Author:
Haworth, Simon

Title:
The use of genetic data in dental epidemiology to explore the causes and consequences of caries and periodontitis

The use of genetic data in dental epidemiology to explore the causes and consequences of caries and periodontitis

Simon Haworth

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of PhD in the Faculty of Health Sciences, September 2019

Word count: 77,472 words

Abstract

The major dental diseases are caries and periodontitis. These two common conditions are important public health problems and have complex, multifactorial aetiology. Study designs which use genetic data can provide evidence about the molecular and biological basis of disease and also help prioritize modifiable risk factors which have causal relevance for disease. To date these approaches have had limited success in dental epidemiology, possibly due to the lack of large studies with genetic data and dental phenotypes.

Through a theoretical review and a pair of applied illustrations, I argue that questionnaire-derived and index-linked dental data provide a valuable resource for the application of modern epidemiological methods. Using these data for genetic association discovery, the main findings are novel risk loci for caries in both adult and paediatric populations, and evidence suggesting that the genetic risk factors for caries and periodontitis are partially overlapping with a range of other health traits.

These newly discovered risk loci can act as proxies for variation in dental disease experience using current methods for causal effect estimation. I test the reciprocal hypotheses that dental diseases have downstream effects on cardio-metabolic traits, and that metabolic traits influence risk of dental diseases, finding some evidence supporting existing beliefs that dental diseases may have undesirable downstream effects on health.

Together, the results suggest that studies which use genetic data will have an important role in the future of dental epidemiology and can help improve understanding of both the molecular and broader health and social aetiology of caries and periodontitis. Unlocking the full potential of these methods will require the community to support still-larger studies and adopt modern working practices in dental epidemiology.

Dedication

To my family – Christopher, Katherine and Jennifer

Acknowledgements

Personal thanks

First, I would like to thank my supervisors for their enthusiasm and energy throughout this project. I am very grateful for the opportunity to work on this project, and would particularly like to thank Ingegerd Johansson and Paul Franks (representing the GLIDE consortium), and Nic Timpson and Steve Thomas (representing the University of Bristol), who encouraged me to plan a collaborative project and supported me at every stage of the project. As well as this day to day support from specific individuals, I am grateful for the institutional support from Wellcome (who funded my research training through the clinical research training fellowship scheme), the University of Bristol, Umeå University and Lund University. I would like to thank Jonathan Sandy, George Davey Smith and Debbie Lawlor for their feedback and guidance both through the formal student review process and informal mentorship. Finally, I would like to thank all the research participants who contributed their data to this project for their generosity.

Scientific input

Many people have helped with specific parts of this project by contributing analysis or providing advice on analysis or interpretation. It is not possible to acknowledge all these personally, but I would like to highlight a small number of people who have been particularly helpful. Dmitry Shungin made major contributions to GWAS meta-analysis in GLIDE, Tom Dudding helped plan and perform analysis in ALSPAC, Ruth Mitchell provided support for analysis of genetic data in UK Biobank, Margaret May provided advice on longitudinal modelling approaches, Min-Jeong Shin provided access to KNHANES data and an analytical team to perform analysis, Erik Ingelssson provided advice on cardiometabolic traits, Justin van der Tas, John Shaffer and Kimon Divaris provided particularly helpful input during manuscript preparation and revision. A full description of the contributions of each person is provided in the publications listed in section 1.6.

Support for underlying resources

Data described in this dissertation were contributed by many studies. Specific acknowledgements for each study are provided in the relevant results chapter or appendix.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ……

DATE:……

Table of contents, list of tables and illustrative material

Table of Contents

List of tables and illustrative material

List of abbreviations

1KG: The 1000 genomes project
ALSPAC: The Avon longitudinal study of parents and children
ARIC: The atherosclerosis risk in communities study
BBJ: Biobank Japan
BMI: Body mass index
CEU: Utah Americans of European descent
CHR: Chromosome
COHRA: The center for oral health research in rural Appalachia
COPSAC: Copenhagen prospective studies on asthma in childhood
CPI: Community periodontal index
DFS: Decayed and filled tooth surfaces
DFSS: Decayed and filled tooth surfaces per available surface
DGA: Dental general anaesthesia
DMFS: Decayed, missing and filled tooth surfaces
DMFT: Decayed, missing and filled teeth
DNBC: Danish national birth cohort
DRDR: The dental registry and DNA repository
EA: Effect allele
EAF: Effect allele frequency
eQTL: Expression quantitative trait locus/loci
ExAC: The exome aggregation consortium
FDR: False discovery rate
FEPT: Filled and extracted permanent teeth
GCTA: Genome-wide complex trait analysis
GENEVA: Gene environment association studies initiative
GENR: The generation R study
GERP: Genomic evolutionary rate profiling
GINIplus: German infant study on the influence of nutrition intervention plus air pollution and genetics on allergy development
GLIDE: Gene-lifestyle interactions in dental endpoints
GSMR: Generalized summary Mendelian randomization

GTEx: The genotype-tissue expression project

GREML: Genome-based restricted maximum likelihood

GWAS: Genome-wide association study

HbA1c: Haemoglobin A1C (glycated haemoglobin)

HCHS/SOL: The Hispanic community health study / study of Latinos

HDL-c: High density lipoprotein cholesterol

HLA: Human leukocyte antigen

HWE: Hardy-Weinberg equilibrium

IBD: Identity by descent

IBS: Identity by state

IFS: The Iowa fluoride study

IHS: The Iowa head start study

HR: Hazard ratio

INDEL: Insertion / deletion

INFO: Imputation quality score

IRR: Incidence rate ratio

IVW: Inverse-variance weighted

KB: Kilobase(s)

KNHANES: The Korea national health and nutrition examination survey

LD: Linkage disequilibrium

LDL-c: Low density lipoprotein cholesterol

LDSR: Linkage disequilibrium score regression

LISA: Lifestyle related factors, immune system and the development in allergies in East and West Germany study

MA: Meta-analysis

MAF: Minor allele frequency

MDC: The Malmö diet cancer study

MDS: Multidimensional scaling

mM: Millimoles per litre

MR: Mendelian randomization

MRI: Magnetic resonance imaging

MSigDB: The molecular signatures database

NFBC1966: The Northern Finland birth cohort 1966

NHS: The national health service

Nteeth: Number of teeth

OR: Odds ratio

PANIC: Physical activity and nutrition in children study

PC: Principal component(s)

PCA: Principal component analysis

PMID: PubMed identifier

POS: Genomic position (using the genome reference consortium human build 37 unless otherwise indicated)

QC: Quality control

RAINE: The Western Australia pregnancy cohort study

RefSeq: The national center for biotechnology information reference sequences database

RSID: dbSNP reference SNP number

SHIP: The study of health in Pomerania

SNP: Single nucleotide polymorphism

STR: The Swedish twin register

TMDUAGP: The Tokyo medical and dental university aggressive periodontitis study

UKB: UK biobank

VIP: Västerbotten intervention program

WGHS: The women's genome health study

WHO: World health organization

WHR *adj* BMI: Waist to hip ratio adjusted for body mass index

WHS: The women's health study

YFS: Young Finns study

ZINB: Zero-inflated negative binomial

Chapter 1: Introduction

1.1: The aetiology of caries and periodontitis

Dental caries, the most prevalent global disease[1], refers to the demineralization of the dental hard tissues of enamel, dentine and cementum by acidic metabolic by-products created by bacteria in dental plaque. These by-products lead to low pH and mineral undersaturation in fluids surrounding the hard tissues, causing reversable damage which can be recovered through remineralization or may progress, through proteolytic destruction of organic tooth structures, to irreversible damage and cavitation[2].

This notion of an interplay between dietary free sugars, dental plaque and cycles of demineralization and remineralization underlies traditional models of caries aetiology and approaches to prevention used in clinical practice. The quantity, frequency and type of free sugars in diet affect the available substrate for bacterial metabolism[3] and are a determinant of the characteristics of dental plaque. For example, sucrose may selectively promote the growth of cariogenic species within plaque[4] and provide substrate for the extracellular matrix of plaque[5], leading to a more cariogenic biofilm. Mechanical removal of plaque in conjunction with fluoride-containing toothpaste is effective in reducing incident caries lesions in a systematic review of randomized controlled trials[6]. Fluoride may help modulate the dynamics of caries by altering the chemical[7,8] or physical characteristics of dental hard tissues[9] so that these tissues are less susceptible to damage from acidic by-products produced by dental plaque[10]. Topical fluoride interventions for caries prevention are well supported by systematic reviews of randomized controlled trials[11].

While this traditional aetiological model is well-understood, it is incomplete, and interventions based on this model have not resolved caries as a public health problem. Caries continues to confer a major burden of morbidity[12] and healthcare spending, and is a major contributor to the global cost of dental diseases which was estimated to exceed 540 billion US dollars in 2015[13]. In recent years a range of alternative aetiological models have been proposed which consider factors which are upstream of the chemical events occurring in plaque biofilm and attempt to place caries in a broader health and social context. One approach to obtain this broader context involves examining the association between caries scores and a range of factors aiming to capture biological, behavioural, socio-economic or

psychological conditions across the life course[14,15]. Other authors have studied specific parts of these emerging aetiological models, for example investigating how and why socio-economic status might influence dental caries[16] and factors upstream of caries such as accessing preventive dental services[17]. The shift in focus from viewing caries in chemical to social terms is mirrored by shifts in clinical perception, for example regarding caries as a person-level disease rather than a tooth-level lesion. This revised focus has been described as a 'complete changeover' in disease perception[18]. This change in disease understanding theoretically provides an opportunity for novel and more effective interventions. Conversely, as the number of putative risk factors for caries increases, prioritizing the modifiable risk factors which are most relevant becomes difficult without evidence for causal association, which is currently unavailable for many traits for example in education[19].

In developing these aetiological models, genetic evidence may have a pivotal role in highlighting previously overlooked molecular and biological processes which contribute to dental caries, and in prioritizing modifiable risk factors which have causal effects on dental health. These motivations are described further in sections 1.2 and 1.4, below.

Periodontitis refers to the progressive destruction of tooth-supporting tissues by a chronic and inappropriate inflammatory response to peri-gingival and sub-gingival dental plaque[20]. The disease is traditionally seen to result from the interaction between the plaque biofilm and the host response in a model formalized by Page & Kornman in 1997[21]. At a similar time, the plaque biofilm was thought to contain periodontal pathogens and commensals, for example a disease-associated microbial complex consisting of *porphyromonas gingivalis*, *tannerella forsythia* and *treponema denticola* was interpreted as a pathogenic microbial complex[22]. In clinical management, concepts tended to focus on attempts to restore a more healthy biofilm by mechanical[23,24] or chemical[25] disruption of plaque and topical or systemic antibiotics[26], rather than attempts to modulate the host. In combination a range of biological mechanisms contributing to periodontal inflammation and destruction in the host have been described, including impaired neutrophil chemotaxis[27] and inappropriate neutrophil activation[28], poor epithelial barrier and defensive functions[29,30] and dysregulation of osteoclastic signalling mechanisms[31].

In recent years authors have argued that this model placed too much emphasis on periodontal pathogens and insufficient emphasis on the upstream determinants of the host response[32]. While there are detectable microbiological differences between clinically healthy and periodontally-compromised sites, no single pathogen is consistently isolated and the notion of a small highly-pathogenic complex is being replaced by a more general model of microbial synergy and dysbiosis[33]. It is becoming clearer that certain oral bacteria are highly specialized to occupy micro-niches within the gingival and periodontal environment[34], so the isolation of microbial complexes which are specific to deep periodontal pockets might partially reflect a consequence rather than a cause of disease. The specific host characteristics which contribute to disease aetiology are themselves likely to be influenced by a range of genetic, lifestyle and environmental features, as well as the presence of co-morbidity[35]. This shift in emphasis from periodontal pathogens to the upstream mediators of host response is now being reflected in newer treatment concepts which aim to modulate the host, for example in randomized controlled trials of dietary interventions which are thought to be anti-inflammatory[36,37].

Analogous to the arguments made for dental caries, the changing perception of periodontitis is creating new opportunities for disease understanding. Conversely, the expanding list of biological pathways which are thought to influence periodontitis and putative risk factors both create challenges in prioritizing the experiments or interventions most likely to be informative to a laboratory scientist or effective in clinical or public health settings. Once again, genetic data may help to prioritize relevant biology, discovering previously un-appreciated aspects of molecular aetiology and distinguishing causal from confounded risk factors.

1.2: Genetic evidence as a method to refine understanding of disease mechanisms and biology

Genetic evidence plays an increasingly important role in the study of health and disease. Knowledge of the genetic variants, gene pathways or regulatory elements which are associated with a disease provides a starting point to explore the underlying biology and mechanisms of disease. Visscher and colleagues[38] highlight obesity[39] and schizophrenia[40] as illustrative examples of traits where novel findings in genetic association studies helped inform detailed laboratory experiments to learn more about the molecular basis of health and disease.

Here genetic associations are used as a flag for relevant biology and the same rationale is used by industry to prioritize potentially relevant proteins or pathways as drug targets. In 2015 it was reported that drugs targeting molecules with genetic evidence for relevance are twice as likely to succeed in trials and obtain an approved indication compared to drugs targeting molecules which lack genetic evidence[41], and this finding is sustained in more recent analysis[42]. As well as suggesting drug targets, genetic evidence can also help understand the mechanisms of drug action, and help distinguish on-target from off-target side effects[43].

In their simplest form these applications only require an association signal as a pointer to a relevant genomic region for wet laboratory follow-up. While genetic association signals do not provide molecular or mechanistic evidence on their own, recent methods have been developed which integrate genetic association signals for a disease with external sources of functional or molecular data in order to generate mechanistic hypotheses. As an example, gene expression data from detailed molecular studies such as the Genotype-Tissue Expression (GTEx)[44] project can be combined with disease associations to prioritize potentially relevant gene transcripts in potentially-relevant tissues[45,46]. The potential utility and limitations of these methods are explored later in this dissertation, but this concept is introduced here to illustrate a point that genetic data can now be used to infer more about the biology of a trait than was previously possible in the candidate gene era.

Genetic data is therefore a valuable tool in dissecting the aetiology of a single disease, but genetic data is increasingly important for broader epidemiological inference such as the relationship between two diseases. Here the ability to estimate genetic correlation[47,48] using

summary statistics of genetic association studies is helping to re-organize the disease phenome by identifying diseases with similar determinants (or differently labelled diseases whose diagnostic criteria capture a similar underlying phenotype). Likewise, it is possible to use genetic data to investigate questions of selection and population history and estimate the degree of bias within an experiment[49]. Notably, these methods require data at a genome-wide scale and have only become available since the shift to examining diseases through genome-wide methods, which have other advantages over candidate gene association studies as described below.

1.3: Genetic association studies for dental caries and periodontitis

Given the imperfect understanding of the aetiology of caries and periodontitis, there is a motivation for using genetic data to learn more about the biological events which lead to disease presentation. In recent years the genome-wide association study (GWAS) has emerged as the primary modality for identifying common genetic variation which is associated with a trait or disease[50]: it has been enormously successful, with approximately 10,000 robust genetic associations identified during the first 10 years of the GWAS method[38]. This approach is well suited to identifying common genetic variants with modest effect sizes, lending itself to application in traits with polygenic architecture[51]. Conversely, there are specific limitations to methods which focus on this portion of the genetic landscape. As examples, identifying the biologically causal variation can be a challenge for association signals identified in GWAS as the tag variants often have modest effects and are located in non-coding regions[52,53], effects of uncommon or rare variation typically cannot be detected unless these have large effect sizes or the sample size is extremely large, and co-incident latent structure in the genotypic and phenotypic landscapes of common genetic variation and complex traits has the potential to mis-lead inference[54,55].

Despite these considerations, the GWAS method has been highly successful for many complex traits. Using psychiatric disorders as an example, far more robust genetic risk variants have been identified in poorly understood regions of the genome than in previously-studied candidate genes[50], supporting an argument that the GWAS method in its short history has rendered candidate gene studies essentially obsolete. To date however, the major dental diseases are poorly represented in this literature. Despite being the most prevalent global disease in 1990, 2007 and 2017[56] there are few reported associations for dental caries (Table

1.1) or periodontitis (Table 1.2), and none are consistently seen at genome-wide significance in non-overlapping studies.

The lack of consistent association signals might suggest that the heritability of caries and periodontitis is low, but this is not supported by family-based studies. Conventional estimates of heritability are between 30 and 70% for both diseases[57-63]. It is likely that both the statistical power of existing GWAS studies and a polygenic genetic architecture[51] of dental diseases explain the low discovery yield. In either case larger sample sizes are needed. The limited success of early GWAS investigations for dental diseases sits in the context of a general tendency for studies in dentistry to be small in scale. The failure of many projects from the first round of 'OMICS' studies in dentistry[64] is possibly due to problems with statistical power. While there is a specific need for large sample sizes for GWAS of caries and periodontitis, there is also a more general rationale for larger sample sizes in dental epidemiology which is discussed in detail in chapter 2.

Table 1.1: Summary of findings from GWAS for dental caries traits

Trait	Study	Participants	Population	Lead single variants at $P<5x10^{-8}$ in discovery sample
Early childhood caries	Ballantine et al, 2017[65]	Multi-ethnic study of 212 children	1 US cohort	0
DMFS / DMFT scores	Morrison et al, 2015[66]	11,754 Hispanic/Latino adults	1 US cohort	2 associations, 1 which was shared across both phenotypes and one only seen for DMFS
2 patterns of caries presentation	Zeng et al, 2014[67]	1,006 European ancestry children	2 US cohorts	1 association, specific to one phenotype
2 patterns of caries presentation	Zeng et al, 2013[68]	1,004 European ancestry children and teenagers	1 US cohort	0
5 patterns of caries presentations in adults	Shaffer et al, 2012[69]	920 European ancestry adults	1 US cohort	2 variants, both of which were associated with a single sub-presentation of caries
DMFS and caries severity scores	Wang et al, 2012[70]	7,443 European ancestry adults	5 US cohorts	0
Caries in children	Shaffer et al, 2011[71]	1305 European ancestry children aged 3 to 12 years	3 US cohorts	0

Table 1.2: Summary of findings from GWAS for periodontitis and related traits

Trait	Study	Participants	Population	Lead single variants at $P<5x10^{-8}$ in discovery sample
Chronic periodontitis	Sanders et al, 2016[72]	10,935 Hispanic/Latino adults	1 US cohort	1 association
Multiple traits representing principal components of inflammatory biomarkers, bacterial colonization and clinical status	Offenbacher et al, 2016[73]	975 European ancestry adults	1 US cohort	10 associations with a principal component.
Alveolar bone loss, clinical periodontitis, severe periodontitis	Hong et al, 2015[74]	677 East Asian ancestry adults	1 Korean cohort	0
Periodontitis	Shimizu et al, 2015[75]	2,760 cases, 15,158 controls of East Asian ancestry	2 Japanese cohorts	0
Probing depth	Shaffer et al, 2014[76]	673 European ancestry adults	1 US cohort	0
Chronic periodontitis	Teumer et al, 2013[77]	4,032 European ancestry	2 German cohorts	0
Periodontitis	Divaris et al, 2013[78]	4,504 European ancestry adults	1 US Cohort	0
Three periodontal pathogen colonization traits	Divaris et al, 2012[79]	1,020 European ancestry adults	1 US cohort	0
Aggressive periodontitis	Schaefer et al, 2010[80]	283 cases, 979 controls of European ancestry	1 German cohort	0

1.4: Causal inference is valuable for prioritizing interventions

In both the clinical and public health settings it is important to identify modifiable risk factors with causal relevance in order to prioritize interventions which are likely to be effective. There is an increasing drive to stop providing ineffective clinical interventions[81] which unethically expose patients to side effects without benefit and represent a miss-allocation of resources. Likewise, population-level interventions on risk factors identified in observational studies will only be effective if those risk factors have causal relevance. Ideally the causal relevance of a risk factor would be established at an early stage. As an example, public health bodies have advocated vitamin D supplements for bone health but a recent meta-analysis of 81 randomized controlled trials shows no meaningful effect on bone mineral density, fractures or falls[82]. It is now too late to allocate the resources invested in this advice and these trials to more effective interventions.

In contrast, the use of epidemiological methods to estimate causality was important in the rapid de-prioritization of c-reactive protein as a causal risk factor for cardiovascular outcomes[83-85] and probably prevented study participants or the public being exposed to unhelpful advice or interventions. Testing for causality in relationships between dental diseases and systemic diseases may similarly prioritize relationships where interventions have the potential to improve population health and prevent miss-allocation of resources towards ineffective interventions. In dental epidemiology, the relationships between dental diseases and cardio metabolic traits represent an area which would benefit from clarification afforded by tests for causality and the justification for this field is explained in detail in chapter 4.

While methods have been developed which aim for causal inference without using genetic data[86,87], briefly reviewed in chapter 4, the unique characteristics of genetic data make it a valuable addition for strengthening epidemiological inference[88]. A common feature across all these methods is that their potential for dental epidemiology is recognised but there are few practical applications in this field[89]. One such approach, Mendelian randomization, involves the use of genetic variants as proxies for a risk factor of interest in order to estimate the causal effect of that risk factor on an outcome[90,91]. This is similar to the concept of genocopy/phenocopy – the notion that genetic and environmental exposures which modify a biological process in the same way can lead to the same end result[92]. By extension, a non-modifiable genotype which leads to disease presentation can be used to mirror or proxy the

modifiable environmental risk factors for disease which act through the same pathway. The identification of common genetic variants is not just a useful exercise to understand disease biology, but also a necessary step towards using methods which use genetic data to understand the relationships between different traits in both disease ontology terms and in tests for potentially causal relationships.

In summary, the use of genetic data in dental epidemiology provides an opportunity to learn more about the biological aetiology of dental diseases, while the intersection of genetic data with epidemiological methods provides opportunities to re-assess the relevance of dental diseases as a modifiable risk factor for other health outcomes. To date, there are few reliable association signals for dental caries and periodontitis, possibly reflecting the small sample sizes in existing GWAS investigations, and future studies will require larger sample sizes than have been achieved to date.

This thesis tests a general hypothesis that genetic data can be used to learn more about the causes and consequences of caries and periodontitis, where ‘causes’ means both the biological and molecular basis of disease and modifiable risk factors. To test this hypothesis, I plan to undertake the largest GWAS for dental caries and periodontitis and apply the results of genetic association discovery in epidemiological analyses to test whether dental diseases are modifiable risk factors for cardiovascular outcomes, and whether cardio-metabolic traits influence risk for dental diseases.

1.5: Commentary

Dental caries and periodontitis are the major dental diseases. Genetic data could be used to identify robust genetic association signals which would help explore the molecular mechanisms of caries and periodontitis, potentially leading to improved understanding of disease biology.

In a separate application, genetic association signals can be used in epidemiological methods to test for causal association between dental diseases and other health traits, leading to improved understanding of the causes and consequences of caries and periodontitis.

These applications may provide a valuable source of evidence for refining aetiological models of disease but will require far larger sample sizes than have been achieved by existing studies. As a starting point, it is therefore necessary to consider ways to obtain dental data in large studies. Chapter 2 starts by considering an appropriate balance between sample size, measurement error and phenotypic resolution.

1.6: Projects contributing to this dissertation

This dissertation represents the cumulation of a series of related projects which I carried out during my 3year research training fellowship. Most of these projects have peer-reviewed output in the form of published articles or conference abstracts. The contributions of each of these projects to this dissertation are summarized graphically in Figure 1.1, while the aims of these projects and my role in each is described in Table 1.3.

Figure 1.1: Overview of projects contributing to this dissertation

Projects contributing data or results which are presented in this dissertation are shaded in purple. Projects which include relevant themes, concepts and methods but do not contribute data to this dissertation are shaded in orange.

Table 1.3: Role in projects contributing to this dissertation

Short project title	Project aims	Roles	Peer-reviewed output
Latent structure project	To examine fine-scale latent structure in the genetic data of UK Biobank	Conception, analysis, interpretation, drafting and revising the manuscript	First author publication in *Nature Communications*[55]
Population stratification position piece	To discuss the ongoing relevance of population stratification in the era of biobank-scale studies with genetic data	Interpretation and revising the manuscript	Co-author publication in *Human Genetics*[93]
Measurement error simulation project	To simulate the impact of different degrees of measurement error on inference in dental epidemiology	Conception, analysis and interpretation	Abstract presentation at the International Association for Dental Research general session, 2019
Dental general anaesthesia project	To assess the longer-term health outcomes of people who had dental treatment under general anaesthesia in childhood	Conception, analysis, interpretation, drafting and revising the manuscript	Joint first author publication in the *British Dental Journal*[94]
Tooth loss project	To assess risk factors for tooth loss in a Swedish population and explore the biasing effects of tooth loss on measurement of caries and periodontitis	Conception, analysis in GLIDE, drafting and revising the manuscript	First author publication in *Community Dentistry and Oral Epidemiology*[95]
Head circumference GWAS project	To identify common and low-frequency variation associated with final head circumference	Analysis in ALSPAC, single-trait and multi-trait meta-analysis, interpretation, drafting and revising the manuscript	Joint first author publication in *Nature Communications*[96]
Mouth ulcers GWAS project	To identify genetic risk loci for mouth ulcers	Conception, analysis in UK Biobank, meta-analysis with 23andMe, bioinformatic follow-up, interpretation, drafting and revising the manuscript	Joint first author publication in *Nature Communications*[97]
GLIDE adult GWAS project	To identify genetic risk loci for caries and periodontitis	Conception, analysis in UK Biobank, multi-trait meta-analysis,	Joint first and sole corresponding author of a

		bioinformatic follow-up, drafting and revising the manuscript	publication in *Nature Communications* which presents findings from both these projects[98]
GLIDE cardio-metabolic Mendelian randomization project	To test for causal association between dental disease and cardio-metabolic traits	Conception, analysis, interpretation, drafting and revising the manuscript	
GLIDE paediatric caries GWAS project	To identify common genetic variants which are associated with dental caries in paediatric populations	Conception, analysis in ALSPAC, meta-analysis, bioinformatic follow-up, interpretation, drafting and revising the manuscript	Joint first author publication in *Human Molecular Genetics*[99]
Oral microbiome pilot projects	To explore the feasibility of obtaining detailed molecular phenotypes for participants in oral health research in Sweden	Member of steering committee for GLIDE, interpretation and revising the manuscripts	Co-author publications in *Scientific Reports*[100], *PLoS ONE*[101] and *Nutrients*[102]
Latent groups in Mendelian randomization project	To evaluate whether heterogeneity in two-sample Mendelian randomization may result from the inclusion of clusters of genetic proxies estimating different causal effects	Inception, analysis	Work ongoing
Swedish Twin Register heritability project	To estimate the heritability of caries scores, caries trajectories and different clinical presentations of caries	Member of steering committee for GLIDE, inception, analysis and interpretation	Abstract accepted for the British Society for Oral and Dental Research conference 2019

Chapter 2: Use of questionnaire-derived and index-linked dental phenotypes

Genome-wide association studies for dental diseases will require large sample sizes with dental data. Obtaining these will require a compromise between detailed sampling and measurement strategies possible in small studies, and more pragmatic approaches to sampling and data acquisition, which might yield larger samples at the expense of phenotypic detail. The aim of the first part of this chapter is to explore theoretical considerations around sample size, sample characteristics, phenotypic resolution and measurement error which might inform decisions about The trade-off between sample size and phenotypic detail.

Following this theoretical review, the use of questionnaire-derived and index-linked approaches to obtaining dental data are illustrated through two examples of observational analysis. Both illustrations aim to demonstrate the value of questionnaire-derived and index-linked dental phenotypes in practical application and inform analytical decisions about the handling of these data in GWAS of caries and periodontitis.

2.1: Theoretical considerations relating to sampling and measurement

2.1.1: The importance of sample size

Large sample sizes are desirable for all epidemiological studies and essential for many. The aim is to estimate a basic parameter such as disease prevalence, and the precision of this estimate is fundamentally tied to both the in-sample prevalence and the square root of the sample size. In the absence of a precise estimate, we have little idea where the true prevalence lies, leading to uninformative statements about any single group and unhelpful comparisons between different groups.

The same is true for the relationship between two traits or a putative risk factor and disease. Here, power to detect a relationship depends on the size of the effect, sample size and the approach taken to account for multiple testing. In the context of GWAS studies, where the effects of common genetic variants are typically small and the multiple testing burden is high, large samples sizes are needed to reliably detect associations between a genetic risk factor and trait, but even in other contexts large sample sizes are desirable. For example, a strong apparent effect might be detectable using a small sample, but that effect might be imprecisely estimated, straddling values that are both clinically relevant and those which are not. For more complex study designs aiming to investigate uncommon diseases in an unselected population, interactions or sub-group effects, large sample sizes are a prerequisite to obtaining informative estimates.

In the absence of adequate statistical power, the combination of chance and imprecise estimates can make inference from different studies challenging. A good example is the relationship between systolic blood pressure and vascular disease mortality. Varying results from individual studies created the impression that blood pressure had a sex-specific effect on vascular disease mortality and that there was a threshold below which hypertension was not clinically important. It took a meta-analysis of 61 prospective studies including 1 million participants and 13 million person-years at risk to collect enough data to change this interpretation[103]. Challenging a further interpretation that there was a ‘j-shaped’ relationship between diastolic blood pressure and cardiovascular disease required a cohort of 1.3 million participants with homogenously-collected data[104]. Obtaining definitive answers to even basic epidemiological questions may need far larger sample sizes than one might expect.

As a separate consideration, it is practical to recruit small sample sizes from highly-selected groups (such as patients attending a specialist clinic in a certain city), but the logistics of recruiting a large sample usually require a parallelised and less-selective recruitment procedure. While there is no inherent reason to believe large samples are more representative, in practice small studies are more likely to over-sample highly-selected groups.

2.1.2: The importance of sampling frame

The characteristics of the study population remain relevant in the era of contemporary epidemiological methods because conclusions about the study population are extrapolated to make inference about the underlying population whom the study represents. This extrapolation can sometimes fall at the first hurdle; in psychology it has been argued that it is not clear which underlying population is represented by healthy volunteers who wish to take part in studies[105]. Here the nebulous sampling frame creates a problem for eternal validity - making general statements about any group outside the study participants may be misleading because we do not know which groups are truly represented by the study participants.

Elsewhere the study sampling frame or exclusion criteria may clearly delineate the underlying population of interest, but the study may not achieve a representative sample of this target population. Participants with a high burden of disease symptoms may be less likely to volunteer for a time consuming and inconvenient study, thereby meaning that the study population has lower burden of disease than the underlying population whom it is intended to represent. A good example is UK Biobank[106], where the study population have markedly different lifestyle, sociodemographic and health characteristics from the underlying population[107], with systematic over-representation of healthy participants. Here, the strategy involved issuing a large number of invitations to compensate for non-response rather than aiming to facilitate participation in those who were invited, a strategy which achieved a large sample size rapidly at the cost of a response rate of less than 6%[108].

In both these situations extrapolating basic epidemiological parameters such as disease prevalence from the study population will produce systematically biased estimates of those characteristics in the intended population because the study group is drawn from a different underlying population. Likewise, basic parameters of interest in genetic epidemiology depend on the characteristics of the study population. For example the apparent heritability of a disease on the observed scale varies with the prevalence of that disease in the study

sample[109]. Thus, extrapolating these (internally-valid) estimates from the study sample to the population whom it is thought to represent can be misleading, a problem of external validity[110].

Sampling phenomena may challenge the internal interpretation of findings as well as external validity. Historically, it has been assumed that valid estimates of the relationship between an exposure and outcome could still be obtained even if the frequency of the exposure and outcome were not representative[107,111]. This notion is problematic if the participants of the study are viewed as 'cases' in a case:control study framework. As early as 1946 it was known that multiple disease states are artefactually correlated within case only analyis[112], by a mechanism we now understand as collider bias[108,113]. Participating in a study requires participants to overcome hurdles such as having the time and means to get to an assessment centre, a favourable attitude towards medical research, and willingness to undergo an elective health examination. Thus, participation in an epidemiological study is itself a complex outcome with many determinants such as geographical proximity to the assessment centre, health status and genotype[114], therefore these features can become artificially correlated within the study population, even if there is no association in the underlying population[55]. The magnitude of associations induced by collider bias is not easily predicted, and the direction of these associations is not intuitive. Again, this phenomenon is not restricted to basic epidemiological inference, but also creates bias and inflates type 1 error rates in methods which leverage genetic data such as Mendelian randomization[115].

While approaches are described to reduce the impact of selection bias (for example, incorporating weighting for probability-of-censoring[116]), these approaches rely on having well-measured covariates and specifying a propensity score model which is a good match to the (unknown) underlying causal model[117]. Adjustment for some biasing pathways can itself introduce bias into other terms[118]. It may therefore be helpful to design studies which minimize the effects of selection bias through high participation rates. Birth cohort designs are one approach which may be effective, as the target population can be clearly defined (for example, children born in a given year in a certain maternity service), recruitment takes place before the development of illnesses which might influence participation and interpolation of study protocols with maternity services can reduce the burden of participation for mothers and their families. Although participants may drop out over time as their life circumstances

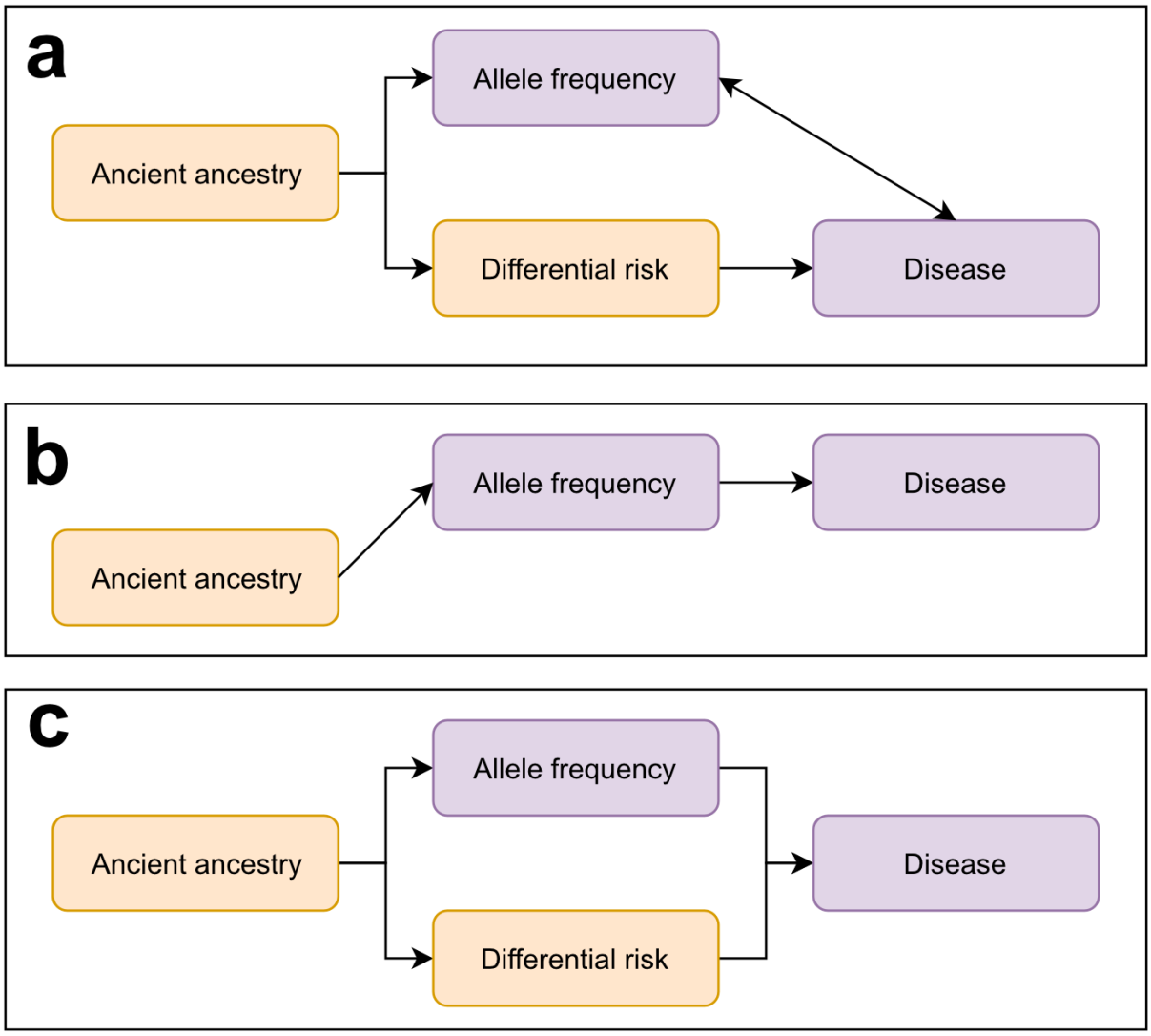
change, data gathered at the outset can help illustrate the magnitude of selection bias, specify a model to reduce biasing effects or even explore the genetic[114] and environmental factors which influence participation or drop out. These applications are only possible if comprehensive data (including genetic data) are obtained early in the study protocol before attrition occurs. As discussed later in this dissertation, enriching existing birth cohorts with dental data may therefore provide an important resource for dental epidemiology, even if dental data is collected many years after initial recruitment. This chapter includes examples of analysis in a birth cohort, and birth cohort designs make a major contribution to analysis included in chapter 5.

Specific to dental epidemiology, these theoretical considerations have implications for the way studies are designed and data are collected. First, there is an argument to design dental data collection in a manner which captures people who are unable to attend an elective dental examination. For example, dental records in primary care, which include participants who usually avoid dentists but do attend for emergency care when needed, may provide a less-selective way of obtaining data than asking participants to attend for an elective dental examination. Second, there is an argument for designing studies in a manner where likely impact of selection bias and determinants of having or not having dental data can be explored. By embedding dental data collection within a pre-existing cohort with a clearly defined sampling frame such as a birth cohort, the properties and likely impact of missing dental data can be better characterized than starting a dental cohort from scratch, and the broader rationale for bringing dental data into mainstream population health studies is discussed in chapter 6.

2.1.3: Importance of subgroups within the study population

Until now the study participants have been described as a single group for simplicity in order to compare them to an underlying population. In truth the study population will contain sub-groups with varying characteristics and for certain analyses it may be important to detect and account for this. Population stratification (i.e the presence of unrecognized subgroups) can lead to spurious genetic associations[119]. Here, differences at subgroup level in the allele frequency of a single genetic variant (due to ancient ancestry) are co-incidentally related to differences in in disease prevalence between sub-groups, inducing non-causal genetic associations (Figure 2.1). Ancient ancestry can be thought of as a confounder of the association between genotype and disease within the study population.

Figure 2.1: Possible relationships between ancient ancestry, allele frequency at a single genotype and disease status



Observed variables are drawn in purple boxes, latent variables are drawn in brown. In panel a) ancestry predicts behaviours which cause disease and allele frequency at a genotype which does not cause disease. The association between allele frequency and disease is spurious and correction for ancestry is desirable. In panel b) ancestry predicts allele frequency at a genotype which causes disease. The association between allele frequency and disease is causal and correction for ancestry is not desirable. In panel c) ancestry predicts both genetic and cultural factors which cause disease. Correction for ancestry is desirable, but over-correction will remove a causal genetic association. Rather than trying to remove structure, approaches such as admixture mapping which exploit this structure may be more appropriate.

In common with approaches for addressing confounders in observational literature, one approach to confounding by ancestry is stratification (aiming to hold the value of ancient

ancestry at a constant within each analytical group), where genetic effect estimates are obtained in each ancestral group within a study. In practice most of the participates within any single study are often of a single ancestral group, meaning that only the largest ancestral group has adequate power for GWAS analysis unless the study size is exceptionally large. It is therefore more common to exclude participants who cluster outside the largest group in principal component analysis[120,121]. This entrenched work practice may have contributed to the ongoing concern that participants of European ancestry are over-represented in the published GWAS literature[122].

This description considers subgroups under broad labels such as 'European ancestry'. These descriptions capture little of the complexity in human origins and migration, where the modern population of Europe arose from a series of migration and admixture events which are still detectable in the genetic information of modern day Europeans[123,124], and the idea that a homogenous population sample is achieved by simply restricting to one ancestral group is likely to be misguided. It is therefore necessary to identify and account for sub-structure even within a population defined in a broad ancestral group. One approach is adjusting for principal components of genetic data[121], which identifies independent axes of variation in genetic data and then treats these as covariates in a multivariable regression framework to try and obtain unbiased estimates of the effect of a genetic variant on the disease. Regardless of whether correction for ancestry is desirable or not (Figure 2.1), this approach has the potential to both under-correct for ancestry (in smaller studies where the principal components have poor signal to noise ratio[54]) and over-correct for biologically causal signals if the principal components are over-fitted and capture the biological signal of interest. There are other concerns; the principal component space estimated from common genetic variation extrapolates poorly to rare variation, resulting in difficulties controlling false positives where uncommon genetic variation is included in analysis[125,126]. Other approaches use long shared haplotypes to infer recent fine-scale ancestry[127] or conceptualize genetic clustering by ancestry in the same framework as genetic clustering by family structure[128].

2.1.4: Importance of environment within a study

The geographical footprint of a study may also be relevant. At population level, risk factors for a disease may cluster around geographical hotspots, such as cardiometabolic risk factors which are known to cluster in urban areas[129]. Here, the risk factors are assumed to be causal for cardiometabolic diseases, but it is also possible that other features which are not causally

related to disease are geographically clustered in disease hotspots. Geographical covariance between putative exposure and outcome can therefore act as a confounder to mislead inference. Given that genetic risk factors are also clustered[130], it is possible that alignment of spatial structure in genetic data, environmental risk and health outcomes could be problematic. In simulations where mate choice is restricted to individuals within a plausible geographic distance, genetic data becomes clustered by geography and this phenomenon is sufficient to induce genetic associations with spatially stratified dummy phenotypes with no real heritable contributions[131]. In real data it is argued that regional clustering of disease and fine-scale ancestry can lead to over-estimation of heritability[132] and may necessitate the use of an environmental covariance matrix or novel techniques such as relatedness disequilibrium regression[133] which are robust to this type of effect. While it may not be possible to detect or account for all environmental effects, an understanding of the socio-geographic differences across the study sampling frame which affect the exposure and outcome variables may be helpful. The notion that geographic clustering of genetic data could be used to explore regional environmental differences in the determinants of health is explored further in chapter 6.

2.1.5: Intersection of studies with different populations

The arguments presented so far imply that persistent, structured variation among people in a group or between groups of people in a study is a problem. This is not always true as all studies rely on systematic structured covariance between putative exposures and outcomes to identify associations. It is not only desirable but essential that there is variation for most epidemiological inference, and stark differences between groups can strengthen rather than complicate inference using the right study designs. Among adults in the Republic of Korea it is common for men to drink and women to abstain from alcohol, providing a natural negative control[87,134] group to strengthen inference about the causal effects of alcohol on outcomes such as stroke[135]. In another example, differences in recombination patterns of genetic information from different ancestral haplotypes within a single but admixed study population can help map genetic associations more precisely than is possible with a single-ancestry study population[136].

In addition to variation within a single study, it can be helpful to examine variation between cohorts drawn from populations with different confounding patterns. Differences in culture, diet, climate, economic development, healthcare systems and genotypes may all act to

'shuffle' networks of confounding features, meaning that inference is strengthened when an association is seen consistently despite these differences. The concept of triangulation in aetiological epidemiology[137] argues that evidence is best assimilated by identifying areas of agreement in multiple studies each with their own strengths and weaknesses rather than trying to design any one 'perfect' study. In dental epidemiology it might be helpful to identify associations which are seen consistently in different populations and using different study designs. This chapter includes observational analysis in two contrasting populations, while chapters 3 and 5 assimilate evidence for genetic association from multiple studies, each of which have their own strengths and limitations.

2.1.6: Distinction between random and non-random measurement error

The ideal study would contain perfect measurements of all putative risk factors, outcomes or covariates. In practice, it is impossible to measure any of these perfectly and there will always be some degree of measurement error. Random measurement error (such as rounding height to the nearest centimetre) can introduce regression dilution bias if this variable is used in regression analysis. Used as an exposure variable in linear regression this error will bias the gradient towards the null; used as an outcome this error will increase the standard error of the regression estimate[138]. This distinction is important – larger sample sizes (with smaller standard error) can recover the problem of imperfect measurement when used as an outcome, but larger sample sizes with poorly measured exposure only lead to estimates which are 'more precisely wrong'[138].

Apart from random error, there may be systematic measurement error, where the source of error correlates with a feature of interest in the study. Here the concern is that error may create a confounded pathway between a putative exposure and outcome. For example, motion of the head (related to certain medication) during magnetic resonance imaging (MRI) scans will systematically alter estimates of grey matter volume and cortical thickness, generating spurious association between medication use and imaging-derived measures of atrophy[139]. Here, the problem can only be recovered by choosing an exposure variable which is uncorrelated with the source of error – for example age may be problematic if age correlates with medication use but using an exposure such as common genetic variation might close the confounding pathway.

In practice it is likely that both random and non-random sources of error influence measurement of most complex traits, including dental disease traits. The possible implications of non-random error in measures of dental caries experience are discussed in detail later in this chapter.

2.1.7: Compromise between phenotypic resolution and study size

As shown by the example of MRI imaging, measurement error is not synonymous with measurement resolution, as both detailed and crude phenotypes can be subject to random and non-random measurement error. Instead, resolution refers to the intended level of detail in a measurement and forms a spectrum from highly detailed (often multi-dimensional) measurements through to cruder measures of an exposure or outcome.

Higher resolution phenotypes typically take more time to obtain and require more specialist equipment than low-resolution phenotypes. Following high dimensional measurement it is often necessary to collapse measurement into a single parameter reflecting variation in a trait of interest in the population. Measuring intracranial volume requires a detailed, whole brain scan followed by careful post-processing[140] to derive a single per-participant value. Conversely, proxy measures such as frontal head circumference[141] can be obtained rapidly with the minimum of equipment and captures variation at population level without further derivation or post-processing. The main attraction of these less resolved phenotypes is that they can be collected in large samples, while the lack of resolution restricts the range of possible scientific questions that could be answered using these data.

The example of MRI and head circumference illustrates a general argument that higher-resolution measures are only ‘better’ at answering research questions which also require that degree of resolution. To proxy variation in cranial dimension at population-level, there is no advantage in the more resolved phenotype, but to investigate specific brain phenotypes an MRI scan is clearly essential. In dental epidemiology, it may be helpful to consider whether high resolution measures have any advantage over lower resolution measures on an application by application basis, rather than making general statements.

2.1.8: Statistical properties

Binary measures do not have a distribution, but most types of data have a statistical distribution. Parametric approaches to modelling data may include an assumption about the

distribution of a variable under the null hypothesis, meaning that understanding the statistical properties of a phenotype can guide an informed decision about modelling strategy. The rationale for different modelling approaches used in this chapter (logistic, Poisson, zero-inflated negative binomial and Weibull regression) is described in detail later in this chapter, but each of these could only be selected after first understanding the nature of the phenotypic measurement. As discussed in chapter 5, the statistical properties of measures of the same trait (in this case dental caries) might change across the lifetime and require different modelling approaches at different life stages.

2.1.9: Compatibility of data and research question

After understanding the sources of error, resolution and distribution of a phenotype it is necessary to judge which research questions they may be suitable for. A low-resolution dental disease phenotype with considerable error but available in a large sample size may be suitable for inclusion in a genome-wide association study. Here, the study design requires large sample sizes, random error has an impact on power but does not bias genetic effect estimates and non-random error is not anticipated to correlate with genotype. The aim is to test for difference in phenotype between participants – meaning a low-resolution measure which captures this without additional transformation is as valid as a high-resolution phenotype which has been collapsed to a person-level score. Indeed, genome-wide association studies for relatively crude measures of adiposity[142], birth weight[143] and type 2 diabetes[144] have all been highly successful, although it has also been noted that the measurement strategy will influence the apparent genetic architecture of a trait[51]. Testing nuanced hypotheses through a mediation analysis framework has different requirements – this might be feasible in much smaller samples, but needs high resolution phenotypes to untangle the potential effects of different mediators, and near-perfect measurement of potential mediators, as error in these measurements will result in invalid inference[145].

In dental epidemiology it may be time to move away from notions of ‘reliable’ or ‘unreliable’ measurement and instead assess the suitability of a dental disease trait on a case-by- case basis, considering the intended inference, planned modelling approach, available sample size and whether sources of error in the outcome are potentially correlated with the exposure. Considering all of the above, the first practical illustration in this chapter explores whether questionnaire-derived dental data obtained in adolescence can be used to investigate differences in dental health between two groups of participants.

2.2: Analysis of questionnaire data in the Avon Longitudinal Study of Parents and Children

Questionnaires can be used to obtain dental data in large collections at low cost. The Avon Longitudinal Study of Parents and Children (ALSPAC) used this approach to obtain data on the oral and dental health of children. Examining these data in observational analysis provides a test bed for describing their characteristics and helps justify the inclusion of questionnaire-derived data for genetic association discovery in chapters 3 and 5. The aim of this analysis is to describe the long-term dental health of people who had dental treatment under general anaesthesia in childhood, and test whether previous dental treatment under general anaesthesia predicts caries or anxiety in adolescence. Results presented in this section have previously been published in the *British Dental Journal*[94].

2.2.1: Introduction

Caries as a public health problem in children

Caries in children remains an important public health problem in the UK. In the 2013 child dental health survey (the most recent available at time of writing) 49% of children had clinical decay experience (including visual enamel caries) by age 5 years[146], with higher prevalence in groups such as boys (52%) and children who are eligible for free school meals (58%). The absolute prevalence decreases when using more stringent criteria such as caries into dentine, but the interpretation is unchanged: dental caries is one of the most prevalent diseases affecting children. In addition to preventing caries and reducing disease burden, reduction in oral health inequality is a stated public health goal in the UK, for example the Children's Oral Health Improvement Programme Board chose 'a reduction in the oral health gap for disadvantaged families' as one of the 6 most important measures of success for their 2016-2020 action plan (source: https://www.gov.uk/government/news/launch-of-the-childrens-oral-health-improvement-programme-board, accessed August 2019).

Disease burden is considered to reflect the medical, financial or socio-economic impact of disease rather than prevalence alone[147]. Children with caries have impaired quality of life and the families of these children are also affected by the children's pain and time away from school among other features[148-151]. Collectively, the high burden of disease in children and the impact on quality of life suggests that preventing and treating dental caries in children should be a public health priority.

General anaesthesia in paediatric dentistry

In recent years dental treatment under general anaesthesia (DGA) has emerged as a commonly used modality for treating younger children with widespread disease. DGA has become the leading cause of hospital admission for children aged 5-9 years old in the UK,[152] with trends for increasing admissions and younger age at admission[153-155]. Used effectively, the interventions which are made possible by DGA can alleviate pain and infection due to dental caries, resulting in improved quality of life for children and their families[156]. A recent meta-analysis of case series showed improvement in the whole family impacts on quality of life after comprehensive treatment[157]. Thus, there is an opportunity to achieve a rapid and meaningful resolution of the symptoms of caries.

It is intuitive that the increase in children undergoing DGA may reflect increasing prevalence or severity of caries in high risk children, but this view is simplistic as several factors may contribute to children having DGA. Dental fear is the most common self-reported indication for DGA in children and adolescents in a cross-sectional case series in Finland, with need for extensive care in second place[158]. Other authors argue that situational factors such as the usual reason for visiting the dentist, region of residence within Britain and sociodemographic characteristics are more strongly associated with DGA than either anxiety or dental status[159]. Thus, having DGA may act as a proxy for a range of social and behavioural risk factors for dental caries, not just the disease itself. This analysis aimed to explore whether questionnaire data are useful in understanding the dental health characteristics of adolescents who underwent DGA 10 or more years previously, and whether asking about previous DGA would be a useful adjunct to caries risk assessment in clinical practice.

2.2.2: Methods

This analysis was conducted using the Avon Longitudinal Study of Parents and Children (ALSPAC), a prospective birth cohort study which was originally conceived as one of a network of pan-European birth cohorts[160]. Analysis was conducted as a nested cohort study, where children who underwent DGA were defined from ALSPAC records and then compared to children who did not have DGA. These definitions were created retrospectively but all data analysed were prospectively obtained.

Sampling frame and recruitment

Avon refers to the former non-metropolitan and ceremonial county which was created in 1974 under the Local Government Act of 1972[161]. Avon contained 6 non-metropolitan districts covering the cities of Bristol, Bath and surrounding semi-urban areas in south-west Bristol, which were broadly aligned to administrative health districts. Of these districts, 3 were located in the greater Bristol area (Southmead district, Frenchay district, Bristol & Weston district), which collectively served a population of around 0.9 million people in 1991[160]. Pregnant women with estimated delivery dates between 1st April 1991 and 31st December 1992 in one of these 3 districts were eligible to participate in the study.

ALSPAC aimed to recruit women early in pregnancy and targeted eligible women through conventional media, information at antenatal clinics and by visiting community locations. By comparing ALSPAC study records with maternity records it appears that recruitment rates were high, with 14,541 pregnancies enrolled in ALSPAC from a target population of 20,248 pregnancies. This recruitment ratio may be slightly flattering as the eligibility criteria and target population were defined retrospectively and some pregnancies in ALSPAC did not strictly meet the eligibility criteria[160]. Subsequent recruitment outside this initial sampling period with 2 additional rounds increased the total number of pregnancies recruited into ALSPAC to 15,247, from which were 14,701 children alive at 1 year. As of 2019, follow-up of mothers[162], children and children of these children is ongoing. Parents gave written informed consent for the participation of their children in the study. Younger children gave assent, whilst older children gave consent for participation. Approval for data collection was obtained from the ALSPAC Ethics and Law Committee, which is registered as an Institutional Review Board, while approval for the analysis of this specific project was obtained from the ALSPAC executive.

Exposure variable and covariates

DGA status was assigned based on parent-completed questionnaires at age 3, 4, 5 and 6 years and a parent-and-child-completed questionnaire at age 7 years. If the family reported their child had a DGA at any one of these questionnaires, the child who had the DGA was classified as a case. If the family completed this section of the questionnaire on at least one occasion but reported that the child had not received a DGA then the child was classified as a control. If the family had not completed this section of the questionnaire on any occasion (for

example if they were no longer active in the ALSPAC study by the time of the first questionnaire which asked about DGA) then the child was excluded from analysis. Severity of childhood caries was taken from the parent-and-child-completed questionnaire at age 7 years. Here, children were asked (with help from a parent) to look in their own mouth with a mirror, compare this to a drawing and then mark teeth with holes, fillings or missing teeth on the drawing.

Information on sex was taken from midwife-reported sex. The mothers' self-reported highest educational achievement was used as a proxy for child socio-economic status.

Outcome variables

A subset of 5,214 ALSPAC participants attended a 'transitioning to adulthood' focus clinic which targeted active participants at age 17 years. Of the young people who attended, 2,643 answered a questionnaire about their dental health, views on dental prevention and attitudes towards seeking dental care. Participants were asked to report the number of teeth they thought had fillings and the number of teeth that had been 'taken out because they were bad', with a follow-up question about the timing of the most recent dental extraction. The number of filled teeth was added to the reported number of extractions (provided the extractions were within the last 2 years) to create an index of filled or extracted permanent teeth (FEPT). This index aimed to capture variation in recent treated dental caries in permanent teeth, as required for the analytical question, and did not aim to capture untreated dental caries or caries in primary teeth.

Participants were asked to complete a Corah dental anxiety scale[163] which asks participants how they would feel at various stages of a dental procedure if they were attending a dentist for 4 different scenarios. Responses to each question stem were scored, and collated into a single index, with a score of 13 or higher used to identify participants with dental anxiety as suggested by Corah *et al*[164]. If participants completed 3 or more domains, then the average of the completed domains was used to estimate values for missing parts of the anxiety scale. If participants completed 2 or fewer domains then they were excluded from analysis of dental anxiety.

The coding of all variables used in analysis is provided in Appendix 2.1.

Modelling approach

Comparison of socio-economic status between DGA and non-DGA groups was tested using a nonparametric equality-of-medians approach. For analysis of 3 binary outcomes at age 17 years (caries-free versus caries-affected, regular dental attendance versus irregular dental attendance, dental anxiety versus no dental anxiety) crude analysis was performed by comparing proportions in the DGA and non-DGA groups and 95% confidence intervals for those proportions. Adjusted analysis was performed using logistic regression modelling.

For modelling the count function of FEPT, the characteristics of FEPT were reviewed as follows;

a) it is a count variable, which can only take discrete values

b) it is zero-inflated with many participants reporting scores of 0 (indicating no recent treated caries in permanent teeth) and a range of values among those with FEPT > 0, indicating extent / severity of treated caries experience in participants who do have recent treated caries.

c) it may contain both random and non-random measurement error

A modelling approach which is robust to these features was required. While Poisson models may be used for count variables these assume no over-dispersion i.e. the variance and mean of the count variable are assumed to be the same. Here, the excess of zeros and the low mean of FEPT might be problematic, as count variables with mean near zero are often overdispersed[165]. The negative binomial distribution relaxes this assumption, allowing the mean and variance to take different values, but is not well-suited if the major source of overdispersion is excess zero values[166].

Specifically for paediatric dental traits, it has been argued that negative-binomial hurdle, zero inflated negative binomial and marginalized zero-inflated negative binomial models give broadly similar interpretation[167], while another author working with data from children at age 5 and in young adults at age 18 and 26 argued that zero-inflated negative binomial (ZINB) was preferred to zero-inflated Poisson in these data[168]. A marginalized adapted ZINB model has been successfully applied in evaluating the impact of a fluoride mouth rinse program on caries in the primary dentition[169]. After reviewing these approaches and the characteristics of FEPT, a ZINB modelling approach was chosen for the count function of FEPT as an outcome.

2.2.3: Results

Demographics and baseline characteristics

The final sample with complete data included 1,695 young people, of whom 128 (7.6%) had received DGA before age 7.7 years. Mean age at follow-up was 17.7 years, and 60% of participants were female with comparable age and sex between participants who had and had not received DGA. At baseline (childhood self-report), young people in the DGA group had worse dental health than those who had not received DGA.

The median reported educational attainment for mothers of young people who received DGA was O level, while the median educational attainment for mothers of young people who had not received DGA was A level. The difference in medians would be unlikely if both groups were drawn from the same underlying population (p=0.014).

FEPT at age 17.7 years

Overall, 54% of the study population reported FEPT greater than 0 at age 17.7 (95% CI: 52%, 56%). In crude analysis participants who had previous received DGA were more likely to have caries (73% with FEPT greater than 0 (95% CI 64%, 80%)) than those who had not previously received DGA (52% with FEPT greater than 0 (95% CI 50%, 55%)). This difference in proportions was not compatible with the null hypothesis that the DGA and non-DGA groups were drawn from the same underlying population (p<0.001).

Exploring the relationship between DGA status and FEPT in a logistic regression framework, having DGA was associated with increased odds of a FEPT score greater than 0 at age 17.7 in all models (Table 2.1). In a model adjusted only for age and sex, the odds ratio was 2.4 (95% CI 1.6,3.6). Adjustment for maternal socio-economic status resulted in a better fitting model but there was little change in the association between DGA and FEPT, suggesting this association was not heavily confounded by socio-economic status. The addition of dental anxiety at age 17 led to improved model fit with some attenuation in the effect estimate of DGA. In the fully adjusted model, which had the best fit, caries status in childhood had a larger effect on caries at age 17.7 than DGA status, however DGA remained conditionally-predictive with an attenuated effect estimate (OR 1.7, 95% CI 1.1, 2.5).

Table 2.1: Association between putative caries risk factors and odds of FEPT>0 at age 17.7 years.

Exposure	Reference	Model 1		Model 2		Model 3		Model 4	
		OR (CI)	P	OR (CI)	P	OR (CI)	P	OR (CI)	P
DGA	No DGA	2.4 (1.6,3.6)	<0.001	2.3 (1.5 3.4)	<0.001	2.1 (1.4, 3.2)	<0.001	1.7 (1.08,2.53)	0.021
Age (years)		0.89 (0.67,1.17)	0.41	0.88 (0.68,1.15)	0.36	0.88 (0.67,1.15)	0.35	0.85 (0.65,1.12)	0.25
Sex (female)	male	1.1 (0.90,1.3)	0.38	1.1 (0.90,1.3)	0.86	1.00 (0.83, 1.23)	0.93	0.98 (0.79,1.2)	0.81
Vocational level qualification	CSE qualification			1.5 (0.86,2.5)	0.16	1.44 (0.85, 2.46)	0.18	1.4 (0.83,2.49)	0.19
O level qualification				0.84 (0.57,1.2)	0.36	0.86 (0.58, 1.27)	0.44	0.84 (0.56,1.26)	0.40
A level qualification				0.76 (0.52,1.1)	0.18	0.79 (0.53, 1.17)	0.24	0.82 (0.54,1.23)	0.33
Degree qualification				0.75 (0.50,1.1)	0.17	0.79 (0.53, 1.18)	0.25	0.82 (0.54,1.24)	0.34
Dental anxiety	No dental anxiety					2.9 (1.86, 4.43)	<0.001	2.8 (1.8, 4.4)	<0.001
Caries at age 7	No reported caries							3.0 (2.4,3.8)	<0.001
Model LR Chi squared$_1$		21.8		33.0		58.7		161.2	
LR test P$_2$				0.03 versus model 1		<0.001 versus model 2		<0.001 versus model 3	

N=1,695 for all models. $_1$Chi squared statistic for the likelihood ratio. $_2$P value tests the null hypothesis that the model indicated in the header of the table is not better fitting than the next-simplest nested model.

In addition to being associated with having non-zero FEPT, it is also possible that one or more variables are associated with extent or severity of caries proxied by FEPT within participants with non-zero FEPT. In ZINB regression analyses adjusted only for age and sex, adolescents with a non-zero FEPT who had received a DGA had 0.52 (95% CI: 0.39, 0.72) higher units of FEPT at age 17 years than those who did not have a DGA. Addition of terms for maternal socio-economic status resulted in little change to this association and little improvement in model fit. By contrast, addition of a term for dental anxiety resulted in substantially improved model fit, and some attenuation in the association between DGA and FEPT count. The best fitting model was a fully adjusted model which also included baseline caries status as a potential influence on FEPT count and the addition of this term resulted in attenuation of the DGA-FEPT count association (Fully adjusted estimate: 0.35 additional units of FEPT (95% CI 0.16,0.54)(Table 2.2)). Overall, these findings had similar interpretation to the logistic regression results presented above.

Table 2.2: Association between FEPT count and putative caries risk factors.

Exposure	Reference	Model 1		Model 2		Model 3		Model 4	
		Beta (SE)	**P**	**Beta (SE)**	**P**	**Beta (SE)**	**P**	**Beta (SE)**	**P**
DGA	No DGA	0.52 (0.099)	<0.001	0.50 (0.099)	<0.001	0.43 (0.098)	<0.001	0.35 (0.097)	<0.001
Age (years)		0.094 (0.082)	0.25	0.089 (0.081)	0.27	0.081 (0.080)	0.31	0.098 (0.079)	0.22
Sex (female)	male	0.004 (0.06)	0.95	0.0038 (0.064)	0.95	-0.076 (0.064)	0.24	-0.090 (0.063)	0.15
Vocational level qualification	CSE qualification			0.10 (0.15)	0.49	0.075 (0.15)	0.61	0.023 (0.14)	0.87
O level qualificiation				-0.10 (0.11)	0.36	-0.098 (0.11)	0.39	-0.11 (0.11)	0.32
A level qualification				-0.13 (0.12)	0.27	-0.12 (0.12)	0.27	-0.12 (0.11)	0.28
Degree qualification				-0.21 (0.12)	0.083	-0.18 (0.12)	0.14	-0.20 (0.12)	0.087
Dental anxiety	No dental anxiety					0.60 (0.098)	<0.001	0.59 (0.096)	<0.001
Caries at age 7	No caries reported							0.44 (0.070	<0.001
LR chi squared		30.6		38.2		77.1		116.50	
Likelihood ratio test				P=0.11 versus model 1		P < 0.001 versus model 2		P <0.001 versus model 3.	

N=914 observations with non-zero FEPT, 781 observations with zero FEPT for all models.

Dental anxiety at age 17

In crude analysis, 7% of the study group had dental anxiety (95% CI: 6%, 9%). Dental anxiety was more frequent in the DGA group (20%; 95% CI: 14%, 28%) compared to the non-DGA group (7%; 95% CI: 6%, 8%). This distribution of responses would be unlikely if the true prevalence of dental anxiety was the same in DGA and non-DGA groups (P<0.001). In logistic regression analysis adjusted only for age and sex, participants in the DGA group had greater odds (OR = 3.6; 95% CI: 2.2, 5.9; P<0.001) of being dentally anxious compared to those in the no DGA group, and female participants had higher odds of anxiety than male participants. By contrast to models treating dental caries as an outcome, addition of baseline dental caries status resulted in little improvement in model fit or attenuation of the relationships between DGA status and anxiety (Table 2.3).

Table 2.3: Association between dental anxiety and putative risk factors

Exposure	Comparison	Model 1		Model 2		Model 3	
		Odds Ratio (CI)	**P**	**Odds Ratio (CI)**	**P**	**Odds Ratio (CI)**	**P**
DGA	No DGA	3.6 (2.2, 5.9)	<0.001	3.3 (2.00, 5.38)	<0.001	3.1 (1.9,5.1)	<0.001
Age (years)		1.04 (0.64, 1.71)	0.87	1.03 (0.62,1.70)	0.91	1.02 (0.62,1.69)	0.93
Sex (female)	male	5.2 (3.04, 8.87)	<0.001	5.2 (3.04, 8.92)	<0.001	5.2 (3.0,8.9)	<0.001
Vocational	CSE			1.33 (0.62, 2.85)	0.46	1.31 (0.61,2.8)	0.48
O level				0.67 (0.37, 1.24)	0.21	0.67 (0.36, 1.23)	0.19
A level				0.60 (0.32, 1.14)	0.12	0.61 (0.32, 1.15)	0.13
Degree				0.46 (0.23, 0.91)	0.026	0.46 (0.23, 0.91)	0.026
Caries at age 7	Caries-free at age 7					1.31 (0.90, 1.91)	0.16
LR chi squared		74.04		84.6		86.5	
LR test p				0.032		0.16	

N=1,695 for all models

Child-reported experience of DGA

Seven hundred and twenty-five children responded to a question at age 7 years asking how they found the experience of DGA. Of these, 290 (40%) reported negative experiences of DGA by saying they "hated it".

Relationships between DGA and dental attendance, and dental anxiety and dental attendance.

Most adolescents (93%) reported attending a dentist for "regular routine checkups". In crude analysis, 11.8% of participants who had a DGA reported only going to a dentist irregularly (defined as any response apart from "regular routine checkups"), compared to 6.4% of participants who did not have a DGA (P value for difference in proportions<0.001).
In a logistic regression adjusted only for age and sex, DGA status at age 7 was associated with increased odds of irregular dental attendance at age 17 (OR 1.9) however this association was imprecisely estimated due to the small number of participants in the DGA group (95% CI 1.07, 3.41, p=0.029), n=1,680.

In crude analysis, adolescents who reported dental anxiety were more likely to report being irregular dental attenders (11.7% in dental anxiety group versus 6.4% in no dental anxiety

group), but this difference in proportions was compatible with chance (P=0.21). In logistic regression analysis adjusted only for age and sex there was weak evidence supporting an association between dental anxiety and dental attendance pattern (OR 1.8 for irregular attendance, 95% CI 1.02, 3.3, p=0.042), n=1,680

2.2.4: Discussion

DGA as a predictor of oral health trajectory

This analysis explored features relating to dental health in adolescents who had DGA in childhood, using adolescents who did not need DGA in childhood as a control group. The main findings are that children who have DGA remain at high risk for future caries and anxiety, and this is detectable by age 17 years even using relatively crude measures of dental status.

The main findings are consistent with other studies examining relapse (re-appearance of signs or symptoms of active dental caries following DGA) or re-treatment (additional treatment under DGA). In one dental teaching hospital in the UK, 34% of children required additional dental treatment during the six years after DGA[170], although some of this treatment could be provided under local anaesthesia. Other studies report repeat DGA rates between 4.2% and 17.0%[171,172], implying high levels of new disease, and recent studies in Saudi Arabia and the Republic of Ireland report similar findings[173,174], suggesting the main findings are not exclusively driven by patterns of DGA use or service provision in the UK.

One interpretation is that children referred for DGA represent a group with high extrinsic risk factors for caries, and these risk factors continue to act after caries-affected teeth are removed. In the UK efforts are made to control extrinsic risk factors for caries during DGA assessment, for example guidelines state that the assessing dentist should ideally be a specialist in paediatric dentistry[175], who is an expert in caries risk assessment, who often proposes different treatment plans from general dental practitioners[176] and whose targeted input is an opportunity to improve children's dental health[177]. Despite these efforts it is possible that attempts to control extrinsic risk factors during DGA are not fully effective, for example parents of this high-risk group of children are likely to lack oral health knowledge even after DGA, and that there may be low levels of ongoing oral health support for children who have had DGA[178,179]. Switching the emphasis from families to their general dental practitioners, a focus group of dental practitioners in London boroughs find caring for families undergoing DGA challenging. These dentists report that cultural barriers between dental services and families, inadequate communication and engagement of secondary care services and failures at national policy level all contribute to problems around the DGA process[180].

Another interpretation is that children referred for DGA have high genetic risk for caries, and this risk manifests in caries in permanent teeth despite efforts to control extrinsic risk factors. It is likely that the genetic risk factors for caries in the primary and permanent teeth are correlated[59] and it is argued that genetic risk explains why interventions on extrinsic risk factors are ineffective for some children[181]. The role of genetic risk factors for caries in childhood is explored further in chapter 5 but the findings of this chapter support the notion that caries should be treated as a person-level disease[182] which is not cured by simply addressing the manifestations of caries, such as removing decayed teeth.

While persistent environmental or genetic risk factors are a plausible explanation for the relationship seen in this study, it is possible (but untested) that the DGA process itself has an unhelpful effect on oral health trajectory, either by re-enforcing existing beliefs by the parent and family that treatment under local anaesthesia is impractical or impossible, or by exposing the child to an unfamiliar and threatening series of experiences which in could even predispose to future anxiety. Small scale qualitative studies have explored some of these more challenging aspects of DGA, for example identifying that children can find the DGA process itself unsettling[183]. Regarding dental anxiety and DGA, existing studies show little difference in dental anxiety between DGA and control groups but are limited by very short follow up times (up to 4 weeks).[184,185] Few studies with longer-term follow up have been undertaken, but a 5-year follow up study of children undergoing DGA in Helsinki identified that, as a group, these children continue to have high levels of dental fear and continue to find dental treatment difficult after DGA[186].

Regardless of the mechanism, DGA status continued to predict large difference in FEPT, suggesting that DGA status may a useful proxy for risk assessment when previous caries experience is unknown. For example, if a teenager attends a new practice for examination where primary teeth are missing and previous records are unavailable, then asking young people and their families about DGA may be a way to help assess risk, or a useful adjunct with incremental predictive value even when previous caries experience is known.

Suitability of birth cohorts as a sampling frame for dental epidemiology

This analysis used the ALSPAC birth cohort as a sampling frame for dental epidemiology. Compared to establishing a bespoke research cohort for dental diseases this strategy has several advantages. First, the study has a clearly defined target population, making it easier to assess the external validity of findings to any given population. Next, the study achieved a high recruitment rate and recruitment occurred before several of the potential barriers to participation in epidemiological research (such as poor health) could develop, probably reducing the impact of selection bias. Although selective drop-out will occur over time[187], it is theoretically possible to account for this as both the number and characteristics of participants who dropped out at various stages are known, unlike other study designs. Using this sampling frame, the study was able to capture participants who would be unlikely to volunteer for a traditional dental examination, such as those who reported high levels of dental anxiety. Finally, the ALSPAC study already has a range of longitudinal measures of putative environmental risk factors and detailed phenotypes including genetic data, so newly obtained dental data can be used for a range of analytical questions without additional data acquisition. Together, these arguments suggest that investment in obtaining dental data in existing studies with rich phenotypic information may represent a better allocation of resources than investment in creating bespoke cohorts for dental diseases.

The utility and limitations of questionnaire-derived dental data.

One motivation for this analysis was to explore the value of self-reported measures of dental status. Here there are several lines of evidence suggesting these measures capture variation in the intended traits; dental caries and dental anxiety.

Starting with disease prevalence, the 2013 child dental health survey reported 32% of 12 year olds and 44% of 15 year olds in England had obvious caries experience[146]. Using the questionnaire-derived FEPT measure, 54% of participants have FEPT > 0, in the range which would be anticipated considering the age of these participants. Equally, the 8% of participants with dental anxiety in this study (defined as Corah's anxiety scale >= 13) is in line with the reported prevalence of dental anxiety in the UK using slightly different criteria, for example 19% in females aged 16-34 and 8% of males aged 16-34 are reported to have dental anxiety (defined as MDAS >= 19)[188].

Next, the associations between FEPT and predictors including previous DGA, socio-economic status and caries at age 7 were in the anticipated direction and of plausible magnitude. The measure of dental anxiety was likewise associated with previous dental treatment and current dental avoidance in the anticipated directions and with plausible effect sizes. These findings are similar to those seen from other studies reporting that self-reported number of teeth is a satisfactory proxy for number of teeth[189] and that self-reported poor oral health identifies a group of people who are at high risk for tooth loss during a 5 and 10 year timescale[190].

It seems likely that these measures, although imperfect, capture variation in dental caries and anxiety when gathered in sufficiently large samples. While the measures contain both random and non-random error, this does not appear to preclude valid inference in observational analysis and is likely to be less of a concern in genetic association studies where the exposure variable is unlikely to correlate with sources of error in measurement of oral health outcomes. This analysis therefore helps support the inclusion of questionnaire-derived dental data for gene discovery in chapters 3 and 5. Although the ZINB models used in this analysis theoretically preserve more information than a logistic regression model using caries-free or caries-affected groups as outcomes, the interpretation of both modelling approaches was similar, helping to justify the decision to model caries as a person-level disease in paediatric populations using logistic regression in chapter 5.

While highlighting the utility of these data, the analysis also highlighted some of the limitations including the lack of resolution which makes these data unsuitable for some research questions. Here it may be valuable to look at other efficient ways of obtaining dental phenotypes in large collections.

