Epigenetic biomarkers of smoking, inflammation, and social differences

Alexandria Andrayas

A thesis submitted for the degree of Doctor of Philosophy in Biosocial Research

Institute of Social and Economic Research and School of Life Sciences

University of Essex

2022

Summary

This thesis aims to investigate the interplay between smoking, DNA methylation, inflammation, and socioeconomic position. First, 16 different methylation-based biomarkers of smoking are compared in their explanation of smoking status, pack years, and cessation. The predictor with the best class separation and that explained the most variation in self-reported smoking was proposed by McCartney et al (2018), however using methylation measured at a single locus in the AHRR gene worked almost as well. Secondly, factors including sex, age, cell type composition, education and socioeconomic classification were investigated to see if these influenced the agreement between self-reported and methylation-based smoking. This showed that more misclassifications occurred in self-reported ex-smokers compared to other smoking groups, and more affluent people compared to individuals not achieving any qualifications or working more routine occupations. Self-reported and DNAme-predicted smoking were also compared in terms of inflammation, and this suggested that DNAme-predicted smoking measures may more closely relate to inflammation than self-reports. Lastly, epigenetic signatures of inflammation were investigated. This showed that many factors influence DNA methylation changes that occur with inflammation, including educational attainment and suggests that the social environment could play a role in epigenetic signatures of inflammation. In EWAS models where self-reported smoking was used, the addition of educational attainment had more of an impact on findings compared to methylation-based measures of smoking. An overarching aim of this thesis is to communicate the importance of interdisciplinary approaches to health research to fully consider how the health of an occurs as part of a greater whole.

Acknowledgements

I would like to thank Professor Leonard Schalkwyk and Professor Meena Kumari for all their insight and encouragement through the duration of my PhD. I would also like to thank all those in the Soc-B Centre for Doctoral Training, and within the Understanding Society, ISER, Life Sciences and Genomics research groups. Lastly, thank you to Shanze, family and friends for their unending encouragement.

I declare that the work hereby submitted is my own original work.

Table of Contents

Table of Tables

Table of Figures

1. Introduction

The aims of this thesis are to investigate and compare epigenetic biomarkers of smoking, evaluate their accuracy and utility in estimating smoking and gain an understanding into the biological and social underpinnings of inflammation in relation to DNA methylation. Firstly, three novel DNA methylation-based predictors of smoking status, pack years, and cessation years were constructed and then compared to existing DNA methylation-based predictors of smoking. Within this chapter the relationship between methylation-derived smoking histories and age is also discussed. Secondly, different demographic and socioeconomic factors are explored to identify if any may influence the agreement between self-reported and DNA methylation-based measures of smoking. In this chapter self-reported and methylation-based smoking status, as well as a smoking methylation score, are also compared in their relationship with commonly measured markers of inflammation and in their impact on the socioeconomic gradient observed in inflammation using different adjustments for smoking. Lastly, the association of DNA methylation measured in blood with two circulatory inflammation markers, fibrinogen and C-reactive protein, is investigated. Within this chapter the impact of cell type composition, age, sex, body mass index, educational attainment, and both self-reported and methylation-derived smoking measures, on differentially methylated loci significantly associated with inflammation is examined. The work carried out throughout this thesis uses DNA methylation resources collected as part of two UK social panel studies, the 1958 National Child Development Study (NCDS) and the UK Household Longitudinal Study (UKHLS), alongside other biological and sociodemographic questionnaire data collected periodically by the studies.

1.1. DNA methylation

Deoxyribonucleic acid (DNA) is a molecule consisting of two polynucleotide chains that coil around each other to form a double helix. This carries the instructions for the development and functioning of all known organisms and many viruses. Each nucleotide is composed of one of four nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a sugar called deoxyribose, and a functional phosphate group. Genetics involves the study of heredity and inherited characteristics while epigenetics refers to the study of changes that affect gene architecture and expression without altering the genetic sequence itself. Epigenetics often involves the study of chromatin structure. Chromatin refers to a mixture of DNA and proteins that form chromosomes and DNA is packaged in nucleosomes, the basic repeating subunit of chromatin packaged inside the cell's nucleus. The three most widely studied epigenetic processes are DNA methylation, histone modifications and non-coding RNA species. Epigenetics modifications affect almost all nuclear processes, including gene transcription and silencing, DNA repair and replication and telomere function. DNA methylation (DNAme) is one example of an epigenetic mechanism where methyl groups, a carbon atom bonded to three hydrogen atoms (CH_3), are added to a DNA molecule. In mammals DNAme widely operates at CpG sites where a cytosine nucleotide is followed by a guanine nucleotide and typically cytosines on both DNA strands become methylated. However, DNAme also occurs in different sequence contexts such as at adenine nucleotides as well as cytosines followed by nucleotides other than guanine. Non-CpG methylation has also been observed in embryonic stem cells (Lister et al., 2009) and hematopoietic progenitor cells (Kulis et al., 2015). DNAme can change the activity of the DNA segment without changing the underlying sequence and as such can influence gene expression, but this is dependent on the genetic location and context in which DNAme occurs.

With the recent completion of a telomere-to-telomere human reference genome, T2T-CHM13, it is now known that even more CpG nucleotides exist in mammals than previously thought, at over 32 million sites along the genome (Gershman et al., 2022). Within the human genome most CpG sites are located inside of clusters, called CpG islands, which are generally unmethylated whereas CpG sites outside of this context

often remain methylated. Changes to chromatin structure can act to separate the genome into transcriptionally active and inactive regions. Roughly 15% of the genome's CpGs are found within CpG islands which in turn make up approximately 1-2% of the human genome. Roughly half of all CpG islands are found at transcription start sites (TSSs), often at ubiquitously expressed ‘housekeeping’ genes. Negative correlations between DNAme and gene expression is enriched in regions with marks of regulatory activity. Noted exceptions to this include imprinted genes and during X chromosome inactivation (XCI) (Moore et al., 2012). High concordance of X inactivation status is observed across tissues, with most TSSs subject to XCI and few escaping from XCI in all tested tissues (Cotton et al., 2015). DNAme in the body of highly transcribed genes is mostly positively correlated with gene expression. Unmethylated CpG islands often lie in intergenic regions and the stability of DNA methylation means that distal regulatory elements, DNA sequences that can regulate genes many kilobases from said gene, and transposable elements, DNA sequences that move from one location on the genome to another, can be controlled and this helps maintains the integrity of the genome (Dahlet et al., 2020). Distal regulatory regions can be enhancers (increasing expression) or silencers (decreasing expression). Regulatory regions often demarcated by DNase hypersensitivity sites, regions of chromatin that has lost its condensed structure exposing the DNA and making it accessible, are shown to be enriched for epigenetically variable loci (Wagner et al., 2014). This provides support that variability in DNAme underlies altered expression patterns and thus modulates disease. Transcriptionally silent genes do not necessarily carry an unmethylated region however, as DNAme does not have the flexibility to properly fine-tune gene expression alone (Weber et al., 2007). Transcription of genes is thus achieved by chromatin organisation at multiple levels. For example, transcription factors are proteins involved in the process of converting DNA into RNA and their binding to eukaryotic chromosomes is strongly restricted by complex chromatin structures. Local nucleosome structures must be reorganised for transcription factors to gain access to regulatory elements, and these changes can occur in time frames ranging from milliseconds to minutes or hours (Voss and Hager, 2014).

DNA methylation may impact gene expression in two ways, one being that methylation itself acts to physically obstruct transcriptional binding proteins and the other involving methyl-CpG-binding domain proteins (MBDs). MBDs bind to methylated DNA enabling the recruitment of other proteins such as histone deacetylases and other chromatin remodelling proteins which in turn forms heterochromatin, a compact and inactive form of chromatin. Methyl groups are added to DNA by a family of enzymes called DNA methyltransferases (DNMTs) that catalyse the transfer of methyl groups from S-adenosylmethionine. These enzymes are involved in both maintenance and de novo methylation. DNMT1 is necessary to maintain already established DNA methylation patterns and ensure DNA methylation patterns are copied to daughter strands during DNA replication and are not lost during passive demethylation. DNMT3a and DNMT3b are required for the establishment of de novo DNA methylation patterns alongside DNMT3L which has no catalytic activity but aids DNMT3s in binding to DNA. Deletion of any DNMT is lethal in murine and human cells showing the indispensable function methylation plays in mammals (Bestor et al., 2000). DNA methylation appears as the default state where signatures must be specifically removed through DNA demethylation which involves the removal of the methyl group (Lister et al., 2009). This process is required for epigenetic reprogramming of genes and has been implicated in multiple disease mechanisms such as tumour progression (Ehrlich, 2009). Demethylation has even been shown to occur in peripheral blood mononuclear cells after surgery at sites annotated to immune system genes (Sadahiro et al., 2020).

It was originally thought that DNA demethylation only occurs passively through dilution of methylation marks, but it is now widely known that methylation marks can be actively erased through a combination of passive dilution and the direct enzymatic removal of the methyl group (Ohno et al., 2013). In mammals, direct excision of 5'-methylcytosine (5mC) paired with G does not seem possible so instead the methylated base undergoes sequential modifications through enzyme-mediated oxidation by ten-eleven translocation (TET). This family of methylcytosine dioxygenases include TET1, TET2 and TET3. TET enzymes may promote DNA demethylation by binding to CpG rich regions preventing DNMT activity but for a TET

enzyme to initiate demethylation it must first be recruited to a methylated CpG site in DNA. They then work by producing 5-hydroxymethylcytosine (5-hmC) as the first intermediate and then further hydroxylating this intermediate to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). Thymine DNA glycosylase (TDG) can recognize these intermediate bases and excises the glycosidic bond resulting in an apyrimidinic site (AP site) which is followed by the base excision repair (BER) pathway to convert the modified cytosine back to its unmodified state (Bochtler et al., 2016). 5mC can also be directly converted to thymine and followed by the BER pathway. The biological significance of 5mC has been widely recognized and may reflect a global decrease of DNA methylation where quantification of global 5-mC could act as a molecular marker for disease. This could be a consequence of many factors such as methyl-deficiency caused by several different environmental influences (Robertson, 2005). More recent studies now show small changes to intermediate DNA methylation may be associated with complex disease phenotypes (Leenen et al., 2016).

1.2.Studying variation in DNA methylation

The study of differences in DNA methylation associated with health-related phenotypes have generally involved one of three techniques. These include the study of global DNA methylation, DNA methylation at specific candidate genes and genome-wide approaches like those used in epigenome-wide association studies (EWAS). Global DNA methylation specifically refers to the level of 5-methylcytosine (5mC) content in a sample relative to the amount of total cytosine. Proxy measures of global DNAme are often used assuming that they accurately reflect 5mC content and include generating an average methylation value from either the average measure in highly repetitive genomic elements or unique CpG sites throughout the genome (Vryer, and Saffery, 2017). Although global DNA methylation has been indicated in many health-related phenotypes and disease it is unable to reflect all changes to epigenetic modifications that occur at a gene-specific level. Candidate genes represent specific and biologically relevant regions of

the genome whose chromosomal location is associated with a particular phenotype and is thus investigated based on a priori knowledge (Kwon and Goate, 2000). Findings from candidate gene approaches often produce high rates of false positives, have not been easily replicated, and are impacted by issues of power and population stratification (Tabor et al., 2002).

Epigenome-wide association studies involve the examination of genome-wide or genome-scale sets of quantifiable epigenetic marks in different individuals to derive associations between epigenetic variation and an identifiable phenotype and offers a 'hypothesis-free' approach (Rakyan et al., 2011). The Illumina microarrays used in EWAS are however not representative of the entire human epigenome with the EPIC array only covering up to 3% of the over 27-32 million CpG sites found in the human genome. Other methods such as whole genome bisulfite sequencing (WGBS) or single molecule real time (SMRT) sequencing technology may offer more efficient methods for determining the methylation status of the genome. Despite this however microarrays offer cost-effective and consistent analysis of many biologically relevant genomic regions. Attempts are made to cover all known genes, especially CpG sites and islands in gene promotors, and coverage is enriched for genes with health-related functions (Bibikova, 2016). Even though the epigenome consists of a multitude of chemical compounds that all act to shape chromatin structure and regulate the genome, EWAS most commonly investigate DNA methylation due to its chemical stability and the fact it is not lost during the DNA extraction process. Variations in DNA methylation can cause disease but can also arise because of disease and thus EWAS are unable to determine the direction of causation. However, it has recently been suggested from a study using transcriptomic data that when comparing diseased and healthy participants findings are more likely to reveal gene expression changes induced by the disease rather than causing the disease (Porcu et al., 2021). Preferably longitudinal studies should be used where DNAme is measured before and after any symptoms of a disease become present. Another issue of EWAS is that DNA methylation is often measured in blood which may not reflect epigenetic variation in tissues specific to a phenotype of interest and blood constitutes multiple cell types

each of which may show distinct DNA methylation profiles making it difficult to decipher if differences are due to the phenotype of interest or sample heterogeneity. Validation is therefore necessary if using blood as a proxy for exposure in other tissues. To aid this issue methylation-derived estimates of cell type proportions are often included as covariates in EWAS (Houseman et al., 2012).

It has been more than 10 years since the conception of EWAS and thousands of publications using this method have since been circulated and this number continues to increase every year. The reduction in costs and innovation of scientific methods to measure epigenetic modifications have made this possible. Illumina microarrays have been the most widely used thus far with the 27K array first identifying smoker-specific hypermethylation at CpG site cg03636183 (*F2RL3*), the 450K array implicating CpG loci (*gene*) cg05575921 (*AHRR*), cg03636183 (*F2RL3*), and cg19859270 (*GPR15*) to smoking and a variety of inflammation-induced diseases and the EPIC array introducing many more novel loci, particularly in regulatory regions, associated within human development and disease. EWAS are also now being used to estimate disease risk by identifying specific DNA methylation loci as biomarkers (McCartney et al., 2018). Polyepigenetic biomarkers may become valuable predictors of susceptibility to human disease such as one study that used a methylation risk score based on levels of methylation change within 187 CpG loci associated with obesity to predict the risk of developing type 2 diabetes in the future (Wahl et al., 2017). Another study was able to detect differentially methylated regions associated with autism spectrum disorders in EWAS carried out using cord blood within new-borns who were later diagnosed (Mordaunt et al., 2020). Another identified a risk score using just three CpG sites with high sensitivity for early detection of colorectal cancer (Heiss and Brenner, 2017). This all suggests a potential for early diagnoses of disease than can improve or prevent disease progression. Epigenetic drugs also offer a novel therapeutic tool where one way to fight cancer is to inhibit methylation by using drugs that impact DNA methylation patterns by targeting histone and DNA methyltransferases. EWAS analysing differential DNA methylation associated with childhood asthma found several loci that were confirmed as drug targets (Reese et al., 2019). Drug-

induced epigenetic changes are a novel way to measure drug response and evaluate prognostic ability where another study looked at the association between drug responses to 526 pharmaceutical agents and DNA methylation in small cell lung cancer (Krushkal et al., 2020). In the study of complex diseases, EWAS-related databases are also now available and provide researchers with a powerful tool to enables searches of specific DNA methylation markers, KEGG pathways, and GO categories and allow easier collation of metadata and good evidence synthesis (Xiong et al., 2020). Many EWAS-related tools now exist to automate the identification of differentially methylated regions or loci, investigate epigenetic variation with diseases and phenotypes, comprehensively process, normalise and examine DNA methylation data, predict histone modifications and DNA methylation levels, as well as complex traits and differential cell types (Wei et al., 2021). This enormous body of research has been made possible by using EWAS and other related methods yet there is still more to be discovered.

1.3. Smoking and DNA methylation

In studying the impact of environmental exposures on DNA methylation the causative role of cigarette smoke in driving epigenetic modifications across the genome has become well established. The first hypothesis-free search for genome-wide significant loci implicated in smoking identified changes in DNA methylation at the *F2RL3* gene (Breitling et al, 2011) and this has been closely followed by the identification of thousands of CpG sites displaying differential methylation between smokers and non-smokers. Smoking associated CpG sites have been found to span all 23 chromosome pairs of the human genome with varying effect sizes. These associations were mostly identified through epigenome-wide association studies (EWAS), and epidemiological studies have worked to further strengthen these findings by identifying plausible underlying biological mechanisms with smoking-related disease. The *F2RL3* gene for example codes for the coagulation factor II (thrombin) receptor-like 3 protein which is vital in

haemostasis and thrombosis through its role in platelet activation and is expressed in several tissues including leukocytes and lung tissue. Changes in DNAme at this gene have been implicated in heart disease (Breitling et al., 2012) and lung cancer and mortality (Zhang et al., 2015).

To date the strongest differences in DNAme between smokers and non-smokers occur at sites located in *AHRR* and the 2q37.1 genetic regions. The *AHRR* gene codes for the aryl hydrocarbon receptor repressor which is a member of the aryl hydrocarbon receptor (AhR) signalling cascade alongside the aryl hydrocarbon receptor nuclear translocator (AhRNT). These proteins belong to the bHLH-PAS protein superfamily consisting of signalling molecules known to participate in the regulation of their own expression via transcription of specific repressor molecules that terminate signal transduction (Schmidt and Bradfield, 1996). A proposed mechanism of this pathway suggests AhRR prevents signal transduction and subsequent transcription activity of AhR by various AhR ligands. In the inactive form AhR remains in the cytoplasm as a multiprotein complex and upon ligand binding this complex dissociates and AhR then translocates into the nucleus and then dimerizes with AhRNT. This heterodimer binds to xenobiotic responsive elements (XREs) located in enhancer regions of target genes and regulates their transcription (Haarmann-Stemmann et al., 2007). The best characterized AhR target gene is *CYP1A1* which encodes a member of the cytochrome P450 superfamily of enzymes involved in the synthesis of cholesterol, steroids and other lipids. AhR ligands include dioxins and polycyclic aromatic hydrocarbons (PAHs) found in substances many individuals are commonly exposed to such as tobacco combustion, secondary plant metabolites, pharmaceuticals, and the by-products of industrialization (Evans et al., 2008). PAHs also induce the expression of *CYP1A1,* and this gene is itself able to metabolize some PAHs to carcinogenic intermediates (Shimada, and Fujii-Kuriyama, 2004). AhRR may also recruit co-repressor molecules and histone deacetylases to XRE gene promoters whereby XRE-bound AhRR recruits the transcriptional co-repressor molecule Ankyrin repeat family A protein 2 (Ankra2), and histone deacetylases 4 and 5, to the Cyp1a1 promoter (Oshima et al., 2004). This causes subsequent condensation of the local chromatin

structure hindering further binding of transcription factors and thus transcription of AhR target genes (Haberland et al., 2009). This suggests an important role of the AhR pathway in environmentally induced toxicity and adaptive xenobiotic metabolism. Expression of AhRR is high in testis, lung, ovary, spleen and pancreas in adults and low in all tissues in foetuses (Yamamoto et al., 2004). It has been found that DNA methylation at *AHRR* in monocytes is correlated with *AHRR* mRNA profiles and with carotid plaque scores. This remained significant even after controlling for self-reported smoking, urinary cotinine, and CVD risk factors, and was replicated in an independent sample (Reynolds et al., 2015). *AHRR* hypomethylation has also been found to be associated with the smoking-related CHRN3A genotype, COPD exacerbations, lung cancer and all-cause mortality.The association of the CHRN3A genotype, used to evaluate smoking heaviness, is of interest given distributions among ever and never smokers were similar suggesting selection could not preclude these findings (Bojesen et al., 2017).

Smoking has also been linked to a small global decrease in DNA methylation (Ambatipudi et al., 2016) and many gene-specific sites are hypomethylated in smokers. Notable exceptions to this are observed within the gene body of *MYO1G* gene. This gene codes for plasma membrane-associated class I myosin which is abundant in T and B lymphocytes and aids in cell elasticity and could relate to smoking-related fibrosis that occurs in several tissues (Olety et al., 2010). Other smoking related CpG sites were not located in transcriptional regions of known genes and are instead located in other regions within the gene or unannotated regions. DNA methylation at these loci is often less closely linked to transcriptional silencing than elements further upstream where DNA methylation often impacts the magnitude of gene expression (Brenet et al., 2011). Two examples of such sites known to be associated with smoking are several in the intergenic q37.1 region on chromosome 2 and loci in the 1st exon of G-protein coupled receptor 15 (*GPR15*). The loci on chromosome 2q37.1 are adjacent to an alkaline phosphatase gene cluster. One alkaline phosphatase gene called *ALPPL2* is responsible for dephosphorylation of many proteins and nucleotides and may offer benefits as a biomarker for many cancers as it is a well-established tumour marker

in ovarian and testicular cancers and seminoma (Albrecht et al., 2004). ALPPL2 enzyme serum concentrations can increase up to tenfold in smokers hinting at DNA methylation changes as a mechanism that increases smokers' risk to cancer (Schmoll et al., 2004). DNA methylation at CpG sites in the *GPR15* gene correlate with current and long-term smoking (Wan et al., 2012) and is one of few genes implicated in epigenetic modifications driven by smoking that showed a negative correlation between gene expression and DNA methylation as well as an increase in gene expression in smokers compared to non-smokers (Tsaprouni et al., 2014). *GPR15* codes for a class A orphan G protein-coupled receptor that is expressed in epithelial and endothelial cells, synovial macrophages and lymphocytes, but mainly T-cells, and regulates T-cell migration and immunity. Other strong smoking signals, existing outside of gene bodies, are observed in the 5' untranslated regions (5'UTR) of the *PRSS23* and *RARA* genes. 5'UTRs are cis- regulatory elements required to regulate translation. *PRSS23* codes for serine protease 23 and is a member of the trypsin family. *RARA* codes for retinoic acid receptor alpha which is a nuclear receptor and transcription factor that works alongside the retinoid X receptor (*RXR*) forming RXR/RAR heterodimers. In the absence of ligand this represses transcription by recruiting co-repressors or enabling the recruitment of histone acetyltransferase and co-activators to encourage gene expression if ligands bind. The genes and genetic regions mentioned represent just a small fraction of the genome where DNA methylation has been shown to vary between smokers and non-smokers.

Tobacco smoke exposure dramatically alters DNA methylation in blood cells, but it is also important to understand smoking effects on DNAme in specific leukocyte subtypes. Genome-wide approaches have found that CpGs have distinct methylation patterns in various tissues and in smoking-associated methylation and gene expression. This shows distinctive cell-type responses to tobacco smoke exposure that may not be apparent in whole blood DNA. Hematopoietic lineage-specific changes may then play a role in disease etiology and explain how DNA methylation in blood cells may mediate complex diseases associated with smoking (Su et al., 2016). This is supported by cell-type deconvolution algorithms which

have shown highly reproducible smoking-associated hypomethylation signatures appear more prominent in the myeloid lineage (You et al., 2020).

Above shows the multitude of biological and pathophysiological processes implicated in smoking and the utility and importance of DNA methylation in smoking-related disease. Deciphering epigenetic signatures of smoking have enabled a reliable and empirical approach to differentiating smokers from non-smokers via the construction of epigenetic biomarkers. A biomarker can involve any naturally occurring molecule, gene, or characteristic by which a particular physiological process or disease can be identified. Biomarkers are often used in clinical settings with the goal of either aiding physicians in the diagnosis of a given disease or informing treatment decisions by providing patient and disease characteristics. Epigenetic biomarkers can offer an objective way to measure health-related characteristics such as smoking and essentially involve two measurements, one being an assay of quantifiable epigenetic modifications for thousands of genomic locations per participant, and the other being a classification of each sample used to translate the experimental read-out into the biomarker outcome (Bock, 2009). Although many mechanisms are involved in epigenetic regulation, from histone acetylation to micro RNAs, epigenetic-based predictions have thus far largely involved DNA methylation due to its stability, ease to measure and established role in health and disease.

Many biomarkers previously used in detecting smoking are limited in their specificity and long-term stability. Cotinine for example is a predominant metabolite of nicotine that has been frequently used to objectively measure smoking, however this has a half-life of less than a day and is therefore unable to fully reflect past exposures (Zhang et al., 2016). Epigenetic biomarkers could then provide an even better objective measure of smoking given that certain smoking related changes to DNA methylation can persist years after cessation (Wan et al., 2012). The known associations between DNA methylation and smoking

can be leveraged to construct accurate and replicable methylation-based predictors of smoking status. This has been demonstrated through the construction of a DNA methylation index using bisulphite pyrosequencing of four genomic loci, located in the *AHRR*, 6p21, and 2q37 regions, that were differentially methylated between smokers and non-smokers. This index provided a strong and positive prediction for previous smoking with an area under the curve (AUC) of 0.83 (Shenker et al., 2013). It has also been demonstrated that DNA methylation can distinguish between active and nascent smokers in individuals with as little as half of a pack-year of tobacco use (Philibert et al., 2016). Variation in DNAme within the *AHRR* locus alone can reliably detect smoking status and intensity in both blood and saliva (Philibert et al., 2020). Further, DNA methylation-based biomarkers of smoking may be especially useful in populations with less precise self-reporting. DNA methylation measurements at the cg05575921 CpG site within the *AHRR* gene has been demonstrated to detect smoking status in populations with varied rates of false-negative self-reports. DNAme alone detected smoking behaviour almost as well as when weighting self-report and AHRR methylation per the different demographic characteristics of the participants. Also, while the reliability of self-reports impacted the accuracy of this biomarker, variation in DNA methylation between populations at the *AHRR* locus did not (Andersen et al., 2017). Few systematic comparisons of epigenetic biomarkers of smoking have been carried out and most existing biomarkers of smoking have only measured DNA methylation at a percentage of genetic loci now available to be analysed in relation to smoking. This thesis aims to critically evaluate multiple DNAme-based biomarkers of smoking.

1.4. Social differences in health

Poor health in disadvantaged groups is a ubiquitous finding (Marmot, 2015). While attempting to explain socioeconomic inequalities in health a large body of research has focused on the relationship between socio-economic position (SEP) and education with behavioural factors such as smoking, diet and heavy drinking.

SEP is known to be associated with various factors such as health behaviours, diet, access to health care, exposure to infection, and stress throughout life and these in turn influence an individual's systemic inflammatory burden (Pollit et al., 2008). This thinking is also in line with allostatic load, a cumulative exposure of the wear and tear bodies experience in response to acute and chronic stressors throughout the life course, that can consider the combined associations between perceived stress and physiologic responses (McEwen, 1998). A recent systematic review of systematic reviews showed there is good evidence to suggest social capital predicts better mental and physical health, and indicators of social capital are protective against mortality (Ehsan et al., 2019). With this said, the pathways in which socioeconomic factors impact health are still not fully understood despite many known associations of SEP with many biomarkers of health and disease. Coronary heart disease (CHD) is one condition that has been studied extensively and where both genetic endowment and socioeconomic factors have been shown to play key roles. CHD risk varies greatly by ethnicity whereby the death rate from CHD in African Americans is 37% higher than for white participants (Cruz-Flores et al., 2011). With this said the usual risk factors of obesity, hypertension and diabetes have failed to fully explain all CHD inequalities and some evidence has pointed to the role of chronic stress in CHD development through inflammatory processes (Kornerup et al., 2010). Psychobiological determinants like this offer a more complete explanation and explain more variance in disease progression and subsequent health outcomes by acknowledging the individual as part of a dynamic environment consisting of many social, psychological and cultural influences. It has been demonstrated that social stressors influence health through multiple biological pathways and that early life adversity can prime individuals to produce greater proinflammatory responses to stressors in later life (Packard et al., 2011). Further, these experiences lead to changes in DNA methylation levels and a proinflammatory epigenetic signature that impacts stress reactivity and cytokine production (Roth and Sweatt, 2011). The susceptibility of DNA methylation to changes in response to stressors thus makes it an excellent tool to better understand how the environment may impact physiological function and may inform social policy to identify the most influential sociodemographic factors driving health disparities. Levels of biomarkers like C-reactive protein

and fibrinogen in systemic inflammation for example are known to be associated with various measures of social position and have been implicated in socio-economic inequalities in inflammation that decrease with age and were not fully explained by smoking status, BMI or diet (Jousilahti et al., 2003 and Davillas et al., 2017).

There is good evidence of the causal role epigenetics in the establishment and progression of disease and cancer, however the social context in which health outcomes are entwined also seems to be important. Given the malleable nature of DNA methylation to environmental exposures, epigenetic biomarkers may aid in defining disparities in health. DNA methylation is one biological mechanism that may provide a link between social environments and health. Recently researchers have investigated the dynamics of stress reactivity and inflammation using life course measures of SEP and repeat measures of DNA methylation (Needham et al., 2015). Both childhood and adult SEP, as well as measures of social mobility, were found to be associated with DNA methylation at several pro-inflammatory genes, although studies have differed in the specific genes found to be differentially methylated (Stringhini et al., 2015). Nevertheless, both studies demonstrate that measurable changes to epigenetic modifications can reflect important, health-related aspects of our social environment. It also suggests that the response seen in the epigenome of individuals with different SEP classifications may relate to genes regulating inflammation. DNA methylation of one gene, nuclear factor of activated T-cells, cytoplasmic 1 (*NTFATC1*), was consistently lower in those with disadvantaged SEP and this occurred in a dose-dependent manner. *NTFATC1* plays a role in inducible gene transcription during immune response, specifically in the activation, differentiation and programmed death of T lymphocytes (Northrop et al., 1994). Interestingly expression patterns of the same gene have also been shown to relate to social rank in macaques (Tung et al., 2012).

Many researchers have now investigated epigenetic signatures related to important social factors such as SEP and educational attainment however often these traits are related to much smaller effect sizes in DNA methylation measurements than other environmental factors such as health behaviours. Unfortunately, these studies often do not provide strong evidence for socioeconomic drivers in modifying DNA methylation. In fact, one meta-analysis revealed 9 CpG sites that were differentially methylated with educational attainment (EA) however only two associations remained after sensitivity analysis and these only explained 0.3–0.7% of variance in EA (Linnér et al., 2017). All sites were also previously associated with smoking, and with much larger effect sizes. An over 20% difference in DNA methylation between smokers and non-smokers was observed for some probes like those found in *AHRR* (Zeilinger et al., 2013). The combined effect of these nine, education-related probes was also highly correlated with the effects of smoking.

Studies carried out over many years have uncovered a clear class gradient in smoking prevalence, with disadvantaged socio-economic position associated with high prevalence of tobacco use (Graham and Hunt, 1994) and higher resistance to changing said behaviours compared to individuals of higher SEP (Syme, 1992). In terms of the prevalence of smoking behaviours, educational inequalities appeared to be larger than income-based discrepancies (Escobedo and Peddicord, 1996). Despite this known social gradient, health behaviours are still to this day largely seen to involve free choice outside of their social context placing the responsibility for health on the individual (Knowles, 1977). This simplification may in part explain the 1 billion people today who still smoke (Doll et al., 2004) despite hundreds of anti-smoking campaigns and efforts since Doll and Hill (1950) and later the US Surgeon General (NIH, 1964) first elucidated the dangers of tobacco use on health.

Frequently only the impact of poverty-level SEP on health behaviours receives focus despite evidence of a graded association at all points on the socioeconomic scale and this should be considered when

investigating pathways between SEP and health endpoints, especially if those pathways may be mediated through health behaviours like smoking. This is especially important given that health behaviours are often subject to measurement error which varies across SEP with the most affluent being more likely to give false-negative results (Patrick et al., 1994). Additionally, the role of SEP is often relegated to a control variable to provide more strength to other aetiologic factors related to health. However, this may lead to incorrect estimates between biological processes such as DNA methylation and smoking where health behaviours and SEP are closely related (Adler et al., 1994). This is supported by the fact many associations between SEP and health fall short when health behaviours are included in regression analysis and the difficulty in quantifying SEP effects independent of smoking, alcohol, diet and physical activity. With this said the diminished effect of SEP observed after adjustment for health behaviours may be due to imprecise measures and it is therefore important to investigate the underlying measurement error of health behaviours to disentangle the role of one etiologic factor independent of another (Lynch et al., 1997). The extent to which the effect of SEP on health is mediated by health behaviours may be better understood by using epigenetic markers of tobacco use. Few studies have aimed to compare socioeconomic gradients in smoking between self-reports and methylation derived estimates and comparisons made between self-reports and other biomarkers like cotinine occurred many decades ago. Few studies have also aimed to better understand factors that influence discrepancies between epigenetic and self-reported health behaviours and how these measures compare in their explanation of markers of health such as in inflammation. This thesis aims to investigate if a socioeconomic gradient exists in the congruence between self-reports and DNAme predictions of smoking. It also aims to investigate how different measures of smoking compare in their association with inflammatory markers.

1.5. Inflammation and DNA methylation

Inflammation describes the protective biological response of tissues against infection, injuries and toxins. Epigenetic mechanisms have been suggested to influence the genetic regulation of pathways related to inflammation. There is some suggestion that decreases to global DNA methylation may occur with increasing levels of inflammation and many EWAS find the majority of differentially methylated genes are hypomethylated in inflammatory processes (Gonzalez-Jaramillo et al., 2019). Immune cells involve a complex network of different cell types and interactions that require differentiation to determine cell phenotype and function. The latter is highly dependent on epigenetic profiles that in turn establish transcriptional programs and bridge the gap between the environment and genome regulation. Recent advances in genome-wide DNA methylation data have provided insights into the roles of DNA methylation in mitigating environmental cues in health and disease (Calle-Fabregat et al., 2020).

Most EWAS investigating smoking-induced changes to DNA methylation have used microarray technology based on whole blood samples. These cells are derived from haematopoietic stem cells (HSC) whereby HSCs differentiate to form all blood cell lineages during haematopoiesis (Birbrai and Frenette, 2016). Such cells are turned over at a high rate and it is of interest how epigenetic alterations in blood persist over time such as those caused by smoking which may endure long after a person has ceased tobacco use. However recently associations between DNA methylation in a panel of CpG sites related to cigarette smoking and lung function levels that were originally identified in whole blood were replicated in lung tissue (de Vries et al., 2018). Blood borne smoke components that also affect HSC stem cells, and this would explain the presence of the AhRR signal across multiple tissues. HSCs expresses AhR and when activated AhR drives expansion of HSCs and directs cell fate, with chronic AhR agonism denoting erythroid differentiation and acute antagonism favouring megakaryocyte specification (Smith et al., 2013). Active smokers often suffer chronic leucocytosis whereby neutrophils are released in response to inflammatory signals from the lung (Van Eeden and Hogg, 2000). It is also known that cigarette smoking is a reversible

cause of elevated white blood cell count in both cross-sectional and longitudinal studies (Higuchi et al., 2016).

Inflammatory markers show a dose-dependent and temporal relationship to smoking and smoking cessation (Wannamethee et al., 2005). Two commonly measured markers of inflammation available in large cohort and panel studies are fibrinogen and C-reactive protein (CRP). Fibrinogen is a blood plasma protein and biomarker of systemic inflammation made in the liver. It is a positive acute phase protein and coagulation factor that traps invading microbes in blood clots and is enzymatically converted to fibrin and then to a fibrin-based blood clot during vascular injury. Fibrin also mediates blood platelet and endothelial cell spreading (Mosesson, 2005). Plasma concentration of fibrinogen positively correlates with inflammation and erythrocyte sedimentation rate, a test to see how erythrocytes settle at the bottom of a test tube, where faster-than-normal rate is indicative of inflammation. Fibrinogen has a half-life of approximately one week and levels vary between health and disease but may remain high despite removal of the inflammatory stimuli. CRP is another acute-phase protein synthesized by the liver where plasma concentration increases following secretion of factors such as interleukin 6 (IL-6) released by macrophages, T cells and adipocytes and is observed in many acute and chronic inflammatory conditions (Lau et al., 2005). CRP binds the surface of dead or dying cells, and some microbes, to activate the complement system via C1q (Thompson et al., 1999). CRP acts as a pattern recognition receptor (PRR) that functions in the innate immune system. It identifies microbial pathogens and components of cells that are released during cell damage and mediates the initiation of antigen-specific adaptive immune response and release of inflammatory cytokines (Kumar et al., 2011). IL-6 is both a pro- and anti- inflammatory myokine. Cytokines are small proteins secreted by the cells of the immune system, such as T cells, that are important in cell signalling (Zhang and An, 2007). Interleukins are a subset of cytokines made by one leukocyte that act on other leukocytes. A myokine is a specific kind of cytokine secreted by skeletal muscle cells. Cytokine activity is pleiotropic and redundant

meaning many different cell types can produce the same cytokine and a single cytokine may act on many different cell types. The most common producers of cytokines include T helper (Th) cells and macrophages.

Many studies have looked at the association of fibrinogen, CRP and IL-6 with epigenetic modifications. Gene-specific approaches have found higher levels of CRP to be associated with higher degree of methylation of *LY86* (Su et al., 2014), and *EEF2* (Arpon et al., 2016) and lower degree of methylation of *AIM2* (Miller et al., 2018) and the *IL-6* promoter gene (Wei et al., 2016). As for IL-6, higher concentrations are found to be associated with a higher degree of methylation of *MGMT*, *RARβ*, *RASSF1A*, and *CDH13* in tumour specimens and of *SOCS1* in peripheral blood. Increased IL-6 levels were also associated with reduced DNA methylation of *USP2*, *TMEM49*, *SMAD3* and *DTNB* (Piperi et al., 2010 and Jhun et al., 2017). Methylation at *LY86* was also associated with fibrinogen (Su et al., 2014). Epigenome-wide analysis have found higher levels of CRP associated with genes enriched in pathways related to atherosclerosis and IL-6. Among the reported genes that were differentially methylated with higher CRP levels, methylation at *SOCS3* and *BCL3* is significantly reduced and the *SOCS3* association remained significant after replication (Ligthart et al. and Marzi et al., 2016). *SOCS3* has been previously reported to play an important role in atherosclerosis and codes for the suppressor of cytokine signalling 3 gene and plays a pivotal role in the innate immune system as a regulator of cytokine signalling along the JAK/STAT pathway (Rottenberg and Carow, 2014). *AIM2* is a key regulator of human innate immune response and is implicated in defence mechanisms against bacterial and viral pathogens (Hornung et al., 2009). Many CpG sites significantly associated with inflammation have also been previously linked with smoking and to future incidence of heart disease (Ligthart et al. 2016). Many replicated CpG sites were also associated with different cardiometabolic phenotypes such as body mass index, fasting glucose and insulin, triglycerides, total cholesterol and HDL-cholesterol, highlighting the pleiotropic network of epigenetics across various related phenotypes. For example, it is known that cytokines are regulators of adipose tissue metabolism (Coppack,

2001) and adiposity influences lipid and glucose homeostasis and simultaneously promotes many different diseases (Chait and den Hartigh, 2020).

The EWAS literature studying the association of DNA methylation and inflammatory markers vary greatly in terms of included covariates. Given that smoking, BMI, cell type composition and inflammatory markers are intricately linked, it is important to understand how adjustment of these factors impact differentially methylated loci identified in relation to inflammation. This project aims to do this in terms of fibrinogen and CRP and investigate if adjustment for smoking using methylation-based biomarkers and adjustment for educational attainment alter findings.

It is clear DNA methylation plays an important role in health and disease (Tost, 2010). In cancer global DNA methylation is linked to chromatin maintenance and chromosomal stability and implicated through different mechanisms. Typically altered methylation at CpG islands within promoters in protein coding genes and microRNAs, small single-stranded non-coding RNA molecules, lead to altered gene expression. Often tumour suppressor genes and DNA repair genes become methylated and oncogenes demethylated (Craig and Wong, 2011). Epigenetic modifications are also implicated in cardiovascular disease and atherosclerosis. Vascular tissue and mononuclear blood cells such as monocytes and lymphocytes in individuals with these diseases exhibit decreases in global methylation with gene-specific areas of methylation. One potential mechanism explaining the overall decrease in global DNA methylation may be linked to increased levels of homocysteine in the blood which is a known risk factor for cardiovascular disease. High levels of homocysteine can inhibit the appropriate function of DNMTs leading to changes in DNA methylation. These changes have been shown to occur at genes related to smooth muscle cell proliferation which in turn lead to dysregulation of epithelial cell function and increased inflammation and atherosclerotic lesions (Castro et al., 2003). High levels of homocysteine are also shown to down regulate

the oestrogen receptor alpha (*ERα*) gene by increasing DNAme at CpG islands within the gene promotor. The *Erα* gene may protect against atherosclerosis by acting as a growth suppressor meaning smooth muscle cells remain in a quiescent state where cells do not proliferate (Huang et al., 2009). DNA methylation at the monocarboxylate transporter (*MCT3*) gene has also been implicated in atherosclerosis. This gene codes for a protein responsible for the transport of lactate and other ketones out of cells. In atherosclerosis, expression of this gene product is downregulated leading to further increases in smooth muscle cell proliferation (Zhu et al., 2005). Epigenetic changes have also been identified in heart failure and may vary depending on the aetiology of heart disease in question. For example, ischemic heart failure is the clinical endpoint of coronary heart disease and genome-wide analysis of DNA methylation marks has shown transcriptional reprogramming leading to suppression of oxidative metabolism (Pepin et al., 2019). In heart failure, changes in cardiac DNA methylation correspond with racial differences in all-cause mortality leading to changes in the activity of metabolic signalling pathways (Pepin et al., 2021). A global loss of DNA methylation also occurs during aging (Gonzalo, 2010) and this loss in methylation is proportional to age and occurs across the whole genome at promoters, intergenic, intronic and exonic regions (Heyn at al., 2012). Increases in DNA methylation with age is shown to occur in some genes including those coding the oestrogen receptor, p16, and insulin-like growth factor 2 (Gonzalo, 2010). DNAme levels have thus been used to accurately estimate age in many human cell types and tissues (Horvath, 2013). Longitudinal analyses of epigenetic variation have shown divergence in DNA methylation patterns between twins from age 5 onwards that were due to environmental influences (Wong et al., 2010). These findings together show the diverse and multifaceted role of DNA methylation in many biological processes related to health and disease. DNA methylation resources are then a valuable tool within epidemiology and can allow for non-deterministic insights into the genetic programming of disease and enable better understanding in how our lifestyles and health behaviours influence our health.

In this thesis first epigenetic biomarkers of smoking are developed, compared to existing methods, and used to examine how best to estimate smoking status, pack years and cessation years from DNA methylation. Secondly, two biomarkers of smoking, a classifier of smoking status called 'smokp SSt' and a methylation score of smoking from McCartney et al (2018), are utilised to examine predictors of agreement between smoking assessed using self-reported and methylation-based measures. The predictors in question include age, sex, smoking status, methylation-derived cell type composition, educational attainment and socioeconomic classification. Self-reported and methylation-based smoking measures were also compared in their relationship with two inflammatory markers, fibrinogen and C-reactive protein, and how adjustment for smoking using these two different methods also impacts the known association between socioeconomic position and inflammation. Lastly, multiple EWAS of two inflammatory markers were carried out with a view to understand how adjustment for smoking using self-reports or methylation-based estimates, and adjustment for education, influences the epigenetic signatures of inflammation.

2. **Methods**

This chapter introduces the datasets and provides an overview of analytic methods used in this thesis. More details are given in each chapter.

2.1. Studies

2.1.1. Understanding Society (UKHLS)

Understanding Society is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales and Northern Ireland that started in 2009 and collects information about people's health, behaviours, attitudes and social and economic circumstances (Lynn, 2009). This has been funded primarily by the Economic and Social Research Council (ESRC) and builds on the success of the British Household Panel Survey (BHPS) that was heavily used by researchers who have generated hundreds of scientific publications since BHPS started in 1991. Understanding Society aims to support a wider range of research than BHPS and DNA methylation resources aim to help this by opening the study to more biological or health-related researchers (Buck and McFall, 2011). Longitudinal studies of this nature can provide understanding of the trajectories of individual life histories. This project makes use of two genome-wide DNA methylation resources created as part of Understanding Society and this initiative primarily focuses on the detailed recorded smoking information available, spanning more than a decade, from yearly mainstage questionnaires. This allows a more precise smoking history to be deduced such as the differences in consistent, low smoking effects and those with lots of variation in the number of cigarettes smoked. Understanding Society then provides a great resource in the analysis of smoking and DNA methylation marks in this project. Other variables from questionnaire data used throughout this project come from the

main survey in waves 2-3 to enable the closest proximity between blood collection, used to obtain DNA methylation profiles, and measures of current lifestyle factors and socioeconomic position.

In 2010-2012 (Waves 2 or 3), after the annual survey, adult respondents were invited to take part in a nurse health assessment interview, which included a range of physical measures and blood samples. With consent the blood samples were frozen for future analysis and DNA extracted. A genome-wide scan using the Illumina human core exome array has been conducted on DNA samples from approximately 10,000 people. DNA methylation was later profiled from the collected DNA samples using the Illumina EPIC methylation array from approximately 3650 participants, consisting of 1425 individuals from the British Household Panel Survey (BHPS) component of Understanding Society and another 2230 from the General Population Sample. The BHPS participants are on average healthier and more affluent than the General Population Sample (GPS) and the youngest participant included in the BPHS samples used in this study is 27 years old whereas the GPS samples are from participants where the youngest member is 18 years old. Bloods from the BHPS sample were collected during Wave 3 while the GPS samples were collected during Wave 2. Understanding Society have also produced a set of biomarkers that either represent key risk factors for diseases that represent major public health problems or reflect key biological pathways between social and environmental factors and health such as inflammatory markers fibrinogen and C-reactive protein (CRP). Understanding Society has been approved by the University of Essex Ethics Committee and the nurse data collection by the National Research Ethics Service (10/H0604/2). All experimental methods performed comply with the Helsinki Declaration.

2.1.2. NCDS (1958 Birth Cohort)

The 1958 National Child Development Study (NCDS) is the second oldest of the British birth cohort studies. The initial sample of 17,415 individuals (8,411 females) consisted of all babies born in Great Britain in a single week in March 1958. These participants have had multiple follow-ups in childhood at 7, 11 and 16 years and in adulthood at 23, 33, 42 and 45 years. This provides high quality prospective data on social, biological, physical, and psychological phenotypes at every sweep. Epigenetic profiles were obtained from DNA samples collected from 529 NCDS subjects at age 44-45, at the same time as intensive phenotyping during this biomedical follow-up which included measures of many biomarkers such as inflammatory markers (Power and Bell, 2006). Epigenetic profiles were generated for two NCDS samples. NCDS1, consisting of 234 subjects, was selected to minimise data missingness for a wide range of exposures in the life course and phenotypes related to healthy ageing. Missingness here refers to missing questionnaire data and information related to participants modifiable and sociodemographic characteristics. NCDS1 subjects were not selected for exposures or outcomes, or for extremes of phenotype distribution. On the other hand, NCDS2, consisted of 294 subjects selected for extremes of adversity exposures in child and adulthood (Fuller et al, 2006). Ethical approval was given by the South-East Multi-Centre Research Ethics Committee (Power and Elliott, 2006).

2.2. EPIC Methylation Array

Most studies investigating epigenetic differences by a phenotype of interest have utilised BeadChip technology using Illumina methylation arrays due to their low cost and high-throughput capabilities enabling genome-wide profiling of DNA methylation marks at single-nucleotide resolution. The genomic coverage of Illumina methylation arrays has increased in size over the years from the Illumina Infinium HumanMethylation27 BeadChip that measures methylation levels at roughly 27,000 CpG dinucleotides (Weisenberger et al., 2008) to the most recent Infinium Methylation EPIC BeadChip with a coverage of

over 850,000 CpG sites. The EPIC array is almost double the size of its predecessor, the Infinium HumanMethylation450 BeadChip, while still covering over 90% of sites found in the previous array. The EPIC array maintains comprehensive coverage of CpG islands and gene promoter regions but also adds better probe coverage of enhancer and gene coding regions and regulatory elements (Pidsley et al., 2016). DNA from whole blood samples was prepared and arrays processed using the protocol detailed by the Illumina manufacturer. The human reference genome build hg19 was used to obtain genetic coordinates and annotate CpG sites to gene names, functions and regions.

