VITAMIN D AND HEAD AND NECK CANCER: AN EXAMINATION OF CAUSALITY AND MECHANISMS.

Tom Dudding

"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"

Bristol Medical School: Population Health Sciences, September 2019

Word count: 57,841

Abstract

There is strong evidence for a protective association of vitamin D on head and neck cancer (HNC) incidence and prognosis from mechanistic and observational epidemiological studies, with the most recent estimate suggesting a 30% reduction in odds of HNC (95% confidence interval (CI): 44%, 12%) even after correcting for smoking and alcohol use. This thesis aims to identify the causal effect of vitamin D on HNC risk and prognosis by triangulating estimates from the most robust observational sources and Mendelian randomization (MR).

I have undertaken the most recent observational study of vitamin D on HNC risk and prognosis in a case-cohort study in UK Biobank. Incident oral, oropharyngeal and laryngeal cancer cases were identified from cancer registry data and participants with no cancer diagnosis were used as controls. In survival analyses, death was identified from linked death registry data. The association between pre-diagnosis serum 25 hydroxyvitamin D (25(OH)D), the commonly measured blood metabolite, and HNC risk and prognosis was assessed using logistic and Cox proportional hazards regression respectively. A doubling in vitamin D was related to a reduction in oral (OR = 0.61 (95% CI: 0.50, 0.76)) and oropharyngeal (OR = 0.75 (95% CI: 0.61, 0.91)) cancer risk as well as a reduction in hazards of death for all HNC sites combined (HR = 0.64 (85% CI: 0.53, 0.77)).

As part of a small team, I also performed genome-wide association study of 25(OH)D in 17 cohort studies, including two with whole genome sequencing data, and meta-analysed. This study identified three novel associations including an association with a low-frequency genetic variant (rs117913124) with large effects on 25(OH)D (-0.43 standard deviation change per A allele). Genetic variants identified as being robustly related to 25(OH)D were combined into a genetic proxy for serum 25(OH)D. This genetic proxy was compared to the serum 25(OH)D measure in a phenome-wide association study of 25(OH)D to assess the confounding structure of the two measures.

Further, I used MR to assess the causal association of 25(OH)D on HNC risk in the GAME-ON study and HNC prognosis in Head and Neck 5000 using the genetic proxy identified above. For risk the MR estimate for a doubling of 25(OH)D was OR = 0.74 (95% CI: 0.50, 1.10) and 1.12 (95% CI: 0.52,2.39) for oral and oropharyngeal cancer respectively. For HNC prognosis, a doubling of 25(OH)D corresponded to a HR of 0.90 (95% CI: 0.51, 1.57) for all HNC sites combined.

In summary, this thesis cannot rule out small effects of 25(OH)D on cancer risk but suggests that the effects as large as those reported in observational studies are unlikely. The genetic proxy for 25(OH)D has been improved by the work of this thesis but questions remain about how appropriate a genetic proxy for 25(OH)D may be in studies of cancer.

Dedication and Acknowledgements

I would like to acknowledge The Wellcome Trust for their generous funding and support which has encouraged me to travel, learn and develop collaborations alongside the exciting task of becoming a scientist.

To the participants of the studies I have used in this project, I cannot thank you enough for your time and dedication.

Thank you to my supervisors, Richard Martin, Nic Timpson and Steve Thomas, without your unfailing encouragement, enthusiasm, morning cycling meetings and knowledge this project and many of the other opportunities that have come with it would not have been possible.

I would also like to thank the IEU for the fantastic environment in which to do research. There is a genius around every corner willing to help or discuss an interesting problem. Particular thanks to my office mates Ashley, Ruth and Simon who have put up with my endless questions (and smelly sports kit) and Louise who wrote the PHESANT program and helped interpret the PheWAS results.

Paul Brennan, Mattias Johansson and everyone else in the Genetic Epidemiology group at IARC, thank you for making me feel so welcome. The combination of science, coffee and views of Mont Blanc are very special.

Finally, thank you to my family. Ellie, your brilliance as a mum, doctor and friend has been a constant inspiration throughout these three years. You have always been there to tell me to take a break at exactly the right time and never complain when I disappear off to the office for hours on end. Amelie and Flora, you never fail to put a smile on my face and give me the most amazing excuse to leave work each day. I dedicate this to you.

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

ABSTRACT ..... **I**

DEDICATION AND ACKNOWLEDGEMENTS ..... **II**

AUTHOR'S DECLARATION ..... **III**

TABLE OF CONTENTS ..... **IV**

LIST OF TABLES ..... **X**

LIST OF FIGURES ..... **XI**

GLOSSARY OF TERMS ..... **XIV**

CHAPTER 1 : OVERVIEW **1**

1.1 Chapter objectives **1**

1.2 Thesis overview **1**

1.3 Brief background and rationale **2**

1.4 Thesis aims and objectives **3**

1.4.1 Objectives 3

1.5 Peer-review and publication of work presented in this thesis **4**

CHAPTER 2 : INTRODUCTION **6**

2.1 Chapter objectives **6**

2.2 Head and neck cancer **6**

2.2.1 Background 6

2.2.2 HNC genetic epidemiology 10

2.2.3 Cancer heterogeneity 11

2.2.4 Major risk factors 11

2.2.5 Changing aetiology 13

2.3 Vitamin D 16

2.3.1 Background and history 16

2.3.2 Vitamin D synthesis and metabolism 17

2.3.3 Measurement of vitamin D 18

2.3.4 Optimal vitamin D levels 19

2.3.5 Factors that affect vitamin D levels 19

2.3.6 Vitamin D supplementation 24

2.4 The association between vitamin D and head and neck cancer 25

2.4.1 Vitamin D and cancer 25

2.4.2 Vitamin D and head and neck cancer 26

2.4.3 Vitamin D cancer protective mechanisms 33

2.5 Summary 35

CHAPTER 3 : METHODOLOGICAL OVERVIEW 37

3.1 Chapter objectives 37

3.2 Causality 37

3.2.1 What is causality 37

3.2.2 How do we investigate causality? 37

3.3 Statistical approaches used in this thesis 42

3.3.1 Multivariable regression 42

3.3.2 Stratification 44

3.3.3 Genome-wide association studies 45

3.3.4 Mendelian randomization 47

3.3.5 Meta-analysis 54

3.4 Studies and participants 55

3.4.1 Cohort study 56

3.4.2 Case-control studies 56

3.4.3 UK Biobank 57

3.4.4 Avon Longitudinal Study of Parents and Children 58

3.4.5 GAME ON 60

3.4.6 Head and Neck 5000 61

CHAPTER 4 : REVISITING THE OBSERVATIONAL VITAMIN D HNC ASSOCIATION 63

4.1 Chapter objectives 63

4.2 Introduction 63

4.3 Methods 64

4.3.1 Overview 64

4.3.2 Studies 64

4.3.3 Identification of incident cancer cases 64

4.3.4 Identification of cancer control participants 65

4.3.5 Follow-up of cancer cases 66

4.3.6 25 hydroxyvitamin D 66

4.3.7 Covariables 66

4.3.8 Statistical analysis 67

4.4 Results 70

4.4.1 Baseline characteristics 70

4.4.2 25(OH)D variation by baseline characteristics 73

4.4.3 Association between 25(OH)D and HNC risk 74

4.4.4 Stratification of results by confounders 75

4.4.5 Sensitivity analyses 78

4.4.6 Comparison and meta-analysis of effect sizes with previously published estimates 78

4.4.7 The association between 25(OH)D and HNC prognosis 79

4.4.8 Stratification of survival analyses 81

4.5 Discussion 82

4.5.1 Head and neck cancer risk 82

4.5.2 Head and neck cancer prognosis 84

4.5.3 Conclusions 85

CHAPTER 5 : IDENTIFYING VITAMIN D GENETIC VARIANTS 87

5.1 Chapter objectives 87

5.2 Introduction 87

5.3 Methods 88

5.3.1 Overview 88

5.3.2 Studies 88

5.3.3 Measurement of serum 25(OH)D 89
5.3.4 Whole-genome sequencing, genotyping and imputation 90
5.3.5 Statistical analysis 90

