group of previous smokers compared to current smokers. Importantly, I found an inverse association between pSIN and Parkinson's disease. The reliability of this inverse association was demonstrated in the 60-year follow-up of the British Doctors' Study[237] with molecular evidence suggesting that nicotine and related chemicals associated with smoking may reduce MPTP-induced dopaminergic toxicity or inhibit the enzymatic oxidation of dopamine[238,239].

Comprehensive analyses of omics and questionnaire data in UKB enabled me to explore the associations of genetic and environmental factors associated with pSIN. Comparisons of current smoking with behaviour and environmental factors indicated that the impact of the genome on

pSIN was limited in smokers and previous smokers and the related genetic correlation analysis indicated a substantial overlap in the genes associated with pSIN and smoking habits. This suggests most of the disease risks associated with smoking are modifiable and preventable. Importantly, the findings that some participants self-reported as never-smokers exhibited pSIN levels comparable with those of current smokers have potentially important health relevance. I hypothesise that this group may include individuals who did not accurately report their smoking status perhaps reflecting a "social desirability" bias[208]. Indeed, in CKB the self-reported smoking history was validated by measurements of exhaled carbon monoxide (CO) among all participants at baseline, 20.3% of this population who reported as never smokers had a CO level exceeding 10 ppm, a widely recognised threshold to indicate smoking[211]. Alternatively, they may be exposed to other factors that are associated with the same protein pathology as pSIN. Based on the genetic correlation, BMI emerged as a risk factor for diseases that share a common pathogenesis with smoking, potentially involving processes such as epithelial cell proliferation, regulation of the immune system, cell growth and T-cell activation. Additionally, our study demonstrated robust associations between pSIN and maternal smoking, supported by both direct questionnaire-based assessments and genetic correlation analyses. Previous reports have consistently demonstrated that cigarette smoking during pregnancy has been linked with disease outcomes in offspring, including neurodevelopmental and behavioural issues, obesity, hypertension, type 2 diabetes, and impaired lung function[240] as well as epigenetic changes[241]. In addition, the findings of the present study provide long-term molecular evidence of these effects by assessment of pSIN. Last but not least, while it is well established that air pollution, especially fine particles, results in higher risks of lung cancer, COPD and cardiovascular

diseases[242,243], the present study provides details of biological instruments to quantify the long-term effects of smoking in large-scale blood-based population studies.

Despite being the largest study conducted to date, the present study was constrained by a limited number of proteins compared with the total number of known proteins. Secondly, Olink measurements are relative quantifications suitable for large cohort studies due to high throughput and cost considerations, any translation of such assays into clinical practice would require replication of these findings using absolute quantification. Nevertheless, the present study provides a comprehensive analysis of the associations of pSIN with health outcomes. However, because the study was limited only to the subset of UKB where plasma proteome was measured, the study had only a limited statistical power to detect associations between pSIN and common diseases and to examine associations between pSIN and less common diseases including cancers at individual sites that would require analyses incorporating specific diseases in multiple cohorts. Further, the causal relationships between environmental exposures and smoking are inherently complex. While our analysis aimed to leverage protein signatures within pSIN to understand the downstream pathology of smoking, it is essential to acknowledge that some associations identified may not be directly attributable to smoking behaviour. Instead, these associations could arise from independent effects of environmental exposures and smoking-related behaviours on proteins that are part of the pSIN, thereby confounding the interpretation of smoking-specific effects. This limitation underscores the challenges of disentangling causal pathways in observational data and highlights the necessity for cautious interpretation when utilising scores based on protein networks to elucidate aetiological

mechanisms. Future studies should consider complementary approaches, such as Mendelian randomization or experimental validation, to disentangle these independent effects and strengthen causal inferences.

I acknowledge that while my study provides a strong foundation for understanding the molecular signatures of smoking and its associations with disease risk, further steps are needed to explore its potential clinical applications. To bridge the gap between discovery and translational research, I propose that future studies should evaluate the feasibility of integrating pSIN into primary care settings. This could involve testing its utility in a clinical trial designed to assess whether proteomic-based risk stratification improves early detection and targeted interventions for high-risk individuals. Key considerations include the cost-effectiveness of measuring the 51 proteomic biomarkers using Olink or alternative platforms, as well as the logistical challenges of incorporating these assessments into routine clinical workflows. Additionally, implementing pSIN in practice would require careful evaluation of whether it provides actionable information beyond traditional risk factors, and whether its use could justify the costs associated with proteomic profiling and the collection of comprehensive clinical, epidemiological, and environmental data. While high-throughput proteomics is currently expensive, costs are expected to decline with technological advancements, potentially making it a viable tool for personalised risk assessment in the future.

Overall, the present study demonstrated that the blood proteome is a powerful tool to measure downstream molecular changes associated with cigarette smoking and a reliable measure to quantify risks of smoking-related diseases. This study adds to the available evidence linking

molecular signatures of smoking with the risk of common diseases associated with smoking and all-cause mortality and allows informing the hazards of current and previous smoking for risk of morbidity and mortality in population studies.

Figures

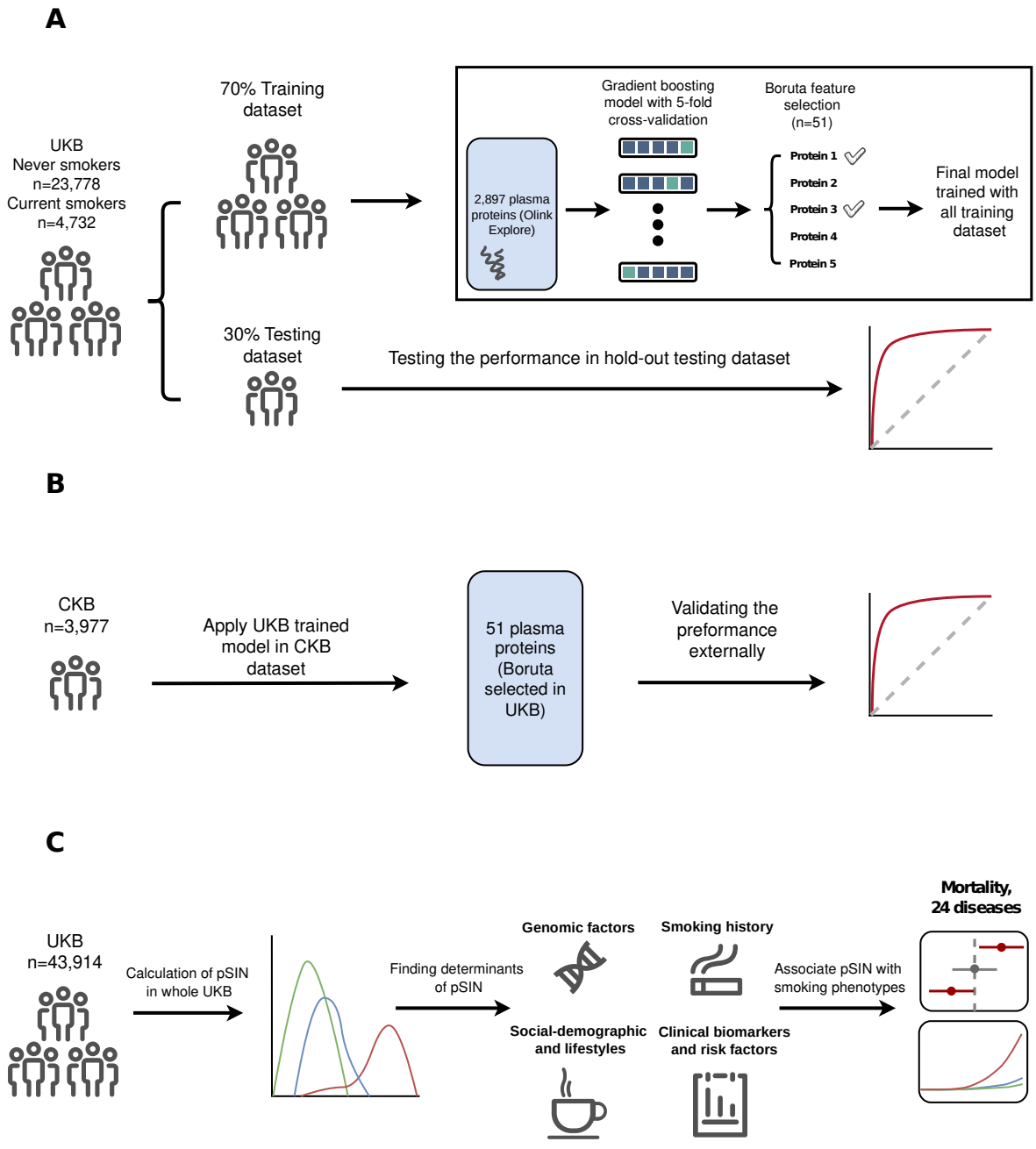


Figure 1: Overview of the study design and analytic approach. (a) gradient boosting classification model was built in the 70% randomly selected UKB population differentiating current smokers and never smokers. Boruta feature selection algorithm was then used to select only relevant features for downstream analysis. **(b)** The model trained in UKB training dataset was further validated externally in CKB male population. **(c)** proteomic Smoking INdex (pSIN) was calculated for the whole UKB cohort. Smoking behavioural, genomic, and exposome determinants of pSIN was

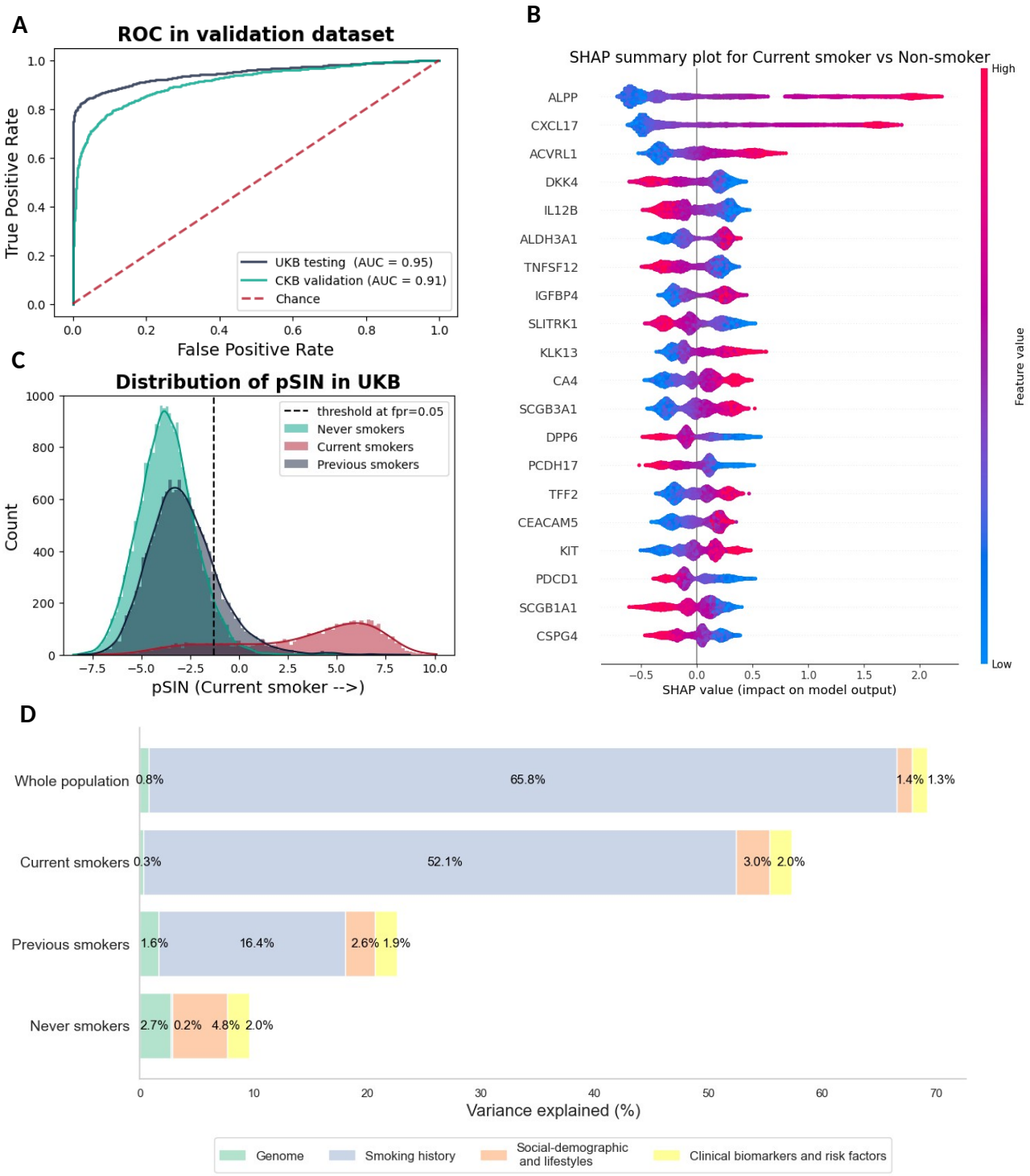


Figure 2: Protein profile differentiating current and never smokers and its determinants. (a) shows the performance of the model in UKB 30% left-out testing dataset and CKB external validation dataset. **(b)** shows summary plot for SHAP value of the top 20 selected proteins. Each dot denotes a participants, colour of the dots denotes the protein expression level and X-axis denotes its contribution to the model decision. Proteins were ranked by mean of the absolute SHAP value. **(c)** shows the distribution of pSIN in the whole UKB population. Dotted line denotes the cut-off value at FPR 0.05 when differentiating current and never smokers. **(d)** Four gradient boosting models were employed to assess the contribution of intrinsic and extrinsic factors to pSIN. The additional variance explained by each category was calculated by subtracting the variance explained by the previous model from the variance explained by the current model.

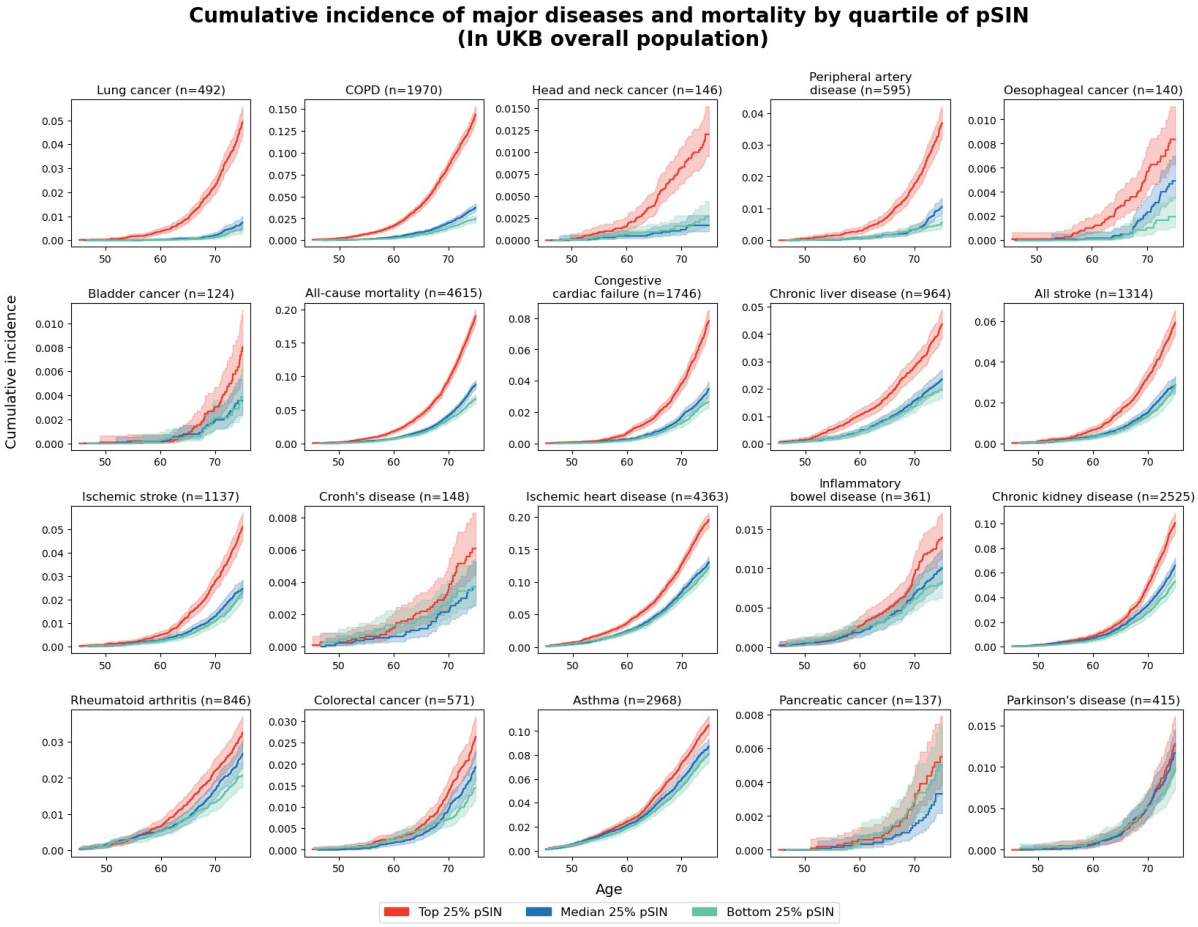
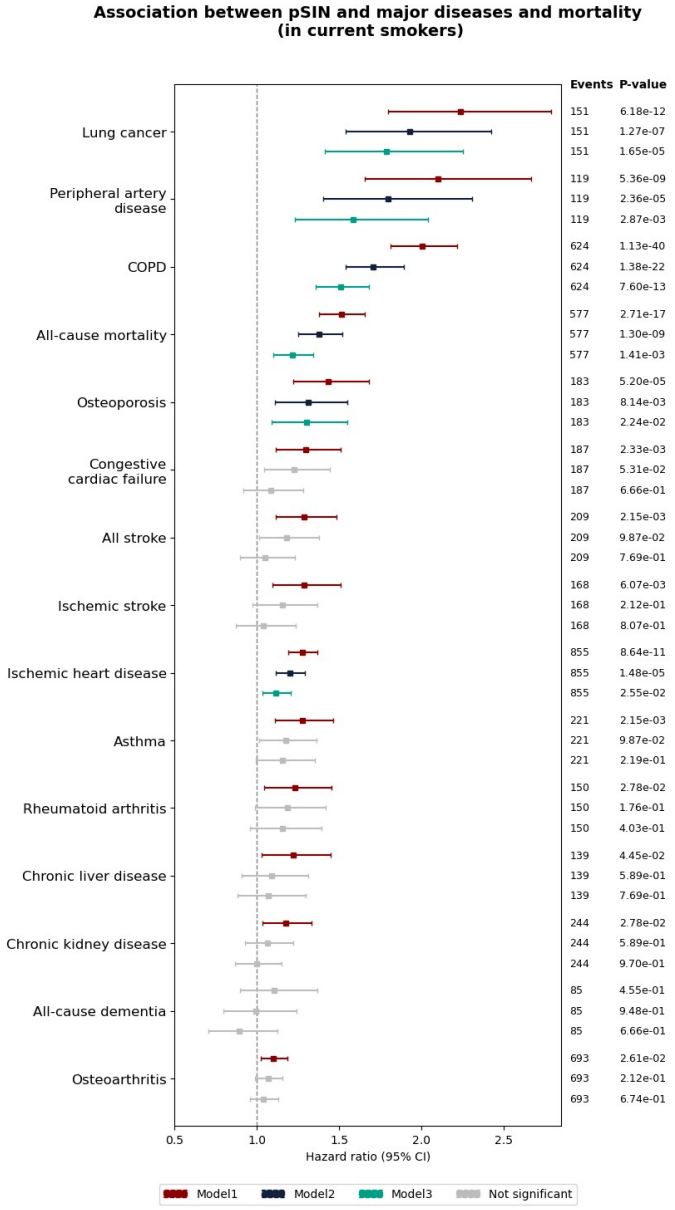
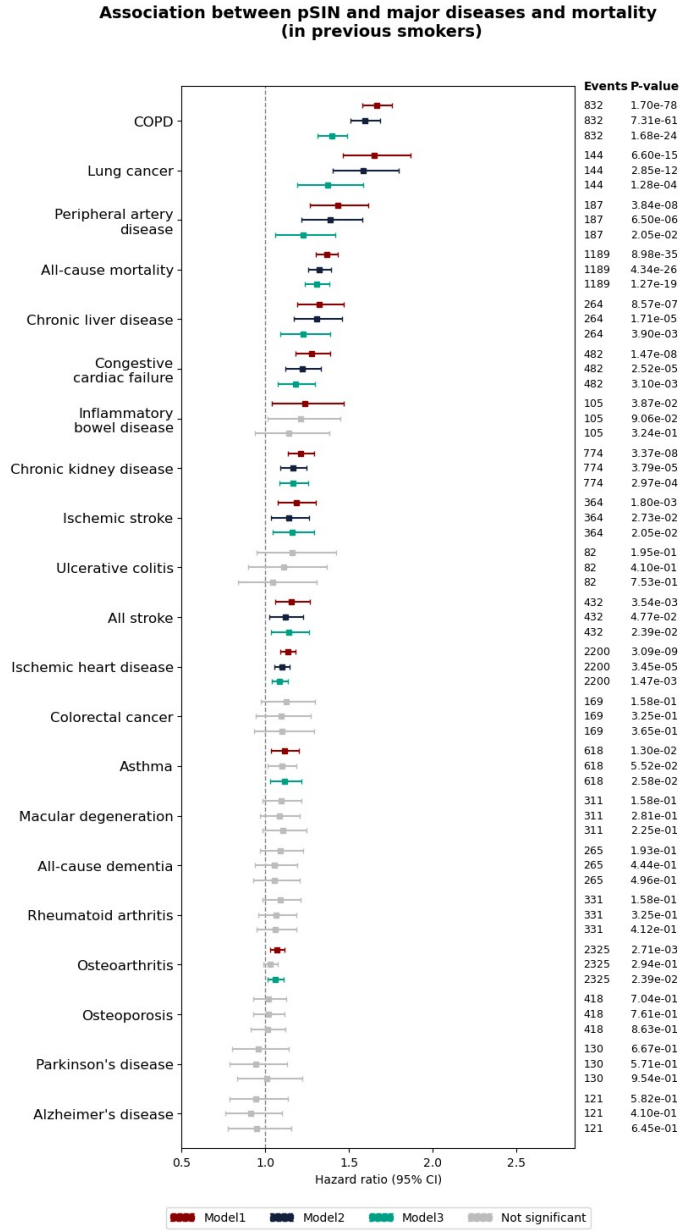


Figure 3: Quartiles of pSIN lead to strongly diverging cumulative incidence of major incident diseases, and mortality. Cumulative incidence plot of top, median and bottom 25% of the pSIN in the whole UKB population with 95% confidence interval shown as lighter shading. X-axis denotes the chronological age and Y-axis denotes the cumulative incidence. Cumulative incidence and number at risk at each age point is shown in ***Table s15*** and ***Table s16***.