The technology used in microarrays first involves a bisulfite conversion of the genomic DNA which converts unmethylated cytosine to uracil and is then subjected to whole genome amplification (WGA) using hexamer priming and Phi29 DNA polymerase. Following this DNA is enzymatically fragmented and purified primers, enzymes and dNTPs are applied to a chip. This chip contains two bead types for each CpG locus, and each bead type is attached to a single stranded 50-mer DNA oligonucleotide that differ in sequence at the free end making them allele specific. One bead type corresponds to the methylated cytosine and the other to the unmethylated cytosine which after conversion to uracil is amplified as thymine in previous steps (Weisenberger et al., 2008). Illumina MethylationEPIC BeadChips employ both Infinium I and Infinium II assays. Type I assay design employs 2 bead types per CpG locus, 1 each for the methylated and unmethylated states and the Type II design uses 1 bead type, with the methylated state determined at the single base extension step after hybridization. The fragmented DNA products are then denatured to single strands and hybridized to the chip via allele specific annealing to either the methylation specific probe or the non-methylation probe. This step is followed by single-base extension with hapten labelled dideoxynucleotides where ddCTP is labeled with biotin and the others (ddATP, ddUTP and ddGTP) are labeled with 2,4-dinitrophenol (Steemers et al., 2006). At this point multi-layered immunohistochemical assays are performed by repeatedly staining with a combination of antibodies that differentiate between the two types. The chip is then scanned to obtain intensities of the unmethylated and methylated bead types

(Bibikova et al., 2011). The system further analyses this microarray data to normalize the raw data and reduce experimental variation effects (Staaf et al., 2008). The ratio between methylated and unmethylated intensities is used to obtain an estimate of the methylation level for each probe or CpG site. This is calculated by dividing the methylated intensity with the sum of unmethylated and methylated intensity plus 100 (M/U+M+100). This is called a beta-value where a value of 0 equates to non-methylation, 1 denotes total methylation and a 0.5 value suggests one copy is methylated but not the other in the diploid human genome (Du et al., 2010). The methylation data is referred to as a beta matrix with *n* columns and *m* rows where *n* refers to the number of samples or participants and *m* refers to the number of CpG sites where DNA methylation is measured.

2.3. Pre-processing and normalisation

2.3.1. UKHLS

Pre-processing, quality control and normalization were carried out in the statistical environment of R (R Core Team, 2017). The bioconductor R package bigmelon was used which has many methods for working with Illumina BeadChip arrays. This package extends the capabilities outlined in the wateRmelon R package (Shalkwyk et al., 2013) by adapting methods from the gdsfmt R package for efficient memory use and management and to overcome the overheads associated with data handing in R (Gorrie-Stone et al., 2017).

Each UKHLS sample was first normalized using the ‘dasen’ function. By doing so technical variation can be more simply dealt with by adjusting intensities rather than the derived “raw” methylation level estimates calculated in the Illumina protocol with little normalization and adjustment. The function involves a

combination of background adjustment using four separate between-array quantile normalizations of methylated Type I, unmethylated Type I, methylated Type II and unmethylated Type II intensities (Pidsley et al., 2013). To elucidate any samples which are grossly affected by this process the function ‘qual’ was used to assess the degree to which the normalized and raw beta values differ. It calculates several performance metrics including the sum of squared differences (SSD) and root mean square error (RMSD) for each sample. A value greater than 0.05 for either SSD or RMSD was used to identify any samples which were noticeably altered when normalised.

Once samples with large discrepancies between original and normalized intensities are removed, the function ‘outlyx’ was utilised to further elucidate data-outliers using a subset of probes from the large data set. This first involves specifying the number of inter-quantile ranges to be discriminated from the upper and lower quantiles which in this case was 2 and are identified from principal component analysis. The computed principal components are used to determine distance measures for each observation and then weights for location and scatter outliers are computed based on these distances and used to determine outliers using an arbitrary threshold of 0.15 (Filzmoser et al., 2008). The next step of quality control checked for sample quality using the ‘bscon’ function which uses the green and red channel readings of the type I and type II bisulfite conversion data to return the median bisulfite conversion percentage value for each array. This quantity shows average conversion of unmethylated cytosine to uracil which is important given that complete conversion is necessary for further study. It uses the intensities of sample-dependent controls included in the array to evaluate performance across arrays. Type I chemistry beta values are calculated by first dividing the first three control probes of the green channel and the second three control probes of the red channel by the sum of all six probes and the unconverted green and red channel probes. Type II chemistry beta values are calculated by simply dividing the methylated red channels by the sum of methylated red and unmethylated green channels. This then outputs a percentage value for bisulfite

conversion. A value of at least 85% conversion was used as a cut-off where samples with noticeably lower values than the rest of the data set were removed from further analysis.

Another quality check consisted of an R implementation of Horvath's epigenetic age clock (Horvath, 2013) using a function named 'agep' to predict the age of each sample. It forms a weighted average of DNA methylation at 353 CpG sites on the human genome that were elucidated using an elastic net regression. This results in a linear regression model whose coefficients correspond to transformed age and used to predict "DNA methylation age" from beta values. The Horvath clock was created using the older 450K microarray and thus 17 of the 353 CpG probe sites used in the DNA methylation-based age predictor are absent in the newer EPIC array. Age prediction was however still fairly accurate but any samples with very large age discrepancies were removed from further analyses. Another quality check visualized sex differences between samples by plotting principal component 1 against principal component 2 to identify if the sex of all participants was correctly matched. Raw intensities per rack were also plotted to identify any obvious batch effects between the different plates. This showed some differences between plates suggesting some technical variation. Although normalization with 'dasen' seems to correct for most of this variation. The two UKHLS DNA methylation resources available have undergone all outlined steps.

2.3.2. NCDS

Samples with a low percentages of bisulfite conversion and any samples that were grossly outlying, or were mismatched based on genotype, sex, and age were removed. The NCDS DNA methylation resource consists of raw beta-values that have not been normalised. As such 'dasen' quantile normalisation could not be used as separate methylated and unmethylated intensities were not available. To overcome this raw beta-values were quantile normalised using 'betaqn'.

2.4. Epigenome-wide association studies

Linear models were carried out using the Bioconductor software-based R package limma. This package allows differential methylation analysis of large-scale microarray data and the identification of differentially methylated CpG sites. The package operates on a matrix of methylation values where rows represent a probe for each genomic feature or CpG site in this case and each column represents the participant sample. The function 'lmFit' fits a generalized, least squares or weighted linear model to each row of data, considering a specified design matrix. This details relevant information related to each sample array and specifies the hypothesis to be tested. Here the treatment-contrasts parametrization method was used to construct design matrices using the 'model.matrix' function. This includes a coefficient for the comparison of interest and any other included covariates. In linear models aimed at comparing DNA methylation between two groups this method is effectively the same as analysis of variance (ANOVA) or multiple regression with a continuous predictor where a model is fitted for each probe (Ritchie et al., 2015).

2.5. Elastic net regression

To build epigenetic biomarkers of smoking one common technique is to use an elastic net regression model using methodology previously described (Horvath, 2013). The elastic net model is designed for high dimensional datasets with more features than samples and where the features are potentially highly correlated (Zou and Hastie, 2005). The model selects the subset of CpG sites that cumulatively produce the best predictor of a provided outcome. Elastic net was implemented in the R package glmnet (Friedman et al., 2010). It uses a combination of Ridge and least absolute shrinkage and selection operator (LASSO) regression. Ridge regression penalizes the sum of squared coefficients and has an (alpha) parameter of 0

while LASSO regression penalises the sum of the absolute values of the coefficients and has an α parameter of 1. Elastic net is a convex combination of ridge and LASSO and therefore the elastic net α parameter was set to 0.5. The lambda value (the shrinkage parameter) was derived using 10-fold cross-validation on the training dataset. The resulting coefficients are then used to predict smoking behaviour using DNA methylation values at the CpG sites selected during the elastic net regularisation. These shrinkage-type regression methods accept biased coefficient estimates in return for lower variance thus obtaining improved prediction accuracy. However, the explanatory variables selected are influenced by every other variable selected and as such the selected predictors may not necessarily be the ones with the strongest association with the outcome and it is not guaranteed that the selected set of variables is correct or truly related to the outcome (Engebretsen and Bohlin, 2019). The aim of a biomarker is to reflect the exposure of interest as closely as possible and is in some way impartial to the biological mechanism at play. Biomarkers simply aim to accurately estimate the trait of interest however this does mean that loci identified using methods like elastic net may not actually reflect the underlying biological mechanism causing the phenotype.

2.6. R packages

All data manipulation, analysis, and visualisation were carried out in R (Version 4.1.1) and the ceres HPC cluster at the University of Essex. The R packages used include bigmelon, haven, misty, data.table, readxl, glmnet, limma, EpiSmokEr, IlluminaHumanMethylationEPICanno.ilm10b2.hg19, tidyverse, pROC, ggtern, grid, gridExtra, ggpubr, labelled, sticky, flextable, gtsummary, GenABEL, ggcorrplot, ggrepel, ggtext, xlsx and report.

3. **Comparison of epigenetic biomarkers of smoking**

3.1. Introduction

An overwhelming amount of research since the first links between smoking and cancer were identified (Doll and Hill, 1950) clearly implicates smoking as one of the most pervasive causes of disease and mortality worldwide. Despite the plethora of diseases caused by smoking and many world-wide policy efforts to reduce smoking behaviours, over one billion people alive today still smoke (WHO, 2020). To better understand the impact of smoking on disease outcomes and socioeconomic inequalities in health, it is important to develop objective measures of health behaviours. Biomarkers are naturally occurring genes, molecules or characteristics by which a physiological process, disease, or lack thereof can be characterised. Previous biomarkers of smoking behaviours have several limitations. Cotinine for example is a metabolite of nicotine and currently the most prevalent biomarker of smoking behaviour but has a half-life of less than a day making it unable to reflect long-term past exposures. DNA methylation-based biomarkers of smoking behaviour may offer an improvement to cotinine by providing a long-term, sensitive and accessible indicator of smoking (Zhang et al., 2016).

Advances in BeadChip microarray technologies have allowed researchers to generate genome-wide datasets of quantifiable epigenetic marks, most commonly measuring DNA methylation at CpG sites. This mechanism can regulate gene activity and cellular function without changing the underlying DNA sequence however the impact of DNA methylation on gene expression in highly context dependent (Schübeler, 2015). DNA methylation in relation to smoking reflects the response of individual's biology to toxic exposures within cigarette smoke. DNA methylation at many sites has been shown to significantly vary between smokers and non-smokers (Joehanes et al., 2016) and this variation has already provided researchers with

sensitive and accurate biomarkers of smoking that can distinguish between active and nascent smokers even in individuals with as little as half of a pack-year of tobacco use (Philibert et al., 2016). Findings suggest that DNA methylation changes observed between smokers and non-smokers occur due to prolonged exposure to cigarette smoke and these changes decay following cessation, the rate at which may be dose-dependent (McCartney et al., 2018). This research has led to robust classifiers of smoking status (Bollepalli et al., 2018) and have provided validated inference of smoking habits such as pack-years in smokers and cessation time in ex-smokers (Maas et al., 2019). Given that epigenetic biomarkers offer an objective way to measure smoking they could also present an opportunity to better understand the impact of observational error in self-reported smoking data. Epigenetic biomarkers of smoking also significantly correlate with mental and physical health outcomes (Corley et al., 2019). It is then vital that the links between smoking status, cumulative exposure and cessation are fully understood in relation to DNA methylation to ensure risk to adverse health outcomes are stratified appropriately.

Many DNA methylation-based biomarkers of smoking have been proposed (Zhang et al., 2016a; Christiansen et al., 2021; Teschendorff et al., 2015; Gao et al., 2016; Yang et al., 2019; Yu et al., 2020; McCartney et al., 2018; Sugden et al., 2019; Odintsova et al., 2021; Elliot et al., 2014; Zhang et al., 2016b; Bollepalli et al., 2019). However, these biomarkers have not been systematically compared in how well they reflect smoking phenotypes such as smoking status, pack years, or years since quitting across different studies. Pack years are a way to measure lifetime cumulative exposure to smoking and is calculated by multiplying the number of packs of cigarettes smoked per day by the number of years a person has smoked. Furthermore, current epigenetic biomarkers of smoking have relied predominantly on retrospectively collected smoking information where data is often collected at a single time point. Recalled smoking data is less reliable than longitudinal data, errors are greatest in ex-smokers, and the extent of inaccuracy increases with time (Krall et al., 1989). DNA methylation has also been almost exclusively measured using 450K Illumina microarrays until recently. This array has also now been superseded by the EPIC array which

can offer wider exploration of the altered epigenetic landscapes seen in smokers compared to non-smokers and offers greater coverage of regulatory elements (Illumina, 2016). This chapter aims to compares published methylation-based biomarkers of smoking to a novel method developed using repeated smoking measures and the newer EPIC microarray. Participants with discrepant smoking information across multiple years of data collection (Table 3.1) were excluded from training data.

Most DNA methylation-based biomarkers of smoking generally employ one of two methods: a smoking index or a methylation score. A smoking index gives a degree of variation in DNA methylation from a never smoker reference across multiple sites. A methylation score represents the average DNA methylation across several CpG sites, each weighted by the effect size related to the phenotype of interest.

$$Smoking\ Index\ (SI) = \frac{1}{n}\sum_{c}^{n} W_c \frac{\beta_{cs-}\mu_c}{\sigma_c}$$

$$Methylation\ Score\ (MS) = (\beta_1 * CpG_1) + (\beta_2 * CpG_2) \ldots + \ (\beta_i * CpG_i)$$

This comparison of methylation-based biomarkers of smoking uses data from Understanding Society, the UK Household Longitudinal Study, a nationally representative panel survey, and the National Child Development Study (NCDS), a birth cohort that began in 1958. We first investigate the characteristics and differences between the available biomarkers. Secondly, we examine how accurately DNA methylation predicts smoking status, pack years and years since quitting. This includes a comparison with novel epigenetic biomarkers of smoking implemented via the 'smokp' function. trained using one of two (USM1 and USM2) DNA methylation resources from Understanding Society. USM1 provides DNA methylation measures from whole blood samples from more than 1000 participants. The capability of each biomarker to predict smoking from DNA methylation was assessed in three samples including NCDS, USM1 and

USM2. The smokp methods were not assessed in USM1 as these were trained using the same samples. All participants within the USM1 sample had previously taken part in the British Household Panel Survey (BHPS) before integration into Understanding Society, coupling large-scale DNA methylation profiles to over ten years of smoking information. Lastly, we investigate the bias in epigenetic biomarkers of pack years and cessation years with age. The aim is to further examine changes to DNA methylation that are associated with smoking, offer recommendations when predicting smoking behaviours from DNA methylation, and better understand the relationship of methylation-based estimates of pack years and cessation with age.

The smokp function is available to use by downloading the watermelon R package here: https://github.com/alexandrayas/wateRmelon. The coefficients used in the smokp function can be found here: https://github.com/alexandrayas/wateRmelon/blob/master/data/smokp_cpgs.rda.

3.2. Methods

3.2.1. Samples

Understanding Society (UKHLS) is an annual household-based panel study which started collecting information about the social, economic, and health status of its participants in 2009. UKHLS collected additional biological information, including blood samples for genetic and epigenetic analysis at wave 3 (2011-2013) for these participants (www.understandingsociety.ac.uk). This meant phenotypic smoking data was available from the many years leading to blood collection. Participants were also asked by the nurse if they had smoked the day their blood was collected.

USM1 refers to the first batch of approximately ~1000 methylation samples produced by Understanding Society. The USM1 analytic data set is drawn from one of the arms of UKHLS, the British Household Panel Survey (BHPS), which began in 1991 and in 2010 was incorporated into UKHLS at the start of its wave 2 (2010-2012). Information on smoking behaviour is collected every year and as such over a decade of longitudinal smoking data, prior to bloods being collected, was available for USM1 participants. USM2 refers to the second batch of UKHLS methylation samples and consists of approximately ~250 of the remaining BHPS samples and the remaining ~2250 come from the General Population Study (GPS) which contains a clustered and stratified, probability sample of households living in Great Britain in 2009-10 that is nationally representative. USM1 consists of 1174 samples and USM2 consists of 2480 making it one of the largest single DNA methylation resources currently available.

The National Child Development Study (NCDS) initial sample consisted of all babies born in Great Britain in a single week in March 1958 and have had multiple follow-ups in childhood at 7, 11 and 16 years and in adulthood at 23, 33, 42 and 45 years. This provides high quality prospective data on social, biological, physical, and psychological phenotypes at every sweep. Epigenetic profiles were obtained from DNA samples collected from 541 NCDS subjects at age 44-45, at the same time as intensive phenotyping during a biomedical follow-up which included measures of many biomarkers such as inflammatory markers (Power and Elliot, 2006). In NCDS two smoking variables were used to classify participants by smoking status which were asked at age 42. One variable coded smoking using 3 levels to define never, ex/occasional, and current smokers and the other coded smoking using 7 levels including those who have never smoked, ex-smokers who smoked at least one cigarette a day, ex-smokers who quit more than 5 years ago, ex-smokers who quit less than 5 years ago, current smokers who smoke less than 10 cigarettes per day, current smokers who smoke 10-20, and current smokers who smoke more than 20 cigarettes per day.

3.2.2. Construction of smoking variables

Smoking status was derived in 1009 participants from substantive interview data utilising responses from two questions to classify each participant into one of three categories: current, former and never smokers. Those answering “Yes” to “Do you smoke cigarettes now?” were assigned as current smokers. Former smokers consisted of participants who answered “No” to smoking now but “Yes” to “Have you ever smoked cigarettes?”. Never smoker participants answered “No” to both questions. To further validate self-reported smoking status, a smoking classification based on repeated smoking measures was obtained where participants were classified as current smokers if they had reported smoking in 2010-12 or also if they had reported smoking in the last 24 hours leading up to blood collection during the nurse visit. Classification of never smokers required participants to have stated not smoking in all waves of data available. Former smokers were classified as such when not smoking in 2010-12 and but smoking regularly or occasionally in previous waves. Data collected at the nurse visit when bloods were collected was also used to further validate smoking status.

To derive more informative smoking phenotypes responses to two further questions, “How old were you when you first started to smoke cigarettes regularly?”, asked in 2010 and 1999 and “How old were you when you last stopped smoking?”, asked in 2010 and 2002, alongside data on number of cigarettes smoked or used to smoke per day, were used to estimate pack years. In current smokers smoking duration was calculated as the difference between one’s age at the nurse visit and their stated age when first started smoking. In former smokers this was calculated as the difference between their age when they first started smoking and their reported age when they last stopped smoking. Years since quitting were estimated in former smokers as the difference between their age at blood collection and the age they reported to have last stopped smoking. Participants who reported to have started smoking before 10 years of age, and participants who reported an age at cessation that was younger than their age when starting smoking were

excluded from analyses. Number of cigarettes smoked per day was asked of current smokers at every year and from former smokers at two time-points (1999 and 2010-12). Pack years estimates cumulative lifetime exposure to cigarette smoking by multiplying the number of packs of cigarettes (number of cigarettes/20) by the number of years a person has smoked. Cessation years refers to the time in years since a former smoker quit smoking or last stopped smoking.

Table 3.1: Timeline of smoking variables utilised from the UK Household Longitudinal Study

Study	**BHPS**										**UKHLS**	
Year	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2010 -12**	**2011 -13**
Wave	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**2**	**3**
Do you smoke cigarettes?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Do you smoke now?	✓										✓	
Have you ever smoked?	✓										✓	
Ever smoked regularly?	✓			✓							✓	
Age started smoking?	✓										✓	
Age stopped smoking?				✓							✓	
N cigs smoked per day?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
N cigs smoked in past?	✓										✓	
Nurse visit, bloods taken												✓
Smoked in the last 24 hours?												✓

3.2.3. DNA methylation and array pre-processing

All participants used in this study provided a blood sample during the UKHLS wave 3 nurse visit approximately 3 months before the main survey and these were sent for storage at -80C before subsequent processing. DNA was then isolated using standard DNA extraction procedures and followed by genome-wide DNA methylome profiling for each participant using the Infinium MethylationEPIC platform (Illumina, 2016). This quantifies DNA methylation at over 850,000 CpG sites in the form of Beta values,

a statistic ranging from 0 to 1 that corresponds to the ratio of methylated signals over the sum of the methylated and unmethylated signals at each site. DNA methylation tends to be biphasic showing a bimodal distribution where a beta-value of 0.5 would suggest one copy is methylated but not the other in the diploid human genome or that the underlying cells that were sampled are highly variable. Quality controls and pre-processing were carried out using the bigmelon Bioconductor package (Gorrie-Stone et al., 2018) in the statistical environment of R (R Core Team, 2017). Beta-values were normalized to control for technical variation while any samples grossly impacted by this process were removed from the dataset. Samples with largely outlying DNA methylation data or those with low bisulfite conversion were also removed.

3.2.4. Data analyses

Generalised linear models (GLMs) were fitted via penalized maximum likelihood to form a weighted average of DNA methylation at multiple CpG loci associated with either smoking status, pack years or cessation in the USM1 participants who were also previously part of BHPS, termed the training set. GLMs with elastic net regularisation were built using the glmnet CRAN package (Friedman et al., 2010) and used to make predictions of smoking status and histories in the remaining USM1 samples as well as USM2 and NCDS, the testing sets. Optimal values for the penalty parameter (lambda) per model were obtained using k-fold (default k = 10) cross validation in each training set. The ‘glmnet’ function automatically selects the coefficients obtained from CpG sites most strongly associated with each smoking phenotype respectively, in turn regulating the selection by which each fit includes only probes that contribute most to the prediction and shrinks the coefficients for all other loci to zero. The regularisation path is computed for the elastic net penalty to estimate smoking, either removing or shrinking correlated model parameters in groups.

GLMs of the multinomial family were used to make DNA methylation-based predictions of smoking status, while gaussian models were fitted for estimating pack years or cessation years. In the multinomial GLMs a probability ranging from 0 to 1 for each of the three smoking status categories is obtained and whichever category shows the largest probability is reported as the best estimate of smoking status per participant. The gaussian GLMs instead predicts an estimated value of the smoking history of interest, namely pack years or cessation in years. Within the models of smoking histories, never smokers were included and coded as having 0 pack years or cessation years. Current smokers were included in the cessation model where their smoking duration was recoded as 'negative cessation'.

All methylation-based biomarkers of smoking on a continuous scale were standardised. This meant these variables were centred by subtracting the mean and then divided by the standard deviation. As all biomarker values seem to have significantly differed between the studies, standardisation was done separately for each dataset.

3.3. Results

3.3.1. Methylation-based biomarkers of smoking

A systematic search for relevant literature related to predicting smoking from DNA methylation was undertaken using PubMed on the 5th of November 2021. The search term used was "(DNA Methylation[Mesh] OR methylation) AND (Smoking[Mesh] OR smoking) AND (predict*)". 408 potentially relevant citations were identified from this literature search. Abstracts were scanned and citations were then limited to those that directly referred to one or more DNA methylation-based biomarkers of adult smoking, were based on Illumina arrays, had available metadata, tested in human blood samples,

and were written in English. 12 distinct and testable methylation-based biomarkers of smoking remained to be investigated. CpG methylation at a single locus (cg05575921) within the AHRR gene was also utilised as a biomarker of smoking given its strong and replicated association with smoking behaviours (Philibert et al., 2013). Alongside these biomarkers 3 novel biomarkers of smoking were developed using Understanding Society Batch 1 samples (USM1). The three novel biomarkers were trained on data indicating smoking status, pack years and cessation years respectively. Smoking status classifies participants into three smoking strata: current, former, and never smokers, pack years give a cumulative lifetime exposure estimate of tobacco use, and cessation years refers to time since quitting smoking. The novel biomarkers are referred to as 'SSt', 'Packyears', and 'Cessation'. The other published biomarkers are referred to by the name of the first author. To note, the biomarkers called 'Elliot', 'Zhang2', and 'Bollepalli' refer to those implemented in the EpiSmokEr R package (Bollepalli et al., 2019). In total 16 different biomarkers for estimating smoking from DNA methylation were compared (Table 3.2). All biomarkers, except for the EpiSmokEr biomarkers, are implemented in the 'smokp' R function. This function takes a beta matrix of DNA methylation measurements and the method to be used as the input. If the specified method is a smoking index a further input is required which specifies the smoking status of each participant. This is needed to obtain an average DNA methylation measurement at each CpG site within the reference 'never smoker' samples from which deviation by smoking is measured. The 'smokp' function then outputs a data frame containing a DNAme-based smoking estimate per sample, calculated via the method specified.

Table 3.2: DNA methylation-based biomarkers of smoking

First author	Year	Studies used	N samples	N CpGs	N available
Quartiles					
Zhang[1]	2016a	ESTHER	1000	2	1
Smoking Index					
Teschendorff[2]	2015	NSHD	Discovery = 400, Validation = 390	1,501	1,399
Gao[3]	2016	ESTHER	Discovery = 1,000, Validation = 548	66	61
Yang[4]	2019	NAS	692	52	48
Yu[5]	2020	ESTHER	1603 (143 LC cases and 1,460 controls)	151	139
Methylation score					
Christiansen[6]	2021	TwinsUK, BCS70, NCDS, NSHD	Discovery = 1,407, Validation = 3,425	2	2
McCartney[7]	2018	GS, LBC1936	Discovery = 5,087, Replication = 895	233	228
Sugden[8]	2019	Dunedin, E-Risk	1,037, 2,232 twins	2,623	2,430
Odintsova[9]	2021	NTR	Discovery = 2,431, Replication = 1,128	24	20
EpiSmokEr					
Elliot[10] **(SSc*)**	2014	SABRE	192	187	173
Zhang[11] **(MS*)**	2016b	ESTHER	Discovery = 500, Replication = 500	4	2
Bollepalli[12] **(SSt*)**	2019	FINRISK, FTC, EIRA, CARDIOGENICS	Discovery = 514, Rep. = 408 twins, 687, 464	121	111
smokp					
SSt	2021	Understanding Society, NCDS	Training = 1,009, Testing = 3,141	87	-
Packyears			Training = 585, Testing = 1,678	41	-
Cessation			Training = 670, Testing = 979	292	-

[1] Zhang, Y., Florath, I., Saum, K. U., & Brenner, H. (2016). Self-reported smoking, serum cotinine, and blood DNA methylation. Environmental research, 146, 395-403.

[2] Teschendorff, A.E., Yang, Z., Wong, A., Pipinikas, C.P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H.B., Thirlwell, C. and Janes, S.M. (2015) Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. JAMA oncology, 1(4), pp.476-485.

[3] Gao, X., Zhang, Y., Breitling, L.P. and Brenner, H. (2016) Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget, 7(30), p.46878.

[4] Yang, Y., Gao, X., Just, A.C., Colicino, E., Wang, C., Coull, B.A., Hou, L., Zheng, Y., Vokonas, P., Schwartz, J. and Baccarelli, A.A. (2019) Smoking-related DNA methylation is associated with DNA methylation phenotypic age acceleration: The veterans affairs normative aging study. International journal of environmental research and public health, 16(13), p.2356.

[5] Yu, H., Raut, J.R., Schöttker, B., Holleczek, B., Zhang, Y. and Brenner, H. (2020) Individual and joint contributions of genetic and methylation risk scores for enhancing lung cancer risk stratification: data from a population-based cohort in Germany. Clinical epigenetics, 12(1), pp.1-11.

[6] Christiansen, C., Castillo-Fernandez, J.E., Domingo-Relloso, A., Zhao, W., Moustafa, J.E.S., Tsai, P.C., Maddock, J., Haack, K., Cole, S.A., Kardia, S.L.R. and Molokhia, M. (2021) Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clinical epigenetics, 13(1), pp.1-13.

[7] McCartney, D.L., Hillary, R.F., Stevenson, A.J., Ritchie, S.J., Walker, R.M., Zhang, Q., Morris, S.W., Bermingham, M.L., Campbell, A., Murray, A.D. and Whalley, H.C. (2018) Epigenetic prediction of complex traits and death. Genome biology, 19(1), pp.1-11.

[8] Sugden, K., Hannon, E.J., Arseneault, L., Belsky, D.W., Broadbent, J.M., Corcoran, D.L., Hancox, R.J., Houts, R.M., Moffitt, T.E., Poulton, R. and Prinz, J.A. (2019) Establishing a generalized polyepigenetic biomarker for tobacco smoking. Translational psychiatry, 9(1), pp.1-12.

[9] Odintsova, V.V., Rebattu, V., Hagenbeek, F.A., Pool, R., Beck, J.J., Ehli, E.A., van Beijsterveldt, C.E., Ligthart, L., Willemsen, G., De Geus, E.J. and Hottenga, J.J. (2021) Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. Frontiers in Psychiatry, 12.

[10] Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Smith, G.D., Hughes, A.D., Chaturvedi, N. and Relton, C.L. (2014) Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clinical epigenetics, 6(1), pp.1-10.

[11] Zhang, Y., Schöttker, B., Florath, I., Stock, C., Butterbach, K., Holleczek, B. & Brenner, H. (2016). Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environmental health perspectives, 124(1), 67-74.

[12] Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S., & Ollikainen, M. (2019). EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. Epigenomics, 11(13), 1469-1486.

[*] EpiSmokEr method names (Bollepalli et al., 2019)

The 12 biomarkers identified from the literature search were published between 2014 and 2021. 4 biomarkers (Zhang et al., 2016a, Zhang et al., 2016b, Gao et al., 2016, Yu et al., 2020) used data from the "Epidemiological investigations on chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population" (ESTHER) study. ESTHER is a large prospective observational cohort study initially consisting of 1000 Germans aged between 50 to 75 years at blood collection. One method proposed by Zhang et al (2016a) involves a score based on 2 loci (cg05575921 and cg06126421) where individuals are given a tally of 0, 1, or 2 depending on whether CpG methylation at the two loci fell within the lowest quartile. The 2 CpG sites were chosen as these showed the strongest associations with all-cause, cardiovascular, and cancer mortality out of nine tested. The nine tested CpGs were selected due to their replicated association with both current and lifetime smoking. Unfortunately, only 1 of the 2 CpG sites used in this method were available in methylation data used this study. As such only scores of 0 or 1 were used. The EpiSmokEr MS and second Zhang et al (2016b) method uses a methylation score consisting of 4 loci (cg05575921*(-10.94) + cg05951221*(-12.04) + cg02451831*(16.01) + cg06126421*(-8.45)). This methylation score was shown to provide the best discrimination between former and never smokers compared to using single CpG methylation measures at the AHRR locus (cg05575921) or cotinine where a cut-off of 4.85 ng/ml was used. The smoking estimate returned by this method is the sum of methylation measured at the 4 sites weighted by their effect size. The effect size in this case refers to the β-regression coefficient of each CpG in relation to its association with cotinine. 2 out of 4 of the CpG sites used in this method were available in this study. This method is also implemented in EpiSmokEr (Bollepalli et al., 2019).

Gao et al (2016) made use of two independent subsamples of ESTHER including the initial 1000 participants and an additional 548 participants, used for validation, who joined the study after the initial recruitment. Yu et al (2020) made use of a case-control study nested within ESTHER consisting of 1460 lung cancer free participants and 143 incident lung cancer (LC) patients. Yang et al (2019) utilised data

from 692 male participants within the Veterans Affairs Normative Aging Study (NAS), a closed longitudinal study of aging in men from eastern Massachusetts. All three studies constructed a smoking index to estimate smoking behaviour from DNA methylation. Here a smoking index simply indicates the degree of deviation in DNA methylation from a reference of never smokers. All three biomarkers also used smoking associated CpGs identified at least twice within a previous systematic review made up of 14 epigenome-wide association studies (EWAS) and 3 gene specific methylation studies (GSMS) (Gao et al., 2015). While Yu et al (2020) made use of all 151 replicated loci in their smoking index, Gao et al (2016) further restricted these to 66 sites that were significantly associated with age acceleration. Age acceleration is a term used to describe the residuals when DNA 'methylation age' (Horvath, 2013) is regressed on to chronological age. The resulting index from Gao et al (2016) showed a monotonic dose-response relationship with age acceleration. Yang et al (2019) also restricted their smoking index to only include CpG sites associated with another methylation-based biomarker. In this case 52 smoking associated CpG sites were found to be significantly associated with DNAmPhenoAge acceleration. DNAmPhenoAge is another aging biomarker but also a predictor of healthspan and chronic disease risk (Levine et al., 2018). Of the 151 replicated smoking associated CpGs, 139 were available in the test datasets including 61 of the 66 sites significantly associated with age acceleration and 48 out of 52 sites significantly associated with DNAmPhenoAge acceleration. Teschendorff et al (2015) first developed the algorithm used to construct a smoking index based on DNA methylation. Their study consisted of 790 women from the MRC National Survey for Health and Development (NSHD) study who all gave a buccal sample when aged 53. These women were split into two groups consisting of 400 and 390 participants respectively for discovery and replication purposes. 152 had matched whole blood samples. Teschendorff et al (2015) identified 1501 validated CpG sites significantly correlated with pack years, and these were used in this smoking index. Of these sites 1399 were available within this study.

Christiansen et al (2021) also developed their biomarker of smoking using NSHD alongside other UK population-based cohorts including the 1958 National Child Development Study (NCDS), the 1970 British Cohort Study (BCS70), and the TwinsUK cohort, totalling 1407 participants. Trans-ethnic replication was carried out using the Strong Heart Study (SHS) and Genetic Epidemiology Network of Arteriopathy Study (GENOA). Here 2 CpGs, cg05575921 (AHRR) and cg00045592 (SLAMF7), were utilised as a biomarker of smoking to distinguish between current and never smokers and this biomarker was utilised in our study. However, different combinations of the top 5 ex-smoking related sites were also used by Christiansen et al (2021) to predict smoke exposure but these were not used in our study. The 2 CpG sites used were available and used in a methylation score.

McCartney et al (2018) trained their methylation score, consisting of 233 CpG sites, using data from 5087 participants, aged 18–99 years, from Generation Scotland. Generation Scotland is a family-structured, population-based longitudinal cohort study. These sites were selected by using penalized regression models where residuals from smoking regressed on age, sex, and ten genetic principal components were used. The McCartney et al (2018) biomarker was tested in 895 participants within the 1936 Lothian Birth Cohort who were aged approximately 70 when bloods were collected. Specifically, data from the Stratifying Resilience and Depression Longitudinally (STRADL) sub-study was used. 228 of the 233 CpGs used in this biomarker were available in our study.

Both Sugden et al (2019) and Odintsova et al (2021) also used a methylation score to estimate smoking from DNA methylation. Both studies used effect sizes from a previously published meta-analysis (Joehanes et al, 2016). This meta-analysis consisted of 16 cohorts within the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium. CHARGE consists of 15,907 participants in total. These 16 cohorts include the Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study

European Ancestry (CHS EA), Cardiovascular Health Study African Ancestry (CHS AA), European Prospective Investigation into Cancer (EPIC), European Prospective Investigation into Cancer and Nutrition-Norfolk (EPIC Norfolk), Framingham Heart Study (FHS), Genetic Epidemiology Network of Arteriopathy (GENOA), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN), Grady Trauma Project (GTP), "Invecchiare in Chianti" (InCHIANTI), Cooperative health research in the Region of Augsburg follow-up survey 4 (KORA F4), Lothian Birth Cohorts of 1921 and 1936 (LBC 1921 and LBC 1936), the Multi Ethnic Study of Atherosclerosis (MESA), Normative Aging Study (NAS) and the Rotterdam Study (RS). Sugden et al (2019) tested in the Dunedin Longitudinal Study birth cohort where bloods at both age 26 and 38 were collected, and the Environmental Risk (E-Risk) Longitudinal Twin Study where bloods were collected at age 18. This methylation score consisted of 2623 CpG sites and 2430 were available in our study. Instead of a sum of methylation measures, weighted by the effect size, Sugden et al (2019) used an average. Odintsova et al (2021) calculated multiple methylation scores based on different subsets of CpGs according to their significance level. Subsets of CpGs were selected based on p-value thresholds and pruned in a stepwise selection of the most significant CpG sites while discounting any highly correlated probes. The best biomarker involved 24 pruned CpGs at $p < 1 \times 10{-}7$ significance threshold and the top 24 smoking associated sites from Joehanes et al (2016) were used here. 20 CpGs were available in our study.

The readily available R package EpiSmokEr (Bollepalli et al., 2019) includes three different methylation-based predictors of smoking. Elliot et al (2014) used CpGs from a previously published EWAS using the Cooperative Health Research in the Region Augsburg (KORA) study (Zeilinger et al, 2013). In this EWAS a discovery and replication of 1793 and 479 participants respectively was used. Elliot et al (2014) tested their method in the Southall And Brent REvisited (SABRE) study and is also investigated within the EpiSmokEr publication. 173 out of 187 CpGs were available in our study. Bollepalli et al (2019), the authors of EpiSmokEr, used the Dietary, Lifestyle and Genetic determinants of Obesity and Metabolic

syndrome (DILGOM) data from the Finnish population based FINRISK 2007 study. Their training dataset consisted of 514 participants, and this was used to develop an epigenetic predictor of smoking status and is the only method other than smokp described here that readily outputs and categorises participants into nominal smoking groups by utilising DNAme-based methylation scores. This was tested on 408 twins from the Finnish Twin Cohort (FTC), 687 participants from the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study with 354 rheumatoid arthritis cases, and 464 samples from the CARDIOGENICS consortium, comprising samples from healthy subjects and from subjects with coronary artery disease. The EpiSmokEr SSt method predicts smoking status from 121 CpG sites. 111 were available in our study.

All studies, other than Christiansen et al (2021), Odintsova et al (2021) and McCartney et al (2018), measured DNA methylation using 450K microarray technology. However, McCartney et al (2018) subset EPIC microarray data to include only probes available on both arrays. 4 out of the 12 previously published biomarkers used an out of sample study for replication purposes. 6 of the mentioned studies suggested that methylation-based measures of smoking were predictive of health or aging related phenotypes. Zhang et al (2016a) showed that adding methylation measures from two smoking-related sites improved fatal cardiovascular risk prediction. Yu et al (2020) showed that methylation scores related to smoking improve lung cancer risk stratification. Sugden et al (2019) showed that their generalized polyepigenetic biomarker of smoking correlates with changes in lung function and gum health and could predict changes in gene expression in pathways related to inflammation and immunity. McCartney et al (2018) showed that their DNAme smoking biomarker significantly predicted mortality. Gao et al (2016) showed that DNAme indicators of smoking, but not self-reported smoking, were associated with age acceleration and similarly Yang et al (2019) showed that DNAme-based smoking indicators and self-reported pack years, but not self-reported smoking status nor cotinine, were significantly associated with DNAmPhenoAge. This in turn suggests a utility for DNAme-based biomarkers of smoking in predicting health outcomes and within epidemiology in general.

Within all 16 tested DNAme-based biomarkers of smoking, a total of 5607 different loci were used, and this included 667 loci used in more than one method 4940 distinct CpG sites. Only 1 CpG, (cg05575921) located within the *AHRR* locus was used in more than 9 biomarkers. The only biomarker to not use this AHRR site were the smokp Packyears biomarker and the Odintsova et al (2021) biomarker. 19 CpG were seen in at least 5 biomarkers, and 4 of these were located at the AHRR locus and a further 4 were in close proximity in the intergenic region on chromosome 2q37.1. 161 (3%) of unique CpG sites were used in at least 2 biomarkers. 4641 (94%) of these CpG sites were available in methylation datasets used throughout this study, meaning 299 CpG sites were missing.

All studies had assessed smoking by comparing current or ever smokers to never smokers. Some studies (Zhang et al., 2016a, Teschendorff et al., 2015, Gao et al., 2016, Yang et al., 2019, McCartney et al., 2018, Zhang et al., 2016b, and Bollepalli et al., 2018) all had data on cumulative lifetime smoking such as pack years, or past smoking, or cotinine. All studies had controlled for age and sex or looked at men and women separately or used a birth cohort. Summary statistics from meta-analyses from Joehanes et al (2016) and Zeilinger et al (2013) were often used. Both assessed smoking using smoking status, and both controlled for sex, age and blood count while Joehanes et al (2016) controlled for technical covariates, and Zeilinger et al (2013) also included BMI and alcohol consumption as covariates. Few studies (Elliot et al (2014), Christiansen et al., 2021) looked at ethnicity. Most studies but not all further controlled for random batch effects, technical variation, leukocyte distribution and bisulfite conversion efficiency. Christiansen et al (2021) also controlled for BMI and Zhang et al (2016a, 2016b), Gao et al (2016), and Yang et al (2019) also controlled for BMI, physical activity, prevalence of CVD, diabetes, and cancer. Zhang et al (2016b) further adjusted for total cholesterol and systolic blood pressure. Gao et al (2016) and Yang et al (2019) further adjusted for alcohol consumption, and the latter also added alcohol consumption, years of education,

hypertension and stroke. A few studies went on to look at how smoking assessed using DNA methylation may associate with cancer and lung lesions (Teschendorff et al., 2015), aging-related diseases, including cardiovascular diseases, diabetes and cancer (Gao et al, 2016), lung cancer incidence (Yu et al., 2020), mortality risk (McCartney et al., 2018), gum health, lung function and the interplay with adverse childhood experiences (ACEs) (Sugden et al., 2019).

3.3.2. Training data

The 3 main biomarkers of smoking outlined in this chapter (SSt, Packyears, Cessation) were trained using questionnaire and methylation data from Understanding Society, Batch 1 (USM1) participants (Table 3.3). These biomarkers aim to estimate smoking status (SSt), defined as one of three categories including current, former and never smokers, pack years (Packyears), a cumulative lifetime measure of tobacco use, and years since quitting (Cessation) respectively. All biomarkers, alongside the 12 previously published biomarkers described above, are implemented in the smokp function. The training data (USM1) used when training the SSt smokp biomarker were limited to participants whose self-reported smoking status were coherent across multiple years of data collection, as outlined in the biomarkers section. By doing so the smoking data used for training is more reliable compared to measures collected at a single time point. Participants in all three smoking categories were included when training the three smokp biomarkers. Never smokers were coded as having 0 pack years and 0 cessation years within the Packyears and Cessation biomarkers. Smoking duration reported by current smokers was subtracted from 0 and recoded as ‘negative cessation’ in the Cessation biomarker.

1,009 participants were included in the training dataset used in the smokp SSt biomarker. This consisted of 445 never, 410 former, and 154 current smokers. 58% of this dataset were female and the mean age

measured 58.5 (± 14.9) years. The proportion of women varied across smoking status where 64%, 51% and 60% of never, former and current smokers respectively identified as female. The mean age of participants in the smokp SSt training dataset also varied by smoking status and measured 57.1 (± 14.8), 62.3 (± 14.8) and 52.3 (±13.1) in never, former and current smokers respectively. This shows a higher proportion of women reported currently or never smoking compared to former smoking. This also shows that former smokers were on average older than current smokers and current smokers were on average younger than never smokers. Further, male former smokers were on average older (64.3 ± 13.9) than female former smokers (60.5 ± 13.3) while male current smokers were on average younger (50.0 ± 12.6) than female current smokers (53.8 ± 13.3). There was little difference in age between male (56.4 ± 14.9) and female (57.5 ± 14.8) never smokers.

976 participants were included in the training dataset used in the smokp Packyears biomarker. This consisted of 445 never smokers, coded as having 0 smoking pack years, and 351 former and 180 current smokers. Across both current and former smokers, the median number of pack years measured 28.5 and ranged from 0 to 152.5. In current smokers, median pack years measured 23.3 and ranged from 0 to 84. In former smokers the median number of pack years were on average greater than current smokers and measured 33.5, ranging from 1.6 to 152.5. Across all 976 participants the mean age was 57.7 (± 15.0) and 58% were female. Within the 351 included former smokers the mean age in years measured 61.5 (± 15.2) and 49% were female. The mean age of the 180 included current smokers measured 51.69 (± 13.32) and 62% were female. The mean age of the 445 included never smokers measured 57.1 (± 14.8) and 64% were female. Male participants in this training dataset reported on average a greater number of pack years (40.4 ± 28.4) compared to female participants (26.9 ± 21.1). This difference is much more pronounced between male (45.6 ± 29.6) and female (29.3 ± 23.4) former smokers compared to male (26.9 ± 19.3) and female (24.9 ± 16.8) current smokers. There is also a greater amount of variation in pack years in male compared to female participants within the dataset being described.

958 participants were included when training the smokp Cessation biomarker. This consisted of 445 never smokers, coded as having 0 cessation years, 362 former smokers, and 151 current smokers whose smoking duration were recoded as 'negative cessation' years. Within the 362 former smokers the median number of years since quitting was 21 and ranged from 1 to 66 years. The median number of 'negative cessation' years in current smokers was -35 and ranged from -64 to -7. Across all 958 participants the mean age measured 58.1 (± 15.0) and 58% were female. The mean age of 177 included former smokers measured 61.7 (± 15.1) years and 48% were female. The mean age of the 151 included current smokers measured 52.2 (± 13.2) and 61% were female. The mean age of the 445 included never smokers measured 57.1 (± 14.8) and 64% were female. Male participants reported on average a greater number of cessation years (24.6 ± 14.9) compared to female participants (19.1 ± 14.0).

3.3.3. Testing data

CpG methylation at the frequently reported *AHRR* locus, the three smokp biomarkers mentioned above (SSt, Packyears and Cessation) and the 12 other previously published biomarkers of smoking meant a total of 16 different biomarkers were compared in this chapter. Each biomarker was compared in how well they estimated smoking behaviours in two independent testing datasets. The two testing datasets consisted of the data from the National Childhood Development Study (NCDS) and Understanding Society, Batch 2 (USM2). In total 2,980 participants had at least reported their smoking status and were included in the testing data.

In USM2 roughly 90% of participants were not part of the proceeding BHPS study and as such only smoking data collected at a single time point was used in testing the described biomarkers of smoking.

2,478 participants stated their smoking status during the 2010-12 main survey. This included 978 never, 992 former, and 508 current smokers. The mean age across all participants measured 50.5 (± 15.4) years and 54% were female. Age and sex also varied by smoking status. The mean ages measured 49.4 (± 15.9), 53.1 (± 15.0), and 47.4 (± 14.4) in never, former and current smokers respectively. The proportion of women was 60%, 49% and 52% in never, former and current smokers respectively. 1,160 current and former smoking participants also reported the age at which they started smoking and the number of cigarettes they did or do smoke per day and as such pack years were calculated for these participants. The median number of pack years across all USM2 participants measured 27.9 and ranged from 0.15 to 132.5. The 1,160 participants included 679 former and 481 current smokers. The median number of pack years in former smokers measured 27.8 and ranged from 0.35 to 132.5. This was greater than the median number of pack years in current smokers which measured 17.0 and ranged from 0.15 to 92. Also, on average male USM2 participants reported a greater number of pack years (32.3 ± 25.6) compared to female participants (23.9 ± 19.5). 681 former smokers in USM2 reported the age that they last quit smoking and cessation years were calculated. The median number of cessation years measured 16 and ranged from 1 to 61. On average the number of cessation years reported by male participants was greater (21.2 ± 14.6) than that reported by female participants (17.1 ± 13.2).

In NCDS a total of 502 participants answered two questions categorising their smoking behaviour and did not display discrepancies between the two where one classified smoking into 3 categories, as used here, and the other categorised smoking into 4 categories where everyday smokers and occasional smokers are distinguished from each other. This includes 200 never, 140 former, and 162 current smokers. 98% of participants were aged 44 when bloods were collected. 52% of participants identified as female however this varied by smoking status where 57% of never, 46% of former, and 51% of current smokers identified as female. 127 current smokers further reported the age they started smoking and how many cigarettes they smoked per day meaning pack years could be calculated. The median number of pack years in NCDS

measured 26 and ranged from 6.8 to 66. There was minimal difference between male (24.5 ± 12.4) and female (26.4 ± 11.3) NCDS participants in terms of pack years although female participants tended to report a greater number of pack years in contrast to USM participants. 113 former smokers further reported the age when they quit smoking meaning years since quitting was calculated. The median number of cessation years in NCDS measured 13 and ranged from 2 to 34. Again, there was little difference in reported cessation years between male (13.1 ± 7.6) and female (13.6 ± 6.7) participants in although on average NCDS females reported slightly greater number of years since quitting, in contrast to USM participants.