5.4 Results 92
5.4.1 Low-frequency 25(OH)D-associated SNVs 92
5.4.2 Common 25(OH)D-associated SNVs 100

5.5 Discussion 101
5.5.1 GWAS meta-analysis 101
5.5.2 Vitamin D genetic proxy 103

CHAPTER 6 : ASSESSING GENETIC VARIATION AS A POTENTIAL VITAMIN D PROXY 105

6.1 Chapter objectives 105

6.2 Introduction 105

6.3 Methods 107
6.3.1 Overview 107
6.3.2 Study cohort 108
6.3.3 Generation of exposures 108
6.3.4 Outcome measures 110
6.3.5 Covariates 110
6.3.6 Statistical analysis 110

6.4 Results 112
6.4.1 Assessing the GRS 112
6.4.2 Phenome scan analysis 113

6.5 Discussion 118
6.5.1 Conclusions 120

CHAPTER 7 : ASSESSING THE CAUSAL ASSOCIATION BETWEEN VITAMIN D AND HNC 122

7.1 Chapter objectives 122

7.2 Introduction 122

7.3 Methods 123
7.3.1 Overview 123

7.3.2 Proxy variant selection 123
7.3.3 Studies 126
7.3.4 DNA extraction genotyping and imputation 126
7.3.5 Assessment of potential confounders 126
7.3.6 Statistical analysis 127

7.4 Results 129
7.4.1 Study participants 129
7.4.2 SNP Selection 131
7.4.3 Assessment of confounders 131
7.4.4 Causal association between 25(OH)D and oral and oropharyngeal cancer 133
7.4.5 Stratification by smoking status 134
7.4.6 Sensitivity analyses 135

7.5 Discussion 137

7.6 Conclusion 140

CHAPTER 8 : ASSESSING THE CAUSAL ASSOCIATION OF VITAMIN D WITH HNC PROGRESSION 142

8.1 Chapter objectives 142

8.2 Introduction 142

8.3 Methods 144
8.3.1 Overview 144
8.3.2 Studies 145
8.3.3 DNA extraction genotyping and imputation 145
8.3.4 Follow-up of HNC cases 145
8.3.5 Vitamin D genetic variants for MR analysis 146
8.3.6 GWAS of HNC prognosis to generate summary level outcome data 146
8.3.7 Statistical analysis 146

8.4 Results 148
8.4.1 Study participants 148
8.4.2 SNP selection 148
8.4.3 Assessment of GRS 151
8.4.4 Assessment of confounding in case only setting 151
8.4.5 Survival analyses 153

8.4.6 Sensitivity analyses 154

8.5 Discussion 155

8.6 Conclusion 157

CHAPTER 9 : DISCUSSION 159

9.1 Synthesis of findings 159

9.1.1 Vitamin D and HNC risk 160

9.1.2 Vitamin D and HNC prognosis 165

9.2 Strengths and limitations 166

9.2.1 Power 166

9.2.2 Head and neck cancer data 166

9.2.3 Vitamin D genetic proxy 167

9.3 Next steps 171

9.3.1 GWAS of 25(OH)D in UK Biobank 171

9.3.2 Genotyped HNC resources 171

9.4 Concluding remarks 172

REFERENCES 173

APPENDIX 187

List of tables

Table 2-1: ICD10 codes of head and neck cancer sites used throughout this thesis. 7
Table 2-2: Effect of vitamin D on HNC risk from observational studies. 29
Table 4-1: ICD codes for major head and neck cancer sites 65
Table 4-2: Baseline characteristics of cases and controls 72
Table 4-3: Stratification by sex and smoking status. 77
Table 5-1: Participating cohorts and number of DNA samples per cohort 89
Table 5-2: Summary statistics results for the *CYP2R1* low-frequency variant, rs117913124, from 19 studies 94
Table 5-3: Association results for genome-wide significant low-frequency variants from discovery 25(OH)D meta-analysis, before and after conditioning on the lead common *CYP2R1* SNP, rs10741657, and the lead low-frequency *CYP2R1* variant, rs117913124. 97
Table 5-4: Effect of different haplotype combinations of the low frequency (rs117913124) and the common (rs10741657) *CYP2R1* variants on 25OHD levels. 98
Table 5-5: Main novel findings of the GWAS meta-analysis 101
Table 6-1: Effect estimates from GWAS for the genetic variants included in the genetic proxy. 109
Table 6-2: Effect of SNPs on seasonally adjusted 25(OH)D in UK Biobank and Jiang *et al.* (2018). 113
Table 7-1: 25-Hydroxyvitamin D genetic variant details. 125
Table 7-2: GAME-ON and UK Biobank participant summaries 130
Table 8-1: 25-Hydroxyvitamin D genetic variant details. 150
Table 8-2: Association of 25(OH)D and smoking in full and selected UK Biobank samples. 152
Table 8-3: Association of 25(OH)D and number of sexual partners in full and selected UK Biobank samples. 152
Table 9-1: Summary of results. 159

List of figures

Figure 2-1: Change and projected incidence of head and neck cancers in the UK (from [6] with permission from Elsevier) 8
Figure 2-2: Oral cavity cancer survival by stage at diagnosis (from [12] with permission from Elsevier) . 8
Figure 2-3: UK smoking prevalence, 1970-2017. 14
Figure 2-4: Recorded adult per capita consumption, in litres of pure alcohol, by WHO region and the world, 1990-2005 (Reprinted from[45] with permission of WHO). 14
Figure 2-5: Volume of pure alcohol (litres) sold per adult (16+) 15
Figure 2-6: Percentage of oropharyngeal tumours that were HPV positive in studies from 1975 – 2010 (from[48] with permission from Elsevier) 16
Figure 2-7: Diagrammatic representation of the synthesis and degradation of vitamin D detailing the important enzymes involved in the process. 18
Figure 2-8: The association of solar radiation with colon cancer mortality rate in US states from 1959 -1961. (from [120] by permission of Oxford University Press) 25
Figure 2-9: Meta-analysis of studies assessing association of 25(OH)D on (A) colorectal (B) breast and (B) prostate cancer incidence (Reprinted from[9] with permission of WHO). 26
Figure 2-10: Key cancer-related signalling pathways targeted by $1,25(OH)_2D$ (from [133] Reprinted by permission from Springer Nature) 34
Figure 3-1: Directed acyclic graph representing a causal association between vitamin D and HNC incidence 41
Figure 3-2: Directed acyclic graph representing a complete confounded association between vitamin D and HNC incidence. 41
Figure 3-3: Directed acyclic graph representing a reverse causal association between vitamin D and HNC incidence. 42
Figure 3-4: Ascertainment bias by selecting or conditioning on case status. 51
Figure 3-5: Collider bias in the context of studies assessing the causal association between an exposure and disease prognosis. 51
Figure 3-6: Fictional example of how outlier genetic variants can influence the causal estimate. 53
Figure 4-1: Seasonal variation in unadjusted 25-hydroxyvitamin D levels. 68
Figure 4-2: Study profile 71
Figure 4-3: Effects of potential confounders on circulating 25(OH)D. 74
Figure 4-4: Odds ratio for developing HNC per log2 increase in 25(OH)D using conditional and unconditional logistic regression. 75

Figure 4-5: Odds ratio for developing HNC per log2 increase in 25(OH)D stratified by site and known confounders. 76
Figure 4-6: The modelled probability of developing HNC within smoking status strata when an interaction between 25(OH)D and smoking status is included in the model. 77
Figure 4-7: Forest plots showing the meta-analysis of results with Fanidi *et al.* (2016). 79
Figure 4-8: Kaplan Meier plots of disease-free survival for low and high 25(OH)D. 80
Figure 4-9: Hazard ratio for all-cause mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression. 80
Figure 4-10: Hazard ratio for all-cause mortality or recurrence per log2 increase in 25(OH)D stratified by site and known confounders. 81
Figure 4-11: Hazard ratio for HNC specific mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression. 82
Figure 4-12: Hazard ratio for all-cause mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression in participants who survived greater than six months post-diagnosis. 82
Figure 5-1: Schematic of the discovery single variant meta-analysis 89
Figure 5-2: Discovery single-variant meta-analysis. 92
Figure 5-3: Forest plot detailing the per allele effect of rs117913124 on standardised log25(OH)D by cohort. 93
Figure 5-4: Comparison of 25(OH)D levels in homozygous (AA), heterozygote (AG) and non-carriers (GG) of the rs117913124 variant in the ALSPAC WGS participants. 95
Figure 5-5: Forest-plot of the effect of the four common SUNLIGHT variants and of the *CYP2R1* low-frequency variant rs117913124 on standardised log-transformed 25(OH)D levels. 99
Figure 5-6: Association Signals from 11p.15.2 100
Figure 6-1: Plot of effect size estimates of genetic variants included in both Jiang *et al.* and Chapter 5. 109
Figure 6-2: Distribution of seasonally adjusted 25(OH)D with genetic risk score. 112
Figure 6-3: Flow diagram showing details of PHESANT analysis algorithm for both the seasonally adjusted 25(OH)D and exposures. 114
Figure 6-4: Overall graphical view of observational PheWAS results. 115
Figure 6-5: Overall graphical view of MR PheWAS results. 116
Figure 6-6: Effect of individual variants in the GRS on having a shigella infection at the time of IBS diagnosis. 118