Association between pSIN and major diseases and mortality
(in current smokers)
Events P-value
Lung cancer
151 6.18e-12
151 1.27e-07
151 1.65e-05
Peripheral artery disease
119 5.36e-09
119 2.36e-05
119 2.87e-03
COPD
624 1.13e-40
624 1.38e-22
624 7.60e-13
All-cause mortality
577 2.71e-17
577 1.30e-09
577 1.41e-03
Osteoporosis
183 5.20e-05
183 8.14e-03
183 2.24e-02
Congestive cardiac failure
187 2.33e-03
187 5.31e-02
187 6.66e-01
All stroke
209 2.15e-03
209 9.87e-02
209 7.69e-01
Ischemic stroke
168 6.07e-03
168 2.12e-01
168 8.07e-01
Ischemic heart disease
855 8.64e-11
855 1.48e-05
855 2.55e-02
Asthma
221 2.15e-03
221 9.87e-02
221 2.19e-01
Rheumatoid arthritis
150 2.78e-02
150 1.76e-01
150 4.03e-01
Chronic liver disease
139 4.45e-02
139 5.89e-01
139 7.69e-01
Chronic kidney disease
244 2.78e-02
244 5.89e-01
244 9.70e-01
All-cause dementia
85 4.55e-01
85 9.48e-01
85 6.66e-01
Osteoarthritis
693 2.61e-02
693 2.12e-01
693 6.74e-01
0.5 1.0 1.5 2.0 2.5
Hazard ratio (95% CI)
Model1 Model2 Model3 Not significant


Association between pSIN and major diseases and mortality
(in previous smokers)
Events P-value
COPD
832 1.70e-78
832 7.31e-61
832 1.68e-24
Lung cancer
144 6.60e-15
144 2.85e-12
144 1.28e-04
Peripheral artery disease
187 3.84e-08
187 6.50e-06
187 2.05e-02
All-cause mortality
1189 8.98e-35
1189 4.34e-26
1189 1.27e-19
Chronic liver disease
264 8.57e-07
264 1.71e-05
264 3.90e-03
Congestive cardiac failure
482 1.47e-08
482 2.52e-05
482 3.10e-03
Inflammatory bowel disease
105 3.87e-02
105 9.06e-02
105 3.24e-01
Chronic kidney disease
774 3.37e-08
774 3.79e-05
774 2.97e-04
Ischemic stroke
364 1.80e-03
364 2.73e-02
364 2.05e-02
Ulcerative colitis
82 1.95e-01
82 4.10e-01
82 7.53e-01
All stroke
432 3.54e-03
432 4.77e-02
432 2.39e-02
Ischemic heart disease
2200 3.09e-09
2200 3.45e-05
2200 1.47e-03
Colorectal cancer
169 1.58e-01
169 3.25e-01
169 3.65e-01
Asthma
618 1.30e-02
618 5.52e-02
618 2.58e-02
Macular degeneration
311 1.58e-01
311 2.81e-01
311 2.25e-01
All-cause dementia
265 1.93e-01
265 4.44e-01
265 4.96e-01
Rheumatoid arthritis
331 1.58e-01
331 3.25e-01
331 4.12e-01
Osteoarthritis
2325 2.71e-03
2325 2.94e-01
2325 2.39e-02
Osteoporosis
418 7.04e-01
418 7.61e-01
418 8.63e-01
Parkinson's disease
130 6.67e-01
130 5.71e-01
130 9.54e-01
Alzheimer's disease
121 5.82e-01
121 4.10e-01
121 6.45e-01
0.5 1.0 1.5 2.0 2.5
Hazard ratio (95% CI)
Model1 Model2 Model3 Not significant

A **B**

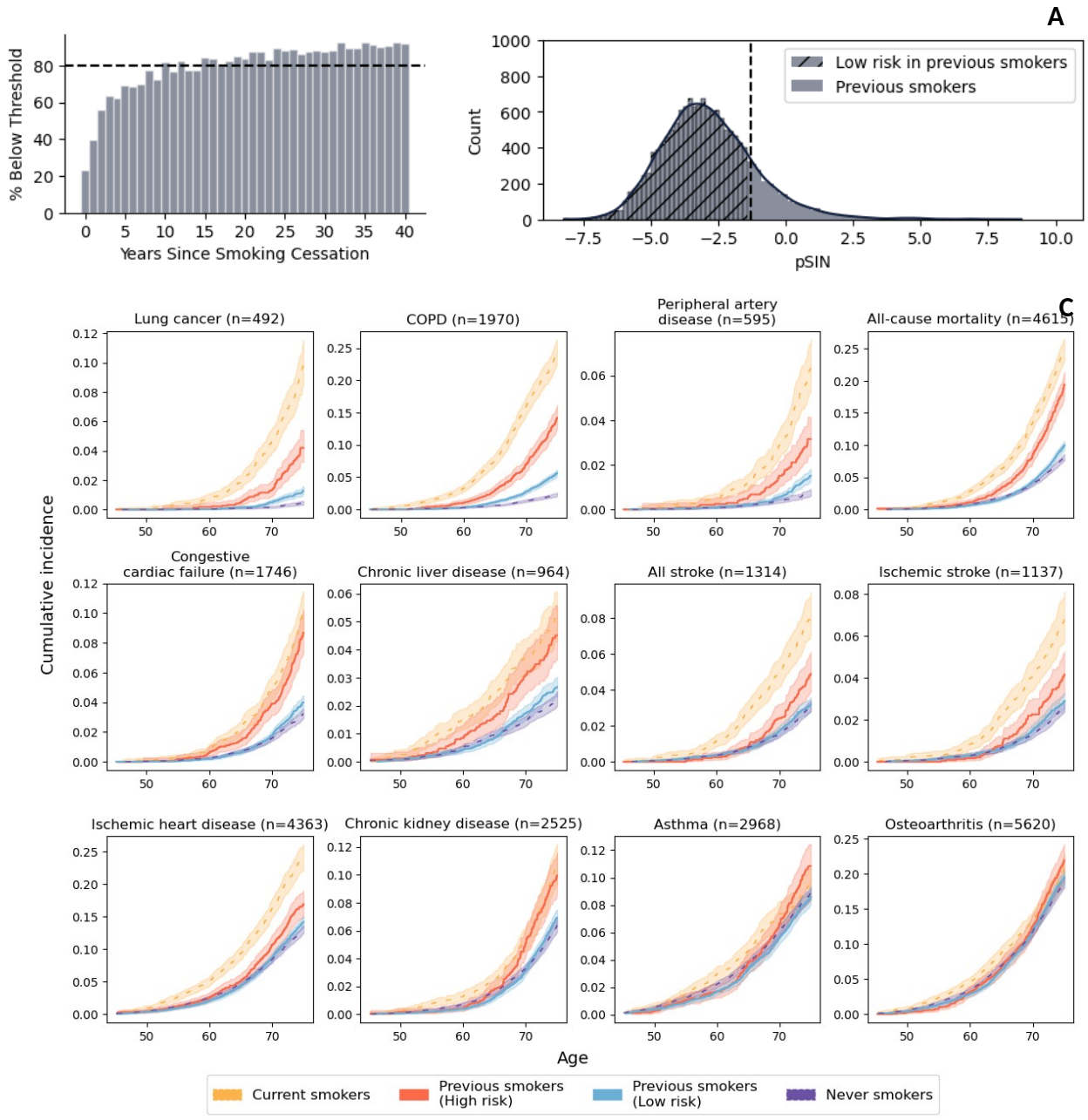


Figure 5: Previous smokers with pSIN similar to never smokers (recovery) show lower morbidity and mortality risks.

(a) shows the percentage of the previous smoker population in each year of smoking cessation bin who have a pSIN lower than the cutoff value used to differentiate current smokers. This percentage first reaches 80% after 10 years of smoking cessation. **(b)** shows the distribution of pSIN in previous smokers. Dotted line denotes the cut-off when differentiating current smokers from never smokers at FPR of 0.05 dividing previous smokers into two groups. Hashed part denotes the group in previous smokers with a similar pSIN as never smokers (recovery). **(c)** shows cumulative incidence plot of low and high-risk group defined by pSIN in previous smokers (orange and blue) with self-reported current smokers as positive control (yellow) and self-reported never smokers as negative control (purple). Cumulative incidence and number at risk charts are shown in **Table S24, S25** respectively.

Supplementary figures

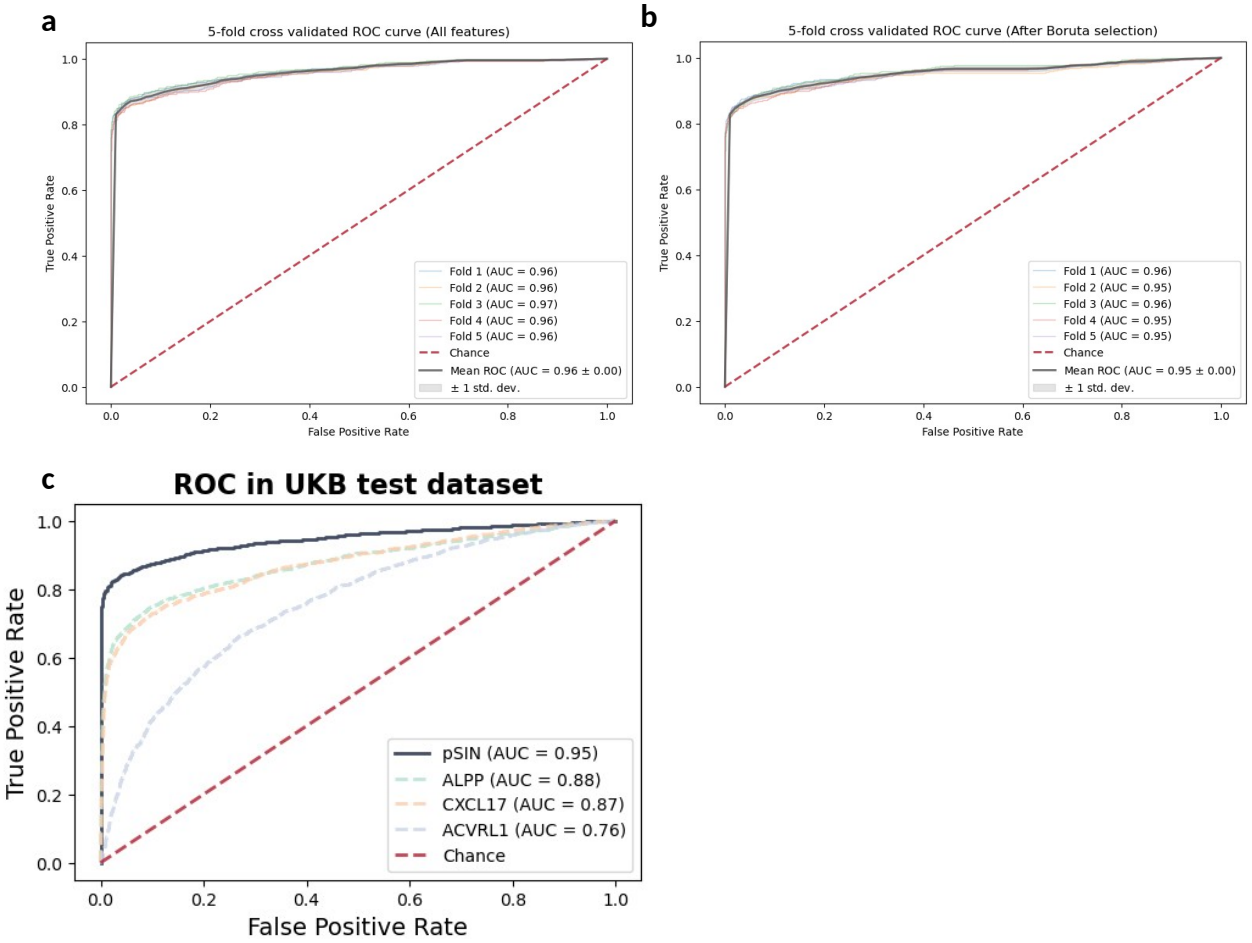


Figure s1 Performance of the classification model.

a)5-fold cross validated ROC curve of models trained using all proteins and **b)** 5-fold cross validated ROC curve of models trained using Boruta selected 51 proteins only c) ROC curve comparison of the gradient boosting model comprising 51 proteins in the UKB test dataset comparing to the performance of the top3 single protein when differentiating current smokers from never smokers.

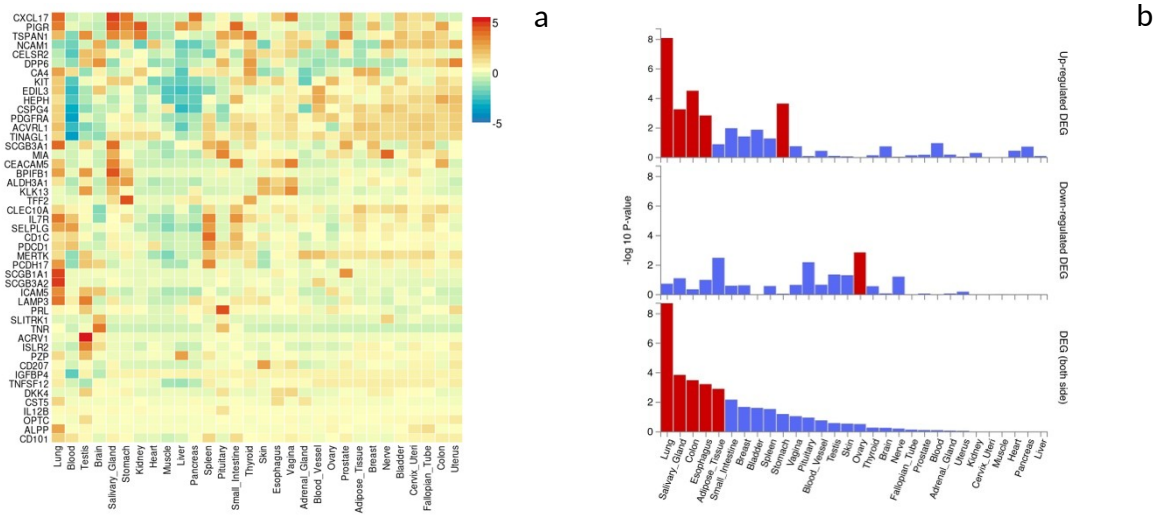
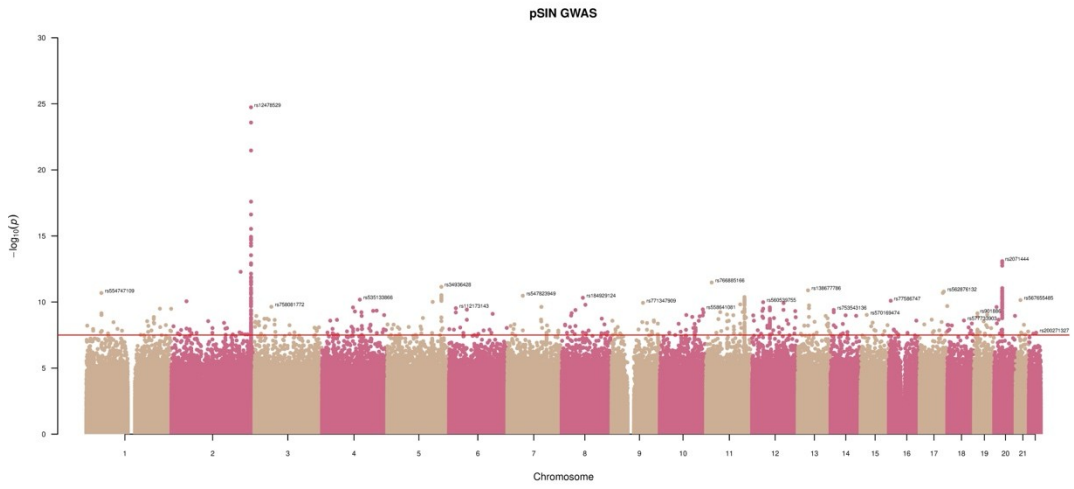


Figure s2 Tissue specific expression of the Boruta selected proteins.