2.3: Tooth loss in adult populations

The previous examples illustrated one way that questionnaire data might be useful in dental epidemiology. For some research questions more resolved phenotypes may be needed. This next example marks a shift from children to adults, from a single population to a pair of contrasting populations, from questionnaire data to index-linked data and from a clinical question to a methodological question. Nevertheless, the underlying motivation is similar – to demonstrate the utility of a flexible approach to gathering dental data in large scale studies. The aim of this analysis is to estimate the relative importance of caries and periodontitis as risk factors for tooth loss, and to test for covariance between disease estimates for caries and periodontitis in two contrasting populations. Results presented in this section have previously been published in *Community Dentistry and Oral Epidemiology*[95]

2.3.1: Introduction

Missing teeth as a measurement problem in GWAS studies for dental caries

Previous GWAS investigations for caries have used the indices of decayed, missing and filled teeth or decayed, missing and filled tooth surfaces (DMFT/DMFS) as measures of dental caries as advocated by the World Health Organization[191]. The justification for including missing teeth in this index is that advanced dental caries may be treated by dental extraction. Conversely, dental extraction may be used to treat periodontitis and to make space in patients with dental crowding as part of a course of orthodontic treatment[192]. The threshold for committing to tooth extraction rather than trying other treatments may depend on patient preferences for treatment, dentist technical skill, patient motivation, and resources (time and money) that a patient can commit to complex treatment plans with uncertain prognosis. Teeth may also be missing due to unplanned tooth loss (for example following dental trauma, or as an end-stage symptom of periodontal disease) or may be developmentally absent or unerupted.

This creates possible problems for studies (including this dissertation) which aim to use DMFS in GWAS. The first is that including missing teeth in DMFS scores results in over-estimation of caries exposure, while stripping out missing teeth under-estimates caries exposure[193,194]. If periodontitis is a major cause of tooth loss in the population included in GWAS analysis and periodontitis is a heritable trait, then measurement error in DMFS is unlikely to be random but correlated with genotypes which predispose to periodontitis. Conversely, the other factors mentioned above such as socio-economic circumstances, patient

and dentist preferences for treatment would not typically be thought of as heritable traits, so, while relevant in clinical practice, might not present a measurement problem for GWAS studies unless the study had exceptionally high statistical power.

Apart from a methodological problem, counts of missing teeth may be a rapid and cost-effective way of assessing dental status in large studies. This approach is becoming an increasingly common way to measure dental status, and missing teeth are associated with a range of adverse health outcomes in epidemiological studies[195-201]. Understanding the dynamics of tooth loss may help aid interpretation of these studies and shed additional light on the properties and limitations of tooth loss as a measure of dental diseases.

As a starting point in thinking about missing teeth in the context of measurement, it may be helpful to estimate the relative importance of periodontitis and caries as causes of tooth loss in the study population. Existing studies do not completely address this question, tending to use data obtained at the time of extraction rather than prospectively[202-206], using repeated cross-sectional samples rather than longitudinal data[207,208], using self-reported tooth loss[207,209,210], or having only dental diagnosis codes rather than dental charting[211]. Other studies have high quality data but in selected patient groups who may not be representative[212,213]. Against this background, the Dunedin study[214-216] is an unusual exception, as this is a prospective birth cohort of approximately 1,000 participants with serial dental examination. Even here there are limitations, as study follow-up has (to date) only reported on dental status to age 38 years, where levels of tooth loss are low[214] compared to participants aged 40 years and older[217,218]. The first aim of this analysis is to use longitudinal data to estimate the relative importance of caries and periodontitis as risk factors for tooth loss in a study population mirroring the population included in GWAS.

Missing teeth as a measurement problem in GWAS studies for periodontitis

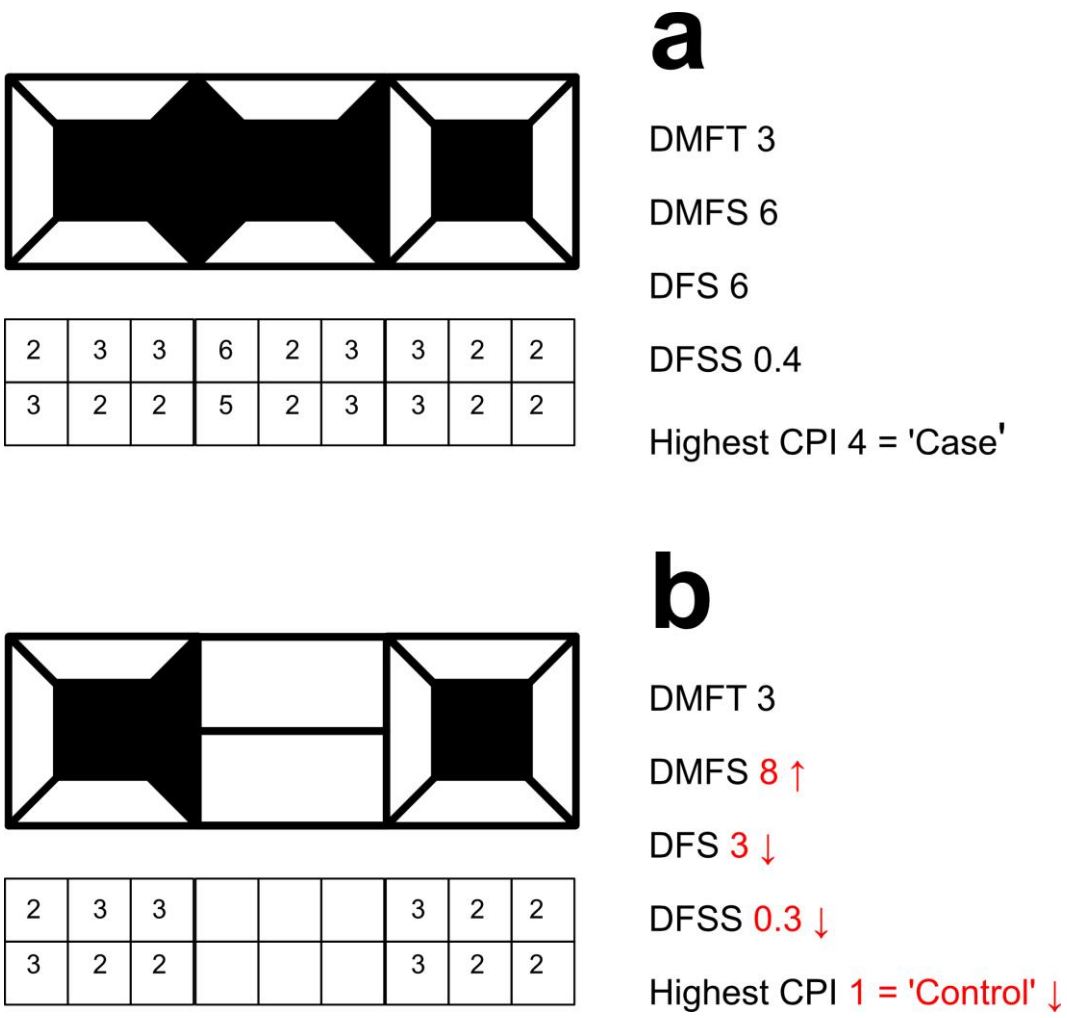
While arguments about missing teeth in DMFS scores are well-described in the literature[193,194], there is little recognition of the reciprocal problem that tooth loss due to dental caries affects measurement of periodontitis. The probability of observing a disease-affected periodontal pocket is related to the number of sampling opportunities, itself a linear derivative of the number of teeth. As illustrated in Figure 2.1, tooth loss therefore affects measurement of both dental caries and periodontal disease. Error in measurement of

periodontitis might therefore be non-random with respect to genotypes which influence dental caries.

As a censoring event on a common outcome of multiple diseases, tooth loss establishes the conditions for collider bias[113,219] which was introduced earlier in this chapter, now acting to induce association between different disease processes leading to tooth loss.

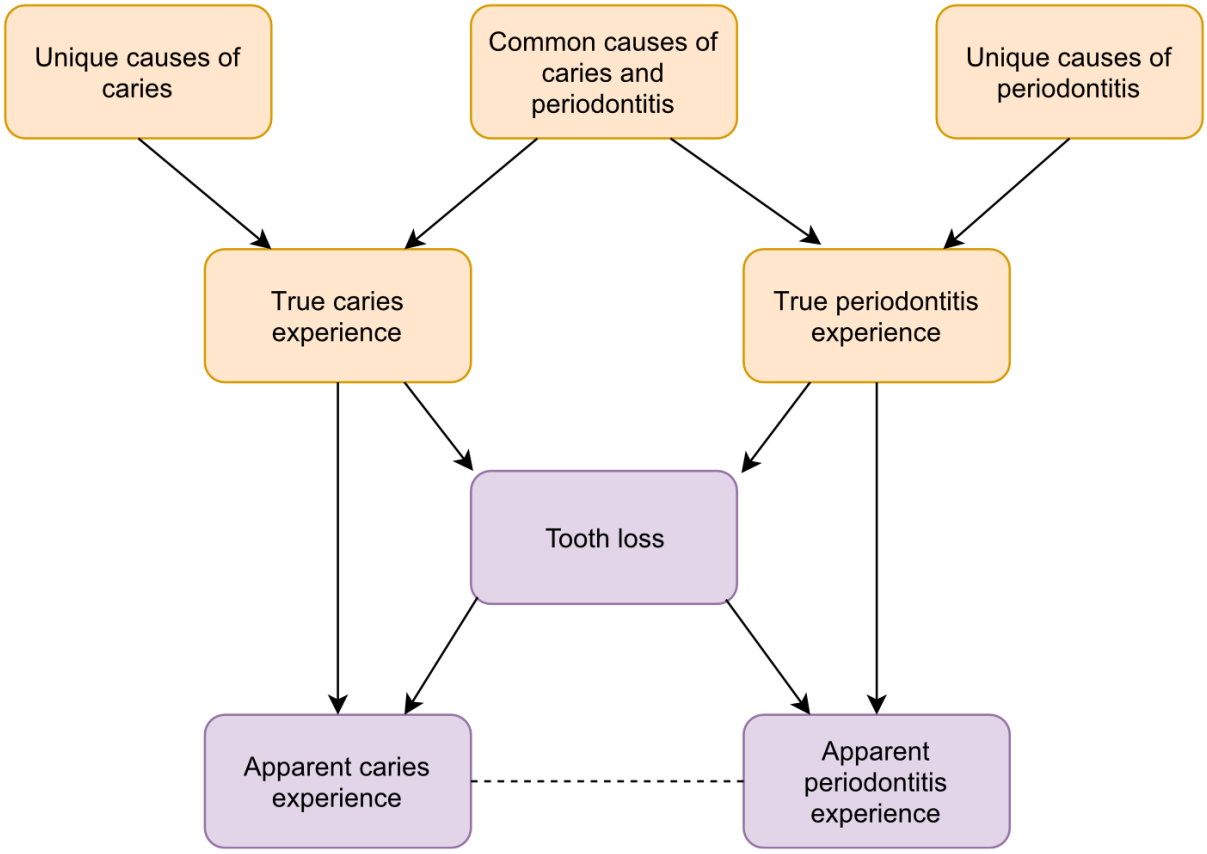
A key testable deduction from these statements is that there should be detectable covariance between disease estimates of dental caries and periodontitis in sufficiently large samples, and that the magnitude of this covariance should vary between populations with different levels of tooth loss, for example between older and younger participants. Conversely, covariance between caries and periodontitis which is created by common causes (such as diet or oral hygiene) should not vary across strata of tooth loss. By examining the magnitude of covariance between caries and periodontitis in populations with varying levels of tooth loss it should be possible to gauge the likely magnitude of these effects and determine whether they are large enough to mislead inference in GWAS for caries or periodontitis. Tooth loss may also serve as a model for illustrating a more general problem around obtaining specific estimates for different processes which share a common endpoint (Figure 2.2)

Figure 2.1: Missing teeth affect measurement of both caries and periodontitis.



The diagram represents restorative and periodontal charting of teeth 45, 46 and 47 for a hypothetical patient before (panel a) and after (panel b) extraction of tooth 46 for management of periodontitis. While DMFT is unaffected, all other summary measures of caries or periodontitis change after tooth 46 is removed.

Figure 2.2: Potential sources of covariance between estimates of caries and periodontitis.



Latent variables are drawn in beige boxes, measured variables are drawn in purple boxes. Pathway effects between true caries experience and true periodontitis experience are omitted for simplicity. Disease estimates for caries and periodontitis are expected to be correlated (dashed line) both because of common causes for these diseases and because of measurement artefacts relating to tooth loss (see Figure 2.1). The magnitude of this correlation might vary between populations with different levels of tooth loss. In addition, tooth loss opens a path for apparent caries experience to capture unique causes of periodontitis, and the magnitude of this effect is partially dependant on the relative importance of caries and periodontitis as causes of tooth loss.

2.3.2: Methods

GLIDE background, sampling frame and protocol

In 1924 the Swedish Parliament commissioned a report on dental care of the nation. This led, in 1938 to the establishment of Folktandvården, a public dental service which continues to this day[220]. The core services are regular dental care for children and teenagers aged between 3 and 19 years (with 95% to 98% uptake), dental care for adults (used by approximately 40% of Swedish adults) and specialist referral services (source http://www.folktandvarden.se/om-folktandvarden/). The Folktandvården has 21 regional divisions, each covering the region of a county council.

Adults attending this service have examination following a centrally determined protocol and dentists working in Folktandvården undergo training to ensure service provision is standardized as far as possible. Assessment of caries and restorative status involves tooth level charting with good lighting, mirrors, probes and radiographic examination where indicated. If teeth are present, they are charted at surface level to indicate carious surfaces, restored surfaces and sound surfaces. Carious surfaces are scored by the extent of caries using D1, D2 and D3 diagnostic thresholds. For restored surfaces, information is recorded on the type of restoration, such as a filling, inlay or onlay. Fissure sealants are not recorded as fillings. Where a tooth is missing, the type of restoration used to replace it (such as dental implant or bridge) is recorded, but the reason for tooth loss is not specified.

Assessment of periodontal status involves screening using the Community Periodontal Index (CPI) , a 5-level screening score which is assessed and recorded in 6 regions (sextants) of the mouth[191]. If there is a need for additional information (for example in diagnosis or monitoring of periodontal disease) then full mouth pocket charting is performed and recorded.

Since 1st January 2000, data from these examinations has been stored as an electronic dental record. These are a full substitute for a paper record and are stored in databases at the Folktandvården division serving that county. Information is logged and updated each time a dentist performs a diagnostic procedure or treatment. All information is stored in conjunction with a unique personal identification number, meaning that, under the right circumstances, data can be retrieved and merged with other health records for people who have consented to participate in health research.

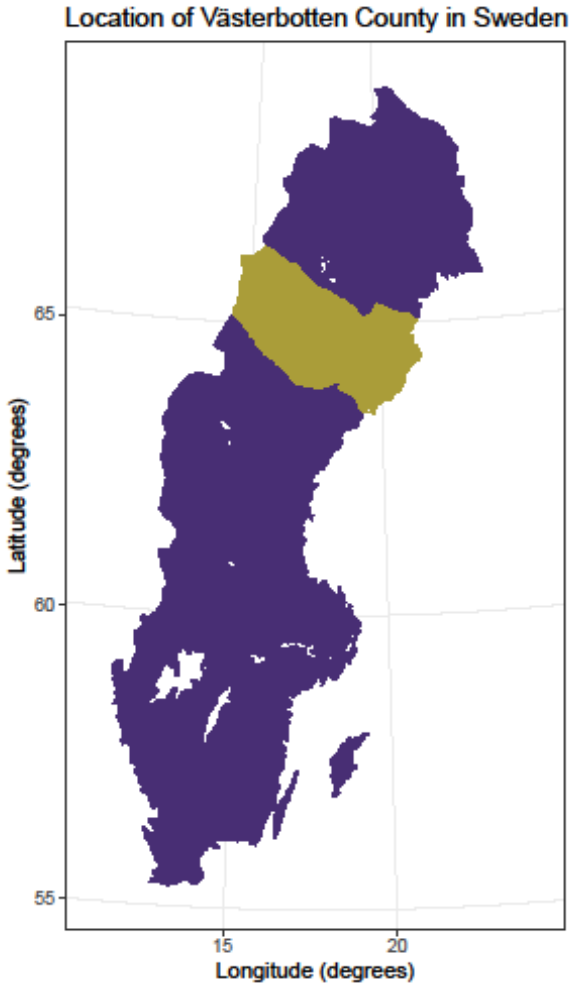
The Gene-Lifestyle Interactions in Dental Endpoints (GLIDE) project is an ongoing research initiative which is managed by Umeå University and receives core funding from Vetenskapsrådet (the Swedish Research Council) and Patentmedelsfonden för Odontologisk Profylaxforskning (the Patent Revenue Fund for Research in Preventive Odontology) (source https://www.umu.se/en/research/projects/gene-lifestyle-interactions-in-dental-endpoints-glide/). GLIDE serves as a hub for co-ordinating genetic analysis across a consortium of studies (the GLIDE consortium, used in chapter 3) but has also created a resource for observational dental epidemiology in Sweden (the GLIDE database) by merging dental records from Folktandvården with health screening records of participants in Swedish epidemiological studies. In this chapter, 'GLIDE' or 'Swedish GLIDE' are used to refer to the database, and to the combination of the GLIDE database with the Västerbotten Intervention Program.

Västerbotten Intervention Program

For this analysis, GLIDE dental data were merged with health screening records from the Västerbotten Intervention Program (VIP), a population based health screening and intervention study in Västerbotten county, Northern Sweden[221] (Figure 2.3). During the mid-20th century cardiovascular diseases became an increasingly important public health problem in Sweden, and by 1985 the highest mortality rates were in the Västerbotten county of Northern Sweden[222]. A population-based screening and intervention program was first piloted in Norsjö, northern Västerbotten in 1985, before being implemented across the entire county, and continues as of 2019, now as a nested study within the broader Northern Sweden Health and Disease Study. Adults living in Västerbotten county are invited to attend the programme at ages 40, 50 and 60 years and can participate by attending their primary care provider. Participants are offered a systematic health assessment by a Swedish district nurse, which includes screening for major cardiovascular disease risk factors through questionnaires, physical measurement and blood tests. Following this, participants are provided with personalized advice on how they could improve their health. Participants are given the option for their screening records to be used in health research, with broad consent for future research and long-term storage of biosamples in the department of biobank research at Umeå University. Participation rates have generally been high, and increased from 56% to 65% between 1990 and 2006[223]. As of August 2018, 107,500 people have participated in VIP on

160,000 occasions (source: https://www.umu.se/forskning/infrastruktur/northern-sweden-health-and-disease-study-nshds/).

Figure 2.3: Location of Västerbotten county in Sweden



Västerbotten county is shaded in beige, remaining Swedish counties are shaded in purple. Other countries are omitted. The map was generated in R using polygon data provided by the 'swemaps' package[224].

The health screening and intervention aims of VIP appear to have been effective in achieving long term behaviour change and reducing mortality from cardiovascular and other causes[225]. This efficacy creates potential challenges for investigating the relationship between a baseline (pre-intervention) exposure and subsequent disease events, as the value of the exposure may change after counselling, meaning associations between the exposures under intervention and cardiovascular outcomes may need careful interpretation.

Apart from the research questions originally envisaged, the study facilitates a range of other questions by acting as a sampling frame to obtain a large and representative sample of participants who have consented to take part in health research provided that data are presented in an anonymized, untraceable manner. For example, the Northern Sweden Diet Database was able to use the VIP infrastructure and sampling frame in conjunction with infrastructure from another study (MONICA) to achieve a large sample with diet data and has used this to investigate the influence of diet on risk of first myocardial infarction[226]. Another example is the nested GLACIER[227,228] sub-study, which used the 10-year repeated sample design of VIP to create a nested longitudinal study evaluating the genetic determinants of change in lipid traits over a 10 year interval.

Data merge

Each Swedish citizen has a unique 12-digit personal identification number which allows multiple sources of data to be linked. Following approval by the Regional Ethical Review Board at Umeå University, Sweden (Dnr 2010-387-31M and 2011-74-32M), dental records for VIP participants who had consented to take part in health research and had attended a dentist at a Folktandvården clinic in Västerbotten country for a full mouth examination with surface-level documentation at least once from 1st January 2000 to 16th November 2015 were requested. Many participants in Swedish GLIDE have dental data from many clinical examinations spread over several years, and some participants in VIP have participated in health screening on more than one occasion 10 years apart and therefore have health data available from multiple dates. For cross-sectional analysis, a single set of health screening, dental charting and periodontal data was used. First, the time difference between screening date and dental charting was calculated, and the single most contemporaneous pair of health screening and dental charting data was selected. Next, the periodontal data from CPI or pocket charting which was obtained most contemporaneously with dental charting data were obtained. For longitudinal analysis, this single best set of data was extended to include all available dental charts, regardless of whether obtained before or after health screening, and the first measure was considered baseline for that participant.

GLIDE inclusion and exclusion criteria

Participants were eligible for inclusion in analysis if a) they were aged 18 years or older with no exclusion for maximum age b) dental data were available within a 5 year window from a

VIP screening visit and c) data were available on demographic characteristics and smoking behaviour. For longitudinal analysis participants also needed to have dental charting data on at least 2 occasions spanning at least 1 year of follow-up from baseline.

Korea National Health and Nutrition Examination Study sampling frame and protocol.
As a contrast to the Swedish GLIDE population, results were also obtained in the Korea National Health and Nutrition Examination Survey (KNHANES) and compared. KNHANES is a national initiative which was established in 1998 to survey the health, quality of life and nutrition of the population of the Republic of Korea[229]. The study has a cross-sectional survey design and aims to recruit approximately 10,000 participants a year and achieve a sample which is representative of the civilian, non-institutionalized population of Korea. KNHANES can reach participants who do not attend a dentist, provided they agree to participate in a dental examination as part of the study, and these people represent a substantial portion of the population. This is explored by asking participants whether they had experienced dental care during the previous year. The proportion who had accessed dental treatment were 60.3%, 30.7%, 63.3% and 39.9% in 2010, 2011, 2012 and 2013, respectively[230]. The KNHANES was approved by the Institutional Review Board of the Korean Centres for Disease Control and Prevention (IRB: 2008-04EXP-01C, 20089-01CON-03-2C,2010-02CON-21C,2012-01EXP-01-2C,2013-07CON-03-4C, 2013-12EXP-03-5C). This population was selected as there are marked contrasts in diet, healthcare systems and social patterns between northern Sweden and the Republic of Korea, and KNHANES adopted a different approach to sampling from Swedish GLIDE. Collectively, these contrasts provide an opportunity to assess patterns of association which are seen consistently despite the different sampling approaches and underlying populations.

Assessment in KNHANES includes two visits. A health interview and health examination are carried out in a mobile examination centre. One week after this a nutrition survey is carried out by a nutritionist who visits participants' homes. Between 2008 and 2014 the KNHANES included dental examination as part of the health examination protocol. The dental portion of the health examination is performed in the mobile examination centre and is staffed by a team of 4 public health dentists. These dentists are provided by a combination of the Centres for Disease Control and 16 cities and provinces. Site conduct quality control is undertaken by the Korean Academy of Preventive Dentistry and Oral Health, who are responsible for surveyor

training, site inspection and conformity assessment and ensuring the examination protocol is reproducible.

The dental examination includes surface-level dental charting. Each tooth surface is classified as sound, active caries, filled, missing (due to caries), missing (for any other reason), sealed, treated (due to any other reasons) or unerupted. The presence of prostheses such as bridges, partial dentures, complete dentures and implants are recorded. Periodontal status is assessed using CPI scores for each sextant which has two or more natural teeth. If they are present, scores were obtained from index teeth in each sextant (17/16, 11, 26/27, 36/37, 31, 46/47). If index teeth were lost then adjacent teeth were examined instead[231].

KNHANES exclusion and inclusion criteria

Participants were eligible if they were aged 18 years or older at time of screening and had complete exposure, outcome and covariate data available. There were no exclusions for maximum age.

Derivation of per-mouth summary data in GLIDE and KNHANES

Caries experience was summarized using World Health Organization caries indices, which were derived centrally from surface-level data. First, third molar teeth were excluded, leaving 28 teeth. Next, carious lesions requiring operative intervention (D2 or deeper) were re-coded as decayed, while initial carious lesions were re-coded as sound. Teeth with a crown were coded as 4 or 5 filled surfaces. Following World Health Organization guidelines[191], all missing teeth were coded as 4 or 5 missing surfaces regardless of recorded cause. Finally, the total of decayed, missing and filled tooth surfaces (DMFS), decayed missing and filled teeth (DMFT), decayed and filled tooth surfaces (DFS) and decayed and filled tooth surfaces divided by total number of non-missing tooth surfaces (DFS per surface or standardized DFS) were derived.

In the primary analysis periodontal status was summarized by categorizing participants into better periodontal health (CPI 2 or lower in all sextants) and worse periodontal health (CPI 3 or higher in any sextant) groups. In GLIDE some participants had full mouth pocket charting data rather than CPI scores. These participants were classified as having impaired periodontal

health if four teeth in the mouth had pocketing of at least 4 mm, or if any tooth in the mouth had pocketing of at least 5 mm.

These definitions might include participants with false pocketing related to gingivitis rather than periodontal attachment loss and periodontitis[232]. To explore this, sensitivity analysis was also performed which used a more stringent definition of worse periodontal health (CPI4 or higher in one or more sextants, one or more teeth with 6mm pocketing).

Longitudinal modelling approach

To relate baseline characteristics to incident tooth loss a survival modelling approach was adopted, which models the time between baseline and tooth loss in order to estimate hazard attributable to exposure variables. Broadly, survival models can be divided into fully parametric models (including those with an accelerated failure time function such as Weibull, exponential and log-logistic), semi-parametric models (such as Cox proportional hazard) and non-parametric models. Fully parametric models have advantages in that they allow clinically interpretable estimates such as of survival time and its distribution. In addition, the residuals in a fully parametric model represent the different between observed and estimated values of time, meaning these models can be used in full maximum likelihood[233] for comparison of model fit. The decision was therefore made to fit a Weibull survival model using the 'streg, distribution(weibull)' function in the statistical package Stata.

The exposure variables were baseline demographic and dental variables (including number of teeth) and the outcome was prospective tooth loss. All participants were considered to have entered the study at their first dental examination. At each subsequent dental examination, the number of teeth was compared to the number of teeth at baseline, and participants who lost 1 or more teeth were classified as having undergone failure. The failure time was entered as the time elapsed (in years) between entering the study and the first tooth loss event. Individuals who had not lost a tooth by the last available dental examination were censored, and time was entered as the time elapsed (in years) between entering the study and the final dental examination. All models were fitted with three age strata (<45 years, ≥45 - <55 years, ≥55 years at baseline) to allow influence of predictors to vary across strata of age. Univariate models were used to obtain unadjusted hazard ratios for potential predictors, before multiple predictors were fitted simultaneously in a multivariate model.

Cross-sectional modelling and transformation of effect estimates.

Cross-sectional models were fitted considering periodontal status as exposure and WHO caries indices as an outcome. Similar to the FEPT example earlier in this chapter, these WHO indices represent count data which can only increase in discrete units assessed at the level of a tooth or tooth surface. Unlike the FEPT example, the distribution of DMFS corresponded approximately to a normal distribution with no excess of zero values. Therefore, a count model was required without a zero-inflated link, and Poisson regression was performed using the stata function 'Poisson'.

The beta coefficients from this analysis are difficult to interpret without transformation and were therefore exponentiated. Following standard terminology, the transformed effect estimates are described as incidence risk ratios (IRR) in the results section. However, in the context of this cross-sectional analysis the represents fully adjusted ratio of a given caries index in exposed individuals compared to non-exposed individuals. For example, an incidence risk ratio of 1.5 for DMFS means the expected DMFS index count in participants with CPI≥ 3 is 1.5 times the expected DMFS index count in participants with CPI < 3.

2.3.2: Results

Study population

Cross-sectional analysis included 62,522 adults with approximately a 46% / 54% split between GLIDE and KNHANES. The GLIDE and KNHANES populations were comparable in terms of mean age, proportion of female participants and proportion with university level education. A higher proportion of participants reported smoking in KNHANES than GLIDE but mean DMFS, mean number of missing teeth and proportion with CPI≥3 were all lower in KNHANES compared to GLIDE (Table 2.4). Longitudinal analysis in GLIDE included 28,244 participants of whom around one third experienced tooth loss during study follow-up. The median time at risk was approximately 12 years.

Table 2.4: Demographic characteristics of the final samples included in the analysis

	GLIDE		KNHANES
	Cross-sectional	**Longitudinal**	
Included participants with baseline observations (N)	28,691	28,244	33,831
Follow-up observations (N)		232,905	
Total observations(N)[1]		261,149	
Median per-participant time at risk (years)		11.7	
Total[1]		298,120	
Sex, % male	49.1	49.2	49.4
Age, years at study baseline (mean (SD))	49.7 (8.6)	45.0 (8.9)	49.2 (16.5)
Proportion in age group			
Under 45 years	35.1	35.1	42.8
45-54.9 years	34.1	34.1	19.2
55 years and older	30.8	30.8	38.0
Dental status at baseline			
Missing teeth (mean (SD))	2.0 (3.2)	1.8 (3.0)	0.71 (6.0)
DFS (mean (SD))	33.3 (19.3)	31.4 (19.2)	11.8 (12.0)
DMFS (mean (SD))	43.2 (25.9)	40.1 (25.2)	22.5 (22.7)
CPI ≥3 (%)	52.0	51.8	27.7
Smoking status			
Current smoker	12.3	12.2	20.5
Not current smoker	87.7	87.8	79.5
Education, % University level	26.9	26.9	29.6

[1]Total observations refers to the sum of baseline and follow-up observations included in longitudinal modelling.

GLIDE longitudinal analysis.

Treating baseline dental status as a predictor of tooth loss, participants with worse periodontal status (CPI 3 or higher) had greater hazard for tooth loss than participants with CPI less than 3. The unadjusted effect estimate was larger in younger participants (Hazard Ratio (HR) 1.64; 95% CI 1.50, 1.79 for participants aged under 45 years) than in older participants (HR 1.38, 95% CI 1.12, 1.48 for participants aged 55 years or older; Table 2.5). The estimated effect of periodontal status on tooth loss was slightly attenuated in all age groups in the fully adjusted model which included adjustment for baseline DFS, baseline missing teeth, age, sex, age group, smoking status and socio-economic status.

Each additional decayed or filled tooth surface at study baseline was associated with slightly increased hazard for tooth loss in individuals aged under 45 years (fully adjusted HR 1.01, 95% CI 1.01, 1.02) but was modelled to have minimal effect on hazard in individuals aged 55 years or older.

Hazard for tooth loss was higher in participants who had previously lost teeth than individuals who had not previously lost any teeth, with the largest effect in young participants who had already lost many teeth at the start of follow-up (unadjusted HR 5.14; 95% CI 3.27, 8.09 for participants aged under 45 years missing 7 or more missing teeth at baseline compared to participants aged under 45 with no missing teeth at baseline). This effect estimate attenuated in the fully adjusted model incorporating terms for baseline periodontal status, baseline DFS and lifestyle and demographic factors described in table 2.7.

Table 2.5: Unadjusted and fully adjusted hazard ratios for dental status as a predictor of tooth loss.

Exposure	Age Group	Reference (for categorical exposures)	Unadjusted hazard Ratio (95% CI)	Unadjusted P	Fully adjusted hazard ratio (95% CI)	Fully adjusted P
CPI3 or higher	Under 45	CPI less than 3	1.64 (1.50, 1.79)	<0.0001	1.41 (1.28, 1.55)	<0.0001
	45 to 54.9		1.51 (1.40, 1.69)	<0.0001	1.28 (1.18, 1.38)	<0.0001
	55 or older		1.38 (1.28, 1.48)	<0.0001	1.22 (1.14, 1.32)	<0.0001
	Combined		1.48 (1.42, 1.55)	<0.0001	1.29 (1.24, 1.36)	<0.0001
Baseline DFS	Under 45		1.02 (1.02, 1.03)	<0.0001	1.01 (1.01, 1.02)	<0.0001
	45 to 54.9		1.01 (1.01, 1.01)	<0.0001	1.005 $(1.003, 1.07)_1$	<0.0001
	55 or older		1.002 (1.001, 1.004)*	0.004	1.002 $(1.0005, 1.004)_1$	0.013
	Combined		1.01 (1.01, 1.01)	<0.0001	1.005 $(1.004, 1.006)_1$	<0.0001
1-2 missing teeth at baseline	Under 45	No missing teeth at baseline	1.59 (1.43, 1.78)	<0.0001	1.38 (1.24,1.54)	<0.0001
	45 to 54.9		1.39 (1.28, 1.51)	<0.0001	1.18 (1.08, 1.28)	<0.0001
	55 or older		1.15 (1.05, 1.26)	<0.0001	1.04 (0.95, 1.14)	0.411
	Combined		1.34 (1.27, 1.41)	<0.0001	1.17 (1.10, 1.23)	<0.0001
3-6 missing teeth at baseline	Under 45		1.25 (1.09, 1.44)	<0.0001	1.17 (1.02, 1.35)	0.026
	45 to 54.9		1.55 (1.41, 1.70)	<0.0001	1.21 (1.10, 1.33)	<0.0001
	55 or older		1.46 (1.34, 1.60)	<0.0001	1.16 (1.06, 1.27)	0.001
	Combined		1.52 (1.44, 1.61)	<0.0001	1.22 (1.15, 1.29)	<0.0001
7 or more missing teeth at baseline	Under 45		5.14 (3.27, 8.09)	<0.0001	2.50 (1.57, 3.99)	<0.0001
	45 to 54.9		3.87 (3.25, 4.62)	<0.0001	2.12 (1.75, 2.56)	<0.0001
	55 or older		1.85 (1.68, 2.03)	<0.0001	1.22 (1.10, 1.36)	<0.0001
	Combined		2.18 (2.02, 2.35)	<0.0001	1.43 (1.32, 1.56)	<0.0001

The fully adjusted estimates are obtained from a multivariable model which also includes adjustment for parameters given in Table 2.7. $_1$Estimates where the upper or lower confidence interval rounded to 1.00 at 3 significant figures are provided with additional significant figures to resolve ambiguity

In sensitivity analysis using a more stringent definition of periodontal status, the estimated effect of periodontal status on tooth loss was greater than in the main analysis, potentially suggesting a dose-response relationship for periodontitis. Having a score of CPI 4 was associated with increased hazard for tooth loss, with largest effect in younger participants (unadjusted HR 2.07; 95% CI 1.85, 2.32) for CPI 4 versus CPI less than 4 in participants aged under 45 years (Table 2.3.3).

Table 2.6: Unadjusted and fully adjusted hazard ratios in sensitivity analysis

Exposure	Age Group	Reference	Unadjusted hazard Ratio (95% CI)	Unadjusted P	Fully adjusted hazard ratio (95% Ci)	Fully adjusted P
CPI 4	Under 45	CPI < 4	2.07 (1.85, 2.32)	<0.0001	1.69 (1.50, 1.90)	<0.0001
	45 to 54.9		1.76 (1.64, 1.89)	<0.0001	1.35 (1.25, 1.46)	<0.0001
	55 or older		1.50 (1.41, 1.60)	<0.0001	1.28 (1.21, 1.37)	<0.0001
	Combined		1.66 (1.59, 1.73)	<0.0001	1.36 (1.30, 1.42)	<0.0001

The fully adjusted model includes adjustment for age and sociodemographic parameters given in Table 2.7 and for dental parameters given in table 2.5 apart from periodontal status using the primary definition, which was omitted from the model.

Treating age, sociodemographic and lifestyle factors as predictors, each 1-year increase in baseline age was associated with increasing hazard for tooth loss during follow-up (unadjusted HR 1.09; 95% CI 1.08, 1.10), with an estimated effect which changed little in the fully adjusted model.

Higher educational attainment was protective against tooth loss in unadjusted analysis, with the strongest effect size in participants aged under 45 years with university level education (HR 0.46; 95% CI 0.39, 0.55) compared to participants aged under 45 years with elementary school level education. These effect estimates attenuated in the fully adjusted model, where, overall, university level education was modelled to have only a modest conditionally-independent effect on hazard for tooth loss when considering baseline dental and smoking status (HR 0.91; 95% 0.85, 0.97 for university level education versus elementary school level education combined across all age groups.

Current smokers had higher hazard for tooth loss than non-smokers in all age groups, with the strongest effect estimate in smokers aged under 45 years compared to non-smokers aged under 45 years (HR 2.53; 95% CI 2.22, 2.88 in unadjusted analysis, HR 1.80; 95% CI 1.57, 2.07 in fully adjusted analysis).

Table 2.7: Unadjusted and fully adjusted hazard ratios for age and sociodemographic variables as a predictor of tooth loss.

Exposure	Age group	Reference	Unadjusted hazard Ratio (95% CI)	Unadjusted P	Fully adjusted hazard ratio (95% CI)	Fully adjusted P
Age 45 - 54.9 y		Under 45	1.37 (1.29, 1.45)	<0.0001	1.79 (1.71, 1.81)	<0.0001
Age ≥55 y			2.13 (2.01, 2.24)	<0.0001	1.48 (1.42, 1.55)	<0.0001
Baseline age (1 year increment within age group)	Under 45		1.08 (1.07, 1.10)	<0.0001	1.04 (1.03, 1.06)	<0.0001
	45 to 54.9		1.10 (1.09, 1.11)	<0.0001	1.07 (1.06, 1.08)	<0.0001
	55 or older		1.09 (1.08, 1.10)	<0.0001	1.08 (1.07, 1.09)	<0.0001
	Combined		1.09 (1.08, 1.10)	<0.0001	1.07 (1.07, 1.08)	<0.0001
Current smoker	Under 45	Non-smoker	2.53 (2.22, 2.88)	<0.0001	1.80 (1.57, 2.07)	<0.0001
	45 to 54.9		2.08 (1.89, 2.28)	<0.0001	1.57 (1.43, 1.74)	<0.0001
	55 or older		1.74 (1.60, 1.89)	<0.0001	1.56 (1.43, 1.70)	<0.0001
	Combined		1.87 (1.98, 2.10)	<0.0001	1.62 (1.53, 1.71)	<0.0001
Comprehensive school or equivalent (education to age 15)	Under 45	Elementary school	0.65 (0.51, 0.84)	0.001	0.79 (0.62, 1.02)	0.067
	45 to 54.9		0.85 (0.75, 0.95)	0.006	0.82 (0.73, 0.92)	0.001
	55 or older		0.87 (0.80, 0.93)	<0.0001	1.01 (0.94,1.09)	0.754
	Combined		0.84 (0.79, 0.90)	<0.0001	0.92 (0.87, 0.98)	0.011
Highschool or equivalent (education to age 16-19)	Under 45	Elementary school	0.55 (0.47, 0.64)	<0.0001	0.76 (0.65, 0.90)	0.001
	45 to 54.9		0.65 (0.59, 0.72)	<0.0001	0.91 (0.82, 1.01)	0.072
	55 or older		0.83 (0.76, 0.91)	<0.0001	1.10 (1.01, 1.21)	0.033
	Combined		0.83 (0.73, 0.91)	<0.0001	0.98 (0.92, 1.04)	0.467
University /College or equivalent (education beyond age 19)	Under 45	Elementary school	0.46 (0.39, 0.55)	<0.0001	0.70 (0.58, 0.84)	<0.0001
	45 to 54.9		0.75 (0.69, 0.82)	<0.0001	0.85 (0.76, 0.96)	0.01
	55 or older		0.61 (0.54, 0.68)	<0.0001	0.99 (0.90, 1.09)	0.856
	Combined		0.66 (0.62, 0.70)	<0.0001	0.91 (0.85, 0.97)	0.006
Female sex	Under 45	Male sex	1.093 $(0.998, 1.179)_1$	0.055	1.11 (1.01, 1.22)	0.032
	45 to 54.9		1.16 (1.08, 1.25)	<0.0001	1.11 (1.03, 1.19)	0.005
	55 or older		1.00 (0.94, 1.06)	0.962	0.97 (0.91, 1.03)	0.355
	Combined		1.07 (1.03, 1.12)	0.001	1.039 $(0.997, 1.083)_1$	0.071

The fully adjusted estimates are obtained from a multivariable model which also includes adjustment for parameters given in Table 2.5. [1]Estimates where the upper or lower confidence interval rounded to 1.00 at 2 decimal places are provided with additional figures to resolve ambiguity.

Analysis of cross-sectional data in GLIDE and KNHANES

In GLIDE, incidence risk for caries traits was higher in participants with CPI 3 or above than in participants with better periodontal health, for example DMFS (IRR 1.06; 95% CI 1.05, 1.07). The association between periodontal status and WHO caries indices was consistent in all age groups, but stronger in younger participants than in older participants, for example DMFS with IRR 1.09 (95% CI 1.06,1.12 in participants aged under 45 years and IRR 1.03 (95% CI 1.01, 1.05) in participants aged 55 years and older.

In KNHANES the was evidence for age-specific association patterns between periodontal status and DMFS. KNHANES participants aged under 45 years with CPI of 3 or higher had higher DMFS scores than participants with CPI less than 3 (IRR 1.11; 95% CI 1.06, 1.17), while KNHANES participants aged 55 years and older with CPI of 3 or higher had lower DMFS scores than participants with CPI less than 3 (IRR 0.94, 95% CI 0.91, 0.96).

Apart from these age-specific effects, some relationships appeared population-specific in direction or magnitude. Impaired periodontal health (CPI 3 or higher) was associated with higher standardized DFS scores in GLIDE (IRR 1.05; 95% CI 1.04, 1.07) but lower standardized DFS scores in KNHANES (IRR 0.95; 95% CI 0.91, 0.96).

There was little evidence for association between periodontal status and either number of teeth or number of tooth surfaces in GLIDE, while in KNHANES there was weak evidence for an age-specific pattern of association (Table 2.8).

Table 2.8: Fitted differences in caries indices between periodontal cases and controls

GLIDE Cross-sectional analysis - Cases defined as CPI ≥3 in one or more sextants or 4mm pocketing around 4 or more teeth or 5mm pocketing around 1 or more teeth								
Age group	**% of cases**	**N**	**Fully adjusted ratio of index in cases:controls (95% CI)**					
			DMFS	**DMFT**	**DFS**	**Standardized DFS**	**Number of surfaces**	**Number of teeth**
Under 45	36.4	10,073	1.09 (1.06,1.12)	1.05 (1.03,1.07)	1.07 (1.04,1.10)	1.08 (1.05,1.11)	0.99 (0.99,1.00)	0.99 (0.99, 1.00)
45 to 54.9	53.1	9,794	1.07 (1.04,1.09)	1.02 (1.01,1.04)	1.03 (1.01,1.05)	1.04 (1.02,1.07)	0.99 (0.98,0.99)	0.99 (0.98, 0.99)
55 or older	68.6	8,824	1.03 (1.01, 1.05)	1.02 (1.01,1.03)	1.05 (1.03,1.07)	1.04 (1.02,1.06)	1.00 (0.99,1.01)	1.00 (1.00, 1.01)
Overall	52.0	28,691	1.06 (1.05,1.07)	1.03 (1.02,1.04)	1.05 (1.04,1.06)	1.05 (1.04,1.07)	0.99 (0.99,1.00)	0.99 (0.991,1.00)
KNHANES Cross Sectional analysis - Cases defined as CPI ≥3 in one or more sextants								
Age group	**% of cases**	**N**	**Fully adjusted ratio of index in cases:controls (95% CI)**					
			DMFS	**DMFT**	**DFS**	**Standardized DFS**	**Number of surfaces**	**Number of teeth**
Under 45	14.0	14,486	1.11 (1.06,1.17)	1.01 (0.97, 1.05)	0.96 (0.91,1.02)	0.98 (0.92, 1.03)	0.98 (0.98,0.99)	0.99 (0.98,0.99)
45 to 54.9	38.8	6,500	1.15 (1.08,1.21)	1.08 (1.03,1.13)	0.96 (0.90,1.02)	0.97 (0.91,1.04)	0.97 (0.97,0.98)	0.97 (0.97, 0.98)
55 or older	46.2	12,845	0.94 (0.91,0.96)	0.94 (0.92, 0.96)	1.02 (0.97, 1.07)	0.98 (0.93, 1.03)	1.04 (1.02,1.05)	1.03 (1.02, 1.05)
Overall	27.7	33,831	0.99 (0.97,1.01)	0.95 (0.93, 0.97)	0.97 (0.94,1.00)	0.95 (0.92,0.98)	1.01 (1.01,1.02)	1.01 (1.01, 1.02)

Fitted differences were obtained from a multivariable Poisson regression model incorporating adjustment for age, sex, smoking status and highest educational level (as a proxy for socio-economic status). Analysis in KNHANES also incorporated adjustment for year of participation in KNHANES.

In sensitivity analysis, participants with CPI4 had greater incidence risk for DMFS than participants with CPI<4 in both GLIDE and KNHAENS, with generally larger effect estimates than those seen in the main analysis (Table 2.3.6). There continued to be evidence for population-specific patterns of association, for example between periodontal status and standardized DFS, where the pattern of association was reversed in KNHANES compared to GLIDE.

Table 2.9: Fitted differences in caries indices between periodontal cases and controls in sensitivity

analysis.

GLIDE Cross-sectional analysis - Cases defined as CPI 4 in one or more sextants or 6mm pocketing around one or more teeth								
Age group	**% of cases**	**N**	**Fully adjusted ratio of index in cases:controls (95% CI)**					
			DMFS	**DMFT**	**DFS**	**Standardized DFS**	**Number of surfaces**	**Number of teeth**
Under 45	12.67	10,073	1.05 (1.01,1.09)	1.03 (1.00,1.05)	1.01 (0.97,1.05)	1.02 (0.99,1.06)	0.99 (0.97,1.00)	0.99 (0.99,1.00)
45 to 54.9	25.55	9,794	1.09 (1.07,1.12)	1.04 (1.03,1.06)	1.03 (1.00,1.05)	1.06 (1.03,1.08)	0.98 (0.97,0.99)	0.98 (0.97,0.98)
55 or older	42.54	8,824	1.04 (1.02,1.05)	1.01 (1.00,1.02)	1.01 (0.99,1.03)	1.02 (1.01,1.04)	0.98 (0.98,0.99)	0.99 (0.98,0.99)
Overall	26.25	28,691	1.05 (1.04,1.07)	1.02 (1.01,1.03)	1.01 (1.00,1.03)	1.03 (1.02,1.05)	0.98 (0.98,0.98)	0.98 (0.98,0.98)
KNHANES Cross-sectional analysis – Cases defined as CPI 4								
Age group	**% of cases**	**N**	**Fully adjusted ratio of index in cases:controls (95% CI)**					
			DMFS	**DMFT**	**DFS**	**Standardized DFS**	**Number of surfaces**	**Number of teeth**
Under 45	2.3	14,486	1.12 (1.00,1.26)	1.00 (0.91,1.11)	0.85 (0.76,0.95)	0.86 (0.77,0.97)	0.97 (0.96,0.99)	0.97 (0.96,0.99)
45 to 54.9	10.2	6,500	1.21 (1.12,1.30)	1.14 (1.07,1.21)	0.93 (0.84,1.04)	0.95 (0.86,1.05)	0.96 (0.94,0.97)	0.96 (0.95,0.97)
55 or older	12.1	12,845	1.01 (0.97,1.05)	1.00 (0.96,1.04)	0.96 (0.89,1.02)	0.96 (0.90,1.03)	0.99 (0.970,1.01)	0.99 (0.97,1.01)
Overall	6.6	33,831	1.06 (1.02,1.09)	1.01 (0.97,1.04)	0.91 (0.87,0.97)	0.92 (0.87,0.97)	0.99 (0.98,1.00)	0.99 (0.98,1.00)

Fitted differences are obtained from a multivariable Poisson regression model incorporating adjustment for age, sex, smoking status and highest educational level (as a proxy for socio-economic status). Analysis in KNHANES also incorporated adjustment for year of participation in KNHANES

2.3.4: Discussion

GWAS studies for dental diseases will need large sample sizes. Index linkage to dental records may provide one way to obtain dental data at large scale, and these data can also be used to explore the characteristics of different measures of caries and periodontitis to make informed decisions about the design and interpretation of GWAS studies. Large studies with index-linked longitudinal records may provide an opportunity to explore clinical questions such as the relative importance of caries and periodontitis as risk factors for tooth loss in different age groups and whether previous treatment decisions predict future tooth loss. To illustrate the utility of index-linked data in these questions, this analysis explored the related themes of the characteristics of tooth loss in a population in Northern Sweden and the covariance between estimates of caries and periodontitis.

Given that both questionnaire-derived and index-linked dental data appear to have utility in observational analysis, it is likely that they are suitable for inclusion in GWAS analysis, where the exposure variable is well-measured and sources of error in the exposure (such as batching effects or imperfect imputation) are unlikely to correlate with sources of error in the outcome measures. The notion that approximate measures in large samples can be used to obtain genetic association signals for a more refined trait was proposed as the 'proxy-phenotype' method published in 2014[234] and formalized more recently. In 2016, a conceptual model for informing decisions about 'quantity-quality trade-off' was proposed and applied using subjective well-being as a proxy for more refined psychological traits[235]. Specifically, this framework considers situations where genome-wide data are available for both a proxy and refined phenotype and aims to find analytical situations which minimize the root mean square error in meta-analysis. While there are several parameters which influence the root mean square error, statistical power to detect genetic associations for the refined phenotype is always improved by adding a proxy phenotype provided there is a high degree of genetic correlation between the two measures (correlation coefficient of 0.8 or higher). Conversely, the addition of proxy phenotypes with modest genetic correlation (correlation coefficient of 0.3) may only be desirable when the sample size is higher for the refined phenotype than the proxy measure, in order to prevent biased estimates of genetic effect (supplementary note 2 in [235]). For dental diseases, the utility of questionnaire data in GWAS will therefore depend not just on the ability to obtain these data in large sample sizes, but also on the ability of these questions to capture traits with similar underlying biology to more refined measures of oral and dental health. This is explored further with a formal test for genetic correlation in the

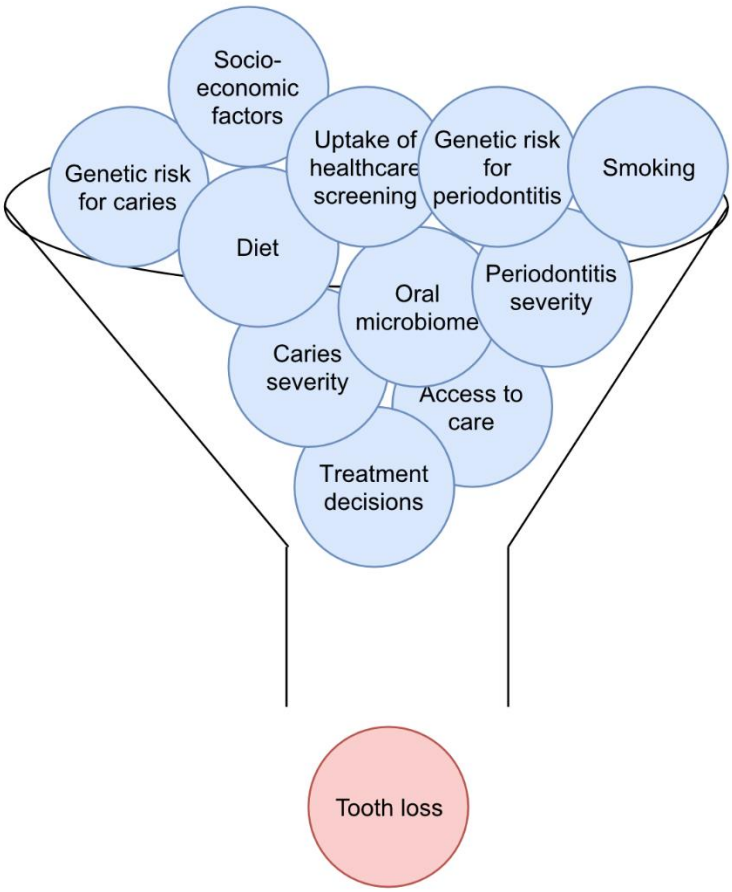
next chapter. Alongside the need to optimize power there is also a need to preserve interpretability, which might be challenged by the combination of phenotypic pairs connected by vertically pleiotropic pathways rather than phenotypic pairs which share similar underlying determinants.

Tooth loss as a grouping variable

In longitudinal analysis all the variables included in survival models were associated with hazard for tooth loss. Although these variables were chosen *a priori* as they were likely to be relevant, this still suggests that, viewed in measurement terms, tooth loss acts as a grouping feature which captures many upstream factors. It is likely that dental caries and periodontal status are the major determinants of tooth loss, but the relative importance of these appears to vary with age. Factors such as age, smoking status and educational attainment are associated with hazard for tooth loss even after adjustment for baseline DFS, periodontal status and previous tooth loss, re-iterating the importance of looking outside the mouth for predicting who is at risk of tooth loss. The association between educational attainment and hazard for tooth loss supports a role for socio-economic status as a risk factor for dental diseases, and this is discussed further in chapter 6.

Although the estimates obtained in this analysis are not estimates of causal effect, this may not matter in the clinical setting where the priority is to identify people at risk of poor oral health outcomes who would benefit from interventions. In this predictive context, the results confirm the relevance of known risk factors for tooth loss, while suggesting that, like findings for DGA earlier in this chapter, previous tooth loss captures a spectrum of features which are relevant to future oral health trajectory (Figure 2.4).

Figure 2.4: Tooth loss as a grouping variable



Results of longitudinal survival analysis suggest that tooth loss acts as a 'grouping variable' which brings together a range of competing and overlapping biological, behavioural and environmental factors which are relevant to oral health into a single measure. This could be visualized as a funnelling event: while the reasons for tooth loss may not be known for any patient, the presence of missing teeth at examination should raise concern about the presence of these upstream risk factors.

While this non-specificity may be advantageous for risk prediction in the clinical setting, it is a double-edged sword for epidemiological research. Missing teeth may be a practical and rapid way to assess a wide range of upstream risk factors which contribute to dental diseases.

Conversely, if missing teeth act as a grouping variable to capture a sizable portion of the human phenome then interpreting results of observational studies which use tooth loss as an exposure presents considerable challenge. Some of these features will be essentially random with regard to genotype, meaning that this may be less of a concern in the context of GWAS analysis. Nevertheless, it may be worth considering whether the environmental and biological events 'funnelled' into tooth loss are partially shared with those manifesting as other diseases: if so then conventional epidemiological studies using tooth loss as an exposure and other diseases may be difficult to interpret.

Methodological considerations for GWAS of dental caries and periodontitis

This analysis tested for covariance between caries indices and measures of periodontal status, showing that these measures are consistently non-independent in northern Sweden and Korea, and that some associations have population-specific magnitude and direction. Measures of dental caries and periodontitis are likely to contain non-random error with respect to the other dental disease, and the magnitude of this error is not predictable. This is potentially a concern for GWAS studies investigating dental diseases, as genotypes which specifically affect one dental disease may theoretically become associated with other dental diseases in association analysis. Conversely, the absolute magnitude of these associations was small in all age groups, in both populations and for all caries indices examined. In GWAS studies with finite statistical power it seems unlikely that these biasing effects are large enough to lead to false discovery, as a single genetic variant would need to have a very large effect on odds of periodontitis before this would be detectable at genome-wide significance for DMFS. Nevertheless, it may be prudent to include sensitivity analysis which compares genetic effects on DMFS across different strata of periodontal status, and to adopt a cautious interpretation acknowledging that these measures are not completely specific to a single dental disease. Understanding of the characteristics of these measures therefore informs the analytical decisions made in chapter 3 and the interpretation of results in chapter 4.

This analysis does not decompose the covariance into 'artefactual covariance' relating to measurement phenomena and 'biological covariance' related to common causes of caries and periodontitis or pathway effects by which dental diseases or their treatment influence other dental diseases. A wide range of common causes have been proposed, of which only a subset were adjusted for in the present study and using imperfect measures such as self-reported

smoking status. Differences in the relative importance of common risk factors or competing causes for tooth loss such as calculus and specific oral groups in the oral microbiome [236-239] between Sweden and Korea may explain the population-specific nature of covariance. Finally, pathway effects between dental diseases likely contribute, for example periodontal inflammation can occur secondary to dental caries in animal models[240] and interventions for caries may affect risk for periodontitis[236,237,241]. While it would be interesting to learn more about the relative importance of 'artefactual covariance' and 'biological covariance', the exposition of the total covariance is more important for the purposes of this dissertation, as this guides analytical decisions for GWAS as described above.

For periodontitis, clinical diagnosis of disease cannot be made using screening tools such as CPI[232], which is intended to be an epidemiolocal tool and has known limitations[242]. In these results, the CPI-based classification performs well in separating groups of participants who are at high risk for tooth loss from groups of participants who are at low risk for tooth loss, with large proportional hazard attributable to CPI-based measures of periodontal status in the longitudinal models. This suggests that, although these measures are not suitable for making person-level inference, they are useful in comparing groups in observational epidemiology and are likely a valid phenotype to explore in a GWAS investigation.

Suitability of adult cohort studies as a sampling frame for dental epidemiology

The Swedish GLIDE dataset described in this analysis used VIP as a source sampling frame. VIP recruits participants from a clearly defined target population which makes assessing the external validity of finding easier. VIP achieves high participation rates (>65%), and it is reported that there is little evidence of selection bias in this cohort[223]. It therefore seems likely that well-designed adult cohort studies like VIP provide a suitable sampling frame for dental epidemiology. Within this sampling frame, dental data were available for participants who had attended a dentist. This will exclude some people but will nevertheless capture a wider range of participants than would be represented using other strategies. For example, over 20% of patients seeking urgent dental care are reported to have high levels of dental fear[243]. These people would be unlikely to voluntarily attend an elective dental examination as part of a bespoke dental cohort, but do attend for urgent dental care when needed, so would be captured in this study.

Utility of dental phenotypes from index linkage

In Swedish GLIDE the use of index-linked phenotypes allowed a larger sample size than would be achieved by setting up a bespoke longitudinal dental research cohort, as reflected by the 300,000 person-years of dental follow-up included in analysis. Compared to questionnaire data, measures derived from dental records have higher resolution and are likely to contain fewer random errors.

Compared to a bespoke dental cohort there may be more variation in diagnostic interpretation as there was no examiner calibration exercise and it is not possible to assess features such as inter-examiner reliability. This is potentially relevant as there can be ambiguity in caries diagnosis of early lesions and there is only moderate agreement between dentists when scoring non-cavitated carious tooth surfaces[244]. Despite this, the dental data were from dentists working to the same protocols within the same organization with mandatory supplementary training, and measurement error in these scores due to diagnostic interpretation is likely to be random with respect to genotype. In addition, effect estimates in cross-sectional analysis had similar precision in Swedish GLIDE and KNHANES, suggesting that any additional random error in GLIDE was not of sufficient magnitude to have a noticeable effect on statistical power.

In conclusion, both questionnaires and index-linkage to dental records in primary care provide practical ways to obtain dental data in large studies, and both approaches have their own strengths and limitations. In the next chapter, both these sources of data will be utilized to explore the role of common genetic variation in the aetiology of caries and periodontitis.

2.4: Commentary

Epidemiological methods which use genetic data require large sample sizes. Questionnaires and index-linkage to dental records provide practical ways of acquiring dental data in large studies. Consideration needs to be given to the resolution and sources of random and non-random error in these data, but with appropriate strategies for analysis and interpretation it is possible to make valid inference from imperfect measures.

Standard clinical measures of caries and periodontitis are theoretically correlated through tooth loss, but in practice there appears to be little covariance between periodontal status and indices of caries exposure, suggesting that these standard clinical measures of caries and periodontitis predominantly capture different diseases and are likely to be suitable for inclusion in GWAS analysis.

The next chapter brings together both index-linked and questionnaire-derived sources of data to explore the role of common genetic variation in the molecular aetiology of caries and periodontitis.

Chapter 3: Genome-wide association studies for dental diseases in adult populations

Reliable genetic association signals can improve understanding of the biological and molecular aetiology of diseases. Used as phenocopies for a modifiable risk factor, these same variants can be applied in epidemiological analyses to help distinguish causal from confounded relationships. Previous GWAS for dental caries and periodontitis have not yielded consistent association signals, possibly because of the limited sample sizes and small effects of single genetic variants on these complex traits. The aim of this chapter is to identify single genetic variants or gene transcripts which are associated with dental diseases in adult populations, and characterize the nature of association signals using bioinformatic resources.. Results presented in this chapter have been published in *Nature Communications*[98].

3.1: Introduction

It is likely that dental caries and periodontitis are manifestations of a complex interplay of innate and environmental risk factors[245]. In adults, the cumulative effects of environment and lifestyle, access to dental treatment and decisions about dental treatment might partially mask genetic effects. In either case, single genetic variants are likely to have small apparent effects on these diseases which will only be detected by adequately powered studies. Considering these factors and the failure of previous GWAS for caries and periodontitis, any new analysis needs to find innovative ways to boost sample size and achieve statistical power.

One option is to expand sample size by including approximate but genetically related phenotypes of interest. A good example is birthweight, where self-reported birthweight is likely prone to random and non-random error compared to midwife-reported birthweight, but nevertheless has similar genetic determinants allowing both measures to be combined for gene discovery[143]. If dental disease can be measured using proxies, similar to clinical measures at a genetic level, it may be worth including these proxies in meta-analysis alongside more refined clinical measures.

This approach relies on a strong prior expectation for which pairs of clinical and self-reported traits should be combined. There may be no strong prior expectation, or the prior expectation might be that the approximate measure captures more than one dental disease. It may be

possible to clarify the meaning of a self-reported measure if clinical data are also available in the same population, but this will not always be possible.

Rather than checking the observational concordance between clinical and self-reported data it might be useful to check the genetic correlation to identify pairs of traits with similar underlying determinants. In recent years summary-statistic based genetic correlation estimators have been developed[47,48] which provide flexible methods for testing the genetic correlation between traits using GWAS summary statistics from overlapping or non-overlapping studies. Provided a clinical trait and a self-reported trait are present in a similar underlying population with similar ancestry, these new methods allow tests for genetic correlation even under situations where it is not possible to assess observational concordance.

The use of genetic correlation to inform genetic association discovery is not new, for example a recent study of Alzheimer's disease used participant-reported parental diagnosis of Alzheimer's as a proxy phenotype in GWAS, but used genetic correlation as a post-hoc test for the validity of this approach[246]. For dental caries and periodontitis, the suggestion is to extend this approach to proxy-trait selection where no single self-reported measure has a clear prior claim, allowing the empirical genetic determinants to guide trait selection.

Provided there is some genetic similarity, this can be exploited to boost statistical power, even if the estimates of genetic effects on these traits are not on the same scale. One approach involves meta-analysis of similar traits, aiming to capture the genetic variation which is relevant to both contributing traits. Examples include the combination of head circumference and intracranial volume to identify variants relevant to cranial dimension[96], and the combination of hay fever, asthma and eczema to identify variants relevant to multiple allergic diseases.[247]

The overall design of this investigation therefore involves identifying pairs of clinical and self-reported proxy phenotypes which are have similar genetic determinants using an empirical approach based on genetic correlation. Following this, these pairs of traits are combined in meta-analysis to expand effective sample size, and the association signals arising from these combined meta-analyses are characterized using bioinformatic follow-up tools.

3.2: Methods

3.2.1: Population

GWAS were performed in two resources. The GLIDE consortium included analysis of 4 clinical dental phenotypes. Each participating cohort performed GWAS analysis, and then single-trait results were combined in meta-analysis. In parallel, GWAS for 6 proxy phenotypes was performed in UK Biobank. The study population for each participating cohort in GLIDE is described below, and the UK Biobank study is described at the end of this section.

Studies in GLIDE

ARIC

The Atherosclerosis Risk in Communities (ARIC) study is a multi-centre prospective cohort study which was established to investigate the aetiology of atherosclerotic vascular diseases[248] with a focus on vascular and lipid phenotypes. Between 1987 and 1989 participants living in four communities in the US who were aged between 45 and 64 years were invited to participate. Initial recruitment took place in Forsyth County, North Carolina; Jackson, Mississippi; suburban districts of Minneapolis, Minnesota and Washington County, Maryland; achieving an initial sample size of 15,792 participants with approximately equal numbers in each centre. Since then, there have been 6 clinical visits and regular remote follow-up of study participants, along with linkage to registers such as a cancer registry[249]. During the fourth clinical follow-up visit in 1996-1998 a dental sub-study was performed including oral examination for around 6000 participants[250]. Dental examination included detailed assessment of caries status and multiple measures of periodontal health[78]. Ethical approval was obtained from the institutional review board at the University of North Carolina, North Carolina.

COHRA

The Center for Oral Health in Appalachia cohort 1 (COHRA1) is a family-based cohort study and part of the GENEVA caries consortium[251] which is described further in chapter 5. The study recruited children and parents of children between 2002 and 2009[69,252]. Examination included full charting of caries status[253] and an abbreviated periodontal examination. If participants had two or more sextants with probing depths of 5.5mm or deeper than full-mouth pocket depth charting was performed on a targeted basis[76]. Ethical approval was obtained from The University of Pittsburgh Institutional Review Board, Philadelphia, and site

approval was obtained from West Virginia University Institutional Review Board, West Virginia.

DRDR

The Dental Registry and DNA Repository (DRDR) is a volunteer-patient cohort which recruited patients seeking care at the School of Dental Medicine, University of Pittsburgh between 2006 and 2007[254]. The study aimed to investigate possible relationships between caries experience and systemic diseases and recruited participates aged 17-84 years[70]. Participants underwent comprehensive dental examination as part of their care at the School of Dental Medicine, and donated DNA for use in research. Ethical approval was obtained from the University of Pittsburgh Institutional Review Board, Pennsylvania.

HCHS/SOL

The Hispanic Community Health Study / Study of Latinos (HCHS/SOL) is a multi-centre cohort study which aimed to describe the health status of Hispanic/Latino people living in communities in the US and investigate the aetiology of a range of diseases which are prevalent in these communities. The study recruited approximately 16,000 self-identified Hispanic/Latino participants between 2008 and 2011 from four centres targeting urban areas of the US with large Hispanic / Latino populations, specifically the Bronx borough of New York, New York; Chicago, Illinois; Miami, Florida and San Diego, California[255]. The study design aimed to account for under-representation of certain age groups and socio-demographic groups by using a stratified two-stage probability design which first defined target postcodes based on census data, then over-sampled participants based on the age characteristics of households which replied[255]. These sampling weights were later log-transformed and used as covariates in GWAS analysis as described later in this section. The study included baseline dental examination with measures of both dental caries and periodontal status[66]. The study was approved by the institutional review boards of all participating entities.

MDC

The Malmö Diet Cancer study (MDC) recruited in the region of Malmö, Sweden between 1991 and 1996, at a time when the city had a population of around 230,000 inhabitants[256]. The study aimed to investigate relationships between diet and cancer incidence and mortality.

People living in Malmö who were born between 1923 and 1950 (aged 45-64 years at time of contact) were invited to participate, aiming to achieve a sample size of around 53,000 participants. Follow-up has included register-linked data and detailed phenotypic assessment on targeted subsets of the study, for example clinical measures of atrial fibrillation for study participants whose linked hospital records indicated a diagnosis of atrial fibrillation[257]. Within MDC, dental data were obtained from the GLIDE database, following the same index linkage procedure described for Swedish GLIDE in chapter 2, meaning that dental charts were originally produced by a dentist working in a Folktandvården clinic. The MDC study received ethical approval from the Ethics Committee at Lund University, Sweden.

NFBC1966

The Northern Finland Birth Cohort 1966 (NFBC1966) is a birth cohort study which was established to investigate the health and development of babies born in 1966, with an emphasis on neurological outcomes. The study recruited in the Oulu and Lapland provinces of Northern Finland, where all babies born in the year 1966 were eligible to participate. 12,058 live births were recruited into the study, representing 96% of the all eligible births in 1966[258]. Longitudinal follow-up has included collection of both clinical and questionnaire data, and between 2012 and 2013 a detailed clinical follow-up clinic took place when the participants were aged 46-47 years which included dental examination[259]. Approval for NFBC1966 was obtained from the Ethical Committee of Northern Ostrobothnia Hospital District, Finland.

SHIP cohorts

The Study of Health in Pomerania (SHIP) is a cohort study which aimed to investigate associations between risk factors, subclinical disorders and clinical diseases for a range of common health outcomes. The study recruited between 1997 and 2001 and targeted participants aged 20-81 years living in West Pomerania, north-eastern Germany, using a two-stage cluster sampling design stratified by sex and age[260] resulting in a target population of approximately 6,300 participants and achieving a 69% response rate of the target population[261]. In 2008-2012 an additional cross-sectional study (SHIP—TREND) was conducted using a stratified random sample from population registers, again in the west Pomerania region. Both SHIP and SHIP-TREND included clinical dental examination

following a similar protocol[262]. Both studies were approved by the Local Ethics committee of the University of Greifswald, Germany.

TWINGENE

The Twingene study (TWINGENE) is a genotyped sub-study of the Swedish Twin Register (STR) which aims to identify effects of specific genetic variants on a range of health traits and diseases. STR is a register which was established in the 1950s and includes more than 170,000 twins born between 1886 and the present day[263]. Since 2009, the parents of twins have been contacted when the twins reach 9 years of age with an invitation to participate in research[264], and 11 other research studies have used the STR as a sampling frame to recruit twins, collectively forming the Swedish Twin Biobank[264]. One of these sub-studies was Twingene, which recruited between 2004 and 2008. All twins born before 1958 who were registered in STR were contacted and invited to participate (aged 46-93 years), resulting in a sample size of approximately 10,600 twins[265]. Of these, dental data on a subset were available in the GLIDE database, following the same procedure as described for MDC and previously in chapter 2. TWINGENE received ethical approval from the Local ethics committee at Karolinska Institutet, Sweden.

WGHS

Much as TWINGENE is a continuation and evolution of existing research infrastructure, the Women's Genome Health Study (WGHS) aimed to create a database of genome-wide genetic data for women who were participating in the pre-existing Women's Health Study (WHS), which initially recruited between 1992 and 1995[266]. WHS initially aimed to evaluate the benefits and risks of aspirin and vitamin E in women with no previous heart disease or major chronic illness, adopting a factorial randomized controlled trial design[267,268]. WHS recruited participants aged 45 and older from the female healthcare workforce population of the US, achieving an initial sample size of around 40,000 participants[266]. In contrast to the other studies in GLIDE, WHS did not include clinical dental examination. Instead, participants were asked whether they had been diagnosed with periodontal disease at baseline, and this question was repeated at 8 intervals during the trial phase of the WHS, and one year after completing the trial phase[269]. Subsequently, WHS was used as a sampling frame for the genotyped study, WGHS. The WGHS received approval from the Brigham and Women's Hospital Institutional Review Board for Human Subjects Research.

BBJ

The Biobank Japan (BBJ) was designed to create common infrastructure for research into many traits in the Japanese population. The design aimed to over-sample patients with major diseases by using hospital-based register data at 12 medical institutions to invite people with any of 47 target disease to participate in research. Initial recruitment occurred between 2003 and 2008[270]. Participants were aged between 20 and 80 years at baseline and agreed to donate DNA at the start of the study and annual serum samples until 2013. For participants with certain diseases, follow-up has included survival data. The initial recruitment phase included 200,000 participants, including approximately 4,000 with a hospital diagnosis code of periodontal disease[75,270]. BBJ received approval from the research ethics committees at the Institute of Medical Science of the University of Tokyo, the RIKEN Yokohama Institute and each of the 12 participating hospitals.

TMDUAGP

The Tokyo Medical and Dental University Aggressive Periodontitis study (TMDUAGP) was established to investigate possible genetic risk factors for aggressive periodontitis in the Japanese population. Patients at the periodontal department of Tokyo Medical and Dental University who had been diagnosed with aggressive periodontitis, lived in the Tokyo urban area of Japan and were aged between 22-46 years were invited to participate in the study[271], with healthy controls drawn from the control group of a previous study[272]. TMDUAGP received ethical approval from the ethical committee of Tokyo Medical and Dental University.

UKB

UK Biobank (UKB) is a large-scale prospective cohort study which was introduced in chapter 2. The study includes approximately 500,000 people, aged between 40 and 69 years, who were recruited between the years 2006 and 2010 from densely-populated regions of England, Scotland and Wales[106,273]. During the baseline visit participants were asked to complete a comprehensive touch-screen questionnaire which included one question stem related to oral and dental health with several possible responses. UK Biobank received ethical approval from the National Health Service Health Research Authority North West Research Ethics Committee (16/NW/0274).

3.2.2: Genotypes and genetic data quality control

In ARIC DNA was extracted from peripheral blood and genotyped using the Affymetrix Human SNP Array 6.0. Genotypes were called using Birdseed[274], with cut-offs of minor-allele frequency (MAF) ≥ 0.01, call rate ≥0.95 and p value for violation of Hardy-Weinberg equilibrium (PHWE) > $1x10^{-5}$. After initial calling, participants who clustered outside the largest ancestral group, who had relatedness by state of second degree of higher, with mismatch between reported and genetic sex or with extreme heterozygosity were excluded. Additionally, participants with an overall call rate less than 0.95 were excluded. Genotypes were pre-phased using SHAPEIT[275], HapiUR[276] and Eagle[277] before imputation to 1000 genomes phase 1 version 3 (1KGP1V3) using Minimac3[278].

In both COHRA1 and DRDR DNA was extracted from saliva and genotyped using the Illumina Human640 Quadv1_B array. Genotypes were called using BeadStudio (v3.3.7, Illumina, San Diego, California), with cutoffs for MAF ≥ 0.02, call rate ≥ 0.95 and PHWE > $1x10^{-4}$. Genotypes were phased using SHAPEIT[275] and imputed using IMPUTE2[279] to 1000 genomes phase 1 version 2 for COHRA1, and 1KGP1V3 for DRDR.

In HCHS/SOL DNA was extracted from peripheral blood then genotyped using a custom array based on an Illumina Omni 2.5-8v1.1 backbone with 150,000 additional custom markers. Genotypes were called using GenomeStudio (Illumina, San Diego, California), with no cutoff for MAF but a strict cutoff for call rate (≥0.98) and PHWE >$1x10^{-5}$. Participant level QC included checks for sex mismatch, gross chromosomal anomalies, relatedness, batch effects and discordance between duplicate samples. Participants with a high Asian ancestry fraction were removed, with no other restrictions for admixture proportion. Genotypes were phased using SHAPEIT2[280] and imputed to 1KGP1V3 using IMPUTE2[279].

In MDC DNA was extracted from peripheral blood and genotyped using the Illumina HumanOmniExpress BeadChip (v1) array. Genotypes were called using GenomeStudio (Illumina, San Diego, California), with cutoffs of MAF ≥ 0.01, call rate ≥ 0.95 and pHWE >$1x10^{-7}$. Genotypes were phased using SHAPEIT2[280] and imputed to 1KGP1V3 using IMPUTE[279].

In NFBC1966 DNA was extracted from peripheral blood and genotyped using the Illumina HumanCNV-370DUO array. Genotype calls were made using BeadStudio (Illumina, San Diego, California), with cutoff values set at MAF ≥ 0.01, call rate ≥ 0.95 and PHWE > $1x10^{-4}$. Genotypes were imputed to 1KGP1V3 using IMPUTE[279].