The USM1 dataset described previously was also used to test how well DNAme-based biomarkers of smoking estimated self-reported measures. USM1 was not used to test the smokp SSt, Packyears, and Cessation biomarkers as these were trained using the same dataset.

Table 3.3: Participant characteristics

Characteristic	N	Never	Former	Current	Overall
Understanding Society, Batch 1 (USM1)					
N	1,170	498	486	186	1,170
Age, median (range)	1,170	57 (28 - 97)	64 (28 - 98)	50 (28 - 83)	59 (28, 98)
Sex, n (%)	1,170				
Male, n (%)	486	177 (36%)	235 (48%)	74 (40%)	486 (42%)
Female, n (%)	684	321 (64%)	251 (52%)	112 (60%)	684 (58%)
Self-reported smoking status (repeated measures), n (%)	1,009	445 (89%)	410 (84%)	154 (83%)	1,009 (86%)
Pack years, median (range)	976	-	33 (2, 152)	23 (0, 84)	445 (44%)
Cessation years, median (range)	958	-	21 (1, 66)	-35 (-64, -7)	410 (41%)
National Child Development Study (NCDS)					
N	502	200	140	162	502
Age, n (%)	502				
44	492	193 (96%)	139 (99%)	160 (99%)	492 (98%)
45	10	7 (3.5%)	1 (0.7%)	2 (1.2%)	10 (2.0%)
Sex, n (%)	502				
Male	241	86 (43%)	76 (54%)	79 (49%)	241 (48%)
Female	261	114 (57%)	64 (46%)	83 (51%)	261 (52%)
Pack years, median (range)	127	-	-	26 (7, 66)	26 (7, 66)
Cessation years, median (range)	113	-	13 (2, 34)	-	13 (2, 34)
Understanding Society, Batch 2 (USM2)					
N	2,478	978	992	508	2,478
Age, median (range)	2,478	50 (16 - 88)	55 (16 - 83)	48 (16 - 81)	51 (16 - 88)
Sex, n (%)	2,478				
Male	1,132	389 (40%)	501 (51%)	242 (48%)	1,132 (46%)
Female	1,346	589 (60%)	491 (49%)	266 (52%)	1,346 (54%)
Pack years, median (range)	1,160	-	28 (0, 132)	17 (0, 92)	22 (0, 132)
Cessation years, median (range)	681	-	16 (1, 61)	-	16 (1, 61)

3.3.4. Prediction of smoking status

Logistic regression models were fitted with one of three comparisons between smoking status (Never vs Current, Former vs Current, Never vs Former) as the dependent variable and methylation-based smoking estimates from each of the 16 biomarkers separately as the independent variable. There was a strong statistically significant difference in DNAme-predicted biomarker values between self-reported never and current smokers when using all 16 methods ($p < 0.001$). Odds ratios (ORs) using the 14 numeric continuous biomarkers ranged from 0.02 when using *AHRR* CpG methylation (CI = 0.01-0.03), or the smokp Cessation method (CI = 0.01-0.03), to 83.6 when using the McCartney et al (2018) method (CI = 55.1-133). As these biomarkers were standardised odd ratios here represent the likelihood of self-reporting current smoking compared to never smoking per one standard deviation increase in the biomarker values. DNAme-predicted current smokers had an OR of 899 (CI = 426-2,318), and DNAme-predicted former smokers had an OR of 7.97 (CI = 5.66-11.2), compared to DNAme-predicted never smokers when using the smokp SSt method (Table 3.4).

There was also a statistically significant difference in biomarker values between former and current smokers when using all methods ($p < 0.001$). Odds ratios (ORs) using the 14 numeric continuous biomarkers ranged from 0.09 when using the smokp Cessation method (CI = 0.07-0.10), to 9.08 when using the McCartney et al (2018) method (CI = 7.74-10.7). A 1 SD increase in the McCartney et al methylation score in these biomarkers meant a 9 times greater likelihood of self-reporting currently smoking compared to former smoking. DNAme-predicted current smokers had an OR of 55.6 (CI = 39.7-79.2), and DNAme-predicted former smokers had an OR of 1.47 (CI = 1.08-2.02), compared to DNAme-predicted never smokers when using the smokp SSt method. 15 out of the 16 tested biomarkers, including *AHRR* CpG methylation and the smokp SSt, smokp Packyears, smokp Cessation, EpiSmokEr MS, Sugden, McCartney, Christiansen,

Odintsova, Yu, Gao, Yang, Teschendorff, and Zhang methods ($p < 0.001$), and the EpiSmokEr SSc method (OR = 1.10, CI = 1.03-1.19, $p = 0.006$) all showed a statistically significant difference in biomarker values between self-reported former and current smokers. ORs ranged from 0.16 (CI = 0.13-0.19) using *AHRR* CpG methylation, to 9.78 (CI = 7.85-12.3) using the McCartney method. There was not a significant difference in DNAme-predict smoking status using the EpiSmokEr SSt method ($p = 0.10$) between self-reported never and former smokers (Table 3.4).

Receiver operating characteristic (ROC) curves were fitted to estimate the area under the curve (AUC) and quantify how well each biomarker distinguished between smoking strata. In total, across all 16 biomarkers, three comparisons and testing datasets, AUC values ranged from 0.474 to 0.999 (Table 3.5). AUCs ranged from 0.474 to 0.999 when distinguishing between never and current smokers. However, the Elliot et al (2014), Zhang et al (2016), and Bollepalli et al (2019) methods appear to have performed particularly badly in the two Understanding Society datasets. Without these methods included DNAme-based biomarkers of smoking were able to differentiate never and current smokers with AUCs ranging from 0.769 to 0.999. The best classifier between never and current smoking across the three testing datasets was the methylation score from McCartney et al (2018) with an average AUC of 0.984, closely followed by using *AHRR* CpG methylation (0.979) and the smokp Cessation biomarker (0.974).

AUC values representing each biomarker's ability to distinguish between former and current smoking ranged from 0.481 to 0.978. When not including the methods implemented via the EpiSmokEr R package AUC values ranged from 0.677 to 0.978. The best classifier between former and current smoking was the McCartney et al (2018) biomarker with an average AUC of 0.935 across the three testing datasets. This was again followed by using *AHRR* CpG methylation with an average AUC of 0.926 and then the smokp Cessation biomarker (0.921). In distinguishing between never and former smokers AUC values ranged

from 0.480 to 0.786. Without the three poorly performing methods, AUCs ranged from 0.531 to 0.786. The best classifiers between never and former smoking was also the McCartney et al (2018) method which showed an average AUC of 0.747, followed by *AHRR* CpG methylation with an average AUC of 0.708. The third best biomarker at differentiating never and former smokers was the smokp Packyears biomarker which showed an average AUC of 0.701. The least effective methylation-based classifiers of smoking strata within the two Understanding Society sub-samples were the Elliot et al (2014), Zhang et al (2016), and Bollepalli et al (2019) methods, none of which achieved an AUC value higher than 0.523 in USM1 and USM2. In NCDS however the EpiSmokEr biomarkers performed well and were able to distinguish between smoking strata with AUC values of up to 0.964.

Across the three comparisons between smoking status, and three testing datasets, the methylation-based biomarker of smoking that showed the best class separation capacity was from McCartney et al (2018) with an average overall AUC of 0.888. This was followed by single CpG methylation measures at the *AHRR* locus with an average AUC of 0.870. The third best overall method was from Yu et al (2020) with an average overall AUC of 0.846. This suggests that the addition of more CpGs does not drastically improve the ability of DNA methylation to accurately predict smoking status outside of the *AHRR* locus. The benefit of the smokp and EpiSmokEr SSt methods over others is that an interpretable smoking measure, namely smoking status, is predicted from DNA methylation and this reflects how epidemiological studies typically measure smoking. Often the output of a methylation score or smoking index is useful however it is unclear what thresholds would denote whether a person smokes or not and these thresholds may not be the same across different populations.

Table 3.4: Binomial logistic regression outputs showing the relationship between each of 16 DNAme-based biomarkers of smoking with three comparisons between smoking status to estimate the effect of biomarker values on smoking across all three datasets

	Never vs Current (N = 2,447/1,848)				Former vs Current (N = 2,366/1,802)				Never vs Former (N = 3,165/2,310)			
Method	**OR**	**95% CI**	**p-value**	**q-value**	**OR**	**95% CI**	**p-value**	**q-value**	**OR**	**95% CI**	**p-value**	**q-value**
AHRR	0.02	0.01, 0.03	<0.001	<0.001	0.13	0.11, 0.15	<0.001	<0.001	0.16	0.13, 0.19	<0.001	<0.001
smokp SSt			<0.001	<0.001			<0.001	<0.001			<0.001	<0.001
Never	—	—			—	—			—	—		
Former	7.97	5.66, 11.2			1.47	1.08, 2.02			5.41	4.34, 6.76		
Current	899	426, 2,318			55.6	39.7, 79.2			16.2	7.53, 42.0		
smokp Packyears	7.01	6.03, 8.20	<0.001	<0.001	1.88	1.71, 2.06	<0.001	<0.001	3.04	2.74, 3.39	<0.001	<0.001
smokp Cessation	0.02	0.01, 0.03	<0.001	<0.001	0.09	0.07, 0.10	<0.001	<0.001	0.68	0.60, 0.77	<0.001	<0.001
EpiSmokEr SSt			<0.001	<0.001			<0.001	<0.001			0.10	0.10
Never	—	—			—	—			—	—		
Former	1.18	0.95, 1.46			1.16	0.93, 1.44			1.02	0.86, 1.20		
Current	2.50	1.98, 3.16			2.06	1.63, 2.60			1.21	1.00, 1.48		
EpiSmokEr SSc	1.36	1.26, 1.48	<0.001	<0.001	1.23	1.14, 1.33	<0.001	<0.001	1.10	1.03, 1.19	0.006	0.007
EpiSmokEr MS	1.36	1.25, 1.48	<0.001	<0.001	1.22	1.12, 1.32	<0.001	<0.001	1.13	1.05, 1.21	<0.001	0.001
Sugden	15.0	12.2, 18.7	<0.001	<0.001	5.03	4.42, 5.76	<0.001	<0.001	2.44	2.18, 2.75	<0.001	<0.001
McCartney	83.6	55.1, 133	<0.001	<0.001	9.08	7.74, 10.7	<0.001	<0.001	9.78	7.85, 12.3	<0.001	<0.001
Christiansen	0.03	0.03, 0.04	<0.001	<0.001	0.16	0.14, 0.18	<0.001	<0.001	0.31	0.26, 0.35	<0.001	<0.001
Odintsova	5.14	4.48, 5.92	<0.001	<0.001	3.27	2.92, 3.67	<0.001	<0.001	1.44	1.33, 1.57	<0.001	<0.001
Teschendorff	3.91	3.46, 4.42	<0.001	<0.001	2.40	2.16, 2.66	<0.001	<0.001	1.58	1.46, 1.72	<0.001	<0.001
Yu	22.1	17.4, 28.7	<0.001	<0.001	5.89	5.13, 6.80	<0.001	<0.001	3.44	2.99, 3.97	<0.001	<0.001
Gao	11.9	9.90, 14.6	<0.001	<0.001	4.21	3.74, 4.77	<0.001	<0.001	2.68	2.38, 3.03	<0.001	<0.001
Yang	18.3	14.6, 23.3	<0.001	<0.001	5.36	4.69, 6.16	<0.001	<0.001	2.57	2.28, 2.91	<0.001	<0.001
Zhang	11.7	9.85, 14.2	<0.001	<0.001	3.92	3.56, 4.33	<0.001	<0.001	2.99	2.53, 3.60	<0.001	<0.001

Table 3.5: Area under the curve (AUC) values distinguishing between self-reported smoking status classes for each of 16 DNAme-based biomarkers of smoking where cell colour represents the strength of classification between smoking status from least (dark green) to most (dark red) strong

	Never vs Current (N = 2,447/1,848)			Former vs Current (N = 2,366/1,802)			Never vs Former (N = 3,165/2,310)		
Method	**NCDS**	**USM1**	**USM2**	**NCDS**	**USM1**	**USM2**	**NCDS**	**USM1**	**USM2**
AHRR	0.970	0.998	0.968	0.904	0.969	0.905	0.733	0.694	0.696
smokp SSt	0.915	-	0.910	0.860	-	0.840	0.564	-	0.670
smokp Packyears	0.879	-	0.840	0.769	-	0.677	0.695	-	0.707
smokp Cessation	0.967	-	0.957	0.920	-	0.922	0.662	-	0.568
EpiSmokEr SSt	0.935	0.486	0.491	0.824	0.506	0.497	0.646	0.480	0.494
EpiSmokEr SSc	0.961	0.474	0.520	0.898	0.484	0.503	0.692	0.490	0.517
EpiSmokEr MS	0.964	0.501	0.506	0.884	0.481	0.483	0.680	0.519	0.523
Sugden	0.940	0.963	0.927	0.878	0.888	0.851	0.674	0.666	0.646
McCartney	0.972	0.999	0.982	0.910	0.978	0.916	0.786	0.728	0.727
Christiansen	0.912	0.990	0.959	0.852	0.941	0.893	0.645	0.652	0.665
Odintsova	0.889	0.875	0.827	0.803	0.822	0.765	0.645	0.574	0.579
Teschendorff	0.769	0.785	0.835	0.691	0.726	0.736	0.591	0.565	0.635
Yu	0.937	0.978	0.948	0.883	0.914	0.877	0.695	0.701	0.678
Gao	0.918	0.956	0.919	0.862	0.869	0.836	0.615	0.702	0.673
Yang	0.947	0.964	0.938	0.885	0.891	0.867	0.674	0.661	0.657
Zhang	0.825	0.976	0.915	0.795	0.883	0.834	0.531	0.592	0.581

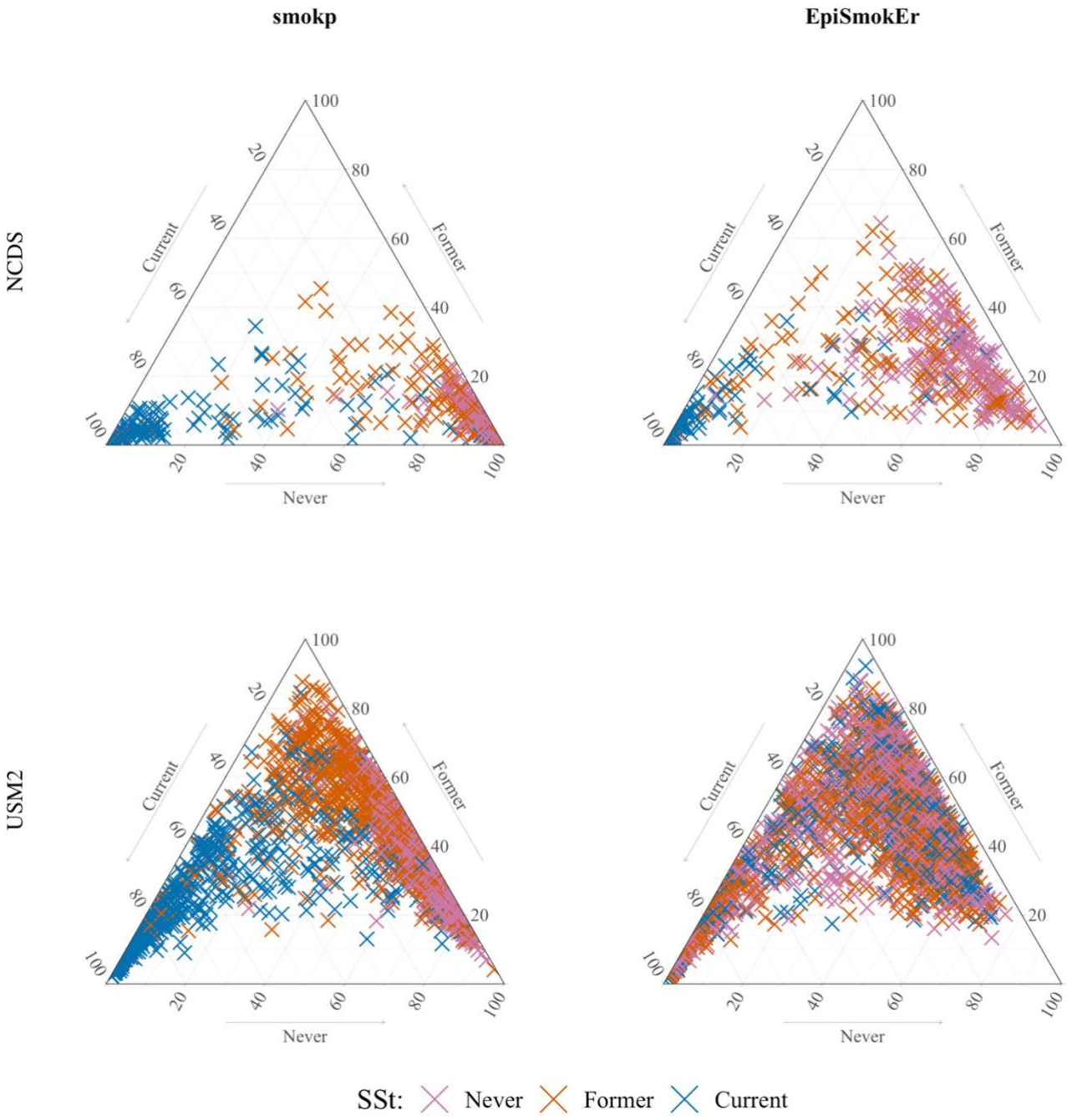

Figure 3.1: Ternary plots showing methylation-based predicted probabilities for each smoking status, including smokp (Left) and EpiSmokEr (Right), coloured by self-reported smoking status (Blue = Current, Orange = Former, Pink = Never)

3.3.5. Prediction of pack years and cessation

Simple linear regression models were fitted with either self-reported pack years or years since quitting as the dependent variable and methylation-based smoking estimates from each of the 16 biomarkers separately as the independent variable (Table 3.5). Across all three datasets, there was a statistically significant effect of the methylation-based biomarker values on self-reported pack years when using 8 out of 16 included biomarkers. This included smokp SSt (Never vs Current: Beta = 17, CI = 14-20; Never vs Former: Beta = 5.9, CI = 3.0-8.8; $p < 0.001$), smokp Packyears (Beta = 11, CI = 10-12, $p < 0.001$), smokp Cessation (Beta = 0.42, CI = 3.1-4.9, $p < 0.001$), Sugden et al (Beta = 1.5, CI = 0.5-2.5, $p = 0.003$), Teschendorff (Beta = 3.1, CI = 2.0-4.1, $p < 0.001$), Yu et al (Beta = 1.3, CI = 0.35-2.3, $p = 0.008$), Gao et al (Beta = 3.0, CI = 2.0-4.0, $p < 0.001$), and Yang et al (Beta = 1.3, CI = 0.34-2.3, $p = 0.009$). The Sugden et al (2018), Yu et al (2020) and Yang et al (2019) biomarker values were less significantly associated with self-reported pack years after false discovery rate (FDR) correction for multiple testing (Table 3.5).

In terms of cessation years, a statistically significant effect of the methylation-based biomarker values on self-reported years since quitting was shown for all but one biomarker, the Gao et al (2016) method (Beta = -0.62, CI = -1.6-0.36; $p = 0.2$). The other 15 biomarkers that did significantly associate with cessation years includes using *AHRR* CpG methylation (Beta = 5.7, CI = 4.6-6.8, $p < 0.001$), smokp SSt (Never vs Current: Beta = -11, CI = -15--7.4; Never vs Former: Beta = 0.85, CI = -1.1, 2.7; $p < 0.001$), smokp Packyears (Beta = 1.6, CI = 0.81-2.5, $p < 0.001$), smokp Cessation (Beta = 12, CI = 11-13, $p < 0.001$), EpiSmokEr SSt (Never vs Current: Beta = 0.45, CI = -1.5, 2.4; Never vs Former: Beta = -3.1, CI = -5.3, -0.9; $p = 0.001$), EpiSmokEr SSc (Beta = -0.84, CI = -1.6—0.03, 10, $p = 0.043$), EpiSmokEr MS (Beta = -1.1, CI = -1.9—0.26, $p = 0.01$), Sugden (Beta = -2, CI =, -3--1, $p < 0.001$), McCartney (Beta = -7.4, CI = -8.5-6.3, $p < 0.001$), Christiansen (Beta = 5.2, CI = 4.2-6.2, 10, $p < 0.001$), Odintsova (Beta = -3.8, CI = -

4.6--2.9, p < 0.001), Teschendorff (Beta = 1.2, CI = 0.31-2.1, p = 0.008), Yu (Beta = -3.3, CI = -4.3—2.2, p < 0.001), Yang (Beta = -2.5, CI = -3.5--1.5, p < 0.001) and Zhang (Beta = -3.7, CI = -4.5--2.9, p < 0.001). Only one biomarker, the Gao et al method (Beta = -0.62, CI = -1.6-0.36, p < 0.001), was not significantly associated with cessation years.

Adjusted R-squared (R^2) values were obtained from these models and used to measure how well each methylation-based smoking biomarker reflects pack years and cessation years within each of the three test datasets (Table 3.6). The amount of variance in pack years explained by the tested methylation-based biomarkers of smoking, across all three test datasets within current smokers ranged from -0.008 to 0.327 suggesting up to a third of the variance seen in pack years may be estimated using DNA methylation. Information on pack years was only collected from current smokers in NCDS, however in the two Understanding Society datasets, USM1 and USM2, pack years could also be derived in former smokers. In current smokers the biomarker that explained the most amount of variance across all datasets was the smokp Packyears method which showed an average adjusted R^2 of 0.176. However, a much greater variance in self-reported pack years is explained by this biomarker within current smokers in USM2 (R^2 = 0.327) compared to NCDS (R^2 = 0.024). The biomarker that explained the second most amount of variance in self-reported pack years was the method specified by Sugden et al (2018). This showed an average adjusted R^2 of 0.109 however this was again greater in USM2 (R^2 = 0.190) compared to NCDS (R^2 = 0.047) and USM1 (R^2 = 0.090). In NCDS the method that explained the most amount of variance in self-reported pack years was the EpiSmokEr SSc (R^2 = 0.055).

In former smokers the amount of variance in self-reported pack years explained by the 16 tested methylation-based biomarkers ranged from -0.003 to 0.281 and in ever smokers adjusted R-squared values ranged from -0.002 to 0.202. The biomarker that explained the greatest amount of variance in self-reported

pack years in former smokers was again the smokp Packyears method which showed an adjusted R^2 of 0.281 in USM2, followed by the smokp SSt method ($R^2 = 0.113$). The Gao et al (2016) method showed the third greatest variance in self-reported pack years on average ($R^2 = 0.079$) across both Understanding Society datasets and this was slightly greater in USM2 ($R^2 = 0.087$) compared to USM1 ($R^2 = 0.071$). However, the biomarker that explained the most amount of variance in self-reported pack years reported by former smokers in USM1 was AHRR CpG methylation showing an adjusted R^2 of 0.089. The biomarker that explained the greatest amount of variance in self-reported pack years in ever smokers was again the smokp Packyears method which showed an adjusted R^2 of 0.202, followed by the smokp SSt method ($R^2 = 0.092$), followed by the smokp Cessation method ($R^2 = 0.040$) in USM2. In USM1 the biomarker that explained the greatest amount of variance in self-reported pack years in ever smokers was the Gao et al (2016) method (R^2 of 0.019) followed by the Sugden et al (2018) method ($R^2 = 0.005$).

In terms of cessation years, the amount of variance (R^2) in self-reported measures explained by the biomarkers in all three datasets ranged from -0.009 to 0.345. In NCDS the biomarker that explained the most variance in self-reported cessation years was EpiSmokEr SSt ($R^2 = 0.31$). This was followed by McCartney et al (2018) ($R^2 = 0.22$). In USM2 the best biomarker of smoking cessation years was the smokp Cessation method showing an adjusted R^2 of 0.35. In USM1 the smokp biomarkers could not be tested however out of the remaining methods the biomarker explaining the largest proportion of variation in cessation years was McCartney et al (2018) which showed an adjusted R^2 of 0.12. In NCDS and USM2 together the biomarker that explained the greatest amount of variance in self-reported cessation years was the smokp Cessation method which showed an average R^2 of 0.25. A greater amount of variance in self-reported cessation was explained by this biomarker in USM2 ($R^2 = 0.35$) compared to NCDS ($R^2 = 0.16$). The biomarker that explained the second largest proportion of variance in cessation was McCartney et al (2018) with an average R^2 of 0.16. This biomarker explained a larger proportion of variance in cessation years reported in NCDS ($R^2 = 0.22$) compared to USM1 ($R^2 = 0.12$) and USM2 ($R^2 = 0.15$).

Table 3.6: Simple linear regression output showing the relationship between each of 16 DNAme-based biomarkers of smoking with self-reported smoking histories to estimate the probability that each biomarker can predict pack years (left) and cessation years (right)

	Pack years (N = 1,818/1,287)				Cessation years (N = 1,156/794)			
Method	**Beta**	**95% CI**	**p-value**	**q-value**	**Beta**	**95% CI**	**p-value**	**q-value**
AHRR	-0.11	-1.1, 0.86	0.8	0.9	5.7	4.6, 6.8	<0.001	<0.001
smokp SSt			<0.001	<0.001			<0.001	<0.001
Never	—	—			—	—		
Former	17	14, 20			0.85	-1.1, 2.7		
Current	5.9	3.0, 8.8			-11	-15, -7.4		
smokp Packyears	11	10, 12	<0.001	<0.001	1.6	0.81, 2.5	<0.001	<0.001
smokp Cessation	4.0	3.1, 4.9	<0.001	<0.001	12	11, 13	<0.001	<0.001
EpiSmokEr SSt			0.11	0.2			0.001	0.002
Never	—	—			—	—		
Former	1.8	-0.93, 4.5			0.45	-1.5, 2.4		
Current	-0.81	-3.8, 2.2			-3.1	-5.3, -0.90		
EpiSmokEr SSc	-0.55	-1.6, 0.49	0.3	0.4	-0.84	-1.6, -0.03	0.043	0.045
EpiSmokEr MS	-0.52	-1.6, 0.52	0.3	0.4	-1.1	-1.9, -0.26	0.010	0.012
Sugden	1.5	0.50, 2.5	0.003	0.009	-2.0	-3.0, -1.0	<0.001	<0.001
McCartney	-0.15	-1.1, 0.85	0.8	0.9	-7.4	-8.5, -6.3	<0.001	<0.001
Christiansen	0.08	-0.90, 1.1	0.9	0.9	5.2	4.2, 6.2	<0.001	<0.001
Odintsova	-0.88	-1.9, 0.16	0.10	0.2	-3.8	-4.6, -2.9	<0.001	<0.001
Teschendorff	3.1	2.0, 4.1	<0.001	<0.001	1.2	0.31, 2.1	0.008	0.010
Yu	1.3	0.35, 2.3	0.008	0.018	-3.3	-4.3, -2.2	<0.001	<0.001
Gao	3.0	2.0, 4.0	<0.001	<0.001	-0.62	-1.6, 0.36	0.2	0.2
Yang	1.3	0.34, 2.3	0.009	0.018	-2.5	-3.5, -1.5	<0.001	<0.001
Zhang	0.42	-0.51, 1.4	0.4	0.5	-3.7	-4.5, -2.9	<0.001	<0.001

Table 3.7: Adjusted R squared (R2) values indicating the proportion of variance each methylation-based biomarker of smoking explains in self-reported smoking histories, including pack years (left) and cessation years (right) from least (dark green) to most (dark red) strong

	Pack years (N = 1,818/1,287)							Cessation years (N = 1,156/794)		
	NCDS	USM1			USM2			NCDS	USM1	USM2
Method	Current	Current	Former	Ever	Current	Former	Ever	Former	Former	Former
AHRR	0.039	0.007	0.089	-0.002	0.156	0.032	-0.001	0.178	0.05	0.098
smokp SSt	0.013	-	-	-	0.14	0.113	0.092	0.136	-	0.035
smokp Packyears	0.024	-	-	-	0.327	0.281	0.202	0.122	-	0.013
smokp Cessation	0.053	-	-	-	0.024	0.019	0.0z4	0.163	-	0.345
EpiSmokEr SSt	-0.007	-0.008	0.001	0.001	-0.004	0.006	0.001	0.312	-0.004	0.009
EpiSmokEr SSc	0.055	0.005	-0.003	-0.002	-0.002	0.001	0	0.168	-0.003	0.009
EpiSmokEr MS	0.038	0.005	-0.002	0	-0.002	0.005	0.001	0.146	-0.002	0.014
Sugden	0.047	0.09	0.048	0.005	0.19	0.031	0.004	0.116	0.005	0.009
McCartney	0.052	0.008	0.076	-0.002	0.171	0.029	-0.001	0.215	0.115	0.152
Christiansen	0.018	0.016	0.046	-0.002	0.135	0.016	-0.001	0.179	0.042	0.104
Odintsova	0.038	-0.006	0	0.003	0.022	0.002	0	0.04	0.043	0.065
Teschendorff	0.017	0	0.001	0	0.189	0.104	0.053	-0.009	-0.003	0.027
Yu	0.04	0.054	0.055	0.003	0.192	0.046	0.003	0.188	0.016	0.035
Gao	0.002	0.068	0.071	0.019	0.227	0.087	0.021	0.153	-0.003	0
Yang	0.037	0.042	0.047	0.003	0.177	0.041	0.003	0.18	0.007	0.022
Zhang	0.003	0.002	0.066	0.002	0.092	0.027	-0.001	0.074	0.044	0.084

3.3.6. Bias and ageing

Bland-Altman plots were used to visualise and assess the agreement between DNA methylation-based estimates from the smokp Packyears and Cessation biomarkers with self-reported pack years and cessation years respectively (Figure 3.2). This was carried out in NCDS and USM2. In both studies, the mean difference between self-reported pack years and methylation-based predictions using the smokp Packyears method measured -6.49 (± 20.35) and the median measured -2.66, with errors ranging from -107.56 to 41.83. In USM2 the mean difference between DNAme-predicted and self-reported pack years was -5.71 (± 20.56), and the median difference was -1.54 with errors in packyears in USM2 ranging from -107.56 to 41.83. In NCDS the mean difference between DNAme-predicted and self-reported pack years was -13.61 (± 16.74) and the median was -12.20, ranging from -68.14 to 19.50. The mean difference between self-reported cessation years and methylation-based estimates, using the smokp Cessation method, measured -17.42 (± 11.27), the median measured -16.14 and errors ranged from -55.89 to 6.23. The negative mean differences observed suggests a systematic underestimation of smoking histories in DNA methylation-based estimates compared to self-reports and this was more pronounced for cessation years than pack years. In USM2 the mean difference between DNAme-predicted and self-reported cessation years was -16.73 (± 11.51), the median difference was -15.02 and errors in packyears in USM2 ranged from -55.89 to 6.23. In NCDS the mean difference between DNAme-predicted and self-reported pack years was -21.59 (± 8.63) and the median was -21.23, ranging from -43.58 to -1.81.

A strong and significant negative proportional bias is observed in USM2 between self-reported and DNAme-derived pack years whereby the effect of increasing pack years on errors between DNAme-based and self-reported pack years measures is statistically significant and negative (beta = -0.92, 95% CI [-0.98, -0.86], $t(1158) = -30.46$, $p < .001$). In NCDS a positive proportional bias is observed instead between self-

reports and DNAme-derived pack years estimates (beta = 0.31, 95% CI [0.02, 0.60], t(125) = 2.08, p = 0.039) and this bias is less significant compared to the negative bias observed in USM2. A statistically significant and negative proportional bias in USM2 was also observed in relation to cessation years (beta = -0.87, 95% CI [-0.93, -0.80], t(679) = -26.27, p < .001). A positive bias with increasing years since quitting was observed in NCDS (beta = 0.27, 95% CI [0.03, 0.51], t(111) = 2.21, p = 0.029).

NCDS is a birth cohort where all participants were aged 44 when bloods were collected however Understanding Society represent a large age range representative of the UK adult population. The stronger bias and different direction of association in USM2 compared to NCDS suggests age may play a role. To explore this, simple linear regressions were used to understand how much of the error between self-reports and methylation-based estimates of smoking histories vary with age in USM2. Age explained a statistically significant proportion of variance in errors between self-reported and DNAme-derived pack years (R^2 = 0.11, F(1, 1158) = 137.76, p < .001) where the effect of age is significant and negative (Beta = -0.45, 95% CI [-0.53, -0.38], t(1158) = -11.74, p < .001). Age explained a weaker proportion of errors between self-reported and DNAme-predicted pack years in current smokers (R^2 = 9.82e-03, F(1, 479) = 4.75, p = 0.030), compared to former smokers (R^2 = 0.11, F(1, 677) = 83.93, p < .001).

Age also explained a statistically significant and moderate proportion of variance in errors between self-reported and DNAme cessation years (R^2 = 0.21, F(1, 792) = 205.11, p < .001) and the effect of age was also negative (Beta = -0.37, 95% CI [-0.42, -0.32], t(792) = -14.32, p < .001). This may suggest age contributes to errors between self-reported and predicted smoking histories. Age did not however fully explain the bias observed in USM2 and cannot explain the positive bias observed in NCDS suggesting other factors play a role.

353 CpG sites were used in the methylation-based predictor of age by Horvath et al (2012). 10 of these sites were also utilised in at least one methylation-based biomarker of smoking tested in this chapter. In smoking indexes outlined by Teschendorff et al (2015) 5 Horvath CpGs were used, 1 by Yu et al (2020), and 1 by Yang et al (2019). Out of the methylation scores tested, Horvath age-related CpGs were only found in the Sugden et al (2019) method. One site (cg25809905) was used in 3 different biomarkers and is located on chromosome 17 (42,467,728bp) in the *ITGA2B* gene and another (cg22947000) was used in 2 biomarkers and is located on chromosome 16 (81,272,281bp) in the *BCMO1* gene. Other genes implicated include *KIAA1199*, *ERG*, *AKT3*, *PRKG2*, *ACOT11*, *MPI*, *PGLYRP2*, and *C10orf99*. *ITGA2B*, *AKT3* and *PRKG2* are involved in platelet activation.

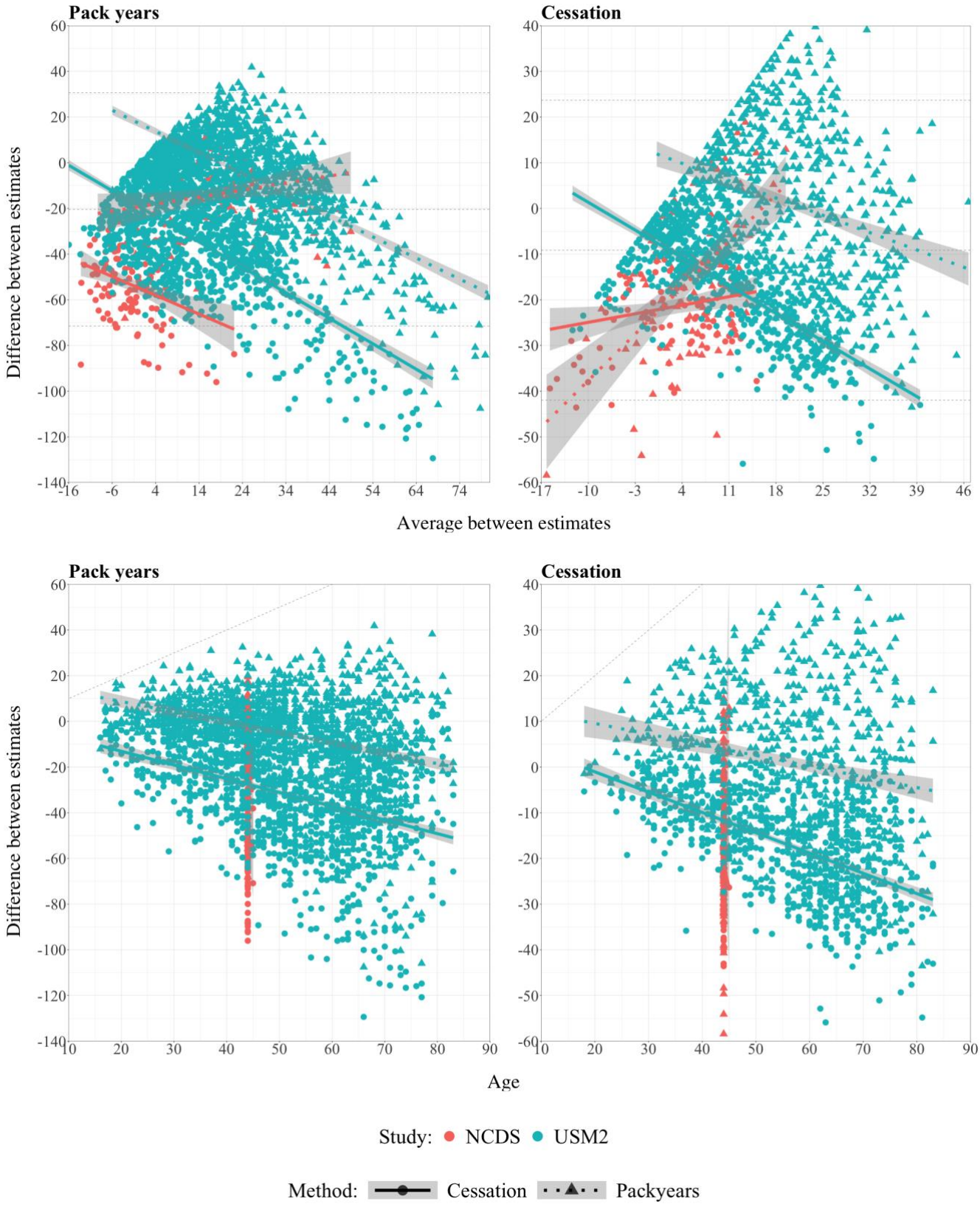

Figure 3.2: Differences between self-reported and unstandardised methylation-based estimates of smoking histories including pack years (Left) and cessation years (Right) by mean averages of self-reported and methylation-based estimates (Top) and age (Bottom)

3.4. Discussion

This study systematically compared 16 different DNA methylation-based biomarkers used to estimate smoking. Each biomarker of smoking was used to measure smoking status (Current, Former, Ever), pack years, and years since quitting smoking. The number and location of loci used in predicting smoking from DNA methylation varied greatly depending on the biomarker being used however the addition of more CpG sites in combination with a single CpG site in the AHRR locus did not greatly improve predictions of self-reported smoking. DNAme-based classification of smoking status was most accurate in current smokers and poorest in former smokers. This reflects the decay of smoking driven changes to DNA methylation upon cessation and could also indicate false negative reporting of smoking behaviours. DNA methylation at the AHRR locus for example was much closer to never smoker levels than current smokers. This suggests an acquisition of DNA methylation changes with increasing pack years that then decay with cessation. Overall, the best biomarkers of smoking were the methylation score from McCartney et al (2018), *AHRR* CpG methylation, and the smokp methods (SSt and Cessation) trained in the USM1 dataset. The use of repeated measures of smoking data while training biomarkers to predict smoking from DNA methylation suggests estimates of smoking status offered by the smokp SSt method may be more reliable compared to other methylation-based biomarkers based on self-reports taken at one time point. This is important given it has been noted that misreports and longitudinal changes provide a source of bias in GWAS experiments of smoking using the UK Biobank as well as other self-reported health behaviours (Xue et al., 2020). A reason for the poor performance of the EpiSmokEr SSt method in UKHLS while it performed well in NCDS may relate to how the DNA methylation data was normalised. In UKHLS DNA methylation was normalised by using methylated and unmethylated probe intensities separately however only beta-values rather than methylated and unmethylated signal intensities were available in NCDS.

DNA methylation-based estimates of smoking pack years and cessation years were highly correlated with age making it difficult to estimate smoking histories outside of changes driven by the ageing process. Average age of smoking initiation was 16.90 (± 4.03) years meaning most participants start smoking in their late teens with very few people starting to smoke past the age of 20. Both age and years spent smoking or quitting will then likely reflect similar variance in DNA methylation. Previously it has been shown that self-reported cigarette consumption effects on intrinsic accelerated DNA methylation-based aging indices may be fully mediated by DNA methylation-based indicators of smoking (Lei et al., 2020). Epigenetic clocks of aging are also age-dependant and the rate of aging-related changes to DNA methylation appears to slow down as some CpG sites reach either full methylation or complete demethylation (El Khoury et al., 2019). A similar process may occur in DNA methylation-based smoking estimates whereby the extent of changes driven by smoking reach saturation. A negative proportional error can be observed across self-reported smoking histories in USM2 but not in NCDS where instead a weaker positive bias occurs. As errors between estimates were not fully explained by age perhaps saturation effects could in part explain why proportional bias in methylation-based estimates of smoking histories occur.

Other studies have made use of pack years as a measure of cumulative exposure to tobacco-related substances and have identified a dose response relationship between tobacco use and DNA methylation (Zhang et al., 2015). Many genes differentially methylated in current smokers relative to never smokers are also significantly associated with duration of smoking (Ambatipudi et al., 2016). Some loci remain differentially methylated even after years since quitting. It is thought that the ratio of lung cancer incidence between current and former smokers increases sharply with time since quitting however the degree of this reduced risk may be overestimated in many studies where risk is calculated by dividing the nearly constant smoker risk rate by the ever-increasing non-smoker rate which varies with age (Peto et al., 2000). When smoking ceases the rate of lung cancer incidence does stop increasing steeply but may still increase with age where risk is often higher in the oldest people (Peto et al., 2011). Pack years and cessation years offer

additional information on smoking behaviours and can help better understand differences in health observed within the same smoking categories and lead to better understanding of smoking aetiology. In this study only a few biomarkers were able to significantly reflect pack years again suggesting that changes to DNA methylation by smoking may reach saturation where little variation occurs within smokers. Most biomarkers of smoking significantly associated with cessation years strengthening the idea that smoking-related changes to DNA methylation decay once quitting.

A huge surplus of hypomethylated over hypermethylated differentially methylated CpGs is noted in relation to smoking. This can perhaps be understood as the activation of biological 'clean up' systems such as is the case for *AHRR*. This gene codes for the aryl hydrocarbon receptor repressor that competes with the aryl hydrocarbon receptor nuclear translocator (*ARNT*) to prevent signal transduction of harmful polycyclic hydrocarbons (PAHs) that tobacco combustion generates (Evans et al., 2008). However, reduced global methylation has also been observed to occur with age (Xiao et al., 2019). Many other genes not necessarily related to xenobiotic toxin responses were also differentially methylated with smoking thus implicating any number of biological pathways in driving DNA methylation changes. Cigarette smoke itself can lead to DNA hypomethylation in several ways. Nicotine when bound to nicotinic acetylcholine receptors activates the cAMP response element-binding protein which has been demonstrated to downregulate DNMT enzymes that catalyse the DNA methylation process (Satta et al., 2008). Smoking may also impact DNA-binding factors that prevent de novo methylation of some CpG sites (Han et al., 2001). Many loci associated with smoking were found in non-coding, intergenic regions. Changes to the epigenetic regulation of such regions play a large role in the aberrant nature of methylome changes and can lead to chromatin instability (Ehrlich, 2008). Global genome-wide hypomethylation is therefore one of the earliest molecular abnormalities seen in cancer and has been described following carcinogen exposure (Lisanti et al., 2013).

DNA methylation and histone modifications are still poorly understood in their response to environmental stressors, and this is further complicated by their interactive nature at the systems level. Although biomarkers should represent objective indicators of normal biological processes that may be measured accurately and reproducibly (Strimbu et al., 2010) these biological processes could in theory respond to any number of environmental stimuli. This idea is strengthened by the interactions observed between DNAme estimates of smoking with age. The value and quality of biomarkers is then dependant on the measurement error of the characteristic being examined and less so to understanding the mechanism of action leading to said biomarker. Instead, better phenotyping of smoking behaviours may improve subject classification as misclassification severely impacts statistical power and this was noted regarding smoking over fifty years ago (Mote and Anderson, 1965).

A further point of contention within findings from EWAS studies of smoking was the persistence of DNA methylation differences following many years of cessation (Guida et al., 2015) and how this may be best utilised as a sensitive and long-term biomarker of tobacco use. Normally studies investigating DNA methylation changes with smoking would correct for cellular heterogeneity (Houseman, 2015) but this may be counterintuitive in the identification of biomarkers as it has been shown to consume degrees of freedom and leads to loss in the statistical power needed to detect meaningful results without necessarily improving findings in some studies (Dogan et al., 2014) thus these were not adjusted for.

3.5. Conclusion

Epigenetic biomarkers were created in this study using repeated measures of smoking data to enable more reliable measures of smoking to train on. In total 16 biomarkers for estimating smoking from DNA methylation were investigated. All methods appear to be able to distinguish current smoking from never

smoking however the prediction of more complex phenotypes such as past smoking, pack years and cessation are more difficult. Errors between estimates are also highly correlated to age. Overall, the smokp and McCartney et al (2018) methods, as well as *AHRR* CpG methylation alone, offer the best estimation of smoking behaviour from DNA methylation. Studies hoping to exploit the potential of epigenetic biomarkers to better explain health outcomes should carefully consider factors that may influence the reliability of methylation-based biomarkers of smoking.

4. **Discrepancies between self-reported and DNAme-predicted smoking**

4.1. Introduction

Smoking leads to a multitude of diseases and is as a major public health concern and focus of public policy, especially since the first international treaty, the WHO Framework Convention on Tobacco Control, which came into force in 2005. Often self-reports are used to classify individuals by their smoking status where participants are asked to report if they have ever smoked and if they smoke now. Self-reports of smoking are accurate in most studies however both the sensitivity, or true positive rate, and the specificity, or true negative rate, can vary greatly between different populations, different interviewer-administered questionnaires, and different observational studies (Patrick et al., 1994). To overcome these issues, it has been suggested that biochemical assessment of smoking may be used to increase the reliability of self-reported smoking data. The most common biochemical assessment of smoking is cotinine, an alkaloid found in tobacco and the most predominant metabolite of nicotine. Cotinine has been frequently used as a biomarker of smoking. A biomarker is a naturally occurring molecule, gene, or characteristic by which a a certain physiological process can be identified. Biomarkers can offer an objective measure of smoking and by using biomarkers of smoking it has been suggested that self-reports may underestimate the true prevalence of smoking in some populations, particularly in studies of cessation and of adolescent smoking (Stookey et al., 1987).

To investigate factors that may influence misclassifications of smoking some studies have investigated discrepancies between self-reported and biologic determination of smoking using serum cotinine. Investigated factors include ethnicity, sex, age, education, past smoking behaviour, smoking intensity and number of household members who smoke. In the CARDIA study larger discrepancies in those with a high

school education or less compared to those with more education, in ex-smokers compared to those reporting never smoking, and in participants who reported spending more time with smokers (Wagenknecht et al., 1992). In the third National Health And Nutrition Examination Survey (NHANES III) self-reported smokers who were misclassified as non-smokers reported smoking fewer daily cigarettes compared to smokers accurately classified using cotinine. In non-smokers participants were more likely to show discrepant findings among persons who reported two or more smokers living in the home compared to those who reported no smokers living in the home. Discrepancies in non-smokers were less likely among participants with ≥12 years of education than among participants with fewer years. In self-reported smokers, younger participants were more likely than persons aged ≥65 years, and self-identified black participants were less likely than white participants to be in discrepancy with cotinine measures. The average number of cigarettes smoked per day in the past 5 days was inversely and highly associated with the probability of discrepancy. Lastly participants who self-reported as ever smokers and who reported not smoking in the previous 5 days were more likely to be in discrepancy than those reported as never smokers (Caraballo et al., 2001). This shows that socioeconomic factors such as education, as well as age, ethnicity, and previous smoking behaviours can all influence the agreement between biochemical assessments of smoking and self-reports.