Tissue specific expression of the selected proteins according to GTEx database. **a)** heatmap of average of normalized expression per gene per tissue. X-axis denotes the corresponding tissue and Y-axis denotes each of the selected proteins. **b)** differential expression genes (DEG) were identified by 2-sided t-test per tissue type versus all other tissue types. Genes with a Bonferroni corrected p-value < 0.05 and absolute log fold change > 0.58 were selected as DEG and were shown as red color

A



B

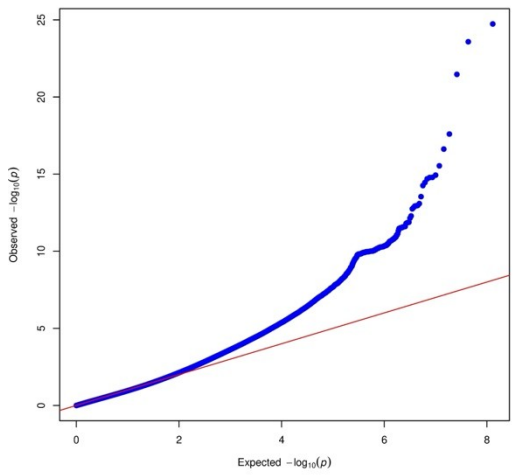
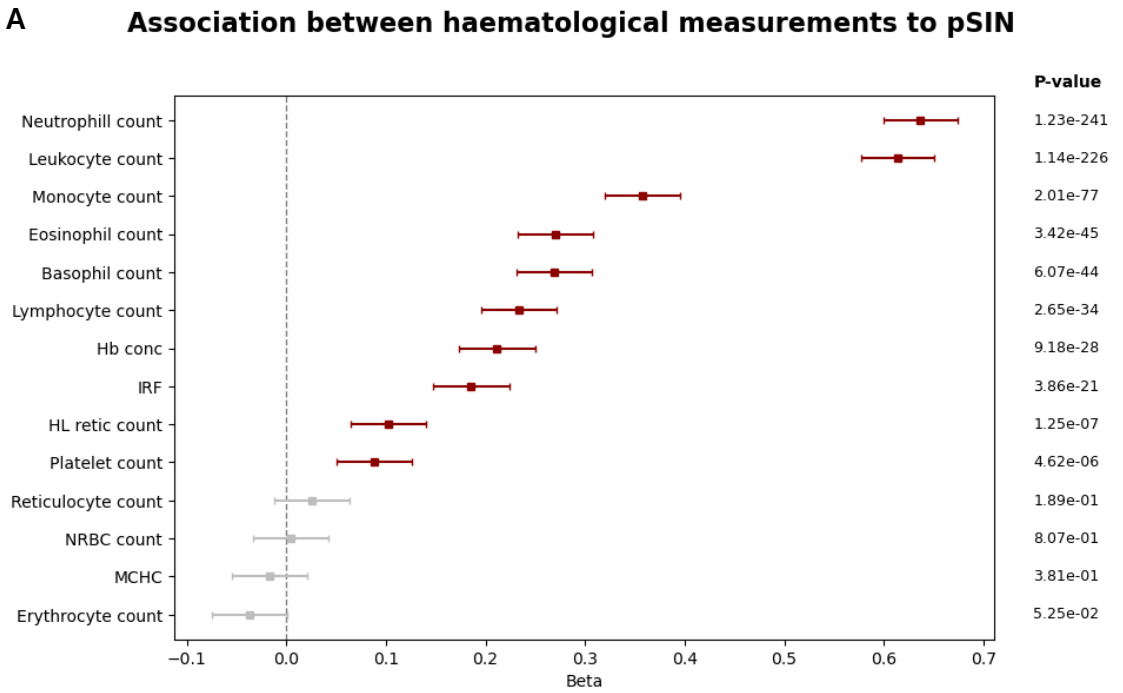


Figure s3 Results of GWAS results on pSIN.

Results of GWAS adjusted for age, sex and 40 PCs **a)** Manhattan plot shows the p-value between SNPs and pSIN. A cut-off p-value of $5*10^{-8}$ was used and shown as red line in the plot. Lead significant SNPs were annotated. **b)** shows the Quantile-quantile plot (QQ-plot) of GWAS analysis



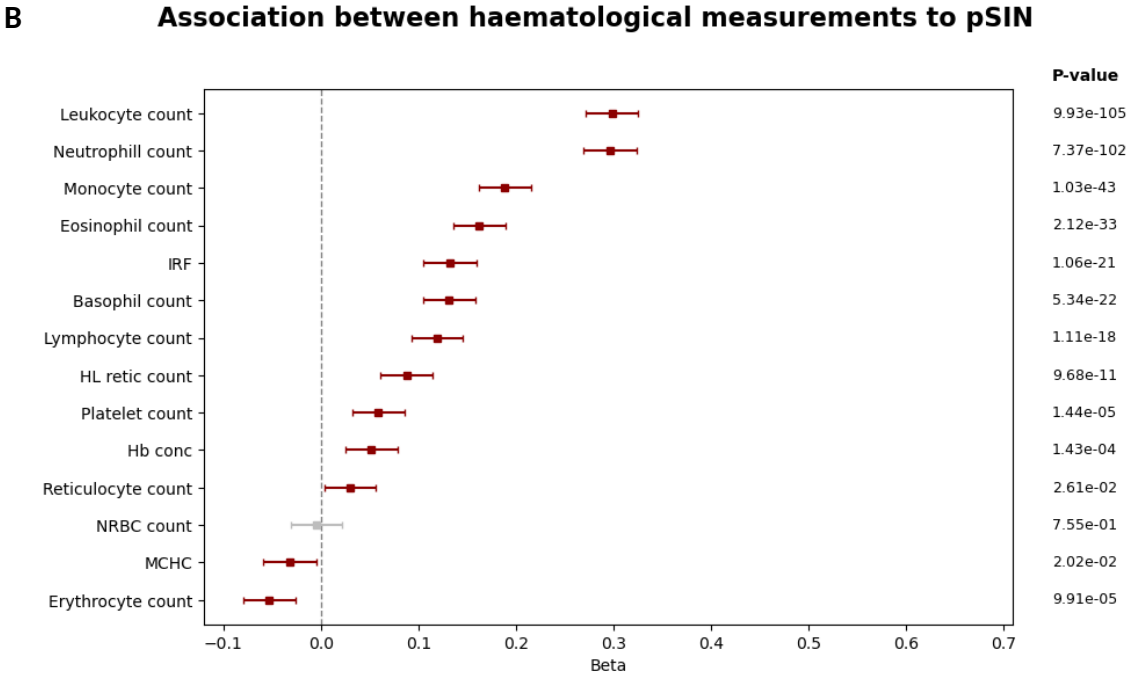


Figure s4 haematological measurements showed significant association with pSIN.

A) Linear regression analysis between individual haematological measurements and pSIN was performed within the whole UKB population adjusting for recruitment center, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, and education. Red colour denotes the association is significant after correcting for FDR multiple testing. Hb, haemoglobin, IRF, immature reticulocyte fraction; HL retic count, reticulocyte (red blood cell) count; NRBC, nucleated red blood cells; MCHC, mean corpuscular haemoglobin concentration. **B)** Model adjust additionally for smoking status.

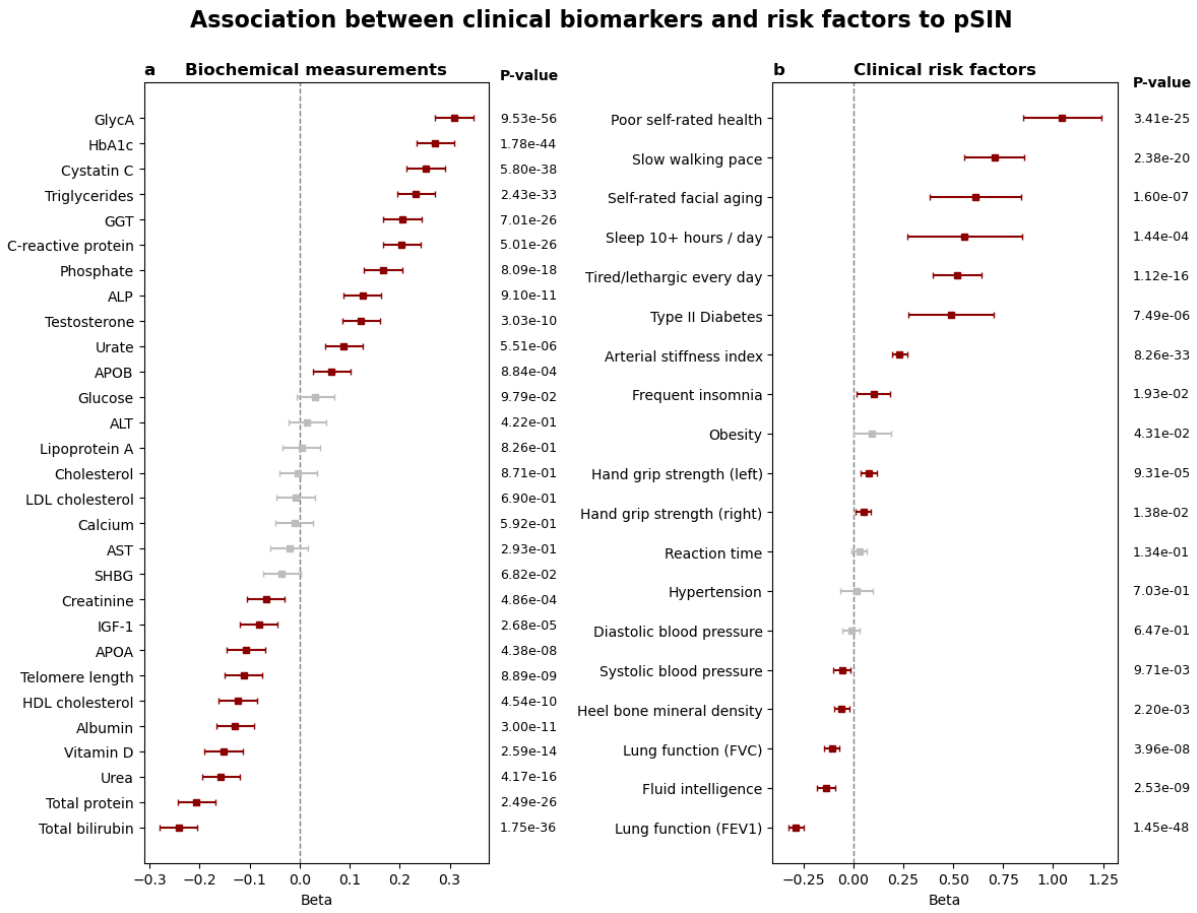


Figure s5 Relationship between blood biomarkers and clinical risk factors with pSIN.

Linear regression analysis was performed within the whole UKB population adjusting for recruitment center, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, and education. Red colour denotes the association is significant after correcting for FDR multiple testing. **a)** Association between biochemical measurements and pSIN. **b)** Association between clinical risk factors and pSIN.

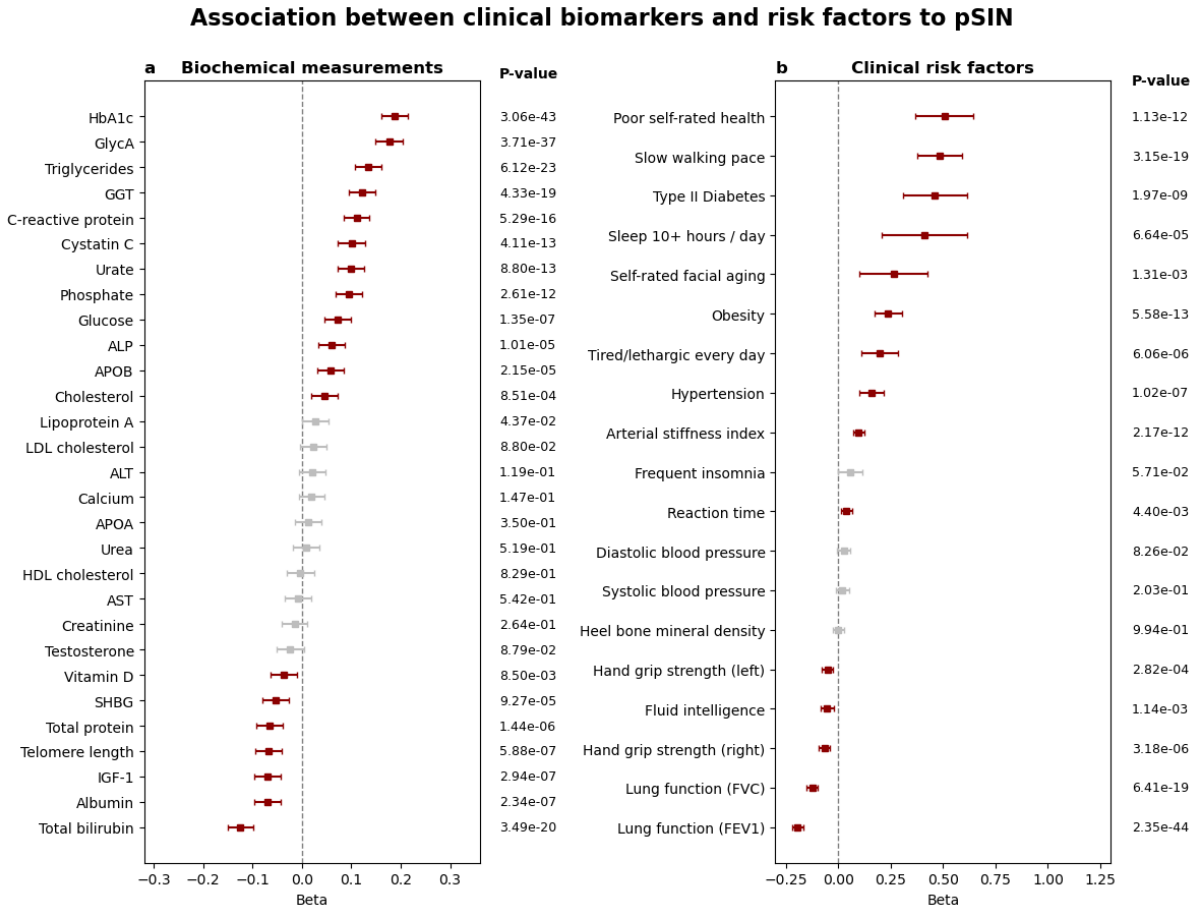


Figure s6 Relationship between blood biomarkers and clinical risk factors with pSIN independent of smoking status.

Linear regression analysis was performed within the whole UKB population adjusting for recruitment center, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, education and smoking status. Red colour denotes the association is significant after correcting for FDR multiple testing. **a)** Association between biochemical measurements and pSIN. **b)** Association between clinical risk factors and pSIN.

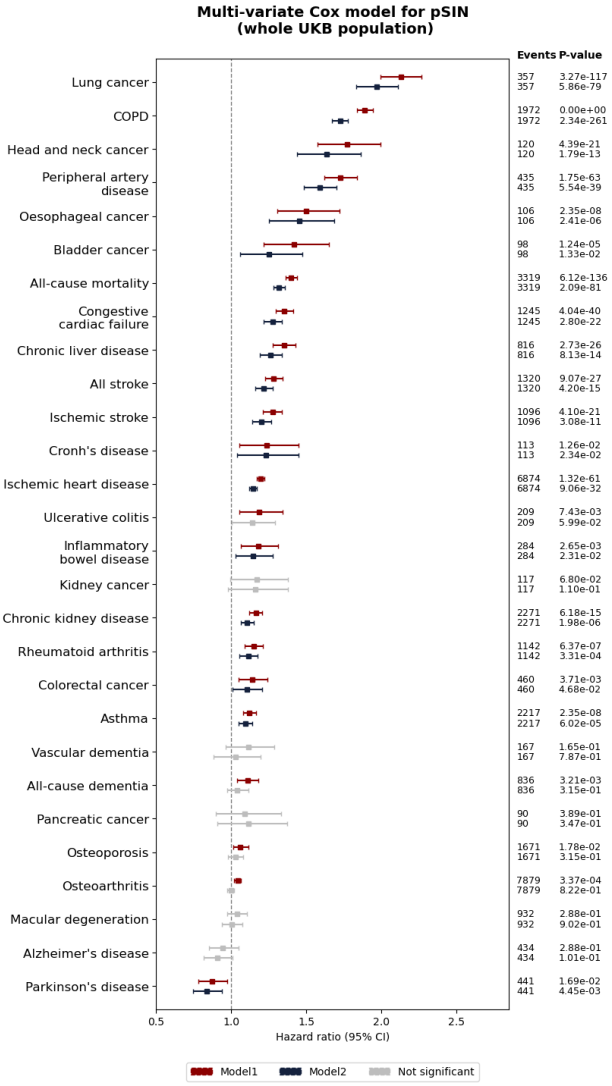


***Figure s7* pSIN is associated with future risks of morbidities and mortality.**

Forest plot shows association between pSIN and 24 morbidities and mortality which has at least 80 cases during follow-up time using multi-variate cox proportional hazard model. In model 1, the exposure was pSIN without adjusting for any covariate as age and sex already been regressed out in protein level. Model 2 was adjusted for recruitment centre, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years. P-values were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour.

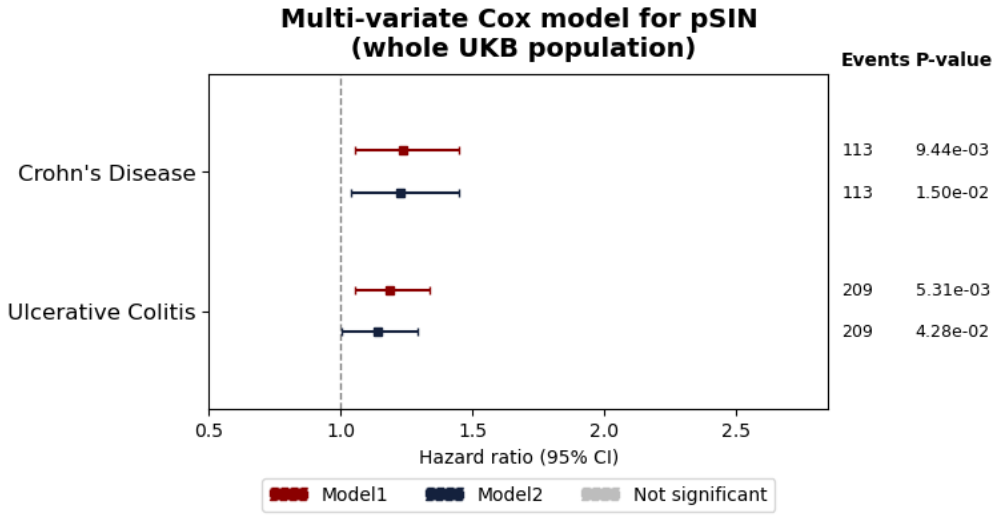


Figure s8 **pSIN is able to capture different hazard ratio of subtypes of incident inflammatory bowel disease.**

Forest plot shows association between pSIN and subtypes of incident inflammatory bowel disease (Crohn's disease and ulcerative colitis) using multi-variate cox proportional hazard model. In model 1, the exposure was pSIN without adjusting for any covariate as age and sex already been regressed out in protein level. Model 2 was adjusted for recruitment centre, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years. P-values were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour.

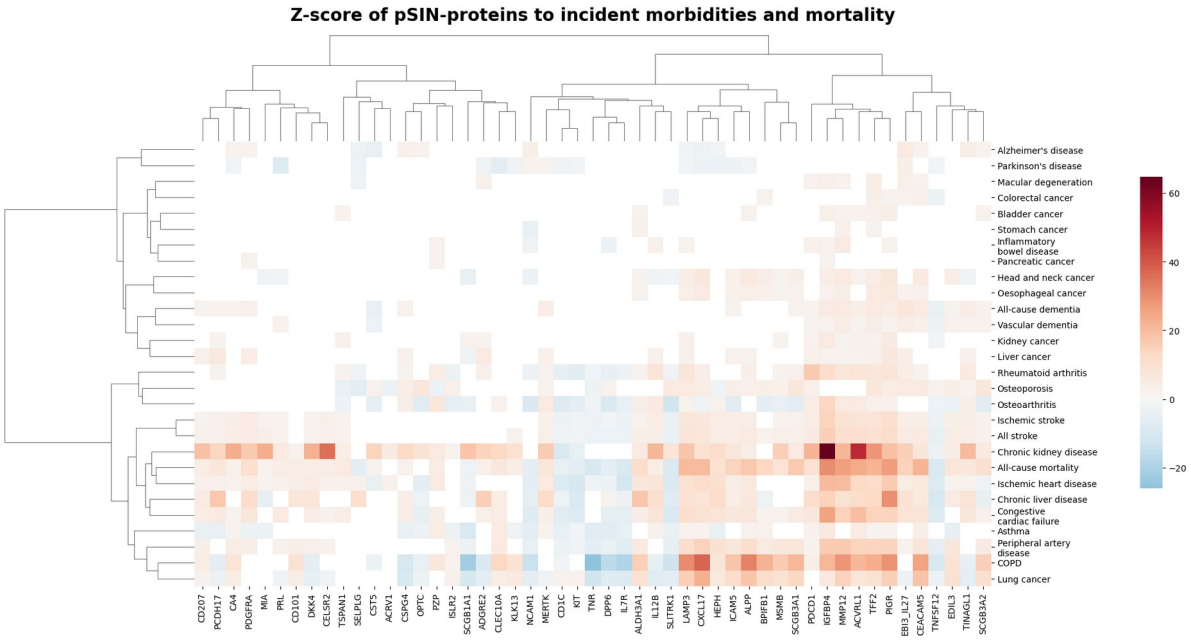


Figure s9 Association between individual proteins and major diseases, and mortality.

COX proportional hazard model was used to assess the association between each of the 51 protein and diseases and mortality. Model was adjusted for recruitment centre, ethnicity, education years, and Townsend deprivation index. Z-score was shown on the heatmap. Association p value was corrected for FDR multiple testing and none-significant associations were shown as white colour.