In SHIP DNA was obtained from peripheral blood and genotyped using the Affymetrix Human SNP array 6.0. Genotypes were called using Birdseed2[274] with no cutoff for MAF but a minimum call rate ≥ 0.8 and PHWE > $1x10^{-4}$. Overall, each participant needed to have a call rate > 0.92 to be carried forward to analysis. Duplicate samples and individuals with mismatch between reported and genetic sex were excluded. Imputation was performed using IMPUTE2[279] and the 1KGP1V3 reference panel.

Genotyping in SHIP-TREND was performed using the Affymetrix Human SNP Array 6.0 and peripheral blood. Variants were called using GenCall (Illumina, San Diego, California) with no cutoff value for MAF but a minimum call rate of ≥ 0.9 and PHWE > $1x10^{-4}$. Participants with a call rate < 0.94 or a mismatch between reported and genetic sex were excluded. Imputation used the 1KGP1V3 reference panel and IMPUTE2[279] software.

In TWINGENE DNA was extracted from peripheral blood and genotyped using the Ilumina OmniExpress array. Genotypes were called using GenomeStudio (v. 2010.3, Illumina, San Diego, California) with no cutoff for MAF but a threshold call rate ≥ 0.97 and PHWE >$1x10^{-5}$. Samples with an overall call rate <0.97 were excluded. Concordance in reported and genetic sex was checked and discordant samples were excluded. Outliers in heterozygosity (with a mean F score more than 5 standard deviations away from the population average mean) were also removed. Pre-phasing used SHAPEIT[275], while imputation used Minimac3[278]. and the 1KGP1V3 reference panel.

In WGHS DNA was extracted from peripheral blood and genotyped using either the Illumina HumanHap300 Duo '+' array or the combination of Illumina HumanHuman300 Duo and Illumina iSelect arrays. Genotypes were called using BeadStudio (v.3.3, Illumina, San Diego, California) with a threshold call rate ≥ 0.98 and PHWE > $1x10^{-6}$. Participants with overall call rate <0.98 were excluded, as were participants who clustered outside the largest ancestral

group in PCA. Phasing was performed using MaCH[281] (v1.0.16) and imputation to 1KGP1V3 was performed using Minimac[278].

In BBJ DNA from peripheral blood was genotyped using either the combined Illumina HumanOmniExpressExomeBeadChip or a combination of Illumina HumanOmniExpress and HumanExomeBeadChip arrays. Genotypes were called using GenomeStudio (Illumina, San Diego, California) with no threshold for MAF but minimum call rate ≥0.99 and PHWE > $1x10^{-6}$. Participants who clustered outside the largest group in PCA analysis, with mismatch between reported and genetic sex and with identity by state greater than second degree relatives were excluded from analysis. Imputation used the 1000 Genomes East Asian ancestry samples.

In TMDUAGP, DNA was extracted from peripheral blood and genotyped using the Illumina HumanCoreExome-12(v1.1) and HumanCoreExome-24(v1.0) arrays. Genotypes were called using GenomeStudio (Illumina, San Diego, California), with thresholds of MAF ≥ 0.01, call rate ≥ 0.95 and PHWE > $1x10^{-7}$. Duplicate samples and individuals with mismatch between reported and genetic sex were identified and excluded. Pre-phasing used SHAPEIT2[280], imputation used Minimac[278] software and 1KGP1V3.

Participants in UKB were genotyped using the UK BiLEVE and UK Biobank axiom arrays. Pre-imputation quality control (QC), phasing and imputation were performed at the Wellcome Trust Centre for Human Genetics prior to data release[273,282]. Single variant analysis used the 2018 (v3) imputed data release including imputation to both Haplotype Reference Consortium and UK10K/1000 genomes combined panel. Additional QC of the imputed data was performed at the University of Bristol, which included restricting analysis to participants of European ancestry (identified as the largest cluster formed after k-means clustering on the first 4 genetic principal components, N=464,708) as well as standard exclusions for mismatch between reported and genetic sex, possible sex chromosome aneuploidy, outliers in heterozygosity and missing rates. An additional 2 participants were removed as they were apparently related to a very large number of participants at 3rd degree or closer; otherwise related participants passing all other quality control were included in single variant analysis. A full description of these additional QC stages has been published[283]. For HLA haplotype analysis imputation of classical HLA haplotypes was performed using

the HLA*IMP:02 algorithm with a multi-population reference panel[273] and analysis was restricted to unrelated participants who self-identified as 'White British' and were located within the largest cluster in k-means analysis.

Following these study-specific measures, central QC of the association result files was performed using EasyQC software[284] and 1KGV1P3 reference data. All genomic positions were harmonized to NCBI Build 37 of the human genome during imputation and alignment was checked centrally by comparing rsid, reported position and reported alleles to reference data. Variants were dropped if the reported alleles did not match those in 1KGV1P3 reference data, or if there was a large discrepancy between reported effect allele frequency and reference allele frequency for the same allele (absolute difference in EAF>0.3). Variants were excluded if they had low imputation quality (below 0.3 for MACH, 0.4 for IMPUTE INFO score), departed from Hardy-Weinberg equilibrium (HWEP $<1x10^{-6}$ in the entire sample for continuous traits or in controls for periodontitis), or had per-file minor allele count <6. Following meta-analysis, variants with MAF < 0.01 or variants present in less than 50% of the total sample were dropped. The decision to restrict analysis to common and low-frequency variation with MAF >= 0.01 was motivated by the prior expectation of a polygenic genetic architecture with small effect sizes which would not be detected with adequate statistical power below MAF of 1%. The decision to restrict analysis to variants represented in at least 50% of the study was motivated by the desire to compare consistency in genetic effects across studies.

3.2.3: Phenotypic derivation

The 4 principal traits in GLIDE were DMFS, DFS per available surface (DFSS), number of teeth (Nteeth) and periodontitis as a case-control trait using a disease-affected or unaffected classification. DMFS was derived as the sum of decayed, missing and filled tooth surfaces (excluding third molar teeth) as described in chapter 2. DFSS was derived as the number of decayed and filled tooth surfaces (excluding third molar teeth) divided by the number of remaining natural tooth surfaces (where bridges, dentures and implants were not considered to be natural tooth surfaces). Nteeth was derived as the total number of natural permanent teeth, excluding wisdom teeth and prosthetic replacement teeth such as bride pontics, dentures and implants. After derivation, these measures were transformed as described in table 3.1

Ideally periodontitis would be defined using the same criteria in all studies. In this consortium however, protocols for periodontal examination varied between studies, meaning a compromise was needed to allow each study to define cases and controls for periodontitis based on measures which had been collected using their protocol. The possible impact of this as a source of heterogeneity in genetic effect estimates is discussed later in this chapter. In ARIC, SHIP, SHIP-Trend and HCHS/SOL, criteria published by the Centers for Disease Control and Prevention/American Academy of Periodontology (CDC/AAP)[285] were used. In COHRA participants were classified as cases if two or more sextants had probing depth of at least 5.5mm, or if participants reported ever having "gum surgery". In TwinGene and MDC, participants were classified as cases if at least two tooth surfaces had probing depth of 5mm or deeper, or at least four tooth surfaces had probing depth of 4 mm or deeper. In BBJ participants were classified as cases or controls based on clinical diagnosis by physicians at participating hospitals, with data retrieved from diagnosis codes in hospital registers. In TMDUAGP, the participants were classified as cases based on the classification developed at the 1999 international workshop for a classification of periodontal disease and conditions[286]. In WGHS the participants reported if they had periodontal disease or not. The question stems were "Since you started the trial (around 3 years ago) / In the past year, were you newly diagnosed with / have you had any of the following", with "Periodontal disease" as one possible response. Participants who selected this response were asked to provide the month and year of diagnosis. Participants with no teeth were excluded from scoring the primary dental disease traits in GLIDE.

In UK Biobank (UKB) the baseline questionnaire stem asked participants "Do you have any of the following? (You can select more than one answer)". The possible answers included "Dentures", "Bleeding gums", "Painful gums", "Loose teeth", "Toothache" or "Ulcers". Participants who selected one of these answers were coded as cases (1) for each respective analysis, while participants who did answer the question but did not select that answer were coded as controls (0). Participants who answered with 'Prefer not to answer' or who did not complete the questionnaire were coded as missing and excluded from analysis.

Table 3.1: Dental disease phenotypes included in GLIDE meta-analysis

Trait name	DMFS	DFSS	Nteeth	Periodontitis
Trait	Decayed, Missing and Filled tooth Surfaces	Decayed and Filled tooth Surfaces standardized to number of tooth surfaces	Number of natural teeth	Presence or absence of periodontitis
Transformation	Residuals generated after regression on age, age squared and study-specific covariates. Residuals standardized to mean of 0 and SD of 1	Residuals generated after regression on age, age squared and study-specific covariates. Inverse normal rank transformation of residuals.		-
Phenotypic assessment	Derived from clinical dental records		Derived from clinical dental records or self-reported (WHGS)	Centers for Disease Control and Prevention/American Academy of Periodontology definitions[285] (4 studies) Two or more tooth surfaces with probing depth ≥ 5mm, or at least four tooth surfaces with probing depth ≥ 4 mm (1 study) Probing depth ≥ 5.5mm in 2 or more sextants (1 study). Participant-reported diagnosis of periodontitis (1 study)

Details in the table, text and figures are reproduced from the published manuscript[98].

3.2.4: Participant-level tests for genetic association

For quantitative traits (DMFS, DFSS and Nteeth), derived variables were created which aimed to reduce variation in the phenotype attributable to age and to improve the statistical distribution of these traits prior to tests for association in GWAS. First, raw phenotype scores were regressed on age, age squared, and other study-specific covariates such as genetic principal components (Table 3.2) and the regression models were used to predict residuals. Predicted residuals were transformed to a mean of zero and standard deviation of 1. For DFSS and Nteeth (but not DMFS), these residuals had a markedly non-normal distribution, so inverse normal rank transformation was applied. These transformation stages were performed separately in male and female participants except for family-based studies, where sex was instead included as an additional covariate in phenotype preparation. For quantitative traits the transformed scores were then regressed on genotype dosage using linear regression and an additive genetic model to estimate the genetic main effect, using a range of publicly available software tools. For chromosome X haploid allele calls for male participants were coded as 0 or 2 to ensure dose equivalence to female participants (coded 0, 1 or 2) in a random inactivation model.

For periodontitis, the raw disease classification codes (0 or 1) were regressed on genotype dosage in a logistic regression framework, where relevant covariates (age, age squared, sex, genetic principal components and other study-specific covariates) were included in the model in a single stage.

In HCHS/SOL, a more complex modelling approach was used to account for admixture, where the participants self-reported ancestry from 6 broad groups (Cuban, Dominican, Mexican, Puerto Rican, Central American, South American), and relatedness, where the study over-sampled from families who responded to invitations and therefore included many related participants. Multi-dimensional clustering was used to generate genetic analysis groups containing participants of similar ancestry. Next, group allocations were then used as covariates in a linear mixed model (partitioned to only fit the proportion of genetic structure due to familial relatedness rather than ancestry) alongside the first 5 genetic principal components, study center and log-transformed sampling weights relating to the sampling strategy described earlier in this chapter[287].

Unlike HCHS/SOL, UKBiobank did not specifically target families but the sample nevertheless includes many related participants. In order to preserve statistical power, GWAS were performed using a linear mixed model (LMM) implemented in BOLT-LMM (v2.3)[128]. Age, age squared, sex and genotyping array were included as covariates in association testing. BOLT-LMM association statistics are on the linear scale. As such, test statistics (beta coefficients and their corresponding standard errors) were transformed to log odds ratios and their corresponding 95% confidence intervals on the liability scale using a Taylor transformation expansion series as detailed in a full description of the GWAS pipeline which has been published online[288].

Table 3.2: Study-specific covariates used in participant-level tests for genetic association in GLIDE

Study	Study-specific covariates	Analysis software used in GWAS
ARIC	Study centre, PC1-PC10	PLINK[289]
COHRA	PC1, PC2	PLINK
DRDR	PC1, PC2	PLINK
HCHS/SOL	Study centre, PC1-PC5, genetic analysis group, log transformed sampling weights	GENESIS[290]
MDC	PC1-PC10	SNPtest[291]
NFBC1966	PC1-PC4	SNPtest
SHIP	None	SNPtest
SHIP-TREND	None	SNPtest
TWINGENE	PC1, PC2	PLINK[292]
WGHS	PC1-PC10	ProbABEL[292]
BBJ	PC1, PC2	Mach2dat[281]
TMDUAGP	PC1, PC2	Mach2dat

3.2.5: Single-trait meta-analysis

For the 4 primary dental traits in GLIDE, fixed-effects genome-wide meta-analysis were performed using METAL[293] and single genomic control[294] for correction of input summary statistics (max λ_{GC}=1.07). For DMFS, DFSS and Nteeth the primary meta-analysis included all available results, and sensitivity analysis excluded results from HCHS/SOL. For periodontitis, the primary meta-analysis included all results apart from the two studies with participants of East Asian ancestry (BBJ, TMDUAGP), and results of these two studies were instead combined in an East Asian ancestry-specific meta-analysis including 17,287 participants.

3.2.6: Identifying clinical and self-reported measures with similar genetic determinants

Genetic correlation coefficient estimates (r_g) and standard errors were obtained using bivariate LDSR. First, summary statistics from single-trait meta-analysis (GLIDE) or GWAS (UKB) were restricted to a subset of approximately 1 million high confidence variants (with MAF > 0.01, present in HapMap3[295]). Pre-computed LD scores and weights (excluding the HLA region) derived in 1KG Phase 3 European ancestry data were downloaded from the LDSCORE repository (https://data.broadinstitute.org/alkesgroup/LDSCORE/). Heritability (h^2_{LDSR}) was estimated for each trait using univariate LD score regression[49] (hereafter 'univariate LDSR') with an unconstrained intercept term to evaluate possible inflationary bias in GWAS summary statistics using the –h2 flag in the LDSC standalone software (v1.0). Genetic correlations between each pair of traits (r_g) were then estimated with bivariate LD score regression[47] (hereafter 'bivariate LDSR') with an unconstrained intercept term to allow

for cryptic sample overlap, implemented using the –rg flag in the LDSR software. With the exception of TMDUAGP, which was not included in the primary analysis, studies in GLIDE and UKB did not recruit participants based on their dental disease status, so heritability estimates on the observed and liability scale were equivalent, and no transformation of heritability estimates was performed.

3.2.7: Multi-trait meta-analysis

Pairs of traits with similar genetic determinants were then combined using a fixed effects z-score meta-analysis implemented in METAL[293]. Z-scores from GLIDE and UKB were weighted, either by the sample size (for continuous traits) or effective sample size for binary traits. If there was detectable inflationary bias in the single-trait results, these were corrected for the estimated magnitude of the inflationary bias using the univariate LDSR intercept term (for dentures $LDSR_I$=1.0784, for loose teeth LDSR=1.0432) prior to multi-trait meta-analysis.

To define associated loci, genome-wide significance was set at $P<5x10^{-8}$ and the variant with the smallest P value for association within a +/- 500 kilobase (kb) window was defined as the lead variant for any given locus using the –indep function of EasyStrata[296]. A stepwise-selection procedure was performed to confirm that lead variants at these loci are conditionally independent of other lead signals and to identify loci containing multiple independent signals of association. This procedure used approximate conditional analysis[297], taking genome-wide summary statistics in conjunction with reference haplotypes for LD estimation from the UK10K project[298] and was implemented using the cojo-slct function of Genome-wide Complex Trait Analysis (GCTA) standalone software (version 1.91.4)[299]. If a locus was defined in the first stage using EasyStrata but contained no signal which was conditionally-independent of other signals on the same chromosome (i.e no variant in that locus reached genome-wide significance in the stepwise selection procedure) then the locus was not reported as a lead signal. If a single locus contained multiple signals or association then the lead variant was reported as the primary finding, and the additional signals were considered secondary findings. The HLA region was defined as chromosome 6: 25-35 megabases. Within this region the LD structure is reported to be variable[300], which precluded use of approximate conditional analysis. Instead, a single lead variant was reported for this entire super-locus and more detailed mapping was performed using imputed HLA haplotypes rather than single variants

Z-scores and effect allele frequencies in combined analysis were then used to derive standardized regression coefficients, where the combined phenotype was arbitrarily assumed to have a standard deviation of 1. The aim of this transformation was to compare the relative magnitude of genetic effects in the combined analysis and provide genetic effect estimates on a continuous scale for subsequent transformations.

To evaluate directional consistency in genetic effects for lead associated variants, the number of SNPs with directionally-consistent effects in GLIDE and UKB was counted. To test for consistency in effect estimates, Pearson correlation coefficient was calculated, including a penalty term for imprecisely-estimated genetic effects by allocating each variant a weight corresponding to inverse-variance (1/SE) of the genetic effect estimate in the GLIDE dataset. For combined analysis of DMFS/dentures tests for consistency in effect direction and size used independently-associated lead variants defined using the criteria described above. For combined analysis of periodontitis/loose teeth only a single locus met the criteria described above, so tests for consistency in effect direction and size used independent suggestively-associated loci ($P<1x10^{-5}$) defined using a stepwise selection procedure.

3.2.8: Follow-up analysis

For each principal multi-trait meta-analysis, inflation in summary statistics attributable to polygenic heritability and inflationary bias were estimated using the LDSR method described previously.

To evaluate enrichment of association signal in genomic regions with specific functional annotations, partitioned heritability[301] was estimated using the –ref-ld-chr argument of LDSR and pre-computed baseline and stratified models derived from 1000 Genomes phase 3 data (baselineLD_v2.1, November 2018 release). To test for enrichment of association signal in genomic regions with tissue-specific patterns of expression, segmented LDSR using tissue-specific annotation files[302] was performed using the –h2-cts flag and multi-tissue gene expression files derived using 1000 Genome phase 3 data (April 2018 release, downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/ in January 2019) . This file annotates regions of the genome which are preferentially expressed in specific tissues based on expression data from both the GTEx project[44] and expression data from the Franke lab (http://ludesign.nl/frankelab/)[303].

Cross-trait comparison at a genome-wide level was tested using hypothesis-free genetic correlation with publicly-available summary results from published GWAS studies. Summary results from combined meta-analysis were uploaded to the LD hub[304] resource (https://ldsc.broadinstitute.org), which uses LDSR to estimate r_g values and standard errors while allowing for known and cryptic sample overlap with an unconstrained bivariate intercept term. All available traits in the LD hub catalogue as of 9th January 2019 were included with no screening to remove duplicate or similar traits, however results of GWAS solely obtained in UKB were not included to prioritize external sources of information.

At a single variant level, cross-trait comparison was made using the PhenoScanner resource[305] (http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner, accessed January 2019), a searchable catalogue of published GWAS summary statistics. For each lead variant in the dental disease combined meta-analysis defined using the procedure described above and potential proxies for that variant (LD $r^2>0.8$ in 1KGP3 European ancestry reference data), summary statistics of tests for association with other traits were searched. If there was evidence for association with a trait in the catalogue (excluding dental diseases and defining association at a conventional threshold of $P<5x10^{-8}$) then this was considered an 'overlap' in genetic architecture at a single variant level, and details of the trait, summary results and publication references were saved. This analysis was summarized by describing the proportion of dental disease variants which 'overlap' with 1 or more non-dental disease traits.

The online functional annotation tool FUMA[52] (http://fuma.ctglab.nl/, accessed January 2019) was used to visualize co-localization of single variant association signal with expression quantitative trait loci (eQTL) and chromatin interaction data. This approach included data on cis-eQTLs (defined as up to 1megabase between SNP and gene) from all available tissues in GTEx v6 and v7. Chromatin interaction data was obtained from HiC assays in 21 tissues.

Where dental disease association signal co-localizes with a known eQTL or chromatin interaction, this might help map the association signal to a biologically-causal gene but might also be co-incidental, given the large number of reported eQTL signals for many transcripts[306]. To help resolve this, a transcriptome-wide association study was performed

using the S-PrediXcan method[307]. First, association between predicted variation in gene transcription and dental disease traits was estimated for 48 tissues in the GTEx project[44], using pre-trained prediction models. Next, tissue-specific predictions were combined using the S-TissueXcan method[45], which integrates information from all tissue-specific predictions to evaluate the overall impact of a transcript on phenotypic variation while accounting for correlation in gene transcription across different tissues. Analysis and post-processing of results was performed using a cloud-based pipeline (https://cloud.hakyimlab.org/, accessed January 2019). Finally, a Bonferroni correction was performed with alpha set at 0.05 divided by the number of transcripts tested.

Gene set and gene pathway analysis was performed using two approaches. DEPICT[308] (version 1.1, release 194), was used to test whether associated loci are enriched in pre-defined sets of genes with similar biological function. DEPICT uses GWAS results for randomly distributed phenotypes to calibrate a null expectation and estimate false discovery rate for enriched gene sets so appears well calibrated compared to other methods and does not need further adjustment for multiple testing. Another advantage of DEPICT is the definition of gene sets, which are derived empirically based on gene co-expression, so may be more biologically informative and unbiased than gene sets based on published literature. Associated loci were defined at ($p<5x10^{-8}$) in combined analysis. Gene sets, tissue expression matrix and gene annotation files were downloaded from the DEPICT repository (https://data.broadinstitute.org/mgp/depict/documentation, accessed January 2019).

A hypergeometric test for gene set enrichment was performed using FUMA[52]. The main advantage of this method is that associated genes can be parsed manually, allowing the results of S-TissueXcan to be used to define associated genes. Because this integrates eQTL data, the definition of associated genes may be more biologically informative than the locus-based definition used by DEPICT. Associated genes were defined as those passing Bonferroni correction in S-TissueXcan analysis as described above. Pathways and gene sets were taken from the Molecular Signatures Database (MSigDB C2), and all protein coding genes in the FUMA database were used as background to evaluate enrichment. Compared to DEPICT, the null expectation for the hypergeometric test is less well defined, meaning a conservative approach to correction of enrichment P values for multiple testing was chosen, and a Bonferroni correction was used.

To test whether the inclusion of participants of Hispanic/Latino ancestry in the primary meta-analysis contributed to heterogeneity in genetic effect estimates for DMFS, genetic effect sizes were compared in HCHS/SOL and other studies in GLIDE. First, meta-analyses in GLIDE were repeated excluding the HCHS/SOL study. Next, estimates from HCHS/SOL and other studies were contrasted using a test for difference in genetic effect implemented using the CALCPDIFF function of EasyStrata[296]. By contrast to other analyses where Bonferroni correction was considered a conservative approach for multiple testing, here the approach is not conservative as the null hypothesis (of no heterogeneity in genetic effect) might be inappropriately accepted by the strict Bonferroni correction. A less stringent approach was therefore chosen and correction for multiple testing used a Benjamini-Hochberg false discovery rate (FDR) adjustment[309] with FDR set to 0.05.

To evaluate whether tooth loss due to periodontitis led to biased estimates of genetic effect on DMFS, GWAS for DMFS was performed separately in periodontitis cases and controls for studies in GLIDE, then combined in two parallel meta-analyses. For lead associated variants, estimates of genetic effect on DMFS were then compared in these two stratified meta-analysis using the CALCPDIFF function of EasyStrata[296]. P values for tests for heterogeneity in genetic effect were corrected using Benjamini-Hochberg false discovery rate (FDR) of 0.05 for the same motivation as described above.

3.3 Results:

3.3.1: Study population

In GLIDE the single-trait meta-analyses included between 7 and 9 studies, and between 26,533 and 45,563 participants (Table 3.3). In UKB all GWAS included 461,031 participants with a variable number of cases (dentures: n=77,714 cases, bleeding gums: n=60,210 cases, loose teeth: n=18,979 cases, toothache: n=18,959 cases, painful gums: n=13,311 cases, ulcers: n=47,091 cases).

Table 3.3: Final sample included in single-trait GLIDE meta-analysis

Trait name	DMFS	DFSS	Nteeth	Periodontitis
Number of studies in primary meta-analysis	9	8	9	7
Number of participants in primary meta- analysis	26,792	26,533	27,949	17,353 cases, 28,210 controls

3.3.2: Aggregate single-variant results and trait selection for multi-trait analysis

In GLIDE, there was detectable polygenic association signal for all 4 traits. The highest heritability estimate was for Nteeth (0.13, SE=0.02), followed by DMFS, DFSS and then periodontitis, where the heritability estimate was 0.01 (SE 0.01). In UKB heritability estimates were estimated with greater precision than in GLIDE, and there was detectable evidence for polygenic heritability for all proxy phenotypes. Heritability estimates ranged between 0.09 and 0.04, with the highest estimate for dentures and the lowest estimate for toothache (Table 3.4).

Table 3.4: Single-trait heritability estimates in GLIDE and UKB

Resource	Trait	h^2_{LDSR}	SE
GLIDE	DMFS	0.090	0.018
	DFSS	0.057	0.017
	Nteeth	0.13	0.019
	Periodontitis	0.0097	0.011
UKB	Ulcers	0.082	0.0088
	Toothache	0.044	0.0072
	Bleeding gums	0.049	0.0033
	Painful gums	0.058	0.0098
	Dentures	0.094	0.0041
	Loose teeth	0.081	0.0091

Within GLIDE there were strong genetic correlations between DMFS and DFSS, DMFS and Nteeth and DFSS and Nteeth (Table 3.5). These genetic correlations might reflect similarity in the underlying genetic determinants of these traits, for example DMFS and DFSS are both considered measures of dental caries. Because number of teeth acts as numerator in DMFS and denominator in DFSS, these correlations might partially relate to the measurement phenomena discussed in chapter 2. For periodontitis the r_g values with DMFS, DFSS and Nteeth were imprecisely estimated, but suggested non-zero genetic correlation.

Table 3.5: Estimated genetic correlations between different clinical measures of dental disease experience in GLIDE

Trait 1	Trait 2	R_g	SE	P
DMFS	DFSS	1.14	0.13	$2.0x10^{-19}$
	Nteeth	-0.46	0.10	$4.7x10^{-6}$
	Periodontitis	-0.25	0.37	0.50
DFSS	Nteeth	-0.63	0.16	$1.1x10^{-4}$
	Periodontitis	0.60	0.60	0.32
Nteeth	Periodontitis	0.21	0.37	0.58

In UKB each self-reported trait showed some degree of similarity with at least one other trait, with the highest estimated genetic correlation between painful gums and toothache. As with the correlations seen within GLIDE, the correlations within UKB might arise because of biological processes which are relevant to multiple facets of dental health. It is also possible

that measurement phenomena contributed, for example if the questions themselves are non-specific or if a symptom can arise from more than one disease. Despite these reasons for similarity, there was also evidence that these measures were non-redundant and captured some distinct processes, for example dentures had little genetic similarity to bleeding gums, with an r_g estimate near zero (Table 3.6).

Table 3.6: Estimated genetic correlations correlations between different self-reported dental disease proxy traits in UKB.

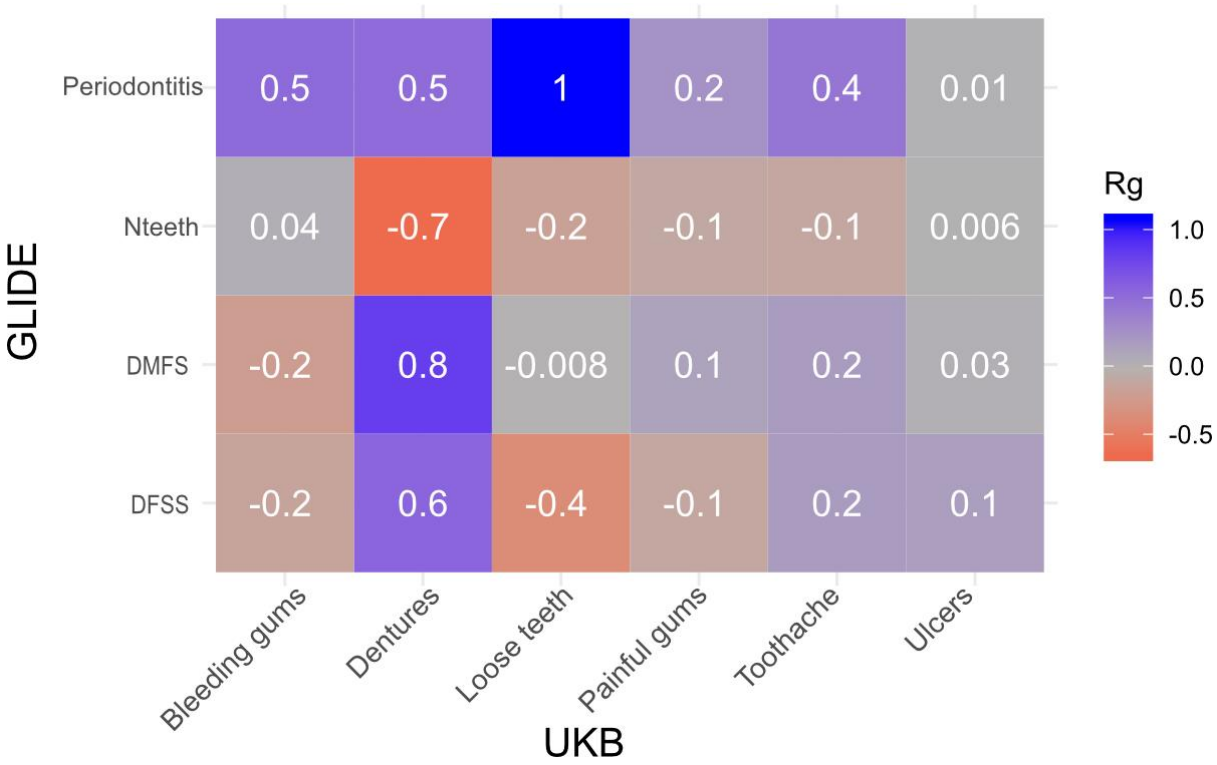
Trait 1	Trait 2	R_g	SE	P
Bleeding gums	Dentures	0.0009	0.035	0.98
	Loose teeth	0.37	0.055	$2.8x10^{-11}$
	Painful gums	0.50	0.069	$2.0x10^{-13}$
	Toothache	0.38	0.072	$1.6x10^{-7}$
	Ulcers	0.20	0.042	$1.3x10^{-6}$
Dentures	Loose teeth	0.46	0.043	$9.6x10^{-27}$
	Painful gums	0.14	0.054	$7.3x10^{-3}$
	Toothache	0.22	0.059	$2.4x10^{-4}$
	Ulcers	-0.082	0.036	0.0024
Loose teeth	Painful gums	0.46	0.088	$1.3x10^{-7}$
	Toothache	0.36	0.089	$6.5x10^{-5}$
	Ulcers	0.071	0.052	0.17
Painful gums	Toothache	0.87	0.11	$3.2x10^{-15}$
	Ulcers	0.59	0.066	$5.9x10^{-19}$
Toothache	Ulcers	0.43	0.065	$3.4x10^{-11}$

Comparing clinical traits to self-reported proxies, the most similar proxy to DMFS was dentures, with an estimated genetic correlation coefficient of 0.82 (SE=0.087) and strong evidence against the null hypothesis of no genetic correlation ($P=4.1x10^{-21}$). For both DFSS and Nteeth the best genetic proxy phenotype was also dentures, but with a weaker extent of genetic correlation than that seen for DMFS (Table 3.7 and Figure 3.1). For periodontitis, the single best genetic proxy was loose teeth, where the genetic correlation was estimated to be near 1 but was imprecisely estimated (SE=0.78) and did not provide evidence against the null hypothesis of no genetic correlation (P=0.17).

Table 3.7: Estimated genetic correlation between single-trait results in GLIDE and UKB

Trait 1	Trait 2	R_g	SE	P
DMFS	Ulcers	0.027	0.073	0.38
	Toothache	0.17	0.12	0.16
	Bleeding gums	-0.22	0.073	0.0029
	Painful gums	0.11	0.13	0.87
	Dentures	0.82	0.087	$4.1x10^{-21}$
	Loose teeth	-0.008	0.090	0.93
DFSS	Ulcers	0.15	0.090	0.097
	Toothache	0.17	0.15	0.26
	Bleeding gums	-0.15	0.099	0.13
	Painful gums	-0.12	0.16	0.44
	Dentures	0.56	0.12	$1.7x10^{-6}$
	Loose teeth	-0.37	0.13	0.0045
Nteeth	Ulcers	0.0057	0.060	0.92
	Toothache	-0.12	0.10	0.26
	Bleeding gums	0.036	0.071	0.51
	Painful gums	-0.11	0.11	0.34
	Dentures	-0.65	0.056	$3.9x10^{-31}$
	Loose teeth	0.061	0.10	0.59
Periodontitis	Ulcers	0.011	0.19	0.95
	Toothache	0.44	0.48	0.93
	Bleeding gums	0.52	0.43	0.23
	Painful gums	0.24	0.39	0.54
	Dentures	0.51	0.42	0.23
	Loose teeth	1.07	0.78	0.17

Figure 3.1: Heatmap showing estimated genetic correlation values between traits in GLIDE and UKB



Cells are shaded by the estimated value of R_g

These genetic correlation estimates were used to select non-redundant pairs of clinical and self-reported traits for combined meta-analysis, allowing each clinical and self-reported trait to appear only once in combined analysis. DMFS and DFSS appeared to capture similar genetic association signals, but heritability was higher for DMFS, so DFSS was not taken forward to combined analysis. DMFS and dentures had similar heritability estimates and a strong genetic correlation, so were selected as the first pair. Nteeth did not have any strong genetic proxies apart from Dentures, which had already been chosen for DMFS, so was eliminated. For periodontitis all genetic correlations were imprecisely-estimated however the point estimate suggested loose teeth as the best proxy phenotype. Given the biological plausibility that untreated periodontitis may present with a symptom of loose teeth, the decision was made to include periodontitis and loose teeth as the second trait pair in multi-trait analysis.

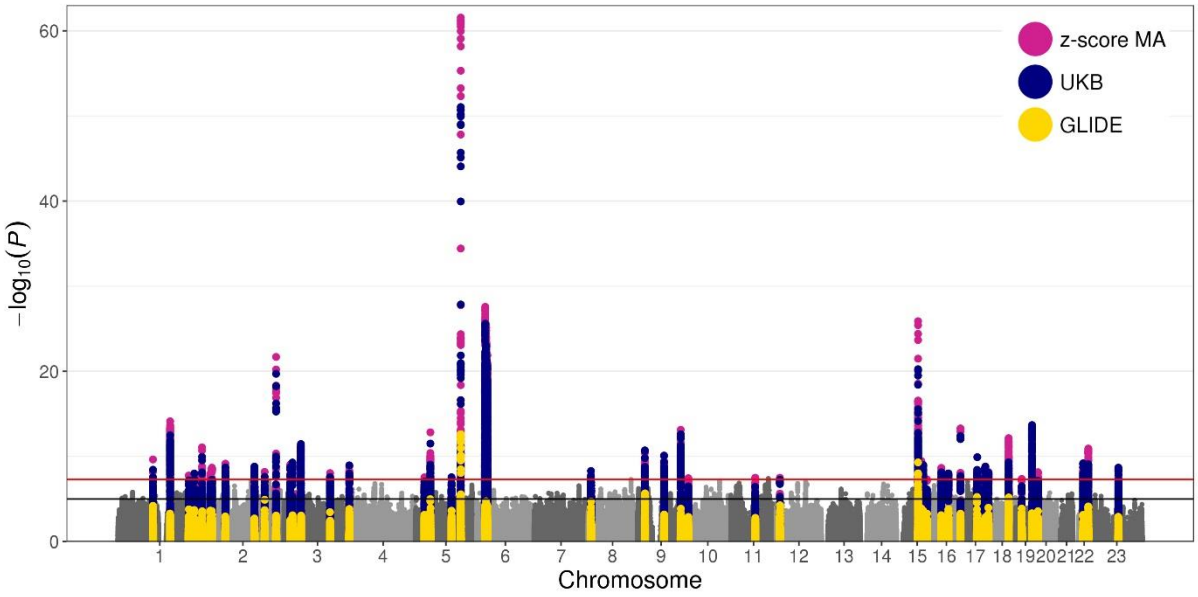
3.3.3: Single variant results in multi-trait meta-analysis.

Combined meta-analysis of DMFS and dentures (hereafter 'DMFS/dentures') included approximately 488,000 participants (n=26,792 from 9 studies for DMFS, n_{cases}=77,714, $n_{controls}$=383,317 for dentures from one study). Combined meta-analysis of periodontitis and loose teeth (hereafter 'periodontitis/loose teeth') included a little over 500,000 participants with similar number of cases in GLIDE and UKB and a greater number of controls in UKB (n_{cases}=17,353, $n_{controls}$=28,210 from 7 studies for periodontitis, n_{cases}=18,979, $n_{controls}$=442,052 for loose teeth).

After final quality control both multi-trait meta-analyses included around 8.9 million SNPs and insertion/deletions (INDELS) with MAF of 0.01 or greater which were present in both GLIDE and UKB. Both multi-trait meta-analyses had inflated values of λ_{GC}, which was attributed to polygenic association signal rather than inflationary bias in univariate LDSR (λ_{GC}=1.37, h^2_{LDSR}=0.085, LDSR intercept $LDSR\_i$=1.00 for DMFS/dentures; λ_{GC}=1.09, h^2_{LDSR}=0.046, $LDSR\_i$=1.00 for periodontal disease/loose teeth.

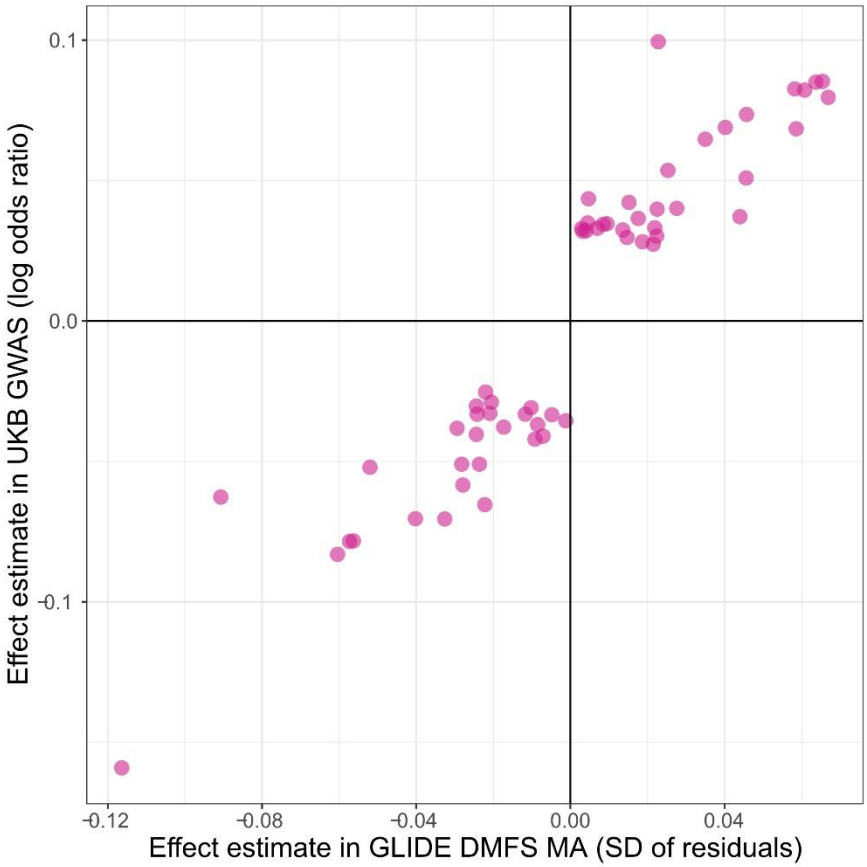
For DMFS/dentures 47 genomic risk loci were identified (Table 3.8a and 3.8b, Fig. 3.2). Three of these risk loci housed more than 1 conditionally-independent signals of association (Table 3.9). There was good concordance in genetic effects in GLIDE and UKB, with all 47 lead variants having directionally-consistent effect direction and the magnitude of genetic effects was closely correlated (Pearson correlation coefficient=0.94) (Figure 3.3). Some of these loci have previously been investigated in candidate gene association studies for dental caries, including the HLA region[310], but none have previously been identified in other GWAS for dental diseases and are therefore considered novel genetic associations.

Figure 3.2: Manhattan plot for DMFS/dentures combined analysis



P-values for all variants in DMFS/dentures meta-analysis are plotted in grey. At loci meeting genome-wide significance, P values for variants within 500KB of the lead associated variant are potted in magenta (DMFS/dentures MA), yellow (GLIDE) and blue (UKB). The red line is drawn at $P=5x10^{-8}$ and the black line is drawn at $P=1x10^{-5}$

Figure 3.3: Plot of concordance of genetic effect estimates in GLIDE and UKB



The lead variant for each novel risk locus in the DMFS/dentures combined meta-analysis is plotted as a single dot. The position on the X axis represents the effect estimate in the GLIDE meta-analysis on a linear scale, while the position on the Y axis represents the effect estimate in the UKB dentures GWAS on a log-odds ratio scale.

Table 3.8a: Genomic risk loci passing genome-wide significance in DMFS/dentures combined meta-analysis (chr 1 – 9)

Locus	Lead variant	Chr:Pos (b37)	EA:EAF	DMFS/dentures		DMFS		Dentures		Ntotal
				Beta (SE)$_1$	P	Beta (SE)	P	Beta (SE)	P	
AMY1C	rs72694438	1:104364878	a:0.21	0.021(0.0033)	2.25E^{-10}	0.028(0.012)	0.018	0.0401(0.0068)	3.66E^{-09}	487822
KRTCAP2	rs4971099	1:155155608	a:0.55	-0.021(0.0027)	7.47E^{-15}	-0.024(0.0089)	0.0060	-0.040(0.0055)	3.24E^{-13}	487821
SYT14	rs2046850	1:210304319	t:0.19	-0.019(0.0033)	1.77E^{-08}	-0.017(0.011)	0.11	-0.038(0.007)	6.62E^{-08}	487821
ITPKB	rs3820640	1:226868918	t:0.84	0.020(0.0037)	2.65E^{-08}	0.0047(0.013)	0.72	0.044(0.0076)	1.01E^{-08}	487822
FAM150B	rs62106258	2:417167	t:0.95	0.042(0.0062)	8.60E^{-12}	0.058(0.026)	0.025	0.083(0.0128)	1.14E^{-10}	487822
ADCY3	rs11676272	2:25141538	a:0.52	-0.015(0.0026)	1.07E^{-08}	-0.021(0.0087)	0.019	-0.029(0.0055)	1.50E^{-07}	487823
ALK	rs80270335	2:29616655	t:0.09	0.027(0.0045)	2.10E^{-09}	0.046(0.017)	0.007	0.051(0.0094)	5.62E^{-08}	487821
AAK1	rs5831974	2:69704336	d:0.45	-0.016(0.0027)	6.97E^{-10}	-0.014(0.0092)	0.14	-0.033(0.0055)	1.93E^{-09}	486706
KCNJ3	rs2652452	2:155670203	a:0.45	-0.016(0.0027)	3.27E^{-09}	-0.0048(0.0087)	0.58	-0.033(0.0055)	1.55E^{-09}	487823
ZNF804A	rs263771	2:185921692	a:0.23	0.018(0.0031)	6.53E^{-09}	0.018(0.011)	0.11	0.037(0.0065)	2.46E^{-08}	487822
WNT10A	rs121908120	2:219755011	a:0.03	-0.081(0.0083)	2.03E^{-22}	-0.12(0.039)	0.0026	-0.16(0.017)	1.94E^{-20}	486339
KCNH8	rs9831002	3:18852697	t:0.49	-0.016(0.0026)	8.79E^{-10}	-0.012(0.0088)	0.19	-0.033(0.0055)	1.81E^{-09}	487822
RARB	rs7429279	3:25118637	a:0.41	0.016(0.0027)	1.28E^{-09}	0.0045(0.009)	0.61	0.035(0.0056)	5.29E^{-10}	487822
RBM5	3:50135699DI	3:50135699	d:0.48	-0.018(0.0027)	4.79E^{-12}	-0.011(0.009)	0.23	-0.038(0.0055)	6.82E^{-12}	486705
STAG1	rs61790808	3:136443008	a:0.64	-0.016(0.0028)	9.10E^{-09}	-0.024(0.011)	0.02	-0.03(0.0057)	1.05E^{-07}	487820
OPA1	rs185566659	3:193394725	a:0.032	0.045(0.0075)	1.54E^{-09}	0.023(0.029)	0.43	0.099(0.016)	1.10E^{-09}	487822
CDH9	rs55769264	5:26928047	a:0.45	0.015(0.0027)	2.84E^{-08}	0.015(0.0096)	0.12	0.030(0.0056)	9.45E^{-08}	487822
FGF10	rs1482698	5:44539453	c:0.38	0.020(0.0027)	1.47E^{-13}	0.023(0.0092)	0.014	0.040(0.0057)	3.10E^{-12}	487823
EFNA5	rs1352724	5:107083487	a:0.22	-0.018(0.0032)	3.64E^{-08}	-0.0085(0.012)	0.48	-0.037(0.0066)	2.76E^{-08}	487823
C5orf66	rs1122171	5:134509987	t:0.59	0.044(0.0027)	2.84E^{-62}	0.064(0.0087)	2.4E^{-13}	0.085(0.0056)	8.96E^{-52}	487823
HLA	rs9366651	6:26336696	t:0.51	-0.029(0.0026)	2.66E^{-28}	-0.028(0.0088)	0.0014	-0.058(0.0055)	4.47E^{-26}	487822
LOC157273	rs898797	8:9229689	t:0.59	0.015(0.0027)	1.52E^{-08}	0.0029(0.009)	0.75	0.033(0.0056)	4.98E^{-09}	487822
DMRTA1	rs10811723	9:22542285	a:0.30	-0.019(0.0029)	3.41E^{-11}	-0.0071(0.0091)	0.43	-0.041(0.0061)	1.94E^{-11}	487821
PRUNE2	rs7852129	9:79346204	a:0.89	-0.025(0.0038)	7.91E^{-11}	-0.028(0.030)	0.34	-0.051(0.0079)	8.38E^{-11}	483297
PBX3	rs10987008	9:128661600	a:0.64	0.021(0.0028)	7.47E^{-14}	0.015(0.009)	0.093	0.042(0.0058)	2.52E^{-13}	487822

Locus – the name of the protein coding gene in the RefSeq database lying nearest to the lead variant; EA – effect allele; EAF – effect allele frequency; $_1$estimates from standardized regression coefficients

Table 3.8b: Genomic risk loci passing genome-wide significance in DMFS/dentures combined meta-analysis (chr 10 – 23)

Locus	Lead variant	Chr:Pos (b37)	EA:EAF	DMFS/dentures		DMFS		Dentures		Ntotal
				Beta (SE)[1]	P	Beta (SE)	P	Beta (SE)	P	
STFA1P	rs7918807	10:10020194	t:0.52	0.015(0.0027)	$3.58E^{-08}$	0.019(0.0087)	0.032	0.028(0.0055)	$3.36E^{-07}$	487821
P2RY2	rs149467613	11:72943483	a:0.05	-0.034(0.0062)	$3.21E^{-08}$	-0.091(0.029)	0.0016	-0.063(0.013)	$1.64E^{-06}$	487823
KLRAP1	rs10772314	12:10704350	a:0.40	-0.015(0.0027)	$3.16E^{-08}$	-0.010(0.0088)	0.25	-0.031(0.0057)	$5.15E^{-08}$	487820
CA12	rs72748935	15:63639416	t:0.46	-0.028(0.0027)	$1.31E^{-26}$	-0.052(0.0092)	$1.6E^{-08}$	-0.052(0.0055)	$5.49E^{-21}$	487820
NEO1	rs6495046	15:73353175	c:0.36	-0.017(0.0028)	$3.47E^{-10}$	-0.024(0.0093)	0.0091	-0.033(0.0058)	$8.70E^{-09}$	487821
CHRNA3	rs10851907	15:78915864	a:0.41	0.016(0.0027)	$1.03E^{-09}$	0.0085(0.0092)	0.36	0.034(0.0056)	$9.88E^{-10}$	486706
RHCG	rs2072693	15:90014945	t:0.48	0.014(0.0026)	$4.92E^{-08}$	0.0215(0.009)	0.015	0.027(0.0055)	$7.53E^{-07}$	487821
TMEM219	rs8054556	16:29958216	a:0.46	0.016(0.0027)	$2.23E^{-09}$	0.0224(0.0089)	0.012	0.030(0.0055)	$4.38E^{-08}$	487821
SALL1	rs1108343	16:51211595	t:0.36	0.016(0.0028)	$1.32E^{-08}$	0.007(0.0095)	0.46	0.033(0.0058)	$9.75E^{-09}$	487819
FOXL1	rs10048146	16:86710660	a:0.81	-0.026(0.0034)	$5.20E^{-14}$	-0.0236(0.012)	0.045	-0.051(0.007)	$3.85E^{-13}$	487820
NPEPPS	rs3865314	17:45669524	a:0.51	0.015(0.0026)	$1.48E^{-08}$	0.0041(0.009)	0.64	0.032(0.0055)	$6.63E^{-09}$	487822
HOXB-AS2	rs9905793	17:46635649	a:0.091	0.027(0.0046)	$6.51E^{-09}$	0.0253(0.013)	0.059	0.054(0.0098)	$3.97E^{-08}$	487820
KCNJ2	rs34559440	17:68399112	t:0.68	-0.016(0.0028)	$1.14E^{-08}$	-0.0012(0.0093)	0.89	-0.036(0.006)	$2.60E^{-09}$	487822
LOC100499467	rs7217268	17:70338127	a:0.38	0.016(0.0027)	$1.48E^{-09}$	0.0095(0.0095)	0.32	0.0346(0.0057)	$1.66E^{-09}$	487821
BAHCC1	rs57067187	17:79361332	t:0.63	0.015(0.0027)	$6.90E^{-09}$	0.0136(0.011)	0.20	0.032(0.0057)	$1.44E^{-08}$	487821
MC4R	rs28822480	18:57924823	a:0.29	0.021(0.0029)	$7.08E^{-13}$	0.044(0.010)	$1.4E^{-05}$	0.037(0.006)	$8.38E^{-10}$	487821
CRLF1	rs2238651	19:18718846	t:0.24	0.017(0.0031)	$4.39E^{-08}$	0.0219(0.011)	0.047	0.033(0.0065)	$3.14E^{-07}$	487821
MAMSTR	rs11672900	19:49220323	a:0.47	-0.020(0.0027)	$4.67E^{-14}$	-0.0092(0.009)	0.31	-0.042(0.0055)	$3.11E^{-14}$	487822
HAO1	rs4816017	20:7654373	a:0.29	-0.017(0.0029)	$7.08E^{-09}$	-0.0209(0.009)	0.026	-0.0329(0.0061)	$8.11E^{-08}$	487820
MTMR3	rs140357883	22:30292811	d:0.84	0.022(0.0036)	$2.33E^{-09}$	0.0248(0.013)	0.055	0.042(0.0075)	$1.51E^{-08}$	486706
FAM118A	rs1569414	22:45727565	t:0.73	-0.020(0.0030)	$1.19E^{-11}$	-0.0294(0.0093)	0.0017	-0.038(0.0063)	$1.00E^{-09}$	487819
HDX	rs5922945	23:83523015	t:0.34	-0.016(0.0028)	$1.55E^{-08}$	-0.022(0.0094)	0.019	-0.025(0.0048)	$1.65E^{-07}$	478256

Locus – the name of the protein coding gene in the RefSeq database lying nearest to the lead variant; EA – effect allele; EAF – effect allele frequency; [1]estimates from standardized regression coefficient.

Table 3.9: Genomic risk loci harboring more than one signal of association

Locus	Lead tag variant			Other variant		
	rsid	Unconditional P value	Conditional P value	rsid	Unconditional P value	Conditional P value
FAM150B	rs62106258	$8.60x10^{-12}$	$3.40x10^{-12}$	rs13028737	$2.18x10^{-8}$	$8.53x10^{-9}$
ALK	rs80270335	$2.10x10^{-9}$	$3.50Ex10^{-11}$	rs4128318	$4.93x10^{-6}$	$4.85x10^{-8}$
CA12	rs72748935	$1.3x10^{-26}$	$2.6x10^{-25}$	rs7180729	$8.00x10^{-10}$	$1.91x10^{-8}$

As it is not possible in the available space to present detailed results for each locus, two variants have been chosen, reflecting the variant with the largest per-allele estimated genetic effect in the DMFS/dentures meta-analysis and the variant with the strongest statistical evidence for association in the DMFS/dentures meta-analysis.

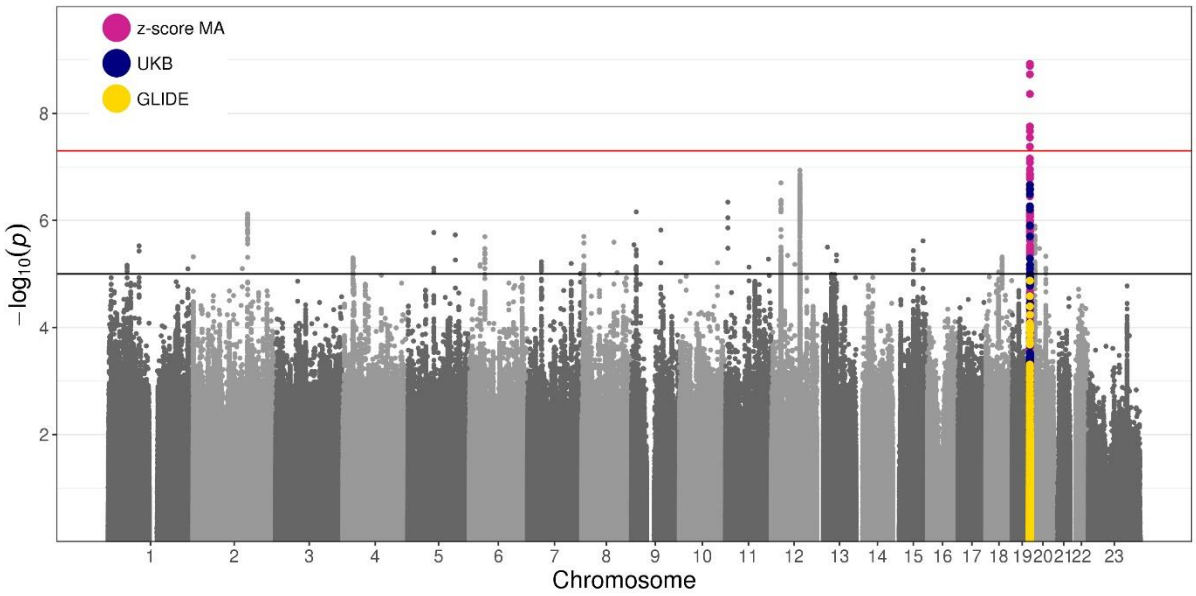
The largest estimated effect in combined meta-analysis (standardized beta=-0.081, $P=2.0x10^{-22}$) was at rs121908120, a low-frequency missense variant within *WNT10A* with EAF=0.026 for A allele. In single-trait analysis rs121908120 also had the largest effects on DMFS (beta= -0.12 standard deviation (SD) change in DMFS residuals; 2.1 [95% CI 1.7, 2.7] fewer decayed, missing or filled tooth surfaces at age 50 years) and on odds of having dentures (OR=0.85; 95% CI 0.82,0.88). rs121908120 results in a phenylalanine to isoleucine substitution and is predicted to have deleterious consequences in multiple *WNT10A* transcripts using the ExAC browser[311].

The strongest statistical evidence for association was seen at rs1122171 ($P=2.8x10^{-62}$), a common variant (EAF=0.59 for T allele) with modest effects in combined analysis (standardized beta=0.044) and single-trait analysis (DMFS beta=0.064, corresponding to 1.2 [95% CI 1.0, 1.4] additional decayed, missing or filled tooth surfaces at age 50 years, OR dentures = 1.09 [95% CI 1.08,1.10]). Based on position only, this variant maps to an uncharacterized protein-coding region, *C5orf66.* Including eQTL and chromatin interaction data in FUMA, this variant was mapped to several possible genes, so the decision was made to prioritize these systematically using S-TisssueXcan results presented below.

For periodontitis/loose teeth a single locus met the pre-defined criteria for genome wide association (Figure 3.4). There was evidence for association at 19q13, where the lead signal was rs12461706 ($P=3.9x10^{-9}$), a common intronic variant within *SIGLEC5,* (EAF=0.40 for T

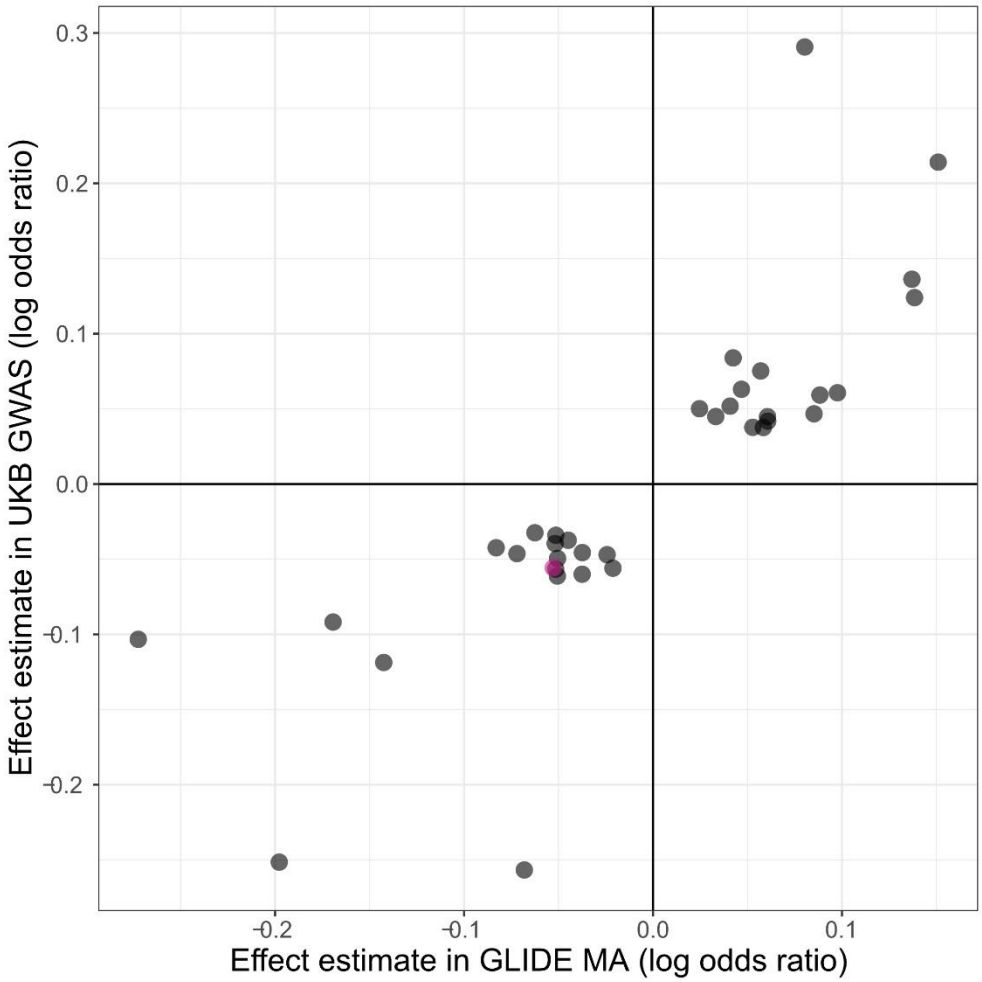
allele) with modest but consistent effect on odds of periodontal disease (OR=1.05) and loose teeth (OR=1.06). *SIGLEC5* was recently reported as a risk locus for aggressive periodontitis[312] so this association is considered a positive control rather than novel discovery. There was good concordance in effect direction and genetic effect estimates are at conditionally-independent single variants passing an arbitrary threshold for suggestive association ($p<1x10^{-5}$)(Figure 3.5).

Figure 3.4: Manhattan plot for periodontitis/loose teeth combined analysis



P values for all variants in periodontitis/loose teeth meta-analysis are plotted in grey. At the 19q13 locus, P values for variants within 500KB of the lead associated variant (previously reported for aggressive periodontitis) are potted in magenta (DMFS/dentures MA), yellow (GLIDE) and blue (UKB). The red line is drawn at $P=5x10^{-8}$ and the black line is drawn at $P=1x10^{-5}$.

Figure 3.5: Concordance of genetic effect estimates in periodontitis/loose teeth combined analysis



Grey dots represent conditionally-independent lead variants with $p<1x10^{-5}$ in combined periodontitis/loose teeth meta-analysis, while the magenta dot represents the lead variant at the 19q13 locus (previously reported for aggressive periodontitis). Effect estimates for these variants in the GLIDE meta-analysis are plotted on the X axis, while effect estimates in the UKB GWAS are plotted on the y axis. Both axes are on a log-odds ratio scale.

3.3.4: Mapping of association signal in HLA region

To help dissect the pattern of single-variant association in the HLA region seen in the DMFS/dentures analysis, imputed HLA haplotypes were examined. There were 336,038 unrelated participants of white British ancestry with non-missing haplotype and phenotypic data. Using each haplotype as an exposure variable in logistic regression models, 10 haplotypes were associated with dentures at a Bonferroni corrected alpha value of 0.05. Associated haplotypes included both HLA class I and HLA class II haplotypes, suggesting

more than one independent signal of association in the HLA region (Table 3.10). Of the associated haplotypes, the strongest evidence for association was for DQB1_201, a common haplotype encoding the DQ-beta 1 chain of the HLA class II complex (haplotype frequency = 0.15, OR=1.07 (95% CI 1.05, 1.09), $P=8.9x10^{-13}$). HLA class II molecules are expressed by antigen presenting cells and alleles of HLA class II are thought to modulate the composition of the oral microbiome, including the cariogenic gram-positive organism *Streptococcus mutans*[313].

Table 3.10: Association between HLA haplotypes and odds of dentures in UKB

Haplotype	Beta	SE	p	Odds Ratio (95% CI)	Haplotype frequency
DRB3_101	0.058	0.009	$1.63x10^{-10}$	1.06 (1.04, 1.08)	0.17
DRB1_301	0.067	0.009	$1.24x10^{-12}$	1.07 (1.05, 1.09)	0.15
DRB1_101	-0.059	0.012	$9.01x10^{-07}$	0.94 (0.92, 0.97)	0.09
DQB1_501	-0.047	0.011	$8.49x10^{-06}$	0.95 (0.93, 0.97)	0.12
DQB1_201	0.068	0.009	$8.87x10^{-13}$	1.07 (1.05, 1.09)	0.15
DQA1_501	0.037	0.008	$6.87x10^{-06}$	1.04 (1.02, 1.05)	0.23
C_701	0.046	0.009	$2.37x10^{-07}$	1.05 (1.03, 1.07)	0.18
B_801	0.061	0.010	$1.72x10^{-10}$	1.06 (1.04, 1.08)	0.14
B_1501	-0.054	0.014	$1.60x10^{-04}$	0.95 (0.92, 0.97)	0.06
A_101	0.058	0.009	$1.04x10^{-11}$	1.06 (1.04, 1.08)	0.19

N=336,038 for all models. Logistic regression incorporated adjustment for age, sex, genotyping array and 40 genetic principal components.

3.3.5: Tests for enrichment in functional categories or tissue-specific annotations

In the DMFS/dentures combined analysis 85 functional annotations were tested for enrichment, including both continuous annotations and binary annotations. Of these, there was evidence for enrichment passing a Bonferroni-corrected P value threshold in 38 annotations. Some of these related to evolutionary conservation – for example DMFS/dentures association signal was enriched for genomic regions with higher Genomic Evolutionary Rate Profiling (GERP) scores ($P=1.5x10^{-28}$). Likewise, variants annotated as highly conserved from primates to humans accounted for only 1.9% of variants included in analysis, yet these variants conferred 30% of the heritability of DMFS/dentures: a 15.5-fold enrichment over baseline which was unlikely to be due to chance ($P=3.0x10^{-15}$; Appendix 3.1).

Examining enrichment of DMFS/dentures association signal in regions of the genome with tissue-specific expression patterns, the highest stratified LDSR coefficients were seen for

gastrointestinal tract (coefficient = $3.4x10^{-9}$, SE=$1.6x10^{-9}$) and minor salivary gland (coefficient = $3.1x10^{-9}$, SE= $1.5x10^{-9}$). Given the wide confidence intervals for these stratified regression coefficients, it was not possible to test for differences in coefficient between tissue types (Appendix 3.2).

In the Periodontitis/loose teeth analysis no functional annotations were considered enriched beyond chance, possibly reflecting reduced statistical power for this analysis compared to DMFS/dentures results (Appendix 3.3). Likewise, tissue-specific regression coefficients were imprecisely-estimated and could not be used to make meaningful comparisons between different tissue types (Appendix 3.4)

3.3.6: Transcript-based tests and gene-set sets

Results of tests for association between predicted gene transcription and DMFS/dentures were available for 15,522 transcripts. There was evidence passing a multiple testing correction that varying transcription of 221 transcripts had effects on DMFS/dentures, including multiple transcripts within the HLA region. Overall, the lead transcript was *ZSCAN9*, a gene within the HLA region which encodes a zinc finger protein, where increasing expression was predicted to associate with increased burden of DMFS/dentures ($P=3.7x10^{-35}$)

Outside the HLA region, the strongest evidence for transcript level association with DMFS/dentures was seen for *CA12*, located in the region of single-variant association signal at 15q22, ($P=2.8x10^{-17}$). *CA12* encodes a member of the carbonic anhydrase family of zinc enzymes which catalyze the hydration of carbon dioxide to form bicarbonate and hydrogen ions to regulate pH. This family has several functions relevant to dental caries, including tooth formation, where multiple carbonic anhydrases are produced by maturation-stage ameloblasts[314], salivary buffering, where defects in *CA12* lead to poor salivary function and xerostomia[315], and regulation of tooth biofilm microbiota, where *CA6* may affect colonization by the cariogenic microorganism *Streptococcus mutans*[100].

Using transcript-level results to characterize the 5q31 region which housed the lead single variant, the most relevant transcript was predicted to be *PITX1,* where increasing transcription was associated with decreased burden of DMFS/dentures, but with a tissue-specific pattern of association ($P=2.9x10^{-12}$). *PITX1* encodes a developmentally expressed

transcription factor with roles in skeletal and mandibular growth and tooth development, and deletion of the *Pitx1* locus in animal models results in abnormal mandibular tooth morphology[316]. Results passing a multiple testing threshold are given in Appendix 3.5.

Increasing transcription of *SIGLEC5* was predicted to associate with increasing odds of periodontitis/ loose teeth ($p=8.7x10^{-07}$, Appendix 3.6). No other results passed a multiple testing correction.

Tests for gene-set enrichment using DEPICT did not identify any gene sets at an acceptable false discovery rate for DMFS/dentures. Similarly, the hypergeometric test did not identify enrichment of DMFS/dentures association signal in any of the Molecular Signature Database (MSigDB C2) curated gene sets which passed a correction for multiple testing.

For periodontitis/loose teeth there was only a single associated locus so tests for gene set enrichment were not performed.

3.3.7: Comparison with other traits and diseases

At a genome-wide level, results of cross-trait comparison were available for 234 results files from published GWAS studies. For DMFS/dentures there was evidence against the null hypothesis of no genetic correlation for 43 results files, representing a smaller number of unique traits, as the same trait could be analysed by several different GWAS studies and the same study might share results in more than one file. There were positive genetic correlations between DMFS/dentures and traits related to smoking including "ever vs. never smoked" (r_g = 0.38, SE=0.042, $P=1.8x10^{-19}$) and diseases which are associated with smoking severity such as lung cancer (r_g = 0.36, SE=0.05, $P= 8.7x10^{-13}$) and coronary artery disease (r_g = 0.19, SE=0.03, $P=2.1x10^{-10}$). There were positive genetic correlations between DMFS/dentures and measures of adiposity including body mass index (BMI, r_g = 0.21, SE=0.027, $P=3.2x10^{-15}$). There were negative genetic correlations between DMFS/dentures and proxy traits capturing longevity (r_g=-0.47, SE=0.064, $P=2.0x10^{-13}$ for father's age at death) and complex traits reflecting socio-demographics and cognition such as educational attainment (r_g=-0.52, SE=0.019, $P=1.8x10^{-163}$ for years of schooling) (Table 3.11). Collectively these patterns mirror the observational associations reported in chapter 2 and associations between dental diseases and cardio-metabolic traits which are reported in the literature and discussed further in chapter 4.

Table 3.11: Genetic correlations between DMFS/dentures and results files in the LD-hub resource.

Trait	PMID[1]	r_g	SE	P
Years of schooling 2016	27225129	-0.5236	0.0192	$1.77x10^{-163}$
Age of first birth	27798627	-0.5026	0.0303	$5.61x10^{-62}$
Years of schooling (proxy cognitive performance)	25201988	-0.5469	0.035	$3.80x10^{-55}$
Years of schooling 2013	23722424	-0.5492	0.0364	$2.09x10^{-51}$
College completion	23722424	-0.5249	0.0382	$6.49x10^{-43}$
Intelligence	28530673	-0.3281	0.0301	$1.41x10^{-27}$
Number of children ever born	27798627	0.3547	0.0347	$1.59x10^{-24}$
Ever vs never smoked	20418890	0.3777	0.0418	$1.79x10^{-19}$
Waist circumference	25673412	0.2321	0.026	$3.93x10^{-19}$
Waist-to-hip ratio	25673412	0.2501	0.0283	$1.10x10^{-18}$
Obesity class 1	23563607	0.2457	0.0287	$1.10x10^{-17}$
Overweight	23563607	0.2399	0.03	$1.32x10^{-15}$
Body mass index	20935630	0.2118	0.0269	$3.24x10^{-15}$
Former vs Current smoker	20418890	-0.5195	0.0687	$3.92x10^{-14}$
Father's age at death	27015805	-0.4665	0.0635	$2.01x10^{-13}$
Lung cancer	27488534	0.3573	0.05	$8.69x10^{-13}$
Mother's age at death	27015805	-0.4926	0.0696	$1.50x10^{-12}$
Body fat	26833246	0.2804	0.04	$2.28x10^{-12}$
Obesity class 2	23563607	0.2549	0.0377	$1.41x10^{-11}$
Coronary artery disease	26343387	0.1939	0.0305	$2.12x10^{-10}$
Age at Menopause	26414677	-0.2145	0.0344	$4.71x10^{-10}$
Parent's age at death	27015805	-0.4726	0.0773	$9.55x10^{-10}$
Childhood IQ	23358156	-0.3688	0.0606	$1.16x10^{-09}$
Hip circumference	25673412	0.1531	0.0254	$1.66x10^{-09}$
Lung cancer (all)	24880342	0.316	0.0541	$5.08x10^{-09}$
Extreme BMI	23563607	0.2444	0.043	$1.28x10^{-08}$
Squamous cell lung cancer	27488534	0.4583	0.0817	$2.02x10^{-08}$
Cigarettes smoked per day	20418890	0.3804	0.0706	$7.11x10^{-08}$
HDL cholesterol	20686565	-0.1854	0.0355	$1.83x10^{-07}$
Childhood obesity	22484627	0.2039	0.0429	$2.02x10^{-06}$
22:6 docosahexaenoic acid	27005778	-0.3402	0.0718	$2.18x10^{-06}$
Depressive symptoms	27089181	0.2163	0.0471	$4.38x10^{-06}$
Anorexia Nervosa	24514567	-0.146	0.0322	$5.97x10^{-06}$
Rheumatoid Arthritis	24390342	0.1751	0.0391	$7.54x10^{-06}$
Fasting glucose main effect	22581228	0.169	0.0381	$9.16x10^{-06}$
Bipolar disorder	21926972	-0.1584	0.036	$1.10x10^{-05}$
Obesity class 3	23563607	0.2058	0.0488	$2.44x10^{-05}$
Insomnia	28604731	0.197	0.0469	$2.72x10^{-05}$
Lung cancer (squamous cell)	24880342	0.431	0.1037	$3.26x10^{-05}$
Leptin_not_adjBMI	26833098	0.2047	0.0556	0.0002
HbA1C	20858683	0.2014	0.0549	0.0002
Attention deficit hyperactivity disorder (GC)	27663945	0.3914	0.1045	0.0002
Attention deficit hyperactivity disorder (No GC)	27663945	0.3913	0.1047	0.0002

Only results passing a correction for multiple testing are provided. [1]PMID refers to the PubMed identifier for the GWAS study which produced the results file.

For periodontitis/loose teeth genetic correlation was estimated with the same 234 results files, of which there was evidence for non-zero genetic correlation with 31. The pattern of correlation was generally similar to that for DMFS/dentures, with positive correlations for smoking and obesity-related traits, and negative correlations for longevity and educational attainment / sociodemographic traits (Table 3.12).

Table 3.12: Estimated genetic correlation between periodontitis/loose teeth and GWAS results files using the LD-hub resource

Trait	$_1$PMID	R_g	SE	P
Years of schooling 2016	27225129	-0.3718	0.0394	$4.00x10^{-21}$
Ever vs never smoked	20418890	0.6429	0.0708	$1.12x10^{-19}$
Age of first birth	27798627	-0.4435	0.0543	$3.06x10^{-16}$
Waist-to-hip ratio	25673412	0.3246	0.0448	$4.42x10^{-13}$
College completion	23722424	-0.35	0.0616	$1.35x10^{-08}$
Years of schooling 2013	23722424	-0.3659	0.0668	$4.30x10^{-08}$
Years of schooling (proxy cognitive performance)	25201988	-0.3409	0.0623	$4.45x10^{-08}$
Waist circumference	25673412	0.266	0.0496	$8.40x10^{-08}$
Body fat	26833246	0.3323	0.0629	$1.27x10^{-07}$
Overweight	23563607	0.2782	0.0541	$2.77x10^{-07}$
Depressive symptoms	27089181	0.3386	0.0659	$2.82x10^{-07}$
Former vs Current smoker	20418890	-0.5551	0.109	$3.52x10^{-07}$
Lung cancer	27488534	0.4318	0.0855	$4.47x10^{-07}$
Obesity class 1	23563607	0.2486	0.0504	$7.95x10^{-07}$
Intelligence	28530673	-0.2482	0.051	$1.11x10^{-06}$
Number of children ever born	27798627	0.2885	0.0615	$2.69x10^{-06}$
Body mass index	20935630	0.2474	0.0538	$4.23x10^{-06}$
Cigarettes smoked per day	20418890	0.4948	0.1087	$5.29x10^{-06}$
Lung cancer (all)	24880342	0.4343	0.0965	$6.80x10^{-06}$
Father's age at death	27015805	-0.4535	0.1021	$8.96x10^{-06}$
Insomnia	28604731	0.341	0.0788	$1.52x10^{-05}$
Obesity class 2	23563607	0.2861	0.0664	$1.66x10^{-05}$
Mother's age at death	27015805	-0.4892	0.1145	$1.94x10^{-05}$
Extreme waist-to-hip ratio	23563607	0.4211	0.0998	$2.47x10^{-05}$
Schizophrenia	25056061	0.1689	0.041	$3.71x10^{-05}$
Extreme BMI	23563607	0.2905	0.0729	$6.66x10^{-05}$
Squamous cell lung cancer	27488534	0.5594	0.1405	$6.86x10^{-05}$
Childhood obesity	22484627	0.2955	0.0761	0.0001
Parent's age at death	27015805	-0.4711	0.1228	0.0001
Subjective well being	27089181	-0.2561	0.0689	0.0002
Lung cancer (squamous cell)	24880342	0.6428	0.172	0.0002

Only results passing a multiple testing threshold are provided. $_1$PMID refers to PubMed identifier for the GWAS study.