Although cotinine can offer a more objective measure of smoking than simply using self-reports, a major issue in using cotinine is that in vivo cotinine has a half-life of approximately 20 hours and can only remain detectable after a maximum of several days after tobacco use. Systematic differences in cotinine levels have also been attributed to variation in CYP2A6 activity, a member of the cytochrome P450 mixed-function oxidase system, which is involved in the metabolism of xenobiotics in the body, and this may result in substantially different cotinine levels between individuals given the same tobacco exposure. In most smokers, cytochrome P450 2A6 (CYP2A6) is the primary enzyme responsible for nicotine metabolism. Individuals carrying inactive CYP2A6 alleles show decreased nicotine metabolism and are less likely to

become smokers and if they do, they smoke fewer cigarettes per day (Pianezza et al., 1998). This has implications in steering smoking behaviour but may also influences drug toxicities and the risk of developing several cancers (Hosono et al., 2017). This information may be utilised to better understand smoking pathology and the integration of genetic, DNAme and metabolomic data could lead to more accurate disease risk stratification.

To overcome this, recently strong differences in DNA methylation have been observed with smoking that may even be able to reflect smoking more than 35 years after smoking cessation (Guida et al., 2015). This has led to several epigenetic biomarkers of smoking that all utilise changes to DNA methylation associated with smoking. DNA methylation is one of several epigenetic processes that controls the architecture of the genome and constitutes the memory of the cell when epigenetic modifications are inherited from the cell from which it descends. This process can influence gene expression without any alterations to the genetic sequence itself and is able to reflect many environmental exposures. DNAme is now heralded as one of the many ways in which our social environment may impact our health and disease risk. DNA methylation-based biomarkers of smoking appear to be highly sensitive and accurate however little work has investigated how socioeconomic factors may influence the agreement between self-reports and epigenetic biomarkers of smoking. The first aim of this chapter is to investigate if age, sex, self-reported smoking, educational attainment and socioeconomic classification are significantly associated with discrepancies between smoking measures using self-reports or DNA methylation-based estimates.

Smoking has long been known to be socially patterned and it is now increasingly seen as socially unacceptable in the modern western world. People in professional occupations tend to be the first socioeconomic groups to quit smoking as it become less socially acceptable or desirable. However, in earlier years at the start of the 20th century smoking was in a way a marker of affluence and remains so in

some areas of the world today. In the UK and most of the western world a “social inversion” has occurred whereby smoking is now associated with social disadvantage. Today the average smoker smokes more than in previous years and is likely to have less money, fewer educational qualifications and work in less prestigious jobs than the average non-smoker. In recent years only a small percentage of medical professionals’ report smoking whereas during World War II as many as 80% of doctors reported smoking while now many disadvantaged groups such as Native Americans show greater smoking rates instead (Houston, 1986). Modern day public health discourse does draw attention to the unequal social distribution of smoking however critics also discuss the importance of agency and embodiment in smoking and how public health measures against smoking could play upon and exacerbate social divisions and inequality (Marron, 2017). Nevertheless, when discussing smoking it is apparent that the social context of tobacco use must also be considered.

Health behaviours such as smoking have a huge influence on individual health and the health of the public. It is also clear that health behaviours are strongly shaped by the socioeconomic environment. Many social and population-based studies now routinely carry out health assessments and perform a range of bio-medical measures through collecting blood samples. By doing so social studies can measure major illnesses as well as offer markers of key physiological systems. A multitude of studies have shown that socioeconomic factors influence many common biologic markers of health, and these findings remain significant even after adjustment for health behaviours like smoking. For example, educational and socioeconomic gradients in inflammation, the defence of the immune system and body from harmful agents, have previously been reported (Muscatell et al., 2020). Two commonly measured inflammatory markers are fibrinogen and C-reactive protein (CRP) which are both positive acute phase proteins meaning their plasma concentrations increase with increasing inflammation.

Systemic inflammation has been proposed as one physiological process linking socioeconomic position to health. Often participants with higher educational attainment and participants classified into more privileged socioeconomic groups exhibit lower levels of peripheral inflammation. This association is heterogenous across the life span and this in turn differs depending on the marker in question. Socioeconomic inequalities in CRP emerged in the 30s and gradually increased with age, peaking up to the late 50s and early 60s and then decreased with age thereafter. Socioeconomic inequalities with fibrinogen decreased with age. In this study body mass index (BMI), smoking, physical activity and healthy diet explained part but not all the socioeconomic inequalities observed in inflammation. Of these factors BMI seemed to attenuate the largest amount of this relationship (Davillas et al., 2017). Previous studies however have shown that in bivariate analyses, inflammatory proteins were inversely associated with both income and education but in multivariate regression models where potential confounders are adjusted for, only low income significantly predicted higher levels of inflammation. This suggests that the reason that higher education is linked to reduced peripheral inflammation is because it reduces the risk for low-income status, which is directly associated with reduced peripheral inflammation (Friedman and Herd, 2010). This study also showed that the association between income and CRP and fibrinogen may be completely mediated by interleukin 6 (IL-6), a pro-inflammatory cytokine and an anti-inflammatory myokine.

Smoking is known to be positively associated with inflammation. Smoking triggers an immunologic response to vascular injury, and this is associated with increased levels of inflammatory markers, such as C-reactive protein and white blood cell count. These markers also predict future cardiovascular events and may be important in driving atherosclerosis. Inflammatory markers show a dose-dependent and temporal relationship to not only smoking but also smoking cessation where the smoking-associated inflammatory response returned to normal within five years after smokers quit (Bakhru and Erlinger, 2005). This suggests vascular effects are reversible. Given that smoking is significantly associated with both education and socioeconomic position, as well as inflammation, it is important to consider the effect of smoking when

investigating socioeconomic gradients in health. The second aim of this chapter is to investigate how adjustment for smoking, using either self-reported or DNAme-predicted smoking status, or a smoking methylation score (McCartney et al., 2018), impacts the relationship between educational attainment or socioeconomic classification and fibrinogen or CRP.

4.2. Methods

4.2.1. Samples

Understanding Society (UKHLS) is an annual household-based panel study which started collecting information about the social, economic, and health status of its participants in 2009. Two methylation datasets from UKHLS are used. The USM1 analytic data set is drawn from the British Household Panel Survey (BHPS), which began in 1991 and in 2010 was incorporated into UKHLS at the start of its wave 2 (2010-2012) when information on smoking behaviour were collected. UKHLS collected additional biological information, including blood samples for genetic and epigenetic analysis at wave 2 (2019-2012) for the USM2 participants (www.understandingsociety.ac.uk). Participants were asked many questions related to their socioeconomic and demographic characteristics and smoking behaviour and were also asked if they had smoked that day during blood collection. USM2 mostly consists of individuals who were not part of BHPS but were instead selected from the General Population Survey (GPS) and consists of approximately 2,500 samples making it one of the largest single DNA methylation resources currently available.

The National Child Development Study (NCDS) initial sample consisted of all babies born in Great Britain in a single week in March 1958 and have had multiple follow-ups in childhood at 7, 11 and 16 years and in

adulthood at 23, 33, 42 and 45 years. This provides high quality prospective data on social, biological, physical, and psychological phenotypes at every sweep. Methylation profiles were obtained from DNA samples collected from 529 NCDS subjects at age 44-45, at the same time as intensive phenotyping during this biomedical follow-up which included measures of many biomarkers such as inflammatory markers (Power and Elliot, 2006).

NCDS used the Registrar General Social Class (RGSC) system to classify participants by socioeconomic position however this has since been superseded by the National Statistics Socio-economic Classification (NSSEC) which was used by UKHLS. To allow for comparison between studies these variables were recoded into three groups: 'Managerial and professional,' 'Intermediate,' and 'Routine'. This means there are some differences between the two studies in terms of how social status is classified. In NCDS the "Intermediate" category contained both manual and non-manual "skilled" workers and the 'Routine' category consisted of those in "partly skilled" or "unskilled" work. In UKHLS the "Routine" category also includes those who have never worked and long-term unemployed individuals. Highest educational qualification obtained was asked at age 33 in NCDS and coded in terms of CSE and O levels or equivalent. However, in UKHLS highest qualification derived in the questionnaire prior to blood collection was used and coded in terms of GCSEs and A-levels or equivalent. Participants who reported "Other" qualifications were not included.

4.2.2. Statistics

This chapter aimed to investigate factors that influence discrepancies between self-reported and DNA methylation-predicted smoking status. It also aimed to see how self-reports and DNAme-based predictions of smoking status compare in their association with inflammatory markers. Factors investigated include

age, sex, self-reported smoking status, educational attainment, socioeconomic classification and methylation-based estimates of cell type composition.

Logistic regression was used to investigate if the factors listed above influence agreement between self-reported and DNAme-predicted smoking. To understand the direction of differences between smoking measures, positive and negative congruence was also investigated separately. Positive congruence refers to either self-reported smokers who were identified as smokers or self-reported non-smokers who were classified as smokers from DNAme. Negative congruence looks at self-reported non-smokers who were classified congruently using DNAme or self-reported smokers who were incorrectly predicted as non-smokers.

Simple linear regression was used to investigate if the composition of any of six cell types (granulocytes, CD8T, CD4T, B cells, monocytes and natural killer) significantly impacted overall congruence. Multivariate linear regression was used to investigate the interplay between smoking and education in explaining measures of two inflammatory markers, fibrinogen and C-reactive protein, and compare adjustment of smoking using self-reports or DNAme-predicted estimates.

4.3. Results

4.3.1. Descriptive statistics

This chapter aims to investigate factors that may influence the congruence or agreement between self-reported and DNAme-based measures of smoking status. Smoking status is defined as one of three groups,

including current, former, and never smokers. The factors examined in this chapter include age, sex, self-reported smoking status, DNAme-predicted smoking status, McCartney methylation score (MS), educational attainment, socioeconomic class and cell type composition estimates (Houseman et al., 2012). The methylation score from McCartney et al (2018) provides an epigenetic estimate of smoking by taking the sum of DNA methylation measures at 233 CpG sites, weighted by their effect size in relation to smoking. A larger MS means a greater likelihood of smoking and 228 loci (98%) were available in this study. DNAme-predicted smoking status and McCartney MS measures were obtained using the 'smokp' R function, described previously, that uses different biomarkers to obtain estimates of smoking behaviours from DNA methylation beta matrices.

This study used data from two studies including 3,011 participants in total. The first study includes 531 samples from the National Child Development Study (NCDS), a birth cohort that started in 1958 where bloods were collected during the age 44 sweep in 2002-2004. This cohort were approximately 52-56 years old in 2010-2012. The second study (USM2) used includes 2,480 participants from the larger epigenetic subsample of the UK Household Longitudinal Study (UKHLS), or Understanding Society, a nationally representative household panel study. USM2 bloods were collected in 2010-2012 when participants were aged between 16 and 88. To disentangle the relationship of age with other covariates such as education and cohort effects, a subset of USM2 aged between 49 and 59 were also investigated throughout as these participants would have been born within 5 years of 1958 and thus would be part of the same or similar birth cohort as NCDS. Data from USM1, the smaller epigenetic subsample of Understanding Society, all participants of which were previously part of the British Household Panel Survey (BHPS), were used to create and train the biomarker of smoking used to predict smoking status from DNA methylation ('smokp SSt') and as such was not used as a testing dataset.

Table 4.1 describes the participant characteristics and estimated cell type composition estimates in NCDS, USM1 and USM2. 46% of USM2 across the full age range (16-88) were male and 44% of USM2 participants aged between 49-59 were male. 48% of NCDS participants were male and 98% were aged 44 when bloods were collected. The median age in USM2 (aged 16-88) measured 51 years. 21%, 40% and 39% of USM2 participants self-reported current, former or never smoking respectively. In NCDS a higher proportion of current smokers (32%) and lower proportion of former smokers (28%) was observed. 40% of NCDS reported never smoking. As for DNAme-predicted smoking status, 58% of participants in USM2 and 66% of NCDS participants were classified as never smokers, while 16% of USM2 and 33% of NCDS participants were classified as current smokers. Former smoking was underestimated in both studies where less than 1% (n = 3) of participants in NCDS, and 26% (N = 641) in USM2 were classified as former smokers via DNA methylation.

The term congruence is used here to refer to whether self-reported smoking status matches DNAme-predicted smoking status. Positive congruence refers to whether smokers were correctly identified as smokers while negative congruence refers to whether non-smokers were correctly classified as non-smokers. There was little difference between overall (65%), positive (67%) and negative (65%) congruence in USM2 however in NCDS positive congruence (87%) was greater than negative congruence (58%) and overall congruence measured 67%. This suggests that in NCDS smokers were correctly identified from DNAme at a greater rate than non-smokers. The proportion of never smokers was overestimated in DNAme-based predictions of smoking status in both NCDS and USM2 while former smoking was underestimated. A similar proportion of current smokers is observed in self-reports compared to DNAme-predicted smoking within NCDS however in USM2 current smoking appears to be underestimated.

In NCDS, USM1 and USM2 39% of participants had obtained a higher qualification or degree (or NVQ 4,5,6) and this included 42% in USM2, 32% in USM1, and 33% in NCDS. 13% of participants across all studies had no formal qualifications and this included 12% of USM2, 19% of USM1, and 9% of NCDS. While a greater proportion of NCDS (44%) participants had obtained a GCSE, CSE 2-5, or O levels (or NVQ 1,2) compared to USM1 (27%) and USM2 (25%), a greater proportion of USM1 (21%) and USM2 (21%) had obtained A levels or equivalent (NVQ3) compared to NCDS (14%). Three groups were used for socioeconomic classification using the Registrar General's Social Class in NCDS and the National Statistics Socio-economic classification (NS-SEC) in UKHLS. For comparison between the two studies these social classes were called 'Management & professional', 'Intermediate,' and 'Routine'. 45% of participants (44% in USM2, 42% in USM1, and 51% in NCDS) reportedly worked within management or had professional occupations and 30% of participants (33% in USM2, 30% in USM1, and 19% in NCDS) worked within routine occupations. In NCDS the 'Routine' category consists of partly skilled or "unskilled" jobs as classified in the old social class scheme and in USM2 the 'Routine' category also includes those who have never worked or are long term unemployed.

The cell type contributing the most to the composition of whole blood was granulocytes ranging from 57% in NCDS, 69% in USM1 and 59% in USM2. This was followed by CD4T cells measuring 18% in NCDS 12% in USM1, and 13% in USM2. A higher proportion of CD8T cells were observed in USM2 (17%) compared to NCDS (2%) and USM1 (7%). Conversely a higher proportion of natural killer cells were observed in NCDS (11%) compared to USM1 (4%) and USM2 (0%). Differences between studies in terms of proportions of B cells and monocytes were less pronounced but slightly higher in NCDS compared to USM1 and USM2 (Table 4.1).

Table 4.2 describes the participant characteristics and estimated cell type composition within NCDS and USM2 by self-reported smoking status. This shows that across the two studies congruence between self-reported and DNAme-predicted smoking status, estimated using the 'smokp SSt' biomarker, is greater in current (89%) smokers compared to never (72%) and former (38%) smokers. Fewer former smokers were classified correctly by smoking status via DNA methylation in NCDS (2%) compared to USM2 (43%). 40% of participants who stated never smoking were male. There were few differences in the proportion of male compared to female participants who self-reported current (48%) or former (51%) smoking. In USM2 the mean age of a former smoker (53.15 ± 15.03) was on average older than both current smokers (47.37 ± 14.39) and never smokers (49.41 ± 15.93). Current smokers were less likely to have achieved a higher qualification or degree (23%) compared to former (44%) and never (47%) smokers. Current smokers were also more likely to have obtained no formal qualifications (17%) compared to former (12%) and never (8%) smokers. Further, current smokers were less likely to have a managerial or professional occupation (35%) compared to former (46%) and never (50%) smokers. Current smokers were also more likely to report working routine occupations (40%) compared to former (31%) and never (24%) smokers. Generally educational and socioeconomic gradients in smoking were more pronounced in USM2 than NCDS. The mean methylation score (McCartney et al., 2018) measured 5.07 (± 0.86) in current smokers, 3.48 (0.62) in former smokers, and 3.07 (0.30) in never smokers across both datasets. Cell type composition estimates derived from DNA methylation did not significantly differ by smoking status.

Table 4.3 describes the participant characteristics per study by overall congruence between self-reported and DNAme-predicted smoking status. DNAme-predicted smoking status was more likely to agree with self-reported smoking status measures in current and never smokers compared to former smokers. In USM2 there was little difference in overall congruence between men and women however in NCDS a higher proportion of women's smoking status was correctly classified via DNA methylation (70%) compared to men (63%). In both studies a greater proportion of participants with no formal educational qualifications

(70%) showed congruent smoking measures compared to those achieving higher qualifications or a degree (63%). Conversely, in NCDS a smaller proportion of participants with no formal qualifications (60%) displayed congruence between self-reports and DNAme-predicted smoking compared to participants achieving higher qualifications or degrees (67%). With this said only 40 participants in NCDS had not achieved any qualifications. In the 49 to 59 years old USM2 subset there was a smaller difference in congruence between the most (67%) and least (70%) educated groups. In USM2 across the full age range (16-88) DNAme estimates were more likely to correctly classify participant's smoking status in those without any qualifications (72%) compared to those who had achieved higher qualifications or degrees (62%). Overall and in USM2 (aged 16-88) there does not appear to be a significant difference in the percentage of participants showing congruence between self-reported and DNAme-predicted smoking status by socioeconomic class. However, in USM2 (aged 49-59) a higher percentage of matches were observed in participants in routine (71%) occupations compared to those in intermediate (68%) and managerial or professional (66%) occupations. Also, in NCDS a higher percentage of matches were observed in participants in routine (69%) and intermediate (69%) occupations compared to those in managerial or professional (65%) occupations.

Table 4.4 describes the participant characteristics per study by the agreement between self-reported and DNAme-predicted smoking status in positive cases, meaning either self-reported smokers who were identified as smokers in DNAme estimates or self-reported non-smokers who were identified as smokers in DNAme estimates. Within both studies 72% of positive cases were correctly identified as smokers while the remaining false positive cases consisted of 19% never smokers and 9% former smokers. In USM2 67% were correctly identified as self-reported smokers, 25% were former smokers and 8% were never smokers. In USM2 (aged 49-59) 71% of positive cases were self-reported current smokers while 9% were former smokers and 20% were never smokers. In NCDS 87% positive cases were correctly classified in self-reported smokers while 11% were former smokers and 2% were never smokers. In USM2 participants with

true positive cases were on average 11 years younger than participants with false positive cases and in USM2 (aged 49-59) true positive cases were on average 2.5 years younger. In USM2 smoking status in women (68%) compared to men (65%) were more likely to be classified correctly. This observed sex difference in positive congruence was more pronounced in USM2 limited to ages 49 to 59 (women = 69%, men = 75%), and in NCDS (women = 92%, men = 83%). Overall, there were fewer true positive cases among participants achieving higher qualifications (55%) compared to participants who had obtained A-levels or equivalent (77%), GCSES or equivalent (82%), and no formal qualification (71%). In USM2 participants achieving higher qualifications (48%) have a lower proportion of true positives compared to participants who had no formal qualification (69%), and this was more pronounced in USM2 subset to ages 49 to 59 and less pronounced in NCDS. Overall, there were fewer true positive cases among participants working in managerial or professional occupations (68%) compared to participants who work in intermediate occupations (74%) or carry out routine work (83%). Congruence between DNAme-predicted and self-reported smoking status in smokers also varied by socioeconomic class in NCDS and USM2 separately.

Table 4.5 describes the participant characteristics per study by the agreement between self-reported and DNAme-predicted smoking status in either self-reported non-smokers who were identified as non-smokers in DNAme estimates or self-reported smokers who were identified as non-smokers via DNAme. Overall, 64% of negative cases were correctly classified as non-smokers. The remaining negative cases consisted of 28% former smokers and 8% current smokers. Overall, 40% of former smokers were correctly identified as a non-smoker, including 44% in USM2 and 2.4% in NCDS. In USM2 participants with true negative cases were on average 6 years older than participants with false negative cases. In USM2 women (64%) were less likely to be correctly classified as non-smokers compared to men (66%). There was also a difference in negative congruence in USM2 limited to ages 49-59 between men (66%) and women (64%). Conversely, in non-smokers within NCDS, women (64%) were more likely than men (53%) to be correctly

classified. Across the two studies a smaller proportion of true negative cases occurred among participants achieving higher qualifications (64%) compared to participants who had achieved A-levels or equivalent (67%) or no qualification (70%) but not compared to those with GCSEs or equivalent (58%). This was also the case in the full USM2 (aged 16-88) dataset but not in USM2 limited to ages 49-59 years old. In USM2 (49-59) a smaller proportion of those with no qualifications (65%) showed true negative congruence compared to participants who had obtained a higher qualification or degree (71%). In terms of socioeconomic class, in NCDS the proportion of true negative cases observed in participants within managerial or professional (60%) occupations was greater compared to those in routine (54%) occupations. This was also true when comparing managerial or professional (65%) and routine (59%) occupations in USM2.

Table 4.1: Participant characteristics and Houseman cell type composition by study

Characteristic	NCDS	USM1	USM2	USM2 (49-59)
N	531	1,174	2,480	578
Male, n (%)	255 (48%)	489 (42%)	1134 (46%)	254 (44%)
Age, median (range)	44 (44-46)	59 (28-98)	51 (16-88)	54 (49-59)
Self-reported SSt, n (%)				
Never	200 (40%)	445 (44%)	978 (39%)	227 (39%)
Former	140 (28%)	410 (41%)	992 (40%)	220 (38%)
Current	162 (32%)	154 (15%)	508 (21%)	131 (23%)
DNAme-predicted SSt, n (%)				
Never	355 (67%)	546 (47%)	1445 (58%)	323 (56%)
Former	3 (0.6%)	460 (39%)	641 (26%)	138 (24%)
Current	173 (33%)	168 (14%)	394 (16%)	117 (20%)
Congruence, n (%)				
Overall	337 (67%)	971 (96%)	1618 (65%)	396 (69%)
Positive	137 (87%)	141 (95%)	348 (67%)	104 (72%)
Negative	200 (58%)	830 (96%)	1270 (65%)	292 (67%)
McCartney MS, median (range)	3.50 (2.56-7.45)	3.23 (2.40-6.87)	3.21 (2.42-6.39)	3.22 (2.56-6.34)
Educational attainment, n (%)				
Higher qualification	160 (33%)	341 (32%)	921 (42%)	220 (43%)
A-level/Equivalent	70 (14%)	223 (21%)	462 (21%)	92 (18%)
GCSE/Equivalent	215 (44%)	284 (27%)	545 (25%)	147 (29%)
No qualification	44 (9.0%)	203 (19%)	271 (12%)	50 (9.8%)
Socioeconomic classification, n (%)				
Management	241 (51%)	259 (42%)	662 (44%)	210 (46%)
Intermediate	140 (30%)	173 (28%)	352 (23%)	108 (24%)
Routine	92 (19%)	181 (30%)	503 (33%)	136 (30%)
Cell type composition, median (range)				
Granulocytes	0.57 (0.18-0.82)	0.69 (0.34-0.97)	0.59 (0.29-0.90)	0.59 (0.31-0.83)
CD4+ T cells	0.18 (0.04-0.44)	0.12 (0.00-0.35)	0.13 (0.00-0.37)	0.14 (0.00-0.37)
CD8+ T cells	0.02 (0.00-0.16)	0.07 (0.00-0.36)	0.17 (0.05-0.44)	0.16 (0.06-0.36)
B cells	0.07 (0.00-0.34)	0.05 (0.00-0.33)	0.04 (0.00-0.23)	0.04 (0.00-0.16)
Monocytes	0.08 (0.00-0.14)	0.04 (0.00-0.20)	0.03 (0.00-0.15)	0.03 (0.00-0.14)
Natural killer (NK) cells	0.11 (0.00-0.26)	0.04 (0.00-0.36)	0.00 (0.00-0.18)	0.00 (0.00-0.18)

Table 4.2: Participant characteristics, smoking measures, and Houseman cell type composition by self-reported smoking status

		NCDS			USM2			Overall		
Characteristic		**Never**	**Former**	**Current**	**Never**	**Former**	**Current**	**Never**	**Former**	**Current**
N		200	140	162	978	992	508	1178	1132	670
Male		86 (43%)	76 (54%)	79 (49%)	389 (40%)	501 (51%)	242 (48%)	475 (40%)	577 (51%)	321 (48%)
Age		44.03 (9.18)	44.01 (9.08)	44.01 (9.11)	49.41 (15.9)	53.15 (15.9)	47.37 (14.4)	48.49 (14.7)	52.02 (14.4)	46.56 (13.6)
McCartney		3.18 (0.43)	3.74 (0.78)	5.55 (0.87)	3.05 (0.26)	3.44 (0.59)	4.92 (0.80)	3.07 (0.30)	3.48 (0.62)	5.07 (0.86)
DNAme	**Never smoker**	197 (98%)	120 (86%)	25 (15%)	848 (87%)	527 (53%)	69 (14%)	1,045 (89%)	647 (57%)	94 (14%)
	Former smoker	0 (0%)	3 (2.1%)	0 (0%)	127 (13%)	422 (43%)	91 (18%)	127 (11%)	425 (38%)	91 (14%)
	Current smoker	3 (1.5%)	17 (12%)	137 (85%)	3 (0.3%)	43 (4.3%)	348 (69%)	6 (0.5%)	60 (5.3%)	485 (72%)
Congruenc	**Overall**	197 (98%)	3 (2.1%)	137 (85%)	848 (87%)	422 (43%)	348 (69%)	1,045 (89%)	425 (38%)	485 (72%)
	Positive	-	-	137 (100%)	-	-	348 (100%)	-	-	485 (100%)
	Negative	197 (100%)	3 (2.4%)	-	848 (100%)	422 (44%)	-	1045 (100%)	425 (40%)	-
Education	**Higher qualification**	73 (39%)	44 (34%)	34 (24%)	424 (49%)	396 (45%)	100 (23%)	497 (47%)	440 (44%)	134 (23%)
	A-level/equivalent	36 (19%)	14 (11%)	19 (13%)	191 (22%)	168 (19%)	103 (23%)	227 (21%)	182 (18%)	122 (21%)
	GCSE/ equivalent	72 (38%)	58 (45%)	72 (50%)	179 (20%)	207 (24%)	159 (36%)	251 (24%)	265 (26%)	231 (39%)
	No qualification	8 (4.2%)	13 (10%)	19 (13%)	80 (9.2%)	109 (12%)	82 (18%)	88 (8.3%)	122 (12%)	101 (17%)
SEC	**Management**	104 (56%)	69 (54%)	57 (39%)	300 (48%)	263 (44%)	97 (33%)	404 (50%)	332 (46%)	154 (35%)
	Intermediate	50 (27%)	36 (28%)	50 (34%)	157 (25%)	134 (22%)	61 (21%)	207 (26%)	170 (23%)	111 (25%)
	Routine	31 (17%)	22 (17%)	38 (26%)	164 (26%)	203 (34%)	136 (46%)	195 (24%)	225 (31%)	174 (40%)
Cell type composition	**Granulocytes**	0.57 (0.10)	0.56 (0.10)	0.56 (0.10)	0.58 (0.08)	0.59 (0.08)	0.59 (0.08)	0.58 (0.09)	0.58 (0.09)	0.58 (0.09)
	CD4+ T cells	0.18 (0.06)	0.18 (0.06)	0.19 (0.06)	0.14 (0.07)	0.13 (0.07)	0.14 (0.07)	0.14 (0.07)	0.14 (0.07)	0.15 (0.07)
	CD8+ T cells	0.03 (0.04)	0.03 (0.03)	0.02 (0.03)	0.18 (0.05)	0.17 (0.05)	0.17 (0.05)	0.15 (0.07)	0.16 (0.07)	0.13 (0.08)
	B cells	0.07 (0.04)	0.08 (0.03)	0.08 (0.03)	0.04 (0.03)	0.04 (0.02)	0.04 (0.03)	0.05 (0.03)	0.05 (0.03)	0.05 (0.03)
	Monocytes	0.07 (0.02)	0.08 (0.02)	0.08 (0.02)	0.04 (0.03)	0.04 (0.03)	0.03 (0.02)	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)
	NK cells	0.11 (0.05)	0.12 (0.05)	0.12 (0.05)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.02 (0.05)	0.02 (0.05)	0.03 (0.06)

Table 4.3: Participant characteristics and cell type composition by overall congruence between self-reported and DNAme-predicted smoking status

	NCDS		USM2		USM2 (aged 47-57)		Overall	
Characteristic	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
N (%)	165 (33%)	337 (67%)	860 (35%)	1,618 (65%)	182 (31%)	396 (69%)	1,025 (34%)	1,955 (66%)
Male, n (%)	88 (37%)	153 (63%)	383 (34%)	749 (66%)	80 (31%)	174 (69%)	471 (34%)	902 (66%)
Age, median (range)	44 (44-45)	44 (44-45)	51 (16-88)	51.50 (16-83)	53.50 (49-59)	54 (49-59)	46 (16-88)	47 (16-83)
Self-reported SSt, n (%)								
Never	3 (1.5%)	197 (98%)	130 (13%)	848 (87%)	27 (12%)	200 (88%)	133 (11%)	1,045 (89%)
Former	137 (98%)	3 (2.1%)	570 (57%)	422 (43%)	128 (58%)	92 (42%)	707 (62%)	425 (38%)
Current	25 (15%)	137 (85%)	160 (31%)	348 (69%)	27 (21%)	104 (79%)	185 (28%)	485 (72%)
McCartney MS, median (range)	3.56 (2.62-6.62)	3.44 (2.56-7.45)	3.18 (2.43-6.07)	3.23 (2.42-6.39)	3.11 (2.59-6.07)	3.31 (2.56-6.34)	3.23 (2.43-6.62)	3.26 (2.42-7.45)
Education, n (%)								
Higher qualification	50 (33%)	101 (67%)	349 (38%)	571 (62%)	73 (33%)	147 (67%)	399 (37%)	672 (63%)
A-level/equivalent	18 (26%)	51 (74%)	148 (32%)	314 (68%)	26 (28%)	66 (72%)	166 (31%)	365 (69%)
GCSE/ equivalent	62 (31%)	140 (69%)	203 (37%)	342 (63%)	51 (35%)	96 (65%)	265 (35%)	482 (65%)
No qualification	16 (40%)	24 (60%)	76 (28%)	195 (72%)	15 (30%)	35 (70%)	92 (30%)	219 (70%)
SEC, n (%)								
Management	80 (35%)	150 (65%)	235 (36%)	425 (64%)	71 (34%)	139 (66%)	315 (35%)	575 (65%)
Intermediate	42 (31%)	94 (69%)	123 (35%)	229 (65%)	35 (32%)	73 (68%)	165 (34%)	323 (66%)
Routine	28 (31%)	63 (69%)	180 (36%)	323 (64%)	40 (29%)	96 (71%)	208 (35%)	386 (65%)
Cell type composition, median (range)								
Granulocytes	0.56 (0.18-0.72)	0.57 (0.24-0.82)	0.59 (0.29-0.90)	0.59 (0.30-0.87)	0.59 (0.36-0.81)	0.58 (0.31-0.83)	0.59 (0.18-0.90)	0.59 (0.24-0.87)
CD4+ T cells	0.18 (0.07-0.44)	0.18 (0.04-0.38)	0.13 (0.00-0.34)	0.13 (0.00-0.37)	0.14 (0.00-0.33)	0.14 (0.00-0.37)	0.14 (0.00-0.44)	0.14 (0.00-0.38)
CD8+ T cells	0.02 (0.00-0.16)	0.02 (0.00-0.16)	0.17 (0.06-0.44)	0.16 (0.05-0.41)	0.15 (0.06-0.35)	0.16 (0.07-0.36)	0.16 (0.00-0.44)	0.15 (0.00-0.41)
B cells	0.08 (0.02-0.21)	0.07 (0.00-0.34)	0.04 (0.00-0.22)	0.04 (0.00-0.23)	0.04 (0.00-0.14)	0.04 (0.00-0.16)	0.04 (0.00-0.22)	0.04 (0.00-0.34)
Monocytes	0.08 (0.00-0.12)	0.08 (0.02-0.14)	0.03 (0.00-0.13)	0.03 (0.00-0.15)	0.03 (0.00-0.13)	0.03 (0.00-0.14)	0.04 (0.00-0.13)	0.04 (0.00-0.15)
NK cells	0.12 (0.02-0.26)	0.11 (0.00-0.26)	0.00 (0.00-0.11)	0.00 (0.00-0.18)	0.00 (0.00-0.11)	0.00 (0.00-0.18)	0.00 (0.00-0.26)	0.00 (0.00-0.26)

Table 4.4: Participant characteristics and cell type composition by positive congruence between self-reported and DNAme-predicted smoking status

	NCDS		USM2		USM2 (aged 47-57)		Overall	
Characteristic	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
N (%)	20 (13%)	137 (87%)	173 (33%)	348 (67%)	40 (28%)	104 (72%)	193 (28%)	485 (72%)
Male, n (%)	14 (17%)	68 (83%)	87 (35%)	165 (65%)	22 (31%)	49 (69%)	101 (30%)	233 (70%)
Age, median (range)	44 (44-44)	44 (44-45)	60 (17-88)	49 (17-81)	54.50 (49-59)	52 (49-59)	58 (17-88)	44 (17-81)
Self-reported SSt, n (%)								
Never	3 (100%)	0 (0%)	130 (100%)	0 (0%)	27 (100%)	0 (0%)	133 (100%)	0 (0%)
Former	17 (100%)	0 (0%)	43 (100%)	0 (0%)	13 (100%)	0 (0%)	60 (100%)	0 (0%)
Current	0 (0%)	137 (100%)	0 (0%)	348 (100%)	0 (0%)	104 (100%)	0 (0%)	485 (100%)
McCartney MS, median (range)	5.65 (4.41-6.62)	5.87 (4.48-7.45)	3.19 (2.53-6.07)	5.35 (3.94-6.39)	3.08 (2.64-6.07)	5.50 (4.20-6.34)	3.31 (2.53-6.62)	5.50 (3.94-7.45)
Education, n (%)								
Higher qualification	6 (18%)	28 (82%)	65 (52%)	59 (48%)	17 (59%)	12 (41%)	71 (45%)	87 (55%)
A-level/equivalent	3 (17%)	15 (83%)	21 (24%)	65 (76%)	6 (25%)	18 (75%)	24 (23%)	80 (77%)
GCSE/ equivalent	6 (8.2%)	67 (92%)	33 (24%)	106 (76%)	7 (17%)	35 (83%)	39 (18%)	173 (82%)
No qualification	4 (19%)	17 (81%)	29 (31%)	64 (69%)	6 (25%)	18 (75%)	33 (29%)	81 (71%)
SEC, n (%)								
Management	13 (21%)	49 (79%)	37 (40%)	56 (60%)	19 (48%)	21 (52%)	50 (32%)	105 (68%)
Intermediate	3 (6.8%)	41 (93%)	26 (38%)	43 (62%)	10 (37%)	17 (63%)	29 (26%)	84 (74%)
Routine	2 (5.9%)	32 (94%)	24 (20%)	96 (80%)	6 (17%)	30 (83%)	26 (17%)	128 (83%)
Cell type composition, median (range)								
Granulocytes	0.58 (0.34-0.71)	0.57 (0.31-0.81)	0.60 (0.31-0.78)	0.60 (0.30-0.87)	0.58 (0.36-0.78)	0.60 (0.37-0.83)	0.60 (0.31-0.78)	0.60 (0.30-0.87)
CD4+ T cells	0.17 (0.09-0.35)	0.18 (0.07-0.35)	0.14 (0.00-0.33)	0.14 (0.00-0.37)	0.15 (0.00-0.33)	0.15 (0.00-0.37)	0.14 (0.00-0.35)	0.15 (0.00-0.37)
CD8+ T cells	0.01 (0.00-0.10)	0.01 (0.00-0.13)	0.16 (0.07-0.34)	0.16 (0.05-0.34)	0.14 (0.07-0.34)	0.15 (0.08-0.29)	0.15 (0.00-0.34)	0.13 (0.00-0.34)
B cells	0.08 (0.04-0.12)	0.07 (0.02-0.18)	0.04 (0.00-0.14)	0.04 (0.00-0.13)	0.03 (0.00-0.14)	0.04 (0.01-0.12)	0.04 (0.00-0.14)	0.05 (0.00-0.18)
Monocytes	0.08 (0.05-0.11)	0.08 (0.03-0.13)	0.04 (0.00-0.12)	0.03 (0.00-0.13)	0.03 (0.00-0.09)	0.03 (0.00-0.13)	0.04 (0.00-0.12)	0.04 (0.00-0.13)
NK cells	0.11 (0.05-0.24)	0.11 (0.02-0.26)	0.00 (0.00-0.11)	0.00 (0.00-0.18)	0.00 (0.00-0.11)	0.00 (0.00-0.18)	0.00 (0.00-0.24)	0.00 (0.00-0.26)

Table 4.5: Participant characteristics and cell type composition by negative congruence between self-reported and DNAme-predicted smoking status

	NCDS		USM2		USM2 (aged 47-57)		Overall	
Characteristic	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
N (%)	145 (42%)	200 (58%)	687 (35%)	1,270 (65%)	142 (33%)	292 (67%)	832 (36%)	1,470 (64%)
Male, n (%)	74 (47%)	85 (53%)	296 (34%)	584 (66%)	58 (32%)	125 (68%)	370 (36%)	669 (64%)
Age, median (range)	44 (44-45)	44 (44-45)	47 (16-81)	53 (16-83)	53 (49-59)	54 (49-59)	44 (16-81)	49 (16-83)
Self-reported SSt, n (%)								
Never	0 (0%)	197 (100%)	0 (0%)	848 (100%)	0 (0%)	200 (100%)	0 (0%)	1,045 (100%)
Former	120 (98%)	3 (2.4%)	527 (56%)	422 (44%)	115 (56%)	92 (44%)	647 (60%)	425 (40%)
Current	25 (100%)	0 (0%)	160 (100%)	0 (0%)	27 (100%)	0 (0%)	185 (100%)	0 (0%)
McCartney MS, median (range)	3.43 (2.62-5.65)	3.10 (2.56-4.57)	3.18 (2.43-5.38)	3.11 (2.42-5.21)	3.11 (2.59-4.88)	3.13 (2.56-4.94)	3.22 (2.43-5.65)	3.11 (2.42-5.21)
Education, n (%)								
Higher qualification	44 (38%)	73 (62%)	284 (36%)	512 (64%)	56 (29%)	135 (71%)	328 (36%)	585 (64%)
A-level/equivalent	15 (29%)	36 (71%)	127 (34%)	249 (66%)	20 (29%)	48 (71%)	142 (33%)	285 (67%)
GCSE/ equivalent	56 (43%)	73 (57%)	170 (42%)	236 (58%)	44 (42%)	61 (58%)	226 (42%)	309 (58%)
No qualification	12 (63%)	7 (37%)	47 (26%)	131 (74%)	9 (35%)	17 (65%)	59 (30%)	138 (70%)
SEC, n (%)								
Management	67 (40%)	101 (60%)	198 (35%)	369 (65%)	52 (31%)	118 (69%)	265 (36%)	470 (64%)
Intermediate	39 (42%)	53 (58%)	97 (34%)	186 (66%)	25 (31%)	56 (69%)	136 (36%)	239 (64%)
Routine	26 (46%)	31 (54%)	156 (41%)	227 (59%)	34 (34%)	66 (66%)	182 (41%)	258 (59%)
Cell type composition, median (range)								
Granulocytes	0.56 (0.18-0.72)	0.58 (0.24-0.82)	0.58 (0.29-0.90)	0.59 (0.31-0.83)	0.59 (0.37-0.81)	0.58 (0.31-0.77)	0.58 (0.18-0.90)	0.59 (0.24-0.83)
CD4+ T cells	0.18 (0.07-0.44)	0.18 (0.04-0.38)	0.13 (0.00-0.34)	0.13 (0.00-0.37)	0.14 (0.00-0.33)	0.14 (0.00-0.37)	0.14 (0.00-0.44)	0.14 (0.00-0.38)
CD8+ T cells	0.02 (0.00-0.16)	0.02 (0.00-0.16)	0.17 (0.06-0.44)	0.17 (0.05-0.41)	0.16 (0.06-0.35)	0.17 (0.07-0.36)	0.16 (0.00-0.44)	0.16 (0.00-0.41)
B cells	0.08 (0.02-0.21)	0.06 (0.00-0.34)	0.04 (0.00-0.22)	0.04 (0.00-0.23)	0.04 (0.00-0.13)	0.04 (0.00-0.16)	0.04 (0.00-0.22)	0.04 (0.00-0.34)
Monocytes	0.08 (0.00-0.12)	0.07 (0.02-0.14)	0.03 (0.00-0.13)	0.04 (0.00-0.15)	0.03 (0.00-0.13)	0.03 (0.00-0.14)	0.04 (0.00-0.13)	0.04 (0.00-0.15)
NK cells	0.12 (0.02-0.26)	0.10 (0.00-0.26)	0.00 (0.00-0.09)	0.00 (0.00-0.11)	0.00 (0.00-0.04)	0.00 (0.00-0.05)	0.00 (0.00-0.26)	0.00 (0.00-0.26)

4.3.2. Logistic regressions explaining congruence

Figures 4.1 and 4.2 show forest plots used to display odds ratios and confidence intervals from logistic regression models looking at the association of sex, age, self-reported and DNAme-predicted smoking status, educational attainment and socioeconomic class on the congruence between self-reported and DNAme-predicted smoking status.

In terms of overall congruence, the effect of sex (Male vs Female) was statistically significant and negative (OR = 0.80, 95% CI [0.66-0.96], p = 0.019), the effect of age was statistically significant and positive (OR = 1.02, 95% CI [1.01, 1.02], p < .001) and the effect of former smoking (OR = 0.06, 95% CI [0.05-0.08], p < .001) and current smoking (OR = 0.29, 95% CI [0.22-0.38], p < .001) compared to never smoking is statistically significant and negative. In comparison to those with higher qualifications, the effect of achieving A-levels or equivalent (OR = 1.30, 95% CI [1.00-1.69], p = 0.049) or no qualifications (OR = 1.51, 95% CI [1.09, 2.08], p = 0.013) was statistically significant and positive. Obtaining GCSEs or equivalent compared to higher qualifications was not significant (Supplementary Table 2).

In terms of positive congruence, the effect of sex was statistically non-significant, and the effect of age was statistically significant and negative (OR = 0.94, 95% CI [0.93-0.96], p < .001). The effect of having A-levels or equivalent (OR = 2.89, 95% CI [1.61-5.33], p < .001), GCSEs or equivalent (OR = 3.11, 95% CI [1.91-5.14], p < .001), or no qualifications (OR = 2.90, 95% CI [1.67-5.13], p < .001) is statistically significant and positive in comparison to having a higher qualification (Supplementary Table 3). In terms of negative congruence, the effect of sex is statistically non-significant and positive, and the effect of age is statistically significant and positive (OR = 1.02, 95% CI [1.01, 1.03], p < .001). The effect of having GCSEs or equivalent (OR = 0.77, 95% CI [-0.48, -0.04], p = 0.02) is statistically significant and negative

in comparison to having higher qualifications but the effect of A-levels or equivalent or having no qualifications was non-significant (Supplementary Table 4).

When looking at socioeconomic class rather than education, overall, the effect of sex remains statistically significant and negative (OR = 1.24, 95% CI [0.63-0.99], p = 0.044) and the effect of age too which was positive (OR = 1.03, 95% CI [1.02, 1.04], p < .001). The effect of former smoking (OR = 0.03, 95% CI [0.02-0.04], p < .001) and current smoking (OR = 0.20, 95% CI [0.14-0.28], p < .001) compared to never smoking was statistically significant and negative. The effect of working in intermediate or routine occupations in comparison to managerial and professional occupations was statistically non-significant and positive (Supplementary Table 5). In terms of positive congruence, the effect of sex is statistically non-significant and positive, the effect of age is statistically significant and negative (beta = 0.93, 95% CI [0.91, 0.96], p < .001). The effect of working in intermediate occupations compared to managerial and professional occupations in positive cases is non-significant, but the effect of routine occupations is statistically significant and positive (OR = 2.65, CI [1.52-4.73], p < .001) (Supplementary Table 6). In terms of negative cases, the effect of sex is statistically non-significant and negative, the effect of age is statistically significant and positive (OR = 1.02, 95% CI [1.01, 1.03], p < .001), and the effect of working in intermediate and routine occupations is non-significant compared to managerial and professional occupations (Supplementary Table 7).

Table 4.6 shows summary statistics from simple logistic regression models (estimated using maximum likelihood) to predict overall congruence between self-reported and DNAme-predicted smoking status with cell type composition using each of six Houseman cell type estimates separately. This shows that DNA methylation-based measures of cell type composition did not appear to influence overall congruence between self-reported and DNAme-predicted smoking status. The only cell type to significantly associate

with congruence between smoking measures were levels of NK cells in NCDS (OR = 0.01, CI [0.00-0.24], p = 0.007).