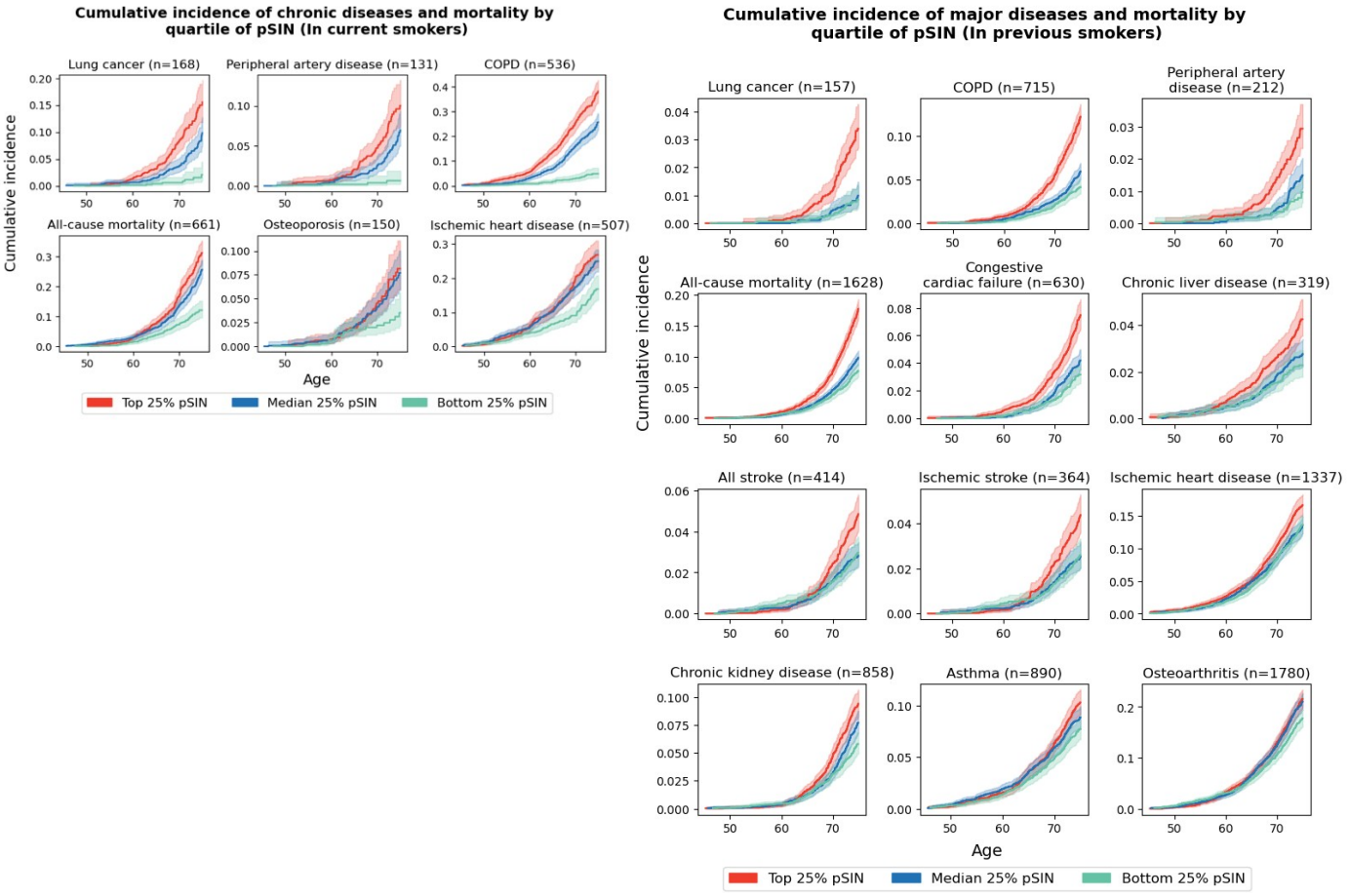


Figure s10* pSIN differentiates future risks of morbidities and mortality in current and previous smokers. a)** shows cumulative incidence plot of top, median and bottom 25% of the pSIN in the current smokers with 95% confidence interval shown as lighter shading. X-axis denotes the chronological age and Y-axis denotes the cumulative incidence. Cumulative incidence and number at risk at each age point is shown in ***Table s17 and ***Table s18***. b) shows cumulative incidence plot of top, median and bottom 25% of the pSIN in the previous smokers with 95% confidence interval shown as lighter shading. Cumulative incidence and number at risk at each age point is shown in ***Table s20*** and ***Table s21***. Only outcomes that were significant in all three cox models were displayed here.

A **B**

C **D**

E F

G **H**

I

Figure s11 conventional smoking history variables still has residual information not fully captured by pSIN

Each forest plot shows association between the corresponding smoking history variables a) smoking status, b) smoking years (current smokers), c) number of cigarettes smoked (current smokers), d) pack years (current smokers), e) smoking cessation years (previous smokers), f) smoking years (previous smokers), g) number of cigarettes smoked (previous smokers), h) pack years (previous smokers), i) has passive smoking exposure (never smokers) and 24 morbidities and mortality which has at least 80 cases during follow-up time using multi-variate cox proportional hazard model. In model 1, the exposure was the corresponding smoking history variable adjusting for recruitment centre, Townsend deprivation index, ethnicity and education years. The model 2 was adjusted additionally for pSIN. P-values were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour.

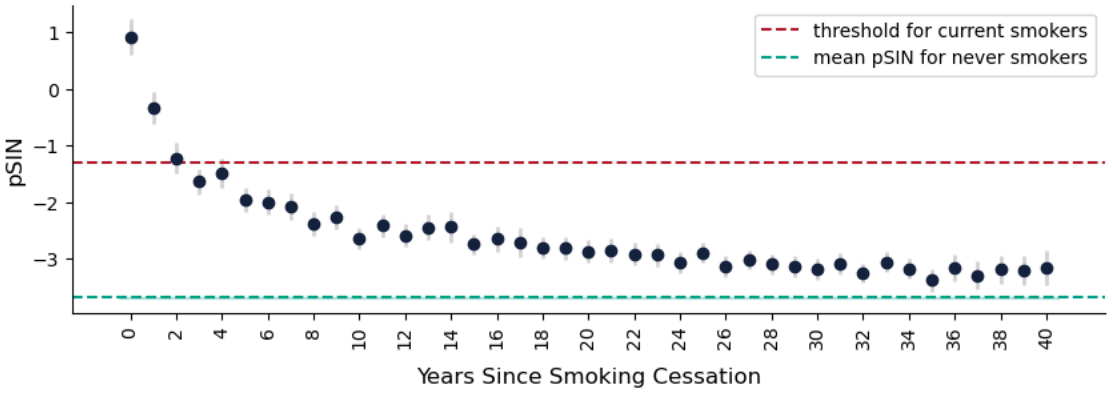


Figure s12 Mean pSIN by years since smoking cessation.

Mean pSIN in each year since cessation bin was shown in the plot with 95%CI. Red dashed line shows the threshold of differentiating current smoker from never smokers with FPR of 0.05. Green dashed line shows the mean pSIN of self-reported never smokers with 95%CI shown in green shadow.

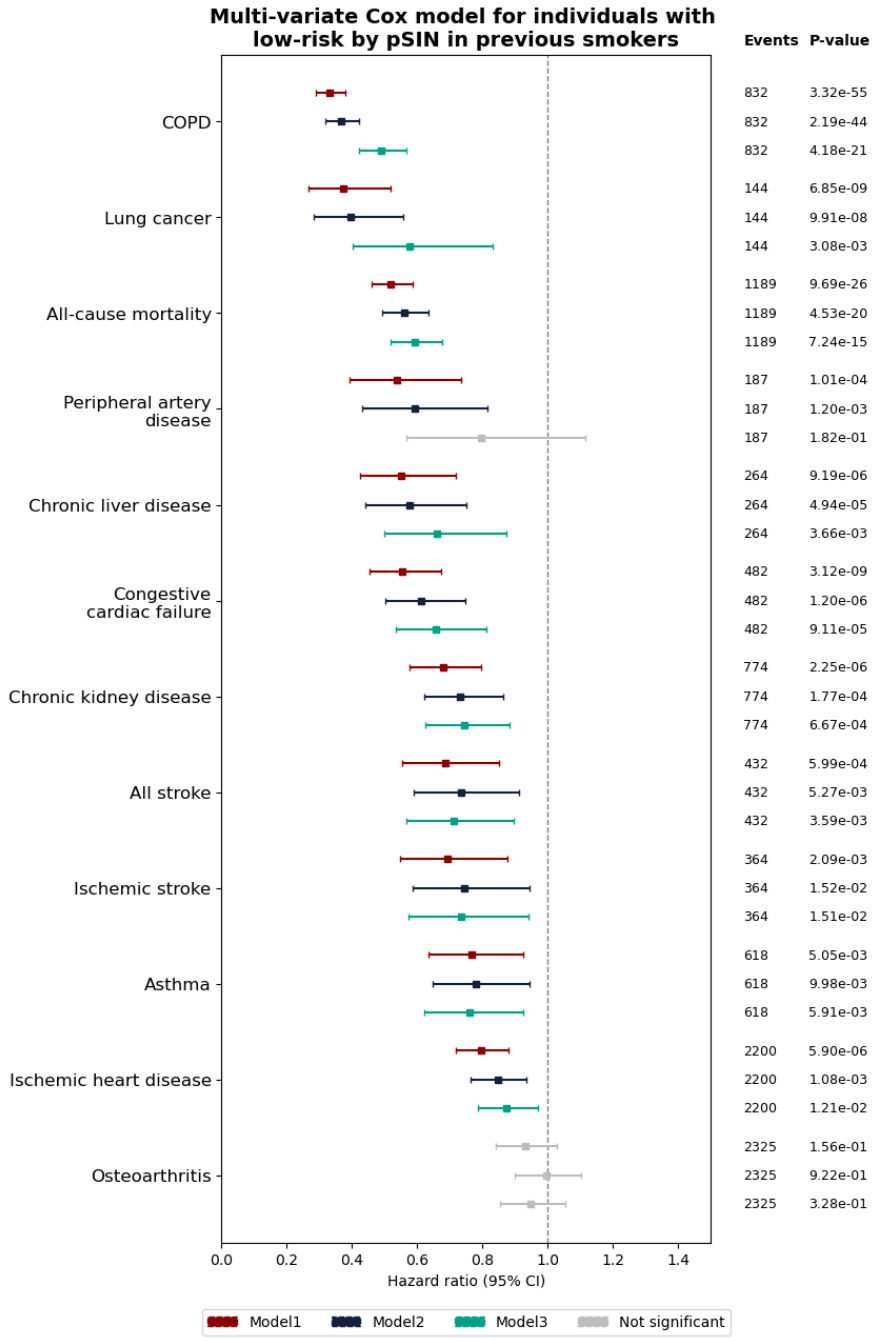


Figure s13 previous smokers with pSIN similar to never smokers were associated with lower morbidity and mortality risks.

Forest plot shows association between low-risk group in previous smokers and health outcomes which has at least 80 cases during follow-up time and were significant in previous smoker model using multi-variate cox proportional hazard model. The comparisons are made between previous smokers with low-risk based on pSIN (i.e., those whose proteomic profiles resemble never smokers) and high-risk based pSIN. In model 1, the exposure was pSIN without adjusting for any covariate as age and sex already been regressed out in protein level. Model 2 was adjusted for recruitment centre, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years. Model 3 further adjusted for smoking pack years and smoking cessation time. P-values were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour.

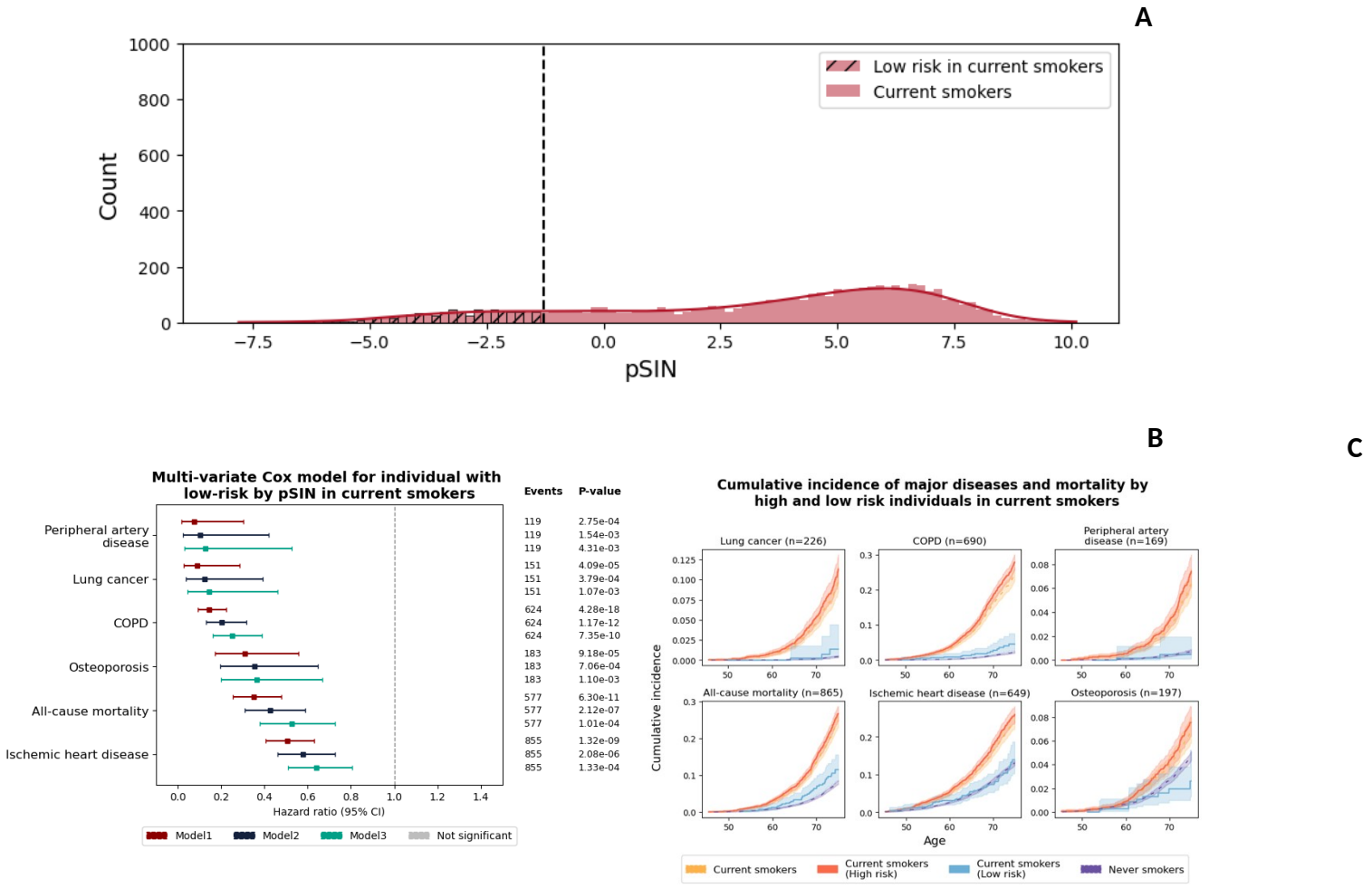


Figure s14 people in current smokers with pSIN similar to never smokers were associated with lower morbidity and mortality risks.

(a) shows the distribution of pSIN in current smokers. Dotted line denotes the cut-off when differentiating current smokers from never smokers at FPR of 0.05 dividing current smokers into two groups. Hashed part denotes the group in current smokers with a similar pSIN as never smokers. **(b)** Forest plot shows association between low-risk group in current smokers and 5 morbidities and mortality which has at least 80 cases during follow-up time and were significant in current smoker model using multi-variate cox proportional hazard model. In model 1, the exposure was pSIN without adjusting for any covariate as age and sex already been regressed out in protein level. Model 2 was adjusted for recruitment centre, Townsend deprivation index, IPAQ physical activity group, ethnicity, alcohol frequency, BMI, and education years. Model 3 further adjusted for smoking pack years. P-values were corrected for FDR multiple testing and non-significant associations after corrections were shown as grey colour. **(c)** shows cumulative incidence plot of low and high-risk group defined by pSIN in current smokers (orange and blue) with self-reported current smokers as positive control (yellow) and self-reported never smokers as negative control (purple). Cumulative incidence and number at risk charts are shown in **Table S27, S28** respectively.

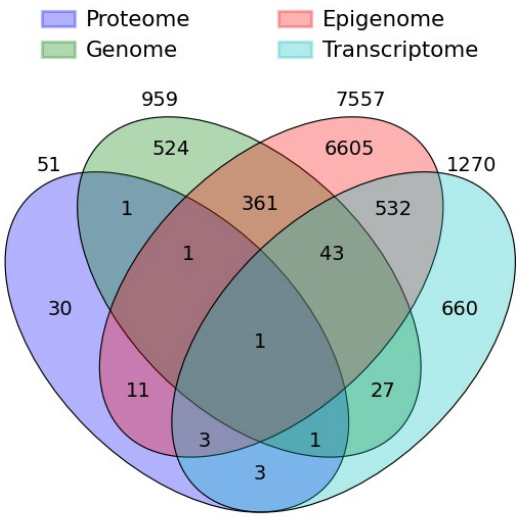


Figure s15 Over lapping genes selected by pSIN and major genomic, transcriptomic and epigenomic studies

pSIN selected 51 proteins were compared against three major studies each for genomic, transcriptomic and epigenomic wide association study of smoking initiation. 30 proteins were not found in previous omics study. Comparing against individual studies of biomarkers of smoking (also including protein biomarkers) 15 of the proteins selected by pSIN were not seen in the previous studies.

Figure s16 Distribution of DNA methylation score from Elliott et al.

DNA methylation score was built in current smokers versus never smokers of the Elliott et al study. DNA methylation score was then calculated for previous smokers. The overall distribution showed similarity comparing to the pSIN in the current smokers, previous smokers and never smokers in UKB.

Chapter 6 Discussion

Gradient boosting method

An essential innovation of this thesis was the adoption of gradient-boosting—in particular, LightGBM —to model both regression (chronological age) and classification (smoking status) tasks in high-dimensional proteomic and metabolomic data. Across all cohorts, LGBM consistently outperformed penalised linear models (LASSO, elastic net) and neural networks, achieving the highest R^2 for proteomic age and superior AUC for smoking classification, while maintaining its performance in external validation cohorts. By contrast, neural networks—though competitive in-sample—proved brittle at the extremes of the age distribution (under 40 and over 75 years), manifesting “edge-clipping” in predicted ages and poor generalisation in cohorts with broader age ranges. This divergence stems from their fundamentally different inductive biases: neural networks learn a smooth, highly parameterised mapping that, in regions of sparse data, must extrapolate and consequently “fall back” toward the population mean; LGBM’s tree-based structure, however, partitions the feature space into leaves that can isolate and fit extreme-age samples without distorting the central region. Moreover, techniques like weight-decay and dropout impose global smoothness priors on neural networks—enhancing extreme predictions to be regularized toward the mean—whereas gradient boosting uses shrinkage and controlled tree complexity yet still permits hard splits and local residual corrections specifically targeting outliers. Finally, neural networks’ sensitivity to feature scaling and distributional shifts exacerbates instability at the tails, while LGBM’s quantile-based splitting

naturally handles skewed inputs. Importantly, LGBM's piecewise-constant fitting and sequential residual learning confers robustness at the fringes of the data, allowing it to retain predictive accuracy where neural nets revert to central predictions. Equally critical is LGBM's compatibility with SHAP, which decomposes each prediction into additive contributions of proteins or metabolites, revealing the biological pathways driving accelerated ageing or smoking-related damage.