Next, comparison was made at a single-variant level. For DMFS/dentures, association statistics for other traits were available for the lead variant or an acceptable proxy for the lead variant for 46 out of the 47 variants queried lead variants. Of these, none have previously been reported at genome-wide significance for dental diseases, 26 variants were associated with one or more non-dental disease traits and the remaining 20 variants have not previously been reported for any trait. Eleven lead variants have previously published associations with adiposity traits, 9 with height traits and 3 with traits representing bone health. A summary of positive findings is included in table 3.13. For periodontitis/loose teeth, rs12461706 has previously-reported association with monocyte count[317] in addition to the association with aggressive periodontitis described earlier in these results.

Table 3.13: Summary of single-variant cross-trait lookup for DMFS/dentures

Target RSID	Traits
rs4971099	Magnesium, urate, adiposity traits
rs2046850	Wheeze or whistling
rs62106258	Adiposity traits and basal metabolic rate
rs11676272	Adiposity traits, height and forced expiratory volume
rs2652452	Physical activity
rs263771	Physical activity
rs121908120	Hair loss
rs7429279	Adiposity traits
rs61790808	Height, adiposity and personality traits
rs55769264	Educational attainment
rs1122171	Height
rs9366651	Height, adiposity traits, educational attainment
rs898797	Bone mineral density, adiposity traits, height and red blood cell traits
rs10987008	Subjective overall health, height
rs149467613	Haematinic traits
rs6495046	Pulse rate, adiposity traits
rs10851907	Smoking traits and smoking-related diseases
rs8054556	Basal metabolic rate and adiposity traits
rs1108343	Bone mineral density
rs10048146	Bone mineral density, height
rs3865314	Height and adiposity traits
rs34559440	Adiposity traits
rs57067187	Height, physical activity, forced expiratory volume and educational attainment
rs28822480	Adiposity traits and basal metabolic rate
rs11672900	Renal, urinary and haematinic traits
rs1569414	Height, facial hair and hair loss, adiposity traits

Only variants with a previously-reported association with a non-oral health trait are presented.

3.3.8: Sensitivity analyses

Concordance in genetic effect estimates between GLIDE and UKB was described earlier in these results. Within GLIDE, concordance in genetic effect estimates in Hispanic/Latino participants and European ancestry participants was assessed. For all lead DMFS/dentures variants, the tests for heterogeneity in DMFS genetic effect estimate between HCHS/SOL and DMFS genetic effect estimates from meta-analysis of other studies in GLIDE did not produce evidence against the null hypothesis of no heterogeneity ($P_{FDR}>0.05$ for all tests for heterogeneity).

Some tooth loss due to periodontitis will be captured in the DMFS/dentures analysis. If genetic association with DMFS/dentures was due to genetic effects on periodontitis, then the estimated genetic effect should be higher in participants with periodontitis than in participants with good periodontal status. For variants identified in DMFS/dentures analysis, estimates of genetic effect on DMFS were obtained separately in periodontitis cases and controls, combined in stratified meta-analysis and then compared. There was little evidence for heterogeneity ($P_{FDR}>0.05$ in all tests), suggesting that periodontal tooth loss was not a major source of bias in genetic effect estimates at lead associated loci.

In a population-specific meta-analysis, effects of rs12461706 on periodontitis were estimated in participants of East Asian ancestry ($N_{controls}$=15,670, N_{cases}=1,680 from 2 studies). This population-specific analysis did not find strong evidence for association at rs12461706 (P=0.80), where the odds ratio was smaller than in the European ancestry meta-analysis (OR=1.03 for A allele vs 1.05 for A allele in European ancestry), and the effect allele frequency is much higher (EAF=0.96) than in the European ancestry population (EAF=0.40).

Genome-wide summary statistics for all single trait and multi-trait analyses are publicly available through the University of Bristol research data repository at (https://data.bris.ac.uk/data/dataset/2j2rqgzedxlq02oqbb4vmycnc2). Regional association plot and a full disclosure of all positive and negative findings from bioinformatic analyses are provided in the Supplementary Information file and Supplementary Datasets of the manuscript at (https://doi.org/10.1038/s41467-019-10630-1).

3.4: Discussion

Until now, the lack of dental data and genetic information in large-scale resources has been a barrier to understanding the role of common genetic variation in dental diseases. The use of index-linked and questionnaire-derived data were introduced as a possible solution to this problem in the last chapter and applied here for genetic association discovery. This demonstrates both the potential and limitations of both clinical and questionnaire-derived dental phenotypes.

At the outset, the largest challenge was understanding the biological meaning of questionnaire data when there is no clinical validation sample. Here, the use of genetic correlation at a genome-wide level was instrumental in helping to anchor interpretation and unlock these data for use in genetic association discovery. The combined DMFS/dentures analysis was successful, identifying variants of which the few (systematically-chosen) examples described in the text have plausibly-relevant biological functions. To return to the *WNT10A*, locus as an illustration, this gene encodes a member of the WNT/β-catenin family of signaling proteins which have documented roles in inducing and regulating tooth formation in animal models[318-320] in addition to their widely-known functions in embryological patterning and oncogenesis. In humans, *WNT10A* mutations have been reported to cause isolated defects in tooth number[321-324] and quality[325,326] and to regulate the cusp architecture and other morphological characteristics of teeth[327]. Genes encoding other proteins in the WNT signaling cascade (but not *WNT10A*) were identified as candidates in an early GWAS for dental caries with some sample overlap with the present study[70]. Likewise, the other 3 regions flagged in the results (*HLA, CA12 and PITX1)* all have plausible biological relevance, and the remaining loci not discussed here will provide substrate for more detailed *in-sillico* or biological follow-up experiments.

Alongside this success there are limitations in the workflow used in this chapter. It was difficult to obtain stable estimates of genetic correlation between periodontitis and other traits, highlighting that this workflow may not be suitable for traits where there is little polygenic association signal or where statistical power to detect polygenic association signal is low. The mathematical derivations of the BOLT-LMM method are based on a quantitative trait model with a univariate normal probability distribution. To apply this method to oral and

dental traits in UK Biobank, these case:control traits were treated as quantitative traits with a post-hoc transformation series to approximate genetic effects on a log odds ratio scale, as suggested by the authors of BOLT[128]. While the BOLT method has advantages in terms of statistical power and control for population substructure, the decision to treat case:control traits as continuous traits can become problematic. In situations where the case fraction is small, the assumption of a univariate normal probability distribution may be unreasonable, and this is of particular concern in situations where the minor allele frequency is low. As illustrated in simulations presented in supplementary table 8 of the BOLT manuscript[128], the combination of a small case fraction and low minor allele frequency is capable of inducing false positive associations. Based on these simulated scenarios and the case fractions for oral and dental disease traits in UK Biobank, a conservative threshold was set and variants with a minor allele frequency of less than 1% were excluded from analysis. Since this analysis was performed, the SAIGE method has been published[328] which uses an alternative approach to model the distribution of score test statistics and calibrate test statistics, meaning that SAIGE can be used even for rare variants with low case fractions.

Analysis in diverse populations can help ensure the findings of a study have external validity for a larger group of people but creates specific methodological challenges. The primary single-trait results in GLIDE included analysis in the HCHS/SOL study which is a study of Hispanic/Latino ancestry. The decision to include or exclude these results needed to consider the increased statistical power which could be gained by adding data balanced against the potential for additional heterogeneity in results which might occur if the biologically causal variants are tagged by different SNP makers in different ancestral groups. Formal tests for heterogeneity were performed to help inform this decision. There was little evidence for heterogeneity in association signals for variants with the greatest association for evidence, , helping to justify the inclusion of HCHS/SOL results in the primary analysis, although it is likely that these tests had low power to detect heterogeneity unless there was a large discrepancy in genetic effect estimates between HCHS/SOL and other studies. In general, the inclusion of multiple ancestral groups in GWAS can increase the risk of population stratification, where there is coincident variation in both allele frequency and outcome within a single dataset By definition, population stratification occurs at the study level (even if it only becomes visible at meta-analysis level), so the inclusion of HCHS/SOL results would not contribute to inflationary bias in meta-analysis provided the within-study modelling

approach was appropriate. For HCHS/SOL specifically, the bespoke modelling approach is reported to achieve good control for population stratification[287], however a conservative approach was taken using genomic control in all studies in GLIDE and tests for inflationary bias as described previously in these results.

The inclusion of non-specific measures creates some challenges for interpretation, for example tooth loss due to periodontitis will be captured in both DMFS analysis in GLIDE and dentures in UKB. Logically this must create biased estimates of genetic effect, but empirically there is little evidence of this happening. The genetic correlations between periodontitis and DMFS (within GLIDE) and between dentures and loose teeth (within UKB) were not strikingly high. The *SIGLEC5* periodontitis risk locus is not a main finding for DMFS/dentures. There is little evidence for heterogeneity in DMFS effect sizes between periodontal cases and controls. Together, these observations support the idea that the two principal multi-trait meta-analyses capture distinct dental diseases, in keeping with results from the last chapter that the observational covariance between periodontitis and DMFS was small, and it is likely that DMFS/dentures combined analysis predominantly captures dental caries. Despite this, the term 'DMFS/dentures' is preferable for transparency and is used instead of caries.

As understanding of genetic associations improves and the research question shifts towards more mechanistic questions it will become necessary to reconsider whether these measures still provide adequate phenotypic resolution. The previous chapter argued that high resolution phenotypes are only 'better' than low-resolution phenotypes when that resolution is required for the research question and exploited by the analysis design. In this chapter, phenotypes are used only to separate carriers and non-carries of risk alleles into groups with higher or lower burden of dental disease. As a next step, we need to consider why these genetic variants are important – i.e which mediators or biological processes lie between allele assignment at conception and signs or symptoms of dental diseases in adulthood. Answering these questions may require more resolved measures, not just of the teeth and clinically manifest dental endpoints but of potential intermediaries such as the characteristics of the oral microbiome. One advantage of starting from large-scale low-resolution measures before moving to refined measures is that the multiple testing burden is greatly reduced – it is now possible to focus on a limited number of genomic regions with good empirical evidence for

biological relevance - a more attractive prospect than trying to undertake GWAS in small-scale collections with detailed phenotypes, or picking a candidate gene which might be irrelevant[50]. The need for detailed follow-up collections to complement large discovery datasets is discussed further in chapter 6.

Until these collections are available it may be possible to use external sources of information to generate hypotheses. As examples, single-variant results were combined with external sources of transcriptomic data, functional annotation and information on a range of health outcomes using summary-level methods. Starting with external sources of transcriptomic data, the S-TissueXcan analysis prioritized 221 gene transcripts predicted to alter the burden of DMFS/dentures, which may be useful targets for understanding functional biology or novel interventions. Compared to examining single-variant results this approach helps improve statistical power to identify relevant genes[329], but is limited by the quality and availability of external data, for example data on transcription in odontoblast or ameloblast tissue are not available in GTEx.

Moving to literature annotations, there was little evidence that DMFS/dentures-associated loci or DMFS/dentures-associated transcripts were enriched for pre-defined gene sets. One possible interpretation is that a wide range of different biological pathways contribute to risk for dental diseases, and each of these pathways are represented by a relatively small number of risk loci or transcripts. This notion of a wide aetiological and biological footprint is similar to results from chapter 2 and is revisited in chapter 6.

One interesting finding was that the heritability of DMFS/dentures was enriched for genomic regions which are conserved through evolution. Here, the estimated fold-enrichment value is similar to that reported for other complex traits with serious health consequences[301], implying that selection pressure acts on biology which regulates the dentition and oral environment. By extension it may be the case that, in evolutionary terms, defects and diseases of the dentition affected reproductive fitness. It is argued that dental appearance acts as an ornamental trait in humans, providing a visible clue about potential mate quality by signalling information about the age, health and environmental and traumatic events which a person has been exposed to[330]. One possible interpretation is therefore that genotypes with negative effects on dental appearance might be selected against. Conversely, the meaning of dental appearance appears

to vary in different historical and social contexts and the attractiveness of 'straight, white teeth' may be a comparatively recent phenomenon[331], which may be relevant as the method used aims to identify signals of selection over an evolutionary timescale. In addition, the genetic correlation estimates presented earlier in this chapter suggested that high dental disease burden is associated with larger rather than smaller family sizes, potentially suggesting that mate selection on dental status is less important than other factors such as diet, nutrition or behavioural phenotypes..

Finally, the integration of DMFS/dentures results with published GWAS studies showed overlap in the genetic determinants of DMFS/dentures and a wide range of non-dental disease traits at both a single variant and genome-wide level. Despite using an entirely different approach from chapter 2, there were striking similarities in the results, for example educational attainment was associated with hazard for tooth loss in observational analysis in the previous chapter and had the single strongest genetic correlation with DMFS/dentures in genetic correlation analysis in this chapter. Although these correlations do not provide causal inference, it seems likely that these results support a role for socio-economic status in the aetiology of caries and tooth loss, as discussed further in chapter 6. A second striking feature is the detectable overlap with measures of poor systemic health, including biomarkers of poor metabolic function, measures of obesity and disease endpoints including coronary artery disease. These correlations might be attributed to a latent causal variable which manifests as multiple measures of health but could also be interpreted as indirect evidence for pathway effects linking dental diseases to systemic health, a hypothesis which is developed further and then tested in the next chapter.

One unexpected finding is the heritability estimates for periodontitis, which are very low compared to estimates reported in the literature[61,332]. This may be due in part to phenotypic misclassification using CPI-based measures, but this alone is unlikely to explain the low heritability estimate. The CPI-based periodontitis classifications are imperfect but, as discussed in the previous chapter, appear valid in observational analysis in GLIDE data, while the measures used in WGHS likewise have plausible associations with other variables in published data[333]. Instead, it is possible that different approaches to disease classification, differing pattern of periodontal treatment, varying age distributions in different GLIDE

cohorts, un-modelled gene x environment interactions and general limitations of SNP-based heritability estimation all contributed to the low observed heritability in meta-analysis.

3.5: Commentary

The combination of self-reported and index-linked dental provides one way to achieve the large sample sizes which are needed for GWAS of dental diseases. Using this approach to explore the biological basis of caries and periodontitis, the main findings are 47 novel risk loci and a shortlist of 221 gene transcripts which may help to refine the understanding of the molecular mechanisms of caries.

Contemporary models of caries aetiology consider the broader health and social context of disease. Here, a hypothesis free cross-trait investigation suggests that both caries and periodontitis have shared genetic determinants with other traits reflecting socio-economic circumstances and poor systemic health, again highlighting how genetic data can help confirm or challenge concepts about disease aetiology.

One widely held belief is that dental diseases have downstream effects on cardiovascular health. The availability of genetic proxies for dental disease experience now provides an opportunity to test this hypothesis and reciprocally, assess the relevance of metabolic health traits as modifiable risk factors for dental diseases. The next chapter therefore considers how genetic data can be used to test for causal association, using cardio-metabolic traits as an exemplar.

Chapter 4: Testing for causal relationships between dental diseases and cardio-metabolic traits.

So far, index-linked and questionnaire-derived dental data have been characterized in descriptive analysis and then used for genetic association discovery. The results of genetic association discovery can theoretically be used in turn to help strengthen epidemiological inference. This chapter exploits genetic information to test for causality in relationships between dental diseases and cardiovascular endpoints or cardio-metabolic traits.
The literature base in this field is extensive and encapsulates numerous hypotheses on the nature of relationships between dental disease and cardio-metabolic traits. To aid the structure of this chapter, a broad distinction is made between studies which treat dental diseases as putative exposure variables (in interpretation or modelling) and studies which consider dental diseases as outcomes. The aim of this chapter is to test for causal associations between dental diseases and cardio-metabolic trait using a bidirectional, two sample Mendelian randomization approach. Results presented in this chapter are also included in the GWAS manuscript published in *Nature Communications*[98]

4.1: Introduction

4.1.1: Dental diseases are associated with cardiovascular disease endpoints.

Associations between dental status and cardiovascular disease events are widely reported in the literature. A recent attempt to summarize published findings through systematic review and meta-analysis included 17 published studies and reported association between the presence or absence of periodontitis and myocardial infarction[334]. This association was recapitulated for signs or sub-features of periodontitis including clinical attachment loss, bleeding on probing and tooth loss[334]. Exploratory cross-sectional studies such as these help generate hypothesises for investigation in a longitudinal setting[335], and in dental epidemiology these longitudinal studies were designed to test the hypothesis that dental status influences risk for cardiovascular disease.

Results of cohort studies started to gain attention in the early 1990s with studies reporting that dental diseases at baseline were associated with increased hazard for coronary heart disease[336]. By 1998 some authors were already interpreting these studies as providing evidence for a causal effect of periodontitis on cardiovascular disease endpoints[337]. The Finrisk 1997 study is a good example of an influential study which adopted a longitudinal design to test for observational association between dental diseases (proxied by missing teeth at study baseline) and cardiovascular disease diagnosis over 13 years of follow up. In this study, missing teeth at baseline were associated with increased hazard for incident coronary heart disease events including acute myocardial infarction and all-cause mortality, with some evidence for a dose-response relationship[338]. As a body of evidence from similar studies has accrued, it has become possible to assimilate results across studies. A recent dose-response meta-analysis examining the relationship between tooth loss and risk for coronary heart disease and stroke included 17 studies. It reported that missing teeth at study baseline were associated with increased risk for both endpoints, with similar risk increment for both endpoints and consistent findings in subgroup design[339].

In Finrisk 1997, the authors interpreted their findings as potential evidence for systemic effects of periodontal disease. As seen in chapter 2 however, missing teeth might capture multiple aspects of dental health, so it is possible that systemic effects (if real) are driven by dental caries. The notion that these associations might not be specific to one compartment of dental disease is seen elsewhere in the literature, for example endodontic lesions on dental panoramic tomography are correlated with evidence for coronary artery disease severity on

coronary angiography[340]. Given that endodontic lesions are manifestations of endodontic infection and a late-stage sequel of dental caries, this observation sets the scene for testing a more general hypothesis around causal effects of dental diseases rather than investigating causal effects of one specific disease process.

4.1.2: Dental diseases are associated with metabolic traits

If dental diseases have effects on cardiovascular events then this might be through a mechanism which is specific to cardiovascular diseases, or through a non-specific metabolic mechanism which might influence risk of several diseases. One way to explore this further is by examining the relationships between dental diseases and metabolic traits. Here, systemic metabolic traits might act as a read-out of disruption of underlying biological events to provide clues regarding the range of biology potentially influenced by dental diseases. Existing studies report associations between dental diseases and groups of traits reflecting glycaemic control, lipid biomarkers and measures of adiposity.

For glycaemic control, both biomarkers of impaired glucose tolerance and overt clinical presentation of impaired glucose homeostasis (as diabetes) have been investigated. In cohort studies, missing teeth and periodontal status at study baseline are associated with subsequent diagnosis of type 2 diabetes[338,341,342]. Otherwise-healthy people with periodontitis appear to have poor glycaemic control and higher risk of developing type 2 diabetes than people without periodontitis[343]. These observations have been interpreted as suggesting that dental diseases have systemic metabolic effects[201,344,345].

If dental diseases can dysregulate systemic metabolic control this might be also be detectable using lipid biomarkers. A recent meta-analysis of 19 studies reported that serum concentrations of low density lipoprotein cholesterol (LDL) and triglycerides are elevated in participants with periodontitis compared to controls, while serum levels of high density lipoprotein cholesterol (HDL) are lower in periodontitis cases compared to controls[346]. Here, the assumed role of decreased serum HDL as a mediator of associations between periodontitis and cardiovascular outcomes is not completely clear as the causal relevance of HDL for cardiovascular disease has been called into question[347].

Altered systemic metabolic control might also be detectable by changes in total adiposity or body composition, meaning it may be helpful to test for effects of dental diseases on traits

such as body mass index (BMI) or waist to hip ratio (WHR). There is some support for this hypothesis, for example a cohort study examining for association between measures of mid-childhood dental status and late-childhood measures of body composition reported that higher DMFT scores at age 15 years were associated with increased odds of over-mean-WHR at 18, adjusted for baseline body composition measures[348]. Apart from metabolic effects, there are other plausible mechanisms by which dental diseases might influence body composition. Loss of oral function due to dental diseases is thought to lead to changes in diet quality or quantity – a hypothesis which has been tested in the literature but with inconsistent results[349]. Walking speed in older adults appears to decline more quickly with tooth loss[350], which might suggest people change their physical behaviour after tooth loss. BMI therefore serves as an example of a complex trait which could plausibly be influenced by dental diseases through several different mechanisms. Each of those mechanisms may itself include intermediaries. Using type 2 diabetes as an example, chronic low-grade inflammation is reportedly associated with onset of type 2 diabetes[351], leading to a hypothesis tested through formal mediation analysis that inflammatory biomarkers might mediate association between periodontitis and glycaemic status[352]. As with the HDL argument earlier, these proposed mechanisms are now being revisited in light of evidence suggesting that inflammatory biomarkers such as C-reactive protein could be a consequence rather than the cause of ill health[83-85].

4.1.3: Reported associations may not be causal.

Although the associations described above might be due to effects of dental diseases, the limitations of conventional observational studies prevent tests for causal effect. In general, confounding and reverse causation are the major limitations of observational epidemiology which affect even well-designed studies. In addition, there may also be sources of bias in recruitment, analysis, presentation or publication of results which systematically distort the available study findings.

Confounding

Confounding occurs when both the exposure and outcome are influenced by a shared causal variable. In the context of dental epidemiology, confounding is likely to be a major concern as many of the major risk factors for dental diseases and cardiovascular disease are shared (such as diet[353] and smoking[354] among others). While most studies attempt to make some degree of statistical adjustment for these features, it is unlikely that all confounding features

are perfectly measured or that these features are modelled in a perfect approach which does not rely on some simplifying assumptions.

Measuring confounding features is challenging, as the confounder might be a complex personality trait reflecting health awareness and self-care[355] or the social, behavioural, demographic or geographic circumstances which are conducive to or barriers against receiving healthcare; for example people with diabetes are reported to utilize dental services at a low rate and have limited oral health knowledge[356]. This is potentially relevant because, as discussed in chapter 2, these abstract, but important features are well-captured by measures of dental status but are not easily measured as a confounding feature in most studies. As another example of a difficult to measure confounder, a hypothetical genetic variant in a vascular pathway which increases risk of both periodontitis and stroke could induce association between these diseases but would be not measured or accounted for in existing studies. There are examples of other complex traits where spurious phenotypic correlation is generated by a small degree of underlying genetic correlation, misleading inference[357], and this idea is starting to be appreciated in dental epidemiology with a recent pilot investigation examining whether a polymorphism is a common risk factor for rheumatoid arthritis and periodontitis in a Japanese population[358].

Reverse causation

Reverse causation occurs when the outcome influences the exposure to mis-lead inference. As discussed later in this chapter, this is a concern in dental epidemiology as there are plausible mechanisms by which metabolic traits might influence dental diseases. Studies in this field have attempted to overcome reverse causation by adopting longitudinal designs but these may not be entirely robust to reverse causation. Complex health outcomes such as cardiovascular disease are likely manifestations of a pre-existing latent disease process rather than an event which begins only at the time of diagnosis. Systemic biological disturbance may pre-date the development of disease by several years, for example a polygenic risk score for coronary artery disease predicts metabolic disturbance in children as young as 7 years of age, well before any clinical manifestation of disease occurs[359]. Systemic inflammation is often suggested as a potential pathway through which dental diseases might influence cardio-metabolic health, but pre-symptomatic cardiovascular disease is itself a cause of inflammatory reactions[360,361] which might conceivably contribute to the aetiology of

periodontitis. Participants in dental cohort studies may already have a latent form of the outcome which is influencing their dental status long before the outcome is clinically manifest.

Other sources of bias

In addition to the inherent limitations of observational studies, there may be situations where inappropriate study design or sampling lead to bias, as discussed in chapter 2. There may be situations where selective reporting, publication or citation strategies lead to the overall literature base becoming misleading. In dental and oral health, it is reported that fewer than 50% of systematic reviews use appropriate methods to examine publication bias[362], and that statistically significant research findings are over-represented in the literature, implying selective publication[363]. Even among published studies, there may be inappropriate attention given to positive findings compared with well-designed studies reporting negative findings[364], a phenomenon termed citation bias. It is now argued that a series of biasing events occur between analytical decisions and citation decisions, with cumulative effects on the apparent relevance of a risk factor or intervention[365].

Cumulative effect on interpretation.

It is possible that these phenomena lead to over- or under-estimation of effect size without altering interpretation, but these bias effects might also lead to incorrectly rejecting the null hypothesis and false discoveries. For studies which treat dental diseases as an exposure there is little evidence to help quantify the extent of the problem, but for other traits there is evidence that the limitations of observational epidemiology are enough to mislead inference. One example is the hypothesis introduced earlier in this chapter, that low-grade levels of inflammatory biomarkers predispose to type 2 diabetes. Although this finding was compelling in observational analysis[351], administration of vitamin D in randomized controlled trials can reduce inflammatory biomarker levels[366] but does not appear to change insulin secretion rate, glucose tolerance or markers of glycaemic status in patients with type 2 diabetes[367-369]. Collectively, the body of evidence in this example suggests that the cohort study identified a confounded rather than causal relationship. For dental diseases it would therefore be helpful to compare estimates using different techniques including those which allow for causal inference.

Methods for causal inference

In contrast with inflammatory biomarkers which could be modulated in a randomized controlled trial, there is no practical or ethical way to randomize participants to having or not having dental diseases. Randomized controlled trials have instead tended to use designs where all participants require treatment and are randomized to high intensity or low intensity treatment, or randomized to immediate or postponed treatment. For ethical reasons, it is necessary to keep these trials short in duration, and the primary endpoint is typically a readily-measured biomarker rather than a clinical endpoint such as cardiovascular disease events. In one such trial, patients with periodontitis and type 2 diabetes randomly allocated to receive more intensive periodontal treatment had more favourable inflammatory biomarker profiles after treatment[370]: an interesting finding which does not help address the hypothesis that periodontitis is a causal risk factor for type 2 diabetes. Other evidence from trials is discussed later in this chapter in the context of results.

The difficulties with performing large-scale trials with long follow-up periods and clinically relevant endpoints mean it may be helpful to look at epidemiological methods for causal inference which do not require intervention. As introduced in chapter 1, approaches including difference in differences analysis, negative control designs, instrumental variable analysis, regression discontinuity and panel data regression aim to make causal inference from epidemiological data[86-89,134]. In recent years there has been increasing interest in using genetic data to aid causal inference[88]. The unique properties of genetic data help in this regard, as the shuffling of genetic data during recombination events mean that common, LD-independent genetic variants appear randomized within an ancestrally-homogenous population, unlike environmental features which are clustered[371]. The stability of germline genetic variation across the life course provides a fixed reference point which is unaffected by the passage of time and therefore not influenced by reverse causation.

Exploiting these unique properties, Mendelian randomization (MR) is a paradigm which extends the notion of phenocopy to evaluate the causal effect of an exposure on an outcome[372,373]. Genetic and environmental exposure which modify the same biological pathway can lead to the same clinical presentation. Here, the genetic phenocopy reflects the impact of the modifiable risk factor on disease[92]. By extension, the genetic phenocopy can be used to mirror or 'proxy' variation in a modifiable environmental risk factor and estimate the

likely impact of that environmental risk on a different disease outcome. To date, the lack of well-characterized genetic association signals for dental diseases has precluded use of MR experiments to test for evidence of causal effects of dental diseases. The results presented in the previous chapter now provide an opportunity to use genetic proxies for dental diseases, while large-scale studies of cardiovascular disease endpoints and cardio-metabolic traits provide a source of outcome data. Collectively, these resources can now be exploited to test for causal effects of dental diseases on cardio-metabolic traits and outcomes.

4.1.4: Theory of MR experiment designs

Assumptions made when using MR in the context of an instrumental variable analysis

All studies using the MR method make three assumptions. First, the genetic variant must be strongly associated with the exposure variable of interest. Second, the genetic variant must not be associated with confounding features of the exposure-outcome association. Third, the genetic variant must not have effects on the outcome through pathways other than its effects on the exposure (sometimes termed the exclusion-restriction assumption[374]).

These assumptions are derived from classical instrumental variable analysis which was first applied in economics and some terminology has also been borrowed from economics. It is common to describe genotypes used in an MR context as 'instruments' and to make a distinction between 'valid instruments' which meet all 3 assumptions and 'invalid instruments' which fail to meet one or more criteria[375]. In truth it is more likely that genetic variants only approximate the characteristics required for an instrument in classical instrumental variable analysis, as pleiotropic effects are widespread[376], some traits have association signal across much of the genome[377] and even common genetic variation will contain fine-scale latent structure[55]. It is therefore likely that many or all genotypes might be 'invalid' in classical instrumental variable analysis, but with a spectrum of properties which are either conducive to estimating causal effect or problematic. For the remainder of this chapter, the terms 'valid instruments' and 'invalid instruments' are used in line with their usual interpretation in the literature and not to suggest that the application of genetic phenocopies in MR analysis is exactly equivalent to instrumental variable analysis.

One sample designs

Early examples[83-85] used genetic, exposure and outcome data in a single study in a design now termed '1 sample MR'. Here, the exposure is regressed on genotype dosage using

ordinary least squares regression, then the regression model is used to predict fitted values of the exposure for each participant. The outcome is then regressed on the variance in the exposure explained by genotype in a second least squares regression model, the gradient of which represents the causal effect estimate. This approach is sometimes termed 'two stage least squares'.

Although this model does not include a direct test for genotype – outcome association, any causal effect of the exposure on outcome must induce association between the genotype and outcome. Thus, examining genotype-exposure and genotype-outcome associations can yield the same inference as the model described above. Intuitively, if the genotype has a very large effect on the exposure but is only weakly associated with the outcome, then the effect of the exposure (proxied by the genotype) on the outcome must be small. This intuition, formalized as a ratio of the coefficients of the genotype-outcome and genotype-exposure association, is termed the 'Wald ratio'[378] and is one way of estimating the causal effect. The main distinction (as compared to the two stage least squares estimation described above) is that only the genotype-exposure and genotype-outcome association statistics are required, meaning this parameter can be estimated from summary level association statistics rather than requiring participant-level genetic data.

Two sample designs

In recent years the potential has risen for studies which examine the genotype-exposure and genotype-outcome associations in different populations, termed '2 sample MR'[90]. These studies have enhanced flexibility in that the exposure and outcome data do not need to be available in the same dataset. These designs can have enhanced power, as the most precise genotype-exposure and genotype-outcome effect estimates available can be used. While there are benefits to the 2 sample approach there are also natural limits. It is difficult to test whether the instrument is associated with confounders of the exposure-outcome association and the 2 sample method introduces an additional (implicit) assumption that both datasets are drawn from the same underlying population, with similar ancestry and comparable distributions of the exposure and outcome. As we saw in chapter 2, the presence of unaddressed population stratification in one study can lead to over- or under-estimation of a SNP-trait association, thus differential control for populations stratification between the two datasets can also yield biased causal effect estimates[54].

It is possible to use more than one genetic proxy in an MR experiment. A well-powered GWAS for an exposure may identify multiple independently-associated variants, which could be used in an MR study. The chief advantage of including multiple genetic variants is that the total variation in the exposure proxied by the genetic variants could increase, resulting in improved statistical power to estimate causal relationships. If participant-level data are available, one method involves aggregating the variants into a composite instrument such as a weighted or unweighted polygenic risk score. Where each individual genotype is only weakly associated with the exposure, this approach may be desirable as a single strong instrument may be less susceptible to biasing effects from weak instrument bias[378,379] than multiple weak instruments. Where summary-level data are available this form of aggregation cannot be used, but it is possible to estimate the causal effect using each genetic proxy in turn and combine these estimates into a pooled estimate. One approach involves inverse variance weighted (IVW) meta-analysis of the causal estimate from each variant, which is equivalent to the 2 stage least squares causal estimate when the instruments are uncorrelated[380].

Detecting and accounting for violations of instrumental variable assumptions in a MR context

Using the terminology introduced earlier in this chapter, these approaches for combining multiple instruments assume that all instruments are valid. This assumption becomes challenging as the number of instruments increases – by probability alone it appears more likely that one or more instruments are invalid when many are included in an MR experiment. It may be helpful to consider ways relax this assumption and examine methods which aim to estimate an unbiased causal effect when one or more instruments are invalid. For dental traits, we have already seen in chapter 3 that the genetic architecture is complex and might include variants involved in pathways related to cardiovascular disease risk (such as vascular pathways) or might be associated with features classically thought to be confounders such as smoking status, motivating the use of a method other than IVW for the primary analysis.

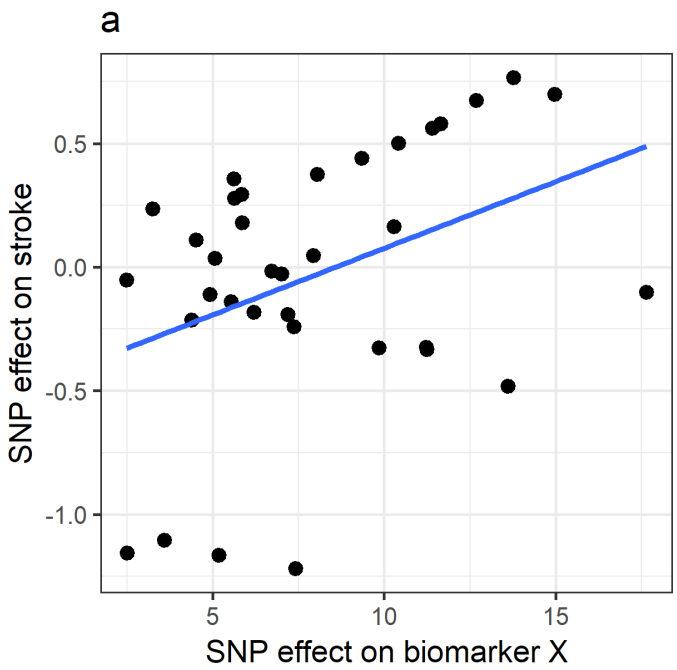
It is theoretically possible to detect invalid instruments by manually screening the list of variants and annotating these with information about their biological function and association with other traits based on the available literature. If these characteristics appear undesirable, the variant could then be excluded from analysis. This method relies on good prior

understanding of the biological mechanism of every variant and published well-powered GWAS studies for all major traits which might be influenced by pleiotropic pathways. This is likely an unrealistic expectation given the challenges around variant annotation, where even identifying the biologically causal gene can be difficult[306] as described in the previous chapter. Similarly, the assumption that the existing literature contains well-powered GWAS for all major confounding traits or biomarkers of pleiotropic pathways is unrealistic. Finally, there is no single set of circumstances under which a variant is either valid or invalid, as the pleiotropic effects might be uncorrelated with one outcome (a valid instrument) and associated with another outcome (an invalid instrument). Manual screening and exclusion of instruments is therefore not considered further in this chapter.

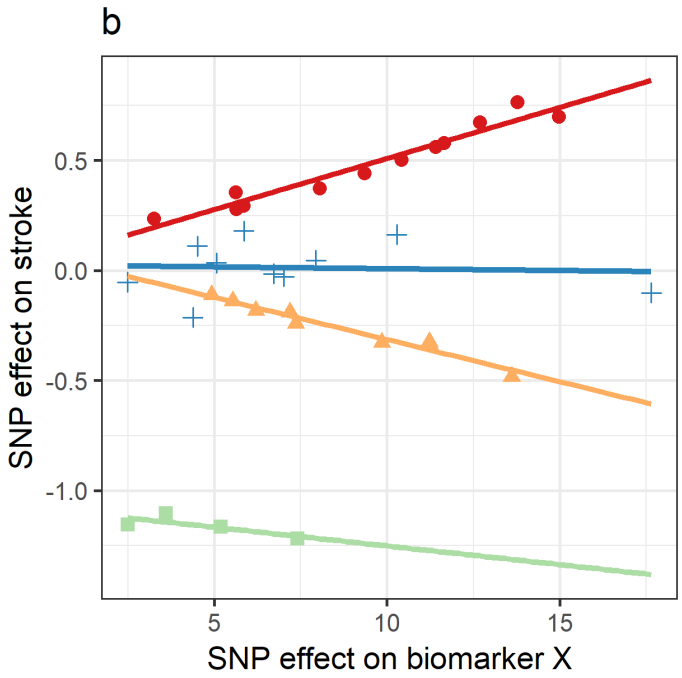
An alternative way to screen and exclude instruments involves studying the distribution of causal effect estimates to empirically identify variants which may be invalid. If a population of genotypes have effects on an outcome only through their effects on an exposure, then the effects on exposure and outcome must be proportionate and the causal effect estimate obtained from each genotype will be similar. Conversely, if a genotype is an invalid instrument with effects on the outcome through some other mechanism other than effects on the exposure, it might have an implausibly large or small effect on the outcome which is disproportionate to its effect on the exposure, producing an outlying estimate of the causal effect. By comparing causal effect estimates across the population of instruments, it should be possible to identify at least some invalid instruments by apparent heterogeneity in their causal effect estimate. This is only possible if information on the distribution of causal effect estimates has been preserved through analysis. This distribution is lost when aggregating multiple variants in a polygenic score and formed part of the motivation for not including polygenic scores in this chapter. While heterogeneity in causal effect estimates may help detect invalid instruments, this makes the assumption that all instruments identify the same causal parameter[381], which might not be the case, e.g. if a non-specific exposure variable contains multiple sub-traits which have different causal effects on the outcome (Figure 4.1). For binary outcomes heterogeneity can also arise from imperfect linear approximation on a log odds ratio scale[382].

Figure 4.1: Heterogeneity in causal effect estimates as an indication of a complex exposure rather than violations of the MR method

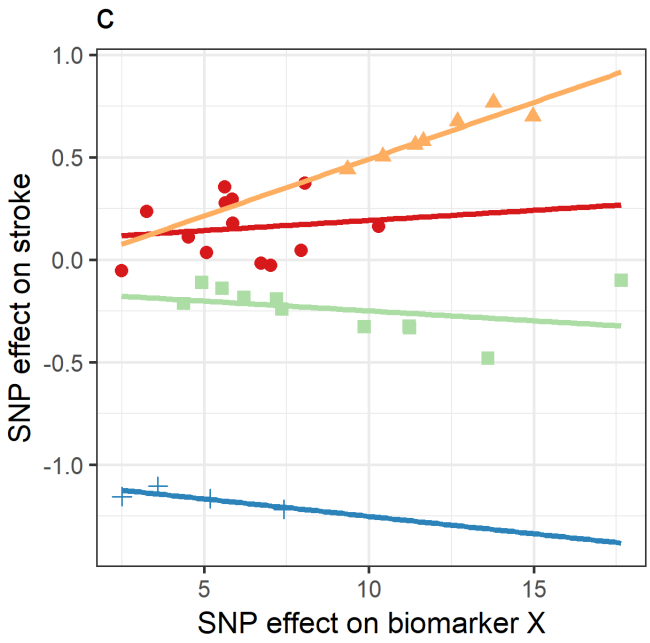
The figures show results of an MR experiment testing the effects of a hypothetical exposure, 'biomarker X', on stroke. This biomarker is upregulated in response to cigarette smoke and vigorous exercise. In an MR experiment using variants identified in GWAS for 'biomarker X' there is considerable heterogeneity in causal effect estimates and overall evidence for unbalanced horizontal pleiotropy (panel a).



Heterogeneity arises because 'biomarker X' is a complex trait which could be influenced through several different pathways, and these pathways have different effects on stroke (panel b). Genotypes which are associated with exposure only through effects on production and transport mechanisms (acting as phenocopies for a drug intervention) have no effect on stroke (blue crosses). Genotypes which are associated with the exposure through vigorous exercise (orange triangles) and smoking (red circles) have divergent effects on the outcome as they estimate the effect of different interventions on the exposure. The remaining variants (green squares) have pleiotropic effects on both the exposure and outcome.



In practical applications the exact properties of each genotype will not be known but it may be possible to cluster genotypes into groups of variants acting as phenocopies for an underlying pathway which affects the exposure. For example, using unsupervised hierarchical clustering it is possible to identify clusters of variants which are similar to the true group allocations in this example (panel c).



By developing and applying this approach it might be possible to infer not just which exposures are relevant in an outcome, but which interventions on that exposure would be most effective with respect to the outcome.

These empirical methods continue to make a distinction between 'valid' and 'invalid' instruments, which are included or excluded from analysis respectively. If instruments are instead considered to have a spectrum of characteristics, it might be helpful to reduce the impact of variants with outlying effect estimates without resorting to exclusions. Modal estimation techniques attempt to estimate causal parameters which are not heavily influenced by outlying variants. Recently, the modal estimation approach was extended to include a heterogeneity-penalized model averaging procedure[383] which estimates the causal effect while iterating over subsets of instruments. The mixture distribution of these estimates is then considered to identify, and down-weight estimates which are contaminated by inclusion of invalid instruments in the iterated subset, producing a final estimate of causal effect which is stable and robust provided at least 50% of variants included in the experiment are valid instruments. The key limitation is that the non-parametric derivation of the most likely causal effect has reduced precision compared to the parametric derivation used by methods such as IVW meta-analysis so may not be suitable for situations where statistical power is a concern.

Another approach involves including a model-level term aiming to capture the average pleiotropic effect. In MR-Egger regression[384] the genotype-exposure and genotype-outcome space is conceptualized as containing both a causal effect (the degree by which increasing exposure associates with increasing outcome) and an intercept term (the apparent effect of a variant on the outcome which has no effect on the exposure, capturing the average unbalanced pleiotropic effect across all variants). A key assumption of this method is that magnitude of pleiotropic effects must be uncorrelated with the magnitude of their effects on the exposure. This assumption may be problematic if certain types of variant (for example low-frequency variation, or protein-coding variants with deleterious consequences) have larger effects than other classes. Violations of this assumption are one reason for inflated type 1 error rates using this method, which also has challenges in terms of statistical power[385].

4.2: Methods

Two sample MR was used to estimate the causal effect of dental diseases on metabolic traits and cardiovascular outcomes. Results of genetic association discovery in chapter 3 were used as a proxy for dental disease exposure, while data on metabolic traits and cardiovascular outcomes were taken from published GWAS consortia.

4.2.1: Exposure

Results from combined meta-analysis of DMFS/dentures were used to select proxies for dental disease exposure. First, variants in the HLA region (chr5: 25-35 Mb) were removed except for a single lead variant. This procedure was performed because the LD structure of the HLA region is reported to be highly variable[300], so LD clumping based on an external reference panel might yield correlated instruments. Next, LD clumping was performed to produce LD-independent variants, using default criteria of $p<5x10^{-8}$ for index variants, an r^2 threshold of 0.01 and reference data from the cohorts arm of the UK10K project (https://www.uk10k.org/data.html).

Results from combined meta-analysis of periodontitis/loose teeth were considered, however the estimation techniques used for primary analysis requires a minimum of 10 independently-associated variants[386]. In addition, binary exposures can violate the exclusion restriction criterion[374] and for these reasons, this trait was not carried forward in the primary analysis.

4.2.2: Outcome data

GWAS literature was searched for studies of cardiovascular outcomes, restricted to studies of European ancestry populations with genome-wide data publicly available. To maximize statistical power and reduce multiple testing burden, the single largest study for each disease was selected (based on sample size) and the same disease could not appear in more than one form. For example, if there were studies of haemorrhagic stroke, ischaemic stroke and a combined analysis of all stroke, the GWAS of all stroke would be chosen. Following these criteria, 2 studies were selected, coronary artery disease taken from the CARDIoGRAM plus C4D meta-analysis[387], and all stroke from the MEGASTROKE consortium[388].

GWAS literature was searched for traits reflecting metabolic status which might plausibly be influenced by dental diseases based on the evidence reviewed earlier in this chapter. Traits related to glucose metabolism, lipid biomarkers and adiposity were reviewed, and for each

trait the single largest study of European ancestry participants with genome-wide data was selected. In total, this strategy yielded 7 traits. There were two traits related to glucose metabolism (type 2 diabetes from the DIAGRAM consortium[144]; fasting glucose from the ENGAGE consortium[389]), three traits related to lipid biomarkers (HDL cholesterol, LDL cholesterol and triglycerides, all from the Global Lipids Genetics Consortium[390]) and two traits related to adiposity (BMI from meta-analysis of the GIANT consortium with UK Biobank[142] waist-to-hip ratio adjusted for BMI from the GIANT consortium[391]).

4.2.3: Causal effect estimation

Selecting methods for primary and sensitivity analysis

Before deciding on an effect estimation strategy, the properties of the genetic proxies for dental diseases were reviewed as follows. First, the overlap analysis in chapter 3 indicated that approximately 50% of these variants were associated with one or more complex traits, meaning that they represent a population of variants with pleiotropic effects. The primary estimation technique therefore needed to include steps to detect and account for bias from horizontal pleiotropy. Second, the effect sizes on DMFS/dentures were generally modest to small, meaning that the instruments would only explain a small proportion of variation in the exposure. The primary estimation technique therefore needed to produce precise causal estimates.

After considering these properties, Generalized Summary Mendelian Randomization with the Heterogeneity in Dependent Instruments test (GSMR-HEIDI) was chosen as the primary estimator. This method takes a causal effect from a variant near the middle of the effect distribution as reference, tests whether other variants have heterogeneous estimates of causal effect and rejects variants with highly heterogeneous causal effect estimates before estimating the causal effect from remaining variants. This procedure yields similar causal effect estimates to MR-Egger analysis, but with greater statistical power in a range of real-life complex trait applications[386]. Apart from the HEIDI filtering, GSMR has other advantages over IVW meta-analysis in that it can account for residual LD between variants not removed by clumping, and can estimate the sampling distribution of genotype-exposure and genotype-outcome estimates from genome-wide data, resulting in slightly improved power compared to IVW estimation[386]

IVW meta-analysis, MR-Egger and heterogeneity-penalized model averaging were selected for sensitivity analyses. These were considered less suitable for the primary analysis because of their limitations in terms of statistical power (MR-Egger, heterogeneity-penalized model averaging) and absence of a test for horizontal pleiotropy (IVW) but were included for comparison purposes.

Estimation of causal effects

Primary analysis was performed using GSMR with HEIDI[386], implemented in GCTA[299] standalone software (v1.92.0). HEIDI-outlier filtering used default criteria to remove variants with heterogenous estimates of causal effect ($p<0.01$ in test for heterogeneity). The GSMR method uses genome-wide data. To ensure that other estimators which use SNP-level data passed through the same process for extracting and harmonizing SNP effects, the GSMR method was performed but omitting the HEIDI test as a wrapper to produce harmonized SNP-level data. Next, these data were imported into R to perform sensitivity analyses. IVW meta-analysis and MR-Egger tests were performed using the 'MendelianRandomization' R package (v0.3). The heterogeneity-penalized model averaging procedure was performed using the R code supplied by the manuscript which describes the method[383].

4.2.4: Interpretation of causal effect estimates

The genotype-exposure estimates used the standardized regression coefficients introduced in the previous chapter, meaning a 1 unit change corresponds to a 1 standard deviation greater dental disease experience as proxied by the combination of DMFS/dentures. To relate this to a tooth surface scale, causal effect estimates (and their standard errors) were rescaled first to standard deviations of tooth surfaces then to tooth surfaces using the transformation series described in the previous chapter.

Causal effect estimates for outcomes on a log-odds scale (coronary heart disease, all stroke and type 2 diabetes) were exponentiated to express the effect estimate as an odds ratio. For outcomes on a continuous scale, the effect estimates were transformed to a clinically tractable unit where possible, for example standard deviations of BMI were converted to a KgM^{-2} scale using the median per-study standard deviation of BMI reported by the GIANT consortium[392].

4.2.5: Multiple testing

Initial interpretation of results was based primarily on estimated effect size and 95% confidence intervals, so untransformed P-values are provided.

4.3: Results

In reference data from the GLIDE cohort introduced in chapter 2, the mean DMFS score at age 50 years was 44 tooth surfaces and a one standard deviation change in DMFS score at age 50 years corresponded to 20 decayed, missing or filled tooth surfaces.

4.3.1: Estimated causal effect of dental disease on cardiometabolic traits in primary analysis

The estimated effect of 1 standard deviation greater exposure to dental disease on cardiovascular disease was an odds ratio of 1.10 (95% CI 0.98, 1.23, P=0.092). For stroke, the estimated effect of 1 standard deviation greater exposure to dental disease was an odds ratio of 1.12 (95% CI 1.01, 1.25, P=0.037). If a Bonferroni approach was used to account for multiple testing, then neither association would be considered statistically significant.

For glucose-related traits, greater exposure to dental disease was estimated to associate with greater odds of type 2 diabetes (OR 1.20, 95% CI 1.04, 1.38, P=0.014). Conversely, there was little evidence for causal association between dental disease and fasting glucose, where the estimated effect was a near-null 0.0026 mM greater fasting glucose (95% CI -0.045, 0.051, P=0.92) for 1SD greater dental disease exposure.

For lipid biomarkers, the estimated causal effect of dental disease was -0.021 (95% CI -0.078, 0.036, P=0.47) standard deviations of inverse normal transformed (SD INT) HDL cholesterol, and 0.016 SD INT (95% CI -0.044, 0.076, P=0.60) for LDL cholesterol. For triglycerides the effect estimate was 0.080 SD INT, with 95% confidence intervals which excluded the null and a test statistic which would pass a Bonferroni-corrected threshold for association (95% CI 0.027, 0.13, P=0.0032).

For adiposity traits, there was little evidence for causal association between dental diseases proxied by DMFS/dentures and total adiposity proxied by BMI, where the effect estimate was -0.02 KgM^{-2}, (95% CI -0.14, 0.10, P=0.75) for 1SD greater exposure to dental disease. Conversely, when using waist to hip ratio adjusted for BMI as a proxy for body composition and visceral fat, there was evidence for unfavourable adiposity increasing effects of dental

disease (effect estimate 0.078 SD INT, 95% CI 0.027, 0.13, P=0.0025). Main results are summarized in table 4.1.

Table 4.1: Causal effect estimates from the principal GSMR-HEIDI analysis

Outcome	Untransformed Beta (SE)	Transformed Beta (95% CI)	Units for outcome	NSNP	P
Coronary artery disease	0.13 (0.079)	1.10 (0.98, 1.23)	OR	54	0.092
All Stroke	0.16 (0.077)	1.12 (1.01, 1.25)	OR	55	0.037
Type 2 diabetes	0.25 (0.10)	1.20 (1.04, 1.38)	OR	51	0.014
Fasting glucose	0.004 (0.034)	0.0026 (-0.045 0.051)	mM	54	0.92
HDL-c	-0.030 (0.041)	-0.021 (-0.078, 0.036)	SD INT	44	0.47
LDL-c	0.023 (0.043)	0.016 (-0.044, 0.076)	SD INT	44	0.60
Triglycerides	0.11 (0.038)	0.080 (0.027, 0.13)	SD INT	45	0.0032
BMI	-0.006 (0.019)	-0.02 (-0.14, 0.10)	KgM^{-2}	35	0.75
Waist Hip Ratio adjusted for BMI	0.11 (0.036)	0.078 (0.027, 0.13)	SD INT	45	0.0025

Untransformed effect estimates are provided for a 1SD greater burden of dental diseases proxied by DMFS/dentures. Transformed effect estimates are provided for a 1SD greater burden of DMFS.

4.3.2: Estimated causal effect of dental disease on cardiometabolic traits in sensitivity analysis

Sensitivity analysis used 3 alternative estimation tools which handle variants with outlying causal effects in a different manner from the GSMR-HEIDI primary analysis. IVW meta-analysis does not include any steps to detect or account for these variants, MR-Egger analysis fits a model term representing the overall estimated impact of unbalanced horizontal pleiotropy and the model averaging procedure down-weights groups of variants whose causal effects are heterogeneous compared to other sets of variants. Comparison of effect estimates using these different procedures was intended to help assess which findings were sensitive or robust under multiple modelling assumptions. All sensitivity analyses produced less precise causal effect estimates than the GSMR primary analysis. This lack of precision was most marked for the MR-Egger estimates, where there were large standard errors and imprecise estimates for all outcomes, preventing meaningful comparison of effect estimates from MR-Egger with other methods.

All estimates from IVW meta-analysis agreed with the GSMR primary results within 95% confidence intervals, and IVW and GSMR results generally had similar interpretation. For type 2 diabetes, coronary artery disease and stroke, the IVW estimates were slightly larger in

magnitude than those for the GSMR primary analysis, however this was not reflected in stronger test statistics due to an accompanying increase in standard error.

The model averaging procedure produced larger causal effect estimates than GSMR for triglycerides and coronary artery disease, and smaller causal effect estimates for type 2 diabetes and all stroke. The confidence intervals for the model averaging estimates were generally wide and were compatible with the GSMR primary results for all traits within 95% intervals (Table 4.2).

Table 4.2: Summary of results of sensitivity analyses using different estimation tools

Outcome	GSMR Beta (SE)	IVW Beta (SE)	MR Egger Beta (SE)	Model averaging Beta	95% CI for model averaging Beta
BMI	-0.006 (0.019)	0.094 (0.070)	-0.38 (0.23)	-0.06	-0.10, -0.01
WHR *adj* BMI	0.11 (0.036)	0.094 (0.047)	0.29 (0.17)	0.12	0.01, 0.22
Type 2 Diabetes	0.25 (0.10)	0.28 (0.15)	0.48 (0.49)	0.16	-0.24, 0.57
Fasting glucose	0.004 (0.034)	0.006 (0.035)	0.084 (0.12)	-0.06	-0.15, 0.05
HDL-c	-0.030 (0.041)	-0.12 (0.070)	0.31 (0.25)	-0.14	-0.25, -0.004
LDL-c	0.023 (0.043)	0.067 (0.064)	-0.053 (0.24)	0.084	-0.08, 0.29
Triglycerides	0.11 (0.038)	0.11 (0.049)	-0.24 (0.17)	0.25	0.12, 0.42
Coronary artery disease	0.13 (0.079)	0.17 (0.11)	-0.51 (0.34)	0.20	-0.40, 0.90
All Stroke	0.16 (0.077)	0.18 (0.081)	0.11 (0.26)	0.046	-0.20, 0.57

All effect estimates are untransformed and provided for a 1SD greater burden of dental diseases proxied by DMFS/dentures

4.4: Discussion

This investigation used genetic proxies for a non-specific measure of dental disease to test a general hypothesis that dental diseases have causal effects on cardiovascular disease or cardio-metabolic traits. In general, MR experiments produce imprecise causal estimates, and this was seen in these results, where several findings are equivocal. Nevertheless, the experiment had sufficient statistical power to confidently reject the null hypothesis for two traits and exclude the presence of large causal effects for the remaining seven. Distinguishing between imprecisely-estimated null findings and modest causal effects is more challenging; while the following discussion section interprets results based on the point estimates, these interpretations may need to be revised as statistical power increases.

In general, there was good agreement between the different estimation techniques used, suggesting that these estimates are not highly sensitive to different approaches in

identification and accounting for potentially invalid instruments. One caution in this interpretation is that the wide confidence intervals for some methods may have masked important differences between methods which will only emerge in future experiments with greater statistical power.

Starting with cardiovascular outcomes, the effect estimates were similar for coronary artery disease and stroke. There was slightly stronger statistical evidence for stroke, possibly reflecting the larger sample size of the MEGASTROKE consortium compared to CARDIoGRAM plus C4D. This similarity in findings is compatible with the hypothesis in the literature that dental diseases are pro-atherosclerotic with proposed mechanisms involving release of heat shock proteins or cytokines from gingival and periodontal tissues in response to dental plaque[393-395]. A range of oral bacteria including the ubiquitous *Streptococcus gordonii* are capable of inducing platelet aggregation[396], meaning that bacteraemia related to dental diseases may be pro-thrombotic. Although much of the evidence base relates to periodontitis, dental caries is also associated with atherosclerosis[397].

Atherosclerotic pathways are therefore a plausible mechanism for the apparent causal effect on stroke seen in this investigation, but systemic effects of the oral microbiome need not be through atherosclerosis. For example, the cariogenic *Streptococcus mutans,* which is plausibly regulated by several of the variants used as instruments in this experiment, may be relevant in haemorrhagic stroke and small vessel micro-bleeds through binding to collagen in damaged cerebral blood vessels and promoting local bleeds[398]. Detection of the periodontal pathogens *Porphyromonas gingivalis*, *Prevotella intermedia* and *Aggregatibacter actinomycetemcomitans* is likewise associated with brain injury with hypotheses for both pro-atherosclerotic and more general pro-inflammatory mechanisms[399-401].

The hypothesis that poorly-controlled dental disease alters glucose control in people with established type 2 diabetes has been suggested in the literature. Small-scale controlled trials which randomize participants with periodontitis and type 2 diabetes to comprehensive or minimal periodontal treatment have reported improved biomarkers of glycaemic control in the group receiving more aggressive treatment[402-405]. This finding is supported by meta-analysis[406] although slightly larger studies including[407] tend to report more conservative findings than the small trials described above, and a more recent systematic review[408]

continued to highlight methodological problems in terms of random sequence generation and lack of blinding for some studies. Systemic spread of inflammatory mediators produced in periodontal tissue has been proposed as a possible mechanism[409,410], although a recent narrative literature review highlighted conflicting evidence in this field and a lack of mechanistic studies to support this hypothesis[411].

The MR estimates presented above test slightly different hypotheses. First, the causal effect of dental diseases on blood glucose in non-diabetic individuals is tested, producing little evidence for causal effects. It is possible to reconcile the MR effect estimate with the reported evidence from trials if this distinction in underlying population is important. For example, the potential effects of dental diseases on blood glucose might be well controlled in people with good underlying glucose homeostasis, but poorly controlled and therefore detectable in people who already have impaired glucose tolerance. Next, the causal effect of dental diseases on odds of type 2 diabetes is tested, finding evidence for an odds-increasing effect. Interpreted in the same light as the finding above, this might suggest that, in people with pre-diabetic states, increasing burden of dental disease has metabolic consequences which accelerate progression to a diagnosis of type 2 diabetes. This interpretation is highly speculative and has little support in the literature but nevertheless may have policy implications – for example a Cochrane review of periodontal interventions for glycaemic control for people with diabetes advocates setting up large-scale long-term studies in cohorts of diabetic patients[412], but there may be greater benefit in preventing diabetes through earlier interventions. While there are few animal studies investigating the topic, one study reported that pre-diabetic mice with induced periodontitis did not appear to have impaired glycose tolerance, earlier onset of diabetes or increased severity of diabetes after onset compared to a control group where periodontitis was not induced[413]. It is possible that the effects of dental disease on type 2 diabetes are not mediated through acceleration of pre-diabetic states, but some other mechanism which is currently unclear.

For lipid biomarkers the MR estimates supported a trait-increasing effect of dental diseases on triglycerides which was not seen for either LDL-c or HDL-c. There is some evidence in the literature that successful management of periodontitis can reduce triglyceride levels, for example a small-scale randomized controlled trial reported patients with periodontitis and hyperlipidaemia randomized to a more aggressive periodontal treatment regime had lower

triglyceride levels 2 and 6 months after treatment compared to patients randomized to a less comprehensive treatment arm[414], a finding which was not recapitulated in a larger randomized controlled trial[415]. Here again, the apparent discrepancy between the MR estimates and trials might be explained if long-term (i.e. lifetime) exposure to dental diseases is more important than short-term variation in periodontitis severity, if the effects of dental diseases on triglycerides relate to caries rather than periodontitis, or if the causal effect detected in MR is too subtle to be seen in these small-scale trials which are typically designed to detect large changes in the primary outcome measures.

For adiposity traits the MR estimates suggest little effect of dental diseases on total adiposity, but that dental diseases might predispose towards unfavourable adiposity, proxied by increased WHR adjusted for BMI. These effects, if verified, might be expected to manifest as altered total adiposity, but in practice it appears cross-sectional associations between dental diseases and measures of adiposity are stronger for measures of body fat percentage and abdominal obesity than measures of total adiposity[416,417].

In a rat model, the morphological characteristics of white adipose tissue appear to change after inducing periodontitis, regardless of whether the rat is obese or normal-weight[418], with the authors suggesting a mechanism of interleukin 6 and tumour necrosis factor α (TNF α) mediated dysregulation of insulin sensitivity. Other mechanisms have been suggested including 'infectobesity'[419], the notion that the microbiome, including the oral microbiome, may modulate host metabolism. Specifically in the oral microbiome, it has been suggested that *Selenomonas noxia,* a common periodontal pathogen, influences host obesity through altered metabolic efficiency in the gastro-intestinal tract, effects on leptin or ghrelin signalling or metabolic effects mediated through adiponectin or TNF α signalling[420,421].

Taken together, the MR results suggest dental diseases have little or no effect on some measures of metabolic health and have a modest detrimental effect on other measures of metabolic health and cardiovascular outcomes. Modest population-level effect estimates might plausibly mask sub-groups of metabolically vulnerable people in whom dental diseases have large downstream effects, and who would benefit from enhanced screening for or management of dental diseases. The monotonicity assumption[422] described here for the outcome measures may be important for the exposure where the modest overall effect of

dental diseases might mask a large effect of longstanding untreated dental diseases or a small or null effect for treated dental caries. Although methods for estimating sub-group or local causal effects in MR are emerging[422], these are not easily adapted to summary-level methods. Finally, even if the true causal effect is modest in all groups of the population it may still be of importance for public health and policy. Considering type 2 diabetes as an example, the context is one of the most morbid and costly global diseases[423] where existing population-level interventions have had limited success in arresting what has been described as a global epidemic[424]. In this context, the apparently modest effects of dental diseases may be of importance for public health and policy if they provide an alternative way to target the global burden of cardio-metabolic disease.

4.5: Testing for reverse-causal effects

The first half of this chapter opened with a statement that dental caries and periodontitis are correlated with cardio-metabolic traits. The MR effect estimates described above suggest that, although causal effects of dental diseases may exist, the effects are generally modest, and the reported correlations are unlikely to be entirely due to causal effects of dental diseases. It therefore seems plausible that there are either reverse-causal relationships or problems with residual confounding in the observational literature base. The MR design provides an opportunity to explore this further by estimating the causal effect of cardio-metabolic traits on dental diseases. The next half of this chapter now describes the rationale, methods, results and possible interpretation of reciprocal analysis.

4.5.1: Metabolic traits as a risk factor for dental diseases

Diabetes and hyperglycaemia are thought to be risk factor for both caries and periodontitis. For dental caries, altered salivary characteristics in people with diabetes have been proposed as a possible mechanism. In cross-sectional studies, participants with diabetes have lower salivary pH, decreased salivary flow rate and decreased salivary calcium concentrations compared to non-diabetic participants[425,426], and these features might unfavourably tip the balance of mineralization and demineralization of dental hard tissues to predispose towards dental caries. Similarly in animal studies, rats injected with alloxan (to model type 1 diabetes) and db/db mice (to model type 2 diabetes) develop spontaneous dental caries after long term hyperglycaemia, while control rodents did not[240].

For periodontitis, impaired fasting glucose and diabetes disease status are associated with periodontitis in a dose-dependent manner in cross-sectional data[427]. In longitudinal studies, a recent meta-regression reported that having a diagnosis of diabetes at study baseline was associated with increased odds of incident diagnosis of periodontitis or progression of pre-existing periodontitis during study follow-up[428]. Elevated levels of glycated haemoglobin (as a biomarker for hyperglycaemia) at study baseline are also associated with increased risk of developing periodontal pocketing during longitudinal follow-up over a 5 year period[429].

Here a range of mechanisms have been suggested, including effects of diabetes on neutrophil function[430], the oral microbiome, microvasculature in periodontal tissues or periodontal repair by gingival fibroblasts. Regulation of human gingival fibroblasts appears important as fibroblast signalling events may be inceptive in sustained periodontal inflammation[431], while

fibroblasts are also important in wound healing and repair. Fibroblast signalling, proliferation and migration appear to be disrupted in a high glucose environment, possibly due to high oxidative stress[432], while diabetic rats have delayed gingival wound healing compared to non-diabetic rats[433].

Finally, it has been suggested that diabetes has effects on both caries and periodontitis through the oral microbiome. Salivary glucose levels are reported to correlate with diabetes diagnosis[426,434] and degree of diabetic control[435], to the extent that salivary glucose has been suggested as a biomarker for screening for type 2 diabetes or monitoring blood glucose levels in diabetic pateients[436,437]. As high levels of salivary glucose (defined here as ≥ 1.0 mg/dL) are associated with alterations in microbiome composition[438], it appears plausible that these changes in saliva composition promote a more pathogenic microbiome, in turn contributing to dental caries risk. Likewise, there are differences in the oral microbiome of diabetic and non-diabetic adults with periodontitis[439] which might have pathological relevance. While these studies are included here to illustrate the motivation for examining possible causal effects of metabolic traits on dental diseases, these studies do not assert a directional effect and are therefore equally relevant to the analysis presented earlier in this chapter.

The literature around lipid biomarkers and dental diseases predominantly explores the hypothesis that dental diseases modulate lipid metabolism rather than the testing for reverse-causal effects. Some authors however have interpreted their findings as suggesting effects of lipid traits on dental diseases, for example a cross-sectional study of Korean adults reported association between lower HDL-c and periodontitis and between higher LDL-c and periodontitis, which the authors interpreted as a risk-increasing effect of lipid traits[440]. In animal models, rats with induced periodontitis fed a high cholesterol diet have more severe periodontal tissue destruction than control rats with induced periodontitis, with the authors suggesting that hypercholesterolemic conditions increase production of pro-inflammatory signalling molecules resulting in in more serve periodontal inflammation[441].

Excess adiposity has been postulated as a risk factor for periodontal disease since at least 1977, when obese-hypertensive rats were reported to have heightened response to local gingival stimuli using a stainless steel wire mechanical irritation model compared to spontaneously hypertensive rats[442]. Cross-sectional studies in humans subsequently identified

an association between obesity and periodontitis, and this evidence has been synthesized in systematic reviews[443,444]. In longitudinal analysis of statutory medical check-up data for employees living near Nagoya, Japan, higher BMI scores at baseline are associated with increased hazard for developing periodontal disease after 5 years[445]. Similarly, students at a Japanese university whose BMI scores increased between entering university and graduation had increased odds of a an increase in CPI score over the same time period compared to students with stable BMI[446].

Aside from obesity as a risk factor for incident dental disease, authors have suggested that obesity modulates dental disease trajectory or response to treatment. In longitudinal analysis the mean number of teeth affected by periodontal disease events (defined as pocketing >3mm) during study follow-up is reported to be higher in participants with higher BMI scores at baseline[447]. Analysis of the same participant group considering a binary endpoint of 'periodontal disease progression events' (alveolar bone loss around 2 or more teeth progressing to ≥40%, two or more teeth progressing to probing pocket depth ≥5 mm, or clinical attachment loss progressing to ≥5 mm) reported a similar finding that BMI was associated with hazard of periodontal disease progression[448]. As with the literature described earlier in this chapter, caution is needed as tooth brushing and flossing frequencies were lower in participants with higher BMI scores[447] in this study population. Like arguments presented earlier in this chapter, it is possible that shared latent traits upstream of both dental and systemic health may have confounded these observed associations.

In summary then, it is possible that cardio-metabolic traits influence dental diseases with plausible mechanisms. Obtaining causal effect estimates from MR may help complement the effect estimates presented earlier in this chapter and provide a more holistic understanding of the relationships between dental diseases and cardio-metabolic traits. Indeed, the relationship between BMI and periodontitis was tested using MR methods in 2015[449]. Since then, improved understanding of genetic associations with BMI and the larger dataset with periodontal outcomes (described in chapter 3) provides an opportunity to re-examine this relationship with greater statistical power, while causal effects of other metabolic traits on either caries or periodontitis remain untested.

4.6: Methods

4.6.1: Exposures

Each trait used as an outcome earlier in this chapter was now considered as a potential exposure. As before, binary traits were not considered as possible exposures, meaning that 3 traits were excluded (coronary artery disease, all stroke and type 2 diabetes) leaving 6 quantitative traits which met the criteria for inclusion in analysis. For these traits, instruments were defined using LD clumping with default criteria of $p<5x10^{-8}$ for index variants, an r^2 threshold of 0.01 and reference data from the cohorts arm of the UK10K project (https://www.uk10k.org/data.html).

4.6.2: Outcome data

GWAS literature was searched for other traits reflecting periodontal or dental caries status. No large-scale studies in independent collections were identified, for example the largest published study for DMFS[66] is a participating study in GLIDE, with results already contained within the DMFS/dentures combined analysis. Outcome data were therefore taken from the combined meta-analyses of DMFS/dentures and periodontitis/loose teeth described in chapter 3.

4.6.3: Causal effect estimation

As before, the primary analysis was performed using GSMR with HEIDI. Sensitivity analysis used IVW meta-analysis, MR-Egger and a model averaging procedure using the same workflow as described earlier in this chapter. If multiple cardio-metabolic traits had a suggestive association with a single outcome measure, then additional sensitivity analysis was performed using multivariable IVW meta-analysis[450]. This analysis aimed to estimate the direct causal effect of each associated trait, for comparison with the other methods which estimate the total causal effect[451].

4.6.4: Interpretation of causal effect estimates

For DMFS/dentures as an outcome, effect estimates were transformed to a tooth surface scale using the transformation series introduced previously. For periodontitis/loose teeth as an outcome, effect estimates were exponentiated to give effects on an odds ratio scale.

4.6.5: Multiple testing

Test results were interpreted based on effect estimates and confidence intervals so untransformed P-values are presented.

4.7: Results

4.7.1: Estimated causal effects of metabolic traits on dental diseases in primary analysis

Using DMFS/dentures as an outcome, the estimated effect of 1mM greater fasting glucose was 1.1 (95% CI 0.2, 2.0) additional decayed, missing and filled tooth surfaces at 50 years of age (P=0.015). For periodontitis/loose teeth, the effect of fasting glucose was imprecisely-estimated, in the direction of an odds-increasing association (OR=1.06) but with wide confidence intervals which crossed the null (95% CI 0.95, 1.20, P=0.28).

For lipid biomarkers, the causal effect estimates for HDL-c, LDL-c and Triglycerides on DMFS/dentures were close to the null, with relatively tight confidence intervals. The estimates of causal effect on Periodontitis/loose teeth were less-precisely estimated but had similar interpretation to the null effects suggested for DMFS/dentures (Tables 4.3, 4.4).

For adiposity traits, the estimated effect of BMI was an additional 0.78 (95% CI 0.70, 0.86) decayed, missing and filled tooth surfaces at 50 years of age for 1 KgM^{-2} greater BMI ($P=6.0x10^{-75}$). For periodontitis/loose teeth the estimated effect was an odds ratio of 1.05 (95% CI 1.04, 1.06, $P=3.0x10^{-18}$) for 1 KgM^{-2} greater BMI. These effects were not recapitulated using WHR*adj*BMI as an exposure and either DMFS/dentures or periodontitis/loose teeth as outcomes, where the effect estimates were imprecise, possibly reflecting the smaller number of instruments for WHRadjBMI compared to BMI.

Table 4.3: Estimated causal effect of metabolic traits on DMFS/dentures in GSMR primary analysis.

Exposure	Untransformed Beta (SE)	Transformed Beta (95% CI)[1]	Units for exposure	NSNP	P
Fasting glucose	0.040 (0.016)	1.1 (0.2, 2.0)	mM	23	0.015
HDL-c	-0.0001 (0.006)	-0.004 (-0.3, 0.3)	SD INT	159	0.98
LDL-c	-0.003 (0.005)	-0.08 (-0.4, 0.2)	SD INT	132	0.58
Triglycerides	-0.007 (0.008)	-0.19 (-0.6, 0.2)	SD INT	95	0.37
BMI	0.13 (0.007)	0.78 (0.70, 0.86)	KgM^{-2}	804	$6.0x10^{-75}$
Waist Hip Ratio adjusted for BMI	-0.017 (0.016)	-0.46 (-1.3, 0.4)	SD INT	47	0.30

[1]All transformed effects are expressed in tooth surfaces per unit of exposure.

Table 4.4: Estimated causal effect of metabolic traits on periodontitis/loose teeth in GSMR primary analysis.

Exposure	Untransformed Beta (SE)	OR (95% CI)$_1$	Units for exposure	NSNP	P
Fasting glucose	0.028 (0.026)	1.06 (0.95, 1.20)	mM	23	0.28
HDL-c	0.015 (0.009)	1.04 (1.00, 1.08)	SD INT	169	0.083
LDL-c	0.006 (0.008)	1.01 (0.97, 1.05)	SD INT	145	0.51
Triglycerides	0.006 (0.011)	1.01 (0.96, 1.07)	SD INT	104	0.58
BMI	0.09 (0.01)	1.05 (1.04, 1.06)	KgM^{-2}	854	$3.0x10^{-18}$
Waist Hip Ratio adjusted for BMI	0.015 (0.025)	1.04 (0.92, 1.16)	SD INT	48	0.55

$_1$All transformed effects are expressed as odds ratios per unit of exposure.

4.7.2: Estimated causal effects of metabolic traits on dental diseases in sensitivity analysis

In sensitivity analysis using alternative estimation tools, the estimated effect of fasting glucose on DMFS/dentures was consistent within 95% confidence intervals in all methods, but with the MR-Egger and the model averaging producing a larger point estimate than the primary technique. Treating periodontitis/loose teeth as an outcome, the estimates from MR-Egger agreed with the primary estimate within 95% confidence, while the estimated effect from the model averaging procedure was more than twice as large as that in the primary analysis and did not agree within 95% confidence intervals. Collectively, these sensitivity analyses suggest that the estimated causal effects of fasting glucose on dental diseases in the primary analysis may be conservative.

For lipid biomarkers as exposure and DMFS/dentures as outcome, sensitivity analyses produced similar effect estimates with consistent interpretation to the primary results. For lipid biomarkers as exposure and periodontitis/loose teeth as outcome, MR-Egger and the model averaging procedure suggested a larger effect of HDL-c than that supported by the primary analysis, but these estimates agreed within 95% confidence intervals.