Figure 7-1: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders for GAMEON. 131
Figure 7-2: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders for GAMEON with the South American studies removed. 132
Figure 7-3: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders UK Biobank. 133
Figure 7-4: Odds ratio (95% CI) for developing cancer for D doubling in 25-hydroxyvitamin D for (A) oral cavity cancer, (B) oropharyngeal cancer and (C) all sites combined in Mendelian randomization analysis. 134
Figure 7-5: Odds ratio (95% CI) for developing cancer for doubling in 25-hydroxyvitamin D stratified by smoking in (A) oral cavity cancer and (B) oropharyngeal cancer. 135
Figure 7-6: Leave-one-out plots for the three GAME-ON regions. 136
Figure 7-7: Histograms of OR for MR analyses repeated for each possible combination of 25-hydroxyvitamin D genetic proxies. 137
Figure 8-1: Collider bias in the context of an MR study assessing the causal association between 25(OH)D and HNC prognosis. 143
Figure 8-2: Scatter plot demonstrating the relationship between GRS and 25(OH)D levels. 151
Figure 8-3: Kaplan Meier curves for high- and low-25(OH)D genetic risk scores in (A) HN5000 and (B) UK Biobank and (C) high- and low-25(OH)D levels in UK Biobank. 153
Figure 8-4: Hazard ratio for death from any cause for a doubling in 25(OH)D from observational and three MR methods 154
Figure 8-5: Leave-one-out sensitivity analysis plots for method 1. 155

Glossary of terms

$1,25(OH)_2D$ – 1,25 di-hydroxyvitamin D
25(OH)D – 25 hydroxyvitamin D
7-DHC – 7-dehydrocholesterol
BMI – Body mass index
CEU - Northern Europeans from Utah (ancestry)
CYP24A1 – Cytochrome P450 family 24 subfamily A member 1
CYP27B1 – Cytochrome P450 family 27 subfamily B member 1
CYP2R1 – Cytochrome P450 family 2 subfamily R member 1
DBP – Vitamin D binding protein
DHCR7 – 7-dehydrocholesterol reductase
EPIC – European Prospective Investigation into Cancer
GRS – Genetic risk score
GWAS – Genome-wide association study
HNC – Head and neck cancer
HPV – Human papilloma virus
HR – Hazard ratio
HRC – Haplotype reference consortium
IBS – Irritable bowel syndrome
INHANCE – International Head and Neck Cancer Epidemiology Consortium
IQR – Interquartile range
IU = International units
LC – Laryngeal cancer
LD – Linkage xivdisequilibrium
MAF – Minor allele frequency
MET – Metabolic equivalent task
MR – Mendelian randomization
MS – Multiple sclerosis
OC – Oral cavity cancer
OPC – Oropharyngeal cancer
OR = Odds ratio
PAR – Population attributable risk
PheWAS – Phenome-wide association study
RCT – Randomised controlled trial
RR – Risk ratio
SCC – Squamous cell carcinoma
SNP – Single nucleotide polymorphism
SNV – Single nucleotide variant
SZA – Solar zenith angle
UVB – Ultraviolet B
VDR – Vitamin D receptor
WGS – Whole genome sequence

Chapter 1 : Overview

1.1 Chapter objectives

The aims of this chapter are to provide an overview of the thesis including a brief background, rationale and aims of the project.

1.2 Thesis overview

In this thesis I have conducted a series of studies to investigate the causal association between vitamin D and head and neck cancer risk and where possible applied the same methods to investigate the causal association between vitamin D and head and neck cancer prognosis.

Descriptive

Chapter 2 provides an introduction to head and neck cancer, the biology of vitamin D and a review of the scientific literature that underpins the association between vitamin D and head and neck cancers. **Chapter 3** introduces the concept of causality and the framework by which causality will be assessed in this thesis. An outline of the main methods and datasets used within this thesis are then given, chapter specific methods are detailed at the beginning of each chapter. Observational epidemiology, genome-wide association study (GWAS), Mendelian randomization and meta-analysis methods are discussed as well as the Avon Longitudinal Study of Parents and Children, UK Biobank, GAME-ON and Head and Neck 5000 studies.

Application

In **chapter 4**, an observational nested case control study is described which investigates the association of vitamin D with head and neck cancer. The results of this are then meta-analysed and compared with previous studies to provide the most precise estimate of the vitamin D effect available.

Chapter 5 describes a genome wide association study of vitamin D which attempts to identify new associations between genetic variants and vitamin D that can act as reliable proxies for vitamin D in MR studies. In **Chapter 6**, the quality of this genetic proxy and its suitability for Mendelian randomization is discussed with reference to the assumptions of Mendelian randomization. A hypothesis free approach is applied to the 30,000 plus measured variables in UK Biobank to identify how the confounding structure of the vitamin D proxy and serum vitamin D measures differ.

In **Chapter 7,** Mendelian randomization is applied, using the improved genetic proxy identified in chapter 5, to assess the vitamin D head and neck cancer association.

Chapter 8 introduces the additional complexities that arise when Mendelian randomization is used to investigate vitamin D associations with head and neck cancer prognosis. Work is presented for a Mendelian randomization study in a case-only setting to estimate the causal association of vitamin D on cancer prognosis.

Synthesis

Chapter 9 aims to synthesise the findings of this thesis and those in the literature to report on the vitamin D-HNC association using the framework set out in Chapter 3. A critique of the methods used in this thesis is given alongside a discussion of how future studies may be able to navigate these issues.

1.3 Brief background and rationale

Head and neck cancer (HNC) is the world's sixth most common cancer and incidence is increasing. Annually, more than 600,000 new patients are diagnosed with HNC and approximately 350,000 die.[1] Furthermore, only 40–50% of patients with HNC will survive beyond five years and those that do often have facial disfigurement, loss of function and pain as result of surgery or radiotherapy.

The majority of HNC are squamous cell carcinomas (SCC) and the oral cavity and oropharynx are the most commonly affected sites.[1] Despite the majority being SCCs, HNCs are heterogeneous with separate sites having slightly different aetiology.[2] The well-established risk factors for developing HNC are tobacco use, alcohol consumption and oral human papilloma virus infection,[3,4] yet factors that affect prognosis are less well understood. Despite the main risk factors being well understood, approximately 30% of population attributable risk cannot be explained by these factors.[5] In addition, there has been a recent shift in characteristics and a worrying increase in the incidence of oral cavity cancer and non-HPV related oropharyngeal cancer.[6] This change is difficult to explain with the parallel reduction in smoking levels and stable levels of alcohol use and there is therefore a requirement to identify novel causal risk factors which may explain these trends.

Vitamin D is predominantly synthesised in the skin on exposure to ultra-violet B radiation and plays a role in calcium and phosphate balance and bone structure.[7] Despite vitamin D being linked to many disease states over recent decades, evidence of strong causal associations is difficult to establish. Vitamin D is easily modifiable and supplementation is becoming more common, despite this there are concerns that changes in working and lifestyle patterns mean populations have chronically low vitamin D levels.