Sex-stratified study of ageing

Gradient boosting models consistently demonstrate high accuracy in predicting chronological age when trained on combined male and female cohorts. Indeed, our previous work on the proteomic ageing clock achieved a very similar R^2 of 0.87 and 0.86 in females and males respectively with sex combined model. However, this pooled-sex approach obscures biological differences: the proteins driving age prediction in males differ substantially from those in females, and the relative importance of shared pathways varies by sex. Consequently, separate male and female models are essential not only for maximising predictive precision but also for probing the distinct molecular mechanisms that underlie ageing trajectories in each sex. Proteins such as FSHB, PAEP and CGA were specifically selected in the female model, highlighting pathways associated with hormonal regulation and reproductive aging[98,174]. FSHB, the follicle-stimulating hormone beta subunit, is closely linked to ovarian function and menopause[98], PEAP, is a glycoprotein secreted by the endometrium and plays a pivotal role in immunomodulation, embryo implantation, and maintenance of early pregnancy[244], while CGA, involved in glycoprotein hormone activity, reflects broader hormonal changes with implications

for bone density, metabolism, and cancer risks in females[174]. Of note, CGA, although ranked in the top 3 in the female model, was not selected at all in the sex-combined ageing model[51]. Conversely, male-specific proteins included markers of structural and vascular integrity such as ENG and ACRV1, which play roles in endothelial function, vascular remodelling, and neurovascular health, pathways critically associated with cardiovascular and cognitive decline in males[245].

Sex-stratified enrichment analyses of metAgeGap and protAgeGap further reveal a shared core of ageing pathways—extracellular matrix organization, anatomical structure development, and cytokine–receptor signalling—reflecting universal tissue remodelling and inflammatory processes. In the female-specific metAgeGap model, cholesterol transport, sterol metabolism, and Liver X Receptor-mediated signalling emerge as dominant signatures, underscoring lipid homeostasis and nuclear receptor regulation in metabolic ageing. The female protAgeGap model similarly highlights hormone-responsive proteins involved in growth factor signalling and lipid-binding functions. Conversely, the male metAgeGap clock is enriched for glycolysis/gluconeogenesis, alcohol-binding proteins, and vitamin digestion pathways, suggesting that energy metabolism and detoxification processes dominate male metabolic ageing. The male protAgeGap model likewise emphasises proteins associated with xenobiotic metabolism and ATP-binding cassette transport, highlighting analogous detoxification networks at the proteomic level.

In females, reproductive history and oestrogen dynamics exert pronounced effects on metAgeGap and protAgeGap estimates. Menopause was identified as a pivotal factor influencing metabolic ageing in females. The hormonal changes accompanying menopause,

particularly the decline in oestrogen levels, exert profound effects on lipid metabolism, inflammation, and body composition, all of which are integral to metabolic aging[200]. The duration since menopause also emerged as a significant factor, with a longer post-menopausal period correlating with higher metAgeGap. This may reflect prolonged exposure to post-menopausal metabolic shifts, including increased visceral fat, reduced insulin sensitivity, and heightened systemic inflammation. The timing and frequency of pregnancies were also significantly associated with protAgeGap, pointing to the long-lasting physiological and cellular effects of pregnancy, including shifts in immune function and metabolic demands[246]. Hormone replacement therapy (HRT) was also found to influence metAgeGap, with HRT users exhibiting older biological ageing. While HRT is recognised for mitigating menopausal symptoms and reducing osteoporosis and cardiovascular disease risks, its complex effects on metabolic pathways highlight the need for further research[247]. Later menarche associates with lower metAgeGap, suggesting lasting metabolic benefits of delayed sexual maturation, while higher parity and older ages at first and last birth correspond to reduced proteomic age acceleration. Current use of oestrogen-progestogen therapy is linked to a younger metAgeGap and a modestly lower protAgeGap, highlighting how exogenous hormones modulate both metabolic profiles and circulating protein signatures long after menopause.

metAgeGap and protAgeGap also exhibit sex-divergent prognostic utility. In females, elevated metAgeGap strongly predicts cardiovascular and neurovascular events—particularly ischemic heart disease and vascular dementia—while its cancer associations are largely confined to colorectal carcinoma. In males, higher metAgeGap forecasts a broader array of conditions, including liver, oesophageal, and lung cancers, and shows stronger hazard ratios for all-cause

mortality and incident heart disease, likely driven by higher exposure to lifestyle risk factors such as smoking and alcohol consumption. The female protAgeGap model likewise correlates with cognitive and vascular outcomes but also identifies specific protein markers linked to breast cancer risk. The male protAgeGap amplifies these associations, with proteins such as ENG and ACRV1 driving links to cardiovascular decline and cognitive disorders. In addition to its strong neurodegenerative associations, protAgeGap demonstrated notable cancer-specific patterns. In males, protAgeGap was linked to incident non-Hodgkin lymphoma, lung cancer, oesophagal cancer, and prostate cancer. These associations align with the well-documented connection between ageing and increased cancer risk, reflecting the accumulation of genetic, epigenetic, and proteomic changes over time that contribute to carcinogenesis[248]. In contrast, protAgeGap in females showed particularly robust associations with hormone-related cancers, such as breast cancer. This highlights the significant influence of reproductive hormones on ageing trajectories and their downstream effects on proteomic profiles. The enrichment of hormone-regulation pathways, including proteins like FSHB, PAEP and CGA, further supports the role of hormonal changes in driving cancer risk and systemically ageing in females.

Cumulative incidence analyses illustrate that females generally develop age-related diseases later than males, across both metAgeGap and protAgeGap frameworks. For non-cancer morbidities—such as osteoarthritis and certain dementias—females in the lowest quartile of metAgeGap maintain a substantially lower cumulative risk until after age 60, a pattern mirrored by protAgeGap estimates. In contrast, males reaching the same low-age-acceleration threshold face elevated risks much earlier, reflecting accelerated disease onset. These sex-specific temporal trajectories underscore the need for longitudinal follow-up extending beyond 75 years

to fully capture female disease onset patterns and validate both metabolomic and proteomic ageing biomarkers across the lifespan.

MetAgeGap vs ProtAgeGap

Compared with metabolic age stratified by sex, proteomic age showed more consistent performance across males and females. This may be because metabolites such as Omega3, and different lipids captured in the Nightingale panel were strongly controlled by menopause in females and hence easier to predict age compared to that in males. Proteomic age also showed far more superior ability to predict age in both sexes with an R^2 of 0.88 in the testing dataset, despite a smaller number of participants included in the study. This could be explained by the Olink plasma proteomic panel included ~3000 proteins and covered more pathways compared to 249 metabolites in the Nightingale panel. However, variance explained by proteomic age still outperformed metabolic age built using lipidomic measured by mass spectrometers with 3,098 lipidome clusters used in the previous study[122].

The comparison of metabolic age and proteomic age highlights nuanced sex-specific differences in the biological ageing process, reflecting both shared and unique pathways identified through GO and KEGG enrichment analyses. Both metabolic age and proteomic age analyses reveal significant enrichment in shared biological pathways, particularly those related to lipid metabolism and extracellular matrix (ECM) regulation. Key pathways such as PI3K/AKT signalling, cytokine-cytokine receptor interaction, and cholesterol metabolism are implicated in both sexes across the models. These shared pathways underscore fundamental processes in cellular signalling, metabolic homeostasis, and tissue maintenance that are central to ageing.

Notably, the PI3K/AKT pathway is recognized for its dual roles in cellular proliferation and apoptosis regulation, linking ageing to cancer development and longevity[161,166]. The identification of ECM-related pathways in both models further emphasizes the structural and functional decline of tissues as hallmarks of ageing, affecting stem cell niches and overall cellular homeostasis[162].

Despite these commonalities, both metabolic age and proteomic age models underscore distinct sex-specific pathways that reflect underlying physiological and hormonal differences.

For metabolic age in males, pathways enriched in lipid oxidation and PPAR signalling highlight the emphasis on triglyceride and fatty acid metabolism, consistent with higher rates of lipid-related cardiovascular diseases observed in males[249]. Conversely, females showed enrichment in pathways tied to organic substance metabolism and inflammatory regulation, potentially modulated by oestrogen's protective effects on lipid metabolism and inflammation[250,251].

On the other hand, proteomic age revealed more prominent distinctions in sex-specific proteins and pathways. For males, ECM degradation and BMP signalling emerge as key contributors, implicating processes like tissue fibrosis, vascular health, and neurogenesis in male-specific ageing trajectories. In females, pathways such as RAS/MAPK and hormone regulation (e.g., FSHB and CGA) are enriched, reflecting the interplay between reproductive health and systemic ageing. These findings align with the observed protective effects of estrogen on cardiovascular and bone health, alongside its role in modulating metabolic and inflammatory processes.

The findings from both models reinforce the complexity of biological ageing and the necessity of sex-specific frameworks in ageing research. By identifying shared and unique pathways, this

analysis contributes to a deeper understanding of how biological ageing manifests differently across sexes, driven by metabolic, hormonal, and proteomic factors.

The differences in the percentage of variance explained and the pathways captured between metabolic age and proteomic age translate into differences in their ability to differentiate risks for incident diseases (***Fig 1***). Proteomic age demonstrated significant correlations with 14 out of 15 non-cancer health outcomes across both sexes, whereas metabolic age was significantly associated with only 7 out of 15 diseases in females and 12 out of 15 in males (***Fig 1A, B***). Shared diseases between the two models, such as chronic kidney disease, COPD, and stroke, underscore the common pathways of ageing-related biological processes captured by both metabolic and proteomic data, including inflammation, metabolic dysfunction, and cellular senescence.

Differences in pathways captured by metabolic age and proteomic age contribute to their differential ability to differentiate disease risks across systems. For example, proteomic age was associated with all tested neurodegenerative diseases in both sexes, whereas metabolic age was significantly linked only to vascular dementia in both sexes and all-cause dementia in males (***Fig 1A, B***). This discrepancy highlights the superior capacity of proteomic age to capture neural ageing and its potential as a robust biomarker for age-related neurological conditions.

Both metabolic age and proteomic age were significantly associated with incident risks of all-cause mortality. However, when comparing sexes and omics models, proteomic age exhibited stronger associations overall. In females, proteomic age demonstrated an HR of 1.24, compared to 1.07 for metabolic age (***Fig 1A***). Similarly, in males, proteomic age yielded a stronger HR of

1.38, compared to 1.20 for metabolic age (***Fig 1B***). These findings suggest that proteomic age provides a more comprehensive assessment of mortality risk and ageing-related disease burden than metabolic age.

For both metabolic and proteomic age models, they showed stronger associations in males compared to females. Further, our results have shown that metAgeGap captured a completely different trajectory of cancer risks compared to protAgeGap (***Fig 1C, D***). In females, while metAgeGap was significantly associated with colorectal cancer, protAgeGap was significantly associated with breast cancer (***Fig 1C***). In males, while metAgeGap was associated with liver, oesophageal, lung and colorectal cancer, protAgeGap was associated with non-Hodgkin lymphoma, lung cancer, oesophageal cancer and prostate cancer (***Fig 1D***). Significant association with lung and oesophageal cancer in both metAgeGap and protAgeGap may suggest the importance of smoking behaviour to males' ageing process as smoking is one of the largest contributors to these two cancer types[252]. Notably, the association between metAgeGap and liver cancer was particularly strong, with an HR of 1.76, showing its relationship with alcohol consumption in males. This finding aligns with metabolic age's robust association with chronic liver disease (HR=1.25) in males, suggesting that it effectively captures markers of liver damage and dysfunction.

To investigate the additive information that protAgeGap has over metAgeGap when predicting incident diseases, we calculated and compared the C-index of the Cox model in participants with both metAgeGap and protAgeGap (n in female = 19,796; n in male = 14,835) (***Fig 2***). In most disease types, cox model with protAgeGap showed a significantly higher C-index compared to the model with metAgeGap. However, the Cox model with both metAgeGap and

protAgeGap only showed an increased C-index compared to the model with protAgeGap in limited diseases including vascular dementia, osteoporosis, chronic kidney diseases, chronic liver diseases in females and vascular dementia, COPD, ischemic stroke, all stroke, all-cause mortality, chronic kidney diseases, Parkinson's disease, chronic liver diseases, and macular degeneration in male (***Fig 2A, B***). The story is similar when combining both omics to predict cancer outcomes as well. Increased C-index over the proteomic model was only seen in kidney cancer, leukaemia in females and oesophageal cancer in males (***Fig 2C, D***). This suggested that combining metabolomic information with an already powerful proteomic ageing clock only added very little information for limited health outcomes. This may be because the Nightingale panel only measured a limited range of metabolites while the Olink proteomic panel already covered almost all biological pathways.

The differential performance of metabolic and proteomic age models in predicting cancer risks underscores the importance of pathway-specific biological insights provided by each model. Metabolic age appears to be more attuned to metabolic alterations that predispose individuals to cancer, while proteomic age reflects broader proteomic disruptions that may contribute to systemic and immune-related cancer pathways. The observed differences in the ability of metabolic and proteomic age models to capture increased risks of incident cancers might also stem from disparities in the size of the population used to construct the models. The metabolic age model was developed using the full cohort of ~500,000 participants from the UK Biobank, providing access to a larger number of cancer cases and thus greater statistical power to detect associations with more rare cancer types. In contrast, the proteomic age model was built on a

smaller sub-cohort of ~50,000 participants, potentially limiting its ability to capture significant associations with rarer cancer outcomes or those requiring higher statistical resolution.

This discrepancy in cohort size underscores the importance of population size in omics studies, particularly when investigating outcomes such as cancer, where incident cases may be relatively infrequent. The larger cohort available to the metabolic model likely increased its sensitivity to detect associations across a broader range of cancers, whereas the proteomic model's reduced sample size may have restricted its statistical power, leading to fewer significant associations despite its robustness in other areas of risk prediction. Future research leveraging larger proteomic datasets or integrative approaches combining multiple omics layers could help address these limitations, enabling more comprehensive evaluations of cancer risks and their underlying biological pathways.

I also observed that when evaluating the association between metabolic age and incident health outcomes using multivariable Cox models, HRs were substantially influenced by BMI, particularly for chronic kidney disease, chronic liver disease, and ischemic heart disease in females. This indicates that a significant proportion of the risk captured by metabolic age for these outcomes may be mediated or confounded by BMI, reflecting the strong interdependence between metabolic pathways and adiposity. In females, this dependency underscores the pivotal role BMI plays as both a determinant and a mediator of metabolic dysregulation and its downstream health effects.

In contrast, proteomic age showed resilience to confounding by BMI or other socio-economic and lifestyle factors when adjusted for in the models. This suggests that proteomic age captures

biological processes related to ageing and disease risk that are less influenced by external factors such as BMI, physical activities and smoking. Its associations with incident health outcomes appeared to remain robust across various adjustment models, reflecting its potential utility as a more direct marker of underlying biological ageing processes rather than those heavily shaped by modifiable risk factors. The stark contrast in how BMI impacts the predictive value of these models highlights differences in the biological pathways captured by metabolic and proteomic age. While metabolic age effectively reflects processes influenced by adiposity and metabolic health, proteomic age seems to offer a broader lens into ageing-related pathophysiology that may be more independent of lifestyle-related factors. This distinction underscores the complementary nature of these models and suggests their combined use could provide a more nuanced understanding of ageing and disease risks.

pSIN and aging

Using plasma proteomics from extensive cohort datasets, this study introduces the proteomic Smoking Index (pSIN)—a biomarker developed through machine learning—to quantify smoking exposure and its associated risks more comprehensively and accurately than previously possible. The key findings of this research offer profound insights into the biological consequences of smoking and hold considerable implications for both scientific inquiry and public health interventions.

The pSIN, derived from 51 proteins, demonstrates exceptional discriminative power, accurately distinguishing current smokers from never smokers in both the UK Biobank and China Kadoorie Biobank cohorts. Its performance, validated across diverse populations, highlights its robustness

and potential utility in global contexts. Unlike traditional biomarkers like exhaled carbon monoxide or plasma cotinine, which are limited to capturing short-term exposure, pSIN captures long-term biological responses to smoking, including systemic oxidative stress, immune dysregulation, and epithelial proliferation[75,211]. These features provide a deeper understanding of the molecular pathways affected by smoking, moving beyond behavioural reporting and enabling a biological characterization of smoking exposure.

The integration of genetic and exposome analyses adds depth to the interpretation of pSIN. The genetic analysis of pSIN revealed 95 significant loci through GWAS, offering critical insights into the heritability and biological pathways influenced by smoking. LDSC analysis demonstrated strong genetic correlations between pSIN and smoking-related traits such as current smoking ($r = 0.78$) and maternal smoking during pregnancy ($r = 0.66$), as well as metabolic and immune-related phenotypes like BMI ($r = 0.33$), diabetes ($r = 0.27$), and white blood cell counts. These findings emphasize the systemic effects of smoking and suggest shared genetic mechanisms underlying smoking exposure and broader health outcomes. The protective genetic correlation with Parkinson's disease ($r = -0.71$) aligns with known inverse epidemiological relationships, while associations with air pollution exposure metrics, such as PM2.5 and NO2, highlight pSIN's broader environmental relevance. Pathway enrichment analyses linked the identified loci to key processes including inflammation, lipid metabolism, and tissue repair, reinforcing the proteomic findings of systemic immune dysregulation and oxidative stress from smoking. Together, these genetic insights not only validate pSIN as a biomarker integrating genetic predispositions and environmental influences but also reveal its potential in identifying at-risk populations and elucidating molecular mechanisms for targeted interventions. Similarly, exposome analysis

demonstrates that factors such as maternal smoking, air pollution, and diet independently contribute to variations in pSIN, reinforcing the multifactorial nature of smoking-related health outcomes.

One of the most significant findings of the study is pSIN's strong association with increased risks of 18 major chronic diseases and all-cause mortality, highlighting its value in capturing the systemic health impacts of smoking. The analysis revealed significant links between smoking, as measured through plasma proteomics, and conditions affecting multiple organ systems, including cardiovascular diseases, chronic obstructive pulmonary disease (COPD), and cancers such as lung and head-and-neck cancers. Beyond serving as a general marker of smoking exposure, pSIN functions as a dynamic tool for stratifying risks across current, former, and never smokers, reflecting the heterogeneity in smoking-related health outcomes.

This stratification capability underscores pSIN's potential as a robust predictive tool in clinical and public health contexts. By identifying high-risk individuals, particularly those whose self-reported smoking history may not fully reflect their cumulative exposure or biological vulnerability, pSIN enables the tailoring of preventive and therapeutic interventions. Its application could improve population-level strategies for smoking-related disease prevention and focus resources on those most in need, ultimately enhancing health outcomes and reducing the burden of smoking-related diseases.

The study also highlights pSIN's relevance in assessing the biological recovery status among previous smokers. By analysing pSIN dynamics post-cessation, the research identifies subgroups of previous smokers whose proteomic profiles resemble those of never-smokers, alongside

others who continue to exhibit elevated pSIN levels and associated risks. Notably, the analysis also reveals that pSIN can identify previous smokers who, even after decades of cessation, exhibit disease risks comparable to current smokers for conditions like asthma, chronic kidney disease, chronic liver disease, and congestive cardiac failure.

Here, I then compared the associations between pSIN and metAgeGap and protAgeGap, adjusting for recruitment centre, ethnicity, education level and Townsend deprivation index (***Fig3***). Significant positive associations were found between pSIN and metAgeGap in both sexes and protAgeGap only in males. This is to compare with self-reported smoking status where significant association was only found in metAgeGap in females. The association between pSIN and metAgeGap was stronger compared to the association between pSIN and protAgeGap. This may be because GWAS results of metAgeGap are associated with deoxidation pathways which are not present in proteins composite protAgeGap.

Further, the overlap between pSIN and protAgeGap is relatively modest, with only 14 out of 51 proteins shared between the two scores (ACRV1, CSPG4, CST5, EDIL3, TSPAN1 exclusive to the male model, CD1C, IL7R, ISLR2, PRL to the female model, and CA4, CELSR2, CXCL17, KIT, MMP12 shared by both sexes) (***Fig 4***). This limited overlap suggests that while smoking and ageing intersect in certain biological pathways, they primarily drive distinct molecular processes. This is further supported by weak associations between self-reported smoking and both metabolic age and proteomic age, indicating that smoking's effects may act through indirect or distinct mechanisms not fully captured by ageing models.

The shared proteins—CA4, CELSR2, CXCL17, KIT, and MMP12—highlight pathways linked to inflammation, lipid metabolism, and oxidative stress. For instance, KIT and MMP12 are associated with immune regulation and tissue remodelling, including endothelial repair, reflecting systemic disruptions common to both smoking and ageing. Similarly, CA4 and CELSR2 are involved in lipid metabolism and vascular integrity, suggesting their role in shared pathways underpinning cardiovascular health. CXCL17, a chemokine involved in immune responses, underscores the intersection of chronic inflammation across the two processes. In general, pSIN incorporates proteins linked to the detoxification of smoking by-products, while protAgeGap includes markers that are mostly specific to cumulative age-related degradation.