For adiposity traits, the estimated effects of BMI on DMFS/dentures and periodontitis/loose teeth were larger in MR-Egger and the model averaging procedure than in the corresponding primary analysis, again suggesting that the primary analysis yielded conservative effect estimates. The effects of WHR*adj*BMI on DMFS/dentures and periodontitis/loose teeth were imprecisely estimated in sensitivity analysis. (Tables 4.5 and 4.6).

The only pairs of exposures and outcomes meeting the prior criteria for multivariable MR were the estimated effects of fasting glucose and BMI on DMFS/dentures. Fitting a multivariable IVW model to these relationships, the estimated effects of fasting glucose and BMI on DMFS/dentures were essentially unchanged (Table 4.7), suggesting the effects of BMI on DMFS/dentures were not mediated through effects of BMI on fasting glucose.

Table 4.5: Estimates of causal effects of metabolic traits on DFMS/dentures using alternative estimation tools.

Exposure	**GSMR Beta (SE)**	**IVW Beta (SE)**	**MR Egger Beta (SE)**	**Model averaging Beta**	**95% CI for model averaging**	**99% CI for model averaging**
Fasting glucose	0.040 (0.016)	0.047 (0.023)	0.091 (0.051)	0.063	0.027, 0.11	0.009, 0.12
HDL-c	-0.0001 (0.006)	-0.05 (0.07)	0.006 (0.012)	-0.0053	-0.016, 0.008	-0.020, 0.012
LDL-c	-0.003 (0.005)	-0.002 (0.008)	0.012 (0.013)	0.0006	-0.010, 0.011	-0.014, 0.015
Triglycerides	-0.007 (0.008)	-0.027 (0.011)	-0.024 (0.018)	-0.015	-0.031, 0.006	-0.036, 0.012
BMI	0.13 (0.007)	0.14 (0.01)	0.14 (0.03)	0.22	0.19,0.24	0.18, 0.25
WHR *adj* BMI	-0.017 (0.016)	-0.012 (0.018)	-0.042 (0.072)	-0.026	-0.08, 0.03	-0.11, 0.05

Table 4.6: Estimates of causal effects of metabolic traits on periodontitis/loose teeth using alternative estimation tools.

Exposure	GSMR Beta (SE)	IVW Beta (SE)	MR Egger Beta (SE)	Model Averaging Beta	95% CI for model averaging	99% CI for model averaging
Fasting glucose	0.028 (0.026)	0.027 (0.035)	0.071 (0.078)	0.11	0.055, 0.18	-0.10, 0.21
HDL-c	0.015 (0.009)	0.016 (0.008)	0.027 (0.013)	0.024	0.002, 0.045	-0.006, 0.05
LDL-c	0.006 (0.008)	0.004 (0.008)	0.004 (0.013)	0.009	-0.008, 0.03	-0.015, 0.033
Triglycerides	0.006 (0.011)	0.009 (0.011)	0.016 (0.019)	0.012	-0.018, 0.043	-0.027, 0.057
BMI	0.09 (0.01)	0.12 (0.01)	0.11 (0.036)	0.14	0.11, 0.18	0.09, 0.19
WHR *adj* BMI	0.015 (0.025)	0.014 (0.029)	0.021 (0.11)	0.005	-0.05, 0.06	-0.08, 0.09

Table 4.7: Comparison of the estimated total and direct causal effects of BMI and fasting glucose on DMFS/dentures from univariable and multivariable MR models.

Exposure	IVW beta (SE)	MV IVW Beta (SE)	95% confidence intervals for MV IVW Beta	MV IVW P
Fasting glucose	0.047 (0.023)	0.061 (0.022)	0.017, 0.11	0.006
BMI	0.14 (0.010)	0.13 (0.010)	0.11, 0.15	4.5×10^{-38}

4.8: Discussion

The second half of this chapter explored the hypothesis that metabolic traits have effects upon dental status, without testing possible mechanisms. Compared to results presented earlier in this chapter, there was generally better statistical power, suggesting that the genotypes proxying metabolic traits explain more variation in these traits than the genotypes proxying dental diseases. Compared to results earlier in this chapter, there were more occasions where the different estimation techniques disagreed, and sensitivity analysis generally suggested that effect estimates from the GSMR primary analysis were conservative.

Earlier in this discussion, a suggestion was made that caution is needed in interpreting MR findings when the confidence intervals are wide. As an illustration of this, the only previous study testing for causal effects of total adiposity on periodontitis[449] concluded that total adiposity was not a risk factor for periodontitis. Conversely, the results of this chapter suggest that biological events which lead to greater total adiposity also increase liability to dental diseases, a finding which already has some support in the literature including evidence from animal models[452,453]. Although these interpretations are apparently contradictory, the effect estimates themselves are not, as the causal effect estimate from the present study sits within the 95% confidence intervals from the previous study, and the authors highlighted that their study lacked statistical power to detect small causal effects[449]. The importance of statistical power is considered further in this discussion.

There was some evidence that impaired glucose homeostasis (proxied by fasting glucose in non-diabetic participants) has effects on the dental diseases proxied by DMFS/dentures. Here, the statistical evidence was weak, in part due to the limited power of this analysis, and in isolation the MR results would not provide good evidence against the null hypothesis. In conjunction with the findings from conventional epidemiology and mechanistic studies reported earlier in this chapter however, the body of evidence tends to favour that impaired glucose homeostasis is a risk factor for dental diseases, an interpretation which is uncontroversial in clinical practice.

In conjunction with results presented in the first half of this chapter, it seems likely that causal pathways explain at least some of the correlation between dental diseases and metabolic traits seen in observational regression or genetic correlation analysis but are

unlikely to explain all the correlation. Taking DMFS/dentures and HDL-c as an example, there was strong evidence for genetic correlation between these traits but little evidence for causal effects in either direction, where the effect estimates were near-null with relatively tight confidence intervals. While the MR findings do not exclude small causal effects, it seems more plausible that common causes of both DMFS/dentures and HDL-c account for most of the phenotypic correlation between these two traits.

In looking for these common causes, it is likely that common environmental risk factors play a part, as mentioned earlier in this chapter. Counter to this, it might be argued that common genetic variation is uncorrelated with most forms of environmental risk[371] and that correlation arising solely from external environment should not have been detected as genetic correlation. Here it may be helpful to consider a spectrum of environment from purely external to purely internal, encompassing the heritable environment, defined as external environmental factors which are related to genetic factors. Here, tobacco smoking may provide an example of a trait which has heritable contributions[454], and alters the internal environment through epigenetic changes[455]. Likewise, genetic association with location[55] means that any local or regional environmental features which are visible in phenotypic scores in GWAS are likely also visible (to a lesser degree) in the association signals arising from the GWAS, meaning that genetic correlations may have similar interpretation to observational correlation for traits with coarse geographic disparity. Collectively, the phenomena of genetic nurture[456], geographic association with common genetic variants and imperfectly-controlled population stratification mean that genetic variants acting as proxies for complex traits may also proxy the trait-causal environment. In the context of MR for complex exposures and complex outcomes, this may mean that that IV-outcome relationship is confounded by unobserved environmental factors. This problem has recently been illustrated in a commentary highlighting the potential for causal estimates to become biased in the direction of observational associations, and that no currently available MR methods can detect and account for unobserved environmental confounders[457]. Groups of genetic variants with similar characteristics can yield different associations with geography[55], meaning that it is difficult to predict the degree of bias and there is currently no statistical framework available to account for this using summary-statistic based methods. Future research using MR in the context of dental diseases may benefit from including within-family analysis, which is theoretically robust to the biasing effects described above, however these studies are

not currently large enough to identify the (likely modest) effects of dental diseases on other health outcomes.

While these challenges are common to all MR experiments using complex traits, there is an additional challenge which is specific to bidirectional analysis. It is possible to imagine scenarios where the genetic determinants of two traits are so similar (or the measurement strategy so imprecise and non-specific) that it is no longer meaningful to obtain MR estimates as the proportional similarity in SNP-exposure and SNP-outcome associations is due to auto-correlation rather than causal effects of the exposure on outcome. This problem was recently formalized in a latent causal variable method[458], which includes a model term for genetic risk shared across exposure and outcome to try to avoid the problem described above.

Given this background, it is important to evaluate whether the strategy used in this experiment produced a set of genetic predictors for DMFS/dentures which are sufficiently different from those used for other traits in reciprocal analysis. First, the strict cut-off criteria for defining instruments in conjunction with finite statistical power means that the instruments should be enriched for variants with (relatively speaking) large effect sizes, which are likely to be biologically proximal to dental caries, and this is reflected in the apparent biological function of the loci described in chapter 4. Next, there is little evidence for enrichment of DMFS/dentures loci in any described gene sets, which runs counter to the hypothesis that this set of genotypes captures variation in a latent process which manifests in multiple traits. Finally, invalid instruments are likely to have outlying causal effect estimates compared to valid instruments, so should either be detected as heterogeneous variants and then excluded in the GSMR-HEIDI analysis or down-weighted in the heterogeneity-penalized model averaging procedure. While no design is perfect, it is unlikely that biasing effects from common causes are enough to generate type 1 errors in this experiment, and this appears to be borne out using the HDL-c and DMFS/dentures example described above.

A separate concern is that the limited statistical power of MR methods may have contributed to type 2 errors and a failure to identify small-magnitude causal associations which will only emerge in larger studies with greater statistical power. Aside from interpretation of the primary effect estimates, the limited power of this experiment contributed to wide confidence intervals in sensitivity analysis and means that discrepancies between different estimation

techniques could only be identified if there was a substantial disagreement between methods. Despite these limitations of power, the MR estimates may still be helpful to place bounds on the likely upper bound of causal effect, meaning it may be possible to exclude large effects. Given these considerations of statistical power, it was necessary to select a limited number of traits, and the traits included here are not exhaustive but aim to capture major facets of cardio-metabolic health. Future applications could involve extending this analysis to other diseases or health outcomes, such as testing the hypothesis that tooth loss influences risk of dementia[459], or more detailed characterization of the apparent metabolic effects of dental diseases through a network-MR approach[460,461] and GWAS results for epigenetic[462], metabolomic[463] or other -omic datasets. In this space, the lack of a large-scale resource with detailed characterization of the oral microbiome may be a barrier to testing mechanistic hypotheses, and the need for large datasets with detailed dental phenotypes is revisited in chapter 6.

Mendelian randomization effect estimates are often interpreted as the effect of an exposure across much of the life course[464]. For DMFS/dentures as an exposure this interpretation is challenging in literal terms as children cannot develop caries in permanent teeth until those teeth erupt and tooth loss is becoming uncommon except in older adults[465]. A less literal interpretation is that the DMFS/dentures variants are proxies for underlying biological processes which manifest as high or low levels of dental disease experience for any given stage in life. For reference, the magnitude of causal effects is given in tooth surfaces in adults in northern Sweden at 50 years of age, but the transformed effect estimate would be different in a population with higher or lower levels of dental disease. While changes in DMFS scores across lifetimes introduce the requirement to recalibrate transformed effect estimates to different populations, these changes in DMFS scores with age are not anticipated to play a large part in either the direction of causal effect or strength of evidence of causal effect. Although these limitations are relevant, they need to be considered in the context of the limitations of the evidence base from observational studies and lack of opportunities to test hypotheses through randomized controlled trials. With appropriate and cautious interpretation, the findings therefore offer useful additional evidence about the nature of associations between dental and cardio-metabolic traits.

4.9: Commentary

Observational associations between dental and systemic diseases are widely reported but do not provide causal inference. Using genetic data to explore a small subset of relationships, there is evidence supporting aspects of forward-causality, reverse-causality and effects of shared causal risk factors, and these aspects have varying importance for different relationships. While dental diseases may be a modifiable risk factor for some health outcomes, the effect sizes appear modest and the trait-specific patterns of association mean it may not be appropriate to extrapolate this small set of findings to make general statements. Nevertheless, the implication that interventions to reduce dental diseases will have more general benefits for population health is exciting and warrants further investigation.

In reciprocal analysis, fasting blood glucose and total adiposity were prioritized as the metabolic traits which are most likely to increase risk of dental diseases, but the effect sizes were modest in magnitude. This chapter provides an illustration of the flexibility of an analytical method which may help feed into future aetiological models for caries and periodontitis and inform future studies which aim to clarify the biological mechanisms and intermediaries underpinning causal relationships. The possible designs of those studies are discussed in chapter 6.

Before then, it may be helpful to consider other dental traits which are poorly represented in the GWAS literature. Dental caries in children and adolescents may have a subtly different molecular aetiology from caries in adult populations, so the next chapter considers whether GWAS for caries in paediatric populations can help explore the aetiology of caries across the life course.

Chapter 5: GWAS for dental caries in children and adolescents

Analysis in chapter 3 was motivated by a desire to learn more about the molecular basis of caries and periodontitis, but also by the need to identify genetic proxies for dental disease exposure for use in epidemiological analysis. To satisfy the assumption that both exposure and outcome studies are drawn from the same underlying population, analysis in chapter 3 was restricted to adult populations similar to those represented in the GWAS literature for cardiovascular outcomes and metabolic traits.

The molecular aetiology of caries in children might be subtly different from adults. Undertaking GWAS studies for caries in paediatric populations may provide an opportunity to learn more about the molecular basis of disease and identify dentition-specific aspects of the molecular aetiology. This chapter borrows many of the same principals used in chapter 3 including a consortium-based approach and appropriate combination of clinical and self-reported traits, but now considers caries traits in children and adolescents as outcomes. The aim of this chapter is to identify genetic variants which are associated with dental caries in paediatric populations. Results presented in this chapter have been published in *Human Molecular Genetics*[99].

5.1: Introduction

5.1.1: Rationale for separate analysis of adults and children

There are general arguments for designing GWAS to allow for age-specific genetic effects. For example, the genetic determinants of head circumference in infancy are closely related to but not identical with those in adolescence[96], and the effects of a single genotype may vary across different developmental windows from birth to adulthood, being detectable by deviated trajectory in a trait rather than absolute value of a trait[466]. Here, there is a putative interaction between genotype and age (i.e the effects of the genetic variant vary with time). In a GWAS meta-analysis this could theoretically be exploited by modelling the interaction term as part of the within-study analysis and bivariate meta-analysis of both the genetic main effect and interaction terms[467]. A simpler approach giving similar inference is to adopt a developmentally-sensitive meta-analysis design stratified by age group which allows the magnitude of the genetic main effect to vary between groups of cohorts where the age of participants differs.

The second major argument for separate analysis of dental caries in adult and paediatric populations is that the statistical characteristics of the ascertained traits are so different in adults and children that it would be inappropriate to adopt the same modelling approach for both populations. Dental caries is a person-level disease[182], and an individual is either affected or unaffected. In populations with near-universal experience of dental caries such as the Swedish GLIDE cohort described in chapter 2, it is uninformative to use this definition and more helpful to investigate severity of disease manifestation using measures such as the counts of caries-affected teeth used in chapter 3. In paediatric populations however, there are typically similar numbers of caries-affected and caries-free participants. Around 50% of children have evidence of caries by age 5 in industrialized nations[468-470], meaning that a binary approach to caries classification is practical. Conversely, this approach is not perfect given the loss of information inherent in dichotomization[471]. Likewise, there are arguments that molecular characteristics[472] are closer to the biofilm imbalance that drives the disease than the clinical endpoints, so might be better measures to understand disease biology, but detailed molecular phenotypes are currently not available in large-enough collections for a GWAS study.

A per-person approach to disease classification is therefore a feasible way to investigate caries in paediatric populations but not for studies which combine adults and children. There

are also practical considerations making the use of a per-person classification desirable for this investigation. Study protocol varied between participating paediatric cohorts, and while the degree of detail in clinical examination varied, these could be harmonized by using a common denominator of caries-free or caries-affected. In the ALSPAC study, clinical data were combined with self-report data. Here there might be limited confidence in the absolute number of self-reported caries-affected teeth, but we have already seen from chapter 2 that these data have face validity in separating participants with high caries liability from participants with low caries liability. A mother or father might know for certain whether they took their child to the dentist to have a filling but be less confident about how many fillings were placed at that appointment. Finally, the ZINB modelling approach used in chapter 2 is not currently scalable to undertake genome-wide analysis, whereas logistic regression (which had similar interpretation to the ZINB modelling results presented previously) is widely implemented in most software tools, again favouring a caries-free versus caries-affected approach.

5.1.2: Rationale for separate analysis of primary and permanent teeth

It is possible that caries in primary teeth may represent an aetiologically-related but subtly distinct disease from caries in permanent teeth. For example, the pattern of association between bacterial species and caries severity differs between the primary and permanent dentition[473], patterns of detection of the cariogenic *Steptococcus Mutans* genomic DNA in human saliva are dentition-specific[474] and the microstructural and chemical attributes of primary teeth are different from those of permanent teeth[475]. It therefore seems plausible that common genetic variation could tag dentition-specific biological processes, or that the relative importance of genomic risk loci might vary between the primary and permanent dentition. Thus, in addition to separating adult participants from children, it may be important to preserve distinction between primary and permanent teeth.

Taking genome-wide genetic effects in aggregate, this hypothesis has been tested using a family-based pathway modelling approach, producing evidence for both shared genetic influences and dentition-specific influences[59]. At a single-variant level, early GWAS investigations for caries in the primary[71] and permanent[70] dentition nominated different candidate loci[71,253,476,477], leading to a side-by side test of a panel of candidate genes, showing inconsistent association patterns between primary and permanent dentition[478]. The authors

concluded that genetic effects on dental caries susceptibility (at least those mediated through enamel) may differ between the primary and permanent dentition[478].

The design of this investigation therefore involved parallel characterization of dental caries in primary and permanent teeth, to allow for dentition-specific genetic effects. Each participating study tested for association between common genetic variation and log-odds of caries, and results were combined in dentition-stratified meta-analysis.

5.2: Methods

5.2.1: Study population

Genome-wide analysis was undertaken in a consortium containing 9 coordinating centres, each of which had access to data from one or more cohort studies. For the purposes of this description, 'clinical dental examination' means that a child was examined in person, whether in a dental clinic or study centre. In this description, 'examiner' refers to a dental professional such as a dentist, dental hygienist, dental therapist, dental nurse, while 'assessor' refers to an individual with training who is not a dental professional, for example a research nurse.

Studies participating in analysis

ALSPAC

The ALSPAC cohort was described in chapter 2. A subset of children attended clinics including clinical dental assessment by a trained assessor at age 31, 43 and 61 months of age, and these were used to derive a clinical phenotype in the primary dentition. Parents were asked to complete questionnaires about their children's health regularly, including comprehensive questions at a mean age of 5.4 and 6.4 years. Parents and children were asked to complete questionnaires about oral health at a mean age of age 7.5 and 10.7 years, and a subset of young people completed a questionnaire at age 17.7 years as described in chapter 2. Both clinical and questionnaire derived data were included in this analysis, with priority given to clinical data where available, as described in the case definition section below. Ethical approval for the ALSPAC study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committee. Full details of ethical approval policies and supporting documentation are available online (http://www.bristol.ac.uk/alspac/researchers/research-ethics/). Approval to undertake analysis of caries traits was granted by the ALSPAC executive committee (project number B2356).

COPSAC cohorts

The Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) centre included analysis in two population based longitudinal birth cohorts in eastern Denmark. COPSAC2000 recruited pregnant women with a history of asthma between 1998 and 2001[479]. Children who developed wheeze in early life were considered for enrollment in a nested randomized trial for asthma prevention. COPSAC2010 recruited pregnant women between 2008 and 2010 and was not selected on asthma status. Both COPSAC2000 and COPSAC2010 studies included regular clinical follow up. Within Denmark clinical dental

assessment is routinely offered to children and adolescents until the age of 18 years and summary data from these examinations are stored in a national register, similar to the Swedish registers described in chapter 2. These data were obtained via index linkage for participants of COPSAC2000 and COPSAC2010. Genetic and phenotypic data from both cohorts were combined and used to perform GWAS in a single dataset. The COPSAC2000 cohort was approved by the Regional Scientific Ethical Committee for Copenhagen and Frederiksberg (KF 01-289/96) and the Danish Data Protection Agency (2008-41-1574). The 2010 cohort (COPSAC2010) was approved by the Danish Ethics Committee (H-B-2008-093) and the Danish Data Protection Agency (2008-41-2599).

Danish National Birth Cohort

The Danish National Birth Cohort (DNBC) is a longitudinal birth cohort which recruited women in mid-pregnancy from 1996 onwards[480]. For this analysis, index linkage was performed to obtain childhood dental records for mothers participating in DNBC. As with the COPSAC studies, these data were originally obtained by a qualified dentist and included surface level dental charting, which was used to derive caries-free or caries-affected status. The DNBC study of caries was approved by the Scientific Ethics Committee for the Capital City Region (Copenhagen), the Danish Data Protection Agency, and the DNBC steering committee.

Generation R study

The Generation R study (GENR) recruited women in early pregnancy with expected delivery dates between 2002 and 2006 living in the city of Rotterdam, the Netherlands. The cohort is multi-ethnic with representation from several non-European ethnic groups. Follow-up has included clinical assessment visits and questionnaires and is ongoing[481]. Intra-oral photography was performed as a part of their study protocol, with surface level coding of these photographs produced by a dental examiner (a specialist in paediatric dentistry)[482]. Analysis in GENR included a) a multi-ethnic association study including all individuals with genetic and phenotypic data[483] and b) analysis including only individuals of European ancestry. The GENR study design and specific data acquisition were approved by the Medical Ethical Committee of the Erasmus University Medical Center, Rotterdam, the Netherlands (MEC-2007-413).

The GENEVA consortium

The Gene, Environment Association Studies (GENEVA) consortium is a group of studies which undertake coordinated analysis across several phenotypes[251] and was first introduced in chapter 3. Within GENEVA, the Center for Oral Health Research in rural Appalachia, West Virginia and Pennsylvania, USA (COHRA), the Iowa Fluoride Study in Iowa, USA (IFS) and the Iowa Head Start study in Iowa, USA (IHS) participated in analysis of dental caries traits in children[67]. COHRA recruited families with at least one child aged between 1 and 18 years of age, with dental examination performed at baseline[252]. IFS recruited mothers and newborn infants in Iowa between 1992 and 1995 with a focus on longitudinal fluoride exposures and dental and bone health outcomes. Clinical dental examination in IFS was performed by trained assessors at age 5, 9 13 and 17 years[484]. IHS recruited children participating in an early childhood education program which included a one-time clinical dental examination[71]. Each participating site in the GENEVA consortium caries analysis received approval from the local university institutional review board (federal wide assurance number for GENEVA caries project: FWA00006790). Within the COHRA arm local approval was provided by the University of Pittsburgh (020703/0506048) and West Virginia University (15620B), whilst the IFS and IHS arms received local approval from the University of Iowa's Institutional Review Board.

GINIplus and LISA

The "German Infant study on the influence of Nutrition Intervention plus air pollution and genetics on allergy development" (GINIplus) is a multi-centre prospective birth cohort study which has an observational and interventional arm. The interventional arm conducted a nutritional intervention during the first four months of life which aimed to investigate the efficacy of a formula nutrition product for infants with a family history of allergy[485]. The study recruited newborn infants with and without family history of allergy in the Munich and Wesel areas, Germany between 1995 and 1998 [486,487]. The "Lifestyle-related factors, Immune System and the development of Allergies in East and West Germany" study (LISA) is a longitudinal birth cohort which recruited between 1997 and 1999 across four sites in Germany[486,488]. For participants living in the Munich area of southern Germany, follow up used similar protocols in both GINIplus and LISA, with questionnaire and clinic data including clinical dental examination by trained examiners at age 10 and 15 years. Analysis for caries in GINIplus and LISA was therefore performed across both studies for participants

at the Munich study centre. The Munich dental examinations for the GINIplus and LISA studies were approved by the ethics committee of the Bavarian Board of Physicians (reference numbers for 10 year follow up: 05100 for GINIplus and 07098 for LISA, for 15 year follow up 10090 for GINIplus, 12067 for LISA).

The PANIC study

The Physical Activity and Nutrition in Children (PANIC) Study is an ongoing controlled single-centre study in Finland. Between 2008 and 2009 PANIC recruited 512 children aged between 6 and 8 years who were living in the city of Kuopio, Finland[489]. The study was established to investigate whether a physical activity and dietary intervention in childhood has long term effects on adiposity, type 2 diabetes, atherosclerotic cardiovascular diseases, musculo-skeletal diseases, psychiatric disorders and oral health problems. The study collected data retrospectively between pregnancy and recruitment and includes prospective follow up between recruitment and adolescence. Clinical dental examinations were performed by a qualified dentist with tooth level charting. The PANIC study protocol was approved by the Research Ethics Committee of the Hospital District of Northern Savo.

The YFS

The Cardiovascular Risk in Young Finns Study (YFS) is a multi-centre investigation which aimed to understand the determinants of cardiovascular risk factors in young people in Finland. The study recruited participants who were aged 3, 6, 9, 12 ,15 and 18 years old in 1980. Eligible participants living in specific regions of Finland were identified at random from a national population register and were invited to participate. Regular follow-up has been performed through physical examination and questionnaires[490]. Clinical dental examination was performed by a qualified dentist with tooth level charting. The YFS study protocol was approved by local ethics committees for contributing sites.

The RAINE cohort

The Western Australian Pregnancy Cohort (RAINE) study is a birth cohort which recruited women between 16th and 20th week of pregnancy living in the Perth area of Western Australia. Recruitment occurred between 1989 and 1991 with regular follow up of mothers and their children through research clinics and questionnaires[491]. The presence or absence of dental caries was recorded by a trained assessor following clinical dental examination at the

year 3 clinic follow up. The RAINE study was approved by the University of Western Australia Human Research Ethics Committee.

5.2.2: Genotypes and genetic data quality control.

For children in ALSPAC genotypes were generated using the Illumina HumanHap550 quad chip and called using GenomeStudio (Illumina, San Diego, California). Prior to imputation, variants with > 5% missing calls, PHWE < $1x10^{-6}$ and variants with MAF < 1% were excluded, leaving a total of 526,688 variants which passed QC measures. Samples with an overall missingness rate > 5%, indeterminate X chromosome or excess heterozygosity were excluded. Following this, phasing was performed in conjunction with genetic data from mothers in ALSPAC using the SHAPEIT2[280] algorithm. Samples which clustered outside of the CEU HapMap2[492] population using multidimensional scaling (MDS) of genome-wide identify by state (IBS) pairwise distances were excluded. Imputation used IMPUTE2[279] (v2.2.2) software and the haplotype reference consortium (HRC) v1.0 imputation panel[493]. Finally, analysis was restricted to unrelated children with estimated identity by descent (IBD) of < 0.1.

In COPSAC genotyping used the Illumina Infinium OmniExpress exome chip and genotypes were called using GenomeStudio (Illumina, San Diego, California). Variants with MAF <1%, missing call rates >5% or PHWE < $1x10^{-6}$ were excluded, leaving 601,374 variants. Samples with >3% overall missingness, with mismatch between genetic and reported sex or with allelic heterogeneity more than 3 standard deviations away from the mean value were removed. For pairs of samples with an IBD estimate of > 0.85 the worst called sample was removed. Samples which clustered outside the CEU HapMap3[295](v2) population using MDS of genome-wide IBS pairwise distances were excluded. Phasing used Eagle2[494] algorithms (implemented in the Sanger Imputation Service: URL https://imputation.sanger.ac.uk/). Imputation used positional Burrows-Wheeler transform methods[495] and the HRC (v1.0) imputation panel, implemented in the Sanger Imputation Service.

In DNBC genotypes were generated using the Illumina Human610-Quad and Illumina Human660W-Quad arrays and called using BeadStudio (Illumina, Dan Diego, California). Variants with MAF <1%, PHWE < $1x10^{-6}$ or >2% missingness were excluded. Samples with > 5% overall missingness, with discordance between genetic and reported sex and low values of heterozygosity were excluded. Samples with an IBD estimate of >0.1875 were excluded.

Samples with large deviation (defined as more than 6 standard deviations) away from the CEU HapMap2 cluster formed in PCA analysis were excluded. A total of 502,119 variants passed quality control and were carried forward to phasing, which used SHAPEIT algorithms[275] implemented in the Michigan imputation server (https://imputationserver.sph.umich.edu/index.html)[278]. Imputation used the HRC panel and minimac (an implementation of the MaCH[281] method) and was performed using the Michigan imputation server.

In GENR genotypes were generated using the Illumina HumanHap610 and Illumina HumanHap660 arrays and called using GenomeStudio (Illumina, San Diego, California). Variants with > 2% missingness, PHWE less than $1x10^{-6}$ or MAF <1% were excluded. Samples with >2.5% missingness, discordance between reported and genetic sex and high heterozygosity (defined as greater than 4 standard deviations away from the population mean) were excluded. For the European ancestry GWAS analysis, samples that clustered more than 4 standard deviations outside the first 4 principal components from the CEU HapMap2 population using MDS were excluded. The remaining samples were then used to generate 10 genetic principal components which were used as covariates in the European ancestry GWAS. Phasing used a total of 459,878 variants and was carried out using MaCH algorithms[281]. Imputation used the 1000 genomes (1KG) reference panel (phase 1, version 3) and the Minimac method (an implementation of MaCH).

Genotyping in GENEVA used the Illumina Human610 Quad(v1B) array and genotypes were called using BeadStudio (version 3.3.7, Illumina, San Diego, California). Sites were removed if there was any suggestion of technical failure of genotyping, if cluster separation was less than 0.2, if sites appeared monomorphic, if the call rate was < 95%, if there were more than 25 Mendelian errors, if PHWE was < 0.0001 or if there was more than 1 discordant genotype call in 90 duplicate pairs. Samples were excluded if the identity was unclear, if the sample appeared duplicated or if the overall missing call rate was greater than 5%. All putative biological relationships in the study were confirmed using genetic data, and a maximum set of unrelated participants was defined and used for GWAS analysis. Phasing used 575,837 variants and was carried out using SHAPEIT[275]. Imputation used the 1KG (phase 1, version 3) reference panel and IMPUTE2[279] algorithms.

In GINI/LISA genetic data were obtained using the Affymetrix 5.0 and Affymetrix 6.0 arrays and called using BRLMM-P (Affymetrix, Santa Clara, California) or Birdseed2[274] algorithms. Variants with a call rate below 95%, MAF < 1% or PHWE < 0.00001 were excluding, leaving 359,648 variants which were carried forward. Samples were excluded if they had an overall call rate below 95%, heterozygosity values greater than 4 standard deviations either side of the mean or had mismatch between reported and genetic sex. MDS of pairwise IBS estimates was performed to identify and exclude outlying samples with excessively high similarity estimates. Phasing used SHAPEITV2[280] algorithms and imputation was performed using IMPUTE2[279] (v2.3.7) and the 1KG phase 1 version 3 panel.

Genotypes in PANIC were generated using the Illumina HumanCoreExome12(V1.0B) chip and called using GenomeStudio (Illumina, San Diego, California). Heterozygous haploid genotypes were set to missing and sites were excluded if the call rate was less than 95%, if MAF was <1% or if PHWE was < $1x10^{-6}$, leaving 390,669 variants. Samples with poor call rates (<95%), excess heterozygosity or mismatch between reported and genetic sex were excluded. Outlying samples in MDS analysis and samples which clustered more than 3 standard deviations away using the first 4 PCs were excluded. Phasing was carried out using SHAPEITV2[280] and imputation used IMPUTE2[279] (v2.3.7) and the 1KG phase 1 version 3 reference panel.

Genotyping in RAINE used the Illumina Human660W Quad BeadChip array and calls were generated using BeadStudio (Illumina, San Diego, California). Variants with >5% missingness, PHWE < $5.7x10^{-7}$ or MAF < 1% were excluded. Samples with an overall call rate <97%, discrepancy between reported and genetic sex and excess heterozygosity (defined as an h score greater than 0.3) were excluded. Analysis was restricted to unrelated participants (defined as a pi statistic < 0.1875). Phasing was performed on a total of 535,632 variants using the MaCH[281] method, and imputation used the 1KG (phase 1 v3) reference panel and Minimac software.

In YFS genetic data were generated using the Illumina670k array and called using Illuminus[496]. Sites with a call rate <95%, MAF <1 or PHWE < $1x10^{-6}$ were excluded, leaving 546,677 variants. Samples with an overall call rate < 95%, with excess heterozygosity, duplicate samples, and with mismatch between reported and genetic sex were excluded.

Analysis was restricted to unrelated participants and samples which clustered outside the largest group in MDS analysis were excluded. Phasing used SHAPEIT[275] algorithms and imputation used IMPUTE2[279] (v2.2.2) and the 1KG (phase 1 v3) reference panel.

The imputed genetic data from these studies includes a mixture of imputation to the HRC and 1KG panels. Where available, the HRC panel was chosen as this offers higher imputation quality on average than smaller panels at the same variant[493]. Ideally all studies would have used the HRC panel, however this was not possible within the timescale of this project. The average improvement in imputation quality for HRC compared to other panels is most marked at low frequency variants with minor allele frequency (MAF) of 0.01 or lower, while there is little difference in panels for variants with MAF of 0.05 or higher. Although the study had a theoretical MAF cut-off point of 0.005, in practice the study aimed to explore common genetic variation where there was enough power to detect genetic effects with plausible effect sizes, where the difference between these panels is minimal. For that reason, both imputation panels were combined in meta-analysis.

Additional central QC

Following tests for genetic association, further QC was performed centrally using the EasyQC R package[284] and accompanying 1KG phase1 v3 reference data or HRC reference data[284] prior to meta-analysis. Variants were dropped which had a per-results file minor allele count (MAC) of 6 or lower, a site-specific sample size of 30 or lower, or an imputation quality score (impute INFO) of less than 0.4. Sites which reported effect and non-effect alleles other than those reported in the reference data were dropped. Following meta-analysis, sites with a weighted minor allele frequency (MAF) of less than 0.005 were dropped, along with variants present in less than 50% of the total sample.

5.2.3: Phenotypic definitions

In line with the decision to treat caries as a person-level disease, all studies used a caries-affected or caries-free approach to classification.

In children aged 2.50 years to 5.99 years any individual with 1 or more decayed or filled tooth was classified as caries affected, with all remaining individuals classified as caries-unaffected. In children aged 6.00 years to 11.99 years of age parallel definitions were created for the primary dentition and permanent dentition respectively. Any individual with at least 1

decayed or filled primary tooth was classified as caries affected for primary teeth, while all remaining participants were classified as unaffected. In parallel, any individual with at least 1 decayed or filled permanent tooth was classified as caries affected for permanent teeth, while all remaining individuals were classified as unaffected. In children and adolescents aged 12.00 to 17.99 years of age any individual with 1 or more decayed or filled tooth or tooth surface (excluding third molar teeth) was classified as caries affected, with remaining individuals classified as unaffected. Analysis was conducted in cross-section, meaning a single participant could only be represented in a single phenotype definition once. Where multiple sources of dental data were available for a single participant within a single phenotypic definition window, the first source of data was selected (reflecting the youngest age at participation) in allocating case status.

For ALSPAC, slightly different criteria were used. The questions asked did not distinguish between primary and permanent teeth. Instead, age at response was used as a proxy to prioritize responses which likely capture caries in either the primary or permanent dentition. Two variables were created, one which prioritized responses from questionnaires before 6.00 years of age (to be combined in meta-analysis with caries in primary teeth) and the other prioritizing responses after 10.00 years of age (to be combined in meta-analysis with caries in permanent teeth). Where clinical data were available in ALSPAC these always took priority over questionnaire data. The final data sweep considered in this analysis targeted adolescents at age 17.5 years. Although the mean age at response was 17.8 years, some participants responded to this after their eighteenth birthday. Data derived from this final questionnaire sweep were not included in the principal meta-analyses but were included in the GCTA heritability analysis.

5.2.4: Statistical approach for single variant association tests

Each cohort preformed GWAS analysis using an additive genetic model. Caries status was modelled against genotype dosage whilst accounting for age at phenotypic assessment, age squared, sex and cryptic relatedness. Sex was accounted for by deriving phenotypic definitions and performing analysis separately within male and female participants, or by including sex as a covariate in association testing. In the GENR study parallel analyses were conducted for participants of European ancestry and the entire study population, using a previously published method[483].

5.2.5: Meta-analysis

Results of GWAS within each study were combined in two principal meta-analyses, representing caries status in primary teeth and caries status in permanent teeth. For primary teeth, parallel meta-analyses were performed, one using results of multi-ethnic analysis in the GENR study and the other using results of European ancestry analysis in the GENR study. The GENR study did not have phenotypic data for permanent teeth, therefore the analysis of permanent teeth contained only studies of European ancestry. Fixed-effects meta-analyses were performed using METAL[293], with study specific estimates of genetic effect (beta) weighted by inverse standard error. Genomic control[120] of input summary statistics was enabled to estimate genomic inflation factor (λ_{GC}) and calibrate input test statistics if λ_{GC} was greater than 1. Tests for heterogeneity in genetic effect were performed using the I^2 statistic[497].

5.2.6: Follow-up analysis

For each principal meta-analysis, inflation in test statistics attributable to inflationary bias from population stratification and polygenic heritability were estimated using linkage disequilibrium score regression (LDSR)[49]. Reference linkage disequilibrium (LD) scores were taken from reference data accompanying the LDSR package, derived in participants of European ancestry in the 1000 genomes project[498]. These data are available online at (https://data.broadinstitute.org/alkesgroup/LDSCORE/).

For comparison, heritability within the ALSPAC study was assessed using the genetic restricted maximum likelihood (GREML) method[499], implemented in the GCTA standalone software package[299]. This aims to estimate heritability attributable to common genetic variation. Analysis used participant level phenotypic data in conjunction with a genetic relatedness matrix estimated from common genetic variants (with MAF> 0.05) present in HapMap3, generated using the –make-grm function of GCTA.

Lead single variants were defined as those with a p value < $5x10^{-8}$ in the meta-analysis. Clumping was performed to identify approximately-independent signals of association using an LD r^2 threshold of 0.2 or lower within a 500 kilobase (kb) window either side of the lead variant. Clumping was performed using PLINK 2.0[500] with LD structure estimated from reference data in the UK10K project[298].

Comparison with other traits was performed using a hypothesis-free cross trait lookup. The lead associated single variant at each independently-associated genomic locus was queried using the SNP lookup function in the MRBase catalogue[501]. This analysis retrieves association statistics for that variant with other traits in GWAS results files. If the target variant was not present in a GWAS results file, then potential proxies (with an LD r^2 of 0.8 or higher in 1KG reference data between the target variant and proxy) were included.

Gene based tests were performed using MAGMA[502], which uses SNP-level summary statistics in conjunction with reference LD data to assess the overall statistical evidence for gene-phenotype association. Reference data was taken from the UK10K project and gene definitions were based on a 50 kilobase window either side of canonical gene start:stop positions. Adjustment for multiple testing used a Bonferroni correction for the number of genes tested.

Transcript level tests for association were performed using the S-PrediXcan method[307], which aims to integrate GWAS results with externally-derived transcriptomic data to infer whether varying gene expression is associated with the GWAS trait. A transcription model trained in whole blood was obtained from the PredictDB data repository (http://predictdb.org) and integrated with meta-analysis summary statistics using the MetaXcan standalone software implementation[503] of the S-PrediXcan method. A Bonferroni adjustment for multiple testing was made using the number of transcript-level tests for association.

Post-hoc power calculations were performed using the Genetic Association Study (GAS) Power Calculator (https://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html, accessed 18/03/19) and the software utility Quanto (v1.2.4)[504].

5.3: Results

5.3.1: Study population

The principal single-variant analysis included children aged between 2.5 and 18.0 years of age (Tables 5.1, 5.2). The lowest prevalence of caries was in the RAINE study (where 6.2% of boys, 7.2% of girls were caries-affected), which was also the youngest study (mean age at participation 3.1 years for boys, 3.1 years for girls). The highest prevalence of caries was in the permanent dentition in the Danish National Birth cohort (93% caries-affected), who were both the oldest cohort included in the principal meta-analysis (mean age at phenotypic assessment 15.8 years), and the cohort whose phenotypic data were collected earliest in time, as these phenotypic data in childhood pre-date enrolment of these participants into the DNBC in 1996-2000 when the participants themselves were pregnant.

Table 5.1: Age and caries prevalence in primary teeth for each participating study (primary teeth)

Study	Phenotype	Male						Female					
		n	mean age (years)	SD age (years)	range age(years)	% cases	n (cases)	n	mean age (years)	SD age (years)	range age(years)	% cases	n (cases)
ALSPAC	Primary_primary clinical	407	5.00	0.46	3.55-5.58	21.87	89	367	5.03	0.43	3.57-5.60	24.25	89
ALSPAC	Probable primary questionnaire	2924	5.23	0.57	3.08-8.58	15.01	439	2767	5.25	0.55	3.17-9.50	14.02	388
YFS	Mixed_primary	145	8.22	1.52	6.00-9.90	80.69	177	175	7.98	1.50	6.00-9.90	79.43	139
GINI/LISA	Mixed_primary	426	10.18	0.20	9.77-11.32	46.00	196	396	10.18	0.22	9.88-11.83	35.00	137
GENEVA	Primary_primary	375	4.57	0.75	2.58-5.99	40.00	150	350	4.56	0.78	2.50-5.97	38.29	134
GENEVA	Mixed_primary	181	8.58	1.88	6.00-11.96	66.30	120	188	8.14	1.80	6.00-11.96	54.79	103
COPSAC	Primary_primary	511	4.86	0.755	2.5-5.99	27.01	138	503	4.85	0.729	2.5-5.99	25.85	130
COPSAC	Mixed_primary	219	10.04	2.135	6.0-11.99	55.25	121	229	10.00	2.097	6.02-11.99	56.33	168
PANIC	Mixed_primary	235	7.91	0.43	6.95-9.44	48.09	113	215	7.88	0.42	6.91-8.92	46.22	99
GenR	Primary_primary (all ethnicities)	729	5.85	0.13	4.87 - 6.00	25.80	118	722	5.85	0.11	4.83-6.00	25.48	184
GenR	Mixed_primary (all ethnicities)	957	6.44	0.53	6.00-8.97	35.63	341	966	6.39	0.48	6.00-8.79	31.99	309
GenR	Primary_primary (European ancestry)	440	5.84	0.12	4.87-6.00	17.27	76	417	5.85	0.12	4.83-6.00	19.18	80
GenR	Mixed_primary (European ancestry)	469	6.33	0.39	2.91-6.00	24.31	114	484	6.31	0.39	6.00-8.72	21.49	104
RAINE	Primary_primary	517	3.11	0.12	2.93-3.95	6.19	32	499	3.11	0.10	2.69-3.99	7.21	36
DNBC	Mixed_primary							4950	7.36	1.15	6.00-11.99	82.06	4062

Details are reproduced from[99]

Table 5.2: Age and caries prevalence in permanent teeth for each participating study (permanent teeth)

Study	Phenotype	Male						Female					
		n	mean age (years)	sd age (years)	Range age (years)	% cases	n (cases)	n	mean age (years)	sd age (years)	range age(years)	% cases	n (cases)
ALSPAC	Probable permanent questionnaire	2881	10.25	1.14	7.59-1.95	42.35	1220	3025	10.36	1.02	7.59-11.99	44.56	1348
ALSPAC	Final questionnaire sweep$_1$	611	17.79	0.43	16.42-19.83	46.32	283	964	17.79	0.42	16.25-19.83	49.90	481
YFS	Mixed_permanent	139	8.38	1.50	6.00-11.90	53.96	75	175	8.20	1.60	6.00-11.90	54.29	95
GINI/LISA	Mixed_permanent	426	10.18	0.20	9.77-11.32	39.00	166	396	10.18	0.22	9.88-11.83	31.00	123
GINI/LISA	Permanent_permanent	418	15.25	0.28	14.80-16.55	40.00	167	389	15.25	0.29	14.84-16.45	38.00	148
GINI/LISA	Permanent_permanent subset$_2$	148	15.31	0.33	14.80-16.55	35.00	52	133	15.25	0.27	14.88-16.23	44.00	59
GENEVA	Mixed_permanent	162	9.11	1.70	6.00-11.96	52.47	85	164	9.05	1.68	6.00-11.96	46.95	77
GENEVA	Permanent_permanent	112	14.54	1.60	12.00-17.89	77.68	87	132	14.61	1.62	12.00-17.97	83.33	110
COPSAC	Mixed_permanent	219	10.04	2.13	6.00-11.99	34.25	75	229	10.00	2.10	6.02-11.99	38.44	88
COPSAC	Permanent_permanent	178	14.29	1.00	12.01-16.82	57.30	102	174	14.29	1.06	12.00-16.80	64.37	112
PANIC	Mixed_permanent	236	7.91	0.43	6.95-9.44	47.88	113	215	7.88	0.42	6.91-8.92	46.22	99
DNBC	Mixed_permanent							4950	11.33	0.59	6.09 - 11.99	84.71	4193
DNBC	Permanent_permanent							5073	15.80	1.26	12.00-17.99	93.30	4733

Details are reproduced from[99]. $_1$This phenotype includes some participants who returned questionnaires after their 18th birthday. This phenotype was used in GCTA heritability analysis to enhance power but was excluded from the principal meta-analysis. $_2$This phenotype includes participants who attended examination at 15 years who had not previously been examined at age 10 years.

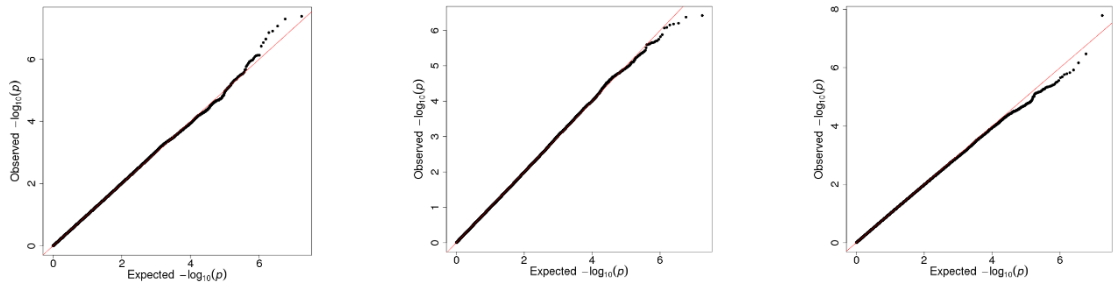
5.3.2: Characteristics of the principal meta-analyses

In children of European ancestry, meta-analysis of caries in primary teeth included 17,037 children (6,922 caries-affected) from 22 results files representing all 9 coordinating centres. After final QC, this meta-analysis included 8,640,819 variants, with mild deflation in summary statistics (λ_{GC} = 0.994)(Figure 5.1a).

Including children from multiple ethnic groups in the GENR study resulted in a slightly larger sample size in meta-analysis (19,003 children, 7,530 affected) and a slightly higher number of variants (8,699,928) passed final QC. There was mild deflation in summary statistics (λ_{GC} = 0.986)(Figure 5.1b).

Analysis of caries in permanent teeth included 13,353 children and adolescents (5,875 caries-affected) from 14 results files representing 7 coordinating centres. The sample size was smaller for permanent teeth than primary teeth because two coordinating centres did not have phenotypic data for permanent teeth (RAINE and GENR) and once centre (COPSAC) only had data for participants in the older birth cohort (COPSAC 2000). The number of variants passing final QC (8,734,121) was similar to meta-analysis of primary teeth, with only mild deflation (λ_{GC} = 0.999)(Figure 5.1c)

Figure 5.1: Quantile quantile plots for GWAS meta-analysis



a b c

a) caries in primary teeth (European ancestry analysis), b) caries in primary teeth (multi-ethnic analysis) c) caries in permanent teeth (European ancestry).

Using the LDSR method, heritability estimates were near-null for all three primary meta-analyses, with wide confidence intervals (Table 5.3). By contrast, heritability estimates using the GREML method in ALSPAC were considerably higher but imprecisely-estimated, with estimates of 0.28 (95% CI 0.09 0.48) for caries in primary teeth and 0.17 (95% CI 0.02:0.31) for caries in permanent teeth.

Table 5.3: Heritability estimates for caries in the primary and permanent dentition.

Phenotype	Method		Estimated h^2 (95% CI)	N
Caries in primary teeth	GCTA GREML		0.28 (0.09:0.48)	7,230
	LDSR	All participants	0.01 (0.00:0.06)	19,003
		European ancestry	0.01 (0.00:0.07)	17,036
Caries in permanent teeth	GCTA GREML		0.17 (0.02:0.31)	6,657
	LDSR		0.06 (0.00:0.12)	13,353

Table reproduced from[99]

5.3.3: Lead single-variant results

In meta-analysis there was therefore little evidence for polygenic association signal at an aggregate, genome-wide level. Likewise, at a single-variant level there were few variants which met the pre-defined threshold for genome-wide significance.

For caries in primary teeth, the strongest single-variant association signal was seen in the European ancestry meta-analysis and a region of 2p25. The lead variant was rs1594318, a common variant (EAF 0.60 for C allele) with modest effects on odds of dental caries (OR 0.85; 95% CI 0.80, 0.90, P =4.1×10^{-8}). Using the dbSNP database (https://www.ncbi.nlm.nih.gov/snp) rs1594318 is annotated as an intronic variant within *ALLC*, a protein-coding gene encoding the enzyme allantoicase which has functions in uric acid degradation and has not previously been reported as a risk locus for caries. No other variants passed genome-wide significance. A regional association plot for the *ALLC* locus is shown in Figure 5.2a.

In meta-analysis of caries in primary teeth which combined participants of all ancestries the genetic effect at rs1594318 was estimated with slightly greater precision but suggested a marginally smaller effect size (OR 0.87; 95% CI 0.82, 0.91), meaning this variant no longer met genome-wide significance (P=3.8×10^{-7}). Taken together, these two meta-analyses

provide some evidence for association at the *ALLC* locus, but this is not considered irrefutable. To help guide inference at this locus, genetic effect estimates from each study were contrasted, showing good consistency (Figure 5.3). While this consistency might be interpreted as supporting evidence for association, it is also likely that, given the modest power of individual studies, the procedure for selecting this variant based on meta-analysis statistics has already filtered out variants with inconsistent effect sizes in different studies and this effect estimate will be prone to winner's curse[505].

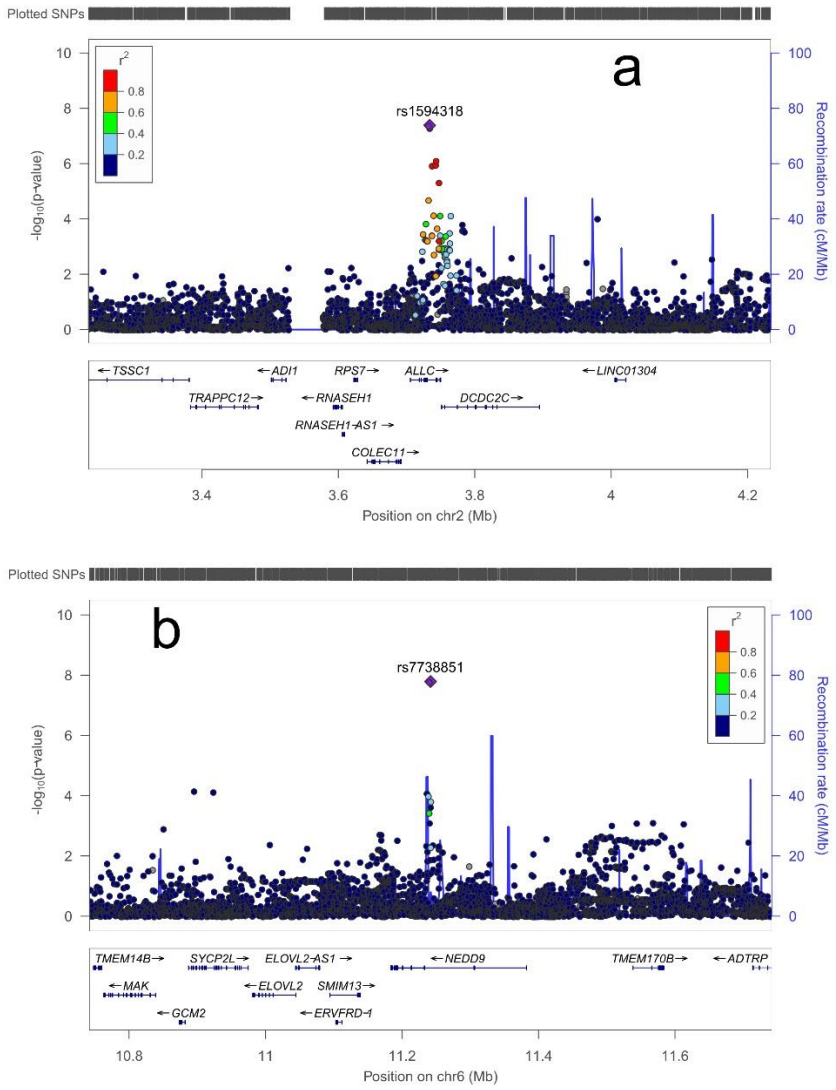
In meta-analysis for caries in permanent teeth the lead single variant was rs7738851, a common variant (EAF=0.85 for A allele) in the 6p24 locus (Figure 3.2b). The estimated effect in meta-analysis was an odds ratio of 1.28 per A allele (95% CI 1.18, 1.40) and there was good consistency in effect estimates between studies (Figure 5.4). rs7738851 is annotated as an intronic variant within Neural Precursor Cell Expressed, Developmentally Down-Regulated protein 9 (*NEDD9*), a scaffolding protein with domains which docks multiple signal transduction proteins.

Table 5.4: Lead single variant for each meta-analysis

Phenotype	Variant	Position	Effect allele: EAF	Beta (SE)	Odds ratio	P value	N	I^2	Phet[1]	Ann[2]
Caries in primary teeth (European ancestry analysis)	rs1594318	chr2 :3733944	C:0.60	-0.17 (0.030)	0.85	$4.1x10^{-8}$	16,994	0.0	0.69	Intronic, *ALLC*
Caries in primary teeth (multi-ethnic analysis)	rs1594318	chr2: 3733944	C:0.60	-0.14 (0.028)	0.87	$3.8x10^{-7}$	18,960	0.0	0.61	Intronic, *ALLC*
Caries in permanent teeth	rs7738851	chr6: :11241788	A:0.85	0.25 (0.044)	1.28	$1.6x10^{-8}$	13,353	13.3	0.20	Intronic, *NEDD9*

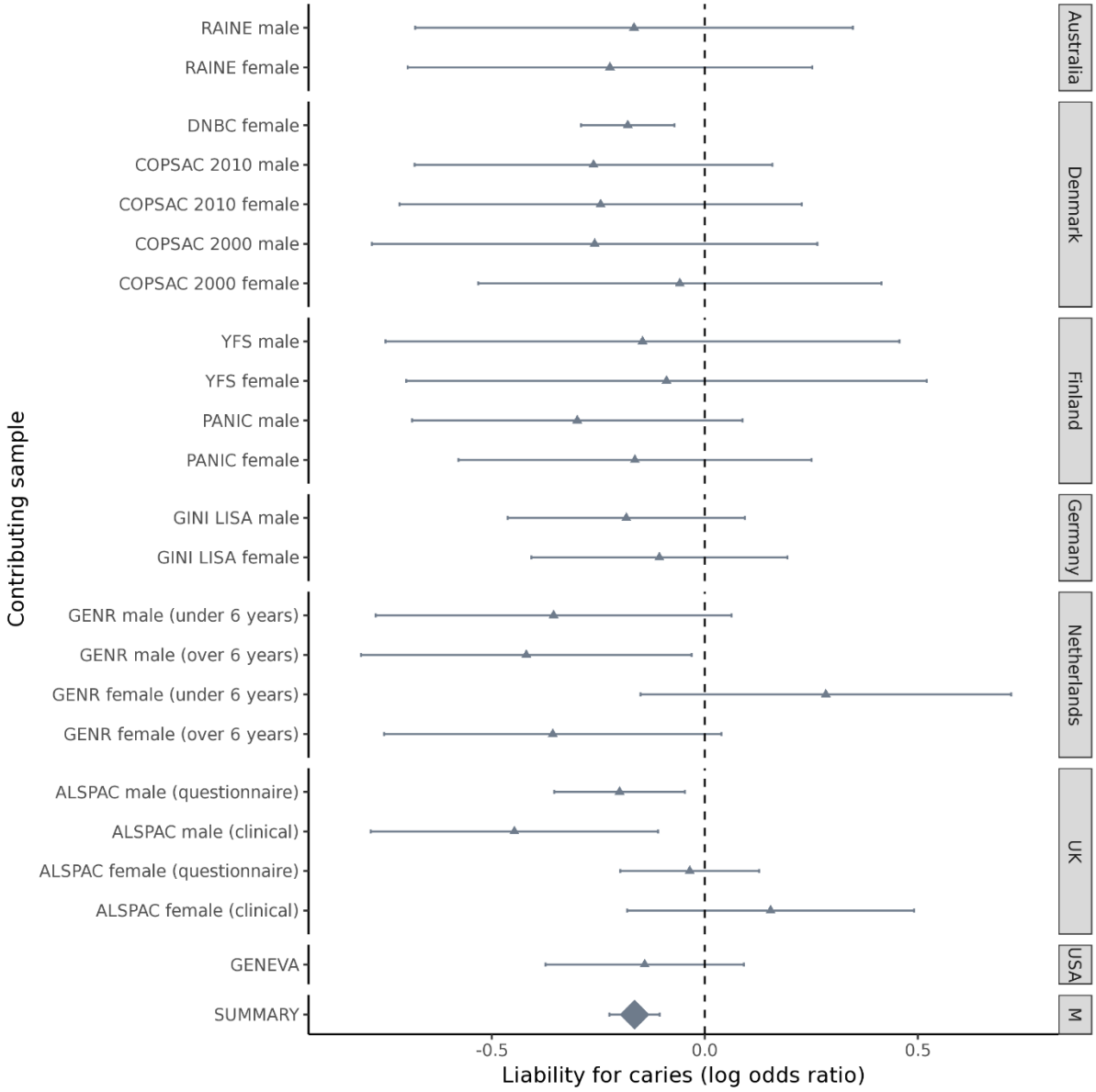
[1]The Phet column contains p values testing the null hypothesis of no heterogeneity in genetic effect estimate in the contributing studies. [2]SNP annotation in the sbSNP database. Table reproduced from[99]

Figure 5.2: Regional association plots for lead associated variants



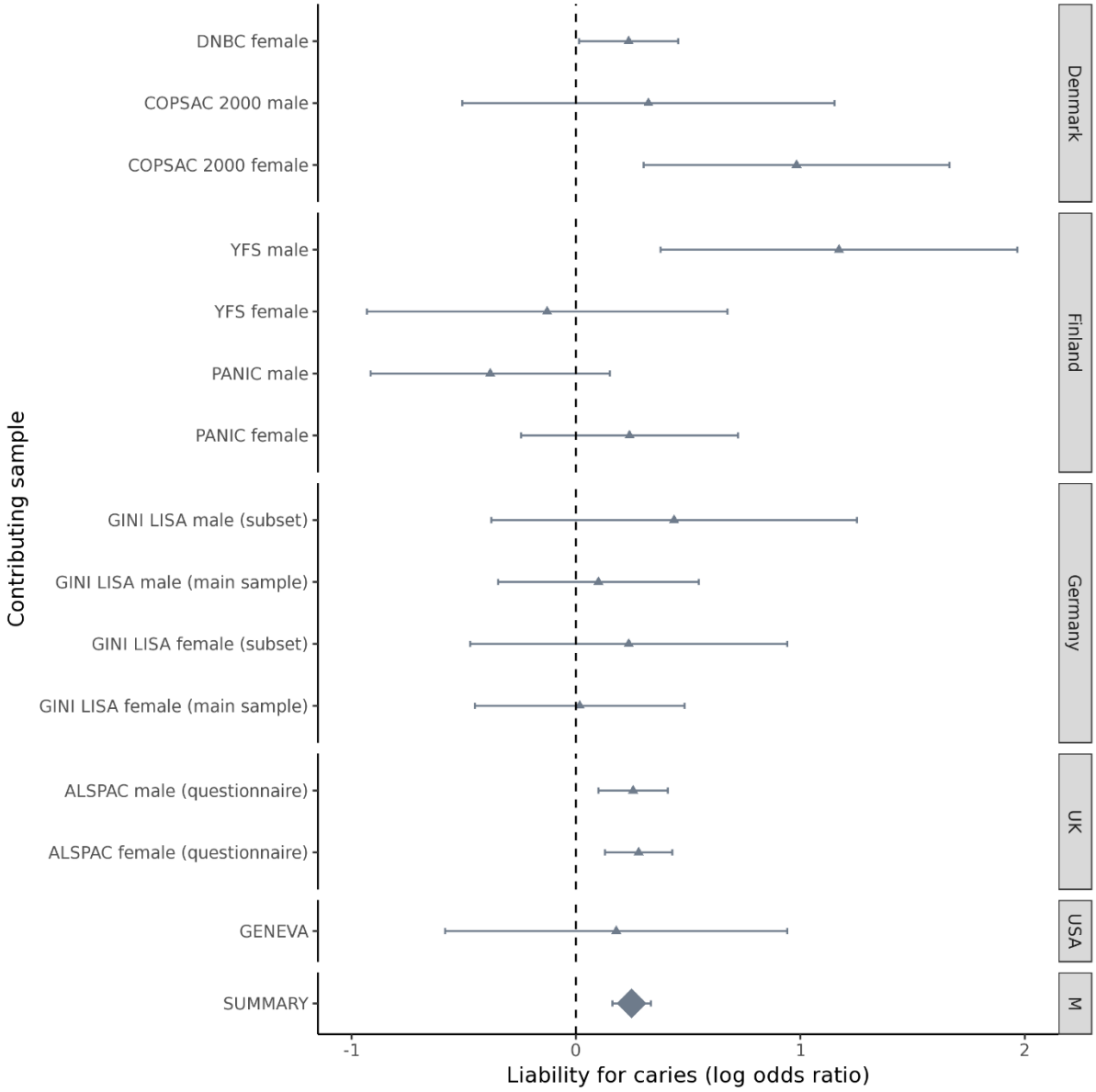
Regional association plots for a) rs1594318 (caries in primary teeth, European ancestry meta-analysis) and b) rs7738851 (caries in permanent teeth). Plots were produced using the LocusZoom resource[506] at (locuszoom.org).

Figure 5.3: Forrest plot of genetic effect estimates for rs1594318



Triangles represent the point estimates of genetic effect in each contributing study. Error bars represent the 95% confidence intervals for the estimated genetic effect. Studies are grouped by geographical location (right hand boxes). The 'summary' represents the overall estimate and confidence intervals from fixed-effects meta-analysis of all contributing studies.

Figure 5.4: Forrest plot of genetic effect estimates for rs7738851



Triangles represent the point estimates of genetic effect in each contributing study. Error bars represent the 95% confidence intervals for the estimated genetic effect. Studies are grouped by geographical location (right hand boxes). The 'summary' represents the overall estimate and confidence intervals from fixed-effects meta-analysis of all contributing studies.

5.3.4: Cross-trait comparisons.

Genome-wide mean chi squared was low for all 3 meta-analysis results files, meaning it was not possible to use the LDSR method used earlier in this dissertation to test for genome-wide genetic correlation between caries and other traits.

As an alternative way to examine genetic overlap with other traits, lead associated variants were compared to published findings in the GWAS literature. Association statistics for 885 published GWAS studies were retrieved for rs1594318 or proxies with $r^2 > 0.8$. None of these traits were associated with rs1594318 at a higher level of evidence than that expected by chance alone (using a Bonferroni-corrected alpha of 0.05), suggesting that rs1594318 does not have widespread functional relevance for human health. Similarly, for rs7738851 there was good representation of this variant in published GWAS studies (662 traits) but none of these traits were associated at a Bonferroni-corrected threshold.

5.3.5: Gene based tests and gene-set approaches

Gene based testing implemented in MAGMA used a Bonferroni correction to adjust for the number of genes tested. Using data from the caries in primary teeth (European ancestry) meta-analysis, there was evidence for association at 3 genes lying within a region of 7q35, namely TRPM8 Channel Associated Factor 1 (*TCAF1),* a ion channel binding protein ($p=1.9x10^{-6}$), Olfactory Receptor Family 2 Subfamily F Member 2 (*OR2F2)* and Olfactory Receptor Family 2 Subfamily F Member 1 (*OR2F1),* a related pair of proteins implicated in recognition and transduction of odour ($P=1.6x10^{-6}$ and $P=1.3x10^{-6}$ respectively).

Using data from the caries in permanent teeth meta-analysis there was no evidence for association passing the multiple testing threshold.

Using predicted transcriptome data, no transcripts were associated with caries in any of the 3 meta-analyses after a Bonferroni correction for multiple testing. The strongest suggestive evidence was seen for CDK5 regulatory subunit-associated protein 3 (*CDK5RAP3),* where increased transcription in whole blood was predicted to increase odds of caries in permanent teeth ($P=3.9x10^{-5}$). *CDK5RAP3* is a regulator of cell proliferation, with a physiological role in neuronal differentiation and oncogenic potential if dysregulated[507,508].

There were too few independently associated loci to apply gene-set tests using the approach described in the methods protocol.

5.3.6: Post-hoc power calculations

For caries in primary teeth in participants of European ancestry (17,037 individuals, 6,922 caries-affected, prevalence 40.6%) the study had 80% power to detect common variants (MAF 0.05 or greater) with an odds ratio of 1.37 or greater at genome-wide significance ($P=5.0x10^{-8}$). Under best-case scenario of MAF between 0.4 and 0.5, the study had power to detect more subtle genetic effects with an odds ratio in the range of 1.13 to 1.15. The absence of strong signals of association in single variant results may therefore imply the genetic architecture[51] of caries in primary teeth is highly polygenic with multiple modest genetic effects rather than an oligogenic architecture with strong individual genotype effects. Other interpretations for the limited single variant discovery yield for primary teeth are discussed later in this chapter.

For caries in permanent teeth (13,353 participants, 5,875 caries-affected, prevalence 44.0%) the statistical power was slightly lower than for caries in primary teeth. The study had 80% power to detect common variants with an odds ratio of 1.43 or greater at genome-wide significance. In ideal circumstances (MAF between 0.4 and 0.5) effects as small as an odds ratio of 1.15 to 1.17 were theoretically detectable with 80% power.

Genome-wide summary statistics from meta-analysis, full results of gene-based tests and scripts used in analysis are publicly available from the University of Bristol Research Data Repository at (https://data.bris.ac.uk/data/dataset/pkqcnil6e9ju2nyreblt3mvwf).

5.4: Discussion

To date, GWAS for dental caries in children and adolescents have included small samples with no prior large-scale consortium-based meta-analysis. This study marks a step change in approach and analysis scale, but results in an incremental rather than transformative discovery yield, illustrating the ongoing need for collection of dental data in cohorts of young people. Despite this, the results of paediatric analysis are a relevant contribution to this thesis. First, the heritability estimates in ALSPAC and apparent stability of genetic effect estimates at lead associated loci when using dental data gathered in different ways provide further evidence for the utility of index-linked and questionnaire-derived data in dental epidemiology. Next, the failure to identify common variants with large effect sizes in this population and fact that the most compelling candidate SNPS are not located within previously studied candidate genes is consistent with the results of chapter 3. Together, these chapters suggest that, contrary to interpretation of published candidate gene studies[35], common variants with large effect sizes are not tolerated for dental caries. It also seems likely that, similar to other traits[50], the most relevant biological variation is outside loci which were studied in the candidate gene era. Results of this chapter therefore add further weight to the argument made in chapter 3 that dental caries is a trait with polygenic genetic architecture which is best explored through genome-wide methods and may have been under recent evolutionary selection.

At a single-variant level, the two novel variants identified in this analysis are potentially interesting and warrant additional investigation to confirm association in an independent population and evaluate the potential mechanisms. For primary teeth, there was evidence of association with rs1594318 which attenuated in the multi-ethnic meta-analysis. This might reflect the variable allele frequency of rs1594318, where the G allele has a frequency of 0.24 in Asian population and 0.42 in European ancestry populations in 1KGP reference data. The allantoicase protein is no longer thought to have meaningful level enzymatic activity in vertebrates, and studies in mice suggest the lack of function is because allantoicase is expressed at low levels and has poor substrate affinity, rather than a complete loss of function[509]. Given the (apparent) limited biological relevance of the allantioicase protein in vertebrates there are few studies investigating the phenotypic consequences of *ALLC* regulation, and there are currently no plans to undertake phenotyping for mouse knockout strains for this gene (source: http://www.mousephenotype.org/data/genes/MGI:2136971 ,

accessed 19/03/19). Intriguingly, there is some evidence that common variants within *ALLC* are associated with response to asthma treatment in humans[510], although no mechanisms has yet been proposed and it is possible that both these polymorphisms and the caries-associated rs1594318 tag a functional gene elsewhere in the locus.

There was evidence for association between rs7738851 (intronic, within *NEDD9)* and caries in permanent teeth. *NEDD9* is reported to mediate integrin-initiated signal transduction pathways and is conserved from gnathostomes into mammals [511,512]. In physiological growth and development *NEDD9* appears to be an important binding site for signal transduction proteins which regulate neuronal growth, differentiation and migration, and impaired functions of *NEDD9* may have effects on these processes[511,513-517]. Neuronal growth and development is likely tied to odontogenesis through neural crest related processes, where disruption of neural crest signalling results in enamel and dentine defects[518,519]. Thus, one hypothesis is that the risk-allele of rs7738851 is associated with caries in permanent teeth by a pathway of altered neural crest signalling by *NEDD9* [515] leading to subtle defects in enamel and dentine formation, but this hypothesis is not tested in this study.

While these two findings may help to understand disease, the overall discovery yield is modest, compared to the results included in chapter 3. The most simple explanation (supported by the power calculations) is that the study is simply underpowered to detect the (likely subtle) effects of single common variation on a highly complex, polygenic[35] and only moderately heritable disease like dental caries, analogous to other comparable complex traits where individual genetic variants confer small effect sizes[520]. To contextualize the likely statistical power of this chapter, no single SNP identified in chapter 3 had a large enough estimated effect on dentures to be detected here for caries, even if the true effect size was the same for both traits, suggesting that the differences in discovery yield between chapters 3 and 5 are predominantly related to statistical power. Early GWAS studies for traits such as type 2 diabetes yielded few association signals and with odds ratios similar for those reported here for caries[521] while a study only 8 years after this early report claims 143 risk variants including both rare variation with large effects and common variation with more subtle effects[522]. Here the progress has been achieved not through better phenotypic resolution or reduced error, nor through radically different statistical methods, but simply by persevering with data collection and incremental improvements in sample size. To mirror this progress in

paediatric dental traits we will need to collect dental data across the whole life course including in children and adolescents, an idea which is explored further in chapter 6.

This explanation based on statistical power is satisfactory for describing the single variant results but less satisfying when thinking about heritability, where the meta-analysis estimates were lower than anticipated from the literature[59]. In isolation, low statistical power does not explain why the heritability estimates from meta-analysis were so low, as the point estimate should (theoretically) not be tied to statistical power. One possible interpretation is that the use of a relatively crude phenotype and inclusion of questionnaire data may have biased heritability estimates towards the null. In the most extreme scenario, where questionnaire responses measured only noise with no ability to distinguish between participants with and without dental caries, this noise would not be expected to correlate with genotype and would yield a near-zero heritability estimate. This interpretation is not compatible with the results of chapter 2, but to explore this argument further, GCTA-GREML estimates were obtained in the ALSPAC sample. In a single sample, these relatively crude measures (including questionnaire-derived data) are nevertheless moderately heritable, supporting the argument in chapter 2 that these questionnaires have face validity, and the heritability estimate only drops to very low values when combining results across studies. Likewise, the argument that the meta-analysis estimate is low because SNP-based estimators produce lower estimates than twin or family based studies[523] does not account for the discrepancy between the GCTA-GREML estimate and the LDSR estimate.

It is possible to imagine scenarios where the heritability estimates from fixed effects meta-analysis (i.e based on genetic risk factors which are common across all populations and have equal importance) are not equivalent to the weighted average of within-study heritability estimates. Allowing the relative importance of any given genetic risk factor (and its association regression coefficient) to vary between populations would achieve this result, as the meta-analysis captures (h^2 common) while the weighted average of within-study estimates captures (h^2 common + h^2 population-specific). For example, if genetic penetrance varied across strata of an environmental risk factor which varied between populations then the meta-analysis estimate would be biased towards the null. There are potential features reported in the literate which might modify genetic effects, including gene-sex interactions[57], fluoride exposure as a modifier of genetic effect[524] and, indirectly, variability in the relevance

of microbiome groups in different populations[525] implies that genetic effects mediated through the oral microbiome will be population-specific in magnitude. If so, the heritability estimates might be interpreted to mean that much of the genetic main effect estimated in individual studies is in truth subject to gene-environment interaction, an explanation which appears plausible given that caries in childhood cannot develop without the right environmental factors and an explanation which may also help explain the lack of consistent associations seen in the literature to date.

This could be explored by modelling gene x environment interactions within studies with bivariate meta-analysis of the genetic main effect and gene x environment interaction term. In practical terms this is unlikely to be viable in small studies, as there must be sound statistical evidence for a main effect before the interaction test becomes meaningful and may be challenging in large consortia as consistently-ascertained data on strata of the exposures described in the paragraph above may not be available. For some exposures (such as fluoride in water supplies) there may be little variation within studies, and the relevant environmental variation may occur between studies. Here, a meta-regression of genetic main effects incorporating a study-level term for fluoridated proportion may be a more viable approach but one which once again will require well-powered studies with contrasting environmental features.

Other sources of heterogeneity may be present in these meta-analysis results but may be less relevant. Sources of variation between study centres (such as nutrition, oral hygiene and access to dental caries) do not necessarily result in heterogeneous genetic effects, as this variability might be uncorrelated with the genetic main effect. In testing this, there is little evidence of heterogeneity at the top associated loci however the I^2 statistic has important limitations and may not detect heterogeneity when there are few contribution studies in meta-analysis and where within-study genetic effect estimates are imprecice[526]. The inability to model interactions, dissect population-specific effects and explore heterogeneity collectively create an argument for collection of paediatric dental data in well-powered studies with comprehensive health and lifestyle data.

5.5: Commentary

This chapter used genetic data to investigate the molecular aetiology of caries in children and adolescents. Although the sample size was larger than any previous genetic investigation of this trait in paediatric populations, there was only statistical power to identify association at common genetic variants with large effects (i.e. greater than those seen in chapter 3). Under these analytical circumstances, the study identified few association signals. Taken in conjunction with chapter 3, the results of this chapter therefore suggest that caries is a polygenic trait, with single genetic variants conferring small effect sizes. Identifying the effects of these variants on caries will require larger studies with paediatric populations. Questionnaire data in ALSPAC and index-linked records in DNBC yielded similar estimates of genetic effect at top associated variants as dental data obtained as part of the study protocol in other cohorts. Index linkage or systematic questionnaire collection from childhood onward may therefore be one way to achieve the larger sample sizes needed. In the next chapter I draw on the whole thesis to highlight common themes and suggest areas for future research, including the need for investment in paediatric studies in dentistry.