There is strong evidence for a protective association of vitamin D on cancer incidence and prognosis from mechanistic and observational epidemiological studies. Laboratory studies have shown vitamin

D and its related metabolites have anticarcinogenic effects with relevance to cancers including HNC[8] and randomised controlled trials (RCTs) of vitamin D supplementation corroborate a potential protective effect (relative risk cancer death = 0.89 (95% CI 0.78, 1.02).[9] Vitamin D has also been associated with improved outcome and reduced tumour aggressiveness in breast, colorectal and prostate cancer.[10]

For HNC, the observational estimates suggest large reductions in risk, particularly for oral cavity cancer. Vitamin D levels, measured before or at diagnosis, have also been strongly associated with improved cancer prognosis. Smoking is likely to be a strong confounder of the relationship with HNC incidence but well-designed prospective studies have shown clinically relevant effects even after robust correction for smoking. Despite this strong observational evidence, causality has not been established in either cancer incidence or prognosis. RCTs are not a practical method to investigate HNC incidence due to the relatively rare disease in question and trials of vitamin D as an adjuvant in HNC patients would require stronger evidence before clinical trials would be indicated.

Alternative methods are required to assess the causal effect of vitamin D on HNC incidence and prognosis to inform whether resources should be used in this area. It is into this unoccupied niche which this thesis fits.

1.4 Thesis aims and objectives

This thesis aims to identify the true causal effect of vitamin D on head and neck cancer risk by triangulating evidence from the most robust observational sources and Mendelian randomization. This information can be used to inform whether future research and health resources should be used in this area. Secondly, the thesis aims to conduct preliminary studies to apply these methods in the study of the true causal association between vitamin D and HNC prognosis.

1.4.1 Objectives

1. Provide the most precise estimate of the effect of vitamin D on HNC risk and prognosis using observational epidemiology
2. Identify a robust genetic proxy for vitamin D and demonstrate that it is minimally affected by confounding and reverse causality
3. Use this genetic proxy in Mendelian randomization studies to estimate the causal effect of vitamin D on HNC risk and prognosis

1.5 Peer-review and publication of work presented in this thesis

Published

- Manousaki, D., Dudding, T. *et al.* Low-Frequency Synonymous Coding Variation in *CYP2R1* Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis. *Am. J. Hum. Genet.* **101,** (2017). (Joint first author) – Appendix Paper 1.
  This paper is based on the GWAS work presented in Chapter 5. I conducted this study as part of a collaboration. Within the collaboration, I was a member of a small team who led the study and writing of the manuscript. I ran the GWAS analyses in the whole genome sequenced and imputed ALSPAC data and helped write and edit the manuscript. The meta-analysis of studies was conducted by colleagues at The University of McGill, however I have conducted similar meta-analyses of GWAS dataset for the HNC prognosis pilot work presented in the Appendix. In the paper, the haplotype analyses are presented for the TwinsUK cohort, for this thesis I present equivalent analyses conducted by myself in the ALSPAC cohort.
- Dudding, T. *et al.* Assessing the causal association between 25-hydroxyvitamin D and the risk of oral and oropharyngeal cancer using Mendelian randomization. *Int. J. Cancer* **143,** 1029–1036 (2018). (First author) – Appendix Paper 2
  This paper is based on the work presented in Chapter 7. For this work I designed the study, conducted all analyses and wrote the manuscript.

Other relevant published research:

GWAS methodology

- Haworth, S. et al. Consortium genome-wide meta-analysis for childhood dental caries traits. Hum. Mol. Genet. (2018). doi:10.1093/hmg/ddy237 (Contributing author)
- Dudding, T. et al. Genome wide analysis for mouth ulcers identifies associations at immune regulatory loci. Nat. Commun. 10, 1052 (2019). (First author)
- Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat. Commun. (2019). doi:10.1038/s41467-018-08219-1 (Contributing author)

MR methodology

- Dudding, T., Thomas, S. J., Duncan, K., Lawlor, D. A. & Timpson, N. J. Re-Examining the Association between Vitamin D and Childhood Caries. PLoS One 10, e0143769 (2015). (First author)

- Tan, V. Y. et al. Reassessing the association between circulating Vitamin D and IGFBP-3: Observational and Mendelian randomization estimates from independent sources. Cancer Epidemiol. Biomarkers Prev. 27, 1462–1471 (2018). (Contributing author)
- Dudding T. et al. Examining the causal relationship between vitamin D and serum metabolic measures: A Mendelian randomization study. American Society of Human Genetics, Orlando 2017

Observational Epidemiology

- Gormley M., Richmond R., Dudding T., & Thomas S. Poster: The changing aetiology of Head and Neck Cancer. International Association of Dental Research 2018.

Planned publications

- Observational association of vitamin D and HNC incidence and prognosis in UK Biobank
- Vitamin D phenome wide association scan in UK Biobank
- Genome-wide association of HNC prognosis
- Mendelian randomization study assessing causal effect of vitamin D on HNC prognosis.

Chapter 2 : Introduction

2.1 Chapter objectives

The aims of this chapter are to provide a detailed outline of head and neck cancer, vitamin D and the association between the two. The concepts and rationale briefly discussed in Chapter 1 will be elaborated and include a summary of the current literature assessing the association between vitamin D and head and neck cancer.

2.2 Head and neck cancer

2.2.1 Background

Head and neck cancer (HNC) is the world's sixth most common cancer and incidence is increasing. Annually, more than 600,000 people are diagnosed with HNC and approximately 350,000 people die from this cancer.[1]

HNCs often present at an advanced stage, particularly in the more occult sites such as the oropharynx. They tend to have a poor prognosis with only 40–50% of patients surviving beyond five years, although this varies across subtypes. Treatment of HNC is dependent on anatomical site and cancer stage and aetiology. The most common treatment modalities are surgery, where anatomical site allows, and radiotherapy often combined with chemotherapy.[11] Despite advances in surgical and radiotherapy techniques, people with HNC often suffer life-long facial disfigurement, loss of function and pain as a result of their treatment.

HNC consists of cancers of the oral, oropharyngeal and nasal cavities, the paranasal sinuses, the major and minor salivary glands, the larynx, the lymphatic tissues of the neck and the thyroid. The vast majority of these cancers (approximately 90%) are squamous cell carcinomas (SCC), other histological types include lymphomas, blastomas, sarcomas and neuroendocrine tumours which are much rarer. Originally, these SCCs were thought to differ only in their anatomical location however more recently they have been shown to be heterogeneous, with different sub-sites having different risk-factors, pathogenesis and clinical behaviours.[2] This thesis will focus on SSC of the oropharynx and oral cavity, with some chapters including data on larynx and hypopharynx cancers where available. Other anatomical sites and histological types will not be studied as they are highly heterogeneous and very rare meaning collections of particular types of these cancer will be very small. The main HNC sites and their corresponding ICD10 codes used throughout this thesis are given in the table below (Table 2-1).

Table 2-1: ICD10 codes of head and neck cancer sites used throughout this thesis.

Cancer site	ICD codes	
Oral cavity cancer	ICD10	C003, C004, C005, C006, C009, C020, C021, C022, ,C028, C029, C030, C031, C039, C040, C041, C048, C049, C050, C058, C059, C060, C061, C062, C068, C069
Oropharynx cancer	ICD10	C01, C024, C090, C091, C098, C099, C100, C101, C102,C103, C104, C108, C109, C140, C142, C051, C052
Larynx cancer	ICD10	C320, C321, C322, C323, C328, C329
Other HNC cancer	ICD10	C148, CD000, C110, C111, C118, C119

2.2.1.1 Oral cavity

Anatomically the oral cavity starts anteriorly from the vermillion borders of the lip and extends posteriorly to the border of the hard and soft palate (superiorly), the circumvallate papilla (inferiorly) and the anterior pillar of fauces laterally. As with other HNC sites, 90% of oral cavity cancers (OC) are SCCs and an array of premalignant legions have been reported to be related to the development of these cancers. These include white (leukoplakia) and red (erythroplakia) patches which cannot be attributed clinically to any other cause.

Despite the accessible nature of these cancers allowing easy self-examination, patients often present late and there is a high incidence of spread to the lymph nodes of the neck. Surgical excision is the first line treatment as anatomical access is favourable. Excision is often accompanied by neck dissection to remove the regional lymph nodes. The method of reconstruction depends on the extent of the excision and can vary from primary closure to the use of microvascular free flaps, with or without bone to reconstruct the bony defect. Adjunctive radiotherapy (and to a lesser extent chemotherapy) is considered, for example in cases with large primary tumours or metastases to lower levels of the neck.