The disease associations of pSIN and protAgeGap reveal both overlaps and distinctions. pSIN is strongly associated with smoking-related diseases, including lung cancer, COPD, and chronic liver disease, while protAgeGap is linked to broader age-related diseases, such as cardiovascular conditions, neurodegenerative disorders, and chronic kidney disease. Both scores are significantly associated with cardiovascular diseases and cancers, but the underlying mechanisms differ: pSIN reflects acute, exposure-driven pathways like inflammation and oxidative stress, while protAgeGap captures gradual, cumulative effects such as mitochondrial dysfunction and metabolic dysregulation. While both models highlight the interplay between smoking and ageing, their unique protein profiles and disease associations emphasize distinct biological processes—pSIN focuses on acute, systemic smoking effects, and protAgeGap on gradual ageing dynamics. Together, they provide a more comprehensive framework for predicting and addressing the health impacts of smoking and ageing across diverse biological contexts.

Limitations of the current study

Several inherent limitations of this study must be acknowledged to contextualise the findings and guide future research. Firstly, while relative quantification methods in proteomics were employed, which are suitable for large-scale cohort studies, these methods have inherent constraints that may limit translational applicability. Relative quantification lacks the precision of absolute quantification techniques, which could better facilitate the clinical application of ageing clocks. Future studies should prioritise integrating absolute quantification strategies to enhance the predictive accuracy and utility of these biomarkers in clinical settings.

Secondly, the cohort's size and diversity pose significant limitations. This constrained sample size also reduced statistical power, particularly for detecting associations between the pSIN and less common diseases, such as site-specific cancers. While associations with common diseases were identified with high confidence, addressing rarer conditions requires larger datasets or pooled analyses across multiple cohorts. This limitation underscores the need for expanded datasets to enable robust detection of associations across a broader spectrum of diseases.

Thirdly, this study analysed only a fraction of the human proteome and metabolome, utilising 2,917 proteins out of the estimated 20,000+ known proteins and 249 metabolites out of the 220,000+ metabolites[253,254]. This limited coverage, while focused on biologically relevant pathways, restricts the ability to fully capture the complexity of biological systems and may overlook important proteins associated with certain conditions. Expanding both the proteomic and metabolic library in future studies could unveil novel associations and provide a more comprehensive understanding of proteome-wide interactions with ageing and disease.

In summary, addressing these limitations will not only enhance the validity of findings but also pave the way for more precise and inclusive applications of multi-omics biomarkers in understanding and mitigating the biological impacts of ageing and smoking.

Future research

Validation in other cohorts

While our findings in UKB demonstrate robust associations between our omics-based ageing measures and health outcomes, UKB participants are on average healthier, more affluent, and less ethnically diverse than the general population. Future work will therefore validate our metabolic and proteomic ageing clocks in external cohorts with (i) a wider chronological age range—including much younger and much older adults—to test model performance at the extremes of the lifespan; (ii) greater ethnic and socioeconomic diversity to ensure transferability across ancestries and environmental exposures; and (iii) community-based, population-representative samples that capture the full spectrum of health status. These efforts will help us assess generalisability, recalibrate models where necessary, and ultimately produce ageing biomarkers with maximal applicability for global, real-world screening and intervention studies.

Combining multi-omics on a biological ageing model

Using only one omics layer, such as proteomics or metabolomics, to build models limits biological insights, as it captures only a fraction of the complex, interconnected processes driving ageing and disease. This limitation was evident in our study, where metabolic age demonstrated strong predictive power for incident kidney and liver diseases but performed

poorly in predicting neurodegenerative conditions. Conversely, proteomic age effectively predicted neurodegenerative diseases but lacked predictive accuracy for female-specific cancers.

The integration of metabolomics and proteomics may provide a more synergistic framework to investigate the complex interplay between molecular changes and health outcomes, particularly ageing and age-associated diseases. Proteomics offers a comprehensive view of protein-level alterations, which are essential for decoding pathways related to inflammation, immune responses, and cellular signalling. These processes are central to understanding ageing mechanisms and their role in conditions such as respiratory and neurodegenerative diseases, where protein misfolding and chronic inflammation are key pathological features. Complementarily, metabolomics captures dynamic metabolic shifts, reflecting real-time alterations in lipid and glucose metabolism, oxidative stress, and energy production. These processes are pivotal for maintaining organ health, including the brain, cardiovascular system, kidneys, and liver, which are often compromised in ageing populations.

By integrating these datasets, we can also discover molecular linkages between metabolic and proteomic pathways, enabling the identification of cascade effects where metabolic dysfunctions lead to proteomic alterations or vice versa. Such insights are critical for elucidating the mechanisms driving age-related diseases and offering opportunities for early intervention.

In conclusion, the combined use of metabolomics and proteomics may transcend the limitations of individual omics approaches, offering a robust platform to explore the intricacies of ageing and its associated diseases. This integrative strategy will not only enhance our mechanistic

understanding but also pave the way for personalized therapeutic approaches aimed at improving health span and longevity. Through this lens, biological ageing will be reframed as a multi-dimensional process, underpinned by interwoven metabolic and proteomic networks that collectively shape health outcomes.

Validating in longitudinal data

The validation of both the ageing and smoking index models using longitudinal data is essential to ensure their robustness and applicability in real-world settings. Longitudinal data enables the tracking of temporal changes, providing insights into the causal relationships between the models' predictions and health outcomes. For the ageing model, longitudinal validation helps capture the dynamic nature of biological ageing and its association with incident morbidities, mortality, and other age-related biomarkers. Such validation ensures the model's predictions remain consistent over time, reinforcing its use in predicting long-term health risks.

Similarly, for the smoking index model, longitudinal data is critical for evaluating its ability to reflect the cumulative impact of smoking on health outcomes over extended periods. This validation helps in understanding how cessation or continued smoking behaviours influence risk trajectories, allowing the model to identify residual risks among former smokers or accurately stratify risks in current smokers.

Expansion of Omics Research to Lifestyle and Environmental Risk Factors

Building on the success of proteomic smoking indices, the integration of proteomics into the study of broader exposomes such as alcohol intake, diet, and air pollution presents an opportunity to address limitations in self-reported measures of these traits. Alcohol intake, for

instance, exhibits a J-shaped relationship with various health outcomes, where low-to-moderate consumption has been associated with protective effects, while heavy consumption markedly increases disease risks[255]. This relationship complicates epidemiological analyses, especially as self-reported alcohol consumption is prone to recall bias and social desirability effects.

Proteomic approaches offer a promising solution by capturing biological signatures that reflect cumulative and nuanced exposure levels. Proteins can serve as biomarkers of chronic alcohol consumption patterns and their associated physiological impacts, potentially offering clarity to the mechanisms underpinning the J-shaped relationship. For example, specific proteomic markers related to liver function, oxidative stress, and inflammatory pathways could differentiate between low, moderate, and high levels of alcohol consumption and their unique contributions to ageing and disease risks.

Additionally, proteomics can mitigate the inaccuracies inherent in dietary assessments and air pollution exposure estimates[25]. Current methods often rely on self-reports or geographic proxies, which lack precision. By contrast, the plasma proteome could provide direct and integrative measures of nutritional and environmental exposures, capturing their cumulative effects on biological ageing processes. For instance, proteins linked to oxidative stress and inflammation may serve as indicators of long-term air pollution exposure.

By modelling the interaction of these lifestyle and environmental factors within a proteomic framework, we can develop a more comprehensive understanding of their joint impact on ageing pathways. This approach has the potential to improve risk stratification, identify high-risk

populations, and inform personalized interventions targeting modifiable exposures to slow biological ageing and reduce disease burden.

Conclusion

In summary, this thesis demonstrates the power of gradient-boosted models—particularly LightGBM—in uncovering sex-specific and multi-omics profiles of biological ageing and smoking exposure, revealing distinct proteomic and metabolomic signatures that differentially predict disease risk across males and females. By comparing proteomic and metabolic clocks, we highlight the superior robustness of proteomic age in capturing a broader spectrum of age-related outcomes, while metabolic age uniquely reflects adiposity-linked pathways. The development of the pSIN further underscores how targeted protein panels can quantify the long-term effects of lifestyle risk factors that contribute to ageing and stratify chronic disease risk beyond traditional self-report measures.

Figures

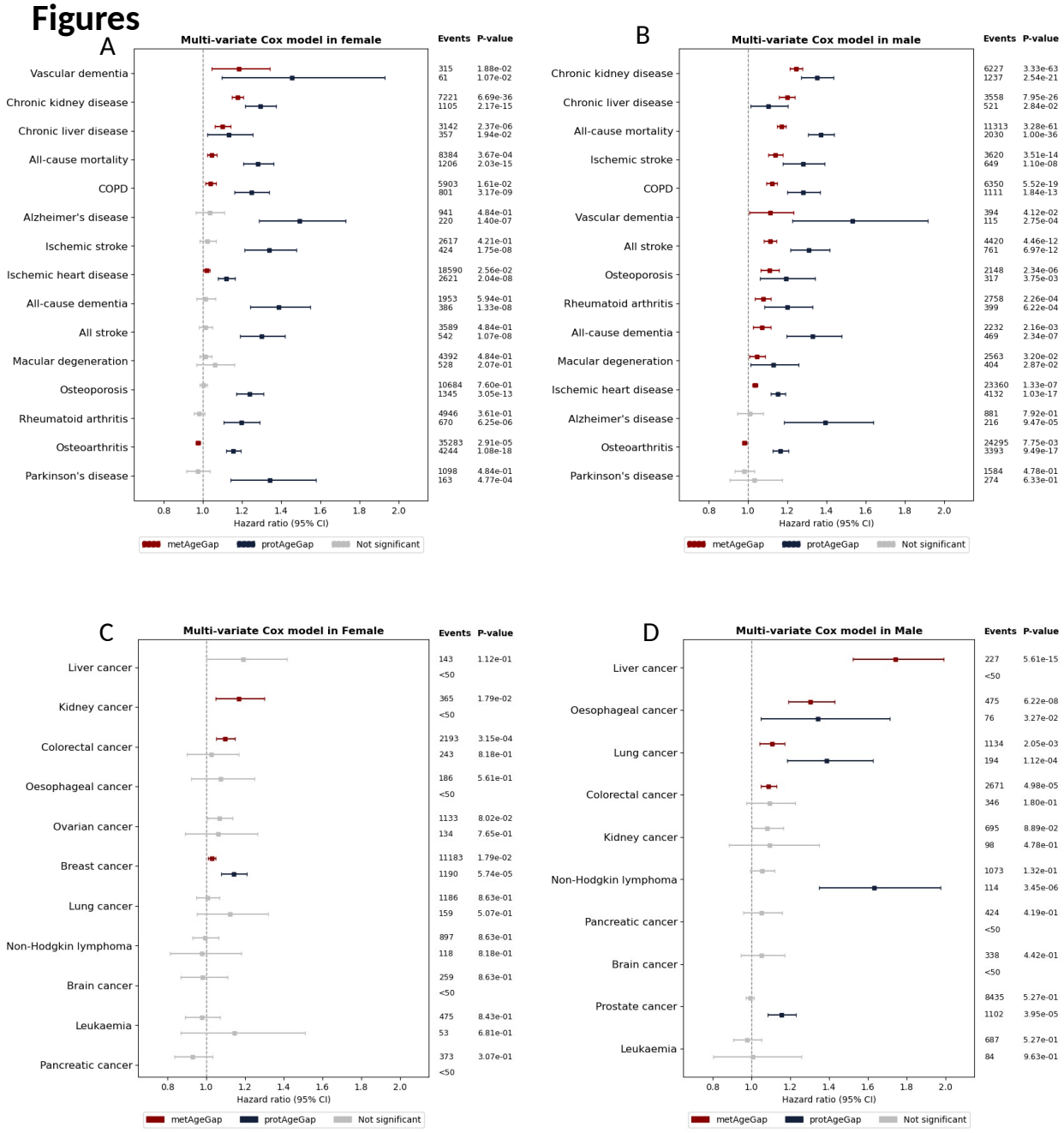


Figure 1: Comparison of associations of incident diseases between metAgeGap and protAgeGap. Forest plot shows associations of metAgeGap (red) and protAgeGap (blue) with incident diseases adjusting for chronological age, recruitment center, Townsend deprivation index, ethnicity, IPAQ activity groups, BMI and smoking status. Grey colour shows if the association is none-significant after FDR adjustments. Associations are only shown if the incident case number is above 50. (A) shows the associations with common diseases in female. (B) shows the associations with common diseases in male. (C) shows the associations with cancers in female. (D) shows the associations with cancers in male.

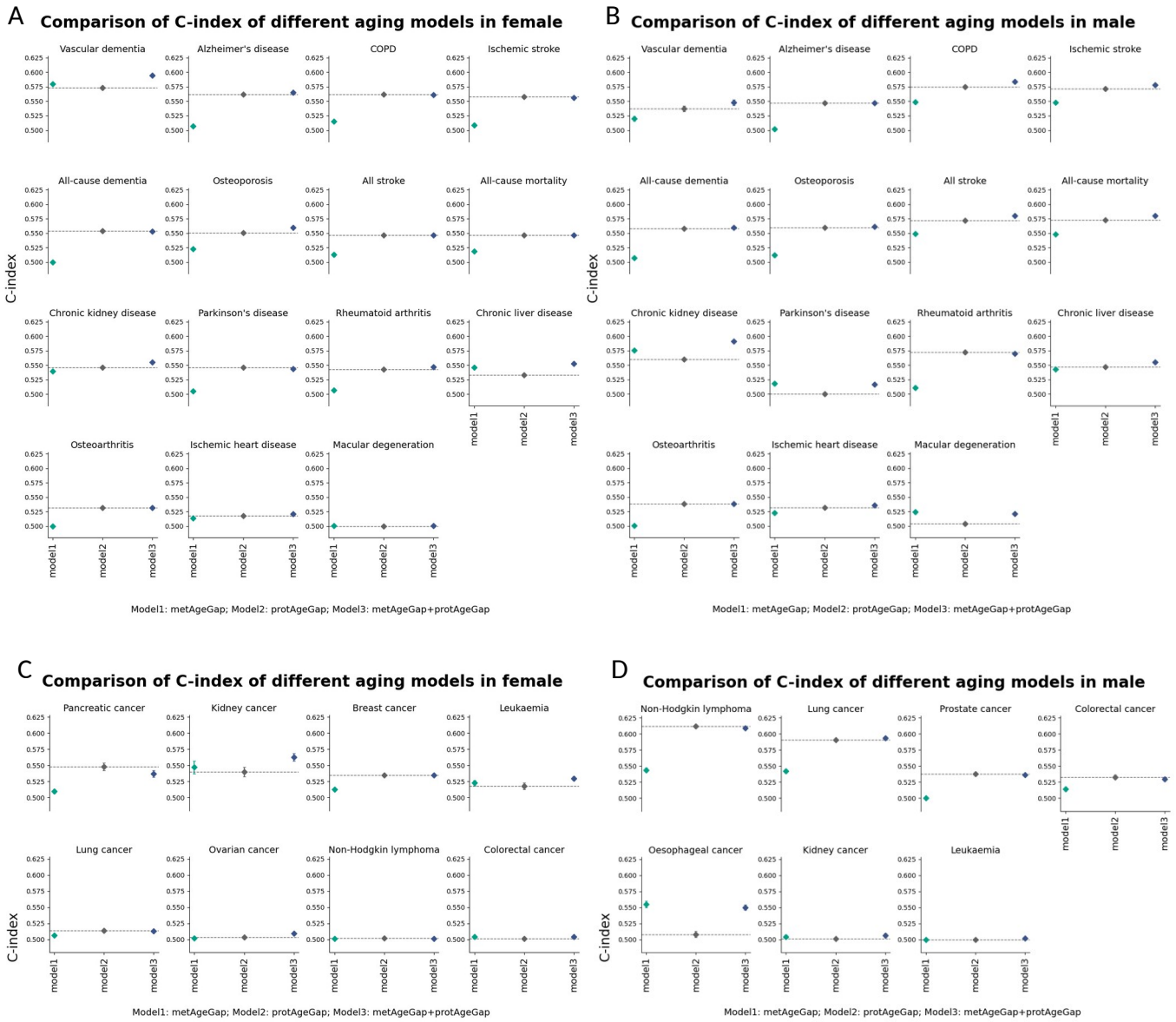


Figure 2: Comparison of C-index between metAgeGap and protAgeGap from cox models. The plot shows c-index of cox model with metAgeGap (green), protAgeGap (grey) and metAgeGap+protAgeGap (blue) to indicate if protAgeGap has additive value comparing to metAgeGap when predict future disease risks. 95% confidence interval was calculated by bootstrapping the cox model for 100 times. (A) shows the C-index when predicting common diseases in female. (B) shows the C-index when predicting common diseases in male. (C) shows the C-index when predicting cancers in female. (D) shows the C-index cancers when predicting cancers in male.

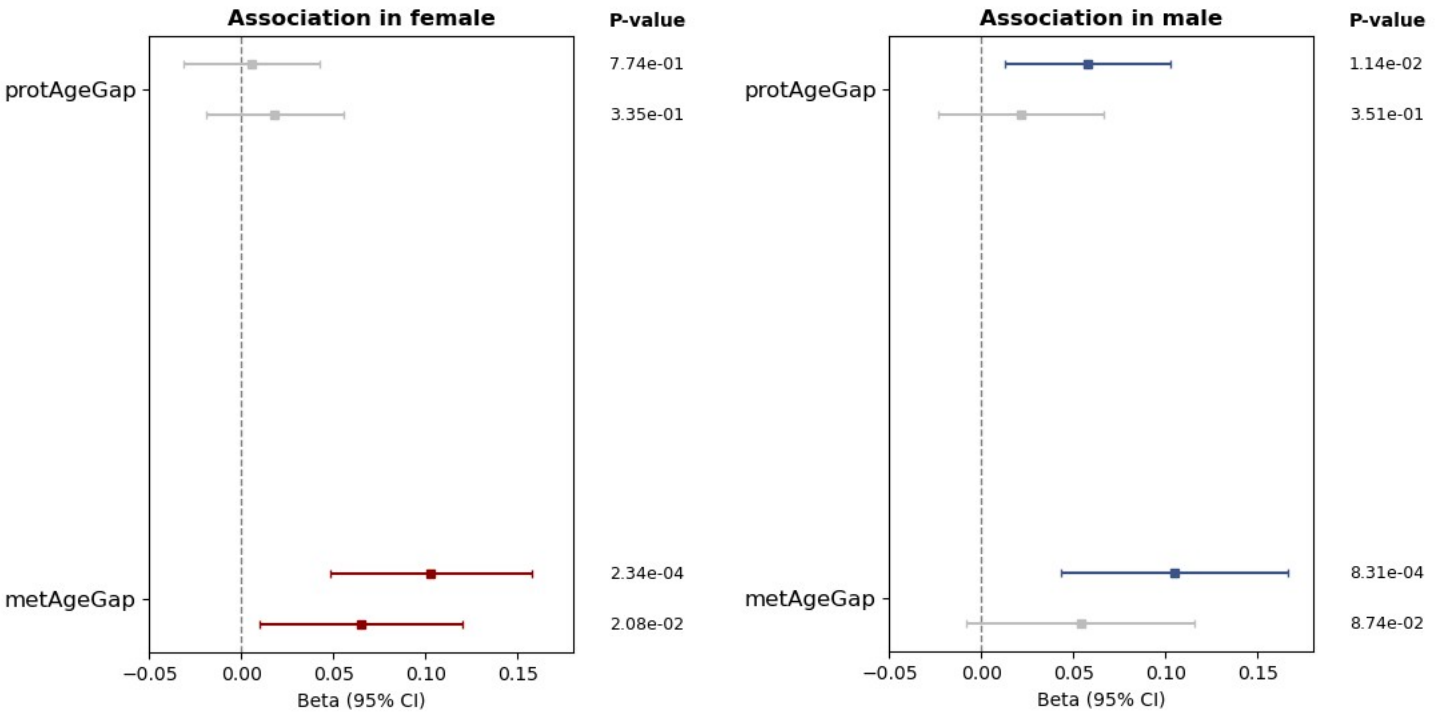


Figure 3: Association between pSIN and metAgeGap and protAgeGap. Linear association was performed adjusting for recruitment center, ethnicity, education years, and Townsend deprivation index. Beta value was plotted with 95% CI. Model 1 denoted the beta value of pSIN predicted smoking status and Model 2 denoted the beta value of self-reported smoking status

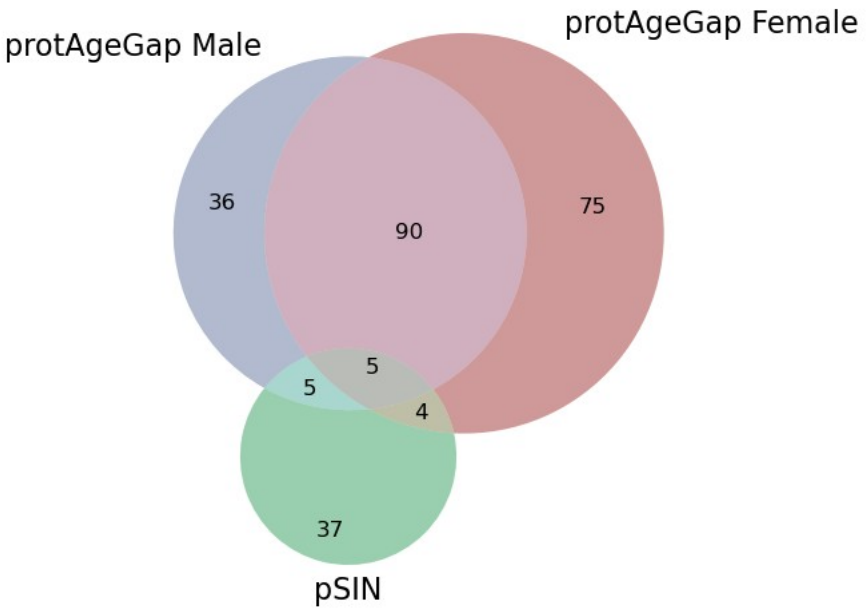


Figure 4: Shared proteins between protAgeGap and pSIN. Venn plot shows the shared proteins between protAgeGap male and female model and pSIN model after Boruta selection.