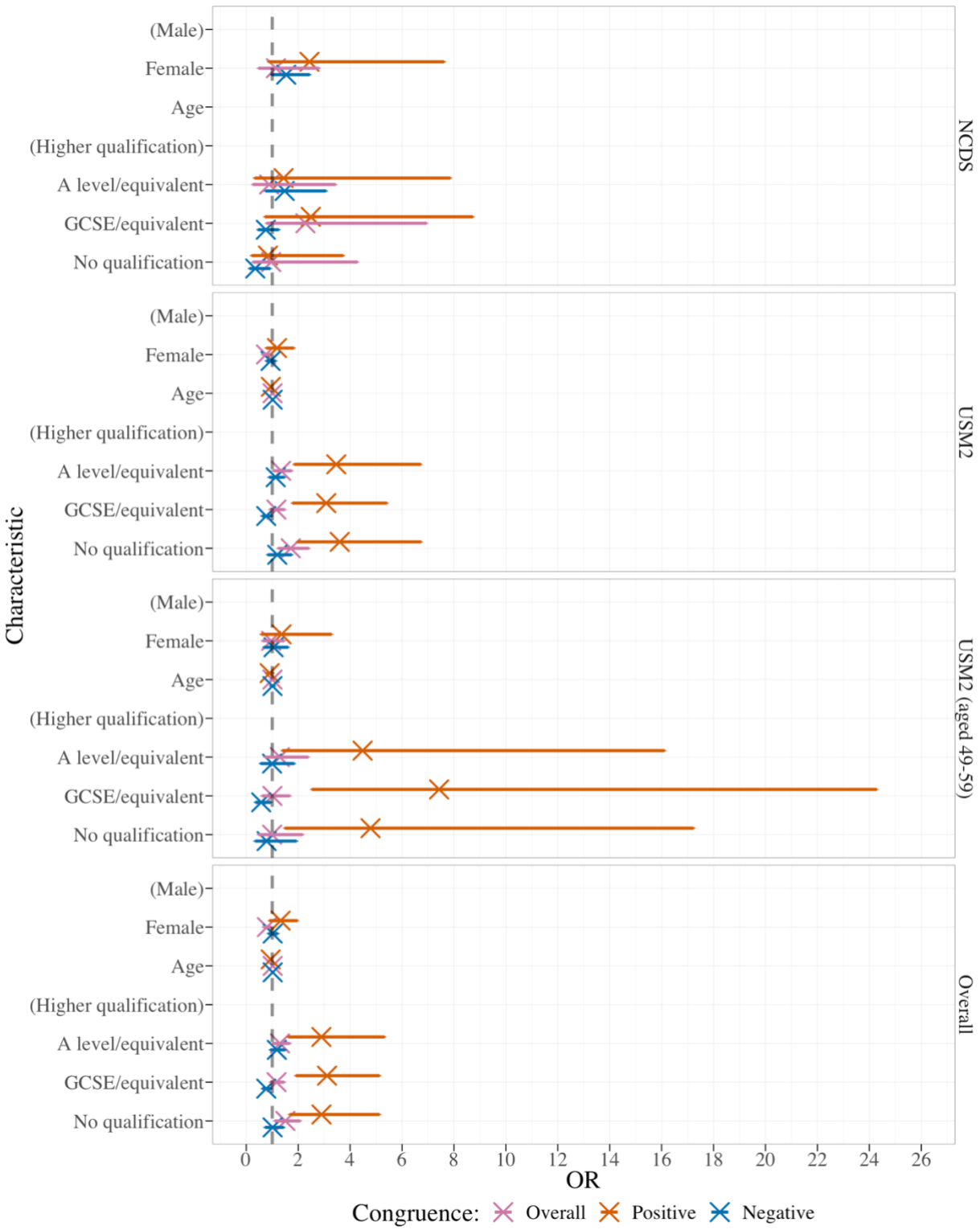

Figure 4.1: Forest plots showing Odds Ratios and 95% confidence intervals from logistic regressions investigating congruence between self-reported and DNAme-predicted smoking status in relation to educational attainment

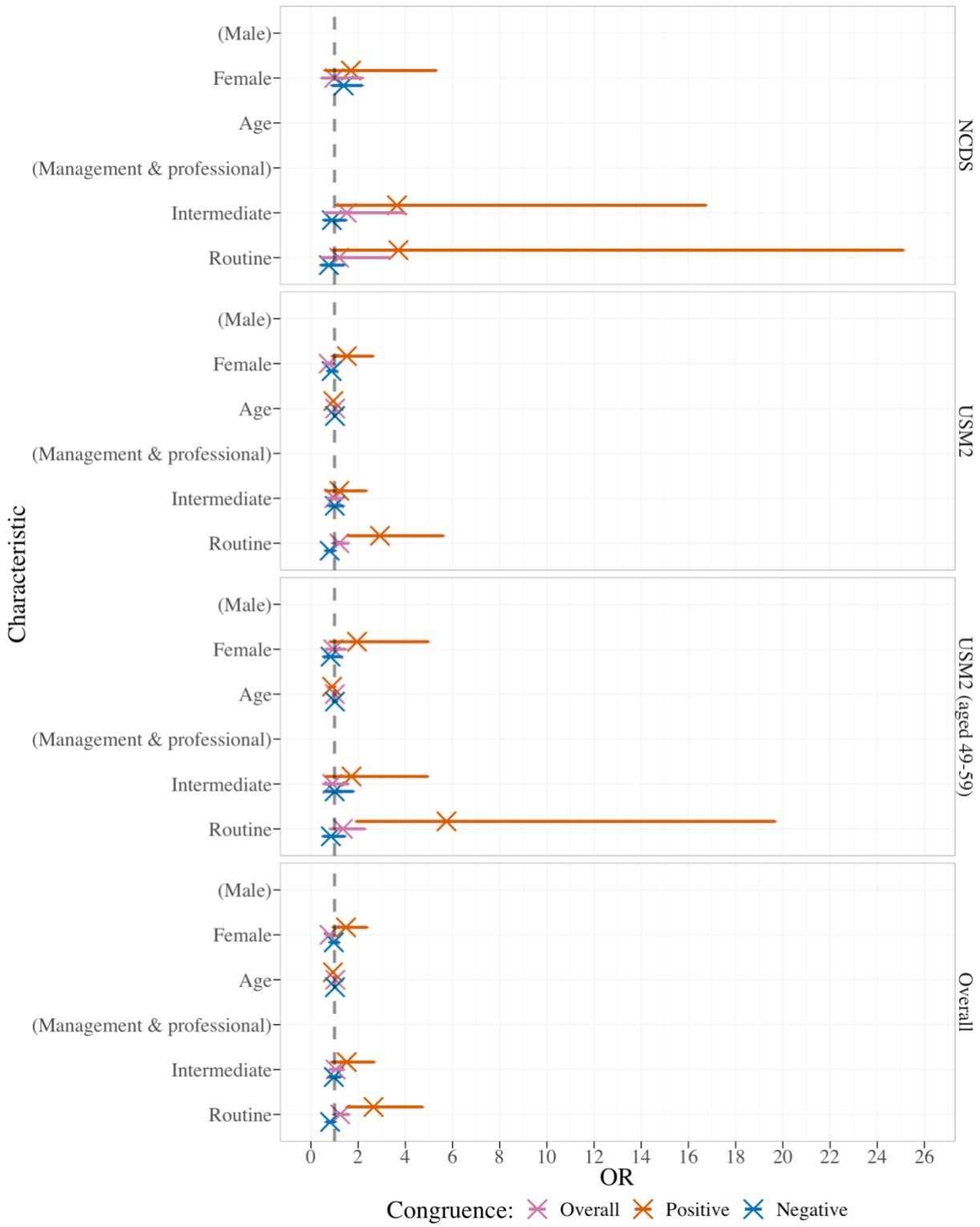

Figure 4.2: Forest plots showing Odds Ratios and 95% confidence intervals from logistic regressions investigating congruence between self-reported and DNAme-predicted smoking status in relation to socioeconomic classification

Table 4.6: Logistic regression showing impact of cell type composition on overall congruence between self-reported and DNAme-predicted smoking status

	NCDS				USM2				USM2 (aged 47-57)				Overall			
Cell type	**N**	**OR**	**CI**	**p**	**N**	**OR**	CI	**p**	**N**	**OR**	**CI**	**p**	**N**	**OR**	**CI**	**p**
Granulocyte	502	4.26	0.63, 29.1	0.14	2478	1.82	0.67, 4.95	0.2	578	0.37	0.04, 3.44	0.4	2980	2.09	0.87, 5.05	0.10
CD4+ T	502	1.21	0.06, 27.3	>0.9	2473	0.95	0.28, 3.25	>0.9	578	1.56	0.13, 19.2	0.7	2975	1.11	0.37, 3.33	0.9
CD8+ T	449	1.53	0.00, 679	0.9	2478	0.28	0.06, 1.34	0.11	578	16.0	0.47, 606	0.13	2927	0.37	0.13, 1.07	0.068
B cell	502	0.03	0.00, 7.42	0.2	2477	1.06	0.04, 29.4	>0.9	578	7.33	0.01, 8,442	0.6	2979	0.72	0.06, 9.64	0.8
Monocyte	502	0.01	0.00, 68.2	0.3	2464	0.82	0.03, 20.7	>0.9	575	0.01	0.00, 6.79	0.2	2966	0.99	0.07, 13.5	>0.9
NK	502	0.01	0.00, 0.24	0.007	2189	1.37	0.00, 66,550	>0.9	505	0.00	0.00, 14,928	0.4	2691	0.55	0.11, 2.76	0.5

4.3.3. Linear regressions explaining inflammatory markers

Figures 4.3 and 4.4 show forest plots used to display beta coefficients and confidence intervals from linear regression models looking at the association of sex, age, educational attainment and socioeconomic class, and compare adjustment of smoking using self-reported smoking status, DNAme-predicted smoking status or a smoking methylation score (MS) (McCartney et al., 2018) in relation to inflammatory markers fibrinogen and C-reactive protein.

Linear regression models were used to compare the relationship between fibrinogen and three different smoking measures including 1. self-reported smoking status, 2. DNAme-predicted smoking status, and 3. a smoking methylation score (McCartney et al., 2018), with educational attainment. Overall, the effect of self-reported former smoking (vs self-reported never smoking) was statistically non-significant while the effect of self-reported current smoking was statistically significant and positive (beta = 0.22, 95% CI [0.16, 0.28], $p < .001$). The effect of educational attainment when comparing participants achieving higher qualifications to those with A-levels or equivalent is statistically non-significant however compared to participants who have achieved GCSEs or equivalent (beta = 0.12, 95% CI [0.07, 0.18], $p < .001$) or no formal qualifications (beta = 0.29, 95% CI [0.21, 0.36], $p < .001$), the effect of educational attainment on fibrinogen is statistically significant and positive. The effect of DNAme-predicted former smoking (beta = 0.10, 95% CI [0.05, 0.16], $p < .001$) and current smoking (beta = 0.30, 95% CI [0.24, 0.36], $p < .001$) compared to never smoking is statistically significant and positive. When adjusting for smoking using DNAme-predicted smoking status, the effect of educational attainment remained statistically non-significant when comparing participants with higher qualifications to those with A-levels or equivalent. However, compared to participants who have achieved GCSEs or equivalent (beta = 0.12, 95% CI [.06, 0.17], $p < .001$) or no formal qualifications (beta = 0.26, 95% CI [0.18, 0.33], $p < .001$), the effect of

educational attainment on fibrinogen was statistically significant and positive. The effect of the McCartney et al (2018) smoking methylation score (MS) measures is statistically significant and positively associated with fibrinogen (beta = 0.12, 95% CI [0.10, 0.15], p < .001). Compared to participants overall who have achieved GCSEs or equivalent (beta = 0.11, 95% CI [.05, 0.16], p < .001) or no formal qualifications (beta = 0.25, 95% CI [0.18, 0.33], p < .001), the effect of educational attainment on fibrinogen was statistically significant and positive. It appears that adjustment for smoking using DNAme may alter the relationship of education and fibrinogen in comparison to using self-reports. In NCDS and the USM2 49-59 age subset there is no statistically significant difference in fibrinogen in DNAme-predicted former smokers compared to never smokers. In USM2 (aged 49-59), no significant difference was observed between participants with higher qualifications vs A levels nor GCSEs (or equivalent) but a significant and positive difference was shown when comparing to participants with no qualifications (0.23, CI [0.06-0.40], p = 0.009). In NCDS there was a statistically significance effect of education when comparing participants with higher qualifications vs GSCE or equivalent (0.17, CI [0.05-0.30], p = 0.007) but not when comparing to those with A levels nor with no qualifications (Supplementary Table 8). The effect of socioeconomic classification on fibrinogen was non-significant (Supplementary Table 9).

The association between C-reactive protein and self-reported smoking status, DNAme-predicted smoking status, the McCartney et al. (2018) smoking methylation score with educational attainment was also investigated. Overall, the effect of self-reported former smoking (vs self-reported never smoking) is statistically non-significant while the effect of self-reported current smoking is statistically significant and positive (beta = 0.32, 95% CI [0.21, 0.44], p < .001). The effect of educational attainment is statistically significant and positive when comparing participants with higher qualifications to participants with no formal qualifications (beta = 0.48, 95% CI [0.34, 0.62], p < .001) and participants with A-levels or equivalent (beta = 0.12, 95% CI [0.00, 0.24], p < 0.043) however the effect of achieving GCSEs or equivalent was non-significant. The effect of DNAme-predicted former smoking (beta = 0.30, 95% CI [.19,

0.41], p < .001) and current smoking (beta = 0.47, 95% CI [0.35, 0.58], p < .001) compared to never smoking is statistically significant and positive. The effect of the smoking methylation score is statistically significant and positively associated with CRP (beta = 0.16, 95% CI [0.11, 0.20], p < .001). When adjusting for smoking using DNA methylation-based methods the effect of educational attainment remained like that of self-reported smoking. The effect of educational attainment on CRP was statistically non-significant when comparing participants with higher qualifications against those with GCSEs or equivalent but significant and positive when comparing to those who have achieved A levels or equivalent (beta = 0.12, CI = [0.00-0.24], p = 0.044) or no formal qualifications (beta = 0.41, CI = [0.27-0.56], p < 0.001) when adjusting for smoking using DNAme-predicted smoking status ('smokp SSt'). The effect of educational attainment on CRP was statistically non-significant when comparing higher qualifications against those with GCSEs or equivalent but significant and positive when comparing to those who have achieved A levels or equivalent (beta = 0.12, CI = [0.00-0.24], p = 0.044) or no formal qualifications (beta = 0.44, CI = [0.30-0.59], p < 0.001) when adjusting for smoking using the smoking methylation score (McCartney et al., 2018). In NCDS and the USM2 49-59 age subset there is no statistically significant difference in CRP in participants who were classified as former smokers using DNAme. It appears that adjustment for smoking using DNAme does not significantly change the relationship of education and CRP in comparison to self-reports (Supplementary Table 10).

When comparing the relationship of CRP with self-reported smoking status, DNAme-predicted smoking status, McCartney MS, and socioeconomic classification the effect of working in intermediate occupations (vs managerial and professional) was non-significant however compared to participants in routine occupations the effect is statistically significant and positive (beta = 0.16, 95% CI [0.04, 0.27], p = 0.007). This significant difference in CRP between those in managerial and professional vs routine occupations remained significant in USM2 using the full age range (16-88) and the USM2 49-49 subset but did not appear in NCDS.

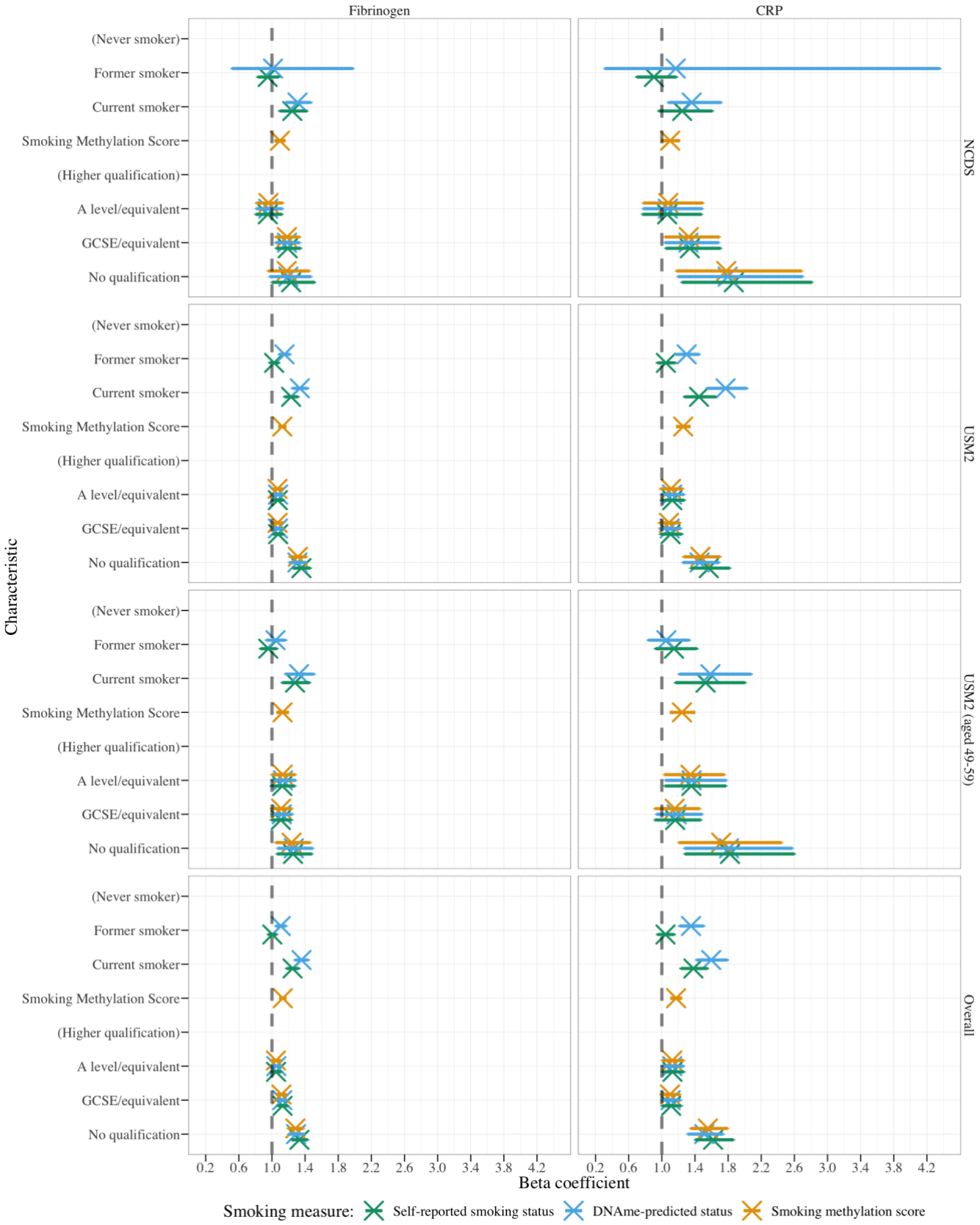

Figure 4.3: Forest plots showing Beta coefficients and 95% confidence intervals from linear regressions investigating associations of fibrinogen and CRP with DNAme-predicted smoking status and educational attainment

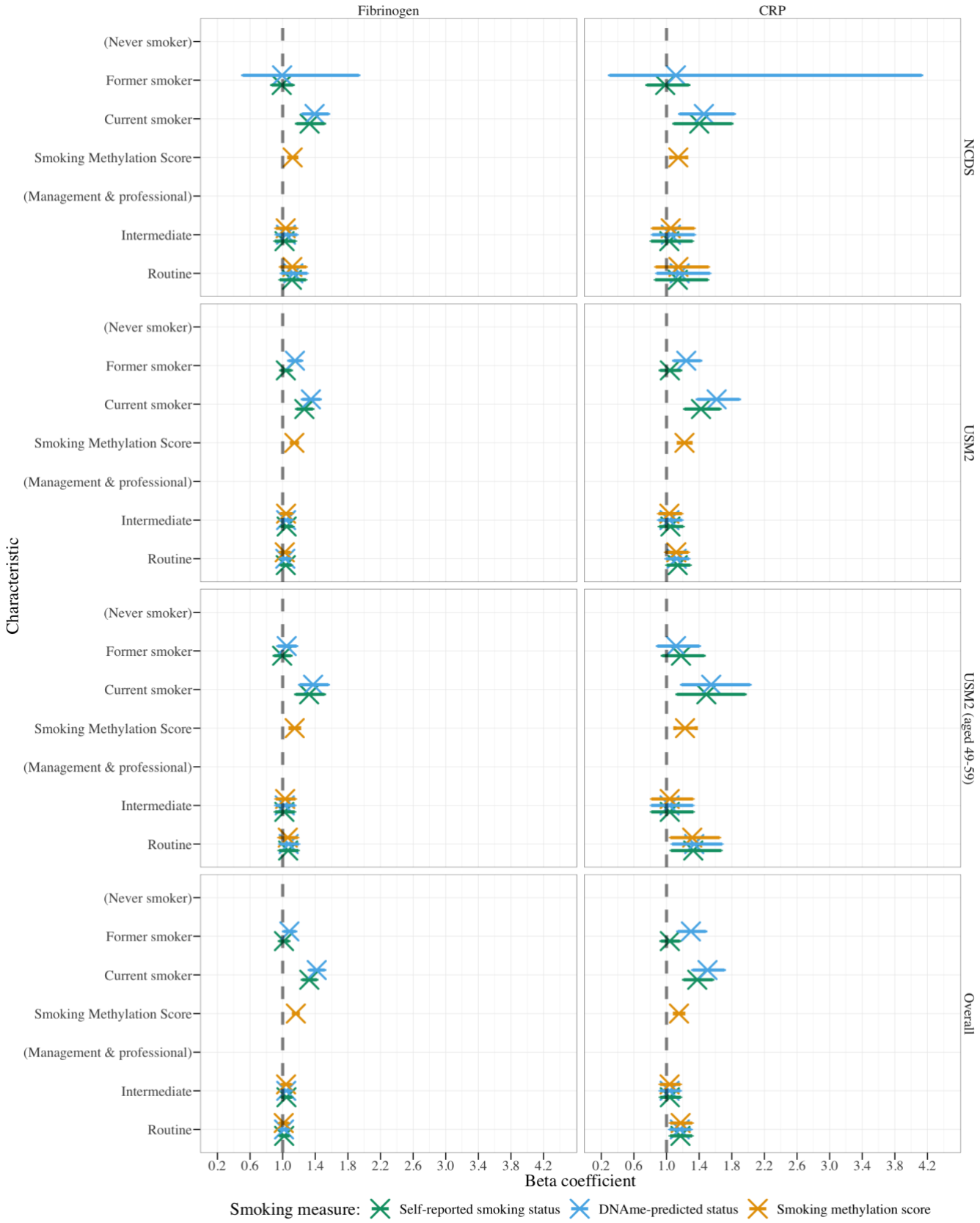

Figure 4.4: Forest plots showing Beta coefficients and 95% confidence intervals from linear regressions investigating congruence associations of fibrinogen and C-reactive protein with DNAme-predicted smoking status and socioeconomic classification

4.4. Discussion

This chapter first aimed to investigate how age, sex, smoking status, education, socioeconomic classification and cell type composition influence the agreement between smoking status measured using self-reports or predicted from DNA methylation. The second aim of this chapter was to compare the use of self-reported and DNAme-predicted smoking as covariates when investigating educational and socioeconomic gradients in inflammatory markers fibrinogen and CRP. The results of this chapter suggest that sex may play a role in the overall congruence between self-reported and DNAme-predicted smoking in NCDS, who were all aged 44 when bloods were collected, and in USM2 limited to 49 to 59 years old, but not in USM2 where the full age range is used. Age may impact overall congruence when including a larger age range. In this case a birth cohort effect can occur where age impacts education as the school leaving age has increased and affluence gradients may differ between generations. Associations therefore vary across the lifespan. It also appears that education and socioeconomic classification impacts overall congruence between self-reported and DNAme-predicted smoking status where misclassification was more common in more affluent participants who had achieved more educational qualifications, or reported working in professional or managerial occupations, compared to less affluent participants. These factors were more influential in driving positive cases, consisting of either smokers who were accurately classified or non-smokers who were misclassified as smokers, compared to negative cases, consisting of non-smokers accurately classified using DNA methylation or smokers inaccurately predicted as non-smokers. This shows both similarities and differences to drivers in discrepancies seen in cotinine, a metabolite of nicotine. Misclassification rate between self-reports and cotinine levels was instead higher in those with a high school education or less, ex-smokers (Lynne et al., 1992). However past smoking, age, and sex also influenced congruence, as in this study (Caraballo et al., 2001).

This chapter also suggests that DNAme-predicted smoking status may more strongly relate to levels of fibrinogen and C-reactive protein than self-reported smoking. Other methylation scores of health behaviours have also been shown to reflect health-related variables such as mortality more closely than self-reported measures of the same behaviours (Corley et al., 2019). Also, in studying the association of educational attainment and socioeconomic classifications with inflammatory markers, the way smoking is adjusted or measured may influence findings related to educational and socioeconomic drivers in inflammation. This all suggests that DNA methylation-based measures of smoking may offer a more objective measure of smoking that is less influenced by educational and socioeconomic factors compared to self-reports.

4.5. Conclusion

In this chapter we investigated factors that influenced the agreement between self-reported and DNAme-predicted smoking measures. This showed that educational attainment and socioeconomic classification (SEC) are possible predictors of congruence between self-reported and DNAme-predicted smoking status. The impact of educational attainment and socioeconomic class was also greater in positive cases consisting of smokers correctly classified via DNA methylation or non-smokers incorrectly classified, compared to negative cases where self—reported non-smokers were classified correctly or self-reported smokers were not. This chapter also investigated whether smoking measured using self-reported or DNA methylation bases measures differed in their association with inflammatory markers fibrinogen and CRP. Generally self-reported smoking status was more strongly associated with inflammation compared to DNAme-based measures. It was also of interest that the association of fibrinogen and CRP with education or SEC differs depending on the adjustment for smoking used. This suggested that when adjusting using self-reports, education is more strongly associated with inflammation compared to adjustments using DNAme-predicted

smoking status or the smoking methylation score. Social drivers in inflammation may then play an even larger role, in comparison to health behaviours such as smoking, than previously thought.

5. **Differences in DNA methylation associated with inflammatory markers**

5.1. Introduction

Inflammation generally refers to the biological response of tissues to injuries, irritants, toxins, hypersensitivities, infection by pathogens, and stress and trauma. This complex process involves different immune cells, blood vessels, many molecular mediators, and the cardiovascular system. Inflammation is a generic response and works as a mechanism of innate immunity. This refers to the activation and coordination of pre-existing mechanisms starting at the body's natural barriers such as skin, mucosa and other secretions. Inflammation can be classified as acute or chronic. Acute inflammation concerns the initial response of the body to harmful stimuli whereas prolonged inflammation, also known as chronic inflammation, refers to an inflammatory response that persists long after the initial cause. This process involves a progressive shift in the type of cells present at the site of inflammation such as mononuclear cells (Abbas et al., 2019). Acute inflammation is accomplished by the increased movement of plasma and leukocytes, white blood cells and particularly granulocytes, from the blood to sites of injury in tissues. A series of biochemical events then propagates and matures this inflammatory response using the local vascular system, immune system, and various cells. Chronic inflammation is characterized by simultaneous destruction and healing of tissue from the inflammatory process. Throughout the inflammatory response chemicals are released that can be measured and used to detect inflammation. Fibrinogen and C-reactive protein (CRP) are two commonly measured inflammatory markers that circulate in the blood and are both produced in the liver. Raised levels of these markers are known risk factors for many chronic conditions and diseases (Liu et al., 2020) and have been implicated in socio-economic inequalities in health across age (Davillas et al., 2017).

Acute inflammation occurs immediately upon injury where cytokines and chemokines promote the migration of neutrophils and mononuclear cells to the site of inflammation (Hannoodee and Nasuruddi, 2020). Cytokines and chemokines are redundant secreted proteins that direct immune cell trafficking. Chronic inflammation however lasts for months or years and, in contrast to neutrophils in acute inflammation, macrophages, lymphocytes, and plasma cells predominate in chronic inflammation. Many diseases are mediated by chronic inflammation and many factors such as obesity, smoking, stress and diet can promote chronic inflammation (Pahwa et al., 2018). Immune cells present at the site of inflammation possess surface receptors known as pattern recognition receptors (PRRs). During acute inflammation PRRs bind to two subclasses of molecules: pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs). This binding leads to the release of inflammatory mediators responsible for the clinical signs of inflammation. Subsequently vasodilation occurs leading to increased blood flow and permeability of the blood vessels, resulting in leakage of plasma proteins and fluid into the tissue leading to swelling. Cellular mediators also permit the migration of leukocytes into the tissue via an acellular chemotactic gradient, but these mediators are short lived. Other biochemical cascade systems also act in parallel during the inflammatory response and include the complement system activated by bacteria, the coagulation system which forms a protective protein mesh over sites of injury, and the fibrinolysis system. The fibrinolysis system acts in opposition to the coagulation system to counterbalance clotting and generate several other inflammatory mediators (Robbins and Cotran, 1979). If an organism or pathogen is not contained by the actions of acute inflammation it can lead to systemic effects on the entire body and evade the tissue at the site of infection by gaining access to the lymphatic system via nearby lymph vessels and lymphatic drainage into the circulatory system. When lymph nodes cannot destroy all pathogens, the infection spreads further. Normally a few hours after an inflammatory response begins a complex coordinated program of resolution occurs. After entering tissues, granulocytes promote the switch of prostaglandins and leukotrienes to lipoxins, anti-inflammatory molecules that initiate a termination sequence. Neutrophil recruitment then ceases and programmed death by apoptosis begins. These events

coincide with the biosynthesis of resolvins and protectins from omega-3 polyunsaturated fatty acids. This process critically shortens the period of neutrophil infiltration by initiating apoptosis. Consequently, apoptotic neutrophils undergo phagocytosis by macrophages, leading to neutrophil clearance and release of anti-inflammatory and reparative cytokines such as transforming growth factor-β1. The anti-inflammatory program ends with the departure of macrophages through the lymphatic system (Serhan and Savill, 2005).

Chronic inflammation is a hallmark of obesity. Many markers of inflammation are elevated in obese people and waist circumference correlates significantly with systemic inflammatory response (Parimisetty et al., 2016). Abnormalities in inflammation are further related to many disorders. The immune system is often involved with inflammatory disorders such as allergic reactions and some myopathies, and some immune diseases have shown links to in inflammatory processes involved in atherosclerosis, ischemic heart disease and cancer (Ungefroren et al, 2011; Coussens and Werb, 2002). Conversely many cells of the immune system also contribute to cancer immunology by suppressing cancer (Gunn et al., 2012). Molecular intersection between receptors of steroid hormones, cellular development, and transcription factors plays key roles in inflammation and cancer. It can impact NF-κB, 'nuclear factor kappa-light-chain-enhancer of activated B cells', which may mediate some of the most critical effects of inflammatory stimuli on cancer cells (Copland et al., 2009). Approximately 15 to 20% of human cancers are associated with chronic inflammation (Mantovani et al. 2008) and 1 in 2 people will develop some type of cancer in their lifetime, with lung cancer being one of the most common (NHS, 2019). Clinical studies have also shown strong links between inflammation and many other diseases too. In patients with atherosclerosis for example elevation in markers of inflammation predicts outcomes of patients with acute coronary syndromes. Low-grade chronic inflammation estimated using CRP also predicts risk of atherosclerotic complications. This can add prognostic information beyond the realm of traditional risk factors. Some treatments that reduce coronary risk also limit inflammation such as with lipid lowering statins (Libby, P, 2012). Another example of a

disorder involving inflammation is hay fever which is caused by a hypersensitive response by mast cells to allergens. Mast cells respond by releasing vasoactive chemicals leading to an inflammatory response and recruitment of leukocytes. Due to the central role of leukocytes in the development and propagation of inflammation, defects in leukocyte functionality often result in a decreased capacity for inflammatory defence leading to increased vulnerability to infection and diseases (Robbins and Cotran, 1979). A large increase in the number of leukocytes in the blood is often also observed in inflammation. Bacterial infection usually results in an increase of neutrophils whereas diseases such as asthma, hay fever, and parasite infestation result in an increase in eosinophils. Other factors influencing inflammation include Vitamin A deficiency which may cause an increase in inflammatory responses (Wiedermann et al., 1996). There is also evidence for a link between inflammation and depression through an increase in cytokines and classical symptoms of being physically sick often show a large overlap with depressive symptoms (Berk et al., 2013). Clinical trials have shown that anti-inflammatory medicines taken in addition to antidepressants significantly improves symptoms and increases positive response to treatment (Müller et al., 2006). Most recently evidence for a link between inflammation and delirium has been proposed based on the results of a recent longitudinal study investigating CRP in COVID-19 patients (Saini et al., 2021). This goes to show that inflammation is of massive interest in population health research as it plays a vital role in a plethora of diseases with shared pathologies.

Fibrinogen and CRP are both acute-phase proteins (APPs). APPs are a class of proteins whose plasma concentrations increase or decrease in response to inflammation. When local inflammatory cells (neutrophil, granulocytes and macrophages) secrete cytokines into the bloodstreams the liver responds by producing many acute-phase reactants. Although APPs are generally beneficial in acute inflammation they can contribute to amyloidosis in chronic inflammation where abnormal proteins known as amyloid fibrils build up in tissue. After stimulation by proinflammatory cytokines, Kupffer cells produce IL-6 in the liver. IL-6 is the major mediator for the hepatocytic secretion of APPs. Synthesis of APP can also be regulated

indirectly by cortisol and cortisol can enhance expression of IL-6 receptors in liver cells and induce IL-6-mediated production of APPs (Jain et al., 2011). Fibrinogen and C-reactive protein are both positive APPs meaning concentration increases with inflammation, but they serve different purposes as part of the innate immune system. CRP acts to destroy or inhibit the growth of microbes by binding opsonins to substances or cells. Opsonins are extracellular proteins that then induce phagocytosis. CRP concentration increases following IL-6 secretions from macrophages and T cells and its role is to activate the complement system via C1q (Thompson et al., 1999). Fibrinogen is a coagulation factor that affects coagulation and is converted enzymatically by thrombin to fibrin and then to a fibrin-based blood clot. Fibrinogen also mediates blood platelet and endothelial cell spreading, tissue fibroblast proliferation, capillary tube formation, and angiogenesis (Mosesson, 2005).

Inflammation can lead to DNA damage due to the generation of reactive oxygen species (ROS) and reactive nitrogen species (RNS) by various intracellular inflammatory mediators, leukocytes and other phagocytic cells. ROS and RNS are normally produced by these cells to fight infection (Coussens and Werb, 2002). Genome-wide analyses of human cancer tissues have also revealed that a single typical cancer cell may possess roughly 100 mutations in coding regions. Chronic inflammation also causes epigenetic alterations such as changes in DNA methylation that are often more common than mutations (Chiba et al, 2012). DNA repair genes are frequently inactivated by methylation in various cancers (Ding et al., 2019). A study recently evaluated the relative importance of mutations and epigenetic alterations in the progression to two different types of cancer. This report showed that epigenetic alterations were much more important than mutations in generating gastric cancers associated with inflammation but were of roughly equal importance in generating esophageal squamous cell cancers associated with tobacco chemicals and alcohol metabolism (Yamashita et al, 2018). It is then of importance to understand the epigenetic underpinnings associated with levels of inflammatory markers.

Epigenetic mechanisms also influence the genetic regulation of pathways related to inflammation. Immune cells involve a complex network of different cell types and interactions that require differentiation to determine cell phenotype and function. The latter is highly dependent on epigenetic profiles that in turn establish transcriptional programs and bridge the gap between the environment and genome regulation. Recent advances in genome-wide DNA methylation data have provided insights into the roles of DNA methylation in health and disease (Calle-Fabregat et al., 2020). There is some suggestion that decreases to global DNA methylation may occur with increasing levels of inflammation and epigenome-wide association studies report most differentially methylated genes are hypomethylated in inflammatory processes (Gonzalez-Jaramillo et al., 2019). CRP has been the most widely studied marker in the epigenetic signatures of inflammation. A recent systematic review showed 17 studies where the relationship between epigenetics and CRP is evaluated. 5 used hypothesis-free EWAS approaches to identify DMPs related to CRP but none for fibrinogen were mentioned in this meta-analysis. In this review only two studies investigated fibrinogen but no EWAS of fibrinogen were mentioned (Gonzalez-Jaramillo et al., 2019). It is therefore of interest to understand the epigenetic signatures of fibrinogen as well as CRP to better understand the role of DNA methylation in inflammation and investigate the similarity or differences between signatures of each marker.

A recent meta-analysis showed hundreds of CpG sites where DNA methylation is significantly associated with chronic low-grade inflammation measured using serum CRP in a large European population. 58 loci were replicated among African Americans and 88% of replicated loci were also associated with at least one related cardiometabolic feature. This study also suggested up to 6% inter-individual variation in CRP could be explained by using a DNA methylation-based additive weighted score even after adjustment for age and sex (Ligthart et al., 2016). The most significant CpG site associated with CRP was located within the *AIM2*

gene and increased DNA methylation at this locus was associated with lower expression of *AIM2* protein and lower CRP levels. *AIM2* stands for Absent in Melanoma 2 and is an inflammasome receptor for double-stranded DNA. This protein activates inflammatory cascades and is implicated in host defence mechanisms against bacterial and viral pathogens (Rottenberg and Carow, 2014). Using a gene specific approach another study also found higher levels of CRP to be associated with lower degree of methylation of *AIM2* (Miller et al., 2018). The *SOCS3* gene has also been implicated which has a role in atherosclerosis where lower DNA methylation was associated with increased expression of *SOCS3*. *SOCS3* plays a pivotal role in the innate immune system as a regulator of cytokine signalling and is referred to as suppressor of cytokine signalling 3 (Hornung et al., 2009). Another study also reported *SOCS3*, among others, to be significantly associated with CRP levels in peripheral blood tissue and human liver tissue (Marzi et al., 2016). This meta-analysis stated a lack of overlap between loci identified in GWAS and EWAS studies related to inflammation suggesting different molecular mechanisms are at play. This is similar to smoking in that the top signals in GWAS tend to be based in genes related to nicotine dependence while many top EWAS loci are involved in toxin clean up systems. An exception to this involves loci in the *TMEM49* gene which was found to be inversely associated with sTNFR2 and IL-6 levels in a separate candidate gene approach study (Smith et al., 2014) and shared the same direction of association with CRP levels in EWAS (Ligthart et al., 2016). Transmembrane Protein 49 (*TMEM49*) is also referred to as Vacuole Membrane Protein 1 (*VMP1*) and is a transmembrane protein that plays a key regulatory role in the process of autophagy. Other findings from previous EWAS of inflammatory markers have implicated many other genes including *AIM*, *RPS6KA2* and *PHOSPHO1* (Ligthart et al., 2016), AQP3 and BCL3 (Marzi et al., 2016), and has shown CRP was positively associated with DNA methylation age using Hannum's approach (Verschoor et al., 2018). The latter is interesting given inflammation, epigenetics, and metabolism all converge and influence cell senescence and ageing (Zhu et al., 2021).

DNA methylation at many replicated CpG sites linked to inflammatory marker concentrations also showed associations with cardiometabolic phenotypes, suggesting an epigenetic overlap with different diseases. This highlights evidence of a pleiotropic network of epigenetic modifications across various phenotypes. Pleiotropy refers to the genetic effect of a single gene on multiple phenotypic traits. Behavioural phenotypes are often regulated by many genes, and the behavioural effects of a gene often dependent on environmental conditions and genetic background however many genes that regulate disease and health-realted phenotypes are themselves very complex with several gene products and functions at play (Anreiter and Sokolowski et al., 2018). It has been shown that methylation may harbour new information in explaining the variation of complex traits such as inflammation characterized by a strong influence of environment. Epigenetic signatures throughout the genome are highly labile due to temporal or spatial factors and in turn inflammatory markers are affected by both genetic and environmental factors. Careful consideration is then critical when considering the inclusion of certain confounders in epigenetic studies and candidate-gene approach studies up to this point have failed to properly control for lifestyle factors that influence inflammation (Gonzalez-Jaramillo et al., 2019). Although most EWAS studies do control for age, sex, cell type composition and smoking, other factors have not always been considered and, up to this point, EWAS studies have mostly been carried out using candidate-gene approaches such as pyrosequencing or the 27K and 450K BeadChip technology from Illumina. The new Infinium EPIC array used in this study is capable of quantifying DNA methylation at almost double the number of CpG sites than the predecessor.

The aim of this study is to investigate the relationship between inflammatory markers and DNA methylation within the UK Household Longitudinal Study (UKHLS) and the National Child Development Study (NCDS) 1958 Birth Cohort. Adjustment for cell type composition, sex, age, smoking status, BMI and educational attainment were also investigated to see how DNA methylation signatures of CRP or fibrinogen may be influenced by these factors. UKHLS ages ranged from 16 to 98 and venous blood samples were collected during the wave 3 (2010-12) nurse visit of Understanding Society. Bloods were collected at age

44-45 in NCDS during the biomedical sweep. 460 NCDS participants, with complete data for the included variables were used in this study. Two subsets of UKHLS were used and referred to as USM1 and USM2. 766 USM1 and 1,826 USM2 participants from the UKHLS study were included in this study where complete data was available.

5.2. Methods

The R package limma was used to identify significantly differentially methylated sites associated with CRP and fibrinogen using large-scale microarray data. The package operates on a matrix of methylation values where the 'lmFit' function fits a linear model to each row of data, considering a specified design matrix that details relevant information related to each sample array, and specifies the hypothesis to be tested. Within this study, the treatment-contrasts parametrization method was used to construct design matrices using the 'model.matrix' function. The resulting object consists of a list of probes from most to least likely to be differentially methylated by inflammatory markers specified in the design matrix and p-values were adjusted as a control for multiple testing. CRP was log transformed as it was not normally distributed. The 'pwod' function in the watermelon R package was used to remove any probe-wise outliers.

5.3. Results

The aim of this chapter was to investigate whether DNA methylation, measured at over 850,000 CpG sites, is significantly associated with two inflammatory markers, fibrinogen and C-reactive protein (CRP). This chapter also aims to enable better understanding of how epigenetic signatures implicated in inflammation differ depending on the epigenome-wide association study (EWAS) model being specified. 10 models per inflammatory marker were investigated and are listed below. Each model uses DNA methylation at each

CpG site (N = 853,973) as the dependent variable and either fibrinogen or CRP alongside any covariates, as the independent variables. Each additional covariate was added iteratively alongside the covariates of the previous model. However, in models 5-10 each smoking measure was added separately so only one adjustment for smoking per model is made. Adjustment for smoking using either self-reported or DNA methylation-based smoking status or a smoking methylation score (McCartney et al., 2018) was then compared. Cell type composition was estimated using DNA methylation (Houseman, 2011) and included granulocytes, CD8T, CD4T and B cells, monocytes, and natural killer cells.

$$\boldsymbol{Model\ 1}: \sim Fibrinogen\ or\ CRP$$

$$\boldsymbol{Model\ 2}: Model\ 1 + DNAme\ estimated\ cell\ types$$

$$\boldsymbol{Model\ 3}: Model\ 2 + age + sex$$

$$\boldsymbol{Model\ 4}: Model\ 3 + BMI$$

$$\boldsymbol{Model\ 5}: Model\ 4\ + self - reported\ smoking\ status$$

$$\boldsymbol{Model\ 6}: Model\ 4 + DNAme - predicated\ smoking\ status$$

$$\boldsymbol{Model\ 7}: Model\ 4 + smoking\ methylation\ score$$

$$\boldsymbol{Model\ 8}: Model\ 5 + educational\ attainment$$

$$\boldsymbol{Model\ 9}: Model\ 6 + educational\ attainment$$

$$\boldsymbol{Model\ 10}: Model\ 7 + educational\ attainment$$

1.1.1. Descriptive statistics

EWAS models were ran separately in three independent datasets which included one from the 1958 National Child Development Study (NCDS) and two from the UK Household Longitudinal Study

(UKHLS). The two UKHLS datasets consist of participants whose bloods were used to construct methylation resources. This was done in two batches and as such these are referred to as USM1 and USM2. All models were restricted to the subset of individuals with complete data for the two inflammatory markers and other stated covariates including age, sex, BMI, self-reported smoking status, and educational attainment. Complete data was available from 460 participants in NCDS, 835 in USM1, and 1,838 in USM2. In total 3,133 participants were included in this investigation. Across all three datasets, fibrinogen ranged from 0.50 to 5.70 g/l and CRP ranged from 0.08 to 152.00 mg/l. In NCDS, fibrinogen ranged from 1.41-5.57, in USM1 from 0.50-5.70, and in USM2 from 1.10-5.70. In NCDS CRP ranged from 0.08-152.00, in USM1 from 0.2-115.5, and in USM2 from 0.2-90.7. This shows that some participants in NCDS and USM1 had above normal CRP levels where measures above 10mg/l is generally considered high. Across the three datasets 44% of participants were male including 47% in NCDS, 42% in USM1 and 45% in USM2. In NCDS over 98% of participants were aged 44 when bloods were collected while the remainder had recently turned 45 years old. In USM1 ages ranged from 28 to 97 and in USM2 ages ranged from 16 to 83 when limiting to participants with complete data (Table 5.1).

31% of NCDS, 15% of USM1 and 20% of USM2 participants reported currently smoking at the time of blood collection. This shows a greater proportion of smokers in NCDS compared to both UKHLS datasets, and more smokers in USM2 compared to USM1. Fewer former smokers were observed in NCDS (28%) compared to USM1 (39%) and USM2 (40%). 32% of NCDS and 15% of USM2 participants were DNAme-predicted as current smokers. 0.7% of NCDS and 25% of USM2 participants were predicted as former smokers. This shows an underestimation of past smoking in both datasets but especially in NCDS. The average smoking methylation score (McCartney et al., 2018) measured 4.09 (± 1.24) in NCDS which was on average greater compared to MS measures in USM1 (3.56 ± 0.93) and USM2 (3.56 ± 0.86). Body mass index (BMI) ranged from 14.50 to 54.04 across the three datasets and did not significantly vary between them. 33% of participants in NCDS, 32% in USM1 and 42% in USM2 had obtained a higher educational

qualification or degree. 8.7% in NCDS, 20% in USM1 and 13% in USM2 had not obtained any formal educational qualifications. A greater proportion of participants in NCDS (44%) had obtained GCSEs or equivalent compared to USM1 (27%) and USM2 (25%). This suggests that while NCDS were more likely to get at least some qualifications, participants in the two UKHLS samples were more likely to take on higher level qualifications. With this said this may be influenced by the large age range in UKHLS compared to NCDS. Granulocyte proportions measured on average 56% in NCDS, 69% in USM1, and 59% in USM2 showing higher proportions in USM1. Natural killer cell proportions measured on average 11% in NCDS, 4% in USM1, and <1% in USM2. CD8T proportions measured on average 3% in NCDS, 7% in USM1, and 17% in USM2. CD4T proportions measured on average 18% in NCDS, 12% in USM1, and 13% in USM2. B cell proportions measured on average 7% in NCDS, 5% in USM1, and 4% in USM2. Monocyte proportions measured on average 8% in NCDS, 4% in USM1, and 4% in USM2. This suggests some differences between datasets in terms of cell type composition.

Table 5.1: Sample characteristics

Characteristic	NCDS	USM1	USM2
N	460	835	1,838
Fibrinogen g/l, mean (SD)	3.02 (0.61)	2.91 (0.61)	2.80 (0.58)
C-reactive protein mg/l, mean (SD)	2.48 (7.54)	3.61 (8.00)	3.15 (6.38)
Sex, n (%)			
Male	218 (47%)	348 (42%)	821 (45%)
Female	242 (53%)	487 (58%)	1,017 (55%)
Age, mean (SD)	44.02 (0.12)	58.25 (14.65)	49.78 (15.39)
BMI (kg/m²), mean (SD)	27.16 (4.88)	28.57 (5.27)	27.91 (5.07)
Self-reported smoking status, n (%)			
Never	188 (41%)	381 (46%)	733 (40%)
Former	129 (28%)	330 (40%)	739 (40%)
Current	143 (31%)	124 (15%)	366 (20%)
DNAme-predicted smoking status, n (%)			
Never	312 (68%)	-	1,104 (60%)
Former	3 (0.7%)	-	460 (25%)
Current	145 (32%)	-	274 (15%)
Smoking methylation score, mean (SD)	4.09 (1.24)	3.56 (0.93)	3.56 (0.86)
Educational attainment, n (%)			
Higher qualification	150 (33%)	264 (32%)	771 (42%)
A-level	69 (15%)	177 (21%)	378 (21%)
GCSE	201 (44%)	228 (27%)	453 (25%)
None	40 (8.7%)	166 (20%)	236 (13%)
Cell type composition, mean (SD)			
Granulocytes	0.56 (0.10)	0.69 (0.08)	0.59 (0.08)
CD4T	0.18 (0.06)	0.12 (0.06)	0.13 (0.07)
CD8T	0.03 (0.03)	0.07 (0.04)	0.17 (0.05)
B cell	0.07 (0.03)	0.05 (0.03)	0.04 (0.02)
Monocytes	0.08 (0.02)	0.04 (0.02)	0.04 (0.03)
NK	0.11 (0.05)	0.04 (0.04)	0.00 (0.01)

5.3.1. Correlations

Figure 5.1 shows correlation plots between continuous variables discussed in this chapter within each study. Fibrinogen and CRP are significantly positively correlated in NCDS ($r = 0.16$, $p < 0.0001$), USM1 ($r = 0.47$, $p < 0.0001$) and USM2 ($r = 0.49$, $p < 0.0001$). Fibrinogen was significantly positively correlated with age in UKHLS within both USM1 ($r = 0.26$, $p < 0.0001$) and USM2 ($r = 0.29$, $p < 0.0001$) but CRP was not significantly associated with age.