OC incidence has increased slightly in the last 10 years and is predicted to continue to do so (Figure 2-1).[6] Prognosis differs according to the stage at diagnosis (Figure 2-2)[12] and in the UK age standardised one- and five-year probability of cancer death is 22.6% and 42.8% respectively.[13]

Figure 2-1: Change and projected incidence of head and neck cancers in the UK (from [6] with permission from Elsevier)

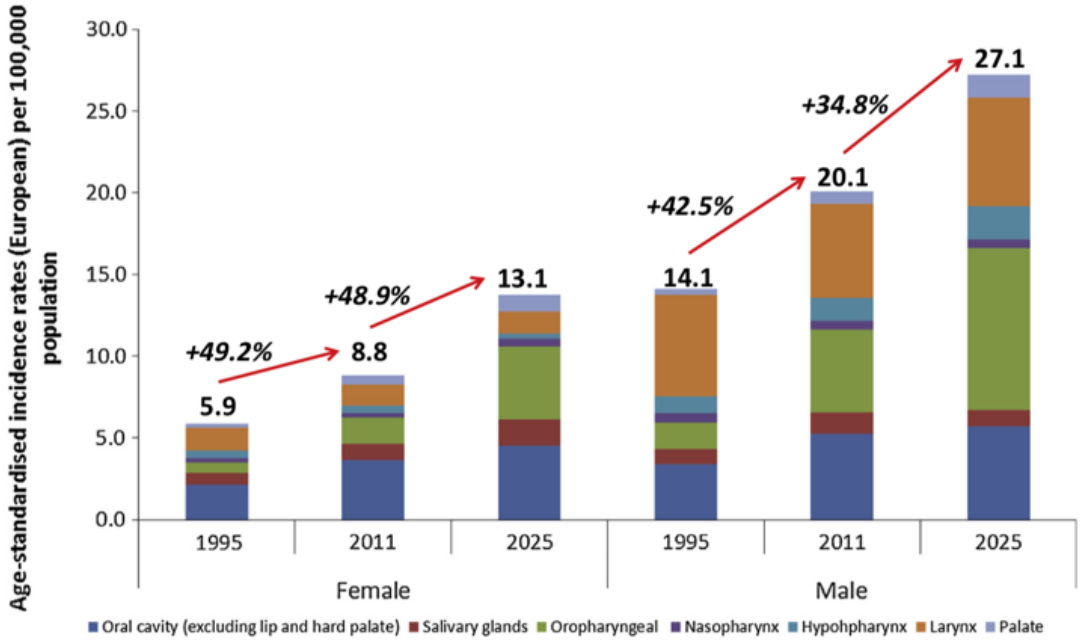
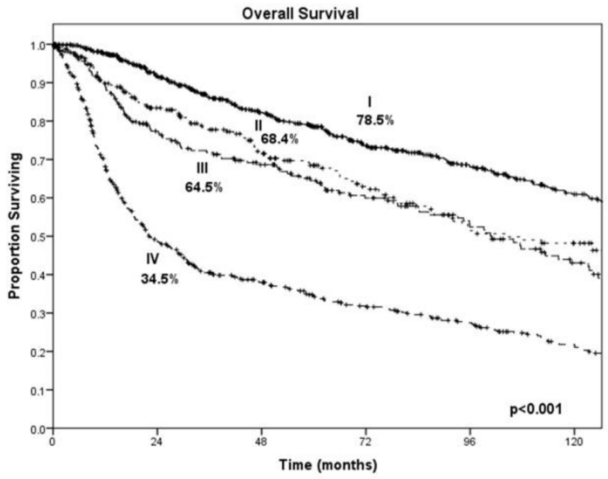


Figure 2-2: Oral cavity cancer survival by stage at diagnosis (from [12] with permission from Elsevier)



2.2.1.2 Oropharynx cancer

The oropharynx constitutes the mid portion of the pharynx and contains the tongue base (posterior to the circumvallate papillae, tonsils (palatine), soft palate and pharyngeal walls. Again 90% of oropharyngeal cancers (OPC) are SCCs. The separation of the oropharynx from the oral cavity is not purely an anatomical one, the tongue base and tonsils contain large amounts of lymphoid tissue whereas the mucosa of the oral cavity does not. This lymphoid tissue forms part of the body's

immuno-surveillance and because of that this tissue encounters many foreign antigens that enter the body via the mouth or nose.

OPCs can be split into Human Papilloma Virus (HPV) positive and HPV negative cancers, with each now considered a separate disease entity.[3] HPV positive OPC patients tend to be younger, more sexually active, generally healthier and of higher socio-economic group whereas HPV-negative OPC patients tend to be heavier smokers and consumers of alcohol and be older.[14]

The treatment of OPC is determined by the p16 status which indicates the aetiological role of HPV. Factors including patient comorbidities, the anatomical position and cancer stage are also important. Treatment consists of either radiotherapy often combined with chemotherapy or surgery. Surgery involves resection of the primary site usually combined with a neck dissection (removal of the draining lymphatics). [15]

OPC incidence is increasing dramatically in the UK (Figure 2-1) and this is largely attributed to rising HPV infection[16] however, non-HPV related OPC is also increasing at a similar rate[17] (see section 2.2.5 Changing aetiology). Oropharyngeal cancer prognosis is better than that of oral cavity cancer with probability of one- and five- year cancer specific deaths being 22.4% and 41.0% respectively in the UK, this is particularly the case for HPV driven cancers.[13]

2.2.1.3 Larynx cancer

The larynx is positioned in the anterior of the neck between the C3 and C7 vertebrae and connects the hypopharynx to the trachea. It extends from the tip of the epiglottis to the inferior border of the cricoid cartilage. The larynx is split into three regions: the supraglottic region which lies above the vocal cords, the glottic region which includes the vocal cords and the area immediately surrounding and inferior to this is the sub-glottic region. Similarly, to both oral and oropharyngeal cancers, laryngeal lancers (LC) are at least 90% SCCs.

LC tend to present earlier than OC and OPC due to voice changes associated with these tumours.[18] They are much more common in men (80%) and this is likely to be due to the higher prevalence of the major risk factors of tobacco and alcohol in males. Together these risk factors are considered to be multiplicative (see 2.2.4 Major risk factors). There is weak evidence linking LC with some dietary and environmental exposures but causal associations have not been established.[19]

Treatment of LC depends predominantly on the stage of the cancer at diagnosis but also on location (supra-glottic, glottic or sub-glottic) with more inferior sites having more difficult access for minimally invasive surgery. Early stage LCs are often treated with minimally invasive trans-oral surgical approaches or radiotherapy. In patients with mid to advanced stage cancers, attempts are

made to preserve the larynx where possible using chemo-radiotherapy, avoiding surgical resection. Where tumours have invaded the cartilage, total laryngectomy is often the treatment of choice. Neck dissection depends on the presence of nodal involvement and the response to the primary treatment.[20]

The incidence of LCs is steadily reducing in the UK (Figure 2-1) and this likely reflects a decline in smoking. Cancer specific death is lower than in OC and OPC with 13% dying within 1 year and 33% dying within 5 years.[13]

2.2.2 HNC genetic epidemiology

There have been several studies which utilise genetic epidemiology to investigate HNC susceptibility over the past few decades. These include various candidate gene studies which have been combined into meta-analyses. These studies tend to focus on genes involved in the metabolism of carcinogens such as *ADH* (the genes for alcohol dehydrogenase isoenzymes) and are reviewed by Cadoni *et al.*[21] In summary, they report some evidence that glutathione S-transferases (*GSTM1*), glutathione S-transferases (*GSTT1*) and human microsomal epoxide hydrolase (*EPHX1*) genes are associated with HNC risk but individual studies are heterogeneous and evidence is statistically weak. The INHANCE consortium of head and neck cancer studies also investigated a list of candidate studies and found some evidence for association with *MGMT* (OR = 0.79, 95% CI = 0.68–0.93), *XRCC1* (OR = 2.3, 95% CI = 1.1–4.7), *ADH1B* (OR = 2.7, 95% CI = 1.9–4.0), *ADH1C* (OR = 1.2, 95% CI = 1.1–1.4), and *GSTM1* (OR = 1.1, 95% CI = 1.0–1.2) although false positives cannot be ruled out,

Two genome wide association studies (GWAS) have been performed to identify genetic variants associated with HNC risk. This method uses a hypothesis free approach across the genome and is discussed further in Chapter 3. These studies have identified several genetic variants with strong evidence for association with HNC including variants in alcohol dehydrogenase genes and in the human leukocyte antigen region of chromosome six which is thought to be related to susceptibility to HPV infection.[22,23]

This thesis will utilise Mendelian randomization, a method which utilises genetic data to help infer causality from observational data. At the start of this theses there had been one study using this method to investigate HNC. This study examined the causal effect of alcohol on HNC using genetic variants in the *ALDH2* gene and found evidence that increased alcohol use increased the risk of HNC.[24]

2.2.3 Cancer heterogeneity

In cancer the term heterogeneity can take many forms beyond that of tumour heterogeneity which, is of current interest to the research community.[25] Despite the fact that the majority of HNC are SCCs, there is heterogeneity across and within anatomical sites as well as more broadly geographically.