References

1. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
2. *Decade of Healthy Ageing: Baseline Report*. (World Health Organization, Geneva, 2021).
3. Brooks, R. C. & Garratt, M. G. Life history evolution, reproduction, and the origins of sex-dependent aging and longevity. *Annals of the New York Academy of Sciences* **1389**, 92–107 (2017).
4. Gilbert, S. F. Aging: The Biology of Senescence. in *Developmental Biology. 6th edition* (Sinauer Associates, 2000).
5. Alberts, S. C. *et al*. Reproductive aging patterns in primates reveal that humans are distinct. *Proc Natl Acad Sci U S A* **110**, 13440–13445 (2013).
6. Wood, B. M. *et al*. Demographic and hormonal evidence for menopause in wild chimpanzees. *Science* https://doi.org/10.1126/science.add5473 (2023) doi:10.1126/science.add5473.
7. Ellis, S., Franks, D. W., Nielsen, M. L. K., Weiss, M. N. & Croft, D. P. The evolution of menopause in toothed whales. *Nature* **627**, 579–585 (2024).
8. Hadza Women's Time Allocation, Offspring Provisioning, and the Evolution of Long Postmenopausal Life Spans | Current Anthropology: Vol 38, No 4. https://www.journals.uchicago.edu/doi/abs/10.1086/204646.
9. Barone, B. *et al*. The Role of Testosterone in the Elderly: What Do We Know? *Int J Mol Sci* **23**, 3535 (2022).

10. Santoro, N. & Randolph, J. F. Reproductive Hormones and the Menopause Transition. *Obstet Gynecol Clin North Am* **38**, 455–466 (2011).
11. National life tables – life expectancy in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2020to2022.
12. Yan, B. W. *et al*. Widening Gender Gap in Life Expectancy in the US, 2010-2021. *JAMA Internal Medicine* **184**, 108–110 (2024).
13. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: An expanding universe. *Cell* **186**, 243–278 (2023).
14. Campisi, J. & d'Adda di Fagagna, F. Cellular senescence: when bad things happen to good cells. *Nat Rev Mol Cell Biol* **8**, 729–740 (2007).
15. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12–23 (2020).
16. Amjad, S. *et al*. Role of NAD+ in regulating cellular and metabolic signaling pathways. *Mol Metab* **49**, 101195 (2021).
17. Kanfi, Y. *et al*. The sirtuin SIRT6 regulates lifespan in male mice. *Nature* **483**, 218–221 (2012).
18. de Souza-Pinto, N. C. *et al*. Repair of 8-oxodeoxyguanosine lesions in mitochondrial dna depends on the oxoguanine dna glycosylase (OGG1) gene and 8-oxoguanine accumulates in the mitochondrial dna of OGG1-defective mice. *Cancer Res* **61**, 5378–5381 (2001).
19. Blackburn, E. H. & Epel, E. S. Too toxic to ignore. *Nature* **490**, 169–171 (2012).
20. Schneider, C. V. *et al*. Association of Telomere Length With Risk of Disease and Mortality. *JAMA Internal Medicine* **182**, 291–300 (2022).

21. Rossiello, F., Jurk, D., Passos, J. F. & d'Adda di Fagagna, F. Telomere dysfunction in ageing and age-related diseases. *Nat Cell Biol* **24**, 135–147 (2022).
22. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, 3156 (2013).
23. Seale, K., Horvath, S., Teschendorff, A., Eynon, N. & Voisin, S. Making sense of the ageing methylome. *Nat Rev Genet* **23**, 585–605 (2022).
24. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* **19**, 371–384 (2018).
25. Kapahi, P., Kaeberlein, M. & Hansen, M. Dietary restriction and lifespan: Lessons from invertebrate models. *Ageing Res Rev* **39**, 3–14 (2017).
26. Amorim, J. A. *et al*. Mitochondrial and metabolic dysfunction in ageing and age-related diseases. *Nat Rev Endocrinol* **18**, 243–258 (2022).
27. Miller, H. A., Dean, E. S., Pletcher, S. D. & Leiser, S. F. Cell non-autonomous regulation of health and longevity. *eLife* **9**, e62659 (2020).
28. Selman, M. & Pardo, A. Fibroageing: An ageing pathological feature driven by dysregulated extracellular matrix-cell mechanobiology. *Ageing Research Reviews* **70**, 101393 (2021).
29. Franceschi, C. *et al*. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mech Ageing Dev* **128**, 92–105 (2007).
30. Calder, P. C. *et al*. Health relevance of the modification of low grade inflammation in ageing (inflammageing) and the role of nutrition. *Ageing Res Rev* **40**, 95–119 (2017).
31. Li, X. *et al*. Inflammation and aging: signaling pathways and intervention therapies. *Sig Transduct Target Ther* **8**, 1–29 (2023).

32. Hipp, M. S., Kasturi, P. & Hartl, F. U. The proteostasis network and its decline in ageing. *Nat Rev Mol Cell Biol* **20**, 421–435 (2019).

33. Fang, S., Holmes, M. V., Gaunt, T. R., Smith, G. D. & Richardson, T. G. *An Atlas of Associations between Polygenic Risk Scores from across the Human Phenome and Circulating Metabolic Biomarkers*. http://medrxiv.org/lookup/doi/10.1101/2021.10.14.21265005 (2021) doi:10.1101/2021.10.14.21265005.

34. Rh, H. & Bjjm, B. Proteostasis in cardiac health and disease. *Nature reviews. Cardiology* **14**, (2017).

35. Ottens, F., Franz, A. & Hoppe, T. Build-UPS and break-downs: metabolism impacts on proteostasis and aging. *Cell Death Differ* **28**, 505–521 (2021).

36. Gressler, A. E., Leng, H., Zinecker, H. & Simon, A. K. Proteostasis in T cell aging. *Semin Immunol* **70**, 101838 (2023).

37. Morimoto, R. I. & Cuervo, A. M. Proteostasis and the Aging Proteome in Health and Disease. *J Gerontol A Biol Sci Med Sci* **69**, S33–S38 (2014).

38. Carter, A. J. *et al*. Target 2035: probing the human proteome. *Drug Discovery Today* **24**, 2111–2115 (2019).

39. Eisenberg, T. *et al*. Cardioprotection and lifespan extension by the natural polyamine spermidine. *Nat Med* **22**, 1428–1438 (2016).

40. Rutledge, J., Oh, H. & Wyss-Coray, T. Measuring biological age using omics data. *Nat Rev Genet* https://doi.org/10.1038/s41576-022-00511-7 (2022) doi:10.1038/s41576-022-00511-7.

41. Lu, A. T. *et al*. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).
42. Ikram, M. A. The use and misuse of ‘biological aging’ in health research. *Nat Med* **30**, 3045–3045 (2024).
43. Chen, W. *et al*. Three-dimensional human facial morphologies as robust aging markers. *Cell Res* **25**, 574–587 (2015).
44. Zhavoronkov, A., Li, R., Ma, C. & Mamoshina, P. Deep biomarkers of aging and longevity: from research to applications. *Aging* **11**, 10771–10780 (2019).
45. Song, H. *et al*. Epigenetic modification in Parkinson’s disease. *Front Cell Dev Biol* **11**, 1123621 (2023).
46. Stoeger, T. *et al*. Aging is associated with a systemic length-associated transcriptome imbalance. *Nat Aging* **2**, 1191–1206 (2022).
47. Hu, W. *et al*. Systematic characterization of cancer transcriptome at transcript resolution. *Nat Commun* **13**, 6803 (2022).
48. Harries, L. W. *et al*. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell* **10**, 868–878 (2011).
49. Kulminski, A. M., Philipp, I., Shu, L. & Culminskaya, I. Definitive Roles of TOMM40-APOE-APOC1 Variants in the Alzheimer’s Risk. *Neurobiol Aging* **110**, 122–131 (2022).
50. Arrell, D. K., Neverova, I. & Van Eyk, J. E. Cardiovascular Proteomics. *Circulation Research* **88**, 763–773 (2001).
51. Argentieri, M. A. *et al. Proteomic Aging Clock Predicts Mortality and Risk of Common Age-Related Diseases in Diverse Populations*.

http://medrxiv.org/lookup/doi/10.1101/2023.09.13.23295486 (2023) doi:10.1101/2023.09.13.23295486.

52. Nagana Gowda, G. A. *et al*. Metabolomics-Based Methods for Early Disease Diagnostics: A Review. *Expert Rev Mol Diagn* **8**, 617–633 (2008).
53. Schmidt, D. R. *et al*. Metabolomics in Cancer Research and Emerging Applications in Clinical Oncology. *CA Cancer J Clin* **71**, 333–358 (2021).
54. Mutz, J., Iniesta, R. & Lewis, C. M. Metabolomic age (MileAge) predicts health and life span: A comparison of multiple machine learning algorithms. *Science Advances* **10**, eadp3743 (2024).
55. Kirkwood, T. B. Evolution of ageing. *Nature* **270**, 301–304 (1977).
56. Kaeberlein, M., Rabinovitch, P. S. & Martin, G. M. Healthy aging: The ultimate preventative medicine. *Science* **350**, 1191–1193 (2015).
57. Schächter, F. *et al*. Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet* **6**, 29–32 (1994).
58. Lumsden, A. L., Mulugeta, A., Zhou, A. & Hyppönen, E. Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank. *eBioMedicine* **59**, (2020).
59. Huang, Y. & Mahley, R. W. Apolipoprotein E: Structure and Function in Lipid Metabolism, Neurobiology, and Alzheimer's Diseases. *Neurobiol Dis* **72PA**, 3–12 (2014).
60. Willcox, B. J. *et al*. FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci U S A* **105**, 13987–13992 (2008).

61. Donlon, T. A. *et al*. FOXO3, a Resilience Gene: Impact on Lifespan, Healthspan, and Deathspan. *J Gerontol A Biol Sci Med Sci* **77**, 1479–1484 (2022).
62. Sanese, P., Forte, G., Disciglio, V., Grossi, V. & Simone, C. FOXO3 on the Road to Longevity: Lessons From SNPs and Chromatin Hubs. *Comput Struct Biotechnol J* **17**, 737–745 (2019).
63. Willcox, B. J. *et al*. The FoxO3 gene and cause-specific mortality. *Aging Cell* **15**, 617–624 (2016).
64. Caruso, C. *et al*. How Important Are Genes to Achieve Longevity? *Int J Mol Sci* **23**, 5635 (2022).
65. Chen, C., Zhou, M., Ge, Y. & Wang, X. SIRT1 and aging related signaling pathways. *Mechanisms of Ageing and Development* **187**, 111215 (2020).
66. Papadopoli, D. *et al*. mTOR as a central regulator of lifespan and aging. *F1000Res* **8**, F1000 Faculty Rev-998 (2019).
67. Castner, S. A. *et al*. Longevity factor klotho enhances cognition in aged nonhuman primates. *Nat Aging* **3**, 931–937 (2023).
68. Shim, H. S. *et al*. TERT activation targets DNA methylation and multiple aging hallmarks. *Cell* **187**, 4030-4042.e13 (2024).
69. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing. *Nature* **561**, 45–56 (2018).
70. Argentieri, M. A. & Amin, N. Integrating the environmental and genetic architectures of aging and mortality. *Nature Medicine* (2025).
71. Zhavoronkov, A. *et al*. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews* **49**, 49–66 (2019).

72. Zuo, L. *et al*. Interrelated role of cigarette smoking, oxidative stress, and immune response in COPD and corresponding treatments. *Am J Physiol Lung Cell Mol Physiol* **307**, L205-218 (2014).

73. Zhou, M. *et al*. Aging and Cardiovascular Disease: Current Status and Challenges. *RCM* **23**, 135 (2022).

74. Lunetta, K. L. *et al*. Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. *BMC Med Genet* **8**, S13 (2007).

75. Montalto, N. J. & Wells, W. O. Validation of Self-Reported Smoking Status Using Saliva Cotinine: A Rapid Semiquantitative Dipstick Method. *Cancer Epidemiology, Biomarkers & Prevention* **16**, 1858–1862 (2007).

76. Deveci, S. E., Deveci, F., Açik, Y. & Ozan, A. T. The measurement of exhaled carbon monoxide in healthy smokers and non-smokers. *Respir Med* **98**, 551–556 (2004).

77. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).

78. BIOS Consortium *et al*. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur J Epidemiol* **34**, 1055–1074 (2019).

79. Palmer, M. *et al*. The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: A systematic review of randomised controlled trials. *PLoS One* **13**, e0189801 (2018).

80. Shen, X. *et al*. Nonlinear dynamics of multi-omics profiles during human aging. *Nat Aging* **4**, 1619–1634 (2024).
81. Bunning, B. J. *et al*. Global metabolic profiling to model biological processes of aging in twins. *Aging Cell* **19**, e13073 (2020).
82. Ke, G. *et al*. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
83. Hägg, S. & Jylhävä, J. Sex differences in biological aging with a focus on human studies. *eLife* **10**, e63425.
84. Merz, A. A. & Cheng, S. Sex differences in cardiovascular ageing. *Heart* **102**, 825–831 (2016).
85. Horvath, S. *et al*. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology* **17**, 171 (2016).
86. Sudlow, C. *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
87. Collins, R. UK Biobank Protocol.
88. Sun, B. B. *et al*. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
89. Chen, Z. *et al*. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652–1666 (2011).
90. Kurki, M. I. *et al*. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
91. Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* 6.

92. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3292500.3330701.
93. Ng, A. Y. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. in *Twenty-first international conference on Machine learning - ICML '04* 78 (ACM Press, Banff, Alberta, Canada, 2004). doi:10.1145/1015330.1015435.
94. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020).
95. Cerliani, M. shap-hypetune. (2022).
96. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the **Boruta** Package. *J. Stat. Soft.* **36**, (2010).
97. Argentieri, M. A. *et al.* Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat Med* **30**, 2450–2460 (2024).
98. El-Hayek, S., Demeestere, I. & Clarke, H. J. Follicle-stimulating hormone regulates expression and activity of epidermal growth factor receptor in the murine ovarian follicle. *Proc Natl Acad Sci U S A* **111**, 16778–16783 (2014).
99. Garcia-Alonso, L. *et al.* Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat Genet* **53**, 1698–1711 (2021).
100. Thorek, D. L., Evans, M. J., Carlsson, S. V., Ulmert, D. & Lilja, H. Prostate Specific Kallikrein-related Peptidases and Their Relation to Prostate Cancer Biology and Detection; Established Relevance and Emerging Roles. *Thromb Haemost* **110**, 484–492 (2013).

101. Dattani, S., Rodés-Guirao, L., Ritchie, H., Ortiz-Ospina, E. & Roser, M. L. E. Our World Data. (2023).

102. Gordon, E. H. Sex differences in frailty: a systematic review and meta-analysis. *Exp. Gerontol* **89**, 30–40 (2017).

103. Peiffer, J. J. Strength and functional characteristics of men and women 65 years and older. *Rejuvenation Res* **13**, 75–82 (2010).

104. Pradhan, A. & Olsson, P.-E. Sex differences in severity and mortality from COVID-19: are males more vulnerable? *Biology of sex Differences* **11**, 53 (2020).

105. Nasiri, M. J. *et al*. COVID-19 clinical characteristics, and sex-specific risk of mortality: systematic review and meta-analysis. *Frontiers in medicine* **7**, 459 (2020).

106. Zarulli, V. *et al*. Women live longer than men even during severe famines and epidemics. *Proceedings of the National Academy of Sciences* **115**, E832–E840 (2018).

107. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L. Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nature Reviews Immunology* **20**, 442–447 (2020).

108. Gillies, G. E. & McArthur, S. Estrogen actions in the brain and the basis for differential action in men and women: a case for sex-specific medicines. *Pharmacological reviews* **62**, 155–198 (2010).

109. Palmisano, B. T., Zhu, L., Eckel, R. H. & Stafford, J. M. Sex differences in lipid and lipoprotein metabolism. *Molecular metabolism* **15**, 45–55 (2018).

110. Lulkiewicz, M., Bajsert, J., Kopczynski, P., Barczak, W. & Rubis, B. Telomere length: how the length makes a difference. *Molecular Biology Reports* **47**, 7181–7188 (2020).

111. Gardner, M. *et al*. Gender and telomere length: Systematic review and meta-analysis. *Experimental gerontology* **51**, 15 (2013).

112. Horstman, A. M., Dillon, E. L., Urban, R. J. & Sheffield-Moore, M. The role of androgens and estrogens on healthy aging and longevity. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **67**, 1140–1152 (2012).

113. Harris, C. R. & Jenkins, M. Gender Differences in Risk Assessment: Why do Women Take Fewer Risksthan Men? *Judgment and Decision making* **1**, 48–63 (2006).

114. Stanton, S. J., Liening, S. H. & Schultheiss, O. C. Testosterone is positively associated with risk taking in the Iowa Gambling Task. *Hormones and behavior* **59**, 252–256 (2011).

115. Lipsky, M. S., Su, S., Crespo, C. J. & Hung, M. Men and oral health: a review of sex and gender differences. *American journal of men's health* **15**, 15579883211016361 (2021).

116. Villiers-Tuthill, A., Copley, A., McGee, H. & Morgan, K. The relationship of tobacco and alcohol use with ageing self-perceptions in older people in Ireland. *BMC public health* **16**, 1–10 (2016).

117. Yusipov, I. *et al*. Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *Aging (Albany NY)* **12**, 24057 (2020).

118. Johnson, A. A. *et al*. The role of DNA methylation in aging, rejuvenation, and age-related disease. *Rejuvenation research* **15**, 483–494 (2012).

119. Van Den Akker, E. B. *et al*. Metabolic age based on the BBMRI-NL 1H-NMR metabolomics repository as biomarker of age-related disease. *Circulation: Genomic and Precision Medicine* **13**, 541–547 (2020).

120. Lehallier, B. *et al*. Undulating changes in human plasma proteome profiles across the lifespan. *Nature medicine* **25**, 1843–1850 (2019).

121. Menni, C. *et al*. Metabolomic markers reveal novel pathways of ageing and early development in human populations. *International Journal of Epidemiology* **42**, 1111–1119 (2013).

122. Reicher, L. *et al*. Phenome-wide associations of human aging uncover sex-specific dynamics. *Nat Aging* **4**, 1643–1655 (2024).

123. Manuel, R. S. J. & Liang, Y. Sexual Dimorphism in Immunometabolism and Autoimmunity: Impact on Personalized Medicine. *Autoimmun Rev* **20**, 102775 (2021).