Chapter 6: Future directions

This dissertation used a series of illustrative analyses to test an overall hypothesis that genetic data can be used to learn more about the causes and consequences of dental caries and periodontitis. This chapter assimilates recurrent themes seen through this series to reflect on overall discussion items and suggest some future directions for dental epidemiology.

6.1: Overall discussion and inference

The utility of index-linked and self-reported measures of dental diseases

Datasets for dental epidemiology could be created using bespoke research cohorts specifically set up to investigate dental diseases, or by bringing dental data into well designed epidemiological cohort studies. New dental disease cohorts would require collection of health, lifestyle, genetic and molecular data from scratch and would therefore require large investment before yielding datasets which can be used in modern epidemiolocal methods. These dental disease cohorts might be subject to selection bias and fail to represent groups of participants who are unable to attend for an elective dental examination. Finally, continuing to study dental diseases in isolation from other complex diseases may not bring about the shift in perception needed to place dental diseases into a new context as an integral part of human health and wellbeing. For these reasons, bringing dental data into well-designed existing epidemiological cohort studies provides practical advantages for resource allocation, theoretical benefits for assessing and minimizing the impact of selection bias and ideological benefits for repositioning the role of dental epidemiology.

One way to enrich existing studies with dental data involves index linkage to records from primary dental care. Dental examination for clinical assessment needs to include comprehensive and accurate information to facilitate the correct diagnosis and treatment plan and to monitor changes in dental status over time with precision[527]. These requirements necessarily set a high bar for the quality of dental examination and record-keeping in clinical practice, and data from routine clinical examination provides a way of obtaining high quality dental data without additional burden on research participants and at low cost. For these reasons, it seems likely that studies like Swedish GLIDE which use index-linkage to make use of clinical records in primary care will be an increasingly important contributor to dental epidemiology. In my role as a member of the GLIDE steering committee I am excited to see recent updates to the index linkage in GLIDE which are extending the value of this resource further, as described later in this chapter.

The specific requirements for dental assessment in clinical practice are sometimes conflated with the requirements for valid inference in dental epidemiology. In truth these requirements are different as measures of dental status in epidemiology need only capture variation in a

dental trait at population level at an acceptable level of resolution for the intended research question. This difference in requirements explains why, under the right circumstances, measures not useful in clinical practice can be extremely helpful in epidemiological research. This is illustrated throughout this dissertation where every results chapter makes use of self-reported dental data to a greater or lesser degree. Given that electronic questionnaire data can be obtained at large scale in research biobanks or consumer-facing initiates like 23andMe, it seems likely that these measures will also become increasingly important in dental epidemiology. I helped pilot, devise and describe a questionnaire for use in ALSPAC at age 23 years[528] and hope to be involved in research which uses this resource to see the impact of decisions in questionnaire design on analysis at first hand. Moving forward, the challenges surrounding use of questionnaire data in dental epidemiology may involve inference (as both analysis and interpretation will need to consider the limitations and sources of random and non-random error in these measures) and standardization, as a consensus questionnaire may be more valuable to the research community than a series of data collection exercises planned in isolation yielding incompatible data. It may be helpful to learn from other diseases, for example multiple sclerosis, where questionnaire tools have been developed through pilot studies[529] and standardized before being implemented at scale.

The importance of longitudinal resources in dental epidemiology

Analysis in chapter 2 showed the importance of previous tooth loss as a predictor of future tooth loss, and previous dental treatment under GA in childhood as a predictor of caries and anxiety in adolescence. High quality observational studies will continue to be important in refining the aetiological models of caries and periodontitis, and the ability to model changes in dental status over time may be an important part of these studies. While growth models have been used to describe changes in dental status over time and identify rapidly-changing and slowly-changing groups within a population[214,215], it may be helpful to re-visit these models in light of newer statistical methods and the desire to relate trajectories to putative risk factors. As one example, the SITAR (Super Imposition by Translation and Rotation) approach[530] can be used to model non-linear trajectories and relate changes in trajectory to specific genetic[96,531] or environmental exposures[532], specific outcomes putatively affected by the trajectory[533], or theoretically be integrated with multi-omic approaches capturing a range of features across the human phenome.

If these methods (or more conventional approaches to observational and life-course epidemiology[14]) are to be widely applied in dental epidemiology, then the community will need to support large cohorts with repeated measures obtained using the same protocol. These cohorts should ideally span disease inception in children, adolescents and young adults as well as disease progression later in adult life. Cohort studies like the Dunedin study[216] provide one model, but resources like the GLIDE cohort described in chapter 2 may also be important. At time of writing, there are 3 major updates underway which will strengthen the utility of GLIDE as a longitudinal resource. First, an updated freeze of data from Folktandvården registers has extended the duration of dental follow-up, which now extends to 18 years for some participants, and this follow-up period will continue to grow with future data merges. Second, participants who attend private dental clinics are now included in the GLIDE database for the first time through index linkage to the Svenskt Kvalitetsregister för Karies och Parodontit, increasing the sample size and unlocking a greater range of research questions such as the factors which affect dental clinic selection. Finally, there are plans underway to engage and recruit children into a paediatric cohort as part of the 'GLIDE-junior' initiative described later in this chapter. Regardless of the approach used to assemble these longitudinal resources, the early age of onset of dental caries compared to other major diseases means that phenotypic data collection may need to start at a younger age than other diseases if we are to understand the full natural history of caries.

The role of socio-economic status in the aetiology of caries

Genetic data provides an opportunity to explore the molecular basis of disease but also allows new ways to challenge or confirm existing hypotheses about the broader context of disease. As an example, the results of genetic correlation analysis in chapter 3 provide a new line of evidence supporting a role of socio-economic status in the aetiology of caries and tooth loss. Like many other complex diseases, oral health outcomes are strongly related to socio-economic status [208,534] and income[535] across multiple populations and using multiple measures of oral health.

A range of conceptual models have been described to explain inequality in health, and these are grouped by Bartley[536] into 4 broad groups of explanations. First, behavioural and cultural explanations have been suggested, where people change their behaviour to match expectations within their social circle, and this may include behaviours which are detrimental

to health (such as smoking) or those which are beneficial to health (such as jogging). For example, some of the inequality in tooth loss is reportedly mediated by differences in smoking and alcohol consumption[537]. These models imply that health inequality can be improved by people changing their behaviour and are now criticized by some authors. For example, Goldberg argues that the decision to place decisions made by individuals as the central factor explaining health inequality is ethically deficient as it implies unwell people have undesirable characteristics, as well as inefficient because individual-level health interventions are rarely effective[538]. In Dentistry, this view is echoed by Watt and Sheiham, who argue that approaches based on behavioural prevention alone are inefficient and may even lead to widening inequities in oral health[539].

Next, a group of explanations have been suggested which Bartley collects under a 'psycho-social' model[536]. These models argue that stressors (such as poor real or perceived financial, job or housing security) create a physiological response to stress which in turn contributes to poor health. Early studies showed that workers who perceived their job to be insecure experienced decreases in job satisfaction and an increase in physical symptoms over time[540], and more recent studies have explored the possible mediators between stressors and ill health, as well as understanding factors such as social support which can help mitigate the effects of stress. For example, salivary cortisol levels (as a proxy for hypothalamic-pituitary adrenocortical axis function) in teenagers who were exposed to poverty in infancy (as a stressor) are associated with quality of early life care[541], suggesting that the effects of a stressor depend on context. In a similar manner, poor housing security in early adult life is reported to have race and sex specific effects on hypertension during 15 years of follow up[542]. For dental caries, it is hypothesized that stress-related changes to saliva are a risk factor for caries, however the evidence testing this hypothesis is currently weak[543]. Apart from physiological responses, stressors may damage oral health in other ways, for example if stressful events mean that people need to prioritize a response to those events over oral health care.

A third group of models focus on the materialist aspects to health. In this group of models, a healthy lifestyle is an expensive lifestyle, because there are costs associated with accessing a high-quality diet, good housing, exercise, leisure and social activities. Since not everyone can afford the time, income and social capital needed to access a healthy lifestyle, inequalities in

health arise over time. In 1997 a model of materialist explanations for inequality in mortality was formalized by Blane, Bartley and Davey Smith[544], however this model as conceived draws attention to social structure and organization, and does not argue that interventions to change the cost of items will address inequality in health. For dental diseases there are costs associated with accessing dental screening and treatment in most countries and regular oral hygiene requires both time and consumable equipment. Conversely, the results of longitudinal analysis in chapter 2 were obtained from a population who accessed dental care repeatedly over a period of several years and still showed an association between socio-economic status and hazard for tooth loss, suggesting that access to professional care may not be an important explanation in this dataset.

Finally, Bartley defines a group of models under a header of 'Macro-social' explanations for health inequality. These argue that socio-economic inequality (rather than deprivation) produces inequality in health outcomes, as proposed by Wilkinson in 1992 following a comparison of health inequalities in countries with differing levels of income inequality[545]. Since then, this theory has been supported by some observations but challenged by other observations, for example the finding that health inequalities often increased before income inequality increased in historical longitudinal datasets[546]. The analyses presented in this dissertation were not designed to test any one of these explanations or group of explanations, but regardless of the origins of inequality in oral and dental health outcomes, there is an ongoing need to both improve oral health in the entire population and reduce inequality, and this is acknowledged as a public health goal in UK policy as described in chapter 2.

Detailed molecular datasets may help improve understanding of dental diseases

This investigation made limited use of functional and gene expression data from external sources of information using summary statistic methods. Apart from gene expression, an increasing range of –'omic' measures are becoming available to summary statistic methods following GWAS for metabolic biomarkers[463] and markers of DNA methylation[462,547]. As these external sources develop further, a similar workflow using genetic proxies for biological intermediaries could be adopted to take advantage of these resources and learn more about caries and periodontitis.

While methods to borrow external sources of information provide an efficient way to enrich a dataset, these methods involve important assumptions and place constraints on the range of

research questions which can be addressed using these methods. One limitation, as discussed in chapter 3, is the inability to explore sub-group effects in these summary statistic based methods. Another, as discussed in chapter 4, is that genetic markers acting as proxies for lifetime variation in an exposure may not reflect the effects of acute fluctuations in that exposure. Another limitation is the lack of omics data in relevant tissues, for example the molecular characteristics of saliva, the composition of the oral microbiome and gene transcription in ameloblasts and odontoblasts are all poorly characterized in publicly available data catalogues.

These limitations might be addressed by de-novo omics data generation in studies with detailed dental phenotypes and comprehensive data on health outcomes. This statement is not a contradiction with the argument for approximate, self-reported phenotypes in large collections, as these resources would be complementary and suited to different research questions. A large resource with approximate phenotypes might be suited to genetic association discovery, while a smaller resource with detailed molecular data would be valuable in understanding the biological mechanisms of a specific genotype. Recent pilot work in GLIDE has explored the feasibility of obtaining detailed molecular phenotypes for a subset of participants, as discussed later in this chapter, and may provide a resource for testing mechanistic hypotheses.

The importance of collaborative research in dental epidemiology

Inference in this dissertation was made possible by the decision of many studies to share data or results. In the case of ALSPAC and UK Biobank, this involved the studies sharing participant-level data as an open-source resource for health research. For the GLIDE consortium and KNHANES this involved the studies agreeing to a consensus analysis plan to share results and work towards a single consortium project. For external studies such as the CARDIoGRAM consortium, this relied on projects being willing to disclose and share complete results of analysis at a genome-wide level.

This level of transparency is unusual in dental epidemiology, but the examples included in this dissertation show that collaborative working is entirely possible for dental research and can yield more interesting results than working in isolation. The benefits of working in collaboration are not restricted to data, but also include the range of perspectives and

expertise which is available when drawing on a larger group of co-authors, and it has been suggested that collaboration is particularly beneficial when researchers have different skills or perspectives[548]. This might be the case when dentists work alongside epidemiologists, statisticians and geneticists. Moving forward, the network of studies, epidemiologists, dentists and analysts who have been brought together by GLIDE may form the basis of a more general collaborating network outside of GLIDE projects. In my role as a member of the GLIDE steering committee I hope to encourage these collaborations and help GLIDE act as a hub for a diverse range of researchers with a shared interest in dental epidemiology. It may be possible to embed collaborative working practices in academic dentists at an early stage in their training, for example using the Swedish GLIDE database as a training resource for student projects which involve collaboration between UK and Swedish universities.

Alongside these benefits of consortium science there are challenges including the lengthy duration of projects and the need to make analytical decisions which suit all studies. As an example, the phenotypic definitions for periodontitis in chapter 3 and caries in children in chapter 5 needed to use the 'lowest common denominator' across all studies. Working in a consortium is therefore not the answer to all problems but an informed choice allowing analysts to gain statistical power and the ability to strengthen inference by comparing results across studies at the expense of flexibility in analytical design. Likewise, the aggregation of results from many small studies with subtle but similar biases can lead to that bias becoming visible at consortium-level[549,550]. The consortium approach is absolutely required in dental epidemiology but does not negate the need for well-designed single studies. Likewise, comparison across studies with similar designs, strengths and limitations does not satisfy the requirements for triangulation[137], which ideally requires contrasting and uncorrelated strengths and limitations, so it will continue to be important to investigate research hypotheses using a range of different methods.

Implications for clinical practice, public health and policy

As understanding of dental caries and periodontitis improves, our ability to identify high-risk groups of the population should also improve. Treatment under dental general anaesthesia at an early age predicts later disease status in adolescents, so should help identify a group of high-risk children and their families, while previous treatment history identifies adults who are at high risk of tooth loss. The results of chapter 2 therefore support existing English[551]

and Scottish[552] guidelines which recommend that dentists should ask about previous dental treatment as part of caries risk assessment.

As proposed by Rose in 1985 (republished 2001), interventions on high risk groups can be effective in targeting groups with high risk and consequentially high motivation, but these approaches need interventions which are appropriate to the individual[553]. To run alongside better detection of high-risk groups it will therefore be necessary to identify interventions which are appropriate to those groups and can modify dental disease trajectories. . At group level, there is an ongoing need to identify interventions which are effective in primary or secondary prevention of dental diseases, and trials which explore the efficacy of different preventative interventions[554-556] are still important. Conversely, Rose highlights inherent limitations with interventions which target high risk groups, for example more cases of disease may arise in 'low risk' than 'high risk' groups simply because there are far fewer people in the 'high risk' group[553]. For that reason, it will continue to be important to also consider effective universal interventions which treat entire populations irrespective of their perceived risk for dental diseases, and this is discussed further in chapter 6.

While prevention of dental diseases will continue to be important, this work suggests that dental diseases may have undesirable downstream effects on health. If so, then primary prevention of dental diseases may have a positive impact on population health more generally, and there may be policy implications for closer alignment of prevention programmes for dental and cardiometabolic disorders. Watt and Sheiham argue that targeting common risk factors through a social determinants framework provides an opportunity to address the causes of both dental and other many diseases simultaneously, representing a more efficient use of resources than interventions which only prevent dental diseases[539]. Given that much of the population already has a high burden of dental disease, it may also be helpful to think about reducing the impact of pre-existing dental diseases on health. Here, the aim would be to identify population-level interventions which mitigate the health consequences of a given burden of dental disease, or targeted treatment which could be offered to people with a high level of dental disease to prevent them developing downstream consequences.

To prioritize these interventions, it might be helpful to characterize a broader range of potentially causal risk factors for dental diseases (for primary prevention) or assess the causal intermediaries downstream of dental diseases (for secondary prevention). This could be tested in a hypothesis-informed manner, for example through causal network analysis including characterization of inflammatory biomarkers, metabolic intermediaries or measures of the oral microbiome composition or function. One risk in this approach is that the broader social and environmental context might be missed. Many people who have lost teeth or have poor oral health report impact on their social and emotional well-being[557-560]. Similar to arguments for BMI[561], It has been suggested that poor dental appearance is interpreted as a reflection of undesirable personal characteristics, leading to people with poor dental appearance being discriminated against and having access to reduced socio-economic opportunities[562]. It is therefore possible that dental diseases exert their apparent effects on other health traits through changes in emotional, social and behavioural circumstances, and the consequences of dental diseases should not be considered purely in biomarker terms. This distinction is important, as the potential intervention for secondary prevention might then involve a social intervention rather than a pharmacological intervention. Instead of a hypothesis-informed causal network it may therefore be desirable to explore the phenotypic consequences of dental diseases in a hypothesis free, phenome-wide manner, where methods such as PHESANT have been described[563] and applied to traits such as BMI[564].

Methodological developments

This dissertation applied emerging methodologies which may be useful in the study of other traits. One development, briefly alluded to in chapter 2, is the depiction of fine-scale latent structure in the genetic data of UK Biobank. While current methods attempt to control for and remove systematic structure, this structure provides a source of variation which could potentially be exploited to gain statistical power or perform natural experiments. For example, genetic predictors of birth location might conceptually be used to explore hypotheses related to geographic determinants of disease, and I am involved in a student project which uses geographic structure in genetic data to infer the likely impact of different local policies on rates of total hip replacement in the UK.

A second development was the use of genome-wide genetic correlation to select proxy phenotypes from an independent collection with imperfect measurement. This might

potentially be useful in other contexts where detailed measurement is only possible in small collections, but has natural limitations, particularly in generating interpretable estimates of genetic effect size. To address this, I used standardized regression coefficients derived from Z scores, but other approaches are emerging. For example, a recent pre-publication article suggests jointly modelling the genetic effects of single variants on two or more correlated traits simultaneously by viewing the observed traits as linear combinations of partially shared underlying latent traits[565].

The genome-wide analysis included in chapters 3 and 5 assumed that the effects of any given genotype on dental caries were the same in each study. While this provides a valid estimate of the weighted average genetic effect size, it seems likely that the absolute-scale effect of a genotype will vary with changes in the distribution of a trait, and the relative importance or penetrance of genetic effects will vary with the availability of co-causal environmental risk factors. In dentistry it has been known since 1954 that gnotobiotic rats are incapable of developing dental caries even in the presence of highly cariogenic diets[566], but can develop caries as little as 3 weeks after monocontamination with oral streptococcus[567]. Here, the cariogenic effect of diet could be completely ameliorated by the absence of the co-causal risk factor. By extension and it seems plausible that the effects of common genetic variation on risk of caries could be reduced or exacerbated by co-causal risk factors[35]. An increasing number of tools are now available to explore gene-environment interactions[568]. As larger datasets emerge and methods for gene by environment interaction analysis improve it may become possible to develop better models which account for or even borrow information from the network of co-causal factors in estimating genetic main effects and interaction terms, and integrate both genetic and environmental factors to understand the circumstances under which genetic risk can be ameliorated.

Another methodological consideration for future research is how to interpret heterogeneity in genetic effects for complex traits like periodontitis. The lack of a consensus definition in the different studies in GLIDE may have contributed to the low apparent heritability of periodontitis, as diagnostic interpretation is a form of environmental influence[569] which might contribute to heterogeneity in effect sizes between different studies. Within one classification system, varying effect sizes due to gene by environment interactions would also result in apparent heterogeneity. Even within one population with homogeneous environment and

disease classification, there may be apparent heterogeneity in genetic effect if there are hidden subgroups with different effect sizes[570], which seems plausible if the definitions used in this study captured more than one sub-type of periodontal disease. Another motivation for resolving within-disease heterogeneity is the possible clinical application in improved diagnosis or management. Aside from the use of genetics to predict risk, precision medicine may also involve the use of genetic tests to identify disease sub-types which respond to specific treatment modalities[571].

If the aim is to use heterogeneity in stratified genetic effect estimates to explore the meaning of genetic risk in different environmental or health contexts as suggested above, then understanding the inherent heterogeneity in the disease or identifying traits which distinguish between different subtypes of the disease will be an important first step, and will require characterization of multiple signatures of periodontitis. Reciprocally, attempts to identify disease subtypes based on heterogeneity in genetic data alone will be difficult to interpret until the role of geography, environmental and health context as sources of heterogeneity are well established.

6.2: Future directions

Next steps for the GLIDE consortium – genetic association discovery

While the genetic risk loci identified in chapter 3 provide a good starting point for understanding the genetic architecture of dental diseases, they explain little of the variation in dental disease traits and only some bioinformatic tools could be usefully applied. For example, there were probably too few loci for gene set enrichment tests to be a useful follow-up tool. Using the GIANT consortium as an example, the sample size used in discovery meta-analysis approximately doubled between 2010 and 2014[572,573], but this led to approximately a 4 fold increase in the number of genetic risk loci identified as well as more meaningful bioinformatic tests. It would therefore be helpful to consider ways to improve sample size for the next round of GLIDE GWAS.

This will include studies which have performed dental examination as part of their protocol since the analysis in chapter 3 was undertaken. As examples, the latest round of the Norwegian Nord-Trøndelag Health (HUNT) Study (URL https://www.ntnu.edu/hunt4) is

currently performing dental examinations for approximately 4,000 participants and anticipates that these data will be ready for analysis in the fourth quarter of 2019, while in Japan the Tohoku Medical Megabank BirThree project[574] plans to perform dental examination on all participants aged 20 years or older (approximately 33,000 participants). Both studies plan to participate in the next round of GLIDE GWAS analysis.

While these studies have the potential to make a major contribution to genetic epidemiology in dentistry, cohorts with the resources to perform a comprehensive dental examination remain a minority. It may also be helpful to use the index-linked approach advocated throughout this dissertation. As an example, GLIDE recently completed an updated data merge with the Swedish Twin Register, providing clinical dental records for more than 50,000 participants of whom many have donated blood samples for genotyping. Outside Sweden, primary care records for National Health Service (NHS) dental services have been linked to genetic data in the Generation Scotland project[575] for around 15,000 participants, and it is theoretically possible (but currently unimplemented) to link summary data stored by the NHS business services authority to records in UK Biobank[576].

It may also be helpful to look for other large-scale collections with dental questionnaires. As an example, the consumer health initiative 23andMe contains a database of customers who have agreed to participate in research. In a recent project[97] (not included in this dissertation) 23andMe replicated association at genetic risk loci for mouth ulcers in a group of 356,000 participants who had answered a questionnaire, and it may be possible to include other questionnaire-derived oral and dental traits in 23andMe as part of a discovery GWAS in GLIDE.

Next steps for the GLIDE consortium – mechanistic refinement

Alongside larger studies for genetic association discovery there need to be resources with detailed phenotypes to explore mechanistic hypotheses. Recent pilot work in the Swedish arm of GLIDE explored the practicality of detailed molecular phenotyping including assessment of the oral microbiota composition using 16S sequencing and cultivation and measures of salivary flow rate and composition[100,101]. These methods are now being scaled up to a larger data collection exercise nested within the sampling frame of the Swedish Twin Register and

will be used to inform the data collection strategy for the 'GLIDE junior' proposal discussed later.

Even without these new measures, the Swedish GLIDE resources can be used to explore questions which would be difficult elsewhere in the field of dental epidemiology. As an example, it has been argued that dental caries lesions on different tooth surfaces represent the endpoints of subtly different disease processes, and that genetic risk factors are only relevant for some clusters of disease presentation[577]. This hypothesis would be difficult to test in most studies but can be assessed using hierarchical clustering of tooth surface data in the Swedish Twin Registry followed by a multivariate variance decomposition model to estimate the heritability and shared heritability of different clusters. Likewise, the longitudinal dental data in Swedish GLIDE provides an unusual opportunity to test the relevance of genetic or environmental risk factors on disease trajectories, and I plan to present initial findings from these analyses at the British Society for Oral and Dental Research meeting this autumn.

Finally, there are very few studies which can explore the interplay between dental diseases in childhood and dental and systemic health in adulthood. The high uptake of Folktandvården dental examination by children living in Sweden (95-98%) provides a potential sampling frame to recruit a population-representative group of children into dental research before some of the reasons for dental avoidance develop and obtain consent and biosamples in the dental clinic. Subsequent follow-up could be performed without any additional burden to these participants, potentially helping to avoid selective drop-out, and index linkage could include dental records, diet databases, and health registers later in life as well as parental data for families who have previously participated in health research. This is the rationale for the 'GLIDE junior' cohort, which will be developed through a pilot study targeting children in 4 northern Swedish counties at age 6 years, with a long-term aspiration to create a nationwide cohort embedding dental research into primary dental care in Sweden.

Future tests for causal effect

Methods for causal effect estimating using the Mendelian randomization paradigm have developed during recent years and are likely to continue to develop. Alongside newer methods, it is likely that larger datasets with both dental and cardio-metabolic phenotypes will become available, which may help generate more precise effect estimates and

demonstrate the similarities or differences between different estimation tools with greater clarity.

As far as I am aware, none of the current methods enable estimation of multiple causal effects from a single exposure, but it may be worth exploring whether there are situations where this would be appropriate. Complex exposures could be conceptualized as a combination of underlying biological traits, each of which may have one or more genetic phenocopies, and it may not be reasonable to assume that each of these underlying traits s has the same effect on the outcome. As introduced by the 'biomarker X' thought experiment in chapter 4, there might be situations where it is more helpful to consider the effect of interventions on the risk factor (such as a drug, exercise intervention and smoking cessation programme) on an outcome than estimating the combined effect of all possible interventions on the risk factor on outcome. Unsupervised hierarchical clustering of phenocopies followed by annotation of each of the clusters to external sources of GWAS results and functional data might provide one way to infer the effects of these different interventions. Alternatively, genetic variants used in an MR experiment could be assigned to different groups based using a bioinformatic approach mining summary statistics of GWAS studies for biomarkers or other complex traits. In either case, this idea requires development before it could be applied in practice but is something I hope to take forward.

Direct application of genetic data to clinical decisions

Theoretically, precision dentistry could use genetic data as an adjunct to routine decision making in clinical practice. For example, it might be possible to identify subtypes of disease presentation and assign treatment tailored to that particular subtype. For this approach to work will be necessary to understand the biological relevance of genetic risk loci and develop interventions to address the specific biological processes which have failed in that patient. This is ambitious, but the mechanistic refinement experiments proposed above may help go some way to closing the gap between genetic association and targeted intervention.

A clinical genetic test for periodontitis was proposed as early as 2002 based on results of candidate gene studies, but rejected due to inconclusive evidence[578]. Nevertheless, the idea of a genetic test for complex dental diseases continued to attract interest and was discussed in detail at a conference in 2015[579], while in 2016, Interleukin Genetics Inc. launched a direct-

to-consumer test based on polymorphisms in *IL1B* following an industry-funded candidate gene association study[580]. This test appears to have been discontinued since Interleukin Genetic Inc was acquired by Orig3n, but the idea is likely to re-appear in one form or another.

While these previous attempts to develop genetic tests have not been entirely successful, changes in technology and methods may lead to a viable test soon. Given the decreasing cost of genome-wide arrays and the modest effects of individual genetic variants, a future test based on a genome-wide predictor may be more effective than previously, based on a limited number of variants. The use of genome-wide genetic risk scores to predict dental diseases is not explored in this dissertation but may be a useful avenue for future research.

Indirect application of genetic data in clinical dentistry

As an alternative or complementary approach, it may be possible to use genetic data to predict risk of disease even when the underlying biology is unclear. Used in the broadest possible sense, genetic data could theoretically predict both biological susceptibility to disease and disease-promoting environmental factors, similar to genetic predictors of educational attainment which can predict a household environment which is conducive to children's learning – termed the 'nature of nurture'[456]. In dentistry, a genetic test predicting the environmental features which give rise to dental disease could theoretically be used as an adjunct to history taking to help identify high risk individuals and intervene before disease develops, as has recently been proposed for cardiovascular diseases[581]. For caries, a complete care pathway describing prevention and then appropriate interventions for every stage of the disease process from inception onward has been described[551], however a large proportion of treatment in practice continues to resolve around treating the late consequences of disease through dental surgery. Unlike the cardiovascular disease example above where pharmacological options such as statins are a natural choice to help reduce risk[582], it is unclear what the appropriate intervention for high risk groups in dentistry might entail, but it could potentially involve an approach such as early escalation to the next level of the care pathway. The partially overlapping heritability of dental diseases and other complex traits suggest that the results of these tests might also have implications for clinical management extending outside dentistry.

Genetic testing as a bridge between clinical dentistry and medicine

The similarities in underlying biology between dental and systemic diseases suggested by chapter 3 of this dissertation provide further evidence that dental diseases cannot be viewed in insolation but as a manifestation of overall health. Currently this is not reflected in the structure of healthcare provision in the UK, where clinical dentistry is provided as an entirely distinct fraction from medicine. Routine dental examinations bring people into contact with a healthcare professional on a regular basis. This may provide an opportunity for primary care dentistry to play a greater role in achieving population health goals, as many of these people are apparently healthy and would not attend a medical practice for health examination. This difference in sampling frames between general medical practice and general dental practice may be instrumental in shifting disease prediction and interventions to an earlier stage in disease natural history.

Currently, information about general health status is already gathered by dentists as part of their medical and social history taking and (indirectly) by dental examination. One way to break down divisions between dentistry and medicine would involve routine sharing of these data. Over time, this could potentially be expanded to a greater range of information collection which could be shared with the medical profession to identify patients who have high genetic or environmental risk for systemic diseases, even if the test was not explicitly designed for systemic diseases. For example, a dental examination protocol might contain a test for salivary cariogenic potential which provides some information about glucose tolerance, and a genome-wide genetic test for dental diseases would likely provide some information about risk of other complex diseases which a general medical practitioner could use in risk assessment. Finally, dental practices could be used as a sampling frame for tests specifically intending to capture risk of systemic diseases, for example by training dental nurses to obtain blood pressure measurement and sharing results of these tests with medical professionals.

Workforce and training

It is clear there are many unanswered questions before genetic data can be used in dental practice as an adjunct to routine decision making. These have been described as analytical and methodological challenges in the paragraphs above, but there are broader considerations around the use of genetic tests. There are complex legal[583] and ethical issues around return of

results from genetic testing, including unclear rationale for testing or interpretation of findings from direct-to-consumer tests, the need for re-contact if interpretation changes and potential for incidental findings. Even in the medical sphere where genetic tests are well-established these tests are creating a major global challenge for the medical professon[584].

Currently the dental workforce is not equipped to deal with these problems in the UK, but it is likely that tests for personal genetic liability to caries and periodontitis will arrive, with or without the workforce being prepared. One way to start preparing for these tests would be to lay the foundations with qualitative studies so that tests are developed which address the priorities of patients and healthcare professionals and respect patient autonomy and choices. Here again the GLIDE consortium may help gather the data needed to address these questions. As an example, the GLIDE cohort included in chapter 2 is a group of adults who are interested in health research, accustomed to completing questionnaires about their health, and attend a dentist, so may provide a key stakeholder group in collecting information about patient views on genetic testing in dentistry.

Another way to prepare the dental workforce would be to invest in education and support mechanisms. In the USA competencies in dental genetics were introduced to the dental curriculum by panel 3 of the Macy study in 2008[585], and have since been adopted as a routine part of providing education in dentistry. Another route may be to provide clear guidance for the dental profession, for example the American Dental Association council on scientific affairs contains a genetic testing workgroup who provide advice on genetic testing in the dental profession. Adopting similar steps to prepare the workforce in the UK and other countries may be a sound long term investment in the dental profession.

6.3: Commentary

At the outset of this dissertation I argued that genetic data could be a key tool to help refine aetiological models of caries and periodontitis. Specifically, I tested the hypothesis that genetic data could be used to investigate the molecular aetiology of caries and periodontitis and test for causal association between dental diseases and cardio-metabolic traits. To explore these questions, large samples with genetic and dental data are needed, and I used a theoretical review and applied illustrations to demonstrate that index-linked and questionnaire-derived dental phenotypes are a practical way to achieve large samples. Following this, both index-linked and questionnaire-derived data were used in GWAS for caries and periodontitis, identifying novel genetic risk loci for caries in both adults and children and suggesting that dental caries has a polygenic genetic architecture. GWAS findings were applied in a Mendelian randomization context, suggesting that dental diseases have undesirable downstream effects on systemic health.

These findings may have implications for clinical practice, public health and policy. Primary prevention of dental diseases may be one way to improve population health and provides an argument for a larger role of primary care dentistry in achieving population health goals. The existence of shared genetic and environmental risk factors for dental and other diseases provides further support for the notion that oral heath should considered an integral part of human health in research, policy and clinical practice. In chairside application, the results re-emphasise the importance of previous treatment history for risk assessment and provide a first step towards use of genetic data in precision dentistry.

For research, the findings will help direct mechanistic studies investigating both the molecular aetiology of caries and the causal intermediaries by which dental and systemic health traits interact. The results suggest additional methodological angles to explore, especially around the instrumentation of complex traits capturing multiple biological processes, and the role of disease subtypes and environment as sources of heterogeneity in genetic effect estimates. Overall, the results are consistent with the initial hypothesis that genetic data will be a key resource in exploring the causes and consequences of dental diseases, while also emphasising the need for still larger sample sizes for some dental traits. To exploit the full potential of new methods in understanding these common and important diseases, the dental research community must invest in collaborative working practices and the creation of large-scale resources for dental epidemiology.

List of References

1. Global Burden of Diseases Injury, Incidence & Prevalence Collabarators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211-1259 (2017).
2. Selwitz, R.H., Ismail, A.I. & Pitts, N.B. Dental caries. *The Lancet* **369,** 51-59 (2007).
3. Sheiham, A. & James, W.P.T. Diet and Dental Caries: The Pivotal Role of Free Sugars Reemphasized. *Journal of Dental Research* **94**, 1341-1347 (2015).
4. Vale, G.C. *et al.* Temporal Relationship between Sucrose-Associated Changes in Dental Biofilm Composition and Enamel Demineralization. *Caries Research* **41**, 406-412 (2007).
5. Paes Leme, A.F. *et al.* Effects of Sucrose on the Extracellular Matrix of Plaque-Like Biofilm Formed in vivo, Studied by Proteomic Analysis. *Caries Research* **42**, 435-443 (2008).
6. Figuero, E. *et al.* Mechanical and chemical plaque control in the simultaneous management of gingivitis and caries: a systematic review. *Journal of Clinical Periodontology* **44**, S116-S134 (2017).
7. Nóbrega, D.F., Fernández, C.E., Del Bel Cury, A.A., Tenuta, L.M.A. & Cury, J.A. Frequency of Fluoride Dentifrice Use and Caries Lesions Inhibition and Repair. *Caries Research* **50**, 133-140 (2016).
8. Tenuta, L.M.A., Zamataro, C.B., Del Bel Cury, A.A., Tabchoury, C.P.M. & Cury, J.A. Mechanism of Fluoride Dentifrice Effect on Enamel Demineralization. *Caries Research* **43**, 278-285 (2009).
9. Chersoni, S. *et al.* In vivo effects of fluoride on enamel permeability. *Clinical Oral Investigations* **15**, 443-449 (2011).
10. Thurnheer, T. & Belibasakis, G.N. Effect of sodium fluoride on oral biofilm microbiota and enamel demineralization. *Archives of Oral Biology* **89**, 77-83 (2018).
11. Bonetti, D. & Clarkson, J.E. Fluoride Varnish for Caries Prevention: Efficacy and Implementation. *Caries Research* **50(suppl 1)**, 45-49 (2016).
12. Vos, T. *et al.* Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* **380**, 2163-2196 (2012).
13. Righolt, A.J., Jevdjevic, M., Marcenes, W. & Listl, S. Global-, Regional-, and Country-Level Economic Impacts of Dental Diseases in 2015. *Journal of Dental Research* **97**, 501-507 (2018).
14. Abreu, L.G. *et al.* Factors associated with the development of dental caries in children and adolescents in studies employing the life course approach: a systematic review. *European Journal of Oral Sciences* **123**, 305-311 (2015)
15. Nicolau, B., Marcenes, W., Bartley, M. & Sheiham, A. A Life Course Approach to Assessing Causes of Dental Caries Experience: The Relationship between Biological, Behavioural, Socio-Economic and Psychological Conditions and Caries in Adolescents. *Caries Research* **37**, 319-326 (2003).
16. Teixeira, A.K.M., Roncalli, A.G. & Noro, L.R.A. Income Trajectories and Oral Health of Young People in a Life Course Study. *Caries Research* **53**, 347-356 (2019).
17. Harris, R.V., Pennington, A. & Whitehead, M. Preventive dental visiting: a critical interpretive synthesis of theory explaining how inequalities arise. *Community Dentistry and Oral Epidemiology* **45**, 120-134 (2017).
18. Carounanidy, U. & Sathyanarayanan, R. Dental caries - A complete changeover (Part I). *Journal of Conservative Dentistry* **12**, 46-54 (2009).
19. Grytten, J. The impact of education on dental health — Ways to measure causal effects. *Community Dentistry and Oral Epidemiology* **45**, 485-495 (2017).

20. Chapple, I.L.C. *et al.* Primary prevention of periodontitis: managing gingivitis. *Journal of Clinical Periodontology* **42**, S71-S76 (2015).
21. Page, R.C. & Kornman, K.S. The pathogenesis of human periodontitis: an introduction. *Periodontology 2000* **14**, 9-11 (1997).
22. Socransky, S.S., Haffajee, A.D., Cugini, M.A., Smith, C. & Kent Jr, R.L. Microbial complexes in subgingival plaque. *Journal of Clinical Periodontology* **25**, 134-144 (1998).
23. Sanz, I., Alonso, B., Carasol, M., Herrera, D. & Sanz, M. Nonsurgical Treatment of Periodontitis. *Journal of Evidence Based Dental Practice* **12**, 76-86 (2012).
24. Smiley, C.J. *et al.* Systematic review and meta-analysis on the nonsurgical treatment of chronic periodontitis by means of scaling and root planing with or without adjuncts. *The Journal of the American Dental Association* **146**, 508-524.e5 (2015).
25. da Costa, L.F.N.P., Amaral, C.d.S.F., Barbirato, D.d.S., Leão, A.T.T. & Fogacci, M.F. Chlorhexidine mouthwash as an adjunct to mechanical therapy in chronic periodontitis: A meta-analysis. *The Journal of the American Dental Association* **148**, 308-318 (2017).
26. Jepsen, K. & Jepsen, S. Antibiotics/antimicrobials: systemic and local administration in the therapy of mild to moderately advanced periodontitis. *Periodontology 2000* **71**, 82-112 (2016).
27. Roberts, H.M. *et al.* Impaired neutrophil directional chemotactic accuracy in chronic periodontitis patients. *Journal of Clinical Periodontology* **42**, 1-11 (2015).
28. Herrmann, J.M. & Meyle, J. Neutrophil activation and periodontal tissue injury. *Periodontology 2000* **69**, 111-127 (2015).
29. Dommisch, H. & Jepsen, S. Diverse functions of defensins and other antimicrobial peptides in periodontal tissues. *Periodontology 2000* **69**, 96-110 (2015).
30. Fujita, T. *et al.* Regulation of defensive function on gingival epithelial cells can prevent periodontal disease. *The Japanese Dental Science Review* **54**, 66-75 (2018).
31. Belibasakis, G.N. & Bostanci, N. The RANKL-OPG system in clinical periodontology. *Journal of Clinical Periodontology* **39**, 239-248 (2012).
32. Meyle, J. & Chapple, I. Molecular aspects of the pathogenesis of periodontitis. *Periodontology 2000* **69**, 7-17 (2015).
33. Hajishengallis, G. & Lamont, R.J. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. *Molecular Oral Microbiology* **27**, 409-419 (2012).
34. Lamont, R.J., Koo, H. & Hajishengallis, G. The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology* **16**, 745-759 (2018).
35. Chapple, I.L.C. *et al.* Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. *Journal of Clinical Periodontology* **44**, S39-S51 (2017).
36. Woelber, J.P. *et al.* An oral health optimized diet can reduce gingival and periodontal inflammation in humans - a randomized controlled pilot study. *BMC Oral Health* **17**, 28 (2016).
37. Woelber, J.P. *et al.* The influence of an anti-inflammatory diet on gingivitis. A randomized controlled trial. *Journal of Clinical Periodontology* **46**, 481-490 (2019).
38. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, 5-22 (2017).
39. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine* **373**, 895-907 (2015).
40. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177-183 (2016).
41. Nelson, M.R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856-860 (2015).

42. King, E.A., Davis, J.W. & Degner, J.F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. Preprint at *bioRxiv* https://doi.org/10.1101/513945 (2019).
43. Yang, L. *et al.* Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome--clozapine-induced agranulocytosis as a case study. *PLoS Computational Biology* **7**, e1002016-e1002016 (2011).
44. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013).
45. Barbeira, A.N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genetics* **15**, e1007889 (2019).
46. Barbeira, A.N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018).
47. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236-1241 (2015).
48. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *The American Journal of Human Genetics* **101**, 737-751 (2017).
49. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015).
50. Duncan, L.E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* (2019).
51. Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J. & Richards, J.B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics* **19**, 110 (2017).
52. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).
53. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Human molecular genetics* **24**, R102-R110 (2015).
54. Lawson, D.J. *et al.* Is Population Stratification in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics* (2019).
55. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications* **10**, 333 (2019).
56. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990;2013;2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211-1259 (2017).
57. Shaffer, J.R. *et al.* Genetic Susceptibility to Dental Caries Differs between the Sexes: A Family-Based Study. *Caries Research* **49**, 133-140 (2015).
58. Shaffer, J.R. *et al.* Genetic Susceptibility to Dental Caries on Pit and Fissure and Smooth Surfaces. *Caries Research* **46**, 38-46 (2012).
59. Wang, X. *et al.* Genes and Their Effects on Dental Caries May Differ between Primary and Permanent Dentitions. *Caries Research* **44**, 277-284 (2010).
60. Mucci, L.A., Björkman, L., Douglass, C.W. & Pedersen, N.L. Environmental and Heritable Factors in the Etiology of Oral Diseases—A Population-based Study of Swedish Twins. *Journal of Dental Research* **84**, 800-805 (2005).
61. Michalowicz, B.S. *et al.* Evidence of a Substantial Genetic Basis for Risk of Adult Periodontitis. *Journal of Periodontology* **71**, 1699-1707 (2000).
62. Bretz, W.A. *et al.* Heritability estimates for dental caries and sucrose sweetness preference. *Archives of Oral Biology* **51**, 1156-1160 (2006).
63. Bretz, W.A. *et al.* Longitudinal analysis of heritability for dental caries traits. *Journal of Dental Research* **84**, 1047-1051 (2005).

64. Nascimento, M.M., Zaura, E., Mira, A., Takahashi, N. & ten Cate, J.M. Second Era of OMICS in Caries Research: Moving Past the Phase of Disillusionment. *Journal of Dental Research* **96**, 733-740 (2017).
65. Ballantine, J.L. *et al.* Exploring the genomic basis of early childhood caries: a pilot study. *International Journal of Paediatric Dentistry* **28**, 217-225 (2018).
66. Morrison, J. *et al.* Genome-wide association study of dental caries in the Hispanic Communities Health Study/Study of Latinos (HCHS/SOL). *Human Molecular Genetics* **25**, 807-816 (2016).
67. Zeng, Z. *et al.* Genome-Wide Association Study of Primary Dentition Pit-and-Fissure and Smooth Surface Caries. *Caries Research* **48**, 330-338 (2014).
68. Zeng, Z. *et al.* Genome-wide Association Studies of Pit-and-Fissure- and Smooth-surface Caries in Permanent Dentition. *Journal of Dental Research* **92**, 432-437 (2013).
69. Shaffer, J.R. *et al.* GWAS of dental caries patterns in the permanent dentition. *Journal of Dental Research* **92**, 38-44 (2013).
70. Wang, X.J. *et al.* Genome-wide association Scan of dental caries in the permanent dentition. *BMC Oral Health* **12**, 10 (2012).
71. Shaffer, J.R. *et al.* Genome-wide Association Scan for Childhood Caries Implicates Novel Genes. *Journal of Dental Research* **90**, 1457-1462 (2011).
72. Sanders, A.E. *et al.* Chronic Periodontitis Genome-wide Association Study in the Hispanic Community Health Study / Study of Latinos. *Journal of Dental Research* **96**, 64-72 (2016).
73. Offenbacher, S. *et al.* Genome-wide association study of biologically informed periodontal complex traits offers novel insights into the genetic basis of periodontal disease. *Human Molecular Genetics* **25**, 2113-2129 (2016).
74. Hong, K.-W., Shin, M.-S., Ahn, Y.-B., Lee, H.-J. & Kim, H.-D. Genome-wide association study on chronic periodontitis in Korean population: results from the Yangpyeong health cohort. *Journal of Clinical Periodontology* **42**, 703-710 (2015).
75. Shimizu, S. *et al.* A Genome-wide Association Study of Periodontitis in a Japanese Population. *Journal of Dental Research* **94**, 555-561 (2015).
76. Shaffer, J.R. *et al.* Genome-Wide Association Study of Periodontal Health Measured by Probing Depth in Adults Ages 18-49 years. *G3-Genes Genomes Genetics* **4**, 307-314 (2014).
77. Teumer, A. *et al.* Genome-wide association study of chronic periodontitis in a general German population. *Journal of Clinical Periodontology* **40**, 977-985 (2013).
78. Divaris, K. *et al.* Exploring the genetic basis of chronic periodontitis: a genome-wide association study. *Human Molecular Genetics* **22**, 2312-2324 (2013).
79. Divaris, K. *et al.* Genome-wide Association Study of Periodontal Pathogen Colonization. *Journal of Dental Research* **91**, S21-S28 (2012).
80. Schaefer, A.S. *et al.* A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis. *Human Molecular Genetics* **19**, 553-562 (2010).
81. Robinson, A. NHS England's plan to pull the plug on ineffective procedures. *British Medical Journal* **362**, k3028 (2018).
82. Bolland, M.J., Grey, A. & Avenell, A. Effects of vitamin D supplementation on musculoskeletal health: a systematic review, meta-analysis, and trial sequential analysis. *The Lancet Diabetes & Endocrinology* **6**, 847-858 (2018).
83. Wensley, F. *et al.* Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *British Medical Journal* **342**(2011).
84. Lawlor, D.A. *et al.* The Association of C-Reactive Protein and CRP Genotype with Coronary Heart Disease: Findings from Five Studies with 4,610 Cases amongst 18,637 Participants. *PLoS ONE* **3**, e3011 (2008).
85. Timpson, N.J. *et al.* C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet* **366**, 1954-1959 (2005).

86. Glass, T.A., Goodman, S.N., Hernán, M.A. & Samet, J.M. Causal inference in public health. *Annual Review of Public Health* **34**, 61-75 (2013).
87. Richmond, R.C., Al-Amin, A., Smith, G.D. & Relton, C.L. Approaches for drawing causal inferences from epidemiological birth cohorts: A review. *Early Human Development* **90**, 769-780 (2014).
88. Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19**, 566-580 (2018).
89. Stefan, L., Hendrik, J. & G., W.R. Causal inference from observational data. *Community Dentistry and Oral Epidemiology* **44**, 409-415 (2016).
90. Lawlor, D.A. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology* **45**, 908-915 (2016).
91. Hartwig, F.P., Davies, N.M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International Journal of Epidemiology* **45**, 1717-1726 (2016).
92. Davey Smith, G. Epigenesis for epidemiologists: does evo-devo have implications for population health research and practice? *International Journal of Epidemiology* **41**, 236-247 (2012).
93. Lawson, D.J. *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics* (2019).
94. Haworth, S., Dudding, T., Waylen, A., Thomas, S.J. & Timpson, N.J. Ten years on: Is dental general anaesthesia in childhood a risk factor for caries and anxiety? *British Dental Journal* **222**, 299-304 (2017).
95. Haworth, S. *et al.* Tooth loss is a complex measure of oral disease: Determinants and methodological considerations. *Community Dentistry and Oral Epidemiology* **46,** 555-562 (2018).
96. Haworth, S. *et al.* Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. *Nature Communications* **10**, 357 (2019).
97. Dudding, T. *et al.* Genome wide analysis for mouth ulcers identifies associations at immune regulatory loci. *Nature Communications* **10**, 1052 (2019).
98. Shungin, D. *et al.* Genome-wide analysis of dental caries and periodontal disease combining clinical and self-reported data. *Nature Communications* **10**, 2773 (2019).
99. Haworth, S. *et al.* Consortium-based genome-wide meta-analysis for childhood dental caries traits. *Human Molecular Genetics* **27**, 3113-3127 (2018).
100. Esberg, A., Haworth, S., Brunius, C., Lif Holgerson, P. & Johansson, I. Carbonic Anhydrase 6 Gene Variation influences Oral Microbiota Composition and Caries Risk in Swedish adolescents. *Scientific Reports* **9**, 452 (2019).
101. Johansson, I., Esberg, A., Eriksson, L., Haworth, S. & Lif Holgerson, P. Self-reported bovine milk intake is associated with oral microbiota composition. *PLoS ONE* **13**, E0193504 (2018).
102. Eriksson, L., Esberg, A., Haworth, S., Holgerson, L.P. & Johansson, I. Allelic Variation in Taste Genes Is Associated with Taste and Diet Preferences and Dental Caries. *Nutrients* **11,** E1491 (2019).
103. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet* **360**, 1903-1913 (2002).
104. Rapsomaniki, E. *et al.* Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. *The Lancet* **383**, 1899-1911 (2014).
105. Lee, J.J. Correlation and Causation in the Study of Personality. *European Journal of Personality* **26**, 372-390 (2012).
106. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).

107. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **186**, 1026-1034 (2017).
108. Swanson, J.M. The UK Biobank and selection bias. *Lancet* **380**, 110-110 (2012).
109. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**, 294-305 (2011).
110. Steckler, A. & McLeroy, K.R. The importance of external validity. *American Journal of Public Health* **98**, 9-10 (2008).
111. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173-1174 (2012).
112. Berkson, J. Limitations of the Application of Fourfold Table Analysis to Hospital Data.*,†. *International Journal of Epidemiology* **43**, 511-515 (2014).
113. Munafò, M.R., Tilling, K., Taylor, A.E., Evans, D.M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology* **47**, 226-235 (2018).
114. Taylor, A.E. *et al.* Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* **47**, 1207-1216 (2018).
115. Gkatzionis, A. & Burgess, S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology* **48,** 691-701 (2019).
116. Howe, C.J., Cole, S.R., Lau, B., Napravnik, S. & Eron, J.J., Jr. Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology (Cambridge, Mass.)* **27**, 91-97 (2016).
117. Bishop, C.D., Leite, W.L. & Snyder, P.A. Using Propensity Score Weighting to Reduce Selection Bias in Large-Scale Data Sets. *Journal of Early Intervention* **40**, 347-362 (2018).
118. Nohr, E.A. & Liew, Z. How to investigate and adjust for selection bias in cohort studies. *Acta Obstetrica et Gynecologica Scandinavica* **97**, 407-416 (2018).
119. Hellwege, J. *et al.* Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics* **95**, 1.22.1-1.22.23 (2017).
120. Bouaziz, M., Ambroise, C. & Guedj, M. Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. *PLoS ONE* **6**, e28845 (2011).
121. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904 (2006).
122. Popejoy, A.B. & Fullerton, S.M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).
123. Belle, E.M.S., Landry, P.-A. & Barbujani, G. Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proceedings. Biological sciences* **273**, 1595-1602 (2006).
124. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics* **21**, 1277-1285 (2013).
125. Zhang, Y., Guan, W. & Pan, W. Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants. *Genetic Epidemiology* **37**, 99-109 (2013).
126. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**, 243-246 (2012).
127. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics* **8** (2012).
128. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed model association for biobank-scale data sets. Preprint at *bioRxiv* https://doi.org/10.1101/194944 (2017).
129. Paquet, C. *et al.* Geographic Clustering of Cardiometabolic Risk Factors in Metropolitan Centres in France and Australia. *International Journal of Environmental Research and Public Health* **13**, 519 (2016).

130. Domingue, B.W., Rehkopf, D.H., Conley, D. & Boardman, J.D. Geographic Clustering of Polygenic Scores at Different Stages of the Life Course. *RSF: The Russell Sage Foundation Journal of the Social Sciences* **4**, 137-149 (2018).
131. Battey, C.J., Ralph, P.L. & Kern, A.D. Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data. Preprint at *bioRxiv* https://doi.org/10.1101/659235 (2019).
132. Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 7377-7382 (2016).
133. Young, A.I. *et al.* Estimating heritability without environmental bias. Preprint at *bioRxiv* https://doi.org/10.1101/218883 (2017).
134. Sanderson, E., Macdonald-Wallis, C. & Davey Smith, G. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology* **47**, 587-596 (2018).
135. Shin, C., Kwack, K., Cho, N.H., Kim, S.H. & Baik, I. Sex-specific differences in the association of a common aldehyde dehydrogenase 2 gene polymorphism and alcohol consumption with stroke risk in a Korean population: a prospective cohort study. *Nutrition Research and Practice* **9**, 79-86 (2015).
136. Sofer, T. *et al.* Admixture mapping in the Hispanic Community Health Study/Study of Latinos reveals regions of genetic associations with blood pressure traits. *PLoS ONE* **12**, e0188400 (2017).
137. Lawlor, D.A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* **45**, 1866-1886 (2016).
138. Hutcheon, J.A., Chiolero, A. & Hanley, J.A. Random measurement error and regression dilution bias. *British Medical Journal* **340**, c2289 (2010).
139. Reuter, M. *et al.* Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage* **107**, 107-115 (2015).
140. Malone, I.B. *et al.* Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. *NeuroImage* **104**, 366-372 (2015).
141. Smit, D.J.A. *et al.* Heritability of Head Size in Dutch and Australian Twin Families at Ages 0-50 Years. *Twin Research and Human Genetics* **13**, 370-380 (2010).
142. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics* **27**, 3641-3649 (2018).
143. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248 (2016).
144. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981-90 (2012).
145. van der Weele, T.J., Valeri, L. & Ogburn, E.L. The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology (Cambridge, Mass.)* **23**, 561-564 (2012).
146. Pitts, N., Chadwick, B. & Anderson, T. Children's Dental Health Survey 2013 Report 2: Dental Disease and Damage in Children England, Wales and Northern Ireland. Published by the Health and Social Care Information Centre (2015).
147. Qiu, R., Hadzikadic, M., Yu, S. & Yao, L. Estimating disease burden using Internet data. *Health Informatics Journal*, 1460458218810743 (2018).
148. Ridell, K. *et al.* Oral health-related quality-of-life in Swedish children before and after dental treatment under general anesthesia. *Acta Odontologica Scandinavica* **73**, 1-7 (2015).
149. Cantekin, K., Yildirim, M.D. & Cantekin, I. Assessing change in quality of life and dental anxiety in young children following dental rehabilitation under general anesthesia. *Pediatric Dentistry* **36**, 12E-17E (2014).

150. Jankauskiene, B., Virtanen, J.I., Kubilius, R. & Narbutaite, J. Oral health-related quality of life after dental general anaesthesia treatment among children: a follow-up study. *BMC Oral Health* **14**, 7 (2014).
151. Gaynor, W.N. & Thomson, W.M. Changes in young children's OHRQoL after dental treatment under general anaesthesia. *International Journal of Paediatric Dentistry* **22**, 258-264 (2012).
152. Health and Social Care Information Centre. Focus on the health and care of young people June 2015. Published by the Health and Social Care Information Centre (2015).
153. Moles, D.R. & Ashley, P. Hospital admissions for dental care in children: England 1997-2006. *British Dental Journal* **206**, 5 (2009).
154. Health and Social Care Information Centre. NHS Outcomes Framework: February 2016 Quarterly Publication. Published by the Health and Social Care Information Centre (2016)
155. Raja, A. *et al.* Characteristics of children undergoing dental extractions under general anaesthesia in Wolverhampton: 2007-2012. *British Dental Journal* **220**, 407-411 (2016).
156. Knapp, R., Gilchrist, F.,, Rodd, H. & Marshmann, Z. Change in children's oral health-related quality of life following dental treatment under general anaesthesia for the management of dental caries: a systematic review. *International Journal of Paediatric Dentistry* **27,** 302-312 (2017).
157. Park, J.S., Anthonappa, R.P., King, N.M. & McGrath, C.P. The family impact of dental general anaesthesia in children: a meta-analysis. *International Journal of Paediatric Dentistry* **29***,* 149-161 (2019).
158. Taskinen, H. *et al.* Self-reported causes for referral to dental treatment under general anaesthesia (DGA): a cross-sectional survey. *European Archives of Paediatric Dentistry* **15**, 105-112 (2014).
159. Ramdaw, A., Hosey, M.T. & Bernabé, E. Factors associated with use of general anaesthesia for dental procedures among British children. *British Dental Journal* **223**, 339 (2017).
160. Boyd, A. *et al.* Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* **42**, 111-127 (2013).
161. Local Government Act 1972. Published by Her Majesty's Stationery Office, London, (1972).
162. Fraser, A. *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* **42**, 97-110 (2013).
163. Corah, N.L. Development of a Dental Anxiety Scale. *Journal of Dental Research* **48**, 596-& (1969).
164. Corah, N.L., Gale, E.N. & Illig, S.J. Assessment of a Dental Anxiety Scale. *Journal of the American Dental Association* **97**, 816-819 (1978).
165. Karazsia, B.T. & van Dulmen, M.H.M. Regression Models for Count Data: Illustrations using Longitudinal Predictors of Childhood Injury. *Journal of Pediatric Psychology* **33**, 1076-1084 (2008).
166. Yang, S., Harlow, L. I., Puggioni, G., & Redding, C. A. A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys. *Journal of Modern Applied Statistical Methods* **16,** 29 (2017).
167. Preisser, J.S., Long, D.L. & Stamm, J.W. Matching the Statistical Model to the Research Question for Dental Caries Indices with Many Zero Counts. *Caries Research* **51**, 198-208 (2017).
168. Lewsey, J.D. & Thomson, W.M. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology* **32**, 183-189 (2004).
169. Preisser, J.S., Das, K., Long, D.L. & Divaris, K. Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine* **35**, 1722-1735 (2016).

170. Tahmassebi, J.F., Achol, L.T. & Fayle, S.A. Analysis of dental care of children receiving comprehensive care under general anaesthesia at a teaching hospital in England. *European Archives of Paediatric Dentistry : Official Journal of the European Academy of Paediatric Dentistry* **15**, 353-60 (2014).
171. Almeida, A.G., Roseman, M., Sheff, M. & Huntington, N. Future caries susceptibility in children with early childhood caries following treatment under general anesthesia. *Pediatric Dentistry* **22,** 302-306 (2000).
172. Thomson, W.M. Day-stay treatment for dental caries at a New Zealand hospital dental unit: a 5-year retrospective audit. *The New Zealand Dental Journal* **90**, 139-142 (1994).
173. McAuliffe, Ú., Kinirons, M. & Harding, M. A retrospective investigation of the oral health records of a cohort of preschool children who received extractions under general anaesthesia including cost analysis of treatment. *Journal of the Irish Dental Association* **63,** 38-44 (2017).
174. Aldossari, G.S. *et al.* The long-term effect of previous dental treatment under general anaesthesia on children's dental fear and anxiety. *International Journal of Paediatric Dentistry* **29,** 177-184 (2019)
175. Adewale, L., Morton, N. & Blayney, M. Guidelines For The Management Of Children Referred For Dental Extractions Under General Anaesthesia. Published by the Association of Paediatric Anaesthetists of Great Britain and Ireland, in collaboration with the Association of Dental Anaesthetists; the British Society of Paediatric Dentistry; the Royal College of Anaesthetists; the Association of Anaesthetists of Great Britain and Ireland and the Royal College of Nursing (2011).
176. Brown, L., Kenny, K. & O'Sullivan, E. Dental general anaesthetic pre-assessments completed by a specialist—does it change patient outcomes? A UK-based study. *International Journal of Paediatric Research* **29,** 162-168 (2019)
177. Fayle, S. Improving oral healthcare for children – a great opportunity. *British Dental Journal* **214**, 547 (2013).
178. Olley, R.C., Hosey, M.T., Renton, T. & Gallagher, J. Why are children still having preventable extractions under general anaesthetic? A service evaluation of the views of parents of a high caries risk group of children. *British Dental Journal* **210,** E13 (2011).
179. Aljafari, A.K., Scambler, S., Gallagher, J.E. & Hosey, M.T. Parental views on delivering preventive advice to children referred for treatment of dental caries under general anaesthesia: A qualitative investigation. *Community Dental Health* **31**, 75-79 (2014).
180. Aljafari, A.K., Gallagher, J.E. & Hosey, M.T. Failure on all fronts: general dental practitioners' views on promoting oral health in high caries risk children- a qualitative study. *BMC Oral Health* **15**, 45 (2015).
181. Divaris, K. & Joshi, A. The building blocks of precision oral health in early childhood: the ZOE 2.0 study. *Journal of Public Health Dentistry* epublication ehead of print (2018).
182. Divaris, K. Predicting Dental Caries Outcomes in Children: A "Risky" Concept. *Journal of Dental Research* **95**, 248-254 (2016).
183. Rodd, H. *et al.* 'I felt weird and wobbly.' Child-reported impacts associated with a dental general anaesthetic. *British Dental Journal* **216,** E17 (2014).
184. Klaassen, M.A., Veerkamp, J.S.J. & Hoogstraten, J. Young children's Oral Health-Related Quality of Life and dental fear after treatment under general anaesthesia: a randomized controlled trial. *European Journal of Oral Sciences* **117**, 273-278 (2009).
185. Balmer, R., O'Sullivan, E.A., Pollard, M.A., Curzon, M.E. Anxiety related to dental general anaesthesia: changes in anxiety in children and their parents. *European Journal of Paediatric Dentistry* **5**, 9-14 (2004).
186. Savanheimo, N. & Vehkalahti, M.M. Five-year follow-up of children receiving comprehensive dental care under general anesthesia. *BMC Oral Health* **14**, 154 (2014).

187. Wolke, D. *et al.* Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *British Journal of Psychiatry* **195**, 249-256 (2009).
188. Humphris, G., Crawford, J.R., Hill, K., Gilbert, A. & Freeman, R. UK population norms for the modified dental anxiety scale with percentile calculator: adult dental health survey 2009 results. *BMC Oral Health* **13**, 11 (2013).
189. Similä, T., Nieminen, P. & Virtanen, J.I. Validity of self-reported number of teeth in middle-aged Finnish adults: the Northern Finland Birth Cohort Study 1966. *BMC Oral Health* **18**, 210 (2018).
190. Meisel, P., Holtfreter, B., Völzke, H. & Kocher, T. Self-reported oral health predicts tooth loss after five and ten years in a population-based study. *Journal of Clincial Periodontology* **45**, 1164-1172 (2018).
191. WHO. Oral health surveys: basic methods - 5th edition. Published by the World Health Organization at http://www.who.int/oral_health/publications/9789241548649/en/ (2013).
192. Mew, J. & Trenouth, M.J.I.J.D.O.S.S. How Many Teeth are Extracted as Part of Orthodontic Treatment? A Survey of 2038 UK Residents. **1**, 1-5 (2018).
193. Broadbent, J.M. & Thomson, W.M. For debate: problems with the DMF index pertinent to dental caries data analysis. *Community Dentistry and Oral Epidemiology* **33**, 400-409 (2005).
194. Lawrence, H.P., Beck, J.D., Hunt, R.J. & Koch, G.G. Adjustment of the M-component of the DMFS index for prevalence studies of older adults. *Community Dentistry and Oral Epidemiology* **24**, 322-331 (1996).
195. Joshy, G., Arora, M., Korda, R.J., Chalmers, J. & Banks, E. Is poor oral health a risk marker for incident cardiovascular disease hospitalisation and all-cause mortality? Findings from 172 630 participants from the prospective 45 and Up Study. *BMJ Open* **6**(2016).
196. Vedin, O. *et al.* Tooth loss is independently associated with poor outcomes in stable coronary heart disease. *European Journal of Preventive Cardiology* **23**, 839-846 (2016).
197. Naorungroj, S. *et al.* Tooth loss, periodontal disease, and cognitive decline in the Atherosclerosis Risk in Communities (ARIC) study. *Community Dentistry and Oral Epidemiology* **43**, 47-57 (2015).
198. Asai, K. *et al.* Tooth Loss and Atherosclerosis: The Nagahama Study. *Journal of Dental Research* **94**, 52S-58S (2015).
199. Liljestrand, J.M. *et al.* Missing Teeth Predict Incident Cardiovascular Events, Diabetes, and Death. *Journal of Dental Research* **94**, 1055-1062 (2015).
200. Tak, I.H. *et al.* The association between periodontal disease, tooth loss and bone mineral density in a Korean population. *Journal of Clinical Periodontology* **41**, 1139-1144 (2014).
201. You, Z., Cushman, M., Jenny, N.S. & Howard, G. Tooth loss, systemic inflammation, and prevalent stroke among participants in the reasons for geographic and racial difference in stroke (REGARDS) study. *Atherosclerosis* **203**, 615-619 (2009).
202. Aida, J. *et al.* Reasons for permanent tooth extractions in Japan. *Journal of Epidemiology* **16**, 214-219 (2006).
203. Murray, H., Locker, D. & Kay, E.J. Patterns of and reasons for tooth extractions in general dental practice in Ontario, Canada. *Community Dentistry and Oral Epidemiology* **24**, 196-200 (1996).
204. Kay, E.J. & Blinkhorn, A.S. The Reasons for the Extraction of Various Tooth Types in Scotland. *Journal of Dentistry* **15**, 30-33 (1987).
205. Chestnutt, I.G., Binnie, V.I. & Taylor, M.M. Reasons for tooth extraction in Scotland. *Journal of Dentistry* **28**, 295-397 (2000).
206. Trovik, T.A., Klock, K.S. & Haugejorden, O. Trends in reasons for tooth extractions in Norway from 1968 to 1998. *Acta Odontologica Scandinavica* **58**, 89-96 (2000).
207. Celeste, R.K., Nadanovsky, P. & Fritzell, J. Trends in socioeconomic disparities in oral health in Brazil and Sweden. *Community Dentistry and Oral Epidemiology* **39**, 204-212 (2011).

208. Elani, H.W. *et al.* Social inequalities in tooth loss: A multinational comparison. *Community Dentistry and Oral Epidemiology* **45**, 266-274 (2017).
209. Naorungroj, S. *et al.* Racial differences in periodontal disease and 10-year self-reported tooth loss among late middle-aged and older adults: the dental ARIC study. *Journal of Public Health Dentistry* **77**, 372-382 (2017).
210. Munoz-Torres, F.J., Mukamal, K.J., Pai, J.K., Willett, W. & Joshipura, K.J. Relationship between tooth loss and peripheral arterial disease among women. *Journal of Clinical Periodontology* **44**, 989-995 (2017).
211. Lee, J.H. *et al.* Trends in the incidence of tooth extraction due to periodontal disease: results of a 12-year longitudinal cohort study in South Korea. *Journal of Periodontal and Implant Science* **47**, 264-272 (2017).
212. Saminsky, M., Halperin-Sternfeld, M., Machtei, E.E. & Horwitz, J. Variables affecting tooth survival and changes in probing depth: a long-term follow-up of periodontitis patients. *Journal of Clinical Periodontology* **42**, 513-519 (2015).
213. Stadker, A.F.M., Marina, Oppermann, R.V. & Gomes, S.C. Tooth Loss in Patients under Periodontal Maintenance in a Private Practice: A Retrospective Study. *Brazilian Dental Journal* **28**, 440-446 (2017).
214. Broadbent, J.M., Foster Page, L.A., Thomson, W.M. & Poulton, R. Permanent dentition caries through the first half of life. *British Dental Journal* **215**(2013).
215. Broadbent, J.M., Thomson, W.M. & Poulton, R. Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. *Journal of Dental Research* **87**, 69-72 (2008).
216. Broadbent, J.M., Thomson, W.M. & Poulton, R. Progression of dental caries and tooth loss between the third and fourth decades of life: A birth cohort study. *Caries Research* **40**, 459-465 (2006).
217. Kassebaum, N.J. *et al.* Global, Regional, and National Prevalence, Incidence, and Disability-Adjusted Life Years for Oral Conditions for 195 Countries, 1990-2015: A Systematic Analysis for the Global Burden of Diseases, Injuries, and Risk Factors. *Journal of Dental Research* **96**, 380-387 (2017).
218. Rozier, R.G., White, A.B. & Slade, G.D. Trends in Oral Diseases in the U.S. Population. *Journal of Dental Education.* **81**, eS97-eS109 (2017).
219. Cole, S.R. *et al.* Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* **39**, 417-420 (2010).
220. Ordell, S.J.T. Från hantverk till akademisk profession. *Tandlakartidningen* **104**, 74-9 (2012).
221. Hallmans, G. *et al.* Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. *Scandinavian Journal of Public Health. Supplement* **61**, 18-24 (2003).
222. Norberg, M., Wall, S., Boman, K. & Weinehall, L. The Vasterbotten Intervention Programme: background, design and implications. *Global Health Action* **3**(2010).
223. Norberg, M. *et al.* Community participation and sustainability--evidence over 25 years in the Västerbotten Intervention Programme. *Global Health Action* **5**, 1-9 (2012).
224. Reinholdsson, T. swemaps R package. Published online at https://github.com/reinholdsson/swemaps (2015).
225. Blomstedt, Y. *et al.* Impact of a combined community and primary care prevention strategy on all-cause and cardiovascular mortality: a cohort analysis based on 1 million person-years of follow-up in Västerbotten County, Sweden, during 1990–2006. *BMJ Open* **5**, e009651 (2015).
226. Bodén, S. *et al.* Dietary inflammatory index and risk of first myocardial infarction; a prospective population-based study. *Nutrition Journal* **16**, 21-21 (2017).
227. Varga, T.V. *et al.* Genetic Determinants of Long-Term Changes in Blood Lipid Concentrations: 10-Year Follow-Up of the GLACIER Study. *PLoS Genetics* **10** (2014).

228. Kurbasic, A. *et al.* Gene-Lifestyle Interactions in Complex Diseases: Design and Description of the GLACIER and VIKING Studies. *Current Nutrition Reports* **3,** 400-411 (2014).
229. Kweon, S. *et al.* Data Resource Profile: The Korea National Health and Nutrition Examination Survey (KNHANES). *International Journal of Epidemiology* **43**, 69-77 (2014).
230. Sejong, C. National Health Statistics 2013: Korean National Health and Nutritional Examination Survey (KNHANES VI-I). Published by the Ministry of Health and Welfare, (2014).
231. Sejong, C. Standardization for oral health survey in KNHANES (2010). Published by the Korea Centers for Disease Control and Prevention (2011).
232. Caton, J.G. *et al.* A new classification scheme for periodontal and peri-implant diseases and conditions – Introduction and key changes from the 1999 classification. *Journal of Clinical Periodontology* **45**, S1-S8 (2018).
233. Zhang, Z. Parametric regression model for survival data: Weibull regression model as an example. *Annals of Translational Medicine* **4**, 484-484 (2016).
234. Rietveld, C.A. *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences* **111**, 13790 (2014).
235. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624-633 (2016).
236. Hirotomi, T., Yoshihara, A., Ogawa, H. & Miyazaki, H. Tooth-related risk factors for periodontal disease in community-dwelling elderly people. *Journal of Clinical Periodontology* **37**, 494-500 (2010).
237. Padbury, A., Eber, R. & Wang, H.L. Interactions between the gingiva and the margin of restorations. *Journal of Clinical Periodontology* **30**, 379-385 (2003).
238. Duckworth, R.M. & Huntington, E. Evidence for putting the calculus: caries inverse relationship to work. *Community Dentistry and Oral Epidemiology* **33**, 349-356 (2005).
239. Griffen, A.L. *et al.* Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *Isme Journal* **6**, 1176-1185 (2012).
240. Nakahara, Y., Ozaki, K. & Matsuura, T. Long-term Hyperglycemia Naturally Induces Dental Caries but Not Periodontal Disease in Type 1 and Type 2 Diabetic Rodents. *Diabetes* **66**, 2868 (2017).
241. Jepson, N.J.A., Moynihan, P.J., Kelly, P.J., Watson, G.W. & Thomason, J.M. Caries incidence following restoration of shortened lower dental arches in a randomized controlled trial. *British Dental Journal* **191**, 140-144 (2001).
242. Leroy, R., Eaton, K.A. & Savage, A. Methodological issues in epidemiological studies of periodontitis - how can it be improved? *BMC Oral Health* **10**(2010).
243. Heyman, R.E. *et al.* Dental Fear and Avoidance in Treatment Seekers at a Large, Urban Dental Clinic. *Oral Health & Preventive Dentistry* **14**, 315-320 (2016).
244. Rechmann, P., Jue, B., Santo, W., Rechmann, B.M.T. & Featherstone, J.D.B. Calibration of dentists for Caries Management by Risk Assessment Research in a Practice Based Research Network - CAMBRA PBRN. *BMC Oral Health* **18**, 2 (2018).
245. Chapple, I.L.C. *et al.* Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. *Journal of Clinical Periodontology* **44**, S39-S51 (2017).
246. Marioni, R.E. *et al.* GWAS on family history of Alzheimer's disease. *Translational psychiatry* **8**, 99-99 (2018).
247. Ferreira, M.A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature Genetics* **49**, 1752 (2017).

248. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *American Journal of Epidemiology* **129**, 687-702 (1989).
249. Joshu, C.E. *et al.* Enhancing the Infrastructure of the Atherosclerosis Risk in Communities (ARIC) Study for Cancer Epidemiology Research: ARIC Cancer. *Cancer Epidemiology Biomarkers & Prevention* **27**, 295-305 (2018).
250. Beck James, D. *et al.* Relationship of Periodontal Disease to Carotid Artery Intima-Media Wall Thickness. *Arteriosclerosis, Thrombosis, and Vascular Biology* **21**, 1816-1822 (2001).
251. Cornelis, M.C. *et al.* The Gene, Environment Association Studies Consortium (GENEVA): Maximizing the Knowledge Obtained from GWAS by Collaboration Across Studies of Multiple Conditions. *Genetic Epidemiology* **34**, 364-372 (2010).
252. Polk, D.E. *et al.* Study protocol of the Center for Oral Health Research in Appalachia (COHRA) etiology study. *BMC Oral Health* **8**, 18 (2008).
253. Wang, X. *et al.* Genome-wide association scan of dental caries in the permanent dentition. *BMC Oral Health* **12**, 57 (2012).
254. Anjomshoaa, I., Cooper, M.E. & Vieira, A.R. Caries is Associated with Asthma and Epilepsy. *European Journal of Dentistry* **3**, 297-303 (2009).
255. Lavange, L.M. *et al.* Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology* **20**, 642-649 (2010).
256. Berglund, G., Elmstahl, S., Janzon, L. & Larsson, S.A. The Malmö Diet and Cancer Study. Design and feasibility. *Journal of Internal Medicine* **233**, 45-51 (1993).
257. Smith, J.G., Platonov, P.G., Hedblad, B., Engström, G. & Melander, O. Atrial fibrillation in the Malmö diet and cancer study: a study of occurrence, risk factors and diagnostic validity. *European Journal of Epidemiology* **25**, 95-102 (2010).
258. Rantakallio, P. The longitudinal study of the Northern Finland birth cohort of 1966. *Paediatric and Perinatal Epidemiology* **2**, 59-88 (1988).
259. Laajala, A. *et al.* Association of indirect restorations with past caries history and present need for restorative treatment in the Northern Finland Birth Cohort 1966. *Clinical Oral Investigations* **22**, 1495-1501 (2018).
260. Völzke, H. *et al.* Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology* **40**, 294-307 (2010).
261. Splieth, C. *et al.* Caries prevalence in an adult population: results of the Study of Health in Pomerania, Germany (SHIP). *Oral Health & Preventive Dentistry* **1**, 149-55 (2003).
262. Schützhold, S. *et al.* Changes in prevalence of periodontitis in two German population-based studies. *Journal of Clinical Periodontology* **42**, 121-130 (2015).
263. Lichtenstein, P. *et al.* The Swedish Twin Registry in the Third Millennium: An Update. *Twin Research and Human Genetics* **9**, 875-882 (2006).
264. Magnusson, P.K.E. *et al.* The Swedish Twin Registry: Establishment of a Biobank and Other Recent Developments. *Twin Research and Human Genetics* **16**, 317-329 (2013).
265. Chen, X. *et al.* Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *American Journal of Human Genetics* **97**, 708-714 (2015).
266. Ridker, P.M. *et al.* Rationale, Design, and Methodology of the Women's Genome Health Study: A Genome-Wide Association Study of More Than 25 000 Initially Healthy American Women. *Clinical Chemistry* **54**, 249 (2008).
267. Ridker, P.M. *et al.* A Randomized Trial of Low-Dose Aspirin in the Primary Prevention of Cardiovascular Disease in Women. *New England Journal of Medicine* **352**, 1293-1304 (2005).
268. Lee, I.M. *et al.* Vitamin E in the Primary Prevention of Cardiovascular Disease and CancerThe Women's Health Study: A Randomized Controlled Trial. *JAMA: The Journal of the American Medical Association* **294**, 56-65 (2005).

269. Yu, Y.-H., Chasman, D.I., Buring, J.E., Rose, L. & Ridker, P.M. Cardiovascular risks associated with incident and prevalent periodontal disease. *Journal of Clinical Periodontology* **42**, 21-28 (2015).
270. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology* **27**, S2-S8 (2017).
271. Sudo, T. *et al.* Association of NOD2 Mutations with Aggressive Periodontitis. *Journal of Dental Research* **96**, 1100-1105 (2017).
272. Hirata, J. *et al.* Variants at HLA-A, HLA-C, and HLA-DQB1 Confer Risk of Psoriasis Vulgaris in Japanese. *The Journal of Investigative Dermatology* **138**, 542-548 (2018).
273. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
274. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253-1260 (2008).
275. Delaneau, O., Coulonges, C. & Zagury, J.-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
276. Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**, 238-251 (2012).
277. Loh, P.-R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* **48**, 811 (2016).
278. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284 (2016).
279. Howie, B.N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* **5**, e1000529 (2009).
280. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics* **10**, e1004234 (2014).
281. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816-834 (2010).
282. Wellcome Trust Centre for Human Genetics. Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped, prospective resource. (Published online by the Wellcome Trust Centre for HUman Genetics at http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580 (2015).
283. Mitchell, R.E. *et al.* UK Biobank Genetic Data : MRC IEU Quality Control, version 2. Published online by the University of Bristol at 10.5523/bris.1ovaau5sxunp2cv8rcy88688v (2019).
284. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* **9**, 1192-1212 (2014).
285. Page, R.C. & Eke, P.I. Case Definitions for Use in Population-Based Surveillance of Periodontitis. *Journal of Periodontology* **78**, 1387-1399 (2007).
286. Armitage, G.C. Development of a Classification System for Periodontal Diseases and Conditions. *Annals of Periodontology* **4**, 1-6 (1999).
287. Conomos, Matthew P. *et al.* Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *American Journal of Human Genetics* **98**, 165-184 (2016).
288. Elsworth, B. *et al.* MRC IEU UK Biobank GWAS pipeline version 1. Published online by The University of Bristol at 10.5523/bris.2fahpksont1zi26xosyamqo8rr (2017).
289. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81** (2007).
290. Conomos, M. *et al.* GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population

structure and/or relatedness. R package published online at https://bioconductor.org/packages/release/bioc/html/GENESIS.html (2019).
291. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499-511 (2010).
292. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134-134 (2010).
293. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010).
294. Steffens, M. *et al.* SNP-Based Analysis of Genetic Substructure in the German Population. *Human Heredity* **62**, 20-29 (2006).
295. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
296. Winkler, T.W. *et al.* EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data. *Bioinformatics* **31**, 259-261 (2015).
297. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369 (2012).
298. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82 (2015).
299. Yang, J.A., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics* **88**, 76-82 (2011).
300. Miretti, M.M. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *American Journal of Human Genetics* **76**, 634-646 (2005).
301. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228 (2015).
302. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50**, 621-629 (2018).
303. Zhernakova, D.V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* **49**, 139 (2016).
304. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
305. Staley, J.R. *et al.* PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207-3209 (2016).
306. e, G.P. *et al.* Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature Genetics* **49**, 1664 (2017).
307. Barbeira, A.N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Preprint at *bioRxiv* https://doi.org/10.1101/045260 (2017).
308. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6**, 9 (2015).
309. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300 (1995).
310. Nibali, L., Di Iorio, A., Tu, Y.-K. & Vieira, A.R. Host genetics role in the pathogenesis of periodontal disease and caries. *Journal of Clinical Periodontology* **44**, S52-S78 (2017).
311. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).
312. Munz, M. *et al.* A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. *Human Molecular Genetics* **26**, 2577-2588 (2018).

313. Ozawa, Y., Chiba, J. & Sakamoto, S. HLA class II alleles and salivary numbers of mutans streptococci and lactobacilli among young adults in Japan. *Oral Microbiology and Immunology* **16**, 353-357 (2001).
314. Reibring, C.-G. *et al.* Expression patterns and subcellular localization of carbonic anhydrases are developmentally regulated during tooth formation. *PloS ONE* **9**, e96007-e96007 (2014).
315. Hong, J.H. *et al.* Essential role of carbonic anhydrase XII in secretory gland fluid and HCO3 (-) secretion revealed by disease causing human mutation. *The Journal of Physiology* **593**, 5299-5312 (2015).
316. Mitsiadis, T.A. & Drouin, J. Deletion of the Pitx1 genomic locus affects mandibular tooth morphogenesis and expression of the Barx1 and Tbx1 genes. *Developmental Biology* **313**, 887-896 (2008).
317. Astle, W.J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
318. Järvinen, E. *et al.* Continuous tooth generation in mouse is induced by activated epithelial Wnt/β-catenin signaling. *Proceedings of the National Academy of Sciences* **103**, 18627-18632 (2006).
319. Han, N. *et al.* β-Catenin Enhances Odontoblastic Differentiation of Dental Pulp Cells through Activation of Runx2. *PLoS ONE* **9**, e88890 (2014).
320. Chen, J., Lan, Y., Baek, J.-A., Gao, Y. & Jiang, R. Wnt/beta-catenin signaling plays an essential role in activation of odontogenic mesenchyme during early tooth development. *Developmental Biology* **334**, 174-185 (2009).
321. Plaisancié, J. *et al.* Mutations in WNT10A are frequently involved in oligodontia associated with minor signs of ectodermal dysplasia. *American Journal of Medical Genetics Part A* **161**, 671-678 (2013).
322. van den Boogaard, M.-J. *et al.* Mutations in WNT10A are present in more than half of isolated hypodontia cases. *Journal of Medical Genetics* **49**, 327-331 (2012).
323. Mostowska, A. *et al.* Nucleotide variants of genes encoding components of the Wnt signalling pathway and the risk of non-syndromic tooth agenesis. *Clinical Genetics* **84**, 429-440 (2013).
324. Jonsson, L. *et al.* Rare and Common Variants Conferring Risk of Tooth Agenesis. *Journal of Dental Research* **97**, 515-522 (2018).
325. Bohring, A. *et al.* WNT10A Mutations Are a Frequent Cause of a Broad Spectrum of Ectodermal Dysplasias with Sex-Biased Manifestation Pattern in Heterozygotes. *The American Journal of Human Genetics* **85**, 97-105 (2009).
326. Tamura, M. & Nemoto, E. Role of the WNT signaling molecules in the tooth. *Japanese Dental Science Review* **52**, 75-83 (2016).
327. Kimura, R. *et al.* Common polymorphisms in WNT10A affect tooth morphology as well as hair shape. *Human Molecular Genetics* **24**, 2673-2680 (2015).
328. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335-1341 (2018).
329. Li, B. *et al.* Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pacific Symposium on Biocomputing* **23**, 448-459 (2018).
330. Hendrie, C.A. & Brewer, G. Evidence to suggest that teeth act as human ornament displays signalling mate quality. *PloS ONE* **7**, e42178-e42178 (2012).
331. Khalid, A. & Quiñonez, C. Straight, white teeth as a social prerogative. *Sociology of Health & Illness* **37**, 782-796 (2015).
332. Mucci, L.A., Bjorkman, L., Douglass, C.W. & Pedersen, N.L. Environmental and heritable factors in the etiology of oral diseases - A population-based study of Swedish twins. *Journal of Dental Research* **84**, 800-805 (2005).
333. Yu, Y.H. *et al.* Family History of MI, Smoking, and Risk of Periodontal Disease. *Journal of Dental Research* **97**, 1106-1113 (2018).

334. Shi, Q. *et al.* Association between Myocardial Infarction and Periodontitis: A Meta-Analysis of Case-Control Studies. *Frontiers in Physiology* **7**(2016).
335. Sedgwick, P. Cross sectional studies: advantages and disadvantages. *British Medical Journal* **348**, g2276 (2014).
336. DeStefano, F., Anda, R.F., Kahn, H.S., Williamson, D.F. & Russell, C.M. Dental disease and risk of coronary heart disease and mortality. *British Medical Journal* **306**, 688-691 (1993).
337. Beck, J.D., Offenbacher, S., Williams, R., Gibbs, P. & Garcia, R. Periodontitis: A Risk Factor for Coronary Heart Disease? *Annals of Periodontology* **3**, 127-141 (1998).
338. Liljestrand, J.M. *et al.* Missing Teeth Predict Incident Cardiovascular Events, Diabetes, and Death. *Journal of Dental Research* **94**, 1055-62 (2015).
339. Cheng, F. *et al.* Tooth loss and risk of cardiovascular disease and stroke: A dose-response meta analysis of prospective cohort studies. *PLoS ONE* **13**, e0194563 (2018).
340. Liljestrand, J.M. *et al.* Association of Endodontic Lesions with Coronary Artery Disease. *Journal of Dental Research* **95**, 1358-1365 (2016).
341. Winning, L., Patterson, C.C., Neville, C.E., Kee, F. & Linden, G.J. Periodontitis and incident type 2 diabetes: a prospective cohort study. *Journal of Clinical Periodontology* **44**, 266-274 (2017).
342. Myllymäki, V. *et al.* Association between periodontal condition and the development of type 2 diabetes mellitus—Results from a 15-year follow-up study. *Journal of Clinical Periodontology* **45**, 1276-1286 (2018).
343. Graziani, F., Gennai, S., Solini, A. & Petrini, M. A systematic review and meta-analysis of epidemiologic observational evidence on the effect of periodontitis on diabetes An update of the EFP-AAP review. *Journal of Clincial Periodontology* **45**, 167-187 (2018).
344. Hirschfeld, J. & Kawai, T. Oral Inflammation and Bacteremia: Implications for Chronic and Acute Systemic Diseases Involving Major Organs. *Cardiovascular & Hematological Disorders - Drug Targets* **15**, 70-84 (2015).
345. Singhrao, S.K. *et al.* Oral Inflammation, Tooth Loss, Risk Factors, and Association with Progression of Alzheimer's Disease. *Journal of Alzheimer's Disease* **42**, 723-737 (2014).
346. Nepomuceno, R. *et al.* Serum lipid levels in patients with periodontal disease: A meta-analysis and meta-regression. *Journal of Clinical Periodontology* **44**, 1192-1207 (2017).
347. Voight, B.F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572-580 (2012).
348. Li, L.-W., Wong, H.M. & McGrath, C.P. Longitudinal Association between Obesity and Dental Caries in Adolescents. *The Journal of Pediatrics* **189**, 149-154.e5 (2017).
349. Gaewkhiew, P., Sabbah, W. & Bernabé, E. Does tooth loss affect dietary intake and nutritional status? A systematic review of longitudinal studies. *Journal of Dentistry* **67**, 1-8 (2017).
350. Welmer, A.-K., Rizzuto, D., Parker, M.G. & Xu, W. Impact of tooth loss on walking speed decline over time in older adults: a population-based cohort study. *Aging Clinical and Experimental Research* **29**, 793-800 (2017).
351. Duncan, B.B. *et al.* Low-Grade Systemic Inflammation and the Development of Type 2 Diabetes: The Atherosclerosis Risk in Communities Study**.** *Diabetes* **52**, 1799-1805 (2003).
352. Torrungruang, K., Ongphiphadhanakul, B., Jitpakdeebordin, S. & Sarujikumjornwatana, S. Mediation analysis of systemic inflammation on the association between periodontitis and glycaemic status. *Journal of Clinical Periodontology* **45**, 548-556 (2018).
353. Wiener, R.C., Shen, C., Findley, P.A., Sambamoorthi, U. & Tan, X. The association between diabetes mellitus, sugar-sweetened beverages, and tooth loss in adults: Evidence from 18 states. *The Journal of the American Dental Association* **148**, 500-509.e4 (2017).
354. Mai, X. *et al.* Associations between smoking and tooth loss according to the reason for tooth loss: the Buffalo OsteoPerio Study. *Journal of the American Dental Association (1939)* **144**, 252-265 (2013).

355. Kim, Y.H., Han, K., Vu, D., Cho, K.-H. & Lee, S.H. Number of remaining teeth and its association with socioeconomic status in South Korean adults: Data from the Korean National Health and Nutrition Examination Survey 2012-2013. *PLoS ONE* **13**, e0196594 (2018).
356. Poudel, P. *et al.* Oral health knowledge, attitudes and care practices of people with diabetes: a systematic review. *BMC Public Health* **18**, 577 (2018).
357. Barbaro, N., Boutwell, B.B., Barnes, J.C. & Shackelford, T.K. Genetic confounding of the relationship between father absence and age at menarche. *Evolution and Human Behavior* **38**, 357-365 (2017).
358. Kobayashi, T. *et al.* The KCNQ1 gene polymorphism as a shared genetic risk for rheumatoid arthritis and chronic periodontitis in Japanese adults: A pilot case-control study. *Journal of Periodontology* **89**, 315-324 (2018).
359. Battram, T. *et al.* Coronary artery disease, genetic risk and the metabolome in young individuals. Epublication in *Wellcome Open Research* **3** (2019).
360. Hansson, G.K. & Hermansson, A. The immune system in atherosclerosis. *Nature Immunology* **12**, 204 (2011).
361. Tulamo, R., Frösen, J., Hernesniemi, J. & Niemelä, M. Inflammatory changes in the aneurysm wall: a review. *Journal of NeuroInterventional Surgery* **10**, i58 (2018).
362. Koletsi, D., Valla, K., Fleming, P.S., Chaimani, A. & Pandis, N. Assessment of publication bias required improvement in oral health systematic reviews. *Journal of Clinical Epidemiology* **76**, 118-124 (2016).
363. Papageorgiou, S.N., Kloukos, D., Petridis, H. & Pandis, N. Publication of statistically significant research findings in prosthodontics & implant dentistry in the context of other dental specialties. *Journal of Dentistry* **43**, 1195-1202 (2015).
364. de Vries, Y.A., Roest, A.M., Franzen, M., Munafò, M.R. & Bastiaansen, J.A. Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene (5-HTTLPR), life stress and depression. *Psychological Medicine* **46**, 2971-2979 (2016).
365. de Vries, Y.A. *et al.* The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine* **48**, 2453-2455 (2018).
366. Mousa, A., Teede, H., Naderpoor, N., de Courten, B. & Scragg, R. Vitamin D supplementation for improvement of chronic low-grade inflammation in patients with type 2 diabetes: a systematic review and meta-analysis of randomized controlled trials. *Nutrition Reviews* **76**, 380-394 (2018).
367. Angellotti, E. *et al.* Vitamin D Supplementation in Patients With Type 2 Diabetes: The Vitamin D for Established Type 2 Diabetes (DDM2) Study. *Journal of the Endocrine Society* **2**, 310-321 (2018).
368. Boeke, A.J.P. *et al.* Effect of moderate-dose vitamin D supplementation on insulin sensitivity in vitamin D–deficient non-Western immigrants in the Netherlands: a randomized placebo-controlled trial. *The American Journal of Clinical Nutrition* **100**, 152-160 (2014).
369. Moreira-Lucas, T.S. *et al.* Effect of vitamin D supplementation on oral glucose tolerance in individuals with low vitamin D status and increased risk for developing type 2 diabetes (EVIDENCE): A double-blind, randomized, placebo-controlled clinical trial. *Diabetes, Obesity and Metabolism* **19**, 133-141 (2017).
370. Artese, H.P.C. *et al.* Periodontal Therapy and Systemic Inflammation in Type 2 Diabetes Mellitus: A Meta-Analysis. *PLoS ONE* **10**, e0128344 (2015).
371. Davey Smith, G. *et al.* Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology. *PLoS Medicine* **4**, e352 (2007).
372. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133-1163 (2008).

373. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**, R89-R98 (2014).
374. Burgess, S. & Labrecque, J.A. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *European Journal of Epidemiology* **33**, 947-952 (2018).
375. Davies, N.M., Holmes, M.V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *British Medical Journal* **362**, k601 (2018).
376. Verbanck, M., Chen, C.Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693-698 (2018).
377. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
378. Burgess, S., Small, D.S. & Thompson, S.G. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26**, 2333-2355 (2015).
379. Davies, N.M. *et al.* The many weak instruments problem and Mendelian randomization. *Statistics in Medicine* **34**, 454-468 (2015).
380. Bowden, J., Davey Smith, G., Haycock, P.C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40**, 304-314 (2016).
381. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S.G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology (Cambridge, Mass.)* **28**, 30-42 (2017).
382. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics* **27**, R195-R208 (2018).
383. Burgess, S., Zuber, V., Gkatzionis, A. & Foley, C.N. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *International Journal of Epidemiology* **47** 1242-1254 (2018).
384. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512-525 (2015).
385. Burgess, S. & Thompson, S.G. Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology* **32**, 377-389 (2017).
386. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018).
387. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121-+ (2015).
388. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics* **50**, 524-537 (2018).
389. Horikoshi, M. *et al.* Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation. *PLoS Genetics* **11**, e1005230 (2015).
390. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274 (2013).
391. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196 (2015).
392. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-U401 (2015).
393. Dietrich, T. *et al.* Evidence summary: the relationship between oral and cardiovascular disease. *British Dental Journal* **222**, 381-385 (2017).

394. Tabeta, K., Yoshie, H. & Yamazaki, K. Current evidence and biological plausibility linking periodontitis to atherosclerotic cardiovascular disease. *Japanese Dental Science Review* **50**, 55-62 (2014).
395. Bartova, J. *et al.* Periodontitis as a risk factor of atherosclerosis. *Journal of Immunology Research* **2014**, 636893-636893 (2014).
396. Haworth, J.A. *et al.* Concerted functions of Streptococcus gordonii surface proteins PadA and Hsa mediate activation of human platelets and interactions with extracellular matrix. *Cellular Microbiology* **19**, e12667 (2017).
397. Glodny, B. *et al.* The occurrence of dental caries is associated with atherosclerosis. *Clinics* **68**, 946-953 (2013).
398. Tonomura, S. *et al.* Intracerebral hemorrhage and deep microbleeds associated with cnm-positive Streptococcus mutans; a hospital cohort study. *Scientific Reports* **6**, 20074 (2016).
399. Hosomi, N. *et al.* Association of Serum Anti-Periodontal Pathogen Antibody with Ischemic Stroke. *Cerebrovascular Diseases* **34**, 385-392 (2012).
400. Pillai, R.S. *et al.* Oral Health and Brain Injury: Causal or Casual Relation? *Cerebrovascular Diseases Extra* **8**, 1-15 (2018).
401. Pussinen Pirkko, J. *et al.* Antibodies to Periodontal Pathogens and Stroke Risk. *Stroke* **35**, 2020-2023 (2004).
402. Kıran, M., Arpak, N., Ünsal, E. & Erdoğan, M.F. The effect of improved periodontal health on metabolic control in type 2 diabetes mellitus. *Journal of Clinical Periodontology* **32**, 266-272 (2005).
403. Stewart, J.E., Wager, K.A., Friedlander, A.H. & Zadeh, H.H. The effect of periodontal treatment on glycemic control in patients with type 2 diabetes mellitus. *Journal of Clincial Periodontology* **28**, 306-310 (2001).
404. Navarro-Sanchez, A.B., Faria-Almeida, R. & Bascones-Martinez, A. Effect of non-surgical periodontal therapy on clinical and immunological response and glycaemic control in type 2 diabetic patients with moderate periodontitis. *Journal of Clinical Periodontology* **34**, 835-843 (2007).
405. Moeintaghavi, A., Arab, H., Bozorgnia, Y., Kianoush, K. & Alizadeh, M. Non-surgical periodontal therapy affects metabolic control in diabetics: a randomized controlled clinical trial. *Australian Dental Journal* **57**, 31-37 (2012).
406. Sgolastra, F., Severino, M., Pietropaoli, D., Gatto, R. & Monaco, A. Effectiveness of Periodontal Treatment to Improve Metabolic Control in Patients With Chronic Periodontitis and Type 2 Diabetes: A Meta-Analysis of Randomized Clinical Trials. *Journal of Periodontology* **84**, 958-973 (2013).
407. Katagiri, S. *et al.* Multi-center intervention study on glycohemoglobin (HbA1c) and serum, high-sensitivity CRP (hs-CRP) after local anti-infectious periodontal treatment in type 2 diabetic patients with periodontal disease. *Diabetes Research and Clinical Practice* **83**, 308-315 (2009).
408. Teshome, A. & Yitayeh, A. The effect of periodontal therapy on glycemic control and fasting plasma glucose level in type 2 diabetic patients: systematic review and meta-analysis. *BMC Oral Health* **17**, 31 (2017).
409. Allen, E.M., Matthews, J.B., O' Halloran, D.J., Griffiths, H.R. & Chapple, I.L. Oxidative and inflammatory status in Type 2 diabetes patients with periodontitis. *Journal of Clinical Periodontology* **38**, 894-901 (2011).
410. Bastos, A.S. *et al.* Lipid Peroxidation Is Associated with the Severity of Periodontal Disease and Local Inflammatory Markers in Patients with Type 2 Diabetes. *The Journal of Clinical Endocrinology & Metabolism* **97**, E1353-E1362 (2012).
411. Polak, D. & Shapira, L. An update on the evidence for pathogenic mechanisms that may link periodontitis and diabetes. *Journal of Clinical Periodontology* **45**, 150-166 (2018).

412. Simpson, T.C. *et al.* Treatment of periodontal disease for glycaemic control in people with diabetes mellitus. *The Cochrane Database of Systematic Reviews* **2015**, CD004714-CD004714 (2015).
413. Li, H. *et al.* Experimental periodontitis induced by Porphyromonas gingivalis does not alter the onset or severity of diabetes in mice. *Journal of Periodontal Research* **48**, 582-590 (2013).
414. Fu, Y.W., Li, X.X., Xu, H.Z., Gong, Y.Q. & Yang, Y. Effects of periodontal therapy on serum lipid profile and proinflammatory cytokines in patients with hyperlipidemia: a randomized controlled trial. *Clinical Oral Investigations* **20**, 1263-1269 (2016).
415. Payne, J.B. *et al.* The effect of subantimicrobial-dose-doxycycline periodontal therapy on serum biomarkers of systemic inflammation: a randomized, double-masked, placebo-controlled clinical trial. *Journal of the American Dental Association (1939)* **142**, 262-273 (2011).
416. Muniz, F.W.M.G. *et al.* Body fat rather than body mass index is associated with gingivitis – A southern Brazilian cross-sectional study. *Journal of Periodontology* **89**, 388-396 (2018).
417. Liu, D. *et al.* High Waist Circumference Obesity has a Greater Association with Periodontitis among a Chinese Population. *Dental Oral Biological and Craniofacial Research* **1**, 1-8 (2018).
418. Su, Y. *et al.* Periodontitis as a Novel Contributor of Adipose Tissue Inflammation Promotes Insulin Resistance in a Rat Model. *Journal of Periodontology* **84**, 1617-1626 (2013).
419. Suresh, S. & Mahendra, J. Multifactorial relationship of obesity and periodontal disease. *Journal of Clinical and Diagnostic Research* **8**, ZE01-ZE3 (2014).
420. Cruz, P., Mehretu, A.M., Buttner, M.P., Trice, T. & Howard, K.M. Development of a polymerase chain reaction assay for the rapid detection of the oral pathogenic bacterium, Selenomonas noxia. *BMC Oral Health* **15**, 95-95 (2015).
421. Goodson, J.M., Groppo, D., Halem, S. & Carpino, E. Is obesity an oral bacterial disease? *Journal of Dental Research* **88**, 519-523 (2009).
422. Swanson, S.A. & Hernán, M.A. The challenging interpretation of instrumental variable estimates under monotonicity. *International Journal of Epidemiology* **47**, 1289-1297 (2018).
423. Seuring, T., Archangelidi, O. & Suhrcke, M. The Economic Costs of Type 2 Diabetes: A Global Systematic Review. *PharmacoEconomics* **33**, 811-831 (2015).
424. Zheng, Y., Ley, S.H. & Hu, F.B. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology* **14**, 88 (2017).
425. Singh, I., Singh, P., Singh, A., Singh, T. & Kour, R. Diabetes an inducing factor for dental caries: A case control analysis in Jammu. *Journal of International Society of Preventive & Community Dentistry* **6**, 125-129 (2016).
426. Puttaswamy, K.A., Puttabudhi, J.H. & Raju, S. Correlation between Salivary Glucose and Blood Glucose and the Implications of Salivary Factors on the Oral Health Status in Type 2 Diabetes Mellitus Patients. *Journal of International Society of Preventive & Community Dentistry* **7**, 28-33 (2017).
427. Choi, Y.-H. *et al.* Association between periodontitis and impaired fasting glucose and diabetes. *Diabetes Care* **34**, 381-386 (2011).
428. Nascimento, G.G., Leite, F.R.M., Vestergaard, P., Scheutz, F. & López, R.J.A.D. Does diabetes increase the risk of periodontitis? A systematic review and meta-regression analysis of longitudinal prospective studies. *Acta Diabetologica* **55**, 653-667 (2018).
429. Morita, I. *et al.* Relationship between Periodontal Status and Levels of Glycated Hemoglobin. *Journal of Dental Research* **91**, 161-166 (2011).
430. Cekici, A., Kantarci, A., Hasturk, H. & Van Dyke, T.E. Inflammatory and immune pathways in the pathogenesis of periodontal disease. *Periodontology 2000* **64**, 57-80 (2014).
431. Ara, T. *et al.* Human gingival fibroblasts are critical in sustaining inflammation in periodontal disease. *Journal of Periodontal Research* **44**, 21-27 (2009).

432. Buranasin, P. *et al.* High glucose-induced oxidative stress impairs proliferation and migration of human gingival fibroblasts. *PLoS ONE* **13**, e0201855 (2018).
433. Kido, D. *et al.* Impact of diabetes on gingival wound healing via oxidative stress. *PLoS ONE* **12**, e0189601-e0189601 (2017).
434. Kumar, S., Padmashree, S. & Jayalekshmi, R. Correlation of salivary glucose, blood glucose and oral candidal carriage in the saliva of type 2 diabetics: A case-control study. *Contemporary Clinical Dentistry* **5**, 312-317 (2014).
435. Gupta, S. *et al.* Correlation of salivary glucose level with blood glucose level in diabetes mellitus. *Journal of Oral and Maxillofacial Pathology* **21**, 334-339 (2017).
436. Mascarenhas, P., Fatela, B. & Barahona, I. Effect of diabetes mellitus type 2 on salivary glucose--a systematic review and meta-analysis of observational studies. *PloS one* **9**, e101706-e101706 (2014).
437. Kartheeki, B. *et al.* Salivary glucose levels in Type 2 diabetes mellitus: A tool for monitoring glycemic control. **1**, 7-14 (2017).
438. Goodson, J.M. *et al.* The salivary microbiome is altered in the presence of a high salivary glucose concentration. *PLoS ONE* **12**, e0170437 (2017).
439. Sabharwal, A. *et al.* The salivary microbiome of diabetic and non-diabetic adults with periodontal disease. *Journal of Periodontology* **90**, 26-34 (2019).
440. Lee, S., Im, A., Burm, E. & Ha, M. Association between periodontitis and blood lipid levels in a Korean population. *Journal of Periodontology* **89**, 28-35 (2018).
441. Tomofuji, T. *et al.* Effects of a High-cholesterol Diet on Cell Behavior in Rat Periodontitis. *Journal of Dental Research* **84**, 752-756 (2005).
442. Perlstein, M.I. & Bissada, N.F. Influence of obesity and hypertension on the severity of periodontitis in rats. *Oral Surgery, Oral Medicine, Oral Pathology* **43**, 707-719 (1977).
443. Suvan, J., D'Aiuto, F., Moles, D.R., Petrie, A. & Donos, N. Association between overweight/obesity and periodontitis in adults. A systematic review. *Obesity Reviews* **12**, e381-e404 (2011).
444. Chaffee, B.W. & Weston, S.J. Association between chronic periodontal disease and obesity: a systematic review and meta-analysis. *Journal of Periodontology* **81**, 1708-1724 (2010).
445. Morita, I. *et al.* Five-Year Incidence of Periodontal Disease Is Related to Body Mass Index. *Journal of Dental Research* **90**, 199-202 (2011).
446. Ekuni, D. *et al.* Relationship between increases in BMI and changes in periodontal status: a prospective cohort study. *Journal of Clinical Periodontology* **41**, 772-778 (2014).
447. Gorman, A., Kaye, E.K., Nunn, M. & Garcia, R.I. Changes in Body Weight and Adiposity Predict Periodontitis Progression in Men. *Journal of Dental Research* **91**, 921-926 (2012).
448. Gorman, A. *et al.* Overweight and obesity predict time to periodontal disease progression in men. *Journal of Clinical Periodontology* **39**, 107-114 (2012).
449. Shungin, D. *et al.* Using genetics to test the causal relationship of total adiposity and periodontitis: Mendelian randomization analyses in the Gene-Lifestyle Interactions and Dental Endpoints (GLIDE) Consortium. *International Journal of Epidemiology* **44**, 638-50 (2015).
450. Burgess, S. & Thompson, S.G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology* **181**, 251-260 (2015).
451. Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology* **48,** 713-727 (2018).
452. Zuza, E.P. *et al.* Influence of obesity on experimental periodontitis in rats: histopathological, histometric and immunohistochemical study. *Clinical Oral Investigations* **22**, 1197-1208 (2018).

453. Cavagni, J. *et al.* Obesity and Hyperlipidemia Modulate Alveolar Bone Loss in Wistar Rats. *Journal of Periodontology* **87**, e9-e17 (2016).
454. Justice, A.E. *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature Communications* **8**, 14977 (2017).
455. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circulation and Cardiovascular Genetics* **9**, 436-447 (2016).
456. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424-428 (2018).
457. Koellinger, P.D. & de Vlaming, R. Mendelian randomization: the challenge of unobserved environmental confounds. *International Journal of Epidemiology* **48**, 665-671 (2019).
458. O'Connor, L.J. & Price, A.L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics* **50**, 1728-1734 (2018).
459. Takeuchi, K. *et al.* Tooth Loss and Risk of Dementia in the Community: the Hisayama Study. *Journal of the American Geriatrics Society* **65**, e95-e100 (2017).
460. Jhun, M.A. *et al.* 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**, 1149-1158 (2017).
461. Burgess, S., Daniel, R.M., Butterworth, A.S., Thompson, S.G. & Consortium, E.P.-I. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International Journal of Epidemiology* **44**, 484-495 (2015).
462. Gaunt, T.R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biology* **17**(2016).
463. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications* **7**, 11122 (2016).
464. Lawlor, D. *et al.* Using Mendelian randomization to determine causal effects of maternal pregnancy (intrauterine) exposures on offspring outcomes: Sources of bias and methods for assessing them. epublication in *Wellcome Open Research* **2**, (2017).
465. Carvalho, J.C. & Schiffner, U. Dental Caries in European Adults and Senior Citizens 1996–2016: ORCA Saturday Afternoon Symposium in Greifswald, Germany – Part II. *Caries Research* **53**, 242-252 (2019).
466. Sovio, U. *et al.* Association between Common Variation at the FTO Locus and Changes in Body Mass Index from Infancy to Late Childhood: The Complex Nature of Genetic Association through Growth and Development. *PLoS Genetics* **7**, e1001307 (2011).
467. van Houwelingen, H.C., Arends, L.R. & Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* **21**, 589-624 (2002).
468. Vernazza, C.R., Rolland, S.L., Chadwick, B. & Pitts, N. Caries experience, the caries burden and associated factors in children in England, Wales and Northern Ireland 2013. *British Dental Journal* **221**, 315-320 (2016).
469. van der Tas, J.T. *et al.* Social inequalities and dental caries in six-year-old children from the Netherlands. *Journal of Dentistry* **62**, 18-24 (2017).
470. Schuller, A.A., van Dommelen, P. & Poorterman, J.H.G. Trends in oral health in young people in the Netherlands over the past 20 years: a study in a changing context. *Community Dentistry and Oral Epidemiology* **42**, 178-184 (2014).
471. Altman, D.G. & Royston, P. The cost of dichotomising continuous variables. *British Medical Journal* **332**, 1080-1080 (2006).
472. Nyvad, B., Crielaard, W., Mira, A., Takahashi, N. & Beighton, D. Dental Caries from a Molecular Microbiological Perspective. *Caries Research* **47**, 89-102 (2013).
473. Aas, J.A. *et al.* Bacteria of Dental Caries in Primary and Permanent Teeth in Children and Young Adults. *Journal of Clinical Microbiology* **46**, 1407-1417 (2008).
474. Vieira, A.R. *et al.* Detection of Streptococcus mutans Genomic DNA in Human DNA Samples Extracted from Saliva and Blood. *ISRN Dentistry* **2011**, 543561-543561 (2011).

475. De Menezes Oliveira, M.A.H. *et al.* Microstructure and mineral composition of dental enamel of permanent and deciduous teeth. **73**, 572-577 (2010).
476. Zeng, Z. *et al.* Genome-wide association study of primary dentition pit-and-fissure and smooth surface caries. *Caries Research.* **48**, 330-338 (2014).
477. Zeng, Z. *et al.* Genome-wide association studies of pit-and-fissure- and smooth-surface caries in permanent dentition. *Journal of Dental Research* **92**, 432-437 (2013).
478. Bayram, M. *et al.* Genetic influences on dental enamel that impact caries differ between the primary and permanent dentitions. *European Journal of Oral Sciences* **123**, 327-334 (2015).
479. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Annals of Allergy Asthma & Immunology* **93**, 381-389 (2004).
480. Olsen, J. *et al.* The Danish National Birth Cohort - its background, structure and aim. *Scandinavian Journal of Public Health* **29**, 300-307 (2001).
481. Kooijman, M.N. *et al.* The Generation R Study: design and cohort update 2017. *European Journal of Epidemiology* **31**, 1243-1264 (2016).
482. van der Tas, J.T. *et al.* Ethnic Disparities in Dental Caries among Six-Year-Old Children in the Netherlands. *Caries Research* **50**, 489-497 (2016).
483. Medina-Gomez, C. *et al.* Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *European Journal of Epidemiology* **30**, 317-330 (2015).
484. Levy, S.M., Warren, J.J., Broffitt, B., Hillis, S.L. & Kanellis, M.J. Fluoride, beverages and dental caries in the primary dentition. *Caries Research* **37**, 157-165 (2003).
485. Heinrich, J. *et al.* GINIplus and LISAplus - Design and selected results of two German birth cohorts about natural course of atopic diseases and their determinants. *Allergologie Select* **1**, 85-95 (2017).
486. Taal, H.R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nature Genetics* **44**, 532-+ (2012).
487. von Berg, A. *et al.* Impact of early feeding on childhood eczema: development after nutritional intervention compared with the natural course - the GINIplus study up to the age of 6 years. *Clinical and Experimental Allergy* **40**, 627-636 (2010).
488. Zutavern, A. *et al.* Timing of solid food introduction in relation to atopic dermatitis and atopic sensitization: Results from a prospective birth cohort study. *Pediatrics* **117**, 401-411 (2006).
489. Eloranta, A.M. *et al.* Dietary factors associated with overweight and body adiposity in Finnish children aged 6-8 years: the PANIC Study. *International Journal of Obesity* **36**, 950-955 (2012).
490. Raitakari, O.T. *et al.* Cohort Profile: The Cardiovascular Risk in Young Finns Study. *International Journal of Epidemiology* **37**, 1220-1226 (2008).
491. Straker, L. *et al.* Cohort Profile: The Western Australian Pregnancy Cohort (Raine) Study-Generation 2. *International Journal of Epidemiology* (2017).
492. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851-861 (2007).
493. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279 (2016).
494. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443-1448 (2016).
495. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics (Oxford, England)* **30**, 1266-1272 (2014).
496. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics (Oxford, England)* **23**, 2741-2746 (2007).

497. Higgins, J.P.T., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557-560 (2003).
498. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
499. Yang, J.A. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-U131 (2010).
500. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
501. Hemani, G. *et al.* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. Preprint at *bioRxiv* https://doi.org/10.1101/078972 (2016).
502. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology* **11**(2015).
503. Department of Medicine of the University of Chicago. PredictDB Data Repository. Published online at predictdb.org (2017).
504. Skol, A., Scott, L., Abecasis, G. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38**, 209-213 (2006).
505. Poirier, J.G. *et al.* Resampling to Address the Winner's Curse in Genetic Association Analysis of Time to Event. *Genetic Epidemiology* **39**, 518-528 (2015).
506. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics (Oxford, England)* **26**, 2336-2337 (2010).
507. Mak, G.W.Y. *et al.* CDK5RAP3 Is a Novel Repressor of p14(ARF) in Hepatocellular Carcinoma Cells. *PlosS ONE* **7**(2012)
508. Mak, G.W.Y. *et al.* Overexpression of a Novel Activator of PAK4, the CDK5 Kinase-Associated Protein CDK5RAP3, Promotes Hepatocellular Carcinoma Metastasis. *Cancer Research* **71**, 2949-2958 (2011).
509. Vigetti, D. *et al.* Property comparison of recombinant amphibian and mammalian allantoicases. *Febs Letters* **512**, 323-328 (2002).
510. Park, T.J. *et al.* Genome-wide association study identifies ALLC polymorphisms correlated with FEV1 change by corticosteroid. *Clinica Chimica Acta* **436**, 20-26 (2014).
511. Tikhmyanova, N., Little, J.L. & Golemis, E.A. CAS proteins in normal and pathological cell growth control. *Cellular and Molecular Life Sciences* **67**, 1025-1048 (2010).
512. Singh, M.K. *et al.* A novel Cas family member, HEPL, regulates FAK and cell spreading. *Molecular Biology of the Cell* **19**, 1627-1636 (2008).
513. Kumar, S., Tomooka, Y. & Noda, M. Identification of a set of genes with developmentally down-regulated expression in the mouse-brain. *Biochemical and Biophysical Research Communications* **185**, 1155-1161 (1992).
514. Latasa, M.J., Jimenez-Lara, A.M. & Cosgaya, J.M. Retinoic acid regulates Schwann cell migration via NEDD9 induction by transcriptional and post-translational mechanisms. *Biochimica Et Biophysica Acta-Molecular Cell Research* **1863**, 1510-1518 (2016).
515. Aquino, J.B. *et al.* The Retinoic Acid Inducible Cas-family Signaling Protein Nedd9 Regulates Neural Crest Cell Migration by Modulating Adhesion and Actin Dynamics. *Neuroscience* **162**, 1106-1119 (2009).
516. Nikonova, A.S., Gaponova, A.V., Kudinov, A.E. & Golemis, E.A. CAS Proteins in Health and Disease: An Update. *Iubmb Life* **66**, 387-395 (2014).
517. Riccomagno, M.M. *et al.* Cas Adaptor Proteins Organize the Retinal Ganglion Cell Layer Downstream of Integrin Signaling. *Neuron* **81**, 779-786 (2014).
518. Wang, S.K., Komatsu, Y. & Mishina, Y. Potential contribution of neural crest cells to dental enamel formation. *Biochemical and Biophysical Research Communications* **415**, 114-119 (2011).

519. Duverger, O. *et al.* Neural Crest Deletion of Dlx3 Leads to Major Dentin Defects through Down-regulation of Dspp. *Journal of Biological Chemistry* **287**, 12230-12240 (2012).
520. Kemp, J.P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nature Genetics* **49**, 1468-1475 (2017).
521. Shu, X.O. *et al.* Identification of New Genetic Risk Variants for Type 2 Diabetes. *PLoS Genetics* **6**, e1001127 (2010).
522. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications* **9**, 2941 (2018).
523. Docherty, A.R. *et al.* SNP-based heritability estimates of the personality dimensions and polygenic prediction of both neuroticism and major depression: findings from CONVERGE. epublication in *Translational Psychiatry* **6** (2016).
524. Shaffer, J.R. *et al.* Effects of enamel matrix genes on dental caries are moderated by fluoride exposures. *Human Genetics* **134**, 159-167 (2015).
525. Johansson, I., Witkowska, E., Kaveh, B., Holgerson, P.L. & Tanner, A.C.R. The Microbiome in Populations with a Low and High Prevalence of Caries. *Journal of Dental Research* **95**, 80-86 (2016).
526. von Hippel, P.T. The heterogeneity statistic I(2) can be biased in small meta-analyses. *BMC Medical Research Methodology* **15**, 35 (2015).
527. Devadiga, A. What's the deal with dental records for practicing dentists? Importance in general and forensic dentistry. *Journal of Forensic Dental Sciences* **6**, 9-15 (2014).
528. Dudding, T., Haworth, S., Sandy, J. & Timpson, N. Age 23 years + oral health questionnaire in Avon Longitudinal Study of Parents and Children. [version 2; peer review: 2 approved]. epublcation in *Wellcome Open Research* **3** (2018).
529. Perrin Ross, A. *et al.* Assessing relapse in multiple sclerosis questionnaire: results of a pilot study. *Multiple Sclerosis International* **2013**, 470476-470476 (2013).
530. Cole, T.J., Donaldson, M.D.C. & Ben-Shlomo, Y. SITAR-a useful instrument for growth curve analysis. *International Journal of Epidemiology* **39**, 1558-1566 (2010).
531. Warrington, N.M. *et al.* Modelling BMI trajectories in children for genetic association studies. *PloS ONE* **8**, e53897-e53897 (2013).
532. Johnson, L., van Jaarsveld, C.H.M., Llewellyn, C.H., Cole, T.J. & Wardle, J. Associations between infant feeding and the size, tempo and velocity of infant weight gain: SITAR analysis of the Gemini twin birth cohort. *International Journal of Obesity (2005)* **38**, 980-987 (2014).
533. Jones-Smith, J.C. *et al.* Early life growth trajectories and future risk for overweight. *Nutrition & Diabetes* **3**, e60-e60 (2013).
534. Mejia, G.C. *et al.* Socioeconomic status, oral health and dental disease in Australia, Canada, New Zealand and the United States. *BMC Oral Health* **18**, 176 (2018).
535. Singh, A., Peres, M.A. & Watt, R.G. The Relationship between Income and Oral Health: A Critical Review. *Journal of Dental Research* **98**, 853-860 (2019).
536. Bartley, M. *Health Inequality: An Introduction to Concepts, Theories and Methods (Second edition)*, (Wiley, 2016).
537. Hach, M. *et al.* Social inequality in tooth loss, the mediating role of smoking and alcohol consumption. *Community Dentistry and Oral Epidemiology* epublication ahead of print (2019).
538. Goldberg, D.S. Social Justice, Health Inequalities and Methodological Individualism in US Health Promotion. *Public Health Ethics* **5**, 104-115 (2012).
539. Watt, R.G. & Sheiham, A. Integrating the common risk factor approach into a social determinants framework. *Community Dentistry and Oral Epidemiology* **40**, 289-296 (2012).
540. Heaney, C.A., Israel, B.A. & House, J.S. Chronic job insecurity among automobile workers: Effects on job satisfaction and health. *Social Science & Medicine* **38**, 1431-1437 (1994).

541. Fearon, R.M.P. *et al.* Poverty, early care, and stress reactivity in adolescence: Findings from a prospective, longitudinal study in South Africa. *Development and Psychopathology* **29**, 449-464 (2017).
542. Vijayaraghavan, M. *et al.* Housing Instability and Incident Hypertension in the CARDIA Cohort. *Journal of Urban Health* **90**, 427-441 (2013).
543. Tikhonova, S. *et al.* Investigating the association between stress, saliva and dental caries: a scoping review. *BMC Oral Health* **18**, 41 (2018).
544. Blane, D., Bartley, M.E.L. & Smith, G.D. Disease aetiology and materialist explanations of socioeconomic mortality differentials. *European Journal of Public Health* **7**, 385-391 (1997).
545. Wilkinson, R.G. Income distribution and life expectancy. *British Medical Journal* **304**, 165 (1992).
546. Truesdale, B.C. & Jencks, C. The Health Effects of Income Inequality: Averages and Disparities. *Annual Review of Public Health* **37**, 413-430 (2016).
547. Min, J.L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics (Oxford, England)* **34**, 3983-3989 (2018).
548. Iglič, H., Doreian, P., Kronegger, L. & Ferligoj, A. With whom do researchers collaborate and why? *Scientometrics* **112**, 153-174 (2017).
549. Berg, J.J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/354951 (2018).
550. Sohail, M. *et al.* Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. Preprint at *bioRxiv* https://doi.org/10.1101/355057 (2018).
551. Public Health England. Delivering better oral health: an evidence-based toolkit for prevention Third edition. (www.gov.uk/phe, 2014).
552. Scottish Dental Clinical Effectiveness Programme. Prevention and Management of Dental Caries in Children - Guidance in Brief (Second Edition). (NHS Education for Scotland, 2018).
553. Rose, G. Sick individuals and sick populations. Reprinted in *International Journal of Epidemiology* **30**, 427-432 (2001).
554. Tickle, M. *et al.* A Randomized Controlled Trial of Caries Prevention in Dental Practice. *Journal of Dental Research* **96**, 741-746 (2017).
555. Pukallus, M.L. *et al.* A randomised, controlled clinical trial comparing chlorhexidine gel and low-dose fluoride toothpaste to prevent early childhood caries. *International Journal of Paediatric Dentistry* **23**, 216-224 (2013).
556. Jamieson, L. *et al.* Dental Disease Outcomes Following a 2-Year Oral Health Promotion Program for Australian Aboriginal Children and Their Families: A 2-Arm Parallel, Single-blind, Randomised Controlled Trial. *EClinicalMedicine* **1**, 43-50 (2018).
557. Anjum, M. *et al.* Does tooth loss have an emotional effect? A cross-sectional and comparative study on nondenture wearers and complete denture wearers. *Journal of the Indian Association of Public Health Dentistry* **15**, 247-251 (2017).
558. Jamieson, L.M., Paradies, Y.C., Gunthorpe, W., Cairney, S.J. & Sayers, S.M. Oral health and social and emotional well-being in a birth cohort of Aboriginal Australian young adults. *BMC Public Health* **11**, 656 (2011).
559. Naik, A.V. & Pai, R.C. Study of emotional effects of tooth loss in an aging north Indian community. *ISRN Dentistry* **2011**, 395498-395498 (2011).
560. Rousseau, N., Steele, J., May, C. & Exley, C. 'Your whole life is lived through your teeth': biographical disruption and experiences of tooth loss and replacement. **36**, 462-476 (2014).
561. Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *British Medical Journal* **352** (2016).

562. Moeller, J., Singhal, S., Al-Dajani, M., Gomaa, N. & Quiñonez, C. Assessing the relationship between dental appearance and the potential for discrimination in Ontario, Canada. *SSM - Population Health* **1**, 26-31 (2015).
563. Millard, L.A., Davies, N.M., Gaunt, T.R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *International Journal of Epidemiology* **47**, 29-35 (2017).
564. Millard, L.A.C., Davies, N.M., Tilling, K., Gaunt, T.R. & Davey Smith, G. Searching for the causal effects of body mass index in over 300 000 participants in UK Biobank, using Mendelian randomization. *PLOS Genetics* **15**, e1007951 (2019).
565. Ning, Z. *et al.* Beyond power: Multivariate discovery, replication, and interpretation of pleiotropic loci using summary association statistics. Preprint at *bioRxiv*, https://doi.org/10.1101/022269 (2019).
566. Orland, F.J. *et al.* Use of the Germfree Animal Technic in the Study of Experimental Dental Caries: I. Basic Observations on Rats Reared Free of All Microorganisms. *Journal of Dental Research* **33**, 147-174 (1954).
567. Blackmore, D.K. & Green, R.M. The use of germ-free and specific pathogen-free rats in short-term experiments for the assessment of potentially cariogenic organisms. *Archives of Oral Biology* **15**, 1149-1155 (1970).
568. Ritz, B.R. *et al.* Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *American Journal of Epidemiology* **186**, 753-761 (2017).
569. Wray, N.R. & Gottesman, I.I. Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Frontiers in Genetics* **3**, 118-118 (2012).
570. Han, B. *et al.* A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nature Genetics* **48**, 803-810 (2016).
571. Galli, S.J. Toward precision medicine and health: Opportunities and challenges in allergic diseases. *The Journal of Allergy and Clinical Immunology* **137**, 1289-1300 (2016).
572. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).
573. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173-1186 (2014).
574. Kuriyama, S. *et al.* The Tohoku Medical Megabank Project: Design and Mission. *Journal of Epidemiology* **26**, 493-511 (2016).
575. Smith, B.H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* **42**, 689-700 (2012).
576. Galloway, J. Putting the Teeth into the UK Biobank. *Primary Dental Care* **18**, 6-12 (2011).
577. Shaffer, J.R. *et al.* Clustering Tooth Surfaces into Biologically Informative Caries Outcomes. *Journal of Dental Research* **92**, 32-37 (2013).
578. Greenstein, G. & Hart, T.C. Clinical utility of a genetic susceptibility test for severe chronic periodontitis: A critical evaluation. *The Journal of the American Dental Association* **133**, 452-459 (2002).
579. Pihlstrom, B.L. & Barnett, M.L. Conference summary: Navigating the Sea of Genomic Data, October 28-29, 2015. *The Journal of the American Dental Association* **147**, 207-213 (2016).
580. Wu, X. *et al.* Association of interleukin-1 gene variations with moderate to severe chronic periodontitis in multiple ethnicities. *Journal of Periodontal Research* **50**, 52-61 (2015).
581. Khera, A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219-1224 (2018).

582. Armitage, J. *et al.* Efficacy and safety of statin therapy in older people: a meta-analysis of individual participant data from 28 randomised controlled trials. *The Lancet* **393**, 407-415 (2019).
583. Clayton, E.W. & McGuire, A.L. The legal risks of returning results of genomics research. *Genetics in Medicine : official journal of the American College of Medical Genetics* **14**, 473-477 (2012).
584. Stark, Z. *et al.* Integrating Genomics into Healthcare: A Global Responsibility. *The American Journal of Human Genetics* **104**, 13-20 (2019).
585. Genetics and Its Implications for Clinical Dental Practice and Education. *Journal of Dental Education* **72**, 86 (2008).

Appendix

Appendix 2

2.1: Coding of questionnaire responses used in ALSPAC

The intended exposure or outcome is given in bold font. The person completing the questionnaire and time when the question was asked is given in underlined, plain font. Questionnaire stems are given in italic font. Possible responses are in plain type. Variables coded from these responses are given in red.

DGA status

Parents when their child was 3, 4, 5 and 6 years old

- *Has he/she visited the dentist since [time since last questionnaire]?*
    - Yes for treatment
        - Did they have a general anaesthetic for this?
            - Yes = DGA
            - No = non-DGA
    - Yes, for inspection only = non-DGA
    - No, not at all = non-DGA

Children when they were 7 years old:

- *Have you ever been given something to make you go to sleep (general anaesthetic) before the dentist did something to your teeth?*
    - Yes = DGA
    - No = non-DGA

Dental health at age 17 years

Young person at age 17 years:

- *How many fillings do you have in your mouth?* = A
- *How many teeth have you had taken out because they were bad?* = B
- *Most recent time you have had teeth taken out because they were bad:*
    - Never
    - In the past year
    - More than 2 years ago – B set to 0 if this answered.

Filled extracted permanent teeth (FEPT) = A + B

Dental anxiety at age 17 years

Young person at 17 years:

- *How would you feel about going to the dentist, if you had to go tomorrow?*
    - Would look forward to it = A1

    - Would not care one way or another = A2
    - Would be a little uneasy about it = A3
    - Would be afraid it would be unpleasant and painful = A4
    - Would be very frightened of what dentist might do = A5
- *How would you feel when waiting in the dentist's surgery for your turn in the chair?*
    - Relaxed = B1
    - A little uneasy = B2
    - Tense = B3
    - Anxious = B4
    - So anxious you break out in sweat or feel physically sick = B5
- *How would you feel when waiting in the dentist's chair for the drill to be ready for your treatment?*
    - Relaxed = C1
    - A little uneasy = C2
    - Tense = C3
    - Anxious = C4
    - So anxious you break out in sweat or feel physically sick = C5
    - Never had dental treatment with a drill = C missing
- *How would you feel when waiting in the dentist's chair, while instruments are prepared to scrape teeth around gums?*
    - Relaxed = D1
    - A little uneasy = D2
    - Tense = D3
    - Anxious = D4
    - So anxious you break out in sweat or feel physically sick = D5
    - Never had teeth cleaned by dentist = D missing

Corah anxiety scale = A+B+C+D. If C or D missing, then the missing value was replaced with the mean of the 3 non-missing domains. If both C and D are missing then the anxiety scale was set to missing. Dental anxiety = 1 for participants with Corah anxiety scale ≥ 13, dental anxiety = 0 for participants with Corah anxiety scale < 13

Reason for attendance at age 17 years

Young person at 17 years:

- *What is the reason you usually attend the dentist?*
    - Regular routine checkups (every 6 or 12 months)
    - Occasional check up (less than every 2 years)
    - Only when has trouble with teeth
    - Never

} Combined into one group

Baseline dental health at age 7 years

Child reported questionnaire at 7 years (completed with help from adult):

How many fillings are there in your mouth? (Don't forget the front teeth) = C

How many teeth can you see or feel which have a hole in them? = D

Baseline caries = C + D

2.2: Acknowledgements for ALSPAC

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf).

2.3: Acknowledgements for GLIDE and KNHANES

The Swedish GLIDE study is funded by the Swedish Research Council (Dnr 2011-3372 and 2015-02597), the Västerbotten County Council and Umeå University, Sweden. Analysis in KNHANES was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2015R12A2A1A15054758 to Min-Jeong Shin).

Appendix 3

3.1: Heritability of DMFS/dentures partitioned by functional annotation

Functional annotation	Prop_SNPs	Prop_SNP h2	Fold enrichment	SE for fold enrichment	P value for enrichment
GERP.NSL2_0	1.745	3.450	1.977	0.080	$1.5E^{-28}$
Conserved_Primate_phastCons46wayL2_0	0.019	0.298	15.469	1.769	$3.0E^{-15}$
MAF_Adj_ASMCL2_0	-2.3E-14	-0.384	1.6E+13	-2.11283E+12	$1.6E^{-14}$
MAF_Adj_LLD_AFRL2_0	0.003	-0.363	-129.767	12.839	$2.3E^{-14}$
Conserved_Primate_phastCons46way.extend.500L2_0	0.176	0.541	3.077	0.269	$1.7E^{-13}$
Nucleotide_Diversity_10kbL2_0	4.609	3.927	0.852	0.020	$1.2E^{-12}$
Conserved_Mammal_phastCons46wayL2_0	0.021	0.340	15.853	2.033	$3.8E^{-12}$
Conserved_LindbladTohL2_0	0.026	0.338	13.158	1.745	$2.8E^{-11}$
MAF_Adj_Predicted_Allele_AgeL2_0	0.000	-0.428	-126233	16656	$5.3E^{-11}$
Backgrd_Selection_StatL2_0	0.178	0.216	1.217	0.036	$6.6E^{-11}$
Conserved_Vertebrate_phastCons46wayL2_0	0.029	0.278	9.434	1.398	$4.1E^{-09}$
CpG_Content_50kbL2_0	0.010	0.011	1.125	0.019	$6.0E^{-09}$
Repressed_Hoffman.extend.500L2_0	0.719	0.576	0.802	0.036	$3.5E^{-08}$
H3K27ac_HniszL2_0	0.389	0.533	1.369	0.069	$1.4E^{-07}$
Intron_UCSC.extend.500L2_0	0.397	0.486	1.224	0.045	$5.0E^{-07}$
SuperEnhancer_HniszL2_0	0.167	0.283	1.691	0.138	$1.3E^{-06}$
H3K9ac_Trynka.extend.500L2_0	0.230	0.408	1.777	0.171	$6.6E^{-06}$
SuperEnhancer_Hnisz.extend.500L2_0	0.170	0.278	1.632	0.138	$7.0E^{-06}$
Conserved_LindbladToh.extend.500L2_0	0.330	0.540	1.636	0.139	$1.2E^{-05}$
DHS_Trynka.extend.500L2_0	0.496	0.751	1.514	0.118	$1.6E^{-05}$
Enhancer_Hoffman.extend.500L2_0	0.090	0.218	2.428	0.326	$2.0E^{-05}$
TSS_Hoffman.extend.500L2_0	0.034	0.151	4.394	0.797	$2.6E^{-05}$
H3K27ac_PGC2.extend.500L2_0	0.335	0.501	1.495	0.116	$2.8E^{-05}$
Conserved_Mammal_phastCons46way.extend.500L2_0	0.339	0.583	1.718	0.164	$2.9E^{-05}$
MAFbin1L2_0	0.102	0.055	0.541	0.114	$4.3E^{-05}$
BivFlnkL2_0	0.014	0.112	8.251	1.773	$5.2E^{-05}$
H3K4me1_Trynka.extend.500L2_0	0.606	0.760	1.254	0.061	$5.4E^{-05}$
H3K9ac_peaks_TrynkaL2_0	0.038	0.216	5.611	1.144	$6.3E^{-05}$
MAFbin2L2_0	0.100	0.049	0.494	0.126	$7.3E^{-05}$
TFBS_ENCODE.extend.500L2_0	0.341	0.542	1.589	0.147	$8.3E^{-05}$
BivFlnk.extend.500L2_0	0.031	0.121	3.895	0.729	$8.6E^{-05}$
H3K27ac_PGC2L2_0	0.269	0.447	1.664	0.171	$1.64E^{-04}$
WeakEnhancer_Hoffman.extend.500L2_0	0.089	0.215	2.421	0.377	$2.25E^{-04}$

H3K4me3_TrynkaL2_0	0.133	0.295	2.219	0.330	$2.33E^{-04}$
FetalDHS_Trynka.extend.500L2_0	0.283	0.492	1.736	0.202	$2.83E^{-04}$
H3K9ac_TrynkaL2_0	0.125	0.289	2.307	0.373	$4.12E^{-04}$
H3K4me1_TrynkaL2_0	0.424	0.633	1.495	0.140	$4.59E^{-04}$
GERP.RSsup4L2_0	0.008	0.087	10.705	2.795	$5.61E^{-04}$
Conserved_Vertebrate_phastCons46way.extend.500L2_0	0.407	0.606	1.490	0.147	$1.14E^{-03}$
Enhancer_HoffmanL2_0	0.042	0.151	3.589	0.824	$1.83E^{-03}$
TSS_HoffmanL2_0	0.018	0.086	4.804	1.270	$2.73E^{-03}$
DHS_peaks_TrynkaL2_0	0.111	0.338	3.053	0.692	$3.21E^{-03}$
BLUEPRINT_DNA_methylation_MaxCPPL2_0	0.032	0.094	2.971	0.648	$3.43E^{-03}$
FetalDHS_TrynkaL2_0	0.084	0.278	3.308	0.784	$3.51E^{-03}$
GTEx_eQTL_MaxCPPL2_0	0.010	0.052	5.003	1.384	$3.98E^{-03}$
DHS_TrynkaL2_0	0.166	0.399	2.398	0.481	$4.30E^{-03}$
H3K27ac_Hnisz.extend.500L2_0	0.420	0.519	1.233	0.081	$4.39E^{-03}$
H3K4me3_Trynka.extend.500L2_0	0.255	0.387	1.517	0.182	$5.05E^{-03}$
DGF_ENCODE.extend.500L2_0	0.538	0.670	1.244	0.086	$5.13E^{-03}$
BLUEPRINT_H3K27acQTL_MaxCPPL2_0	0.017	0.056	3.407	0.856	$5.76E^{-03}$
Coding_UCSC.extend.500L2_0	0.064	0.131	2.059	0.383	$6.21E^{-03}$
Coding_UCSCL2_0	0.014	0.078	5.453	1.621	$6.78E^{-03}$
Promoter_UCSC.extend.500L2_0	0.057	0.124	2.173	0.439	$8.04E^{-03}$
UTR_5_UCSC.extend.500L2_0	0.027	0.070	2.600	0.638	$1.31E^{-02}$
H3K4me3_peaks_TrynkaL2_0	0.042	0.139	3.337	0.980	0.017
PromoterFlanking_HoffmanL2_0	0.008	-0.033	-3.919	2.063	0.017
MAFbin8L2_0	0.100	0.143	1.424	0.180	0.018
UTR_3_UCSCL2_0	0.011	0.040	3.592	1.121	0.020
MAFbin10L2_0	0.098	0.141	1.431	0.194	0.028
DGF_ENCODEL2_0	0.136	0.304	2.238	0.572	0.030
non_synonymousL2_0	0.003	0.023	8.424	3.443	0.033
Transcr_Hoffman.extend.500L2_0	0.762	0.686	0.900	0.047	0.035
BLUEPRINT_H3K4me1QTL_MaxCPPL2_0	0.013	0.035	2.591	0.846	0.062
H3K4me1_peaks_TrynkaL2_0	0.170	0.309	1.817	0.453	0.071
Recomb_Rate_10kbL2_0	1.552	1.314	0.847	0.090	0.077
UTR_3_UCSC.extend.500L2_0	0.026	0.056	2.126	0.638	0.077
PromoterFlanking_Hoffman.extend.500L2_0	0.033	0.072	2.182	0.698	0.090
MAFbin9L2_0	0.101	0.133	1.316	0.187	0.093
TFBS_ENCODEL2_0	0.131	0.240	1.828	0.518	0.110
Intron_UCSCL2_0	0.387	0.422	1.089	0.062	0.143
MAFbin3L2_0	0.100	0.081	0.813	0.144	0.193
Enhancer_Andersson.extend.500L2_0	0.019	0.037	1.949	0.835	0.256
UTR_5_UCSCL2_0	0.005	0.016	2.863	1.750	0.290
WeakEnhancer_HoffmanL2_0	0.021	0.054	2.578	1.552	0.310

Transcr_HoffmanL2_0	0.346	0.303	0.876	0.123	0.316
CTCF_HoffmanL2_0	0.024	-0.006	-0.265	1.368	0.356
MAFbin7L2_0	0.100	0.115	1.156	0.171	0.362
Repressed_HoffmanL2_0	0.461	0.404	0.876	0.140	0.377
synonymousL2_0	0.003	0.016	4.977	5.158	0.440
Promoter_UCSCL2_0	0.046	0.070	1.506	0.667	0.447
Enhancer_AnderssonL2_0	0.004	-0.005	-1.124	3.504	0.543
MAFbin5L2_0	0.098	0.092	0.931	0.155	0.658
CTCF_Hoffman.extend.500L2_0	0.071	0.057	0.808	0.507	0.704
MAFbin6L2_0	0.100	0.095	0.949	0.163	0.753
MAFbin4L2_0	0.101	0.096	0.957	0.153	0.780
Baseline	1	1	1	0	NA

Prop_SNPs; the proportion of SNPs included in the model which have the functional annotation given in the annotation column. Prop_SNP_h2; the proportion of SNP based heritability accounted for by SNPs which have the functional annotation. Fold enrichment; enrichment in heritability conferred by SNPs with the functional annotation over the baseline model. P for enrichment; test of the null hypothesis that the fold enrichment value = 1.