It is clear from the descriptions of the three cancer sites above that there is heterogeneity in terms of incidence and aetiology. There is also important within site heterogeneity. As mentioned, OPC cancers can be further stratified into HPV positive and HPV negative cancers with each considered distinct disease entities.[3] A different type of heterogeneity exists in OC. Each location within the oral cavity has distinct tissue compositions, for example the tissue composition of the gingiva and hard palate consists of dense collagenous stroma attached to the periosteum of the bone, the buccal mucosa consists of dense stroma with the presence of minor salivary glands and the tongue is primarily muscular. Furthermore, each of these areas will have different exposure to carcinogens, for example users of betel quid (a combination of areca nut, catechu, slaked lime, and often tobacco wrapped in a betel leaf) will have very high exposure of carcinogens to the buccal mucosa as this is where the folded leaf is placed in the oral cavity. These differences may lead to subtly different cancer pathogenesis across the locations of the oral cavity and also within each tumour, referred to as inter-tumoral heterogeneity.[26]

Geographic heterogeneity is present, for example, the difference in HPV positive OPC in South America compared to North America and Europe. In South America only 4.1% (95% CI: 1.1, 7.1) of OPC tumours are HPV positive compared to 59.3(95% CI: 53.1, 65.5) and 31.1 (95% CI: 22.7, 39.5) in North America and Europe respectively.[27]

These various heterogeneities make investigating HNC complex and highlight the need to stratify by various factors. Without this, associations that are not shared across strata may be missed when analysing all sites or regions together. Even with awareness, this process is difficult in HNC research due to relatively small numbers with each stratum.

2.2.4 Major risk factors

Alcohol and tobacco use are the major risk factors for OC and OPC,[4,28] together explaining approximately 63.7% (95% CI: 44.7, 74.7) and 71.5% (95% CI: 57.6, 80.2) of the population attributable risk (PAR) respectively.[5] Infection with HPV has emerged more recently as an important risk factor, particularly in OPC.[29]

2.2.4.1 Tobacco and Alcohol

Tobacco use is known to be a causal risk factor for HNCs[30] with the most precise estimate being ever smokers having three and a half times greater odds of developing HNC than never smokers (Odds ratio (OR) = 3.46 (95% CI: 3.24, 3.70)). There is a dose response relationship with both intensity (number of cigarettes smoked per day) and duration (number of years smoked) of smoking, with smoking duration being more important than intensity.[31] There was also strong associations exist for cigarette, pipe and cigar smoking.[32] The PAR for tobacco alone (not including the effect of alcohol) in OC and for OPC is approximately 24% for both cancers with the PAR for tobacco alone in LC being more than double this (52.2% (95% CI:77.8, 36.0)).[5]

Alcohol has been considered by the International Agency for Research on Cancer (IARC) as a convincing risk factor for HNCs since the late 1980s[33] and was considered a causal risk factor in their 2010 Monograph.[34] A pooled analysis in the International Head And Neck Cancer Epidemiology (INHANCE) Consortium showed increased odds of OC and OPC with increased intensity and duration of alcohol consumption, but little effect of these factors on LC[35] and this finding is mirrored by several other epidemiological studies. There is currently little evidence to support that different types of alcoholic beverage contribute differing levels of HNC risk, although there is potentially a lower risk attributable to wine compared to beer or liquor at low to moderate alcohol consumption.[36] The PAR for alcohol alone (not including the effect of smoking) in OC is near zero, suggesting most of the effect of alcohol acts through the effect of tobacco. For OPC and LC the PAR for alcohol alone is 5.6 (95% CI: 1.9, 7.3) and 2.9 (95% CI: -0.3, 4.4) respectively.[5]

As alluded to for OC above, tobacco and alcohol use is highly correlated, and it can be difficult to identify independent effects of these risk factors. In a study from the INHANCE consortium, in non-drinkers the OR for HNC for ever versus never smoking was 2.13 (95% CI = 1.52 to 2.98), slightly lower than seen when drinkers and non-drinkers are studied together. Interestingly, in never smokers, the odds of developing HNC was only greater when the frequency of alcohol consumed was high (OR for three or more drinks per day versus never drinking = 2.04, 95% CI = 1.29 to 3.21) with no evidence of increased odds at lower alcohol intake frequencies.

The effect of tobacco and alcohol has been shown to interact in the association with HNCs. This is demonstrated nicely in work by Hashibe and colleagues[5] where they partition the PAR into that which is for alcohol alone, tobacco alone and the portion which is attributed to both risk factors together. For all HNCs, the total PAR is reported as 72.0% (95% CI: 61.2, 79.1) with almost half of this attributable to the synergistic effect of alcohol and tobacco, and less than 40% attributable to

tobacco use alone (PAR=33.0% (95% CI: 42.6, 25.9)) plus alcohol use alone (PAR=4.0% (95% CI: 1.5, 5.3)).

2.2.4.2 Human papilloma virus infection

HPV, the causative infective agent for many cervical and anal cancers, has been strongly linked with oropharyngeal cancers for more than two decades. Research in the early 2000s confirmed it as a causative agent in many OPC cases.[37] HPV16 is the predominant form and is found in the tumour tissue of approximately 36% of oropharyngeal cases,[38] with estimates ranging from 0 to 86%.[39] HPV16 positivity is lower in OC (23.5% (95 CI: 21.9, 25.1)) and LC 24.0% (95% CI: 21.8, 26.3)) tumour tissue where the role of HPV is less clear. There has been a global increase in the incidence of OPC which is predominantly driven by the emergence of HPV.[3]

2.2.4.3 Socio-economic status

Socioeconomic status (SES) is a strong risk factor for head and neck cancer with lower SES being associated with a higher risk. As might be expected, this association was partly but driven by other risk factors[40] but a larger study estimated that approximately one third of the effect of SES was not driven by smoking and alcohol use.[41] More recently, a study found evidence to suggest an interaction between SES and the other major risk factors of smoking and alcohol use, with the effects of these stronger in lower SES groups than in higher SES groups.[42]

2.2.4.4 Age and Sex

HNCs are more common in males than females with the age-standardised incidence rates in males being more than double that in females in the UK (CRUK). This pattern may reflect higher levels of smoking and alcohol use in males than females. The age distribution of HNC does not differ by sex with the peak age of diagnosis being 65 to 69 years in both sexes.