124. Link, J. C. & Reue, K. The Genetic Basis for Sex Differences in Obesity and Lipid Metabolism. *Annu Rev Nutr* **37**, 225–245 (2017).

125. Roshandel, D., Lu, T., Paterson, A. D. & Dash, S. Beyond apples and pears: sex-specific genetics of body fat percentage. *Front. Endocrinol.* **14**, (2023).

126. Mayer, M. missRanger: Fast Imputation of Missing Values. (2023).

127. Zhang, Z., Gayle, A. A., Wang, J., Zhang, H. & Cardinal-Fernández, P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. *Ann Transl Med* **5**, 484 (2017).

128. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in 92–96 (Austin, Texas, 2010). doi:10.25080/Majora-92bf1922-011.

129. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317 (2019).

130. Zhou, W. *et al*. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet* **54**, 1466–1469 (2022).

131. Purcell, S. *et al*. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).

132. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

133. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

134. Wu, T. *et al*. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

135. Deelen, J. *et al*. A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat Commun* **10**, 3669 (2019).

136. Sivanesan, S., Taylor, A., Zhang, J. & Bakovic, M. Betaine and Choline Improve Lipid Homeostasis in Obesity by Participation in Mitochondrial Oxidative Demethylation. *Front. Nutr.* **5**, (2018).

137. Ridgway, N. D. The role of phosphatidylcholine and choline metabolites to cell proliferation and survival. *Critical Reviews in Biochemistry and Molecular Biology* **48**, 20–38 (2013).

138. Miller, M. *et al*. Triglycerides and Cardiovascular Disease. *Circulation* **123**, 2292–2333 (2011).

139. El Harchaoui, K. *et al*. High-density lipoprotein particle size and concentration and coronary risk. *Ann Intern Med* **150**, 84–93 (2009).

140. Traub, M. L. & Santoro, N. Reproductive aging and its consequences for general health. *Ann N Y Acad Sci* **1204**, 179–187 (2010).

141. Fan, G. *et al*. Reproductive factors and biological aging: the association with all-cause and cause-specific premature mortality. *Hum Reprod* **40**, 148–156 (2025).

142. Kuningas, M. *et al*. The relationship between fertility and lifespan in humans. *AGE* **33**, 615–622 (2011).

143. Williams, G. C. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Science of Aging Knowledge Environment* **2001**, cp13–cp13 (2001).

144. Kirkwood, T. B. L. & Rose, M. R. Evolution of senescence: late survival sacrificed for reproduction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **332**, 15–24 (1997).

145. McArdle, P. F. *et al*. Does Having Children Extend Life Span? A Genealogical Study of Parity and Longevity in the Amish. *The Journals of Gerontology: Series A* **61**, 190–195 (2006).

146. Fetal microchimerism and maternal health during and after pregnancy - Keelin O'Donoghue, 2008. https://journals.sagepub.com/doi/full/10.1258/om.2008.080008 (2025).

147. Tauqeer, Z., Gomez, G. & Stanford, F. C. Obesity in women: insights for the clinician. *Journal of Women's Health* **27**, 444–457 (2018).

148. Chaston, T. B. & Dixon, J. B. Factors associated with percent change in visceral versus subcutaneous abdominal fat during weight loss: findings from a systematic review. *Int J Obes* **32**, 619–628 (2008).

149. Agrawal, S. *et al*. Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat Commun* **13**, 3771 (2022).

150. Kuk, J. L. *et al*. Visceral fat is an independent predictor of all-cause mortality in men. *Obesity (Silver Spring)* **14**, 336–341 (2006).

151. Khan, M. A. *et al*. Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. *Cureus* **12**, e9349 (2025).

152. Liberopoulos, E. *et al*. Gender Differences in Familial Hypercholesterolemia: Insight From the HELLAS-FH Registry. *Circulation* **138**, A12838–A12838 (2018).

153. Timmers, P. R. *et al*. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* **8**, e39856 (2019).

154. Yengo, L. *et al*. Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641–3649 (2018).

155. Okbay, A. *et al*. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat Genet* **54**, 437–449 (2022).

156. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. *Nat Genet* **50**, 906–908 (2018).

157. Joseph, N. M., Ramamoorthy, L. & Satheesh, S. Atypical Manifestations of Women Presenting with Myocardial Infarction at Tertiary Health Care Center: An Analytical Study. *J Midlife Health* **12**, 219–224 (2021).

158. Canto, J. G. *et al*. Symptom Presentation of Women With Acute Coronary Syndromes: Myth vs Reality. *Archives of Internal Medicine* **167**, 2405–2413 (2007).

159. Wik, L. *et al*. Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell Proteomics* **20**, 100168 (2021).

160. Szklarczyk, D. *et al*. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).

161. Hemmings, B. A. & Restuccia, D. F. PI3K-PKB/Akt Pathway. *Cold Spring Harb Perspect Biol* **4**, a011189 (2012).

162. Schnellmann, R. & Gerecht, S. Reconstructing the ageing extracellular matrix. *Nat Rev Bioeng* **1**, 458–459 (2023).

163. Guertin, D. A. *et al*. Ablation in Mice of the mTORC Components raptor, rictor, or mLST8 Reveals that mTORC2 Is Required for Signaling to Akt-FOXO and PKCα, but Not S6K1. *Developmental Cell* **11**, 859–871 (2006).

164. Martins, R., Lithgow, G. J. & Link, W. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell* **15**, 196–207 (2016).

165. Lawrence, M. S. *et al*. Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature* **505**, 495–501 (2014).

166. He, Y. *et al*. Targeting PI3K/Akt signal transduction for cancer therapy. *Sig Transduct Target Ther* **6**, 1–17 (2021).

167. Mendoza, M. C., Er, E. E. & Blenis, J. The Ras-ERK and PI3K-mTOR Pathways: Cross-talk and Compensation. *Trends Biochem Sci* **36**, 320–328 (2011).

168. Yu, C. F., Liu, Z.-X. & Cantley, L. G. ERK negatively regulates the epidermal growth factor-mediated interaction of Gab1 and the phosphatidylinositol 3-kinase. *J Biol Chem* **277**, 19382–19388 (2002).

169. Kodaki, T. *et al*. The activation of phosphatidylinositol 3-kinase by Ras. *Curr Biol* **4**, 798–806 (1994).

170. Molina, J. R. & Adjei, A. A. The Ras/Raf/MAPK Pathway. *Journal of Thoracic Oncology* **1**, 7–9 (2006).

171. Kranenburg, O., Gebbink, M. F. B. G. & Voest, E. E. Stimulation of angiogenesis by Ras proteins. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1654**, 23–37 (2004).

172. Schüler, S. C. *et al*. Extensive remodeling of the extracellular matrix during aging contributes to age-dependent impairments of muscle stem cell functionality. *Cell Reports* **35**, (2021).

173. Zaidi, M. *et al*. FSH, Bone Mass, Body Fat, and Biological Aging. *Endocrinology* **159**, 3503 (2018).

174. Bieche, I. *et al*. Identification of CGA as a novel estrogen receptor-responsive gene in breast cancer: an outstanding candidate marker to predict the response to endocrine therapy. *Cancer Res* **61**, 1652–1658 (2001).

175. Rivera, R. T., Pasion, S. G., Wong, D. T., Fei, Y. B. & Biswas, D. K. Loss of tumorigenic potential by human lung tumor cells in the presence of antisense RNA specific to the

ectopically synthesized alpha subunit of human chorionic gonadotropin. *J Cell Biol* **108**, 2423–2434 (1989).

176. Tsai, H.-J. *et al*. The Prognostic and Predictive Role of Chromogranin A in Gastroenteropancreatic Neuroendocrine Tumors – A Single-Center Experience. *Front. Oncol.* **11**, (2021).

177. Matsumoto, T., Asakura, H. & Hayashi, T. Increased salivary chromogranin A in women with severe negative mood states in the premenstrual phase. *J Psychosom Obstet Gynaecol* **33**, 120–128 (2012).

178. Hanif, H. *et al*. Update on the applications and limitations of alpha-fetoprotein for hepatocellular carcinoma. *World Journal of Gastroenterology* **28**, 216 (2022).

179. Conte, M. *et al*. GDF15, an emerging key player in human aging. *Ageing Research Reviews* **75**, 101569 (2022).

180. Pence, B. D. Growth Differentiation Factor-15 in Immunity and Aging. *Frontiers in Aging* **3**, 837575 (2022).

181. Wollert, K. C., Kempf, T. & Wallentin, L. Growth Differentiation Factor 15 as a Biomarker in Cardiovascular Disease. *Clinical Chemistry* **63**, 140–151 (2017).

182. Yatsuga, S. *et al*. Growth differentiation factor 15 as a useful biomarker for mitochondrial disorders. *Annals of Neurology* **78**, 814 (2015).

183. Adela, R. & Banerjee, S. K. GDF-15 as a Target and Biomarker for Diabetes and Cardiovascular Diseases: A Translational Prospective. *Journal of Diabetes Research* **2015**, 490842 (2015).

184. Fuchs, T. *et al*. Macrophage inhibitory cytokine-1 is associated with cognitive impairment and predicts cognitive decline – the Sydney Memory and Aging Study. *Aging Cell* **12**, 882–889 (2013).

185. Akiyama, T., Raftery, L. A. & Wharton, K. A. Bone morphogenetic protein signaling: the pathway and its regulation. *Genetics* **226**, iyad200 (2024).

186. Ponomarev, L. C., Ksiazkiewicz, J., Staring, M. W., Luttun, A. & Zwijsen, A. The BMP Pathway in Blood Vessel and Lymphatic Vessel Biology. *International Journal of Molecular Sciences* **22**, 6364 (2021).

187. Morrell, N. W. *et al*. Targeting BMP signalling in cardiovascular disease and anaemia. *Nat Rev Cardiol* **13**, 106–120 (2016).

188. Yousef, H. *et al*. Age-Associated Increase in BMP Signaling Inhibits Hippocampal Neurogenesis. *Stem Cells* **33**, 1577–1588 (2015).

189. Meyers, E. A. *et al*. Increased bone morphogenetic protein signaling contributes to age-related declines in neurogenesis and cognition. *Neurobiology of Aging* **38**, 164–175 (2016).

190. Ehata, S. & Miyazono, K. Bone Morphogenetic Protein Signaling in Cancer; Some Topics in the Recent 10 Years. *Frontiers in Cell and Developmental Biology* **10**, 883523 (2022).

191. Zylbersztejn, F. *et al*. The BMP pathway: A unique tool to decode the origin and progression of leukemia. *Experimental Hematology* **61**, 36–44 (2018).

192. Augeri, D. J., Langenfeld, E., Castle, M., Gilleran, J. A. & Langenfeld, J. Inhibition of BMP and of TGFβ receptors downregulates expression of XIAP and TAK1 leading to lung cancer cell death. *Molecular Cancer* **15**, 27 (2016).

193. Fan, Y. *et al*. BMP-9 is a novel marker for colorectal tumorigenesis undergoing the normal mucosa-adenoma-adenocarcinoma sequence and is associated with colorectal cancer prognosis. *Oncology Letters* **19**, 271 (2019).

194. Hollis, B. *et al*. Genomic analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. *Nat Commun* **11**, 1536 (2020).

195. Day, F. R., Elks, C. E., Murray, A., Ong, K. K. & Perry, J. R. B. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci Rep* **5**, 11208 (2015).

196. Jáni, M. *et al*. Birth outcomes, puberty onset, and obesity as long-term predictors of biological aging in young adulthood. *Front. Nutr.* **9**, (2023).

197. Ryczkowska, K., Adach, W., Janikowski, K., Banach, M. & Bielecka-Dabrowa, A. Menopause and women's cardiovascular health: is it really an obvious relationship? *Archives of Medical Science : AMS* **19**, 458 (2022).

198. Nerattini, M. *et al*. Systematic review and meta-analysis of the effects of menopause hormone therapy on risk of Alzheimer's disease and dementia. *Frontiers in Aging Neuroscience* **15**, 1260427 (2023).

199. de Villiers, T. J. Bone health and menopause: Osteoporosis prevention and treatment. *Best Practice & Research Clinical Endocrinology & Metabolism* **38**, 101782 (2024).

200. Levine, M. E. *et al*. Menopause accelerates biological aging. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 9327 (2016).

201. Kara, R. J. *et al*. Fetal Cells Traffic to Injured Maternal Myocardium and Undergo Cardiac Differentiation. *Circulation Research* **110**, 82–93 (2012).

202. Ryan, C. P. *et al*. Pregnancy is linked to faster epigenetic aging in young women. *Proceedings of the National Academy of Sciences* **121**, e2317290121 (2024).

203. Poganik, J. R. *et al*. Biological age is increased by stress and restored upon recovery. *Cell Metab* **35**, 807-820.e5 (2023).

204. Doll, R. & Hill, A. B. The Mortality of Doctors in Relation to Their Smoking Habits. *Br Med J* **1**, 1451–1455 (1954).

205. Doll, R. & Hill, A. B. Lung Cancer and Other Causes of Death in Relation to Smoking. *Br Med J* **2**, 1071–1081 (1956).

206. Reitsma, M. B. *et al*. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *The Lancet* **397**, 2337–2360 (2021).

207. Chan, K. H. *et al*. Tobacco smoking and risks of more than 470 diseases in China: a prospective cohort study. *The Lancet Public Health* **7**, e1014–e1026 (2022).

208. Persoskie, A. & Nelson, W. L. Just Blowing Smoke? Social Desirability and Reporting of Intentions to Quit Smoking. *Nicotine Tob Res* **15**, 2088–2093 (2013).

209. Ambrose, J. A. & Barua, R. S. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *J Am Coll Cardiol* **43**, 1731–1737 (2004).

210. Caraballo, R. S., Giovino, G. A., Pechacek, T. F. & Mowery, P. D. Factors associated with discrepancies between self-reports on cigarette smoking and measured serum cotinine levels among persons aged 17 years or older: Third National Health and Nutrition Examination Survey, 1988-1994. *Am J Epidemiol* **153**, 807–814 (2001).

211. Deveci, S. E., Deveci, F., Açik, Y. & Ozan, A. T. The measurement of exhaled carbon monoxide in healthy smokers and non-smokers. *Respiratory Medicine* **98**, 551–556 (2004).

212. Sugden, K. *et al*. Establishing a generalized polyepigenetic biomarker for tobacco smoking. *Transl Psychiatry* **9**, 92 (2019).

213. Joehanes, R. *et al*. Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics* **9**, 436–447 (2016).

214. Bojesen, S. E., Timpson, N., Relton, C., Smith, G. D. & Nordestgaard, B. G. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax* **72**, 646–653 (2017).

215. Corley, J. *et al*. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. *Transl Psychiatry* **9**, 1–15 (2019).

216. Yousefi, P. D. *et al*. DNA methylation-based predictors of health: applications and statistical considerations. *Nat Rev Genet* **23**, 369–383 (2022).

217. Huan, T. *et al*. A Whole-Blood Transcriptome Meta-Analysis Identifies Gene Expression Signatures of Cigarette Smoking. *Hum. Mol. Genet.* ddw288 (2016) doi:10.1093/hmg/ddw288.

218. Argentieri, M. A. *et al*. Integrating the environmental and genetic architectures of aging and mortality. *Nat Med* 1–10 (2025) doi:10.1038/s41591-024-03483-9.

219. Welsh, S. Genotyping of 500,000 UK Biobank participants. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/ukb_dna_processing.pdf (2017).

220. The UK Biobank resource with deep phenotyping and genomic data | Nature. https://www.nature.com/articles/s41586-018-0579-z.

221. Chen, Z. *et al*. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology* **40**, 1652–1666 (2011).

222. Lonsdale, J. *et al*. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

223. CTG-VL Complex Traits Genetics Virtual Lab. https://vl.genoma.io/.

224. Eldjarn, G. H. *et al*. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).

225. Zhou, C., Gao, Y., Ding, P., Wu, T. & Ji, G. The role of CXCL family members in different diseases. *Cell Death Discov.* **9**, 1–12 (2023).

226. Lee, S. *et al*. Genetic or therapeutic neutralization of ALK1 reduces LDL transcytosis and atherosclerosis in mice. *Nat Cardiovasc Res* **2**, 438–448 (2023).

227. Jain, K. *et al*. Pathogenic Variant Frequencies in Hereditary Haemorrhagic Telangiectasia Support Clinical Evidence of Protection from Myocardial Infarction. *J Clin Med* **13**, 250 (2023).

228. Liu, M. *et al*. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237–244 (2019).

229. Noyce, A. J. *et al*. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Annals of Neurology* **72**, 893–901 (2012).

230. Doll, R., Peto, R., Boreham, J. & Sutherland, I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* **328**, 1519 (2004).

231. Jha, P. & Peto, R. Global Effects of Smoking, of Quitting, and of Taxing Tobacco. *New England Journal of Medicine* **370**, 60–68 (2014).

232. Le Foll, B. *et al*. Tobacco and nicotine use. *Nat Rev Dis Primers* **8**, 1–16 (2022).

233. Saunders, G. R. B. *et al*. Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature* **612**, 720–724 (2022).

234. Chen, F. *et al*. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat Genet* **55**, 291–300 (2023).

235. Charlesworth, J. C. *et al*. Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med Genomics* **3**, 29 (2010).

236. Elliott, H. R. *et al*. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenet* **6**, 4 (2014).

237. Mappin-Kasirer, B. *et al*. Tobacco smoking and the risk of Parkinson disease. *Neurology* **94**, e2132–e2138 (2020).

238. Maggio, R. *et al*. Nicotine prevents experimental parkinsonism in rodents and induces striatal increase of neurotrophic factors. *J Neurochem* **71**, 2439–2446 (1998).

239. Castagnoli, K. P., Steyn, S. J., Petzer, J. P., Van der Schyf, C. J. & Castagnoli, N. Neuroprotection in the MPTP Parkinsonian C57BL/6 mouse model by a compound isolated from tobacco. *Chem Res Toxicol* **14**, 523–527 (2001).

240. Banderali, G. *et al*. Short and long term health effects of parental tobacco smoking during pregnancy and lactation: a descriptive review. *J Transl Med* **13**, 327 (2015).

241. Deng, W. Q. *et al*. Maternal smoking DNA methylation risk score associated with health outcomes in offspring of European and South Asian ancestry. *eLife* **13**, (2024).

242. Cohen, A. J. *et al*. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **389**, 1907–1918 (2017).

243. Mallah, M. A. *et al*. Cigarette smoking and air pollution exposure and their effects on cardiovascular diseases. *Front Public Health* **11**, 967047 (2023).

244. Mc, B. *et al*. Effects of aging on menstrual cycle hormones and endometrial maturation. *Fertility and sterility* **64**, (1995).

245. Snellings, D. *et al*. Somatic Mutations in Vascular Malformations of Hereditary Hemorrhagic Telangiectasia Result in Biallelic Loss of ENG or ACVRL1. *bioRxiv* 731588 (2019) doi:10.1101/731588.

246. Pham, H. *et al*. The effects of pregnancy, its progression, and its cessation on human (maternal) biological aging. *Cell Metabolism* **36**, 877–878 (2024).

247. Mayor, S. Review warns that risks of long term HRT outweigh benefits. *BMJ* **325**, 673 (2002).

248. White, M. C. *et al*. Age and Cancer Risk. *Am J Prev Med* **46**, S7-15 (2014).

249. Tyagi, S., Gupta, P., Saini, A. S., Kaushal, C. & Sharma, S. The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. *J Adv Pharm Technol Res* **2**, 236–240 (2011).

250. Monteiro, R., Teixeira, D. & Calhau, C. Estrogen Signaling in Metabolic Inflammation. *Mediators Inflamm* **2014**, 615917 (2014).

251. Palmisano, B. T., Zhu, L. & Stafford, J. M. Estrogens in the Regulation of Liver Lipid Metabolism. *Adv Exp Med Biol* **1043**, 227–256 (2017).

252. Association, A. L. 10 of the Worst Diseases Smoking Causes | State of Tobacco Control. https://www.lung.org/research/sotc/by-the-numbers/10-worst-diseases-smoking-causes.

253. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158–D169 (2017).

254. Wishart, D. S. *et al*. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **50**, D622–D631 (2022).

255. Tsai, M.-K., Gao, W. & Wen, C.-P. The relationship between alcohol consumption and health: J-shaped or less is more? *BMC Medicine* **21**, 228 (2023).