In NCDS fibrinogen was significantly positively associated with BMI ($r = 0.31$, $p < 0.0001$) and granulocytes ($r = 0.11$, $p = 0.02$), and negatively associated with CD4T ($r = -0.09$, $p < 0.05$) and CD8T ($r = -0.11$, $p < 0.05$) cell type proportions. CRP was significantly positively associated with BMI ($r = 0.14$, $p < 0.01$) and CD4T cell type proportions ($r = 0.13$, $p < 0.01$) and significantly negatively associated with natural killer cell type proportions ($r = -0.10$, $p < 0.05$).

In USM1 fibrinogen was also significantly positively correlated with BMI ($r = 0.21$, $p < 0.0001$), and estimated granulocyte ($r = 0.26$, $p < 0.0001$) and monocyte ($r = 0.16$, $p < 0.0001$) proportions. In USM1 fibrinogen was significantly negatively associated with CD4T ($r = -0.21$, $p < 0.0001$), CD8T ($r = -0.15$, $p < 0.0001$) and B cells ($r = -0.13$, $p < 0.001$), but not natural killer (NK) cell type proportions. As for CRP, inflammatory marker levels were significantly positively correlated with BMI ($r = 0.16$, $p < 0.0001$) and estimated granulocyte ($r = 0.18$, $p < 0.0001$) and monocyte ($r = 0.14$, $p < 0.0001$) proportions, and negatively associated with CD4T ($r = -0.13$, $p < 0.0001$), CD8T ($r = -0.10$, $p < 0.0001$), B cells ($r = -0.11$, $p < 0.001$) and also natural killer cells ($r = -0.07$, $p < 0.05$).

In USM2 fibrinogen was positively correlated with BMI ($r = 0.28$, $p < 0.0001$), estimated granulocyte ($r = 0.19$, $p < 0.0001$) and monocyte ($r = 0.20$, $p < 0.0001$) proportions. In USM2 fibrinogen was negatively associated with CD4T ($r = -0.12$, $p < 0.0001$), CD8T ($r = -0.19$, $p < 0.0001$), B cells ($r = -0.11$, $p < 0.0001$) but not natural killer cell proportions. CRP was significantly positively correlated with BMI ($r = 0.20$, $p < 0.0001$), estimated granulocyte ($r = 0.13$, $p < 0.0001$) and monocyte ($r = 0.15$, $p < 0.0001$) proportions and negatively associated with CD8T ($r = -0.13$, $p < 0.0001$), CD4T ($r = -0.08$, $p < 0.001$), B cell ($r = -0.08$, $p < 0.001$), but not natural killer cells.

In all three datasets granulocytes were significantly negatively associated with estimated CD4T, CD8T, B cell and natural killer cell type proportions. In USM1 and USM2 DNAme-estimated granulocyte proportions were also negatively associated with monocyte cell proportions but not in NCDS. In all three datasets CD4T cell proportions were significantly positively associated with B cells and negatively correlated with monocytes. In USM1 and USM2 CD4T cell were negatively associated with NK cells however in NCDS CD4T cell were positively correlated to NK cells. In NCDS CD4T cells were also significantly negatively associated with CD8T but these were positively correlated in USM2. CD8T cells were significantly negatively correlated to B cells, monocytes and NK cells in NCDS. In USM1 CD8T cells were also significantly negatively associated with monocytes but positively associated with B cells. In USM2 CD8T cells were positively associated with B and NK cells. B cells in NCDS were positively associated with monocytes and NK cells. In USM1 and USM2 B cells were negatively associated with monocytes. Monocytes and NK cells were significantly positively associated in NCDS but negatively associated in USM1.

The smoking methylation score (MS) (McCartney et al., 2018) was significantly negatively associated with CD8T cells in NCDS and USM2. The smoking MS was significantly positively correlated to B and NK

cells. In USM1 the smoking MS was only significantly negatively correlated with NK cells. In USM2 the smoking MS was also significantly negatively associated with monocyte and positively correlated to granulocytes.

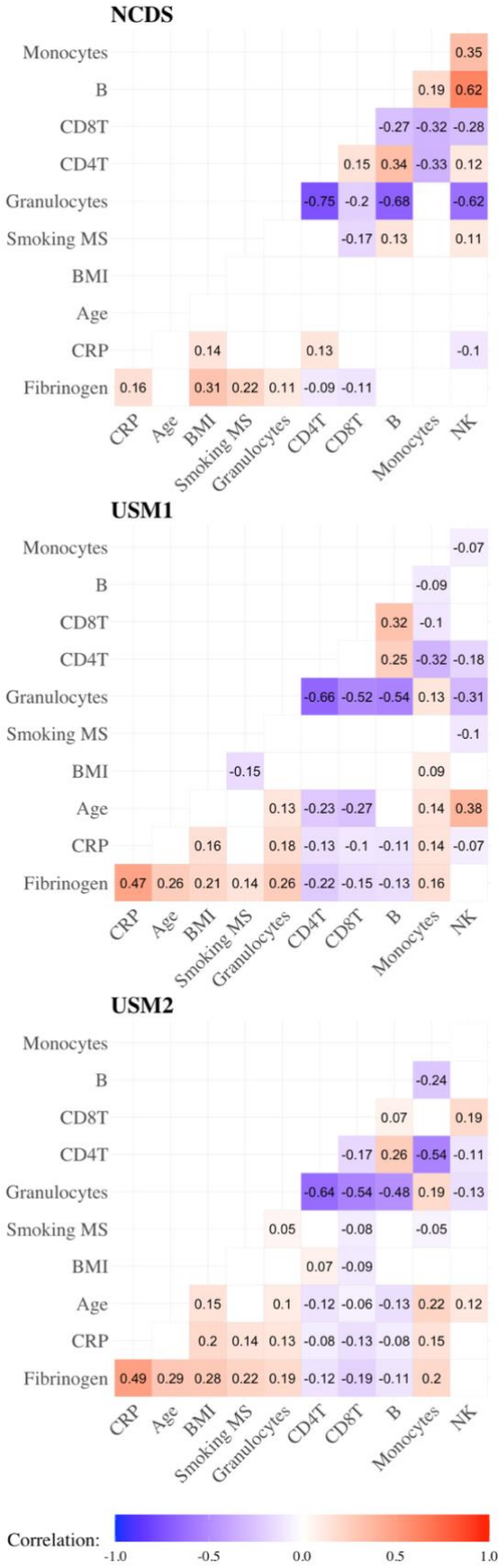

Figure 5.1: Correlation matrix

5.3.2. DMPs

Figures 5.1 and 5.2 show manhattan plots displaying log p-values and chromosomal positions for the top 5000 DMPs associated with inflammatory markers identified from each of 6 EWAS models (Models 5-10). EWAS models 5-7 were used to compare the influence of smoking adjustment, either from self-reported smoking status, DNA methylation-predicted smoking, or a smoking methylation score (McCartney et al, 2018), on epigenetic signatures of inflammation. In models 8-10 further adjustment for educational attainment was carried out to investigate if educational attainment also influenced the epigenetic programming of inflammation and if so whether this differs depending on how smoking is controlled for such as relying self-reports or deriving from DNA methylation. Table 5.2 shows the number of significantly differentially methylated loci associated with inflammatory markers in each EWAS.

In NCDS where no additional covariates were included (Model 1) only 1 differentially methylated probe (DMP) was significantly associated with fibrinogen. This site (cg15194935) is located on chromosome 19 (10,405,955bp) within the *ICAM5* gene. 1 DMP was also significantly associated with fibrinogen after adjustment for cell type composition (Model 2) however this was a different site (cg26416615) located on chromosome 10 (63,751,843bp) within the *ARID5B* gene. 1 DMP was also significantly associated with fibrinogen after further adjustment for age and sex (Model 3) but again this site was different (cg13531315) and located on chromosome 19 (41,354,553bp) within the *CYP2A6* gene. None of these CpG sites remained significant after further adjustment for BMI (Model 4) and none were identified with further adjustment for smoking and education. 0 DMPs were significantly associated with CRP in NCDS (Table 5.2).

In USM1 where no additional covariates were included (Model 1) 25,568 loci were significantly associated with fibrinogen. After adjustment for cell type composition (Model 2), 56 fibrinogen-associated DMPs

(fibDMPs) were identified in USM1. 11 The number of fibDMPs identified in USM1 reduced to 11 after adjustment for age and sex (Model 3) and 8 fibDMPs were identified after adjustment for BMI (Model 4). 5 out of these 8 fibDMPs were also identified without BMI adjustment (Model 3) and this included one CpG site (cg09349128) located on chromosome 22 in the q13.3 region (50,327,986bp), one (cg18608055) located on chromosome 19 (1,130,866bp) within the *SBNO2* gene, one (cg00490406) located on chromosome 1 (159,046,773bp) within the *AIM2* gene, one (cg03067296) located on chromosome 17 (76,274,577bp) within *LOC100996291*, and lastly one site (cg20559943) located on chromosome 4 (185,820,756bp) within an intergenic region. After further adjustment for smoking using self-reports (Model 5), only the chromosome 22q13.3 site remained significant. However, with further adjustment for educational attainment (Model 8) two sites remained significant, one being the chromosome 22q13.3 site and the other being the *AIM2* site located on chromosome 1. When adjusting for smoking using DNA methylation-predicted smoking measures (Models 6-7) both sites were significant and remained significant after adjustment for educational attainment (Models 9-10).

In USM2 57,338 CpG sites were significantly associated with fibrinogen when no covariates are included in the EWAS (Model 1). After adjustment for cell type composition (Model 2) 16,787 fibDMPs were identified and when age and sex were additionally added (Model 3) 213 significant fibDMPs were identified. After further adjustment for BMI (Model 4), 92 loci were significantly associated with fibrinogen and 79 (86%) of these were also identified in the previous model (Model 3). When adjusting for smoking using self-reported smoking status (Model 5) 12 fibDMPs were found, 11 when using DNAme-predicted smoking status (Model 6) and 8 when using the smoking methylation score (Model 7) (McCartney et al., 2018). After further adjustment for educational attainment the number of significant fibDMPs reduced in all three separate smoking adjustments. The loci in question were dependent on the smoking measure used. After controlling for educational attainment, when using self-reported smoking status (Model 8) 9 fibDMPs were found, 8 when using DNAme-predicted smoking status (Model 9) and 7 when using the McCartney

(2018) smoking methylation score (Model 10). 10 unique fibDMPs were identified in total in USM2 after adjustment for all stated covariates (cell composition, age, sex, BMI, educational attainment and either self-reported or DNAme-predicted smoking or the smoking methylation score). 6 out of these 10 unique fibDMPs were identified irrespective of the smoking measure being used. This included, in order of significance, a site (cg26416615) located on chromosome 10 (63,751,843bp) within the *ARID5B* gene, a site (cg09349128) on chromosome 22 in the q13.3 region, a site (cg07252680) within the *SERPINA1* gene on chromosome 14 (94,857,224bp), a site (cg18608055) within the *SBNO2* gene, a site (cg18978030) on chromosome 13 (113,243,542bp) within the *TUBGCP3* gene, and a site (cg24499891) located on chromosome 6 (25,007,637bp) within the *FAM65B* gene.

In USM1 in EWAS models where no additional covariates were included (Model 1) 9,739 loci were significantly associated with CRP. After adjustment for cell type composition (Model 2) 35 CRP associated DMPs (crpDMPs) were identified, and 27 crpDMPs were found in USM1 after adjustment for age and sex (Model 3). 18 crpDMPs were identified in USM1 after adjustment for BMI (Model 4). 16 out of 18 (89%) crpDMPs were also identified without BMI adjustment. After adjustment for smoking, when using self-reported smoking status (Model 5) 13 crpDMPs were found, 9 and 12 when using a smoking methylation score (Model 7). After further adjustment for educational attainment the crpDMPs identified changed and were again dependent on the smoking measures being used. When using self-reported smoking status (Model 8) 9 crpDMPs were found and 11 when using a smoking methylation score (Model 10). 11 unique crpDMPs were identified in USM1 after adjustment for all stated covariates (cell composition, age, sex, BMI, educational attainment and either self-reported smoking or the smoking methylation score). The 2 additional DMPs observed when adjusting for smoking using the McCartney (2018) methylation score were one (cg17501210) on chromosome 17 (17,030,253bp) within the *MPRIP* gene and one (cg23842572) on chromosome 6 (166,970,252bp) within the *RPS6KA2* gene. The 9 crpDMPs identified in USM1 include one site (cg23320029) located on chromosome 3 (171,004,750bp) within the *TNIK* gene, one site

(cg22652934) located on chromosome 21 (36,180,035bp) within the *RUNX1* gene, one site (cg09349128) located on chromosome 22 in the q13.3 region, one site (cg03067296) located within *LOC100996291*, two sites (cg10636246, cg00490406) located on chromosome 1 within the *AIM2* gene, one site (cg15251256) on chromosome 12 (46,101,447bp) in the q13.11 region, one site (cg11551560) located on chromosome 15 (70,528,789bp), and lastly one site (cg26416615) located within the *ARID5B* gene.

In USM2 25,643 crpDMPs were identified without any covariates included in the EWAS (Model 1). After adjustment for cell type composition (Model 2) 1,177 were identified. When age and sex were added (Model 3) 786 significant crpDMPs were identified. After further adjustment for BMI (Model 4) 222 crpDMPs were found and 198 (89%) were also identified in the previous model (Model 3). After adjustment for smoking, when using self-reported smoking status (Model 5) 97 crpDMPs were found, 41 when using DNAme-predicted smoking status (Model 6) and 40 when using a smoking methylation score (Model 7). When controlling for educational attainment the number of crpDMPs identified reduced. This was more pronounced when using self-reported compared to DNAme-predicted smoking status or a methylation score. When using self-reported smoking status (Model 8) 54 crpDMPs were found, 37 when using DNAme-predicted smoking status (Model 9) and 39 when using a smoking methylation score (Model 10). 55 unique fibDMPs were identified in USM2 after adjustment for all stated covariates (cell composition, age, sex, BMI, educational attainment and some measure of smoking). 37 out of these 55 crpDMPs were identified independent of the smoking measure being used. The 37 crpDMPs included 3 sites located in *SOCS3* gene, 2 located in the *AIM2* gene, 2 located in *LOC100996291*, 2 located in the *SNBO2* gene, and 8 were in intergenic regions. Other genes implicated include *ARID5B, C17orf85, CBY3, CD82, CMTM4, DNAJC5B, FAM65B, KIAA0090, KLHL2, LMNB2, LOC645434, MKL2, NACC2, PHOSPHO1*, POC1B, *RNF146, RUNX1, SERPINA1* and *WDR8*.

5.3.3. Inflation

Observed P-values from the 853,973 tested CpG sites were regressed on to expected P-values to obtain estimates of inflation factors for each EWAS model as a measure of genomic control. Table 5.2 shows these inflation estimates. An inflation factor (λ) of 1.00 is considered good and suggests test statistics from the EWAS model are reliable and no inflation has occurred. Across all studies, inflation factors ranged from 10.10 to 0.96 in fibrinogen EWAS and from 5.51 to 0.78 in CRP EWAS. In NCDS inflation factors ranged from 1.66- 0.96 in fibrinogen EWAS and 1.07-0.79 in CRP EWAS. In USM1 λ estimates ranged from 5.22-1.06 in fibrinogen EWAS and from 3.35-1.02 in CRP EWAS. In USM2 λ estimates ranged from 10.10-1.18 in fibrinogen EWAS and from 5.51-1.14 in CRP EWAS.

Test statistics from all EWAS models were overinflated without adjustment for cell type composition, age and sex, and BMI (Models 1-3). After adjusting for these covariates (Model 4) λ estimates based on the fibrinogen EWAS measured 1.06 in NCDS, 1.09 in USM1, and 1.43 in USM2. After further adjusting for self-reported smoking status (Model 5) the inflation factor from the fibrinogen EWAS in NCDS measured 0.96 but did not fall below 1 when using DNA-predicted smoking status (Model 6) nor the smoking methylation score (Model 7) where λ measured 1.03 in both cases. In fibrinogen EWAS after adjusting for self-reported smoking status the inflation factor in USM1 measured 1.09 and 1.25 in USM2. When using self-reported smoking status in conjunction with educational attainment (Model 8) the inflation in the NCDS fibrinogen EWAS measured 1.01, and when using DNAme-predicted smoking status (Model 9) λ measured 1.06 and 1.05 when using the smoking methylation score (Model 10). This suggests that test statistics from EWAS may be less inflated when using methylation-based measures of smoking, especially when educational attainment is not known.

Inflation estimates based on CRP EWAS after adjustment for cell composition, age, sex, and BMI (Model 4) measured 1.11 in USM1 and 1.30 in USM2. In NCDS λ measured 0.84 and suggests the CRP EWAS carried out in NCDS were underpowered once BMI was adjusted for. In CRP EWAS λ estimates when additionally controlling for self-reported smoking (Model 5), inflation factors measured 0.79 in NCDS, 1.24 in USM1, and 1.36 in USM2. When adjusting for DNAme-predicted smoking status instead (Model 6), CRP EWAS λ estimates measured 0.79 in NCDS, and 1.15 in USM2. When adjusting for smoking using a smoking methylation score (Model 7) CRP EWAS λ estimates measured 0.79 in NCDS, 1.03 in USM1, and 1.19 in USM2. This suggests that test statistics are less inflated when using methylation-based measures of smoking compared to self-reports.

In CRP EWAS when further adjusting for educational attainment and smoking using self-reports (Model 8) estimates of inflation (λ) measured 0.81 in NCDS, 1.02 in USM1, and 1.15 in USM2. When using DNAme-predicted smoking status (Model 9) λ estimates measured 0.80 in NCDS, and 1.14 in USM2. When using a smoking methylation score, CRP EWAS λ estimates measured 0.79 in NCDS, 1.02 in USM1, and 1.18 in USM2. This suggests adjustment for socioeconomic factors like education in EWAS models may improve the reliability of test statistics and identified DMPs by reducing inflation. Test statistics from all EWAS models in USM2 were overinflated and this likely relates to the large sample size and suggests more covariates should be considered.

Table 5.2: Number of differentially methylated probes (DMPs) and inflation factors (λ) identified in each EWAS model

	NCDS (N = 460)		USM1 (N = 835)		USM2 (N = 1,838)	
Model	**DMPs**	**λ**	**DMPs**	**λ**	**DMPs**	**λ**
Fibrinogen						
~ Fibrinogen (**Model 1**)	1	1.66	25,568	5.22	57,338	5.22
Model 1 + Cell type estimates (**Model 2**)	1	1.14	56	1.58	16,787	1.58
Model 2 + Age + Sex (**Model 3**)	1	1.06	11	1.13	213	1.13
Model 3 + BMI (**Model 4**)	0	1.06	8	1.09	92	1.09
Model 4 + Self-reported SSt (**Model 5**)	0	0.96	1	1.09	12	1.09
Model 4 + DNAme SSt (**Model 6**)	0	1.03	-	-	11	1.07
Model 4 + Smoking MS (**Model 7**)	0	1.03	2	1.09	8	1.09
Model 5 + Education (**Model 8**)	0	1.01	2	1.06	9	1.06
Model 6 + Education (**Model 9**)	0	1.06	-	-	8	1.07
Model 7 + Education (**Model 10**)	0	1.05	2	1.09	7	1.09
C-reactive protein						
~ CRP (**Model 1**)	0	1.07	9,739	3.35	25,643	3.35
Model 1 + Cell type estimates (**Model 2**)	0	1.07	35	1.29	1,177	1.29
Model 2 + Age + Sex (**Model 3**)	0	1.07	27	1.16	786	1.16
Model 3 + BMI (**Model 4**)	0	0.84	18	1.11	222	1.11
Model 4 + Self-reported SSt (**Model 5**)	0	0.79	13	1.24	97	1.24
Model 4 + DNAme SSt (**Model 6**)	0	0.79	-	-	41	1.04
Model 4 + Smoking MS (**Model 7**)	0	0.79	12	1.03	40	1.03
Model 5 + Education (**Model 8**)	0	0.81	9	1.02	54	1.02
Model 6 + Education (**Model 9**)	0	0.80	-	-	37	1.02
Model 7 + Education (**Model 10**)	0	0.79	11	1.02	39	1.02

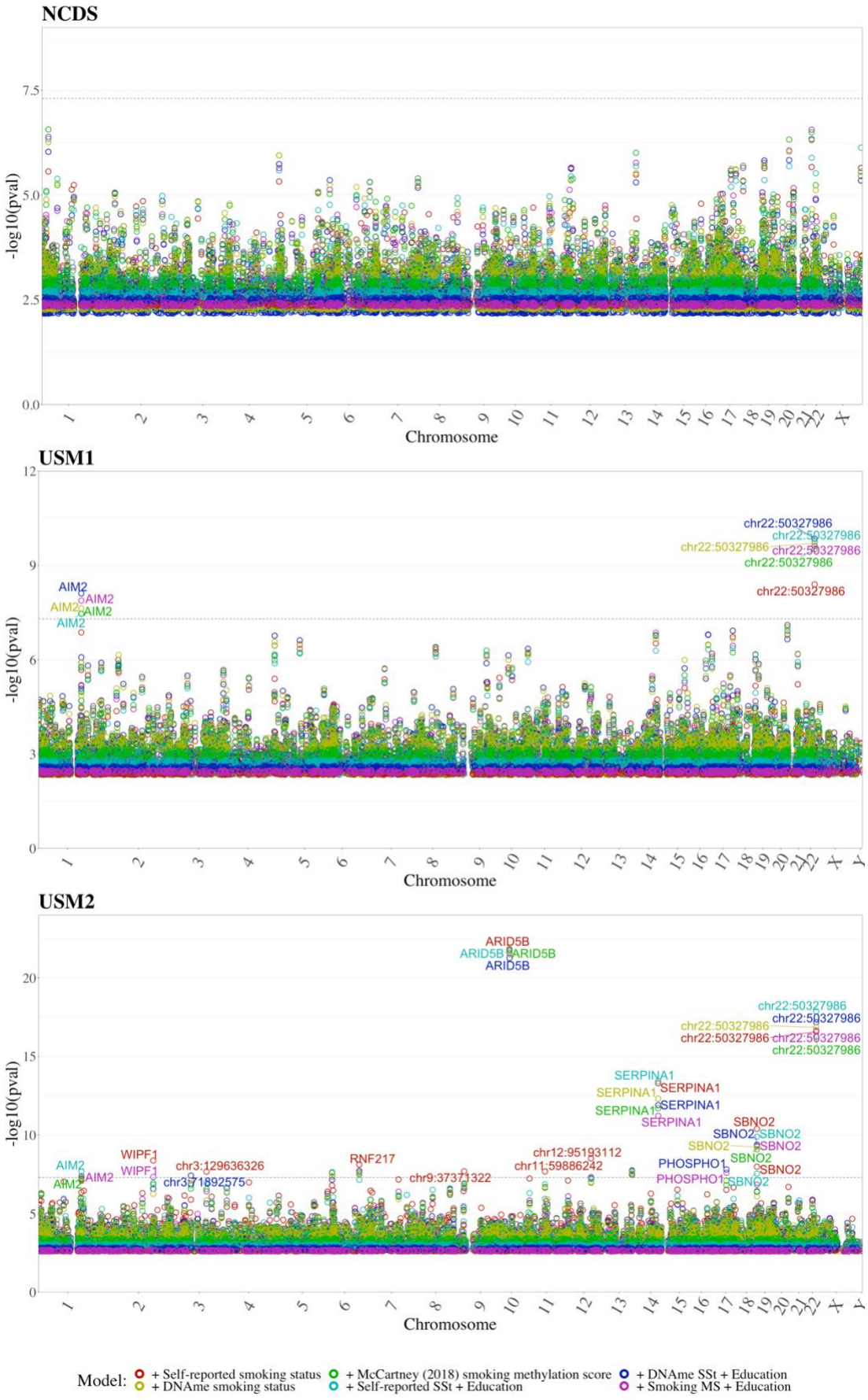

Figure 5.2: Manhattan plots showing top 5000 fibrinogen associated DMPs

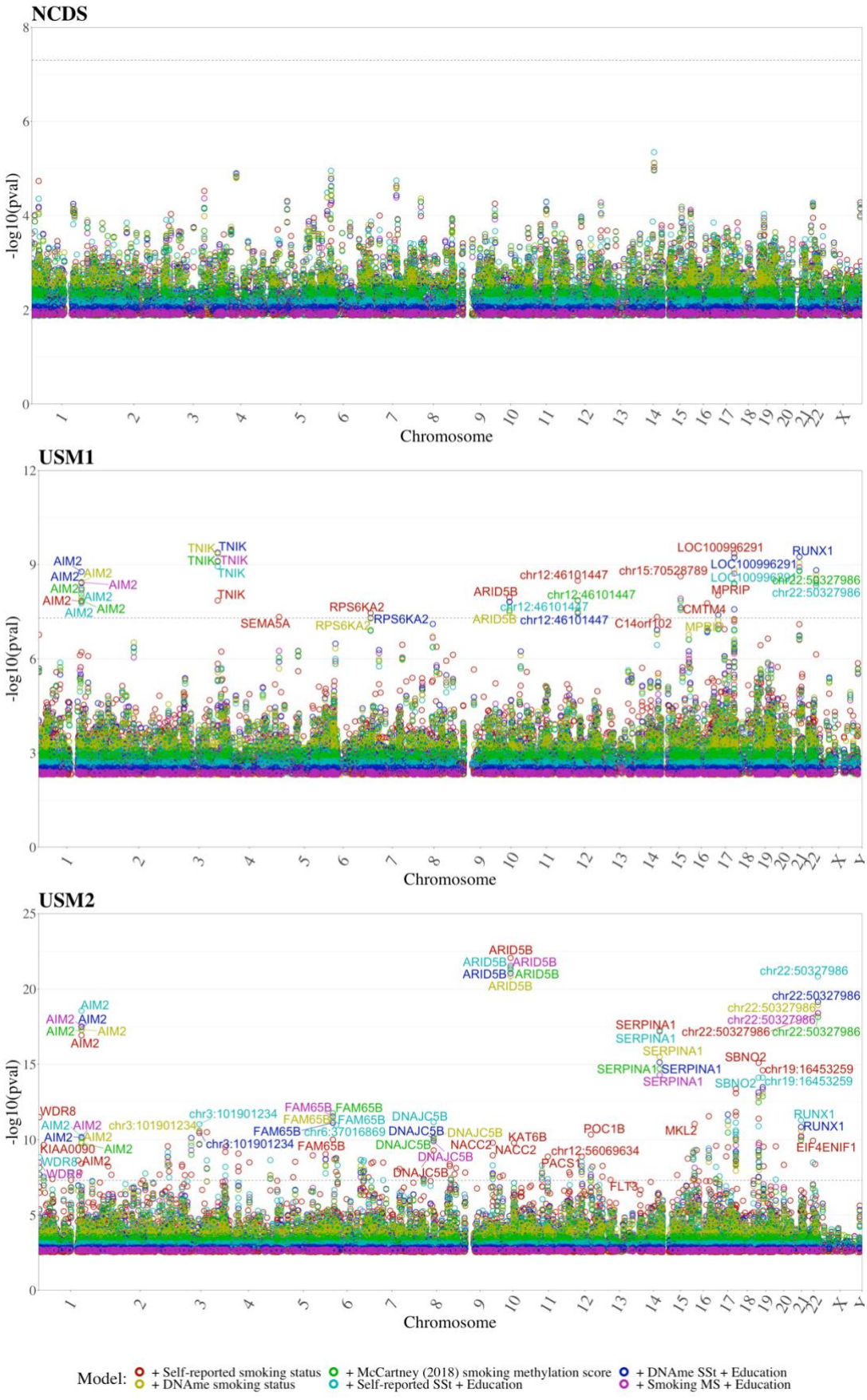

Figure 5.3: Manhattan plots showing top 5000 CRP associated DMPs

5.3.4. Meta-analysis across studies

Figures 5.4 and 5.5 show t statistics related to the top CpG sites of interest when adjusting for smoking (and age, sex, BMI and cell type composition) with and without educational attainment (Models 5-10). Although many inflammation-related sites that reached genome-wide significance in USM2 were not in NCDS and USM1, this figure shows that many identified DMPs show the same direction across all studies, giving strength to these findings. A total of 62 unique CpGs were identified in fibrinogen and CRP EWAS after adjustment for all stated covariates (Models 8-10). When using self-reported smoking status and without controlling for education (Model 5) 22 out of 62 (35%) top inflammation related DMPs showed the same direction of association, meaning a positive or negative t statistic, in NCDS, USM1 and USM2. However, when additionally controlling for education with self-reported smoking (Model 8), 24 out of 62 (39%) DMPs shared the same direct across all three datasets. When using the smoking methylation score (McCartney et al., 2018) without education in the model (Model 7) the number of DMPs in the same direction in NCDS, USM1, and USM2 was 18 and this number did not change when adding education as a covariate (Model 10). In EWAS where DNAme-predicted smoking status is used to control for smoking (Model 6), 41 DMPs shared the same direction of association in NCDS and USM2, and when adding educational attainment this reduced to 37 (Model 9). (Supplementary Tables 12 and 13).

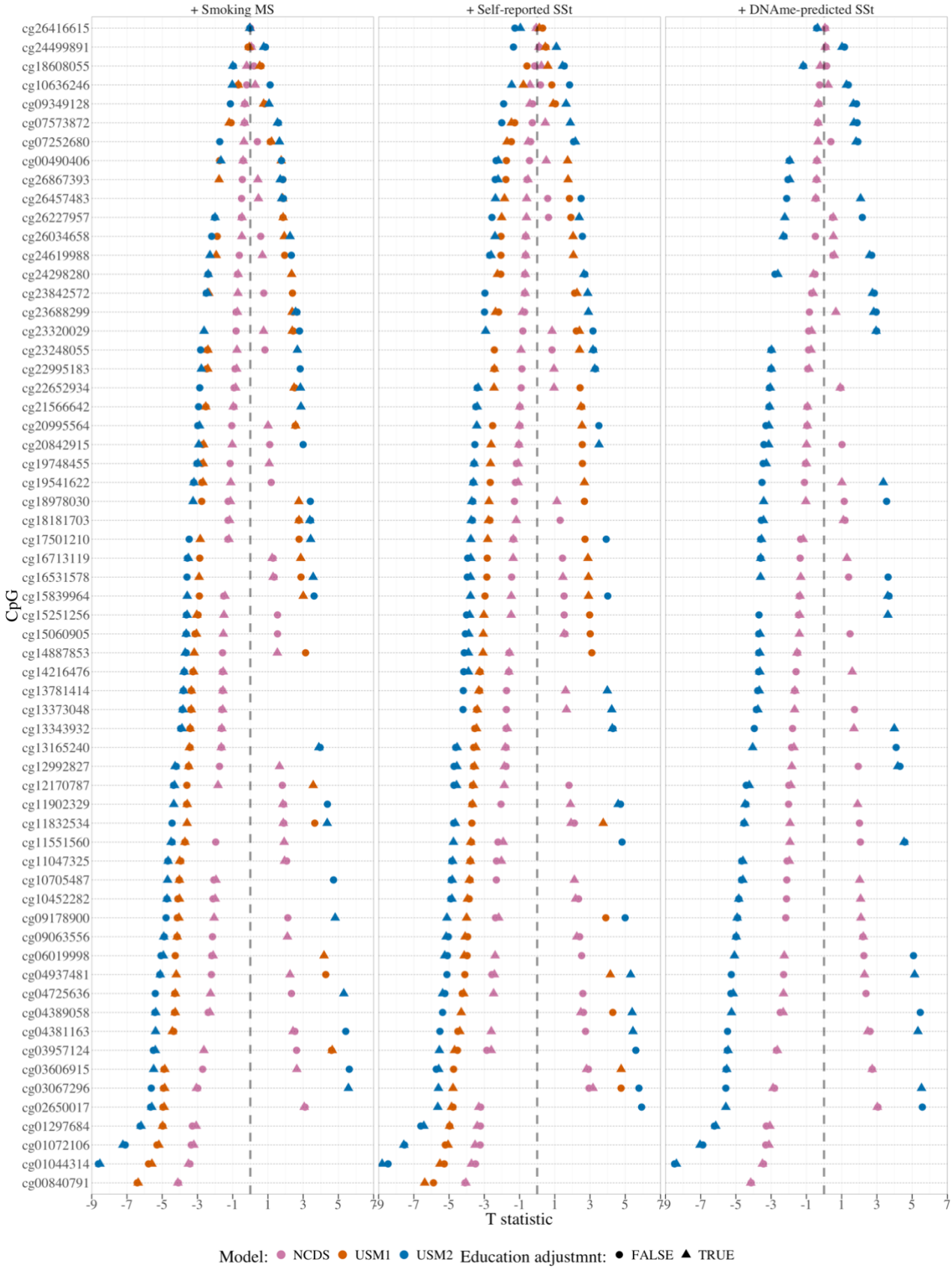

Figure 5.4: Plot showing T statistics for the top 62 CpG sites associated with inflammatory markers across datasets in fibrinogen EWAS after adjustment for smoking and other covariates (Models 5-7) and with or without educational attainment (Models 8-10) included

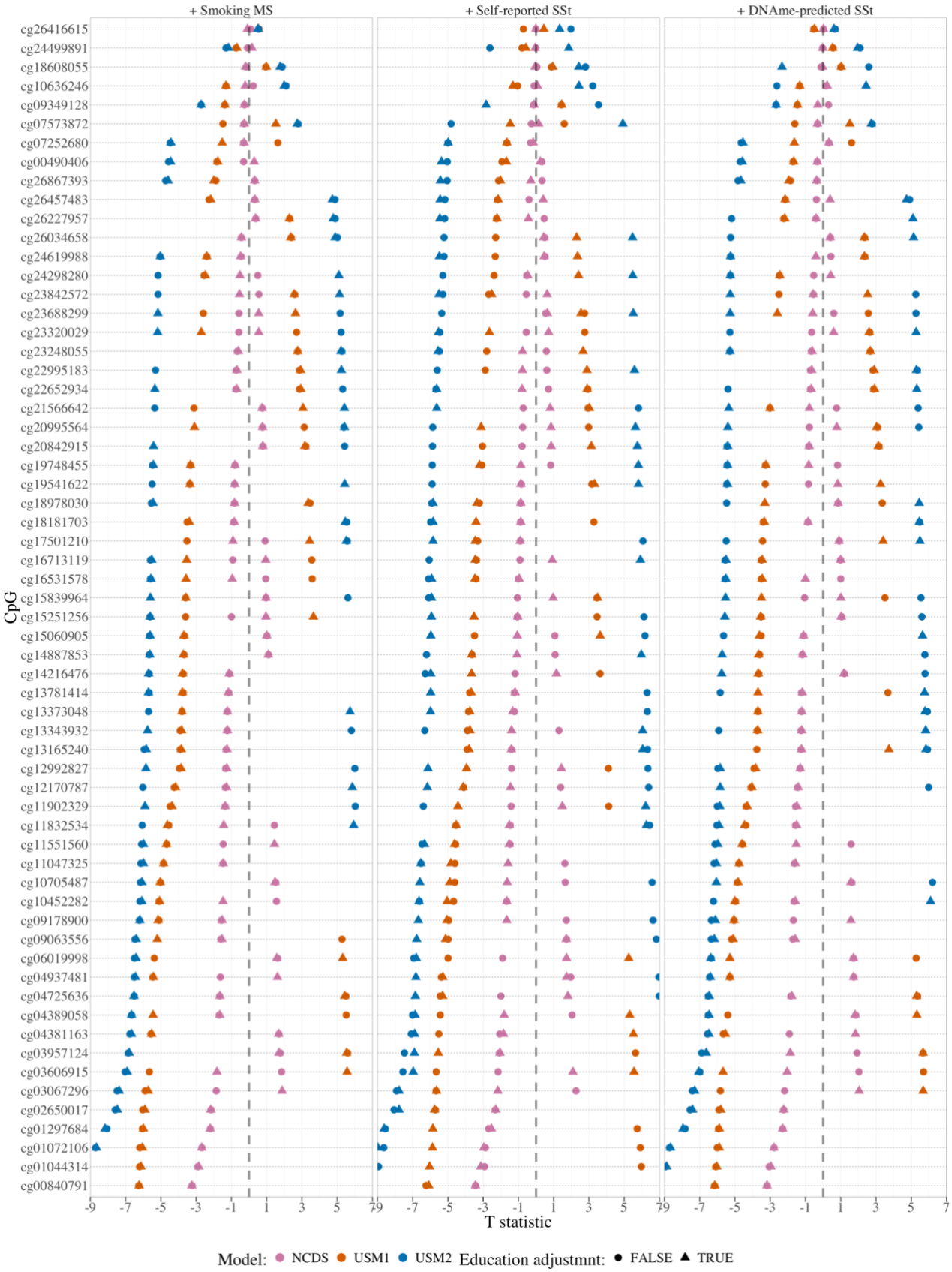

Figure 5.5: Plot showing T statistics for the top 62 CpG sites associated with inflammatory markers across datasets in CRP EWAS after adjustment for smoking and other covariates (Models 5-7) and with or without educational attainment (Models 8-10) included

5.3.5. Gene enrichment

A total of 50 fibrinogen-related and 123 CRP-related coding genes were implicated in epigenetic differences associated with inflammation while adjusting for cell composition, age, sex and BMI (Model 4) including a total of 140 unique inflammation-related genes. After further adjustment for smoking, a total of 9 fibrinogen-related and 67 CRP-related protein coding genes were implicated, including 70 unique inflammation related genes. After further adjustment for education, a total of 7 fibrinogen-related and 37 CRP-related coding genes were implicated.

STRING (https://string-db.org/) analysis was used to identify any protein-protein interaction networks within these three sets of genes and carry out pathway and functional enrichment (Figure 6.4). A total of 154 genes were investigated in total and 145 of these genes were available in the STRING database. Some of the genes not included are non-protein coding genes such as include LOC100996291, LOC101060019, LOXL1-AS1, MIR4505, LINC00299, LOC102724020, LOC645434, MIR646HG. For others the protein name was different from the gene name and as such renamed in analysis and these include C14orf102 renamed as NRDE2, APOB48R as APOBR, KIAA0090 as EMC1, C14orf43 as ELMSAN1, C19orf76 as RIIAD1, TMEM49 as VMP1, WDR8 as WRAP73, C5orf62 as CAMP (Figure 5.6).

Within the network of inflammation-related genes identified without adjustment for smoking and education, consisting of 131 protein coding genes, 67 edges were observed. The expected number of edges was 41 meaning this network had significantly more interactions than expected (PPI enrichment p-value: 0.000163). There were no significant pathway enrichments observed in all tested categories, including Biological Process (Gene Ontology), Molecular Function (Gene Ontology), Cellular Component (Gene Ontology), KEGG Pathways, Reactome Pathways, WikiPathways, Disease-gene associations

(DISEASES), and Protein Domains and features (Pfam, InterPro, SMART). However, there were some functional enrichments within this network. 2 selected terms were enriched and referred to as 'Chromosomal rearrangement' and 'Phosphoprotein'. 8 publications from PubMed were significantly enriched. These publications largely related to smoking and DNA methylation. The 8 publications are 'Tobacco smoking leads to extensive genome-wide changes in DNA methylation' (Zeilinger et al., 2013), 'DNA Methylation Trajectories During Pregnancy' (Gruzieva et al., 2009), 'Smoking-Related DNA Methylation is Associated with DNA Methylation Phenotypic Age Acceleration: The Veterans Affairs Normative Aging Study' (Yang et al., 2019), 'The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes' (Gao et al., 2017), 'Identification of DNA methylation changes in new-borns related to maternal smoking during pregnancy' (Markunas et al., 2014), '450K epigenome-wide scan identifies differential DNA methylation in new-borns related to maternal smoking during pregnancy' (Joubert et al., 2012), 'CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study' (Harlid et al., 2014), and 'The Influences of Genetic and Environmental Factors on Methylome-wide Association Studies for Human Diseases' (Sun, 2014).

The network consisting of the 67 gene products that significantly associated with inflammation after smoking adjustment showed 16 edges, but this was not significantly more than expected (PPI enrichment p-value: 0.255). The network consisting of the 37 gene products significantly associated with inflammation after further adjustment for educational attainment also did not have significantly more edges than expected (PPI enrichment p-value: 0.649). In both gene networks no significant functional enrichment was found (Figure 5.6). This suggests that the bulk of DNA methylation differences observed with inflammatory marker levels are predominately driven by biological processes related to BMI and smoking.

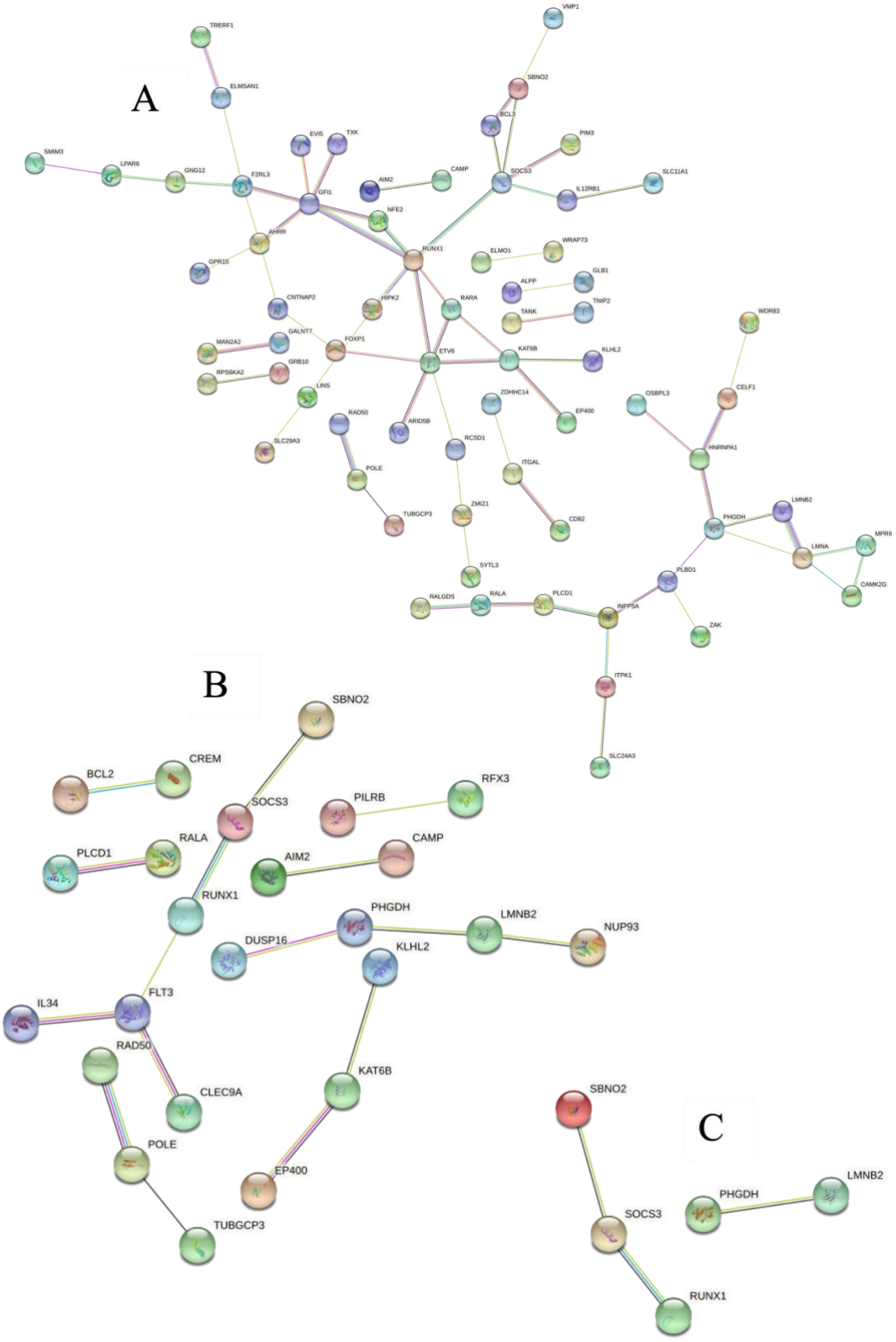

Figure 5.6: STRING analysis showing known interactions between genes differentially methylated with inflammatory markers after adjustment for: A) cell composition, age, sex and BMI, B) smoking, C) educational attainment

5.4. Discussion

This chapter aimed to investigate the relationship between DNA methylation and two inflammatory markers, fibrinogen and CRP, in 460 participants from NCDS and 2673 participants from two UKHLS sub-samples (USM1 = 835, USM2 = 1,838). It also aimed to better understand how adjustment for health-related factors impacts inflammation-related epigenetic signals. These factors included cell type composition, age, sex, smoking, BMI and educational attainment. Self-reported and DNA methylation-based smoking status, and a smoking methylation score (McCartney et al., 2018), were also compared to see if different adjustments for smoking alter epigenetic signatures of inflammation. This chapter was able to replicate findings from previous published EWAS studies (Ligthart et al., 2016; Miller et al., 2018). It also added to the literature of known loci significantly associated with inflammation through the identification of novel loci such as the CpG site on chromosome 22 in the intergenic q13.3 region that was significantly associated with both fibrinogen and CRP in USM1 and USM2. Many genes were implicated in both EWAS of fibrinogen and CRP and these genes include *AIM2*, *SBNO2*, *TMEM49*, *ARID5B*, *RPS6KA2* and *SERPINA1*. This shows that fibrinogen and CRP share similarities in their relationship to DNA methylation and this may relate to their shared function as acute-phase proteins. All mentioned genes have roles within inflammation.

AIM2 codes for a protein found in hematopoietic cells involved in the innate immune response by recognizing cytosolic double-stranded DNA and inducing caspase-1-activating inflammasome formation in macrophages. Upon binding to DNA, it is thought that *AIM2* undergoes oligomerization to associate with *PYCARD* initiating the recruitment of a caspase-1 precursor and processing of interleukin-1 beta and interleukin-18. *AIM2* is involved in the detection of cytosolic dsDNA from viral and bacterial origin in a non-sequence-specific manner and can also trigger PYCARD-dependent, caspase-1-independent cell death.

AIM2 may also act as a tumour suppressor by repressing NF-kappa-B transcriptional activity (Man et al., 2016). *SBNO2* acts as a transcriptional coregulator with both coactivator and corepressor functions. This protein inhibits the DCSTAMP-repressive activity of *TAL1*, enhancing the access of the transcription factor MITF to the DC-STAMP promoter in osteoclast. As such *SBNO2* plays a role in bone homeostasis. It is also thought to be involved in the transcriptional repression of NF-kappaB in macrophages as a regulator in the proinflammatory cascade (El Kasmi et al., 2007). *TMEM49*, also known as *VMP1*, codes for a multi-spanning membrane protein in the endoplasmic reticulum (ER) required for autophagosome formation. This gene controls the disassociation of autophagosomes from the ER through its interaction with *BECN1* and *ATP2A2* and modulates ER contacts with lipid droplets, mitochondria and endosomes. *TMEM49* is required for lipoprotein secretion, cell-cell adhesion, and cell junctions. Upon stress such as with bacterial and viral infection, *TMEM49* promotes the formation of cytoplasmic vacuoles followed by cell death. Recently *TMEM49* has been identified as a host factor required for infection by all flaviviruses tested such as Zika virus and Yellow fever virus (Hoffmann et al., 2021).

ARID5B codes a transcription coactivator that plays a key role in adipogenesis and liver development by regulating the transcription of target genes involved in adipogenesis. This is a DNA binding protein involved in forming the H3K9Me2 demethylase complex. DNAme at this region is inversely associated with ARID8B expression and atherosclerosis, the underlying pathology of CHD, and knockdown of this gene reduces expression of genes in atherosclerosis-related inflammatory pathways. Hence *ARID5B* expression acts as a biomarker of CHD and links DNAme and chromatin function. *ARID5B* can also dysregulate immunometabolism towards chronic inflammatory phenotype. It acts by forming a complex with phosphorylated *PHF2*, which mediates demethylation at Lys-336 and targets the PHF2-ARID5B complex to promoters. Then *PHF2* mediates demethylation of dimethylated 'Lys-9' of histone H3 (H3K9me2) leading t transcription activation of target genes. *ARID5B* may also play a role in adipogenesis through regulation of triglyceride metabolism in adipocytes thereby regulating expression of adipogenic

genes. Overexpression leads to induction of smooth muscle marker genes, suggesting that it may also act as a regulator of smooth muscle cell differentiation and proliferation (Baba et al., 2011).