2.2.4.5 Other risk factors

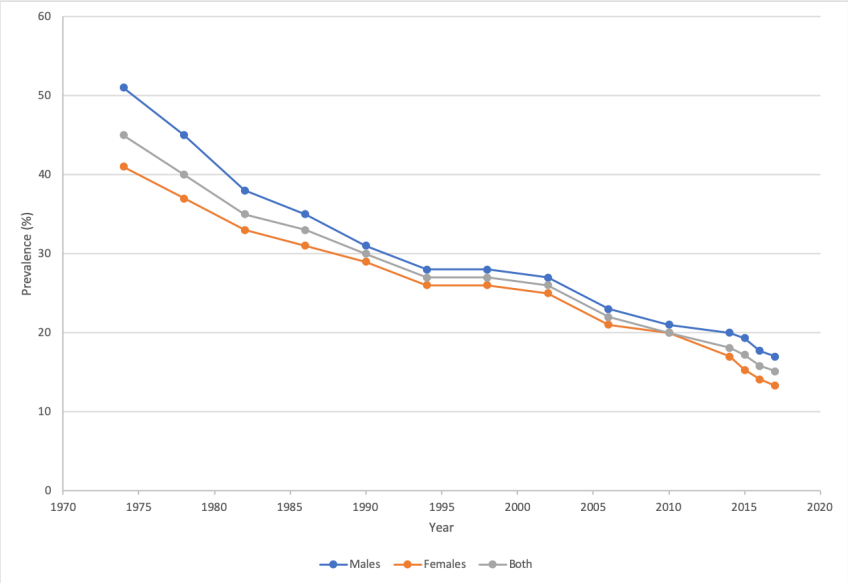
Several other factors have been linked with an increase in the risk of HNC without strong causal links being established. These included a diet low in fresh fruit and vegetables, low BMI, poor oral health, Marijuana use and sexual health.[43]

2.2.5 Changing aetiology

The use of tobacco is generally on the decline across the world. World Health Organisation (WHO) predictions suggest that tobacco use in females is declining for all regions and in males for all regions except Africa and Eastern Mediterranean.[44] This decline is mirrored in the UK with smoking prevalence approximately 15% in 2017, down from 45% in the mid 1970s (Figure 2-3). Alcohol

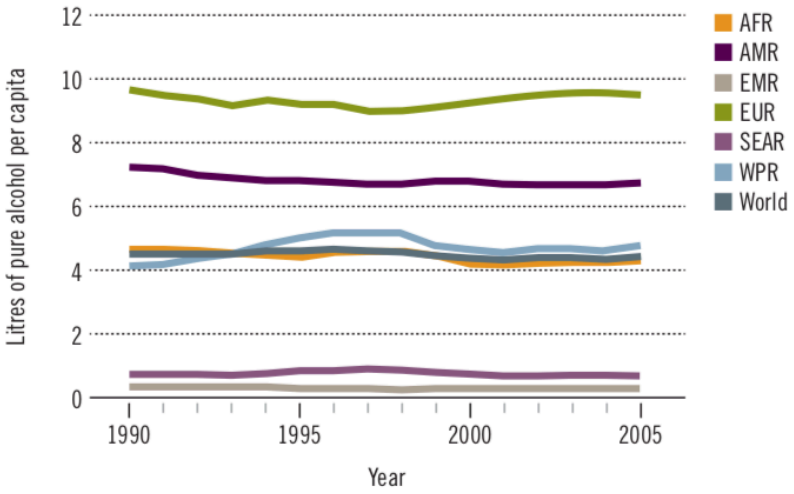
consumption worldwide has been relatively stable[45] (Figure 2-4) and again this is mirrored in the UK (Figure 2-5).

Figure 2-3: UK smoking prevalence, 1970-2017.



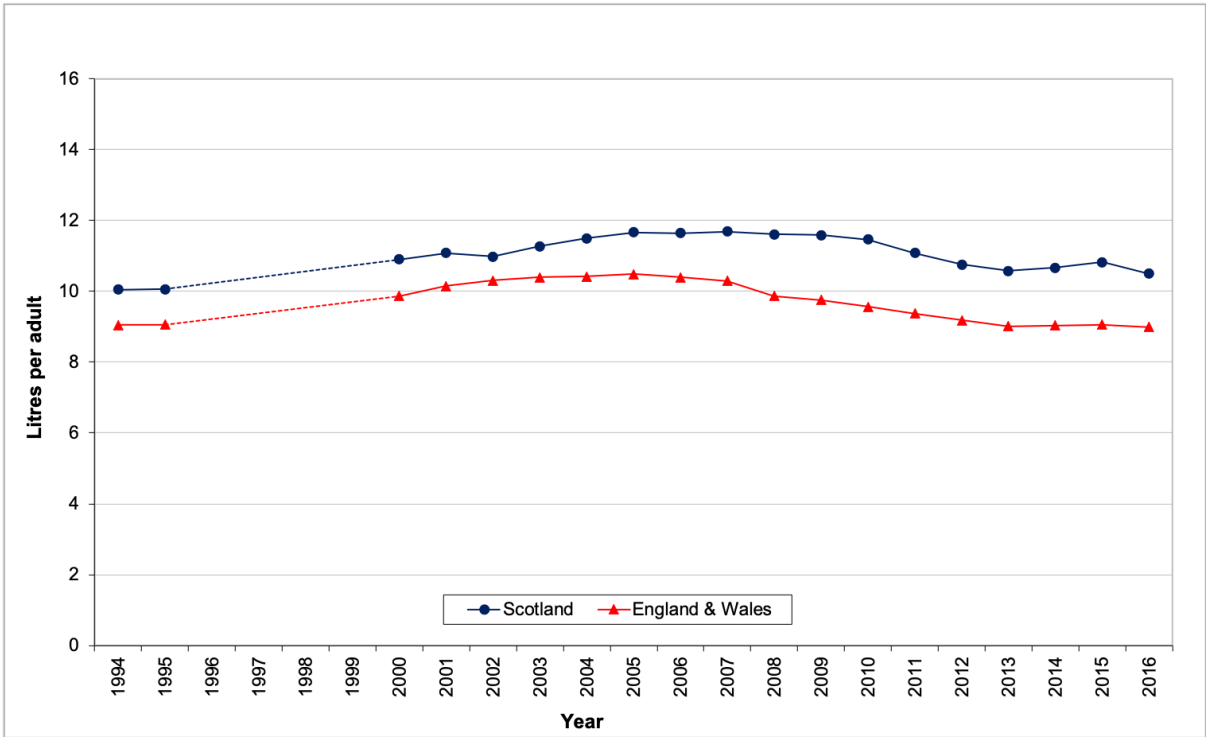
Data from UK Office for national statistics. Adult smoking habits in the UK: 2017 [46]

Figure 2-4: Recorded adult per capita consumption, in litres of pure alcohol, by WHO region and the world, 1990-2005 (Reprinted from[45] with permission of WHO).



AFR= African region, AMR=Region of the Americas, EMR= Eastern Mediterranean region, EUR=European region, SEAR=South-East Asia region, Western Pacific region, WHO=World Heath Organisation

Figure 2-5: Volume of pure alcohol (litres) sold per adult (16+) Scotland and England & Wales, 1994-2016

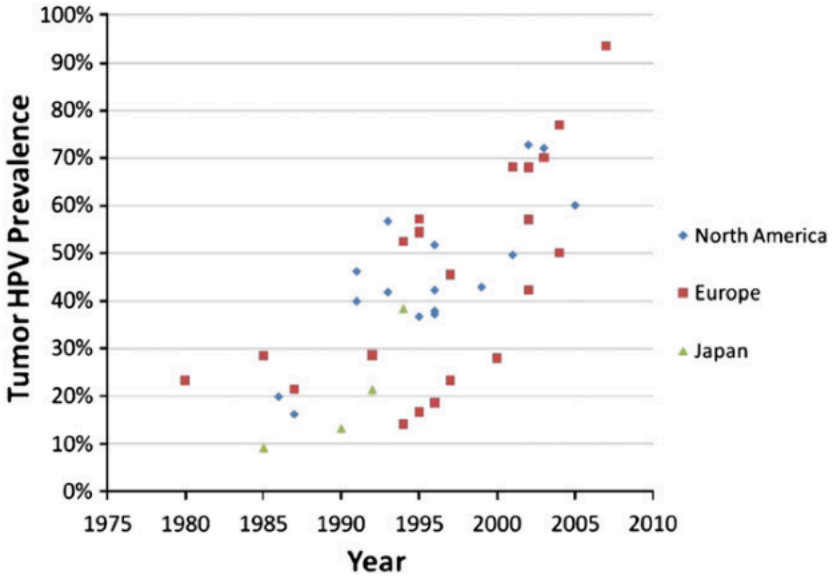


Data from: Monitoring and Evaluating Scotland's Alcohol Strategy: Monitoring Report 2017. http://www.healthscotland.scot/publications/mesas-monitoring-report-2017

Given the burden of HNC is largely related to these two risk factors, particularly tobacco use, the incidence of HNCs would be expected to decline over time, alongside the reduction in tobacco use. This reduction has been demonstrated in Europe for other smoking related cancers such as lung and LC[6] but the incidence of OC and OPC has increased during the same period across Europe[6] and in the UK.[47]

A review by D'Souza and Dempsey[48] demonstrated that tumour HPV positivity is increasing over time (Figure 2-6). This pattern could explain the worrying increase in OPC incidence in developed countries over the past few decades,[29,49,50] despite the decrease in tobacco use. However, within the UK the proportion of HPV related OPC is unchanged and there is also an increase in non-HPV OPC.[17] This suggests that the increase in OPC cases cannot be explained solely by an increase in HPV infection.[17,47] Furthermore, there has been an increase in OC incidence in the UK. Comparing demographic and lifestyle factors in HN5000, a contemporary clinical cohort with previous HNC collections shows a higher proportion of these OCs being in female, non-smoking individuals who do not report engaging in oral sex. This trend which is unlikely to be specific to the UK is concerning and suggests there may be an unknown aetiology causing a proportion of these OC.