3.2: Heritability of DMFS/dentures in genomic regions with tissue-specific annotation

Tissue name	Stratified LDSR coefficient	SE	P
A03.556.Gastrointestinal.Tract	3.39E-09	1.55E-09	0.014
Minor_Salivary_Gland	3.08E-09	1.52E-09	0.022
A03.556.249.249.356.Colon	2.98E-09	1.59E-09	0.030
A03.556.249.249.209.Cecum	2.76E-09	1.58E-09	0.040
A14.549.167.646.Periodontium	2.75E-09	1.68E-09	0.051
A15.145.229.637.555.Leukocytes..Mononuclear	2.57E-09	1.80E-09	0.076
A08.186.211.132.810.428.200.Cerebellum	2.54E-09	1.13E-09	0.012
A08.186.211.865.428.Metencephalon	2.51E-09	1.13E-09	0.013
A03.556.500.760.464.Parotid.Gland	2.41E-09	1.57E-09	0.062
A03.556.124.Intestines	2.40E-09	1.53E-09	0.059
Esophagus_Mucosa	2.37E-09	1.52E-09	0.059
A08.186.211.730.317.357.Hypothalamus	2.34E-09	2.09E-09	0.131
A11.118.637.555.567.569.200.700.T.Lymphocytes..Regulatory	2.32E-09	1.34E-09	0.042
A03.556.875.500.Esophagus	2.31E-09	1.58E-09	0.073
A03.556.500.760.Salivary.Glands	2.28E-09	1.57E-09	0.074
Heart_Left_Ventricle	2.24E-09	1.06E-09	0.018
A05.360.319.679.256.Cervix.Uteri	2.17E-09	1.29E-09	0.046
Adrenal_Gland	2.16E-09	1.17E-09	0.032
A03.556.124.526.767.Rectum	2.07E-09	1.51E-09	0.085
Testis	2.06E-09	1.62E-09	0.103
A10.549.Lymphoid.Tissue	2.05E-09	1.35E-09	0.065
Kidney_Cortex	2.02E-09	1.32E-09	0.062
A11.118.637.555.567.562.B.Lymphocytes	2.02E-09	1.58E-09	0.100
Muscle_Skeletal	1.99E-09	1.27E-09	0.059
A11.627.340.360.Granulocyte.Precursor.Cells	1.92E-09	1.49E-09	0.099
A15.145.229.637.555.567.569.200.CD4.Positive.T.Lymphocytes	1.83E-09	1.42E-09	0.100
A03.556.249.249.356.668.Colon..Sigmoid	1.81E-09	1.45E-09	0.106
A14.549.Mouth	1.78E-09	1.72E-09	0.151
A08.186.211.132.Brain.Stem	1.78E-09	1.28E-09	0.083
A08.186.211.730.317.357.352.435.Hypothalamo.Hypophyseal.System	1.72E-09	1.99E-09	0.194
A08.186.211.464.405.Hippocampus	1.68E-09	1.25E-09	0.090
A05.360.319.887.Vulva	1.64E-09	1.37E-09	0.114
A15.382.520.604.800.Palatine.Tonsil	1.64E-09	1.43E-09	0.126
A11.118.637.555.567.569.T.Lymphocytes	1.64E-09	1.58E-09	0.150
A09.371.Eye	1.64E-09	1.55E-09	0.145
A15.382.490.555.567.Lymphocytes	1.62E-09	1.68E-09	0.167
A10.615.550.Mucous.Membrane	1.53E-09	1.40E-09	0.138
A11.118.637.Leukocytes	1.51E-09	1.88E-09	0.211

A10.615.Membranes	$1.43E^{-09}$	$1.32E^{-09}$	0.140
Bladder	$1.42E^{-09}$	$1.14E^{-09}$	0.106
A03.556.249.124.Ileum	$1.41E^{-09}$	$1.65E^{-09}$	0.195
Pituitary	$1.39E^{-09}$	$1.60E^{-09}$	0.192
A15.382.Immune.System	$1.38E^{-09}$	$1.43E^{-09}$	0.168
A10.272.497.Epidermis	$1.36E^{-09}$	$1.59E^{-09}$	0.195
A03.556.875.875.Stomach	$1.30E^{-09}$	$1.57E^{-09}$	0.204
A08.186.211.730.885.287.249.Basal.Ganglia	$1.30E^{-09}$	$1.19E^{-09}$	0.139
A11.872.700.500.Induced.Pluripotent.Stem.Cells	$1.29E^{-09}$	$1.34E^{-09}$	0.168
A03.556.124.369.Intestinal.Mucosa	$1.29E^{-09}$	$1.45E^{-09}$	0.187
A14.549.167.Dentition	$1.27E^{-09}$	$1.76E^{-09}$	0.236
A08.186.211.730.885.287.500.571.735.Visual.Cortex	$1.22E^{-09}$	$1.20E^{-09}$	0.155
A08.186.211.730.885.287.249.487.Corpus.Striatum	$1.22E^{-09}$	$1.18E^{-09}$	0.152
Cells_EBV-transformed_lymphocytes	$1.20E^{-09}$	$1.35E^{-09}$	0.187
A08.186.211.464.Limbic.System	$1.17E^{-09}$	$1.23E^{-09}$	0.170
Esophagus_Muscularis	$1.17E^{-09}$	$1.19E^{-09}$	0.163
A08.186.211.730.885.287.500.270.Frontal.Lobe	$1.14E^{-09}$	$1.32E^{-09}$	0.193
Cervix_Ectocervix	$1.11E^{-09}$	$1.51E^{-09}$	0.231
Brain_Cerebellar_Hemisphere	$1.10E^{-09}$	$1.16E^{-09}$	0.171
A03.556.875.Upper.Gastrointestinal.Tract	$1.09E^{-09}$	$1.49E^{-09}$	0.232
A05.360.319.679.490.Endometrium	$1.09E^{-09}$	$1.19E^{-09}$	0.180
A08.186.211.Brain	$9.95E^{-10}$	$1.23E^{-09}$	0.209
A05.360.319.114.630.Ovary	$9.39E^{-10}$	$1.28E^{-09}$	0.231
A06.407.071.Adrenal.Glands	$9.24E^{-10}$	$1.48E^{-09}$	0.266
A03.556.124.684.Intestine..Small	$8.57E^{-10}$	$1.46E^{-09}$	0.279
A05.810.453.Kidney	$8.45E^{-10}$	$1.20E^{-09}$	0.242
A02.835.583.443.800.Synovial.Membrane	$8.45E^{-10}$	$1.45E^{-09}$	0.280
Vagina	$8.33E^{-10}$	$1.38E^{-09}$	0.274
A08.186.211.730.885.287.500.Cerebral.Cortex	$7.70E^{-10}$	$1.24E^{-09}$	0.268
A05.810.890.Urinary.Bladder	$7.55E^{-10}$	$1.63E^{-09}$	0.322
A11.436.348.Hepatocytes	$7.48E^{-10}$	$1.52E^{-09}$	0.312
A15.145.229.637.555.567.562.725.Plasma.Cells	$6.70E^{-10}$	$1.41E^{-09}$	0.318
A05.360.490.Germ.Cells	$6.53E^{-10}$	$1.30E^{-09}$	0.308
Heart_Atrial_Appendage	$6.42E^{-10}$	$1.15E^{-09}$	0.288
A03.734.Pancreas	$6.39E^{-10}$	$1.36E^{-09}$	0.319
A11.872.378.590.635.Granulocyte.Macrophage.Progenitor.Cells	$6.22E^{-10}$	$1.53E^{-09}$	0.342
A17.815.Skin	$6.19E^{-10}$	$1.37E^{-09}$	0.326
Brain_Cerebellum	$6.15E^{-10}$	$1.21E^{-09}$	0.306
A08.186.211.730.317.Diencephalon	$5.99E^{-10}$	$1.85E^{-09}$	0.373
A10.272.Epithelium	$5.80E^{-10}$	$1.40E^{-09}$	0.339
A11.329.171.Chondrocytes	$5.68E^{-10}$	$1.50E^{-09}$	0.353
A15.382.520.604.700.Spleen	$5.55E^{-10}$	$1.35E^{-09}$	0.340

A06.407.Endocrine.Glands	$5.13E^{-10}$	$1.23E^{-09}$	0.339
Pancreas	$4.84E^{-10}$	$1.27E^{-09}$	0.352
A10.615.550.599.Mouth.Mucosa	$4.65E^{-10}$	$1.33E^{-09}$	0.363
A05.360.319.Genitalia..Female	$4.60E^{-10}$	$1.20E^{-09}$	0.351
A15.145.229.Blood.Cells	$4.15E^{-10}$	$1.57E^{-09}$	0.396
A08.186.211.464.710.225.Entorhinal.Cortex	$4.10E^{-10}$	$1.22E^{-09}$	0.368
A11.329.372.600.Macrophages..Alveolar	$3.93E^{-10}$	$1.41E^{-09}$	0.390
A14.549.885.Tongue	$3.85E^{-10}$	$1.79E^{-09}$	0.415
A11.627.635.Myeloid.Progenitor.Cells	$3.81E^{-10}$	$1.50E^{-09}$	0.400
A04.531.520.Nasal.Mucosa	$3.72E^{-10}$	$1.40E^{-09}$	0.395
A05.360.319.679.Uterus	$3.43E^{-10}$	$1.27E^{-09}$	0.393
A14.724.Pharynx	$3.36E^{-10}$	$1.61E^{-09}$	0.417
Stomach	$3.12E^{-10}$	$1.34E^{-09}$	0.408
A02.633.567.850.Quadriceps.Muscle	$2.63E^{-10}$	$1.38E^{-09}$	0.424
Skin_Not_Sun_Exposed_(Suprapubic)	$2.47E^{-10}$	$1.37E^{-09}$	0.429
A11.872.040.Adult.Stem.Cells	$2.34E^{-10}$	$1.25E^{-09}$	0.426
Cervix_Endocervix	$2.28E^{-10}$	$1.28E^{-09}$	0.429
Esophagus_Gastroesophageal_Junction	$1.91E^{-10}$	$1.16E^{-09}$	0.435
A05.360.319.114.373.Fallopian.Tubes	$1.63E^{-10}$	$1.39E^{-09}$	0.453
Small_Intestine_Terminal_Ileum	$1.60E^{-10}$	$1.65E^{-09}$	0.461
A11.118.637.555.567.562.440.Precursor.Cells..B.Lymphoid	$1.36E^{-10}$	$1.32E^{-09}$	0.459
A15.382.490.555.567.537.Killer.Cells..Natural	$9.04E^{-11}$	$1.39E^{-09}$	0.474
A10.549.400.Lymph.Nodes	$7.32E^{-11}$	$1.29E^{-09}$	0.477
A15.382.490.555.567.622.Lymphocytes..Null	$5.12E^{-11}$	$1.59E^{-09}$	0.487
A15.145.229.188.Blood.Platelets	$4.73E^{-11}$	$1.69E^{-09}$	0.489
A11.872.190.Embryonic.Stem.Cells	$3.74E^{-11}$	$1.25E^{-09}$	0.488
A05.360.319.679.690.Myometrium	$2.36E^{-11}$	$1.29E^{-09}$	0.493
A15.145.Blood	$1.94E^{-11}$	$1.57E^{-09}$	0.495
Brain_Spinal_cord_(cervical_c-1)	$1.66E^{-11}$	$1.08E^{-09}$	0.494
A06.407.071.140.Adrenal.Cortex	$3.96E^{-12}$	$1.36E^{-09}$	0.499
Colon_Sigmoid	$-5.30E^{-12}$	$1.29E^{-09}$	0.502
A06.407.312.Gonads	$-7.26E^{-12}$	$1.27E^{-09}$	0.502
Lung	$-1.39E^{-11}$	$1.31E^{-09}$	0.504
Skin_Sun_Exposed_(Lower_leg)	$-1.31E^{-10}$	$1.37E^{-09}$	0.538
A05.360.444.492.362.Foreskin	$-1.38E^{-10}$	$1.28E^{-09}$	0.543
A11.329.830.Stromal.Cells	$-1.55E^{-10}$	$1.38E^{-09}$	0.545
A15.382.812.522.Macrophages	$-1.58E^{-10}$	$1.52E^{-09}$	0.542
A11.497.497.600.Oocytes	$-1.68E^{-10}$	$1.24E^{-09}$	0.554
A11.436.294.064.Glucagon.Secreting.Cells	$-1.74E^{-10}$	$1.41E^{-09}$	0.549
A10.690.Muscles	$-2.33E^{-10}$	$1.31E^{-09}$	0.571
Liver	$-2.98E^{-10}$	$1.33E^{-09}$	0.589
A15.382.812.260.Dendritic.Cells	$-3.45E^{-10}$	$1.34E^{-09}$	0.601
Brain_Nucleus_accumbens_(basal_ganglia)	$-3.49E^{-10}$	$1.06E^{-09}$	0.630
Ovary	$-3.52E^{-10}$	$1.30E^{-09}$	0.607

A15.145.300.Fetal.Blood	$-3.92E^{-10}$	$1.64E^{-09}$	0.594
Brain_Anterior_cingulate_cortex_(BA24)	$-3.97E^{-10}$	$1.07E^{-09}$	0.644
A07.541.560.Heart.Ventricles	$-4.19E^{-10}$	$1.14E^{-09}$	0.644
A07.541.358.100.Atrial.Appendage	$-5.26E^{-10}$	$1.21E^{-09}$	0.668
A02.835.583.443.800.800.Synovial.Fluid	$-5.37E^{-10}$	$1.51E^{-09}$	0.639
Fallopian_Tube	$-5.42E^{-10}$	$1.30E^{-09}$	0.662
Brain_Frontal_Cortex_(BA9)	$-5.54E^{-10}$	$1.14E^{-09}$	0.687
A07.541.358.Heart.Atria	$-5.61E^{-10}$	$1.20E^{-09}$	0.680
A10.615.284.473.Chorion	$-6.18E^{-10}$	$1.33E^{-09}$	0.679
A04.411.Lung	$-6.21E^{-10}$	$1.37E^{-09}$	0.674
A09.371.729.Retina	$-6.28E^{-10}$	$1.18E^{-09}$	0.703
Breast_Mammary_Tissue	$-6.64E^{-10}$	$1.13E^{-09}$	0.722
Brain_Putamen_(basal_ganglia)	$-7.00E^{-10}$	$1.12E^{-09}$	0.734
Thyroid	$-7.14E^{-10}$	$1.17E^{-09}$	0.730
A11.872.Stem.Cells	$-7.48E^{-10}$	$1.43E^{-09}$	0.699
A11.872.378.Hematopoietic.Stem.Cells	$-7.51E^{-10}$	$1.55E^{-09}$	0.686
Brain_Caudate_(basal_ganglia)	$-7.66E^{-10}$	$1.09E^{-09}$	0.760
A11.436.397.Keratinocytes	$-8.14E^{-10}$	$1.62E^{-09}$	0.692
A05.360.444.Genitalia..Male	$-8.15E^{-10}$	$1.20E^{-09}$	0.752
Colon_Transverse	$-8.56E^{-10}$	$1.48E^{-09}$	0.718
A11.329.629.Osteoblasts	$-8.62E^{-10}$	$1.38E^{-09}$	0.734
Brain_Hypothalamus	$-8.62E^{-10}$	$1.10E^{-09}$	0.783
A15.382.812.Mononuclear.Phagocyte.System	$-9.08E^{-10}$	$1.51E^{-09}$	0.726
A10.336.707.Prostate	$-9.34E^{-10}$	$1.30E^{-09}$	0.764
Prostate	$-9.78E^{-10}$	$1.41E^{-09}$	0.756
A10.165.450.300.Cicatrix	$-1.00E^{-09}$	$1.36E^{-09}$	0.769
Brain_Amygdala	$-1.02E^{-09}$	$1.10E^{-09}$	0.825
A11.329.Connective.Tissue.Cells	$-1.03E^{-09}$	$1.55E^{-09}$	0.747
A07.541.Heart	$-1.04E^{-09}$	$1.10E^{-09}$	0.828
A06.407.900.Thyroid.Gland	$-1.05E^{-09}$	$1.16E^{-09}$	0.816
Brain_Substantia_nigra	$-1.05E^{-09}$	$1.11E^{-09}$	0.828
Artery_Aorta	$-1.07E^{-09}$	$1.23E^{-09}$	0.808
Brain_Hippocampus	$-1.08E^{-09}$	$1.19E^{-09}$	0.818
A03.620.Liver	$-1.08E^{-09}$	$1.36E^{-09}$	0.788
A05.810.453.324.Kidney.Cortex	$-1.12E^{-09}$	$1.51E^{-09}$	0.771
A05.360.Genitalia	$-1.13E^{-09}$	$1.14E^{-09}$	0.840
A03.734.414.Islets.of.Langerhans	$-1.16E^{-09}$	$1.39E^{-09}$	0.798
A11.872.580.Mesenchymal.Stem.Cells	$-1.18E^{-09}$	$1.37E^{-09}$	0.807
A10.165.450.300.425.Keloid	$-1.23E^{-09}$	$1.31E^{-09}$	0.826
A11.436.329.Granulosa.Cells	$-1.27E^{-09}$	$1.13E^{-09}$	0.870
Whole_Blood	$-1.29E^{-09}$	$1.51E^{-09}$	0.805
A11.872.378.590.817.Megakaryocyte.Erythroid.Progenitor.Cells	$-1.37E^{-09}$	$1.44E^{-09}$	0.829
A08.186.211.730.885.287.500.670.Parietal.Lobe	$-1.39E^{-09}$	$1.24E^{-09}$	0.869

A11.329.228.Fibroblasts	-1.41E^{-09}	1.21E^{-09}	0.878
A10.615.789.Serous.Membrane	-1.43E^{-09}	1.27E^{-09}	0.871
A07.231.114.Arteries	-1.44E^{-09}	1.33E^{-09}	0.860
A15.382.490.315.583.Neutrophils	-1.44E^{-09}	1.35E^{-09}	0.858
A15.378.316.580.Monocytes	-1.47E^{-09}	1.60E^{-09}	0.820
A14.724.557.Nasopharynx	-1.49E^{-09}	1.51E^{-09}	0.839
Artery_Tibial	-1.50E^{-09}	1.15E^{-09}	0.905
Uterus	-1.58E^{-09}	1.22E^{-09}	0.903
Brain_Cortex	-1.63E^{-09}	1.10E^{-09}	0.931
A02.165.Cartilage	-1.64E^{-09}	1.33E^{-09}	0.891
Adipose_Visceral_(Omentum)	-1.65E^{-09}	1.07E^{-09}	0.938
Spleen	-1.66E^{-09}	1.37E^{-09}	0.886
A15.145.846.Serum	-1.66E^{-09}	1.43E^{-09}	0.878
A06.407.312.782.Testis	-1.71E^{-09}	1.11E^{-09}	0.939
Adipose_Subcutaneous	-1.82E^{-09}	1.10E^{-09}	0.951
A15.382.680.Phagocytes	-1.87E^{-09}	1.51E^{-09}	0.892
A11.627.624.249.Monocyte.Macrophage.Precursor.Cells	-1.90E^{-09}	1.41E^{-09}	0.911
A10.165.114.830.500.750.Subcutaneous.Fat..Abdominal	-1.93E^{-09}	1.30E^{-09}	0.931
A02.835.232.834.151.Cervical.Vertebrae	-2.01E^{-09}	1.25E^{-09}	0.946
A11.872.190.260.Embryoid.Bodies	-2.03E^{-09}	1.36E^{-09}	0.932
Nerve_Tibial	-2.05E^{-09}	1.06E^{-09}	0.973
A11.872.653.Neural.Stem.Cells	-2.14E^{-09}	1.41E^{-09}	0.936
A15.378.316.Bone.Marrow.Cells	-2.24E^{-09}	1.56E^{-09}	0.925
A11.443.Erythroid.Cells	-2.31E^{-09}	1.36E^{-09}	0.955
A10.690.467.Muscle..Smooth	-2.34E^{-09}	1.44E^{-09}	0.948
A11.382.Endocrine.Cells	-2.34E^{-09}	1.26E^{-09}	0.969
A11.329.114.Adipocytes	-2.43E^{-09}	1.35E^{-09}	0.965
A07.541.510.110.Aortic.Valve	-2.51E^{-09}	1.34E^{-09}	0.969
A08.186.211.653.Mesencephalon	-2.56E^{-09}	1.35E^{-09}	0.971
Cells_Transformed_fibroblasts	-2.58E^{-09}	1.18E^{-09}	0.985
A11.620.520.Myocytes..Smooth.Muscle	-2.62E^{-09}	1.49E^{-09}	0.961
A10.165.114.830.750.Subcutaneous.Fat	-2.83E^{-09}	1.26E^{-09}	0.988
Artery_Coronary	-2.83E^{-09}	1.26E^{-09}	0.988
A07.231.908.Veins	-3.27E^{-09}	1.25E^{-09}	0.995
A11.436.Epithelial.Cells	-3.29E^{-09}	1.59E^{-09}	0.981
A07.231.Blood.Vessels	-3.40E^{-09}	1.32E^{-09}	0.995
A07.231.908.670.874.Umbilical.Veins	-3.46E^{-09}	1.31E^{-09}	0.996
A11.436.275.Endothelial.Cells	-3.83E^{-09}	1.34E^{-09}	0.998

Each row represents a single LDSR model, fitted only using SNPs which are annotated as specifically expressed in the tissue type given in the first column. The coefficient, standard error and P value represent a single-tissue test for heritability.

3.3: Heritability of Periodontitis/loose teeth partitioned by functional annotation

Functional annotation	Proportion of SNPs	Proportion of SNP heritability	Fold enrichment	SE for fold enrichment	P value for enrichment
CpG_Content_50kbL2_0	0.010	0.012	1.20	0.055	4.78E-04
SuperEnhancer_HniszL2_0	0.167	0.351	2.10	0.395	5.39E-03
SuperEnhancer_Hnisz.extend.500L2_0	0.170	0.336	1.97	0.372	5.76E-03
H3K4me1_peaks_TrynkaL2_0	0.170	-0.518	-3.05	1.723	8.39E-03
MAFbin6L2_0	0.100	-0.032	-0.32	0.573	0.015
Conserved_Primate_phastCons46way.extend.500L2_0	0.176	0.626	3.56	1.102	0.016
Backgrd_Selection_StatL2_0	0.178	0.215	1.21	0.119	0.021
H3K27ac_PGC2.extend.500L2_0	0.335	0.710	2.12	0.466	0.021
GERP.NSL2_0	1.745	2.887	1.65	0.295	0.026
Enhancer_AnderssonL2_0	0.004	0.127	29.45	14.693	0.039
Coding_UCSCL2_0	0.014	0.167	11.73	5.157	0.040
MAFbin7L2_0	0.100	0.253	2.54	0.839	0.048
MAFbin1L2_0	0.102	0.007	0.07	0.559	0.050
Promoter_UCSCL2_0	0.046	-0.149	-3.21	2.336	0.053
H3K27ac_Hnisz.extend.500L2_0	0.420	0.637	1.51	0.284	0.063
BLUEPRINT_H3K27acQTL_MaxCPPL2_0	0.017	0.110	6.67	3.138	0.066
BLUEPRINT_DNA_methylation_MaxCPPL2_0	0.032	0.152	4.79	2.091	0.066
Intron_UCSC.extend.500L2_0	0.397	0.502	1.26	0.161	0.085
CTCF_Hoffman.extend.500L2_0	0.071	0.293	4.14	1.854	0.086
MAFbin2L2_0	0.100	0.014	0.14	0.532	0.092
MAF_Adj_ASMCL2_0	0.000	-0.236	1.01E+13	-6.30E+12	0.093
GERP.RSsup4L2_0	0.008	0.132	16.14	10.068	0.116
H3K4me1_Trynka.extend.500L2_0	0.606	0.789	1.30	0.197	0.134
MAFbin9L2_0	0.101	0.189	1.87	0.633	0.136
TFBS_ENCODE.extend.500L2_0	0.341	0.658	1.93	0.662	0.160
FetalDHS_Trynka.extend.500L2_0	0.283	0.583	2.06	0.764	0.169
MAFbin8L2_0	0.100	0.179	1.78	0.678	0.227
GTEx_eQTL_MaxCPPL2_0	0.010	-0.034	-3.29	3.777	0.232
FetalDHS_TrynkaL2_0	0.084	-0.173	-2.06	2.779	0.243
UTR_3_UCSCL2_0	0.011	0.073	6.53	4.814	0.244
MAFbin3L2_0	0.100	0.047	0.47	0.478	0.256
synonymousL2_0	0.003	0.061	19.60	16.320	0.258
MAFbin10L2_0	0.098	0.158	1.61	0.545	0.262

Nucleotide_Diversity_10kbL2_0	4.609	4.110	0.89	0.105	0.266
H3K27ac_PGC2L2_0	0.269	0.081	0.30	0.670	0.267
Transcr_HoffmanL2_0	0.346	0.592	1.71	0.652	0.268
H3K4me3_TrynkaL2_0	0.133	0.288	2.17	1.219	0.341
Enhancer_Andersson.extend.500L2_0	0.019	0.081	4.24	3.514	0.358
Conserved_Mammal_phastCons46way.extend.500L2_0	0.339	0.488	1.44	0.502	0.375
Intron_UCSCL2_0	0.387	0.316	0.82	0.212	0.397
Conserved_LindbladToh.extend.500L2_0	0.330	0.444	1.34	0.410	0.408
H3K4me3_peaks_TrynkaL2_0	0.042	-0.087	-2.08	3.880	0.413
Conserved_Mammal_phastCons46wayL2_0	0.021	0.124	5.79	6.125	0.437
H3K27ac_HniszL2_0	0.389	0.464	1.19	0.247	0.453
TSS_HoffmanL2_0	0.018	-0.038	-2.15	4.758	0.505
DGF_ENCODEL2_0	0.136	0.341	2.51	2.330	0.513
Conserved_Primate_phastCons46wayL2_0	0.019	0.099	5.14	6.479	0.522
MAF_Adj_Predicted_Allele_AgeL2_0	0.000	-0.158	-46534.90	72459.869	0.543
MAF_Adj_LLD_AFRL2_0	0.003	-0.090	-32.11	52.134	0.558
WeakEnhancer_HoffmanL2_0	0.021	-0.040	-1.89	5.348	0.580
H3K4me3_Trynka.extend.500L2_0	0.255	0.351	1.38	0.679	0.580
TSS_Hoffman.extend.500L2_0	0.034	0.081	2.36	2.459	0.582
UTR_3_UCSC.extend.500L2_0	0.026	0.062	2.35	2.460	0.583
UTR_5_UCSCL2_0	0.005	0.029	5.33	8.065	0.594
Enhancer_Hoffman.extend.500L2_0	0.090	0.149	1.66	1.219	0.595
CTCF_HoffmanL2_0	0.024	0.087	3.64	5.071	0.601
DGF_ENCODE.extend.500L2_0	0.538	0.635	1.18	0.340	0.604
MAFbin5L2_0	0.098	0.070	0.71	0.566	0.611
UTR_5_UCSC.extend.500L2_0	0.027	0.064	2.40	2.845	0.626
BivFlnkL2_0	0.014	0.055	4.07	6.353	0.631
non_synonymousL2_0	0.003	0.020	7.54	13.946	0.640
PromoterFlanking_Hoffman.extend.500L2_0	0.033	-0.009	-0.28	2.780	0.648
TFBS_ENCODEL2_0	0.131	0.030	0.23	1.691	0.649
Promoter_UCSC.extend.500L2_0	0.057	0.089	1.56	1.357	0.681
Conserved_Vertebrate_phastCons46way.extend.500L2_0	0.407	0.337	0.83	0.449	0.702
DHS_TrynkaL2_0	0.166	0.050	0.30	1.959	0.715
Transcr_Hoffman.extend.500L2_0	0.762	0.807	1.06	0.165	0.718
BivFlnk.extend.500L2_0	0.031	0.056	1.81	2.394	0.738

H3K4me1_TrynkaL2_0	0.424	0.353	0.83	0.544	0.757
WeakEnhancer_Hoffman.extend.500L2_0	0.089	0.124	1.39	1.432	0.783
Conserved_LindbladTohL2_0	0.026	-0.011	-0.42	5.367	0.790
MAFbin4L2_0	0.101	0.114	1.13	0.516	0.801
PromoterFlanking_HoffmanL2_0	0.008	0.025	3.06	8.222	0.802
H3K9ac_Trynka.extend.500L2_0	0.230	0.269	1.17	0.682	0.802
H3K9ac_peaks_TrynkaL2_0	0.038	0.001	0.03	4.526	0.829
Recomb_Rate_10kbL2_0	1.552	1.443	0.93	0.384	0.851
Enhancer_HoffmanL2_0	0.042	0.062	1.48	3.046	0.875
DHS_Trynka.extend.500L2_0	0.496	0.460	0.93	0.468	0.876
Conserved_Vertebrate_phastCons46wayL2_0	0.029	0.012	0.41	4.543	0.898
H3K9ac_TrynkaL2_0	0.125	0.104	0.83	1.364	0.902
BLUEPRINT_H3K4me1QTL_MaxCPPL2_0	0.013	0.019	1.42	3.579	0.905
Coding_UCSC.extend.500L2_0	0.064	0.071	1.12	1.446	0.936
Repressed_Hoffman.extend.500L2_0	0.719	0.711	0.99	0.134	0.938
Repressed_HoffmanL2_0	0.461	0.481	1.04	0.559	0.938
DHS_peaks_TrynkaL2_0	0.111	0.119	1.08	2.509	0.975
Baseline	1	1	1	0	NA

Prop_SNPs; the proportion of SNPs included in the model which have the functional annotation given in the annotation column. Prop_SNP_h2; the proportion of SNP based heritability accounted for by SNPs which have the functional annotation. Fold enrichment; enrichment in heritability conferred by SNPs with the functional annotation over the baseline model. P for enrichment; test of the null hypothesis that the fold enrichment value = 1.

3.4: Heritability of Periodontitis/loose teeth in genomic regions with tissue-specific annotation

Tissue name	Stratified LDSR coefficient	SE for coefficient	P_value
A03.734.Pancreas	$5.24E^{-09}$	$2.37E^{-09}$	0.01
A11.872.190.Embryonic.Stem.Cells	$5.18E^{-09}$	$2.28E^{-09}$	0.01
Cells_EBV-transformed_lymphocytes	$4.81E^{-09}$	$2.27E^{-09}$	0.02
Pancreas	$4.24E^{-09}$	$2.06E^{-09}$	0.02
A11.872.Stem.Cells	$4.06E^{-09}$	$2.32E^{-09}$	0.04
A11.118.637.555.567.562.B.Lymphocytes	$3.95E^{-09}$	$2.47E^{-09}$	0.05
Minor_Salivary_Gland	$3.87E^{-09}$	$1.94E^{-09}$	0.02
A05.360.444.492.362.Foreskin	$3.82E^{-09}$	$1.84E^{-09}$	0.02
A11.329.629.Osteoblasts	$3.53E^{-09}$	$2.25E^{-09}$	0.06
A11.329.830.Stromal.Cells	$3.52E^{-09}$	$2.36E^{-09}$	0.07
Adrenal_Gland	$3.33E^{-09}$	$2.15E^{-09}$	0.06
A11.118.637.555.567.569.200.700.T.Lymphocytes..Regulatory	$3.01E^{-09}$	$2.25E^{-09}$	0.09
Brain_Cerebellum	$2.93E^{-09}$	$1.65E^{-09}$	0.04
A11.872.040.Adult.Stem.Cells	$2.91E^{-09}$	$2.22E^{-09}$	0.09
A11.872.378.590.635.Granulocyte.Macrophage.Progenitor.Cells	$2.79E^{-09}$	$2.44E^{-09}$	0.13
A11.872.700.500.Induced.Pluripotent.Stem.Cells	$2.63E^{-09}$	$2.13E^{-09}$	0.11
A06.407.312.782.Testis	$2.62E^{-09}$	$1.87E^{-09}$	0.08
Brain_Cerebellar_Hemisphere	$2.60E^{-09}$	$1.61E^{-09}$	0.05
A15.382.Immune.System	$2.59E^{-09}$	$2.28E^{-09}$	0.13
A11.436.Epithelial.Cells	$2.47E^{-09}$	$2.25E^{-09}$	0.14
Esophagus_Mucosa	$2.22E^{-09}$	$2.28E^{-09}$	0.17
A10.165.450.300.Cicatrix	$2.17E^{-09}$	$2.42E^{-09}$	0.18
A05.810.453.324.Kidney.Cortex	$2.15E^{-09}$	$2.12E^{-09}$	0.16
A03.556.500.760.464.Parotid.Gland	$2.13E^{-09}$	$2.33E^{-09}$	0.18
A11.329.372.600.Macrophages..Alveolar	$2.12E^{-09}$	$2.22E^{-09}$	0.17
A11.436.397.Keratinocytes	$2.08E^{-09}$	$2.24E^{-09}$	0.18
A10.615.284.473.Chorion	$2.00E^{-09}$	$2.11E^{-09}$	0.17
A10.272.Epithelium	$1.98E^{-09}$	$2.28E^{-09}$	0.19
A10.165.450.300.425.Keloid	$1.95E^{-09}$	$2.22E^{-09}$	0.19
A10.615.Membranes	$1.94E^{-09}$	$2.19E^{-09}$	0.19
A05.810.890.Urinary.Bladder	$1.78E^{-09}$	$2.38E^{-09}$	0.23
Uterus	$1.73E^{-09}$	$2.15E^{-09}$	0.21
Pituitary	$1.71E^{-09}$	$2.11E^{-09}$	0.21
Stomach	$1.70E^{-09}$	$1.96E^{-09}$	0.19
A03.556.124.Intestines	$1.68E^{-09}$	$2.29E^{-09}$	0.23
A02.835.232.834.151.Cervical.Vertebrae	$1.64E^{-09}$	$2.16E^{-09}$	0.22
A15.145.229.637.555.567.562.725.Plasma.Cells	$1.64E^{-09}$	$2.30E^{-09}$	0.24
Brain_Hypothalamus	$1.61E^{-09}$	$1.65E^{-09}$	0.16

A15.382.490.555.567.622.Lymphocytes..Null	$1.61E^{-09}$	$2.43E^{-09}$	0.25
A07.541.358.100.Atrial.Appendage	$1.60E^{-09}$	$2.29E^{-09}$	0.24
A04.411.Lung	$1.60E^{-09}$	$2.16E^{-09}$	0.23
A11.382.Endocrine.Cells	$1.53E^{-09}$	$2.65E^{-09}$	0.28
A06.407.071.140.Adrenal.Cortex	$1.50E^{-09}$	$2.11E^{-09}$	0.24
A10.336.707.Prostate	$1.50E^{-09}$	$2.02E^{-09}$	0.23
A11.872.190.260.Embryoid.Bodies	$1.40E^{-09}$	$2.18E^{-09}$	0.26
A03.734.414.Islets.of.Langerhans	$1.34E^{-09}$	$2.21E^{-09}$	0.27
A05.360.319.679.490.Endometrium	$1.34E^{-09}$	$2.08E^{-09}$	0.26
A08.186.211.730.885.287.249.487.Corpus.Striatum	$1.28E^{-09}$	$1.95E^{-09}$	0.26
A15.145.229.637.555.567.569.200.CD4.Positive.T.Lymphocytes	$1.24E^{-09}$	$2.28E^{-09}$	0.29
Small_Intestine_Terminal_Ileum	$1.22E^{-09}$	$2.17E^{-09}$	0.29
A05.360.319.679.256.Cervix.Uteri	$1.20E^{-09}$	$2.17E^{-09}$	0.29
A11.329.228.Fibroblasts	$1.09E^{-09}$	$2.14E^{-09}$	0.31
A15.382.490.555.567.Lymphocytes	$1.08E^{-09}$	$2.31E^{-09}$	0.32
A11.436.294.064.Glucagon.Secreting.Cells	$1.08E^{-09}$	$2.10E^{-09}$	0.30
Brain_Anterior_cingulate_cortex_(BA24)	$1.07E^{-09}$	$1.68E^{-09}$	0.26
A08.186.211.730.885.287.500.571.735.Visual.Cortex	$1.03E^{-09}$	$1.74E^{-09}$	0.28
Brain_Frontal_Cortex_(BA9)	$1.01E^{-09}$	$1.66E^{-09}$	0.27
A11.436.329.Granulosa.Cells	$1.00E^{-09}$	$2.16E^{-09}$	0.32
A11.329.Connective.Tissue.Cells	$9.65E^{-10}$	$2.07E^{-09}$	0.32
Brain_Hippocampus	$9.52E^{-10}$	$1.55E^{-09}$	0.27
Brain_Cortex	$9.49E^{-10}$	$1.68E^{-09}$	0.29
A14.549.Mouth	$9.16E^{-10}$	$2.35E^{-09}$	0.35
A08.186.211.132.Brain.Stem	$9.06E^{-10}$	$1.70E^{-09}$	0.30
A03.556.500.760.Salivary.Glands	$8.80E^{-10}$	$2.27E^{-09}$	0.35
A11.872.378.Hematopoietic.Stem.Cells	$8.37E^{-10}$	$2.55E^{-09}$	0.37
A05.360.319.114.630.Ovary	$8.04E^{-10}$	$1.96E^{-09}$	0.34
Ovary	$8.03E^{-10}$	$2.09E^{-09}$	0.35
A11.872.378.590.817.Megakaryocyte.Erythroid.Progenitor.Cells	$7.67E^{-10}$	$2.21E^{-09}$	0.36
Lung	$7.53E^{-10}$	$2.17E^{-09}$	0.36
A14.549.167.Dentition	$7.43E^{-10}$	$2.35E^{-09}$	0.38
A17.815.Skin	$7.04E^{-10}$	$2.13E^{-09}$	0.37
A03.556.124.526.767.Rectum	$6.76E^{-10}$	$2.17E^{-09}$	0.38
A03.556.875.Upper.Gastrointestinal.Tract	$6.12E^{-10}$	$2.19E^{-09}$	0.39
A14.549.885.Tongue	$5.69E^{-10}$	$2.12E^{-09}$	0.39
A05.360.319.679.Uterus	$5.39E^{-10}$	$2.05E^{-09}$	0.40
Brain_Nucleus_accumbens_(basal_ganglia)	$5.18E^{-10}$	$1.67E^{-09}$	0.38
Skin_Not_Sun_Exposed_(Suprapubic)	$5.12E^{-10}$	$2.04E^{-09}$	0.40
A11.627.635.Myeloid.Progenitor.Cells	$5.05E^{-10}$	$2.53E^{-09}$	0.42
A05.360.319.114.373.Fallopian.Tubes	$4.97E^{-10}$	$1.85E^{-09}$	0.39
A07.541.358.Heart.Atria	$4.14E^{-10}$	$2.26E^{-09}$	0.43
A06.407.071.Adrenal.Glands	$3.98E^{-10}$	$2.16E^{-09}$	0.43

A11.118.637.555.567.569.T.Lymphocytes	$3.67E^{-10}$	$2.08E^{-09}$	0.43
A08.186.211.132.810.428.200.Cerebellum	$3.58E^{-10}$	$1.71E^{-09}$	0.42
A03.556.Gastrointestinal.Tract	$3.25E^{-10}$	$2.39E^{-09}$	0.45
A03.556.249.249.356.Colon	$2.88E^{-10}$	$2.22E^{-09}$	0.45
A15.145.229.637.555.Leukocytes..Mononuclear	$2.83E^{-10}$	$2.34E^{-09}$	0.45
A15.382.520.604.800.Palatine.Tonsil	$2.51E^{-10}$	$2.11E^{-09}$	0.45
Prostate	$1.71E^{-10}$	$2.17E^{-09}$	0.47
A11.627.624.249.Monocyte.Macrophage.Precursor.Cells	$1.66E^{-10}$	$2.24E^{-09}$	0.47
Brain_Caudate_(basal_ganglia)	$1.63E^{-10}$	$1.74E^{-09}$	0.46
Kidney_Cortex	$1.30E^{-10}$	$1.74E^{-09}$	0.47
Artery_Tibial	$1.24E^{-10}$	$1.72E^{-09}$	0.47
Colon_Sigmoid	$1.09E^{-10}$	$1.74E^{-09}$	0.47
A15.382.812.522.Macrophages	$1.02E^{-10}$	$2.21E^{-09}$	0.48
Whole_Blood	$6.88E^{-11}$	$2.01E^{-09}$	0.49
Cervix_Endocervix	$6.77E^{-11}$	$1.98E^{-09}$	0.49
Brain_Amygdala	$5.81E^{-11}$	$1.53E^{-09}$	0.48
A05.360.444.Genitalia..Male	$5.23E^{-11}$	$1.92E^{-09}$	0.49
Artery_Aorta	$4.07E^{-11}$	$1.80E^{-09}$	0.49
Cells_Transformed_fibroblasts	$4.02E^{-11}$	$1.90E^{-09}$	0.49
A10.615.550.Mucous.Membrane	$1.60E^{-11}$	$2.35E^{-09}$	0.50
A15.145.300.Fetal.Blood	$7.25E^{-12}$	$2.51E^{-09}$	0.50
A11.627.340.360.Granulocyte.Precursor.Cells	$-5.21E^{-11}$	$2.51E^{-09}$	0.51
A03.556.875.500.Esophagus	$-6.69E^{-11}$	$2.05E^{-09}$	0.51
A03.556.875.875.Stomach	$-8.32E^{-11}$	$2.12E^{-09}$	0.52
A14.724.Pharynx	$-8.72E^{-11}$	$1.88E^{-09}$	0.52
Brain_Substantia_nigra	$-1.06E^{-10}$	$1.56E^{-09}$	0.53
A07.231.908.Veins	$-1.72E^{-10}$	$2.30E^{-09}$	0.53
A10.615.550.599.Mouth.Mucosa	$-1.77E^{-10}$	$2.30E^{-09}$	0.53
A11.436.275.Endothelial.Cells	$-1.86E^{-10}$	$2.15E^{-09}$	0.53
A02.165.Cartilage	$-2.18E^{-10}$	$2.01E^{-09}$	0.54
A04.531.520.Nasal.Mucosa	$-2.19E^{-10}$	$2.31E^{-09}$	0.54
A11.872.653.Neural.Stem.Cells	$-2.36E^{-10}$	$2.19E^{-09}$	0.54
A11.436.348.Hepatocytes	$-2.80E^{-10}$	$2.40E^{-09}$	0.55
Cervix_Ectocervix	$-3.06E^{-10}$	$2.00E^{-09}$	0.56
A08.186.211.730.885.287.249.Basal.Ganglia	$-3.42E^{-10}$	$1.86E^{-09}$	0.57
A11.118.637.Leukocytes	$-3.49E^{-10}$	$2.57E^{-09}$	0.55
Nerve_Tibial	$-3.75E^{-10}$	$1.84E^{-09}$	0.58
A15.382.490.315.583.Neutrophils	$-3.96E^{-10}$	$2.17E^{-09}$	0.57
A11.872.580.Mesenchymal.Stem.Cells	$-4.17E^{-10}$	$2.08E^{-09}$	0.58
Brain_Putamen_(basal_ganglia)	$-4.20E^{-10}$	$1.71E^{-09}$	0.60
A08.186.211.730.317.Diencephalon	$-4.28E^{-10}$	$2.09E^{-09}$	0.58
A08.186.211.865.428.Metencephalon	$-4.34E^{-10}$	$1.71E^{-09}$	0.60
A14.549.167.646.Periodontium	$-4.65E^{-10}$	$2.46E^{-09}$	0.57
A06.407.312.Gonads	$-4.68E^{-10}$	$1.98E^{-09}$	0.59

A15.145.229.Blood.Cells	$-4.89E^{-10}$	$2.66E^{-09}$	0.57
A05.360.319.Genitalia..Female	$-4.93E^{-10}$	$2.17E^{-09}$	0.59
A05.360.319.887.Vulva	$-5.17E^{-10}$	$2.27E^{-09}$	0.59
A15.145.Blood	$-5.32E^{-10}$	$2.65E^{-09}$	0.58
Liver	$-5.34E^{-10}$	$1.88E^{-09}$	0.61
A07.231.114.Arteries	$-5.56E^{-10}$	$2.45E^{-09}$	0.59
A15.145.846.Serum	$-5.89E^{-10}$	$2.18E^{-09}$	0.61
A03.556.124.684.Intestine..Small	$-6.64E^{-10}$	$2.44E^{-09}$	0.61
A06.407.Endocrine.Glands	$-6.79E^{-10}$	$1.99E^{-09}$	0.63
A11.443.Erythroid.Cells	$-6.83E^{-10}$	$2.08E^{-09}$	0.63
A06.407.900.Thyroid.Gland	$-6.86E^{-10}$	$2.27E^{-09}$	0.62
Artery_Coronary	$-7.04E^{-10}$	$1.72E^{-09}$	0.66
A08.186.211.730.885.287.500.Cerebral.Cortex	$-7.15E^{-10}$	$1.73E^{-09}$	0.66
Spleen	$-7.47E^{-10}$	$2.49E^{-09}$	0.62
A11.497.497.600.Oocytes	$-8.03E^{-10}$	$1.99E^{-09}$	0.66
A07.231.908.670.874.Umbilical.Veins	$-8.11E^{-10}$	$2.24E^{-09}$	0.64
A07.541.510.110.Aortic.Valve	$-8.49E^{-10}$	$2.14E^{-09}$	0.65
A05.360.490.Germ.Cells	$-8.53E^{-10}$	$1.95E^{-09}$	0.67
A08.186.211.464.Limbic.System	$-8.88E^{-10}$	$1.82E^{-09}$	0.69
Breast_Mammary_Tissue	$-8.89E^{-10}$	$1.84E^{-09}$	0.69
Adipose_Subcutaneous	$-9.35E^{-10}$	$1.83E^{-09}$	0.70
A15.382.680.Phagocytes	$-9.35E^{-10}$	$2.21E^{-09}$	0.66
Testis	$-9.37E^{-10}$	$2.03E^{-09}$	0.68
A05.360.Genitalia	$-9.73E^{-10}$	$2.07E^{-09}$	0.68
Colon_Transverse	$-9.75E^{-10}$	$1.96E^{-09}$	0.69
Fallopian_Tube	$-9.87E^{-10}$	$2.04E^{-09}$	0.69
Thyroid	$-9.91E^{-10}$	$2.07E^{-09}$	0.68
Bladder	$-1.00E^{-09}$	$1.63E^{-09}$	0.73
A08.186.211.730.885.287.500.270.Frontal.Lobe	$-1.02E^{-09}$	$1.83E^{-09}$	0.71
A03.556.249.124.Ileum	$-1.04E^{-09}$	$2.46E^{-09}$	0.66
Skin_Sun_Exposed_(Lower_leg)	$-1.04E^{-09}$	$1.95E^{-09}$	0.70
A15.382.520.604.700.Spleen	$-1.10E^{-09}$	$2.36E^{-09}$	0.68
Heart_Atrial_Appendage	$-1.12E^{-09}$	$1.87E^{-09}$	0.73
A02.835.583.443.800.Synovial.Membrane	$-1.12E^{-09}$	$1.97E^{-09}$	0.72
A03.556.124.369.Intestinal.Mucosa	$-1.14E^{-09}$	$2.18E^{-09}$	0.70
Heart_Left_Ventricle	$-1.35E^{-09}$	$1.94E^{-09}$	0.76
A03.556.249.249.209.Cecum	$-1.43E^{-09}$	$2.16E^{-09}$	0.75
A03.556.249.249.356.668.Colon..Sigmoid	$-1.44E^{-09}$	$2.16E^{-09}$	0.75
A15.378.316.Bone.Marrow.Cells	$-1.47E^{-09}$	$2.10E^{-09}$	0.76
Muscle_Skeletal	$-1.47E^{-09}$	$1.93E^{-09}$	0.78
A08.186.211.730.317.357.352.435.Hypothalamo.Hypophyseal.System	$-1.47E^{-09}$	$2.25E^{-09}$	0.74
A09.371.729.Retina	$-1.49E^{-09}$	$1.81E^{-09}$	0.79
A14.724.557.Nasopharynx	$-1.51E^{-09}$	$2.01E^{-09}$	0.77
A11.329.171.Chondrocytes	$-1.59E^{-09}$	$2.13E^{-09}$	0.77

A15.145.229.188.Blood.Platelets	-1.60E-09	2.00E-09	0.79
A15.378.316.580.Monocytes	-1.69E-09	2.10E-09	0.79
Brain_Spinal_cord_(cervical_c-1)	-1.70E-09	1.60E-09	0.86
A08.186.211.Brain	-1.77E-09	1.79E-09	0.84
Esophagus_Gastroesophageal_Junction	-1.82E-09	1.82E-09	0.84
A03.620.Liver	-1.82E-09	2.11E-09	0.81
A08.186.211.464.710.225.Entorhinal.Cortex	-1.92E-09	1.65E-09	0.88
Vagina	-1.97E-09	1.87E-09	0.85
A07.231.Blood.Vessels	-1.99E-09	2.37E-09	0.80
A08.186.211.730.317.357.Hypothalamus	-2.09E-09	2.45E-09	0.80
A15.382.490.555.567.537.Killer.Cells..Natural	-2.14E-09	2.30E-09	0.82
A10.549.Lymphoid.Tissue	-2.17E-09	2.15E-09	0.84
A10.272.497.Epidermis	-2.22E-09	2.16E-09	0.85
A09.371.Eye	-2.23E-09	2.26E-09	0.84
Adipose_Visceral_(Omentum)	-2.36E-09	2.01E-09	0.88
A15.382.812.260.Dendritic.Cells	-2.43E-09	2.18E-09	0.87
A10.165.114.830.500.750.Subcutaneous.Fat..Abdominal	-2.50E-09	1.97E-09	0.90
A05.360.319.679.690.Myometrium	-2.62E-09	2.01E-09	0.90
A10.165.114.830.750.Subcutaneous.Fat	-2.79E-09	1.94E-09	0.92
Esophagus_Muscularis	-2.84E-09	1.79E-09	0.94
A08.186.211.464.405.Hippocampus	-2.91E-09	1.83E-09	0.94
A11.329.114.Adipocytes	-2.92E-09	2.09E-09	0.92
A08.186.211.730.885.287.500.670.Parietal.Lobe	-3.02E-09	1.89E-09	0.94
A07.541.560.Heart.Ventricles	-3.05E-09	1.94E-09	0.94
A10.615.789.Serous.Membrane	-3.25E-09	2.09E-09	0.94
A15.382.812.Mononuclear.Phagocyte.System	-3.28E-09	2.13E-09	0.94
A08.186.211.653.Mesencephalon	-3.32E-09	1.88E-09	0.96
A11.620.520.Myocytes..Smooth.Muscle	-3.33E-09	2.28E-09	0.93
A05.810.453.Kidney	-3.37E-09	1.66E-09	0.98
A10.549.400.Lymph.Nodes	-3.57E-09	2.08E-09	0.96
A02.633.567.850.Quadriceps.Muscle	-3.66E-09	1.84E-09	0.98
A02.835.583.443.800.800.Synovial.Fluid	-3.76E-09	2.27E-09	0.95
A07.541.Heart	-3.82E-09	2.18E-09	0.96
A10.690.467.Muscle..Smooth	-4.27E-09	2.17E-09	0.98
A11.118.637.555.567.562.440.Precursor.Cells..B.Lymphoid	-4.41E-09	2.18E-09	0.98
A10.690.Muscles	-5.40E-09	1.84E-09	1.00

Each row represents a single LDSR model, fitted only using SNPs which are annotated as specifically expressed in the tissue type given in the first column. The coefficient, standard error and P value represent a single-tissue test for heritability.

3.5: Results of S-TissueXcan analysis of DMFS/dentures passing a multiple testing correction

Gene	P	N-tissues	N-indep	z_mean	z_sd	chr	start	stop
ADAM15	2.93E^{-09}	15	11	0.251941	2.384597	1	155050566	155062775
EFNA1	5.39E^{-13}	9	7	-5.4649	1.831997	1	155127460	155134857
SLC50A1	9.15E^{-13}	3	2	5.166324	3.796788	1	155135344	155138857
MUC1	2.68E^{-13}	13	9	3.28801	3.796272	1	155185824	155192916
THBS3	5.35E^{-14}	26	8	4.847069	2.101632	1	155195588	155209051
MTX1	7.09E^{-07}	3	3	0.327615	3.863563	1	155208699	155213824
GBA	1.52E^{-08}	12	7	-3.67951	2.585057	1	155234452	155244699
FAM189B	1.79E^{-08}	6	5	1.40397	3.390339	1	155247205	155255483
RUSC1	1.06E^{-13}	2	2	-4.89512	3.91577	1	155320896	155331114
MRPS14	1.24E^{-06}	7	7	-0.73124	3.021885	1	175010789	175023425
ADCY3	3.49E^{-08}	14	4	0.355393	4.782906	2	24819169	24919839
DTNB	1.99E^{-08}	11	10	-1.15694	1.947262	2	25377198	25673647
C2orf70	2.33E^{-06}	4	4	-2.53654	2.784768	2	26562582	26579532
FNDC4	1.43E^{-06}	3	3	0.544254	3.855838	2	27491883	27495245
GCKR	2.73E^{-07}	5	5	0.14683	3.262635	2	27496842	27523684
ALK	3.90E^{-07}	6	6	-2.00013	2.042634	2	29192774	29921566
GFPT1	1.49E^{-06}	13	10	4.105662	1.512985	2	69319769	69387254
NFU1	2.73E^{-07}	11	6	1.224664	2.898687	2	69395750	69437628
AAK1	1.03E^{-06}	12	9	-3.35472	2.18002	2	69457997	69674349
ASTL	2.94E^{-06}	7	6	1.100384	2.863982	2	96123850	96138436
NEURL3	4.62E^{-07}	3	3	-0.35335	4.248192	2	96497643	96508109
LMAN2L	1.04E^{-06}	25	4	1.364018	1.249396	2	96705929	96740064
CNNM4	3.99E^{-07}	11	8	1.488266	1.674738	2	96760902	96811891
VIL1	1.48E^{-08}	2	2	-1.8059	5.271663	2	218419092	218453295
USP37	1.85E^{-06}	19	5	3.47991	2.470763	2	218450251	218568361
CNOT9	5.96E^{-09}	10	8	1.634806	2.938823	2	218568580	218597080
STK36	2.01E^{-08}	3	3	-3.35756	0.906318	2	218672026	218702716
CNPPD1	1.51E^{-06}	9	9	0.129554	2.327833	2	219171897	219178106
SLC16A14	1.60E^{-06}	5	5	0.31658	2.674233	2	230034974	230068999
RNF123	6.14E^{-11}	33	4	5.260851	0.86952	3	49689499	49721529
GMPPB	1.91E^{-08}	20	3	3.51088	1.151654	3	49716844	49723951
CDHR4	1.81E^{-06}	6	4	1.444559	3.457152	3	49790732	49799835
FAM212A	9.96E^{-08}	13	5	0.941999	4.858038	3	49803254	49805030
UBA7	1.04E^{-08}	17	5	-4.07206	2.74851	3	49805207	49813946
TRAIP	3.68E^{-07}	4	4	3.091897	2.049824	3	49828599	49856574
CAMKV	6.61E^{-09}	4	4	4.772462	2.513174	3	49857988	49870222
MST1R	4.78E^{-08}	24	6	5.57748	0.949591	3	49887002	49903873
RBM6	1.62E^{-11}	47	1	-6.52066	0.310912	3	49940007	50100045
RBM5	4.52E^{-12}	3	3	-3.00256	3.807762	3	50088908	50119021
SEMA3F	3.90E^{-10}	14	5	-3.66232	3.060608	3	50155045	50189075
NPRL2	3.74E^{-08}	3	3	1.324317	3.977378	3	50347330	50351091
ATP13A4	1.99E^{-06}	24	6	-1.1209	1.441766	3	193402077	193593111

CCL28	7.48E^{-07}	5	5	0.421952	2.947342	5	43376645	43412391
PITX1	2.90E^{-12}	9	9	-0.92876	3.741745	5	135027735	135034813
IL9	2.25E^{-06}	1	1	-4.72996		5	135892246	135895827
CARMIL1	3.13E^{-07}	16	11	-0.56849	2.198728	6	25279078	25620530
HIST1H2BA	1.51E^{-08}	1	1	5.66065		6	25726777	25727292
SLC17A4	1.47E^{-12}	3	3	2.232725	6.020267	6	25754699	25781191
SLC17A3	4.53E^{-13}	5	4	-1.98253	2.149642	6	25833066	25882286
TRIM38	1.97E^{-17}	15	10	0.556493	3.119181	6	25962802	25991226
HIST1H1C	1.17E^{-18}	11	8	-2.08478	2.928425	6	26055787	26056428
HFE	1.62E^{-15}	15	8	-1.68956	2.377147	6	26087281	26098343
HIST1H4C	6.02E^{-11}	2	2	-0.07424	8.137953	6	26103876	26104310
HIST1H1T	2.72E^{-06}	1	1	4.690753		6	26107419	26108136
HIST1H1E	5.61E^{-07}	2	2	-1.18209	5.429397	6	26156354	26157107
HIST1H2BD	2.64E^{-30}	5	5	5.142492	5.925571	6	26158146	26171349
HIST1H3D	1.64E^{-12}	3	3	1.745924	5.024016	6	26196840	26197250
HIST1H1D	2.34E^{-16}	7	7	-0.79011	2.981458	6	26234268	26234933
HIST1H4H	1.89E^{-20}	6	6	0.440773	6.216309	6	26277609	26285638
BTN3A2	4.10E^{-18}	48	1	-8.28802	0.87895	6	26365159	26378320
BTN2A2	2.23E^{-21}	37	8	-0.67504	2.407724	6	26383096	26394874
BTN3A1	3.82E^{-07}	33	3	-1.31066	1.37177	6	26402237	26415216
BTN3A3	1.86E^{-11}	17	9	0.177267	2.260897	6	26440472	26453415
BTN2A1	2.48E^{-11}	18	8	-0.76401	2.10598	6	26457904	26476621
BTN1A1	5.86E^{-07}	4	4	2.720774	2.074255	6	26501221	26510422
HMGN4	1.43E^{-23}	21	6	2.371957	2.446113	6	26538405	26546254
ABT1	1.62E^{-15}	18	9	3.183327	2.105255	6	26596952	26600744
ZNF322	1.65E^{-10}	23	5	-2.88664	2.017432	6	26634383	26659752
HIST1H2BJ	3.82E^{-12}	2	2	2.453822	6.586655	6	27125897	27132750
HIST1H2AG	4.02E^{-13}	2	2	-5.46413	2.736518	6	27133042	27135291
HIST1H2BK	1.05E^{-20}	7	6	3.775874	2.084003	6	27146418	27146798
PRSS16	3.95E^{-08}	19	7	4.457505	1.483936	6	27247701	27256624
ZNF184	2.04E^{-19}	5	5	1.666942	4.224954	6	27450743	27473118
HIST1H4J	5.42E^{-19}	3	3	-6.26948	2.743231	6	27824108	27824480
HIST1H3J	2.71E^{-08}	2	2	1.605672	5.448099	6	27890382	27893106
ZNF165	3.26E^{-13}	18	4	-5.0112	3.84148	6	28080975	28089563
ZKSCAN8	1.00E^{-18}	10	9	-0.7717	4.650545	6	28141910	28159472
ZSCAN9	9.69E^{-35}	19	8	2.920388	3.649239	6	28224886	28233482
ZKSCAN4	3.39E^{-21}	4	4	5.474057	3.444257	6	28244623	28252224
NKAPL	3.67E^{-20}	7	7	-0.96662	3.750917	6	28259320	28260958
ZSCAN26	3.08E^{-15}	40	2	-7.00265	1.535033	6	28267010	28278224
PGBD1	8.42E^{-24}	14	11	3.475703	4.535142	6	28281572	28302549
ZSCAN31	1.87E^{-19}	38	5	4.558689	2.617848	6	28324693	28356271
ZKSCAN3	4.69E^{-23}	25	6	-4.48951	3.705752	6	28349914	28369177
ZSCAN12	2.48E^{-21}	9	6	1.276204	3.766074	6	28378955	28399734
ZSCAN23	1.39E^{-15}	43	2	4.617561	1.675619	6	28431930	28443502

ZBED9	1.42E^{-07}	6	6	0.292592	2.251462	6	28571630	28616212
TRIM27	2.31E^{-24}	19	7	-2.16302	2.526513	6	28903002	28923989
ZNF311	2.03E^{-20}	22	7	-0.18797	2.400061	6	28994785	29005316
OR2J3	1.19E^{-10}	1	1	-6.44048		6	29111891	29112826
OR2J2	4.67E^{-07}	1	1	5.039566		6	29173303	29174574
GABBR1	1.78E^{-15}	12	9	-2.03863	3.67285	6	29555629	29633976
OR2H2	8.64E^{-21}	9	9	5.208556	3.271204	6	29587455	29589038
MOG	2.98E^{-15}	9	9	3.720323	4.086387	6	29656981	29672372
ZFP57	3.78E^{-12}	34	2	6.849792	1.307111	6	29672392	29681110
HLA-F	1.01E^{-11}	28	7	0.479277	2.898089	6	29722775	29738528
HLA-G	2.32E^{-23}	37	6	-4.60488	2.017414	6	29826967	29831125
HLA-A	1.87E^{-17}	39	5	-0.04365	3.246746	6	29941260	29945884
ZNRD1	5.35E^{-08}	28	7	-1.90339	1.658699	6	30058899	30064909
RNF39	1.87E^{-19}	27	9	-1.11091	3.176834	6	30070266	30075887
TRIM31	1.24E^{-17}	12	8	-4.01118	4.302959	6	30102897	30113106
TRIM15	6.09E^{-17}	5	5	-3.22756	2.989256	6	30163206	30172696
TRIM26	1.32E^{-11}	15	8	1.503714	2.472252	6	30184455	30213427
TRIM39	1.71E^{-06}	2	2	-0.01982	5.082462	6	30326479	30343729
RPP21	1.39E^{-10}	9	9	-1.2472	2.680248	6	30345131	30346884
HLA-E	2.15E^{-14}	14	10	3.719096	1.801237	6	30489467	30494205
PRR3	4.46E^{-14}	4	4	-2.16397	4.088682	6	30556886	30563723
ABCF1	7.07E^{-09}	21	6	2.996414	1.200235	6	30571376	30597179
C6orf136	1.66E^{-06}	1	1	4.790696		6	30647039	30653210
PPP1R18	9.48E^{-24}	16	10	-6.94352	3.47947	6	30676389	30687895
NRM	1.23E^{-11}	26	8	-1.17453	2.326583	6	30688047	30691420
MDC1	9.81E^{-29}	9	9	-0.75003	4.589329	6	30699807	30717889
FLOT1	3.57E^{-17}	24	10	6.309973	2.640222	6	30727709	30742733
IER3	1.13E^{-20}	10	9	1.228614	3.732468	6	30743199	30744554
DDR1	1.06E^{-15}	13	11	-0.10487	2.956899	6	30876421	30900156
SFTA2	9.42E^{-08}	7	6	1.300718	3.257278	6	30931353	30955636
DPCR1	7.04E^{-07}	5	4	-2.8258	1.298908	6	30934523	30954221
MUC22	7.66E^{-08}	2	2	1.796956	5.444641	6	31010474	31035402
C6orf15	2.14E^{-11}	4	4	2.717007	4.61017	6	31111223	31112559
PSORS1C1	1.28E^{-15}	41	4	-4.77194	1.397113	6	31114750	31140092
CDSN	2.28E^{-06}	2	2	-0.10638	1.813963	6	31115090	31120446
PSORS1C2	6.51E^{-10}	32	5	-2.75917	1.546885	6	31137536	31139350
CCHCR1	1.84E^{-16}	40	5	5.971202	2.853549	6	31142439	31158238
TCF19	4.97E^{-18}	36	6	1.205737	3.020038	6	31158542	31167159
POU5F1	1.77E^{-17}	41	7	1.018697	3.539623	6	31164337	31180731
HLA-C	3.08E^{-10}	46	2	5.916451	0.990761	6	31268749	31272130
HLA-B	3.25E^{-19}	33	8	1.626045	2.422798	6	31269491	31357188
DDX39B	5.94E^{-08}	8	7	-0.18835	2.287757	6	31530219	31542448
ATP6V1G2	2.64E^{-18}	32	5	-3.2573	1.965887	6	31544462	31548427
LTA	3.04E^{-07}	6	6	-0.10932	2.571185	6	31572054	31574324

LST1	$4.57E^{-10}$	10	10	2.761777	2.033456	6	31586124	31588909
PRRC2A	$3.37E^{-07}$	8	8	-0.516	2.606589	6	31620720	31637771
BAG6	$3.15E^{-16}$	31	6	-2.05331	1.99045	6	31639028	31652705
APOM	$4.58E^{-23}$	8	7	-6.05869	3.563013	6	31652416	31658210
GPANK1	$2.37E^{-14}$	14	12	-4.15989	1.862818	6	31661229	31666283
CSNK2B	$6.82E^{-14}$	9	7	0.166044	3.232528	6	31665236	31670343
LY6G5B	$4.90E^{-07}$	39	3	2.882316	1.258882	6	31670167	31673776
ABHD16A	$1.64E^{-11}$	9	9	0.140066	2.879436	6	31686949	31703444
LY6G6F	$4.98E^{-13}$	1	1	7.22585		6	31706885	31710595
MPIG6B	$2.76E^{-16}$	5	5	-0.89927	4.644071	6	31718594	31726714
LY6G6C	$3.15E^{-16}$	4	4	-1.18332	4.866289	6	31718648	31721845
DDAH2	$1.78E^{-18}$	13	7	-3.6862	2.320543	6	31727038	31730617
CLIC1	$2.40E^{-23}$	9	8	-6.01	4.184418	6	31730581	31739763
MSH5	$1.78E^{-17}$	8	8	-4.04092	4.15696	6	31739948	31762834
VWA7	$1.12E^{-06}$	7	6	1.863047	1.832476	6	31765590	31777294
HSPA1B	$2.00E^{-06}$	5	5	2.141879	2.750441	6	31827735	31830255
C6orf48	$9.04E^{-07}$	17	10	1.304761	1.709129	6	31834608	31839766
SLC44A4	$4.02E^{-09}$	6	6	0.521444	3.385357	6	31863192	31879046
C2	$2.96E^{-14}$	7	7	2.671952	2.876154	6	31897785	31945672
CFB	$1.42E^{-18}$	4	4	3.228679	4.17867	6	31945650	31952084
NELFE	$1.92E^{-11}$	16	7	0.567779	2.231454	6	31952087	31959110
DXO	$5.71E^{-15}$	16	9	0.375363	1.771179	6	31969810	31972292
STK19	$6.67E^{-08}$	25	7	-0.72817	1.596113	6	31971091	31982821
C4A	$3.99E^{-21}$	44	1	-8.54648	0.754306	6	31982024	32002681
C4B	$4.01E^{-22}$	31	14	3.987195	3.857333	6	32014762	32035418
CYP21A2	$6.47E^{-16}$	29	11	4.377275	2.840009	6	32038265	32041670
TNXB	$1.32E^{-14}$	4	4	-1.89039	3.822404	6	32041154	32115334
ATF6B	$3.63E^{-13}$	28	2	0.073245	1.975064	6	32098176	32128253
PRRT1	$2.99E^{-15}$	13	8	0.1331	2.462147	6	32148359	32154373
PPT2	$6.43E^{-20}$	4	4	6.434022	3.159223	6	32153441	32163680
AGPAT1	$1.01E^{-09}$	5	5	2.228053	2.230736	6	32168212	32178096
RNF5	$1.07E^{-23}$	27	5	-8.60541	0.973198	6	32178354	32180793
AGER	$4.67E^{-19}$	22	11	4.653121	3.388794	6	32180968	32184324
PBX2	$1.53E^{-08}$	4	4	-2.28682	3.359166	6	32184741	32190186
NOTCH4	$2.31E^{-21}$	36	5	-5.22609	2.281906	6	32194843	32224067
C6orf10	$2.86E^{-06}$	1	1	-4.68041	NA	6	32288526	32371912
BTNL2	$3.18E^{-30}$	6	6	-6.00535	2.727268	6	32393963	32407128
HLA-DRA	$4.28E^{-09}$	7	7	-2.70476	2.317299	6	32439842	32445046
HLA-DRB1	$1.87E^{-18}$	43	5	-1.61793	2.388344	6	32578769	32589848
HLA-DQA1	$1.46E^{-15}$	37	4	-4.19769	2.298061	6	32628179	32647062
HLA-DQB1	$9.08E^{-11}$	47	2	-6.14702	0.994262	6	32659467	32668383
HLA-DQB2	$2.85E^{-16}$	47	3	5.817022	1.305037	6	32756098	32763534
TAP2	$3.84E^{-14}$	35	6	1.855794	3.205718	6	32821833	32838780
PSMB8	$8.97E^{-08}$	4	4	1.941393	2.758011	6	32840717	32844703

PSMB9	3.08E^{-17}	27	3	6.066171	1.105325	6	32844136	32859585
TAP1	5.07E^{-14}	15	11	0.55157	3.50826	6	32845209	32853978
HLA-DMB	2.49E^{-13}	8	7	3.270928	4.718708	6	32934629	32941070
HLA-DMA	2.02E^{-11}	26	4	5.364321	1.908509	6	32948613	32969094
BRD2	6.34E^{-12}	9	8	1.012832	3.374747	6	32968660	32981505
HLA-DPA1	7.86E^{-07}	22	7	-1.11891	1.378812	6	33064569	33080775
HLA-DPB1	9.55E^{-08}	19	17	-0.20115	1.900049	6	33075926	33087201
RXRB	9.45E^{-10}	5	5	-3.21581	2.95856	6	33193588	33200688
RING1	5.26E^{-07}	6	6	-0.83203	2.271586	6	33208495	33212722
VPS52	6.38E^{-08}	7	6	-0.59437	2.277717	6	33250272	33272047
DMRTA1	1.42E^{-06}	18	10	1.366887	2.568891	9	22446841	22455740
RFK	3.21E^{-07}	22	21	-1.65888	1.625194	9	76385517	76394517
PBX3	7.82E^{-08}	16	6	-4.90126	2.123415	9	125747345	125967377
ARAP1	6.27E^{-07}	7	4	1.511052	3.846908	11	72685069	72793599
STARD10	1.15E^{-06}	34	4	-1.98955	1.872194	11	72754729	72794168
FNTB	1.85E^{-06}	11	8	0.172436	2.272999	14	64986720	65062652
LACTB	5.07E^{-11}	31	4	5.115038	1.111403	15	63121800	63142061
RAB8B	4.85E^{-10}	3	3	-0.49883	5.478217	15	63189469	63267782
APH1B	1.13E^{-10}	6	6	4.16588	3.125151	15	63276018	63309126
CA12	2.78E^{-17}	12	9	1.907738	5.271893	15	63321378	63382161
SNX22	1.42E^{-08}	13	6	-0.8993	1.477159	15	64151715	64157481
ACSBG1	9.41E^{-09}	13	8	0.324978	1.868057	15	78167468	78245688
IREB2	4.15E^{-07}	6	5	-2.27819	2.533153	15	78437431	78501456
PSMA4	2.54E^{-06}	21	7	1.87223	1.700884	15	78540405	78552419
NPIPB11	1.17E^{-06}	2	2	4.800154	0.080711	16	29381354	29404029
NPIPB12	5.56E^{-07}	33	4	4.300259	0.774906	16	29483642	29505999
KCTD13	1.80E^{-06}	9	7	2.251119	2.034023	16	29905012	29927035
TMEM219	9.81E^{-07}	3	3	-4.19194	1.853293	16	29940885	29973052
INO80E	1.96E^{-07}	37	2	-4.8427	0.896199	16	29995294	30005793
PPP4C	1.78E^{-06}	3	3	-2.71839	2.154113	16	30075978	30085377
MAPK3	2.10E^{-07}	18	5	-3.69107	1.356344	16	30114105	30123506
KPNB1	2.62E^{-07}	5	5	-3.03176	2.034554	17	47649476	47685505
TBKBP1	1.52E^{-06}	35	3	-4.20623	0.978262	17	47694081	47712050
SP2	1.16E^{-06}	6	6	0.662576	2.852653	17	47896150	47928957
PNPO	1.19E^{-06}	30	6	-3.34289	0.96532	17	47941506	47948288
COPZ2	3.58E^{-07}	13	11	0.540838	2.149376	17	48026167	48038030
CBX1	6.84E^{-09}	11	8	-3.33578	1.328616	17	48070052	48101521
SKAP1	7.14E^{-09}	8	8	2.238794	2.020179	17	48133440	48430275
HOXB3	4.15E^{-07}	13	9	1.568702	2.070148	17	48548870	48604912
NTN5	1.94E^{-10}	19	7	5.703861	1.706547	19	48661407	48673081
FUT2	1.08E^{-10}	18	7	-1.71532	5.603353	19	48695971	48705950
MAMSTR	7.38E^{-11}	16	9	4.848071	1.876859	19	48712742	48719721
RASIP1	1.69E^{-10}	12	6	4.540053	3.201108	19	48720587	48740721
IZUMO1	2.52E^{-08}	29	5	1.825827	1.919832	19	48740852	48746909

TM9SF4	1.61E^{-06}	6	5	-0.05062	3.340518	20	32109506	32167258
MTMR3	7.80E^{-08}	41	2	-4.85276	0.928012	22	29883155	30030866
RIBC2	4.61E^{-07}	31	4	-3.46191	1.051483	22	45413691	45432496

N-tissues; the number of tissue-specific association tests available. N-indep; the number of independent tissue-specific tests taking into account correlation in transcription patterns across different tissues. Z-mean; the mean Z score in tissue-specific association tests. Z-sd; the standard deviation of tissue specific Z-scores.

3.6: Results of S-TissueXcan analysis of Periodontitis/loose teeth passing a multiple testing correction

gene_name	pvalue	N-tissues	N-indep	z_mean	z_sd	chr	start	end
SIGLEC5	8.73E^{-07}	23	13	3.167701	1.941799	19	51611927	51645545

N-tissues; the number of tissue-specific association tests available. N-indep; the number of independent tissue-specific tests taking into account correlation in transcription patterns across different tissues. Z-mean; the mean Z score in tissue-specific association tests. Z-sd; the standard deviation of tissue specific Z-scores.

3.7: Acknowledgements for chapter 3

UK Biobank

This research has been conducted using the UK Biobank Resource (Application; 40644). UK Biobank was established by Wellcome, the Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. UK Biobank has received funding from the Welsh Government British Heart Foundation, Cancer Research UK and Diabetes UK, and has received support from the National Health Service.

ARIC

The Atherosclerosis Risk in Communities Study was carried out as a collaborative study supported by National Heart, Lung and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; National Institutes of Health contract HHSN268200625226C; National Institute of Environmental Health Sciences grant

P30ES010126; and National Institute of Dental and Craniofacial Research grants R01DE11551, R01DE021418 and R01DE023836. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research NCAT grant UL1-RR025747. Analysts working on this project (Kimon Divaris and Cary Agler) are partially supported by U01-DE025046.

COHRA1

Analysis and genotyping in COHRA were funded through NIH grants U01-DE018903, R01-DE014899, R03-DE021425 and R03-DE024264, R56-DE027055. Genotyping was performed as part of the GENEVA consortium by the Center for Inherited Disease Research (www.cidr.jhmi.edu) through an NIH contract. Genome-wide summary statistics are available through the Human Genomics Analysis Interface of the FaceBase consortium (URL: http://FaceBase.sdmgenetics.pitt.edu/, NIH Grant # 5U01-DE024425).

DRDR

The Dental Registry and DNA Repository is supported by the University of Pittsburgh and funded through NIH grant U01-DE018903. Genotyping was performed as part of the GENEVA consortium by the Center for Inherited Disease Research (www.cidr.jhmi.edu) through an NIH contract.

HCHS/SOL

The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung and Blood Institute (NHLBI) to the University of North Carolina (N01- HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236) and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Kelsey Grinde

was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082.

MDC

The Malmö Diet Cancer study has received support from a number of sources including the Swedish Research Council, Swedish Heart-Lung Foundation, Swedish Cancer Foundation Albert Pahlsson Foundation, Lundströms Foundation and the city of Malmö, Sweden.

NFBC1966

The NFBC study was launched by the late Professor Paula Rantakallio. The 46 year follow-up study was made possible by the generosity of NFBC participants who continue to remain active in the study and support from and the NFBC project centre. NFBC1966 received financial support from University of Oulu Grant no. 24000692, Oulu University Hospital Grant no. 24301140 and ERDF European Regional Development Fund Grant no. 539/2010 A31592. Mr. Jari Päkkilä designed the software (electronic patient file) used in NFBC1966.

SHIP and SHIP-TREND

SHIP is part of the Community Medicine Research network (CMR http://www.community-medicine.de) of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103 and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. Generation of genome-wide SNP data has been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg, West Pomerania. The University of Greifswald is a member of the Caché Campus programme of InterSystems GmbH.

TWINGENE

TwinGene is a substudy of the Swedish Twin Registry which is managed by Karolinska Institutet and receives funding through the Swedish Research Council under the grant no 2017-00641.

WGHS

The WGHS is supported by the National Heart, Lung, and Blood Institute (HL043851 and HL080467) and the National Cancer Institute (CA047988 and UM1CA182913) with funding for genotyping from Amgen. Yau-Hua Yu is supported by the National Institute of Dental and Craniofacial Research (1K23DE026804-01A1).

BBJ

The Biobank Japan project was supported by the Ministry of Education, Culture, Sports, Science, and Technology, Japanese government and the Japan Agency for Medical Research and Development.

TMDUAGP

The Tokyo Medical and Dental University Aggressive Periodontitis Study received support from the Japan Society for the Promotion of Science.

Swedish GLIDE

The Swedish GLIDE project is funded by the Swedish Research Council (Dnr 2011-3372 and 2015-02597), the Västerbotten County Council and Umeå University, Sweden. Analysis used the computation facilities of the Advanced Computing Research Centre (http://www.bris.ac.uk/acrc/), the Research Data Storage Facility of the University of Bristol (http://www.bris.ac.uk/acrc/storage/) and the High Performance Computing Center North (HPC2N) at Umeå University (http://www.hpc2n.umu.se).

External resources

The LDHub and PhenoScanner resources are made possible by studies and databases which made GWAS summary data available. These are listed in full online at (http://ldsc.broadinstitute.org/about/) and (http://www.phenoscanner.medschl.cam.ac.uk/information.html).

LDHub gratefully acknowledges the contributions of ADIPOGen (Adiponectin genetics consortium), C4D (Coronary Artery Disease Genetics Consortium), CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis), CKDGen (Chronic Kidney Disease Genetics consortium), dbGAP (database of Genotypes and Phenotypes), DIAGRAM (DIAbetes Genetics Replication And Meta-analysis), ENIGMA (Enhancing

Neuro Imaging Genetics through Meta Analysis), EAGLE (EArly Genetics & Lifecourse Epidemiology Eczema Consortium, excluding 23andMe), EGG (Early Growth Genetics Consortium), GABRIEL (A Multidisciplinary Study to Identify the Genetic and Environmental Causes of Asthma in the European Community), GCAN (Genetic Consortium for Anorexia Nervosa), GEFOS (GEnetic Factors for OSteoporosis Consortium), GIANT (Genetic Investigation of ANthropometric Traits), GIS (Genetics of Iron Status consortium), GLGC (Global Lipids Genetics Consortium), GPC (Genetics of Personality Consortium), GUGC (Global Urate and Gout consortium), HaemGen (haematological and platelet traits genetics consortium), HRgene (Heart Rate consortium), IIBDGC (International Inflammatory Bowel Disease Genetics Consortium), ILCCO (International Lung Cancer Consortium), IMSGC (International Multiple Sclerosis Genetic Consortium), MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium), MESA (Multi-Ethnic Study of Atherosclerosis), PGC (Psychiatric Genomics Consortium), Project MinE consortium, ReproGen (Reproductive Genetics Consortium), SSGAC (Social Science Genetics Association Consortium) and TAG (Tobacco and Genetics Consortium), TRICL (Transdisciplinary Research in Cancer of the Lung consortium), UK Biobank., Alkes Price (the systemic lupus erythematosus GWAS and primary biliary cirrhosis GWAS) and Johannes Kettunen (lipids metabolites GWAS).

Appendix 4

4.1: Acknowledgements for chapter 4

Genome-wide association summary statistics were made available by the GIANT, ENGAGE, DIAGRAM, GLGC, MEGASTROKE and CARDIoGRAMplusC4D consortia. These consortia are supported by numerous sources described in the publications referenced in chapter 4, and at http://megastroke.org/acknowlegements.html for the MEGASTROKE project.

Specifically, data were downloaded from the following sources;

GIANT consortium (adiposity traits);

https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

The ENGAGE consortium (fasting glucose);

http://mccarthy.well.ox.ac.uk/publications/2015/ENGAGE_1KG/.

The DIAGRAM consortium (type 2 diabetes); http://www.diagram-consortium.org/downloads.html.

GLGC (lipid traits); http://csg.sph.umich.edu/willer/public/lipids2013/.

MEGASTROKE (all stroke); http://megastroke.org/download.html.

CARDIoGRAMplusC4D (coronary artery disease/myocardial infarction);

www.cardiogramplusc4d.org

Appendix 5

5.1: Acknowledgements for chapter 5

Mark McCarthy and the Early Growth Genetics consortium helped recruit studies which contributed to this analysis.

For **ALSPAC**, a general acknowledgement statement was included earlier in this appendix. Specific to this analysis, collection of phenotypic data was supported by Wellcome and the UK Medical Research Council [grant number 076467/Z/05/Z]. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf).

The **Young Finns Study** has been financially supported by the Academy of Finland [grant numbers 286284,134309(Eye), 126925,121584,124282,129378 (Salve), 17787 (Gendi), and 41071 (Skidi)] the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility area of Kuopio, Tampere and Turku University Hospitals [grant number X51001]; Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research ; Finnish Cultural Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahnsson Foundation; Signe and Ane Gyllenberg Foundation; and Diabetes Research Foundation of Finnish Diabetes Association.

Analysis within the **GENEVA consortium** was supported by the following USA National Institutes of Health (NIH) grants from the National Institute of Dental and Craniofacial Research (NIDCR): [grant numbers R01-DE014899, U01-DE018903, R03-DE024264, R01-DE09551, R01-DE12101, P60-DE-013076], and a National Institutes for Health contract [contract number HHSN268200782-096C]

Analysis within **Raine** was supported by the National Health and Medical Research Council of Australia [grant numbers 572613 and 40398] and the Canadian Institutes of Health Research [grant number MOP-82893]. Raine is made possible by the contributions of the study participants and their families and the Raine Study research staff. Raine has received long term funding from the NH&MRC and also the following institutes for providing funding for Core Management of the Raine Study: The University of Western Australia (UWA), Curtin University, the Raine Medical Research Foundation, the UWA Faculty of Medicine, Dentistry and Health Sciences, the Telethon Kids Institute, the Women's and Infant's Research Foundation (King Edward Memorial Hospital), Murdoch University, the University of Notre Dame (Australia), and Edith Cowan University. The Western Australian DNA Bank (National Health and Medical Research Council of Australia National Enabling Facility) assisted with preparation of genetic data, and analysis was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and Government of Western Australia.

The **Generation R** Study is conducted by the Erasmus Medical Center in close collaboration with the School of Law and Faculty of Social Sciences of the Erasmus University Rotterdam,

the Municipal Health Service Rotterdam area, Rotterdam, the Rotterdam Homecare Foundation, Rotterdam and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond [STAR-MDC], Rotterdam. The study was made possible by the contribution of children and parents, general practitioners, hospitals, midwives and pharmacies in Rotterdam. The generation and management of GWAS genotype data for the Generation R Study was done at the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, The Netherlands. Pascal Arp, Mila Jhamai, Marijn Verkerk, Manoushka Ganesh, Lizbeth Herrera and Marjolein Peters helped create and manage the GWAS database. The general design of Generation R Study was made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Netherlands Organisation for Scientific Research (NWO), the Ministry of Health, Welfare and Sport and the Ministry of Youth and Families. Additionally, the Netherlands Organization for Health Research and Development supported the Generation R Study [ZonMw 907.00303, ZonMw 916.10159, ZonMw VIDI 016.136.361 and ZonMw VIDI 016.136.367] to Fernando Rivadeneira and Carolina Medina-Gomez. The Generation R project also received funding from the European Union's Horizon 2020 research and innovation programme under the following grant agreements: [No. 633595 (DynaHEALTH) and No. 733206 (LIFECYCLE)]. Generation R received additional funding from the European Research Council [ERC Consolidator Grant, ERC-2014-CoG-648916].