RPS6KA2 codes a serine/threonine-protein kinase that acts downstream of ERK (MAPK1/ERK2 and MAPK3/ERK1) signalling and mediates mitogenic and stress-induced activation of transcription factors, regulates translation, and mediates cellular proliferation, survival, and differentiation (Zhao et al., 1995). *SERPINA1*, or alpha-1-antitrypsin (*A1AT*), is an inhibitor of serine proteases where its primary target is elastase which it also inactivates, and it also has a moderate affinity for plasmin and thrombin. *SERPINA1* irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (Guttman et al., 2015). The aberrant form is found in the plasma of chronic smokers and persists even up to 10 years after smoking is ceased. Multiple Long Intergenic Non-Protein Coding RNA genes were also implicated in the epigenetic programming of inflammation. The most consistent locus, found to be significantly associated with fibrinogen and CRP across almost all models in UKHLS, is chr22;50327986 (chr 22q13.33). Some neighbouring genes in this region include *WNT7B*: Wingless-type MMTV integration site family member 7B, *SHANK3*: SH3 and multiple ankyrin repeat domains 3, *SULT4A1*: sulfotransferase family 4A, member 1, PARVB:parvin beta.

Some replicated genes (Marzi et al, 2016) were only found in CRP EWAS and not significantly associated with fibrinogen. This includes the *SOCS3* and *TNIK* genes. The SOCS family of proteins form part of a classical negative feedback system that regulates cytokine signal transduction. *SOCS3* is involved in negative regulation of cytokines that signal through the JAK/STAT pathway. Binding to *JAK2* inhibits its kinase activity and regulates IL-6 signalling. It has been shown to suppress fetal liver erythropoiesis and

regulate the onset and maintenance of allergic responses mediated by T-helper type 2 cells (Rottenberg and Carow, 2014). *TNIK* is another serine/threonine kinase that acts as an essential activator of the Wnt signalling pathway and is required to activate the expression of Wnt target genes. This gene may play a role in cytoskeletal rearrangements and cell spreading and the response to environmental stress (Fu et al., 1999). Other genes were implicated in fibrinogen EWAS but were not significantly associated with CRP, such as the *ACOX1* gene. This is the first enzyme of the fatty acid beta-oxidation pathway (Oaxaca-Castillo et al., 2007).

All genes discussed show biologically plausible functions that relate to inflammation. This shows that DNA methylation can reflect changes in commonly measured inflammatory markers such as fibrinogen and CRP. However, the loci that are found to be significantly associated with inflammation are highly dependent on not only the sample size but also the covariates included. The larger sample size in USM2 meant the statistical power to detect significant differences associated with inflammation was greater in this dataset compared to USM1 and NCDS. This is likely why no sites were significantly associated with inflammatory markers in NCDS. With this said, more power also increases the chance of identifying spurious associations when important inflammation-related covariates are not controlled. This suggests that when carrying out EWAS in large samples further adjustment for covariates outside of the usual cell composition measures, age, sex, smoking status and BMI may be necessary.

It is difficult to disentangle the relationship of DNAme and inflammation with cell type composition as inflammation is driven by an orchestra of cellular changes where more mononuclear cells become present at the site of inflammation. In EWAS of phenotypes with multiple known drivers, a sensitivity analysis should perhaps be carried out where adjustment for various covariates is explored to ensure that significant associations are in fact associated with the phenotype of interest and not some other related risk factor. On

the other hand, it is also important to carefully consider the relationship between the covariates included in EWAS to avoid issues with collinearity. This may over adjust the model, especially with low sample numbers, and this may mean potentially real signals are missed. This is highlighted by differences observed between EWAS with or without adjustment for educational attainment. Educational gradients in both inflammation (Davillas et al., 2017) and smoking (Escobedo and Peddicord, 1996) have been observed. Fewer CpG sites were significantly associated with inflammation when additionally adjusting for education. However, this appeared to be more noticeable when self-reported measures of smoking are used. This suggests that using DNA methylation-based biomarkers to control for smoking in EWAS may reduce noise in epigenetic differences by inflammation that is caused by the interplay with social factors.

Clonal haematopoiesis (CH) describes the process where hematopoietic stem cells start making cells with the same genetic mutation and is highly prevalent in older people. Acquired leukemic mutations that have a proliferative advantage to these cells can accelerate atherosclerosis and increase IL-6/IL-1β expression and recent findings have shown that genetic IL-6 signalling deficiency influences incident CVD events (Bick et al., 2020). One common mutation that increases JAK-STAT signalling leads to increased expression of *AIM2*, oxidative DNA damage and DNA replication stress in atherosclerotic lesions in mice that express Jak2VF selectively in macrophages. This genetic mutation appears to activate the *AIM2* inflammasome where Aim2 deficiency reduced atherosclerosis (Fidler et al., 2021). Alterations to DNA methylation are associated with airway macrophage differentiation and lung fibrosis phenotypes. In this *ARID5B* was implicated using H3K4me1 chromatin immunoprecipitation sequencing. *ARID5B* DNAme status was shown to mark monocyte-to-macrophage and airway macrophage development and CpGs in this gene overlap with DHS and H3K4me1 enrichment (McErlean et al., 2021). It has also been linked with metabolism in hepatocytes and natural killer cells (Baba et al., 2011) and adipogenesis (Claussnitzer et al., 2015).

Often sociodemographic factors are not controlled for in EWAS however genes related to intracellular trafficking/protein quality control, such as co-chaperone activities involved in glucocorticoid signaling, deubiquitination involved in beta-2 adrenergic receptor recycling and anionic amino-acid transport, have been implicated. Here DNA methylation in the *FKBP5* gene has been implicated in complex diseases and relates to aging and stress-related phenotypes and can increase peripheral inflammation (Nabais et al., 2021). DNA methylation has also been found to be associated with early-life adversity in youth (Sumner et al., 2022). Little was found through gene set enrichment, and this may relate to bias in gene set enrichment array calculations. Specific methods have been developed to overcome this and take into consideration the differential number of probes between genes (Maksimovic et al., 2021).

5.5. Conclusion

In this chapter epigenetic signatures of two inflammatory markers, fibrinogen and C-reactive protein, were investigated. This was carried out using DNA methylation profiles from two population studies, including the National Childhood Development Study (NCDS) and the UK Household Longitudinal Study (UKHLS). The impact of cell type composition, sex, age, BMI, smoking and educational attainment on epigenetic signatures of inflammation was investigated. This showed that all factors contribute in some way to epigenetic signatures of inflammation and factors such as cell type composition and BMI contribute the largest effect on DNA methylation by inflammation. Social factors such as educational attainment do still appear to influence DNA methylation through inflammatory processes. This suggests EWAS should take sociodemographic characteristics into account to avoid spurious associations. Self-reported smoking status was compared to DNAme-predicted measures and a smoking methylation score (McCartney et al, 2018) in their adjustment for smoking in inflammatory-related methylation changes. Findings suggest that

methylation-based smoking measures may more closely relate to inflammatory marker measures compared to self-reports. It also suggested that epigenetic differences in inflammation may be less influenced by education when using methylation-based smoking measures. This all goes to show that the epigenetic landscape of inflammation is far-reaching, complex and influenced by a plethora of factors that should be considered when identifying significant differences in DNA methylation driven by inflammation.

6. Conclusion and limitations

In this thesis different methods for estimating smoking from DNA methylation were compared in three independent datasets, one from NCDS and two from UKHLS. All methylation-based biomarkers of smoking could reliably distinguish smokers from non-smokers however methods differed in their explanation of past smoking and smoking histories such as pack years or cessation years. One CpG site located in the *AHRR* gene was the most used locus in estimating smoking from DNAme and contributed the largest effect sizes. The addition of extra CpG sites only slightly improved accurate estimation of smoking and this was not the case in all methods. Overall, the best performing biomarker of smoking was by McCartney et al. (2018). However, an issue with this methylation score is that it is unclear what threshold values dictate smoking from non-smoking and these thresholds would likely vary in different populations. The 'smokp SSt method' instead uses three separate methylation scores which are converted to log odds and used to classify samples into current, former, or never smokers. The next stage of this study was to investigate what factors influenced agreement between self-reported and DNAme-predicted smoking status. This showed that social differences such as educational attainment and socioeconomic class appear to influence congruence between self-reports and methylation-based measures of smoking. These factors had a stronger impact on positive cases than negative cases of smoking. Self-reported and DNAme-predicted smoking status was also compared in their association with inflammation. This suggested that DNAme-based smoking status more closely reflect levels of inflammatory markers fibrinogen and CRP than self-reports. It also suggests that adjustment for smoking using DNAme-based smoking status may change known social gradients in inflammation. Lastly, epigenetic signatures of inflammation were examined through various EWAS. This shows that cell type composition, sex, age, BMI, and smoking all contribute to inflammatory load and in turn influence DNA methylation changes associated with fibrinogen or CRP. However, these variables are related to one another which could complicate the interpretability and reproducibility of findings. The addition of educational attainment further influenced epigenetic

signatures of smoking and suggests that social differences in health should be considered in epigenome-wide association studies to prevent spurious findings. Adjustment for smoking in inflammation EWAS was carried out using self-reported or DNAme-based measures of smoking. There were differences in the number and location of inflammation associated CpG sites depending on whether smoking is measure via self-reports or DNA methylation. Findings suggest that when sociodemographic data is not available controlling for smoking in EWAS using DNAme-predicted values could help reduce significant differences in DNA methylation observed in inflammation caused by educational differences. Taken together this thesis has looked closely at the interplay between smoking, education, inflammation and DNA methylation. Given that health and disease are influenced by genetics, health behaviours and the social environment this is just one example used to demonstrate the importance of interdisciplinary research in epidemiology.

7. References

Abbas, A.K., Lichtman, A. and Pillai, S. (2019) Basic Immunology: Functions and Disorders of the Immune System, 6e: Sae-E-Book. Elsevier India.

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994) Socioeconomic status and health: The challenge of the gradient. American Psychologist, 49(1), 15-24.

Al Muftah, W.A., Al-Shafai, M., Zaghlool, S.B., Visconti, A., Tsai, P.C., Kumar, P., Spector, T., Bell, J., Falchi, M. and Suhre, K. (2016) Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clinical epigenetics, 8(1), pp.1-10.

Albrecht, W., Santis, M.D. and Dossenbach-Glaninger, A. (2004) Testicular tumour markers: Cornerstones in the management of malignant germ cell tumours/Hoden-Tumor-marker: Eckpfeiler in der Behandlung maligner Keimzelltumoren. LaboratoriumsMedizin, 28(2), 109-115.

Alves, J., Perelman, J., Ramos, E., & Kunst, A. E. (2018). The emergence of socioeconomic inequalities in smoking over the life-course. Revue d'Épidémiologie et de Santé Publique, 66, S409.

Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Calvez-Kelm, L., Kaaks, R., Barrdahl, M., Boeing, H., Aleksandrova, K., Trichopoulou, A., Lagiou, P., Naska, A., Palli, D., Krogh, V., Polidoro, S., Tumino, R., Panico, S., Bueno-de-Mesquita, B., Peeters, P., QuiroÃÅs, J., Navarro, C., Ardanaz, E., Dorronsoro, M., Key, T., Vineis, P., Murphy, N., Riboli, E., Romieu, I. and Herceg, Z. (2016) Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics, 8(5), 599-618.

Andersen, A.M., Philibert, R.A., Gibbons, F.X., Simons, R.L. and Long, J. (2017) Accuracy and utility of an epigenetic biomarker for smoking in populations with varying rates of false self-report. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 174(6), pp.641-650.

Anreiter, I. and Sokolowski, M.B. (2018) Deciphering pleiotropy: How complex genes regulate behavior. Communicative & integrative biology, 11(2), pp.1-4.

Arpon, A., Riezu-Boj, J.I., Milagro, F.I., Marti, A., Razquin, C., Martínez-González, M.A., Corella, D., Estruch, R., Casas, R., Fitó, M. and Ros, E. (2016) Adherence to Mediterranean diet is associated with methylation changes in inflammation-related genes in peripheral blood cells. Journal of physiology and biochemistry, 73(3), pp.445-455.

Baba, A., Ohtake, F., Okuno, Y., Yokota, K., Okada, M., Imai, Y., Ni, M., Meyer, C.A., Igarashi, K., Kanno, J. and Brown, M. (2011) PKA-dependent regulation of the histone lysine demethylase complex PHF2–ARID5B. Nature cell biology, 13(6), pp.668-675.

Bakhru, A. and Erlinger, T.P. (2005). Smoking cessation and cardiovascular disease risk factors: results from the Third National Health and Nutrition Examination Survey. PLoS medicine, 2(6), p.e160.

Berk, M., Williams, L.J., Jacka, F.N., O'Neil, A., Pasco, J.A., Moylan, S., Allen, N.B., Stuart, A.L., Hayley, A.C., Byrne, M.L. and Maes, M. (2013) So depression is an inflammatory disease, but where does the inflammation come from? BMC medicine, 11(1), pp.1-16.

Bestor, T. (2000) The DNA methyltransferases of mammals. Human Molecular Genetics, 9(16), 2395-2402.

Bibikova, M. (2016) DNA Methylation Microarrays. In Epigenomics in Health and Disease (pp. 19-46). Academic Press.

Bick, A.G., Pirruccello, J.P., Griffin, G.K., Gupta, N., Gabriel, S., Saleheen, D., Libby, P., Kathiresan, S. and Natarajan, P. (2020). Genetic interleukin 6 signaling deficiency attenuates cardiovascular risk in clonal hematopoiesis. Circulation, 141(2), pp.124-131.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L. and Fan, J.B. (2011) High density DNA methylation array with single CpG site resolution. Genomics, 98(4), 288-295.

Birbrair, A., & Frenette, P. S. (2016). Niche heterogeneity in the bone marrow. Annals of the New York Academy of Sciences, 1370(1), 82.

Bochtler, M., Kolano, A. and Xu, G. (2016) DNA demethylation pathways: Additional players and regulators. BioEssays. 39(1), 1600178.

Bock, C. (2009) Epigenetic biomarker development. Epigenomics, 1(1), 99-110.

Bojesen, S.E., Timpson, N., Relton, C., Smith, G.D. and Nordestgaard, B.G. (2017). AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax, 72(7), pp.646-653.

Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S., & Ollikainen, M. (2019). EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. Epigenomics, 11(13), 1469-1486.

Braveman, P. and Gottlieb, L. (2014) The social determinants of health: it's time to consider the causes of the causes. Public health reports, 129(1_suppl2), pp.19-31.

Breitling, L., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. American journal of human genetics. 88(4): 450-7.

Breitling, L.P., Salzmann, K., Rothenbacher, D., Burwinkel, B. and Brenner, H. (2012) Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. European heart journal, 33(22), pp.2841-2848.

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A., Socci, N. and Scandura, J. (2011) DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. PLoS ONE. 6(1), e14524.

Buck, N. and McFall, S. (2011) Understanding Society: design overview. Longitudinal and Life Course Studies, 3(1), 5-17.

Calle-Fabregat, C.D.L., Morante-Palacios, O. and Ballestar, E. (2020) Understanding the relevance of DNA methylation changes in immune differentiation and disease. Genes, 11(1), p.110.

Caraballo, R.S., Giovino, G.A., Pechacek, T.F. and Mowery, P.D. (2001). 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. American journal of epidemiology, 153(8), pp.807-814.

Castro, R., Rivera, I., Struys, E.A., Jansen, E.E., Ravasco, P., Camilo, M.E., Blom, H.J., Jakobs, C. and Tavares de Almeida, I. (2003) Increased homocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease. Clinical chemistry, 49(8), pp.1292-1296.

Chait, A. and Den Hartigh, L.J., 2020. Adipose tissue distribution, inflammation and its metabolic consequences, including diabetes and cardiovascular disease. Frontiers in cardiovascular medicine, 7, p.22.

Chiba, T., Marusawa, H. and Ushijima, T. (2012) Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. Gastroenterology, 143(3), pp.550-563.

Christiansen, C., Castillo-Fernandez, J.E., Domingo-Relloso, A., Zhao, W., Moustafa, J.E.S., Tsai, P.C., Maddock, J., Haack, K., Cole, S.A., Kardia, S.L.R. and Molokhia, M. (2021) Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clinical epigenetics, 13(1), pp.1-13.

Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V. and Abdennur, N.A., 2015. FTO obesity variant circuitry and adipocyte browning in humans. New England Journal of Medicine, 373(10), pp.895-907.

Coppack, S.W. (2001). Pro-inflammatory cytokines and adipose tissue. Proceedings of the nutrition society, 60(3), pp.349-356.

Copland, J.A., Sheffield-Moore, M., Koldzic-Zivanovic, N., Gentry, S., Lamprou, G., Tzortzatou-Stathopoulou, F., Zoumpourlis, V., Urban, R.J. and Vlahopoulos, S.A. (2009). Sex steroid receptors in skeletal differentiation and epithelial neoplasia: is tissue-specific intervention possible? Bioessays, 31(6), pp.629-641.

Corley, J., Cox, S.R., Harris, S.E., Hernandez, M.V., Maniega, S.M., Bastin, M.E., Wardlaw, J.M., Starr, J.M., Marioni, R.E. and Deary, I.J. (2019). Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. Translational Psychiatry, 9(1), pp.1-15.

Cotton, A.M., Price, E.M., Jones, M.J., Balaton, B.P., Kobor, M.S. and Brown, C.J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Human molecular genetics, 24(6), pp.1528-1539.

Coussens, L.M. and Werb, Z. (2002). Inflammation and cancer. Nature, 420(6917), pp.860-867.

Craig, J. and Wong, N.C. (2011) Epigenetics: a reference manual. Caister Academic Press.

Cruz-Flores, S., Rabinstein, A., Biller, J., Elkind, M.S., Griffith, P., Gorelick, P.B., Howard, G., Leira, E.C., Morgenstern, L.B., Ovbiagele, B. and Peterson, E. (2011) Racial-ethnic disparities in stroke care: the American experience: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 42(7), 2091-2116.

Dahlet, T., Lleida, A.A., Al Adhami, H., Dumas, M., Bender, A., Ngondo, R.P., Tanguy, M., Vallet, J., Auclair, G., Bardet, A.F. and Weber, M. (2020) Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. Nature communications, 11(1), pp.1-14.

Davillas, A., Benzeval, M. and Kumari, M. (2017) Socio-economic inequalities in C-reactive protein and fibrinogen across the adult age span: Findings from Understanding Society. Scientific Reports, 7(1), pp.1-13.

de Vries, M., van der Plaat, D.A., Nedeljkovic, I., Verkaik-Schakel, R.N., Kooistra, W., Amin, N., van Duijn, C.M., Brandsma, C.A., van Diemen, C.C., Vonk, J.M. and Boezen, H.M. (2018) From blood to lung tissue: effect of cigarette smoke on DNA methylation and lung function. Respiratory research, 19(1), pp.1-9.

Ding, N., Maiuri, A.R. and O'Hagan, H.M. (2019) The emerging role of epigenetic modifiers in repair of DNA damage associated with chronic inflammatory diseases. Mutation Research/Reviews in Mutation Research, 780, pp.69-81.

Dogan, M. V., Shields, B., Cutrona, C., Gao, L., Gibbons, F. X., Simons, R. & Philibert, R. A. (2014). The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC genomics, 15(1), 151.

Doll, R., & Hill, A. B. (1950). Smoking and carcinoma of the lung. British medical journal, 2(4682), 739.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. British Medical Journal, 328(7455), 1519.

Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics, 11(1), pp.1-9.

Ehrlich, M. (2009) DNA hypomethylation in cancer cells. Epigenomics, 1(2), pp.239-259.

El Kasmi, K.C., Smith, A.M., Williams, L., Neale, G., Panopolous, A., Watowich, S.S., Häcker, H., Foxwell, B.M. and Murray, P.J. (2007) Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway. The Journal of Immunology, 179(11), pp.7215-7219.

El Khoury, L. Y., Gorrie-Stone, T., Smart, M., Hughes, A., Bao, Y., Andrayas, A. & Schalkwyk, L. C. (2019). Systematic underestimation of the epigenetic clock and age acceleration in older subjects. Genome biology, 20(1), 283.

Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Smith, G.D., Hughes, A.D., Chaturvedi, N. and Relton, C.L. (2014) Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clinical epigenetics, 6(1), pp.1-10.

Escobedo, L.G. and Peddicord, J.P. (1996) Smoking prevalence in US birth cohorts: the influence of gender and education. American Journal of Public Health, 86(2), 231-236.

Engebretsen, S. and Bohlin, J. (2019). Statistical predictions with glmnet. Clinical epigenetics, 11(1), pp.1-3.

Evans, B. R., Karchner, S. I., Allan, L. L., Pollenz, R. S., Tanguay, R. L., Jenny, M. J., ... & Hahn, M. E. (2008). Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. Molecular pharmacology, 73(2), 387-398.

Fidler, T.P., Xue, C., Yalcinkaya, M., Hardaway, B., Abramowicz, S., Xiao, T., Liu, W., Thomas, D.G., Hajebrahimi, M.A., Pircher, J. and Silvestre-Roig, C., 2021. The AIM2 inflammasome exacerbates atherosclerosis in clonal haematopoiesis. Nature, 592(7853), pp.296-301.

Filzmoser, P., Maronna, R. and Werner, M. (2008) Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3): 1694-1711.

Friedman, E.M. and Herd, P. (2010). Income, education, and inflammation: differential associations in a national probability sample (The MIDUS study). Psychosomatic medicine, 72(3), p.290.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL: http://www.jstatsoft.org/v33/i01/.

Fu, C.A., Shen, M., Huang, B.C., Lasaga, J., Payan, D.G. and Luo, Y. (1999) TNIK, a novel member of the germinal center kinase family that activates the c-Jun N-terminal kinase pathway and regulates the cytoskeleton. Journal of Biological Chemistry, 274(43), pp.30729-30737.

Fuller E, Power C, Shepherd P, Strachan D (2006) "Technical report on the National Child Development Study biomedical survey 2002–2004". https://cls.ucl.ac.uk/wp-content/uploads/2017/07/NCDS-biomed-technical-report.pdf

Gao, X., Jia, M., Zhang, Y., Breitling, L.P. and Brenner, H. (2015) DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clinical epigenetics, 7(1), pp.1-10.

Gao, X., Zhang, Y., Breitling, L.P. and Brenner, H. (2016) Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget, 7(30), p.46878.

Gao, X., Thomsen, H., Zhang, Y., Breitling, L.P. and Brenner, H. (2017) The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. Clinical epigenetics, 9(1), pp.1-13.

Gershman, A., Sauria, M.E., Guitart, X., Vollger, M.R., Hook, P.W., Hoyt, S.J., Jain, M., Shumate, A., Razaghi, R., Koren, S. and Altemose, N. (2022). Epigenetic patterns in a complete human genome. Science, 376(6588), p.eabj5089.

Gonzalez-Jaramillo, V., Portilla-Fernandez, E., Glisic, M., Voortman, T., Ghanbari, M., Bramer, W., Chowdhury, R., Nijsten, T., Dehghan, A., Franco, O.H. and Nano, J. (2019) Epigenetics and inflammatory markers: a systematic review of the current evidence. International journal of inflammation, 2019.

Gonzalo, S. (2010) Epigenetic alterations in aging. Journal of applied physiology, 109(2), pp.586-597.

Gorrie-Stone, T. J., Smart, M. C., Saffari, A., Malki, K., Hannon, E., Burrage, J. & Schalkwyk, L. C. (2019). Bigmelon: tools for analysing large DNA methylation datasets. Bioinformatics, 35(6), 981-986.

Graham, H. and Hunt, S. (1994) Women's smoking and measures of women's socio-economic status in the United Kingdom. Health Promotion International, 9(2), 81-88.

Gruzieva, O., Merid, S.K., Chen, S., Mukherjee, N., Hedman, A.M., Almqvist, C., Andolf, E., Jiang, Y., Kere, J., Scheynius, A. and Söderhäll, C., 2019. DNA methylation trajectories during pregnancy. Epigenetics insights, 12, p.2516865719867090.

Guida, F., Sandanger, T.M., Castagné, R., Campanella, G., Polidoro, S., Palli, D., Krogh, V., Tumino, R., Sacerdote, C., Panico, S. and Severi, G. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Human molecular genetics, 24(8), pp.2349-2359.

Gunn, L., Ding, C., Liu, M., Ma, Y., Qi, C., Cai, Y., Hu, X., Aggarwal, D., Zhang, H.G. and Yan, J., 2012. Opposing roles for complement component C5a in tumor progression and the tumor microenvironment. The Journal of Immunology, 189(6), pp.2985-2994.

Haarmann-Stemmann, T., Bothe, H., Kohli, A., Sydlik, U., Abel, J. and Fritsche, E. (2007) Analysis of the transcriptional regulation and molecular function of the aryl hydrocarbon receptor repressor in human cell lines. Drug metabolism and disposition, 35(12), pp.2262-2269.

Haberland, M., Montgomery, R.L. and Olson, E.N. (2009) The many roles of histone deacetylases in development and physiology: implications for disease and therapy. Nature Reviews Genetics, 10(1), pp.32-42.

Han, L., Lin, I. G., & Hsieh, C. L. (2001). Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. Molecular and cellular biology, 21(10), 3416-3424.

Hannoodee, S. and Nasuruddin, D.N., 2020. Acute inflammatory response. StatPearls [Internet].

Harlid, S., Xu, Z., Panduri, V., Sandler, D.P. and Taylor, J.A. (2014) CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environmental health perspectives, 122(7), pp.673-678.

Heiss, J.A. and Brenner, H. (2017) Epigenome-wide discovery and evaluation of leukocyte DNA methylation markers for the detection of colorectal cancer in a screening setting. Clinical epigenetics, 9(1), pp.1-9.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J. and Puca, A.A. (2012) Distinct DNA methylomes of newborns and centenarians. Proceedings of the National Academy of Sciences, 109(26), pp.10522-10527.

Higuchi, T., Omata, F., Tsuchihashi, K., Higashioka, K., Koyamada, R., & Okada, S. (2016). Current cigarette smoking is a reversible cause of elevated white blood cell count: Cross-sectional and longitudinal studies. Preventive medicine reports, 4, 417-422.

Hornung, V., Ablasser, A., Charrel-Dennis, M., Bauernfeind, F., Horvath, G., Caffrey, D.R., Latz, E. and Fitzgerald, K.A. (2009) AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. Nature, 458(7237), pp.514-518.

Hornung, V., Ablasser, A., Charrel-Dennis, M., Bauernfeind, F., Horvath, G., Caffrey, D.R., Latz, E. and Fitzgerald, K.A. (2009) AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. Nature, 458(7237), pp.514-518.

Horvath, S. (2013) DNA methylation age of human tissues and cell types. Genome biology, 14(10), pp.1-20.

Hosono, H., Kumondai, M., Maekawa, M., Yamaguchi, H., Mano, N., Oda, A., Hirasawa, N. and Hiratsuka, M. (2017). Functional characterization of 34 CYP2A6 allelic variants by assessment of nicotine C-oxidation and coumarin 7-hydroxylation activities. Drug Metabolism and Disposition, 45(3), pp.279-285.

Houseman, E. A. (2015). DNA methylation and cell-type distribution. In Computational and Statistical Epigenomics (pp. 35-50). Springer, Dordrecht.

Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. and Kelsey, K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics, 13(1), pp.1-16.

Houston, C.S. (1986). The sociology of cigarette smoking. CMAJ: Canadian Medical Association Journal, 134(8), p.878.

Huang, Y.S., Zhi, Y.F. and Wang, S.R. (2009) Hypermethylation of estrogen receptor-α gene in atheromatosis patients and its correlation with homocysteine. Pathophysiology, 16(4), pp.259-265.

Hughes A, Smart M, Gorrie-Stone T, Hannon E, Mill J, Bao Y, Burrage J, Schalkwyk L, Kumari M, Benzeval M. (2018) EGAS00001002836 European Genome-phenome Archive. https://www.ebi.ac.uk/ega/studies/EGAS00001002836

Illumina Support. Retrieved from http://support.illumina.com.

Inflammation, epigenetics, and metabolism converge to cell senescence and ageing: the regulation and intervention. Signal Transduction and Targeted Therapy, 6(1), pp.1-29.

Jain, S., Gautam, V. and Naseem, S. (2011) Acute-phase proteins: As diagnostic tool. Journal of Pharmacy and Bioallied Sciences, 3(1), p.118.

Jhun, M.A., Smith, J.A., Ware, E.B., Kardia, S.L., Mosley Jr, T.H., Turner, S.T., Peyser, P.A. and Park, S.K. (2017) Modeling the causal role of DNA methylation in the association between cigarette smoking and inflammation in African Americans: a 2-step epigenetic Mendelian randomization study. American journal of epidemiology, 186(10), pp.1149-1158.

Joubert, B.R., Håberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L.A. and Ueland, P.M. (2012) 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives, 120(10), pp.1425-1431.

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R. & Moreno-Macias, H. (2016). Epigenetic signatures of cigarette smoking. Circulation: cardiovascular genetics, 9(5), 436-447.

Jousilahti, P., Salomaa, V., Rasi, V., Vahtera, E. and Palosuo, T. (2003) Association of markers of systemic inflammation, C reactive protein, serum amyloid A, and fibrinogen, with socioeconomic status. Journal of Epidemiology & Community Health, 57(9), 730-733.

Knowles, J.H. (1977) Introduction: doing better and feeling worse: health in the United States. Daedalus, 1.

Kornerup, H., Osler, M., Boysen, G., Barefoot, J., Schnohr, P. and Prescott, E. (2010) Major life events increase the risk of stroke but not of myocardial infarction: results from the Copenhagen City Heart Study. European Journal of Cardiovascular Prevention & Rehabilitation, 17(1), 113-118.

Krall, E. A., Valadian, I., Dwyer, J. T., & Gardner, J. A. N. E. (1989). Accuracy of recalled smoking data. American Journal of Public Health, 79(2), 200-202.

Krushkal, J., Silvers, T., Reinhold, W.C., Sonkin, D., Vural, S., Connelly, J., Varma, S., Meltzer, P.S., Kunkel, M., Rapisarda, A. and Evans, D. (2020) Epigenome-wide DNA methylation analysis of small cell lung cancer cell lines suggests potential chemotherapy targets. Clinical epigenetics, 12(1), pp.1-28.

Kuhn M (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.Rproject.org/package=caret

Kulis, M., Merkel, A., Heath, S., Queirós, A.C., Schuyler, R.P., Castellano, G., Beekman, R., Raineri, E., Esteve, A., Clot, G. and Verdaguer-Dot, N. (2015) Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nature genetics, 47(7), pp.746-756.

Kumar, H., Kawai, T. and Akira, S. (2011) Pathogen recognition by the innate immune system. International reviews of immunology, 30(1), pp.16-34.

Kwon, J.M. and Goate, A.M. (2000) The candidate gene approach. Alcohol Research & Health, 24(3), p.164.

Lau, D.C., Dhillon, B., Yan, H., Szmitko, P.E. and Verma, S. (2005) Adipokines: molecular links between obesity and atheroslcerosis. American Journal of Physiology-Heart and Circulatory Physiology, 288(5), pp.H2031-H2041.

Leenen, F. A., Muller, C. P., & Turner, J. D. (2016). DNA methylation: conducting the orchestra from exposure to phenotype? Clinical epigenetics, 8(1), 92.

Lei, M. K., Gibbons, F. X., Simons, R. L., Philibert, R. A., & Beach, S. R. (2020). The Effect of Tobacco Smoking Differs across Indices of DNA Methylation-Based Aging in an African American Sample: DNA Methylation-Based Indices of Smoking Capture These Effects. Genes, 11(3), 311.

Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y. and Whitsel, E.A. (2018) An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY), 10(4), p.573.

Libby, P. (2012) Inflammation in atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology, 32(9), pp.2045-2051.

Ligthart, S., Marzi, C., Aslibekyan, S., Mendelson, M.M., Conneely, K.N., Tanaka, T., Colicino, E., Waite, L.L., Joehanes, R., Guan, W. and Brody, J.A. (2016) DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome biology, 17(1), pp.1-15.

Linnér, R.K., Marioni, R.E., Rietveld, C.A., Simpkin, A.J., Davies, N.M., Watanabe, K., Armstrong, N.J., Auro, K., Baumbach, C., Bonder, M.J. and Buchwald, J. (2017) An epigenome-wide association study meta-analysis of educational attainment. Molecular psychiatry, 22(12), 1680.

Lisanti, S., Omar, W. A., Tomaszewski, B., De Prins, S., Jacobs, G., Koppen, G., ... & Langie, S. A. (2013). Comparison of methods for quantification of global DNA methylation in human cells and tissues. PloS one, 8(11), e79044.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M. and Edsall, L. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271), 315-322.

Liu, J., Carnero-Montoro, E., van Dongen, J., Lent, S., Nedeljkovic, I., Ligthart, S., Tsai, P.C., Martin, T.C., Mandaviya, P.R., Jansen, R. and Peters, M.J. (2019) An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nature communications, 10(1), pp.1-11.

Liu, J., Zhang, Y., Lavie, C.J., Tabung, F.K., Xu, J., Hu, Q., He, L. and Zhang, Y., 2020. Associations of C-reactive protein and fibrinogen with mortality from all-causes, cardiovascular disease and cancer among US adults. Preventive medicine, 139, p.106044.

Lynch, J.W., Kaplan, G.A. and Salonen, J.T. (1997) Why do poor people behave poorly? Variation in adult health behaviours and psychosocial characteristics by stages of the socioeconomic lifecourse. Social science & medicine, 44(6), 809-819.

Lynn, P., 2009. Sample design for understanding society. Underst. Soc. Work. Pap. Ser, 2009.

Maas, S. C., Vidaki, A., Wilson, R., Teumer, A., Liu, F., van Meurs, J. B., ... & van Dongen, J. (2019). Validated inference of smoking habits from blood with a finite DNA methylation marker set. European journal of epidemiology, 34(11), 1055-1074.

Mackenbach, J. P., Karanikolos, M., & McKee, M. (2013). The unequal health of Europeans: successes and failures of policies. The Lancet, 381(9872), 1125-1134.

Maksimovic, J., Oshlack, A. and Phipson, B. (2021). Gene set enrichment analysis for genome-wide DNA methylation data. Genome biology, 22(1), pp.1-26.

Man, S.M., Karki, R. and Kanneganti, T.D. (2016) AIM2 inflammasome in infection, cancer, and autoimmunity: Role in DNA sensing, inflammation, and innate immunity. European journal of immunology, 46(2), pp.269-280.

Mantovani, A., Allavena, P., Sica, A. and Balkwill, F. (2008) Cancer-related inflammation. nature, 454(7203), pp.436-444.

Markunas, C.A., Xu, Z., Harlid, S., Wade, P.A., Lie, R.T., Taylor, J.A. and Wilcox, A.J. (2014) Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. Environmental health perspectives, 122(10), pp.1147-1153.

Marmot, M. (2015) The health gap: the challenge of an unequal world. The Lancet, 386(10011), 2442-2444.

Marron, D. (2017). Smoke gets in your eyes: what is sociological about cigarettes? The Sociological Review, 65(4), pp.882-897.

Marzi, C., Holdt, L.M., Fiorito, G., Tsai, P.C., Kretschmer, A., Wahl, S., Guarrera, S., Teupser, D., Spector, T.D., Iacoviello, L. and Sacerdote, C. (2016) Epigenetic signatures at AQP3 and SOCS3 engage in low-grade inflammation across different tissues. PLoS One, 11(11), p.e0166015.

Marzi, C., Holdt, L.M., Fiorito, G., Tsai, P.C., Kretschmer, A., Wahl, S., Guarrera, S., Teupser, D., Spector, T.D., Iacoviello, L. and Sacerdote, C., 2016. Epigenetic signatures at AQP3 and SOCS3 engage in low-grade inflammation across different tissues. PLoS One, 11(11), p.e0166015.

McCartney, D. L., Stevenson, A. J., Hillary, R. F., Walker, R. M., Bermingham, M. L., Morris, S. W. & Porteous, D. J. (2018). Epigenetic signatures of starting and stopping smoking. EBioMedicine, 37, 214-220.

McCartney, D.L., Hillary, R.F., Stevenson, A.J., Ritchie, S.J., Walker, R.M., Zhang, Q., Morris, S.W., Bermingham, M.L., Campbell, A., Murray, A.D. and Whalley, H.C. (2018) Epigenetic prediction of complex traits and death. Genome biology, 19(1), pp.1-11.

McErlean, P., Bell, C.G., Hewitt, R.J., Busharat, Z., Ogger, P.P., Ghai, P., Albers, G.J., Calamita, E., Kingston, S., Molyneaux, P.L. and Beck, S., 2021. DNA methylome alterations are associated with airway macrophage differentiation and phenotype during lung fibrosis. American Journal of Respiratory and Critical Care Medicine, 204(8), pp.954-966.

McEwen, B.S. (1998) Stress, adaptation, and disease: Allostasis and allostatic load. Annals of the New York academy of sciences, 840(1), 33-44.

Mendelson, M.M., Marioni, R.E., Joehanes, R., Liu, C., Hedman, Å.K., Aslibekyan, S., Demerath, E.W., Guan, W., Zhi, D., Yao, C. and Huan, T. (2017) Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian randomization approach. PLoS medicine, 14(1), p.e1002215.

Miller, M.W., Maniates, H., Wolf, E.J., Logue, M.W., Schichman, S.A., Stone, A., Milberg, W. and McGlinchey, R. (2018) CRP polymorphisms and DNA methylation of the AIM2 gene influence associations between trauma exposure, PTSD, and C-reactive protein. Brain, behavior, and immunity, 67, pp.194-202.

Mittelstraß, K. and Waldenberger, M. (2018) DNA methylation in human lipid metabolism and related diseases. Current opinion in lipidology, 29(2), p.116.

Moore, L.D., Le, T. and Fan, G. (2013) DNA methylation and its basic function. Neuropsychopharmacology, 38(1), 23-38.

Mordaunt, C.E., Jianu, J.M., Laufer, B.I., Zhu, Y., Hwang, H., Dunaway, K.W., Bakulski, K.M., Feinberg, J.I., Volk, H.E., Lyall, K. and Croen, L.A. (2020) Cord blood DNA methylome in newborns later diagnosed with autism spectrum disorder reflects early dysregulation of neurodevelopmental and X-linked genes. Genome medicine, 12(1), pp.1-25.

Mosesson, M.W. (2005) Fibrinogen and fibrin structure and functions. Journal of thrombosis and haemostasis, 3(8), pp.1894-1904.

Mote, V. L., & Anderson, R. L. (1965). An Investigation of the Effect of Misclassification on the Properties of χ^2-Tests in the Analysis of Categorical Data. Biometrika, 52(1/2), 95-109.

Müller, N., Schwarz, M.J., Dehning, S., Douhe, A., Cerovecki, A., Goldstein-Müller, B., Spellmann, I., Hetzel, G., Maino, K., Kleindienst, N. and Möller, H.J. (2006). The cyclooxygenase-2 inhibitor celecoxib has therapeutic effects in major depression: results of a double-blind, randomized, placebo controlled, add-on pilot study to reboxetine. Molecular psychiatry, 11(7), pp.680-684.

Muscatell, K.A., Brosso, S.N. and Humphreys, K.L. (2020) Socioeconomic status and inflammation: a meta-analysis. Molecular psychiatry, 25(9), pp.2189-2199.

Nabais, M.F., Laws, S.M., Lin, T., Vallerga, C.L., Armstrong, N.J., Blair, I.P., Kwok, J.B., Mather, K.A., Mellick, G.D., Sachdev, P.S. and Wallace, L. (2021). Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. Genome biology, 22(1), pp.1-30.

Needham, B.L., Smith, J.A., Zhao, W., Wang, X., Mukherjee, B., Kardia, S.L., Shively, C.A., Seeman, T.E., Liu, Y. and Diez Roux, A.V. (2015) Life course socioeconomic status and DNA methylation in genes related to stress reactivity and inflammation: The multi-ethnic study of atherosclerosis. Epigenetics, 10(10), 958-969.

NIH (1964) Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service.

Northrop, J.P., Ho, S.N., Chen, L., Thomas, D.J., Timmerman, L.A., Nolan, G.P., Admon, A. and Crabtree, G.R. (1994) NF-AT components define a family of transcription factors targeted in T-cell activation. Nature, 369(6480), 497.

Odintsova, V.V., Rebattu, V., Hagenbeek, F.A., Pool, R., Beck, J.J., Ehli, E.A., van Beijsterveldt, C.E., Ligthart, L., Willemsen, G., De Geus, E.J. and Hottenga, J.J. (2021) Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. Frontiers in Psychiatry, 12.

Ohno, R., Nakayama, M., Naruse, C., Okashita, N., Takano, O., Tachibana, M., Asano, M., Saitou, M. and Seki, Y. (2013) A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. Development, 140(14), 2892-2903.

Olety, B., WaÃàlte, M., Honnert, U., Schillers, H. and BaÃàhler, M. (2010) Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity. FEBS letters, 584(3), 493-499.

Oshima, M., Mimura, J., Yamamoto, M. and Fujii-Kuriyama, Y. (2007) Molecular mechanism of transcriptional repression of AhR repressor involving ANKRA2, HDAC4, and HDAC5. Biochemical and biophysical research communications, 364(2), pp.276-282.

Packard, C.J., Bezlyak, V., McLean, J.S., Batty, G.D., Ford, I., Burns, H., Cavanagh, J., Deans, K.A., Henderson, M., McGinty, A. and Millar, K. (2011) Early life socioeconomic adversity is associated in adult life with chronic inflammation, carotid atherosclerosis, poorer lung function and decreased cognitive performance: a cross-sectional, population-based study. BMC public health, 11(1), 42.

Pahwa, R., Goyal, A., Bansal, P. and Jialal, I. (2018) Chronic inflammation.

Parimisetty, A., Dorsemans, A.C., Awada, R., Ravanan, P., Diotel, N. and d'Hellencourt, C.L. (2016) Secret talks between adipose tissue and central nervous system via secreted factors—an emerging frontier in the neurodegenerative research. Journal of neuroinflammation, 13(1), pp.1-13.

Patrick, D.L., Cheadle, A., Thompson, D.C., Diehr, P., Koepsell, T. and Kinne, S. (1994) The validity of self-reported smoking: a review and meta-analysis. American journal of public health, 84(7), 1086-1093.

Pepin, M.E., Ha, C.M., Crossman, D.K., Litovsky, S.H., Varambally, S., Barchue, J.P., Pamboukian, S.V., Diakos, N.A., Drakos, S.G., Pogwizd, S.M. and Wende, A.R., 2019. Genome-wide DNA methylation encodes cardiac transcriptional reprogramming in human ischemic heart failure. Laboratory Investigation, 99(3), pp.371-386.

Pepin, M.E., Ha, C.M., Potter, L.A., Bakshi, S., Barchue, J.P., Haj Asaad, A., Pogwizd, S.M., Pamboukian, S.V., Hidalgo, B.A., Vickers, S.M. and Wende, A.R. (2021) Racial and socioeconomic disparity associates with differences in cardiac DNA methylation among men with end-stage heart failure. American Journal of Physiology-Heart and Circulatory Physiology, 320(5), pp.H2066-H2079.

Petersen, A.K., Zeilinger, S., Kastenmüller, G., Römisch-Margl, W., Brugger, M., Peters, A., Meisinger, C., Strauch, K., Hengstenberg, C., Pagel, P. and Huber, F. (2014) Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Human molecular genetics, 23(2), pp.534-545.

Peto, J. (2011). That lung cancer incidence falls in ex-smokers: misconceptions 2. British journal of cancer, 104(3), 389-389.

Peto, J. (2012). That the effects of smoking should be measured in pack-years: misconceptions 4.

Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., & Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Bmj, 321(7257), 323-329.

Philibert, R., Dogan, M., Beach, S.R., Mills, J.A. and Long, J.D. (2020) AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 183(1), pp.51-60.

Philibert, R., Hollenbeck, N., Andersen, E., McElroy, S., Wilson, S., Vercande, K. & Wang, K. (2016). Reversion of AHRR demethylation is a quantitative biomarker of smoking cessation. Frontiers in psychiatry, 7, 55.

Pianezza, M.L., Sellers, E.M. and Tyndale, R.F. (1998). Nicotine metabolism defect reduces smoking. Nature, 393(6687), pp.750-750.

Pidsley, R., Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC genomics, 14(1), 293.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C. and Clark, S.J. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology, 17(1), 208.

Piperi, C., Themistocleous, M.S., Papavassiliou, G.A., Farmaki, E., Levidou, G., Korkolopoulou, P., Adamopoulos, C. and Papavassiliou, A.G. (2010) High incidence of MGMT and RARβ promoter methylation in primary glioblastomas: Association with histopathological characteristics, inflammatory mediators and clinical outcome. Molecular medicine, 16(1), pp.1-9.

Pollitt, R.A., Kaufman, J.S., Rose, K.M., Diez-Roux, A.V., Zeng, D. and Heiss, G. (2008) Cumulative life course and adult socioeconomic status and markers of inflammation in adulthood. Journal of Epidemiology & Community Health, 62(6), 484-491.

Porcu, E., Sadler, M.C., Lepik, K., Auwerx, C., Wood, A.R., Weihs, A., Sleiman, M.S.B., Ribeiro, D.M., Bandinelli, S., Tanaka, T. and Nauck, M., 2021. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nature Communications, 12(1), pp.1-9.

Power, C. and Elliott, J. (2006) Cohort profile: 1958 British birth cohort (national child development study). International journal of epidemiology, 35(1), pp.34-41.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rakyan, V.K., Down, T.A., Balding, D.J. and Beck, S. (2011) Epigenome-wide association studies for common human diseases. Nature Reviews Genetics, 12(8), 529.

Reese, S.E., Xu, C.J., Herman, T., Lee, M.K., Sikdar, S., Ruiz-Arenas, C., Merid, S.K., Rezwan, F.I., Page, C.M., Ullemar, V. and Melton, P.E. (2019) Epigenome-wide meta-analysis of DNA methylation and childhood asthma. Journal of Allergy and Clinical Immunology, 143(6), pp.2062-2074.

Reynolds, L.M., Wan, M., Ding, J., Taylor, J.R., Lohman, K., Su, D., Bennett, B.D., Porter, D.K., Gimple, R., Pittman, G.S. and Wang, X. (2015). DNA methylation of the aryl hydrocarbon receptor repressor associations with cigarette smoking and subclinical atherosclerosis. Circulation: Cardiovascular Genetics, 8(5), pp.707-716.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.

Robbins, S.L. and Cotran, R.S. (1979). Pathologic basis of disease. Saunders.

Robertson, K.D. (2005) DNA methylation and human disease. Nature reviews. Genetics, 6(8), 597.

Roth, T.L. and Sweatt, J.D. (2011) Epigenetic marking of the BDNF gene by early-life adverse experiences. Hormones and behavior, 59(3), 315-320.

Rottenberg, M.E. and Carow, B. (2014) SOCS3, a major regulator of infection and inflammation. Frontiers in immunology, 5, p.58.

Sadahiro, R., Knight, B., James, F., Hannon, E., Charity, J., Daniels, I.R., Burrage, J., Knox, O., Crawford, B., Smart, N.J. and Mill, J. (2020) Major surgery induces acute changes in measured DNA methylation associated with immune response pathways. Scientific reports, 10(1), pp.1-13.

Saini, A., Oh, T.H., Ghanem, D.A., Castro, M., Butler, M., Sin Fai Lam, C.C., Posporelis, S., Lewis, G., David, A.S. and Rogers, J.P., 2021. Inflammatory and blood gas markers of COVID-19 delirium compared to non-COVID-19 delirium: a cross-sectional study. Aging & Mental Health, pp.1-8.

Salameh, Y., Bejaoui, Y. and El Hajj, N. (2020) DNA methylation biomarkers in aging and age-related diseases. Frontiers in genetics, 11, p.171.

Satta, R., Maloku, E., Zhubi, A., Pibiri, F., Hajos, M., Costa, E., & Guidotti, A. (2008). Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. Proceedings of the National Academy of Sciences, 105(42), 16356-16361.

Schmidt, J.V. and Bradfield, C.A. (1996) Ah receptor signaling pathways. Annual review of cell and developmental biology, 12(1), pp.55-89.

Schmoll, H.J., Souchon, R., Krege, S., Albers, P., Beyer, J., Kollmannsberger, C., Fossa, S.D., Skakkebaek, N.E., De Wit, R., Fizazi, K. and Droz, J.P. (2004) European consensus on diagnosis and treatment of germ cell cancer: a report of the European Germ Cell Cancer Consensus Group (EGCCCG). Annals of Oncology, 15(9), 1377-1399.

Schübeler, D. (2015). ESCI award lecture: regulation, function and biomarker potential of DNA methylation. European Journal of Clinical Investigation, 45(3), 288-293.

Serhan, C.N. and Savill, J. (2005) Resolution of inflammation: the beginning programs the end. Nature immunology, 6(12), pp.1191-1197.

Shenker, N.S., Ueland, P.M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J.M. and Vineis, P. (2013) DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology, pp.712-716.

Smith, A.K., Conneely, K.N., Pace, T.W., Mister, D., Felger, J.C., Kilaru, V., Akel, M.J., Vertino, P.M., Miller, A.H. and Torres, M.A. (2014) Epigenetic changes associated with inflammation in breast cancer patients treated with chemotherapy. Brain, behavior, and immunity, 38, pp.227-236.

Smith, B.W., Rozelle, S.S., Leung, A., Ubellacker, J., Parks, A., Nah, S.K., French, D., Gadue, P., Monti, S., Chui, D.H. and Steinberg, M.H. (2013) The aryl hydrocarbon receptor directs hematopoietic progenitor cell expansion and differentiation. Blood, 122(3), pp.376-385.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 3(1).

Staaf, J., Vallon-Christersson, J., Lindgren, D., Juliusson, G., Rosenquist, R., HoÃàglund, M., RingneÃÅr, M. (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC Bioinformatics, 9(1), 409.

Steemers, F., Chang, W., Lee, G., Barker, D., Shen, R. and Gunderson, K. (2006) Whole-genome genotyping with the single-base extension assay. Nature Methods, 3(1): 31-33.

Stookey, G.K., Katz, B.P., Olson, B.L., Drook, C.A. and Cohen, S.J. (1987). Evaluation of biochemical validation measures in determination of smoking status. Journal of Dental Research, 66(10), pp.1597-1601.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? Current Opinion in HIV and AIDS, 5(6), 463.

Stringhini, S., Polidoro, S., Sacerdote, C., Kelly, R.S., Van Veldhoven, K., Agnoli, C., Grioni, S., Tumino, R., Giurdanella, M.C., Panico, S. and Mattiello, A. (2015) Life-course socioeconomic status and DNA methylation of genes regulating inflammation. International journal of epidemiology, 44(4), 1320-1330.

Su, D., Wang, X., Campbell, M.R., Porter, D.K., Pittman, G.S., Bennett, B.D., Wan, M., Englert, N.A., Crowl, C.L., Gimple, R.N. and Adamski, K.N., 2016. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. PloS one, 11(12), p.e0166486.

Su, S., Zhu, H., Xu, X., Wang, X., Dong, Y., Kapuku, G., Treiber, F., Gutin, B., Harshfield, G., Snieder, H. and Wang, X., 2014. DNA methylation of the LY86 gene is associated with obesity, insulin resistance, and inflammation. Twin Research and Human Genetics, 17(3), pp.183-191.

Sugden, K., Hannon, E.J., Arseneault, L., Belsky, D.W., Broadbent, J.M., Corcoran, D.L., Hancox, R.J., Houts, R.M., Moffitt, T.E., Poulton, R. and Prinz, J.A. (2019) Establishing a generalized polyepigenetic biomarker for tobacco smoking. Translational psychiatry, 9(1), pp.1-12.

Sumner, J.A., Gambazza, S., Gao, X., Baccarelli, A.A., Uddin, M. and McLaughlin, K.A. (2022). Epigenetics of early-life adversity in youth: cross-sectional and longitudinal associations. Clinical epigenetics, 14(1), pp.1-12.

Sun, Y.V. (2014) The influences of genetic and environmental factors on methylome-wide association studies for human diseases. Current genetic medicine reports, 2(4), pp.261-270.

Syme, S. L. (1992) Social determinants of disease. Maxcy-Rosenau Public Health and Preventive Medicine, 13, 687-700.

Tabor, H.K., Risch, N.J. and Myers, R.M., 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics, 3(5), pp.391-397.

Thompson, D., Pepys, M.B. and Wood, S.P. (1999) The physiological structure of human C-reactive protein and its complex with phosphocholine. Structure, 7(2), pp.169-177.

Teschendorff, A.E., Yang, Z., Wong, A., Pipinikas, C.P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H.B., Thirlwell, C. and Janes, S.M. (2015) Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. JAMA oncology, 1(4), pp.476-485.

Tost, J. (2010) DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. Molecular biotechnology, 44(1), pp.71-81.

Tsaprouni, L., Yang, T., Bell, J., Dick, K., Kanoni, S., Nisbet, J., VinÃÉuela, A., Grundberg, E., Nelson, C., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A., Ouwehand, W., Dermitzakis, E., Spector, T., Samani, N. and Deloukas, P. (2014) Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics, 9(10), 1382-96.

Tung, J., Barreiro, L.B., Johnson, Z.P., Hansen, K.D., Michopoulos, V., Toufexis, D., Michelini, K., Wilson, M.E. and Gilad, Y. (2012) Social environment is associated with gene regulatory variation in the rhesus macaque immune system. Proceedings of the National Academy of Sciences, 109(17), 6490-6495.

Ungefroren, H., Sebens, S., Seidl, D., Lehnert, H. and Hass, R. (2011). Interaction of tumor cells with the microenvironment. Cell Communication and Signaling, 9(1), pp.1-8.

University of Essex, Institute for Social and Economic Research, NatCen Social Research and Kantar Public. (2018) Understanding Society: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. UK Data Service. SN: 6614.

Van Eeden, S. F., & Hogg, J. C. (2000). The response of human bone marrow to chronic cigarette smoking. European Respiratory Journal, 15(5), 915-921.

Verschoor, C.P., McEwen, L.M., Kobor, M.S., Loeb, M.B. and Bowdish, D.M. (2018) DNA methylation patterns are related to co-morbidity status and circulating C-reactive protein levels in the nursing home elderly. Experimental gerontology, 105, pp.47-52. Zhu, X., Chen, Z., Shen, W., Huang, G., Sedivy, J.M., Wang, H. and Ju, Z., 2021.

Voss, T.C. and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature Reviews Genetics, 15(2), pp.69-81.

Vryer, R. and Saffery, R. (2017) What's in a name? Context-dependent significance of 'global'methylation measures in human health and disease. Clinical epigenetics, 9(1), pp.1-4.

Wagenknecht, L.E., Burke, G.L., Perkins, L.L., Haley, N.J. and Friedman, G.D. (1992). Misclassification of smoking status in the CARDIA study: a comparison of self-report with serum cotinine levels. American Journal of Public Health, 82(1), pp.33-36.

Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T. and Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome biology, 15(2), pp.1-17.

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W.R., Kunze, S., Tsai, P.C., Ried, J.S., Zhang, W., Yang, Y. and Tan, S. (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature, 541(7635), pp.81-86.

Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A. and DeMeo, D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Human molecular genetics, 21(13), pp.3073-3082.

Wannamethee, S.G., Lowe, G.D., Shaper, A.G., Rumley, A., Lennon, L. and Whincup, P.H. (2005) Associations between cigarette smoking, pipe/cigar smoking, and smoking cessation, and haemostatic and inflammatory markers for cardiovascular disease. European heart journal, 26(17), pp.1765-1773.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M. and Schübeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics, 39(4), pp.457-466.

Wei, L., Xia, H., Zhao, Y., Zhang, Z. and Chen, J. (2016) Predictors of white blood cell interleukin-6 DNA methylation levels in healthy subjects. International Journal of Clinical and Experimental Medicine, 9(11), pp.22162-22168.

Wei, S., Tao, J., Xu, J., Chen, X., Wang, Z., Zhang, N., Zuo, L., Jia, Z., Chen, H., Sun, H. and Yan, Y. (2021) Ten Years of EWAS. Advanced Science, p.2100727.

Weisenberger, D.J., Van Den Berg, D., Pan, F., Berman, B.P. and Laird, P.W. (2008) Comprehensive DNA methylation analysis on the Illumina Infinium assay platform. Illumina, San Diego.

World Health Organisation (2005) WHO Framework Convention on Tobacco Control. Retreived from: https://web.archive.org/web/20041117032449/http://www.who.int/tobacco/framework/countrylist/en/

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wiedermann, U., Chen, X.J., Enerbäck, L., Hanson, L.Å., Kahu, H. and Dahlgren, U.I. (1996) Vitamin A deficiency increases inflammatory responses. Scandinavian Journal of Immunology, 44(6), pp.578-584.

Wong, C.C., Pidsley, R. and Schalkwyk, L.C. (2013) The wateRmelon Package.

Wong, C.C.Y., Caspi, A., Williams, B., Craig, I.W., Houts, R., Ambler, A., Moffitt, T.E. and Mill, J. (2010) A longitudinal study of epigenetic variation in twins. Epigenetics, 5(6), pp.516-526.

World Health Organization. (2020). Tobacco. Retrieved from https://www.who.int/news-room/fact-sheets/detail/tobacco.

Xiao, F. H., Wang, H. T., & Kong, Q. P. (2019). Dynamic DNA methylation during aging: a "prophet" of age-related outcomes. Frontiers in genetics, 10, 107.

Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., Li, Z., Zhang, Z. and Bao, Y. (2020) EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic acids research, 48(D1), pp. D890-D895.

Xue, A., Jiang, L., Zhu, Z., Wray, N. R., Visscher, P. M., Zeng, J., & Yang, J. (2020). Genome-wide analyses of behavioural traits biased by misreports and longitudinal changes. medRxiv.

Yang, Y., Gao, X., Just, A.C., Colicino, E., Wang, C., Coull, B.A., Hou, L., Zheng, Y., Vokonas, P., Schwartz, J. and Baccarelli, A.A. (2019) Smoking-related DNA methylation is associated with DNA methylation phenotypic age acceleration: The veterans affairs normative aging study. International journal of environmental research and public health, 16(13), p.2356.

Yamashita, S., Kishino, T., Takahashi, T., Shimazu, T., Charvat, H., Kakugawa, Y., Nakajima, T., Lee, Y.C., Iida, N., Maeda, M. and Hattori, N. (2018). Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. Proceedings of the national academy of sciences, 115(6), pp.1328-1333.

You, C., Wu, S., Zheng, S.C., Zhu, T., Jing, H., Flagg, K., Wang, G., Jin, L., Wang, S. and Teschendorff, A.E. (2020). A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. Nature communications, 11(1), pp.1-13.

Yu, H., Raut, J.R., Schöttker, B., Holleczek, B., Zhang, Y. and Brenner, H. (2020) Individual and joint contributions of genetic and methylation risk scores for enhancing lung cancer risk stratification: data from a population-based cohort in Germany. Clinical epigenetics, 12(1), pp.1-11.

Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A. and Strauch, K. (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one, 8(5), 63812.

Zhang, J.M. and An, J. (2007) Cytokines, inflammation and pain. International anesthesiology clinics, 45(2), 27.

Zhang, Y., Florath, I., Saum, K. U., & Brenner, H. (2016). Self-reported smoking, serum cotinine, and blood DNA methylation. Environmental research, 146, 395-403.

Zhang, Y., Schöttker, B., Florath, I., Stock, C., Butterbach, K., Holleczek, B. & Brenner, H. (2016). Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environmental health perspectives, 124(1), 67-74.

Zhang, Y., Schöttker, B., Ordóñez-Mena, J., Holleczek, B., Yang, R., Burwinkel, B., Butterbach, K. and Brenner, H. (2015) F2RL3 methylation, lung cancer incidence and mortality. International journal of cancer, 137(7), pp.1739-1748.

Zhao, Y., Bjørbaek, C.H.R.I.S.T.I.A.N., Weremowicz, S., Morton, C.C. and Moller, D.E. (1995) RSK3 encodes a novel pp90rsk isoform with a unique N-terminal sequence: growth factor-stimulated kinase function and nuclear translocation. Molecular and cellular biology, 15(8), pp.4353-4363.

Zhu, S., Goldschmidt-Clermont, P.J. and Dong, C., 2005. Inactivation of monocarboxylate transporter MCT3 by DNA methylation in atherosclerosis. Circulation, 112(9), pp.1353-1361.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), pp.301-320.

8. Appendix

Supplementary Table 1

	NCDS			USM1			USM2		
Characteristic	**Never**	**Former**	**Current**	**Never**	**Former**	**Current**	**Never**	**Former**	**Current**
AHRR	0.80 (0.63, 0.94)	0.55 (0.21, 0.77)	-1.26 (-1.61, -0.71)	0.61 (0.41, 0.76)	0.34 (-0.09, 0.62)	-2.03 (-2.49, -1.42)	0.64 (0.45, 0.78)	0.40 (-0.12, 0.66)	-1.66 (-2.20, -0.82)
smokp SSt									
Never	197 (98%)	120 (86%)	25 (15%)	0 (NA%)	0 (NA%)	0 (NA%)	848 (87%)	527 (53%)	69 (14%)
Former	0 (0%)	3 (2.1%)	0 (0%)	0 (NA%)	0 (NA%)	0 (NA%)	127 (13%)	422 (43%)	91 (18%)
Current	3 (1.5%)	17 (12%)	137 (85%)	0 (NA%)	0 (NA%)	0 (NA%)	3 (0.3%)	43 (4.3%)	348 (69%)
smokp Packyears	-0.49 (-0.86, -0.16)	-0.08 (-0.54, 0.42)	0.93 (0.42, 1.43)	-0.67 (-1.00, -0.29)	0.16 (-0.49, 1.08)	0.98 (0.32, 1.48)	-0.57 (-0.97, -0.14)	0.02 (-0.58, 0.79)	0.86 (0.02, 1.54)
smokp Cessation	0.76 (0.54, 0.96)	0.57 (0.08, 0.82)	-1.18 (-1.58, -0.70)	0.31 (0.06, 0.60)	0.42 (0.10, 0.81)	-2.01 (-2.46, -1.61)	0.46 (0.17, 0.77)	0.36 (-0.07, 0.74)	-1.51 (-2.11, -0.84)
EpiSmokEr SSt									
Never	172 (86%)	81 (58%)	9 (5.6%)	98 (22%)	113 (28%)	40 (26%)	210 (21%)	239 (24%)	125 (25%)
Former	14 (7.0%)	19 (14%)	1 (0.6%)	244 (55%)	201 (49%)	78 (51%)	546 (56%)	518 (52%)	264 (52%)
Current	14 (7.0%)	40 (29%)	152 (94%)	103 (23%)	96 (23%)	36 (23%)	222 (23%)	235 (24%)	119 (23%)
EpiSmokEr SSc	-0.72 (-0.95, -0.47)	-0.42 (-0.76, -0.02)	1.07 (0.59, 1.61)	-0.27 (-0.65, 0.39)	-0.31 (-0.70, 0.34)	-0.34 (-0.73, 0.31)	-0.37 (-0.69, 0.33)	-0.30 (-0.68, 0.40)	-0.30 (-0.65, 0.41)
EpiSmokEr MS	-0.74 (-0.98, -0.51)	-0.45 (-0.84, -0.04)	1.11 (0.63, 1.51)	-0.24 (-0.71, 0.50)	-0.19 (-0.65, 0.36)	-0.27 (-0.77, 0.52)	-0.25 (-0.72, 0.46)	-0.17 (-0.67, 0.52)	-0.24 (-0.72, 0.49)

	NCDS			USM1			USM2		
Characteristic	**Never**	**Former**	**Current**	**Never**	**Former**	**Current**	**Never**	**Former**	**Current**
Sugden	-0.65 (-0.99, -0.37)	-0.35 (-0.71, 0.06)	1.02 (0.50, 1.62)	-0.56 (-0.88, -0.17)	-0.17 (-0.63, 0.47)	1.39 (0.87, 2.00)	-0.55 (-0.89, -0.17)	-0.24 (-0.69, 0.31)	1.22 (0.48, 1.97)
McCartney MS	-0.83 (-0.95, -0.66)	-0.54 (-0.74, -0.14)	1.31 (0.83, 1.56)	-0.61 (-0.76, -0.42)	-0.32 (-0.58, 0.15)	2.09 (1.48, 2.44)	-0.65 (-0.79, -0.48)	-0.37 (-0.64, 0.16)	1.72 (0.90, 2.24)
Christiansen	0.61 (0.34, 0.90)	0.45 (-0.04, 0.71)	-1.11 (-1.65, -0.49)	0.53 (0.24, 0.83)	0.28 (-0.14, 0.62)	-1.82 (-2.39, -1.11)	0.59 (0.30, 0.84)	0.32 (-0.15, 0.65)	-1.52 (-2.12, -0.71)
Odintsova	-0.63 (-1.02, -0.11)	-0.18 (-0.71, 0.25)	0.82 (0.13, 1.49)	-0.31 (-0.91, 0.22)	-0.12 (-0.68, 0.50)	1.08 (0.41, 1.72)	-0.32 (-0.91, 0.20)	-0.11 (-0.72, 0.53)	0.86 (0.19, 1.51)
Teschendorff	-0.79 (-1.13, 0.44)	0.17 (-1.05, 0.72)	0.78 (-0.44, 1.19)	-0.16 (-0.82, 0.40)	0.09 (-0.65, 0.61)	0.87 (0.19, 1.38)	-0.40 (-0.93, 0.15)	-0.01 (-0.61, 0.60)	0.89 (0.21, 1.51)
Yu	-0.66 (-0.93, -0.48)	-0.41 (-0.67, -0.05)	1.08 (0.50, 1.64)	-0.58 (-0.85, -0.29)	-0.18 (-0.56, 0.36)	1.63 (1.04, 2.24)	-0.60 (-0.85, -0.29)	-0.29 (-0.65, 0.23)	1.31 (0.62, 2.07)
Gao	-0.51 (-0.89, -0.27)	-0.41 (-0.77, -0.03)	0.95 (0.31, 1.54)	-0.58 (-0.91, -0.23)	-0.14 (-0.54, 0.42)	1.36 (0.73, 2.06)	-0.60 (-0.90, -0.23)	-0.22 (-0.63, 0.30)	1.22 (0.42, 1.89)
Yang	-0.68 (-0.91, -0.41)	-0.46 (-0.64, -0.09)	1.01 (0.41, 1.62)	-0.49 (-0.88, -0.13)	-0.15 (-0.58, 0.42)	1.46 (0.91, 2.09)	-0.53 (-0.86, -0.23)	-0.23 (-0.62, 0.27)	1.29 (0.54, 2.03)
Zhang (Lower Quartile), n (%)	2 (1.0%)	10 (7.1%)	107 (66%)	10 (2.2%)	85 (21%)	150 (97%)	15 (1.5%)	175 (18%)	429 (84%)

Supplementary Table 2:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p-**	**N**	**OR**	**95% CI**	**p**
Sex	462				2,198				509				2,660			
Male		—	—			—	—			—	—			—	—	
Female		1.14	0.47, 2.80	0.8		0.76	0.62, 0.93	0.007		0.96	0.63, 1.46	0.8		0.80	0.66, 0.96	0.019
Age					2,198	1.01	1.01, 1.02	<0.001	509	1.01	0.94, 1.08	0.8	2,660	1.02	1.01, 1.02	<0.001
Self-reported SSt	462				2,198				509				2,660			
Never		—	—			—	—			—	—			—	—	
Former		0.00	0.00, 0.00	<0.001		0.09	0.07, 0.12	<0.001		0.10	0.06, 0.16	<0.001		0.06	0.05, 0.08	<0.001
Current		0.11	0.02, 0.34	<0.001		0.26	0.20, 0.35	<0.001		0.45	0.24, 0.85	0.014		0.29	0.22, 0.38	<0.001
Educational attainment	462				2,198				509				2,660			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		0.88	0.25, 3.44	0.8		1.34	1.03, 1.76	0.031		1.29	0.71, 2.39	0.4		1.30	1.00, 1.69	0.049
GCSE/equivalent		2.28	0.77, 6.95	0.14		1.16	0.90, 1.49	0.2		1.01	0.60, 1.69	>0.9		1.17	0.93, 1.47	0.2
No qualification		0.96	0.25, 4.29	>0.9		1.71	1.22, 2.42	0.002		1.00	0.47, 2.18	>0.9		1.51	1.09, 2.08	0.013

Supplementary Table 3:

	NCDS			USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p**
Sex	146			442				119				588			
Male		—	—		—	—			—	—			—	—	

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**		**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p**	**N**	**OR**	**95% CI**	**p**
Female		2.44	0.86, 7.62	0.10		1.18	0.76, 1.84	0.4		1.36	0.57, 3.29	0.5		1.32	0.89, 1.98	0.2
Age					442	0.95	0.93, 0.96	<0.001	119	0.90	0.78, 1.03	0.14	588	0.94	0.93, 0.96	<0.001
Educational attainment	146				442				119				588			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		1.44	0.31, 7.86	0.7		3.47	1.84, 6.73	<0.001		4.48	1.38, 16.1	0.016		2.89	1.61, 5.33	<0.001
GCSE/equivalent		2.48	0.71, 8.73	0.15		3.08	1.78, 5.42	<0.001		7.43	2.52, 24.3	<0.001		3.11	1.91, 5.14	<0.001
No qualification		0.84	0.20, 3.74	0.8		3.60	1.97, 6.74	<0.001		4.79	1.48, 17.2	0.012		2.90	1.67, 5.13	<0.001

Supplementary Table 4:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
Sex	316				1,756				390				2,072			
Male		—	—			—	—			—	—			—	—	
Female		1.54	0.97, 2.45	0.068		0.94	0.77, 1.14	0.5		1.05	0.68, 1.62	0.8		1.02	0.85, 1.22	0.9
Age					1,756	1.02	1.01, 1.02	<0.001	390	1.01	0.94, 1.08	0.8	2,072	1.02	1.01, 1.03	<0.001
Educational attainment	316				1,756				390				2,072			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		1.48	0.73, 3.08	0.3		1.13	0.87, 1.48	0.3		0.99	0.55, 1.85	>0.9		1.18	0.92, 1.51	0.2

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
GCSE/equivalent		0.75	0.45, 1.26	0.3		0.77	0.60, 0.99	0.037		0.57	0.35, 0.94	0.029		0.77	0.62, 0.96	0.020
No qualification		0.35	0.12, 0.93	0.039		1.19	0.82, 1.76	0.4		0.78	0.34, 1.94	0.6		1.01	0.72, 1.44	>0.9

Supplementary Table 5:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
Sex	457				1,515				454				1,972			
Male		—	—			—	—			—	—			—	—	
Female		0.99	0.44, 2.21	>0.9		0.75	0.58, 0.96	0.023		0.95	0.60, 1.48	0.8		0.79	0.63, 0.99	0.044
Age					1,515	1.02	1.01, 1.03	<0.001	454	1.01	0.94, 1.09	0.8	1,972	1.03	1.02, 1.04	<0.001
Self-reported SSt	457				1,515				454				1,972			
Never		—	—			—	—			—	—			—	—	
Former		0.00	0.00, 0.00	<0.001		0.05	0.04, 0.07	<0.001		0.11	0.06, 0.18	<0.001		0.03	0.02, 0.04	<0.001
Current		0.08	0.02, 0.25	<0.001		0.19	0.13, 0.27	<0.001		0.51	0.26, 1.03	0.058		0.20	0.14, 0.28	<0.001
Socioeconomic classification	457				1,515				454				1,972			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		1.51	0.61, 3.96	0.4		0.98	0.71, 1.36	>0.9		0.91	0.52, 1.59	0.7		1.05	0.78, 1.40	0.8
Routine		1.19	0.45, 3.36	0.7		1.20	0.90, 1.60	0.2		1.35	0.80, 2.30	0.3		1.24	0.95, 1.62	0.12

Supplementary Table 6:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
Sex	140				282				103				422			
Male		—	—			—	—			—	—			—	—	
Female		1.70	0.59, 5.31	0.3		1.52	0.89, 2.64	0.13		1.95	0.80, 4.98	0.2		1.48	0.93, 2.38	0.10
Age					282	0.95	0.92, 0.97	<0.001	103	0.89	0.78, 1.03	0.11	422	0.93	0.91, 0.96	<0.001
Socioeconomic classification	140				282				103				422			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		3.65	1.08, 16.7	0.056		1.20	0.62, 2.35	0.6		1.72	0.62, 4.96	0.3		1.52	0.87, 2.68	0.15
Routine		3.71	0.91, 25.1	0.10		2.93	1.56, 5.61	<0.001		5.74	1.92, 19.7	0.003		2.65	1.52, 4.73	<0.001

Supplementary Table 7:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
Sex	317				1,233				351				1,550			
Male		—	—			—	—			—	—			—	—	
Female		1.38	0.88, 2.18	0.2		0.87	0.69, 1.11	0.3		0.83	0.52, 1.31	0.4		0.97	0.79, 1.19	0.8
Age					1,233	1.02	1.01, 1.03	<0.001	351	1.01	0.94, 1.09	0.7	1,550	1.02	1.01, 1.03	<0.001
Socioeconomic classification	317				1,233				351				1,550			

	NCDS			USM2			USM2 (aged 49-59)			Overall		
Characteristic	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**	**N**	**OR**	**95% CI**	**p-value**
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.88	0.53, 1.49	0.6		1.01	0.75, 1.37	>0.9		1.01	0.57, 1.80	>0.9		0.97	0.75, 1.26	0.8
Routine		0.76	0.41, 1.40	0.4		0.79	0.61, 1.04	0.091		0.85	0.50, 1.44	0.5		0.81	0.64, 1.04	0.092

Supplementary Table 8:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**
Self-reported SSt	462				2,171				506				2,633			
Never		—	—			—	—			—	—			—	—	
Former		-0.05	-0.19, 0.08	0.4		0.03	-0.03, 0.08	0.3		-0.05	-0.15, 0.06	0.4		0.00	-0.05, 0.05	0.9
Current		0.22	0.09, 0.35	0.001		0.21	0.14, 0.27	<0.001		0.25	0.12, 0.37	<0.001		0.22	0.16, 0.28	<0.001
Educational attainment	462				2,171				506				2,633			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		-0.05	-0.22, 0.12	0.5		0.07	0.00, 0.13	0.037		0.12	-0.01, 0.25	0.074		0.05	-0.01, 0.11	0.12
GCSE/equivalent		0.17	0.05, 0.30	0.007		0.07	0.01, 0.13	0.022		0.10	-0.01, 0.21	0.079		0.12	0.07, 0.18	<0.001
No qualification		0.21	0.00, 0.42	0.051		0.30	0.23, 0.38	<0.001		0.23	0.06, 0.40	0.009		0.29	0.21, 0.36	<0.001

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**	**N**	**Beta**	**95% CI**	**p-value**
DNAme smokp SSt	462				2,171				506				2,633			
Never		—	—			—	—			—	—			—	—	
Former		0.01	-0.66, 0.68	>0.9		0.14	0.08, 0.20	<0.001		0.04	-0.07, 0.15	0.5		0.10	0.05, 0.16	<0.001
Current		0.27	0.15, 0.39	<0.001		0.29	0.22, 0.36	<0.001		0.28	0.15, 0.41	<0.001		0.30	0.24, 0.36	<0.001
Educational attainment	462				2,171				506				2,633			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		-0.05	-0.21, 0.12	0.6		0.07	0.01, 0.13	0.029		0.12	0.00, 0.25	0.058		0.05	-0.01, 0.11	0.088
GCSE/equivalent		0.17	0.04, 0.29	0.009		0.07	0.01, 0.13	0.024		0.11	0.00, 0.22	0.050		0.12	0.06, 0.17	<0.001
No qualification		0.18	-0.03, 0.39	0.090		0.27	0.19, 0.35	<0.001		0.23	0.07, 0.40	0.007		0.26	0.18, 0.33	<0.001
McCartney MS	462	0.09	0.05, 0.14	<0.001	2,171	0.12	0.09, 0.15	<0.001	506	0.12	0.07, 0.17	<0.001	2,633	0.12	0.10, 0.15	<0.001
Educational attainment	462				2,171				506				2,633			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		-0.04	-0.21, 0.12	0.6		0.06	0.00, 0.13	0.047		0.12	-0.01, 0.25	0.063		0.05	-0.01, 0.11	0.12
GCSE/equivalent		0.17	0.04, 0.29	0.010		0.07	0.00, 0.13	0.034		0.11	-0.01, 0.22	0.066		0.11	0.05, 0.16	<0.001
No qualification		0.16	-0.05, 0.37	0.13		0.27	0.19, 0.35	<0.001		0.21	0.04, 0.38	0.015		0.25	0.18, 0.32	<0.001

Supplementary Table 9:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
Self-reported SSt	462				2,120				500				2,582			
Never		—	—			—	—			—	—			—	—	
Former		-0.10	-0.36, 0.16	0.5		0.05	-0.06, 0.15	0.4		0.14	-0.08, 0.35	0.2		0.04	-0.05, 0.14	0.4
Current		0.22	-0.04, 0.48	0.094		0.37	0.24, 0.50	<0.001		0.42	0.15, 0.70	0.002		0.32	0.21, 0.44	<0.001
Educational attainment	462				2,120				500				2,582			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		0.06	-0.27, 0.39	0.7		0.12	-0.01, 0.24	0.062		0.30	0.03, 0.57	0.027		0.12	0.00, 0.24	0.043
GCSE/equivalent		0.29	0.04, 0.54	0.021		0.10	-0.02, 0.22	0.10		0.15	-0.09, 0.39	0.2		0.11	0.00, 0.21	0.051
No qualification		0.63	0.22, 1.0	0.003		0.45	0.30, 0.60	<0.001		0.60	0.24, 1.0	0.001		0.48	0.34, 0.62	<0.001
DNAme-predicted smokp SSt	462				2,120				500				2,582			
Never		—	—			—	—			—	—			—	—	
Former		0.16	-1.2, 1.5	0.8		0.26	0.15, 0.37	<0.001		0.05	-0.18, 0.29	0.6		0.30	0.19, 0.41	<0.001
Current		0.31	0.08, 0.54	0.009		0.57	0.44, 0.71	<0.001		0.46	0.19, 0.73	<0.001		0.47	0.35, 0.58	<0.001
Educational attainment	462				2,120				500				2,582			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		0.07	-0.26, 0.40	0.7		0.12	-0.01, 0.24	0.063		0.31	0.04, 0.58	0.025		0.12	0.00, 0.24	0.044
GCSE/equivalent		0.28	0.03, 0.52	0.027		0.09	-0.03, 0.21	0.13		0.16	-0.07, 0.40	0.2		0.10	0.00, 0.21	0.059

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
No qualification		0.59	0.18, 1.0	0.005		0.38	0.22, 0.53	<0.001		0.59	0.24, 0.95	0.001		0.41	0.27, 0.56	<0.001
McCartney	462	0.10	0.01, 0.18	0.034	2,120	0.23	0.17, 0.28	<0.001	500	0.22	0.11, 0.33	<0.001	2,582	0.16	0.11, 0.20	<0.001
Educational attainment	462				2,120				500				2,582			
Higher qualification		—	—			—	—			—	—			—	—	
A-level/equivalent		0.07	-0.26, 0.40	0.7		0.10	-0.02, 0.23	0.10		0.30	0.03, 0.56	0.030		0.12	0.00, 0.24	0.047
GCSE/equivalent		0.28	0.04, 0.53	0.026		0.08	-0.04, 0.20	0.2		0.14	-0.09, 0.38	0.2		0.09	-0.01, 0.20	0.083
No qualification		0.58	0.16, 1.0	0.007		0.38	0.23, 0.53	<0.001		0.54	0.19, 0.89	0.003		0.44	0.30, 0.59	<0.001

Supplementary Table 10:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
Self-reported SSt	457				1,497				453				1,954			
Never		—	—			—	—			—	—			—	—	
Former		-0.01	-0.14, 0.12	0.9		0.04	-0.03, 0.10	0.2		-0.01	-0.11, 0.10	>0.9		0.02	-0.04, 0.07	0.6
Current		0.29	0.16, 0.41	<0.001		0.24	0.16, 0.31	<0.001		0.28	0.15, 0.41	<0.001		0.28	0.21, 0.35	<0.001
Socioeconomic classification	457				1,497				453				1,954			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.02	-0.10, 0.15	0.7		0.05	-0.03, 0.12	0.2		0.02	-0.10, 0.14	0.7		0.05	-0.02, 0.11	0.2

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
Routine		0.11	-0.04, 0.25	0.14		0.04	-0.03, 0.10	0.3		0.06	-0.05, 0.18	0.3		0.02	-0.04, 0.08	0.6
DNAme-predicted smokp SSt	457				1,497				453				1,954			
Never		—	—			—	—			—	—			—	—	
Former		-0.01	-0.67, 0.66	>0.9		0.14	0.07, 0.21	<0.001		0.05	-0.06, 0.16	0.4		0.08	0.01, 0.15	0.025
Current		0.33	0.21, 0.45	<0.001		0.30	0.22, 0.38	<0.001		0.32	0.19, 0.44	<0.001		0.35	0.28, 0.42	<0.001
Socioeconomic classification	457				1,497				453				1,954			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.04	-0.08, 0.16	0.5		0.04	-0.03, 0.11	0.3		0.03	-0.09, 0.14	0.7		0.04	-0.02, 0.11	0.2
Routine		0.12	-0.02, 0.26	0.091		0.03	-0.03, 0.10	0.3		0.07	-0.04, 0.18	0.2		0.02	-0.04, 0.08	0.6
McCartney MS	457	0.12	0.07, 0.16	<0.001	1,497	0.13	0.10, 0.17	<0.001	453	0.14	0.09, 0.19	<0.001	1,954	0.15	0.12, 0.18	<0.001
Socioeconomic classification	457				1,497				453				1,954			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.04	-0.09, 0.16	0.6		0.04	-0.03, 0.11	0.3		0.03	-0.09, 0.15	0.6		0.04	-0.02, 0.10	0.2
Routine		0.11	-0.03, 0.25	0.12		0.03	-0.04, 0.09	0.4		0.06	-0.05, 0.17	0.3		0.01	-0.05, 0.07	0.7

Supplementary Table 11:

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
Self-reported SSt	457				1,461				446				1,918			
Never		—	—			—	—			—	—			—	—	
Former		-0.02	-0.28, 0.24	0.9		0.04	-0.08, 0.16	0.5		0.16	-0.05, 0.38	0.14		0.04	-0.07, 0.15	0.5
Current		0.34	0.09, 0.59	0.009		0.35	0.20, 0.50	<0.001		0.40	0.12, 0.67	0.005		0.32	0.19, 0.45	<0.001
Socioeconomic classification	457				1,461				446				1,918			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.03	-0.21, 0.28	0.8		0.05	-0.09, 0.19	0.5		0.04	-0.21, 0.29	0.8		0.04	-0.08, 0.16	0.5
Routine		0.13	-0.15, 0.41	0.4		0.13	0.00, 0.26	0.047		0.28	0.05, 0.51	0.016		0.16	0.04, 0.27	0.007
DNAme-predicted smokp SSt	457				1,461				446				1,918			
Never		—	—			—	—			—	—			—	—	
Former		0.11	-1.2, 1.4	0.9		0.22	0.08, 0.35	0.002		0.11	-0.12, 0.34	0.3		0.26	0.13, 0.39	<0.001
Current		0.38	0.15, 0.61	0.001		0.48	0.32, 0.64	<0.001		0.44	0.17, 0.71	0.002		0.41	0.28, 0.54	<0.001
Socioeconomic classification	457				1,461				446				1,918			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.05	-0.19, 0.30	0.7		0.04	-0.10, 0.17	0.6		0.04	-0.21, 0.28	0.8		0.04	-0.08, 0.16	0.5
Routine		0.15	-0.13, 0.43	0.3		0.12	-0.01, 0.24	0.068		0.29	0.06, 0.52	0.012		0.15	0.04, 0.27	0.010

	NCDS				USM2				USM2 (aged 49-59)				Overall			
Characteristic	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**	**N**	**Beta**	**95% CI**	**p**
McCartney MS	457	0.13	0.05, 0.22	0.002	1,461	0.20	0.14, 0.26	<0.001	446	0.20	0.10, 0.31	<0.001	1,918	0.14	0.09, 0.19	<0.001
Socioeconomic classification	457				1,461				446				1,918			
Management & professional		—	—			—	—			—	—			—	—	
Intermediate		0.05	-0.19, 0.29	0.7		0.03	-0.10, 0.17	0.6		0.04	-0.21, 0.28	0.8		0.04	-0.08, 0.16	0.5
Routine		0.14	-0.14, 0.42	0.3		0.11	-0.01, 0.24	0.081		0.27	0.05, 0.50	0.019		0.16	0.04, 0.27	0.007

Supplementary Table 12:

	NCDS						USM1						USM2					
Model	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
CpG																		
cg00840791	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-
cg01044314	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-
cg01072106	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-
cg01297684	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-
cg02650017	-	+	+	-	-	+	+		-	-		-	+	+	-	-	-	-
cg03067296	+	-	-	+	+	-	-		-	-		-	+	-	-	-	+	+
cg03606915	+	+	-	+	+	+	+		-	+		-	-	-	+	-	-	-
cg03957124	-	-	+	-	-	-	-		+	-		+	+	-	-	-	-	-
cg04381163	+	+	+	-	-	+	+		-	-		-	-	-	+	+	+	-
cg04389058	+	-	-	+	+	-	-		-	-		-	-	+	-	+	-	-
cg04725636	+	+	+	-	-	-	-		-	-		-	-	-	-	-	-	+
cg04937481	-	-	-	-	-	+	+		+	+		-	-	-	-	+	+	-
cg06019998	+	+	-	-	-	-	-		-	-		+	-	+	-	-	-	-
cg09063556	+	+	-	+	+	+	+		-	-		-	-	-	-	-	-	-
cg09178900	-	-	+	-	-	+	-		-	-		-	+	-	-	-	-	+
cg10452282	+	-	-	+	+	+	-		-	-		-	-	-	-	-	-	-
cg10705487	-	-	-	+	+	+	-		-	-		-	-	-	+	-	-	-
cg11047325	-	-	+	-	-	-	+		-	-		-	-	-	-	-	-	-
cg11551560	-	+	-	-	-	-	+		-	-		-	+	+	-	-	+	-
cg11832534	+	+	+	+	+	-	+		+	+		-	-	-	-	-	-	+
cg11902329	-	-	+	+	+	+	+		-	-		-	+	-	+	+	-	-
cg12170787	+	-	+	-	-	-	-		-	-		+	-	-	-	-	-	-
cg12992827	-	+	-	-	-	-	+		-	-		-	-	+	-	-	+	-
cg13165240	-	-	-	-	-	-	-		-	-		-	-	+	+	-	-	+

	NCDS						USM1						USM2					
Model	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
cg13343932	-	-	-	-	-	+	-		-	-		-	+	-	-	+	+	-
cg13373048	-	+	-	+	+	-	-		-	-		-	-	-	-	+	-	-
cg13781414	-	-	-	+	+	-	-		-	-		-	-	-	-	+	-	-
cg14216476	-	-	-	-	-	+	-		-	-		-	-	-	-	-	-	-
cg14887853	-	-	-	-	-	-	+		+	-		-	-	-	-	-	-	-
cg15060905	+	+	+	+	+	-	-		-	-		-	-	-	-	-	-	-
cg15251256	+	-	+	-	-	-	-		-	-		-	-	-	-	-	+	-
cg15839964	+	-	-	-	-	-	-		-	+		+	+	+	+	-	+	-
cg16531578	-	+	+	+	+	-	+		+	+		-	-	+	-	-	-	+
cg16713119	+	-	+	-	-	+	+		-	+		+	-	-	-	-	-	-
cg17501210	-	-	-	-	-	-	-		+	-		-	+	-	-	-	-	+
cg18181703	+	+	-	-	-	+	-		+	-		+	-	-	+	-	-	+
cg18978030	-	+	-	+	+	-	-		-	-		+	-	+	+	-	-	-
cg19541622	-	-	+	-	-	+	-		-	+		-	-	-	-	-	+	-
cg19748455	-	-	-	-	-	-	+		-	-		-	-	-	-	-	-	-
cg20842915	-	+	+	-	-	-	-		-	-		-	-	-	+	+	-	-
cg20995564	-	-	-	-	-	-	+		+	+		+	+	-	-	-	-	-
cg21566642	-	-	-	-	-	-	-		-	+		-	-	-	-	-	-	+
cg22652934	-	+	-	+	+	+	-		+	-		+	-	-	-	-	-	+
cg22995183	-	-	-	+	+	-	-		-	-		-	+	-	+	+	-	-
cg23248055	+	-	+	-	-	-	-		-	+		-	+	-	-	+	-	+
cg23320029	-	-	-	+	+	-	+		+	+		+	+	+	+	-	+	-
cg23688299	-	-	-	-	-	+	-		+	-		+	-	+	+	+	+	+
cg23842572	-	-	+	-	-	-	-		+	+		-	-	+	-	+	+	-
cg24298280	-	-	-	-	-	-	-		-	-		+	+	-	-	+	-	-
cg24619988	-	+	-	-	-	+	+		+	+		-	-	+	+	-	+	-

	NCDS						USM1						USM2					
Model	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
cg26034658	-	-	+	-	-	+	-		-	+		+	+	-	-	-	-	+
cg26227957	+	+	-	-	-	+	-		+	-		+	-	+	-	+	-	-
cg26457483	+	-	-	-	-	-	+		+	-		+	+	-	+	-	+	+
cg26867393	-	-	-	-	-	-	+		+	+		-	-	-	+	-	-	+
cg00490406	-	-	-	+	+	-	-		-	+		+	-	-	+	-	-	-
cg07252680	-	+	+	-	-	-	-		+	-		+	+	+	-	+	+	+
cg07573872	-	-	-	+	+	-	-		-	-		-	-	+	+	+	+	+
cg09349128	-	-	-	-	-	-	-		+	+		+	-	+	-	+	+	+
cg10636246	+	-	-	-	-	+	+		-	-		-	+	+	+	-	+	-
cg18608055	-	+	+	+	+	-	-		+	+		+	+	-	-	+	-	-
cg24499891	+	+	+	+	+	+	+		-	+		-	-	+	+	+	+	+
cg26416615	-	+	+	-	-	+	+		-	+		-	-	-	-	-	-	-

Supplementary Table 13:

	NCDS						USM1						USM2					
Model	**5**	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
CpG																		
cg00840791	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
cg01044314	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
cg01072106	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
cg01297684	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
cg02650017	-	+	+	-	-	+	+	+	-	-	-	-	+	+	-	-	-	-
cg03067296	+	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-	+	+
cg03606915	+	+	-	+	+	+	+	-	-	+	-	-	-	-	+	-	-	-
cg03957124	-	-	+	-	-	-	-	-	+	-	-	+	+	-	-	-	-	-
cg04381163	+	+	+	-	-	+	+	+	-	-	+	-	-	-	+	+	+	-
cg04389058	+	-	-	+	+	-	-	-	-	-	-	-	-	+	-	+	-	-
cg04725636	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
cg04937481	-	-	-	-	-	+	+	-	+	+	-	-	-	-	-	+	+	-
cg06019998	+	+	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-
cg09063556	+	+	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
cg09178900	-	-	+	-	-	+	-	-	-	-	-	-	+	-	-	-	-	+
cg10452282	+	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
cg10705487	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-
cg11047325	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
cg11551560	-	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	+	-
cg11832534	+	+	+	+	+	-	+	-	+	+	+	-	-	-	-	-	-	+
cg11902329	-	-	+	+	+	+	+	-	-	-	-	-	+	-	+	+	-	-
cg12170787	+	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
cg12992827	-	+	-	-	-	-	+	-	-	-	-	-	-	+	-	-	+	-
cg13165240	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+

	NCDS						USM1						USM2					
Model	**5**	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
cg13343932	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	+	+	-
cg13373048	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	+	-	-
cg13781414	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	-	-
cg14216476	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-
cg14887853	-	-	-	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-
cg15060905	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
cg15251256	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-
cg15839964	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	+	-
cg16531578	-	+	+	+	+	-	+	+	+	+	-	-	-	+	-	-	-	+
cg16713119	+	-	+	-	-	+	+	-	-	+	+	+	-	-	-	-	-	-
cg17501210	-	-	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-	+
cg18181703	+	+	-	-	-	+	-	-	+	-	+	+	-	-	+	-	-	+
cg18978030	-	+	-	+	+	-	-	-	-	-	-	+	-	+	+	-	-	-
cg19541622	-	-	+	-	-	+	-	-	-	+	-	-	-	-	-	-	+	-
cg19748455	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
cg20842915	-	+	+	-	-	-	-	+	-	-	-	-	-	-	+	+	-	-
cg20995564	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	-
cg21566642	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	+
cg22652934	-	+	-	+	+	+	-	-	+	-	-	+	-	-	-	-	-	+
cg22995183	-	-	-	+	+	-	-	+	-	-	+	-	+	-	+	+	-	-
cg23248055	+	-	+	-	-	-	-	-	-	+	-	-	+	-	-	+	-	+
cg23320029	-	-	-	+	+	-	+	+	+	+	-	+	+	+	+	-	+	-
cg23688299	-	-	-	-	-	+	-	-	+	-	+	+	-	+	+	+	+	+
cg23842572	-	-	+	-	-	-	-	+	+	+	+	-	-	+	-	+	+	-
cg24298280	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	-	-
cg24619988	-	+	-	-	-	+	+	+	+	+	+	-	-	+	+	-	+	-

	NCDS						USM1						USM2					
Model	**5**	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**	5	**6**	**7**	**8**	**9**	**10**
cg26034658	-	-	+	-	-	+	-	+	-	+	+	+	+	-	-	-	-	+
cg26227957	+	+	-	-	-	+	-	+	+	-	+	+	-	+	-	+	-	-
cg26457483	+	-	-	-	-	-	+	+	+	-	-	+	+	-	+	-	+	+
cg26867393	-	-	-	-	-	-	+	-	+	+	-	-	-	-	+	-	-	+
cg00490406	-	-	-	+	+	-	-	-	-	+	+	+	-	-	+	-	-	-
cg07252680	-	+	+	-	-	-	-	-	+	-	-	+	+	+	-	+	+	+
cg07573872	-	-	-	+	+	-	-	-	-	-	-	-	-	+	+	+	+	+
cg09349128	-	-	-	-	-	-	-	+	+	+	+	+	-	+	-	+	+	+
cg10636246	+	-	-	-	-	+	+	-	-	-	-	-	+	+	+	-	+	-
cg18608055	-	+	+	+	+	-	-	+	+	+	+	+	+	-	-	+	-	-
cg24499891	+	+	+	+	+	+	+	+	-	+	+	-	-	+	+	+	+	+
cg26416615	-	+	+	-	-	+	+	+	-	+	+	-	-	-	-	-	-	-