Figure 2-6: Percentage of oropharyngeal tumours that were HPV positive in studies from 1975 – 2010 (from[48] with permission from Elsevier)



In light of there being a portion of HNC risk not attributable to tobacco, alcohol or HPV and the recent changes in the profile of the three main HNCs it is important that research is focused on identifying potential novel risk factors.

2.3 Vitamin D

2.3.1 Background and history

Vitamin D is a steroid related compound that has important roles in calcium homeostasis and is therefore important for skeletal health.[51] More recently however it has also been linked to non-skeletal diseases such as cardiovascular disease,[52] autoimmune diseases[53] and cancer.[54]

Vitamin D was identified in the early 20th century principally due to its ability to cure rickets. Sir Edward Mellanby had identified that rickets in dogs could be cured by feeding them cod-liver oil, [55] known to contain vitamin A. Elmer McCollum meanwhile destroyed the vitamin A in cod-liver oil by oxygenation and found that the oil could no longer cure vitamin A deficiency but could still cure rickets.[56] He correctly identified a new vitamin was present in cod-liver oil, which was subsequently named vitamin D. Vitamin D_2, or ergocalciferol, was isolated and its structure identified in 1930[57] with vitamin D_3, or cholecalciferol, identified in 1937 (reviewed in[58]). It was hypothesised that vitamin D_3 was formed by exposure to sunlight but this was not confirmed until 1972.[59] This raised the question of whether vitamin D_3 was a true vitamin in the classical sense as it can be produced by the body in sufficient quantities and therefore behaves more like a hormone than an essential nutrient.

2.3.2 Vitamin D synthesis and metabolism

Vitamin D_3 is predominantly synthesised in the skin on direct exposure to ultra-violet B (UVB) radiation. 7-dehydrocholesterol (7-DHC) is photoisomerised to pre-vitamin D on absorption of UVB radiation of wavelengths between 280 and 315 nm.[60] This pre-vitamin is then converted into vitamin D_3 via a heat isomerisation reaction or further photoisomerised back to 7-DHC or to inactive metabolites (lumisterol and tachysterol).[60] Availability of 7-DHC in the skin is determined by the activity of 7-dehydrocholesterol reductase (DHCR7) that converts 7-DHC into cholesterol. Increased activity of the DHCR7 enzyme reduces the availability of the 7-DHC substrate for the initial photoisomerization reaction to form pre-vitamin D.

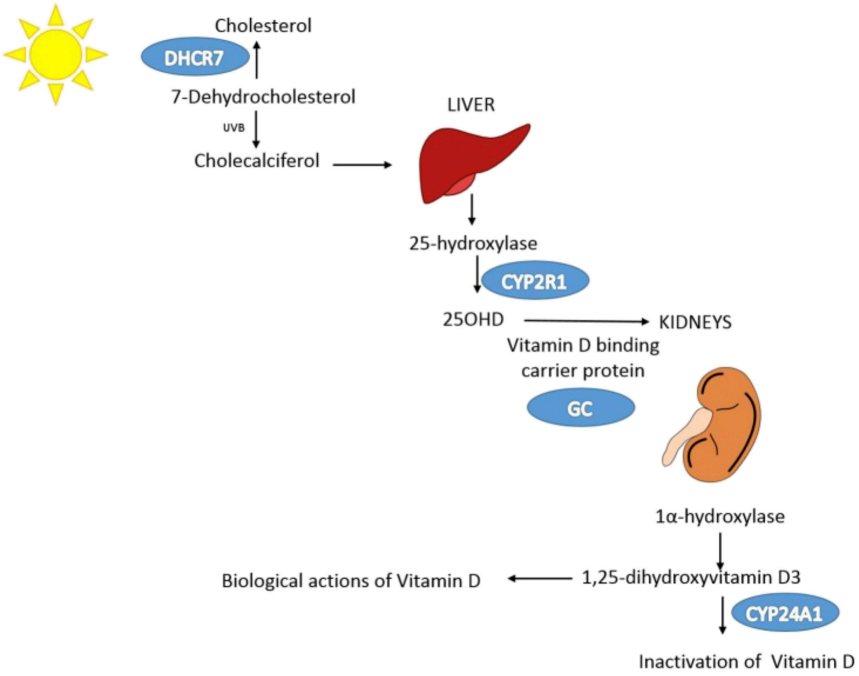
Vitamin D can be absorbed from the diet. Vitamin D_2 is absorbed from plants or artificially fortified foods and vitamin D_3 from oily fish. Supplementation with vitamin D is common and can be in either of the two forms, concentrations vary dramatically across different formulations and there is no consensus on the ideal form or concentration.

Vitamin D_2 and D_3 are lipid soluble and stored in adipose tissue[61] whereas the circulating form is 25-hydroxyvitamin D (25(OH)D). This metabolite is formed in the liver by hydroxylation of either vitamin D_2 or D_3 by the cytochrome P450 family 2 subfamily R member 1 (CYP2R1) enzyme, also known as vitamin D 25-hydroxylase. 25(OH)D is transported in the blood predominantly bound to the vitamin D binding protein (DBP) (80-90%) or albumin (10-20%) with a tiny fraction unbound, which is referred to as free 25(OH)D.[62]

The active form of vitamin D, 1,25-di-hydroxyvitamin D ($1,25(OH)_2D$), also known as calcitriol, is formed by a second hydroxylation step. 25(OH)D is hydroxylated by the cytochrome P450 family 27 subfamily B member 1 (CYP27B1) enzyme, also known as 25-hydroxyvitamin D 1-alpha hydroxylase, to $1,25(OH)_2D$ predominantly in the kidneys. 25(OH)D is transported into the renal cells attached to DBP via a megalin-mediated exocytosis. Extrarenal conversion has also been reported in skin, placenta, and lymph nodes with some activity also likely in adrenal medulla, brain, pancreas, and colon.[63] Like the inactive form, $1,25(OH)_2D$ is predominantly transported to where it is required in the circulation, bound to DBP. At these tissues $1,25(OH)_2D$ exerts its effects by binding the vitamin D receptor (VDR) intracellularly.

Breakdown of $1,25(OH)_2D$ occurs in the kidney and target tissues where cytochrome P450 family 24 subfamily A member 1 (CYP24A1) is expressed.[64] This enzyme catalyses the hydroxylation of both $1,25(OH)_2D$ and 25(OH)D at an additional carbon, a step which marks the metabolites for further breakdown and excretion predominantly through bile in the faeces.[65]

Figure 2-7: Diagrammatic representation of the synthesis and degradation of vitamin D detailing the important enzymes involved in the process.



2.3.3 Measurement of vitamin D

Vitamin D is measured using an assay of 25(OH)D rather than a measure of the active metabolite, $1,25(OH)_2D$. Although it may seem that $1,25(OH)_2D$ concentration would provide a better measure of vitamin D status this is not the case. $1,25(OH)_2D$ is buffered by an increase in parathyroid hormone in vitamin D deficient patients. This would lead to near-normal $1,25(OH)_2D$ assay results even in deficient patients. Furthermore, the half-life of $1,25(OH)_2D$ is four to six hours and is therefore highly susceptible to short term fluctuations whereas the half-life of 25(OH)D is several weeks. Finally, the concentration of the active vitamin D form is 1000-fold lower than its inactive precursor, making the practicality of its measurement more prone to error.[66]

The 25(OH)D assay measures the total 25(OH)D in circulation, this includes the bound component (to either DBP or albumin) and the free component. There is controversy within the literature as to whether the total or free component of 25(OH)D better reflects vitamin D status.[67,68] There is strong evidence that the levels are highly correlated, suggesting 25(OH)D is an adequate measure. However, the same study also shows that in vitamin D deficient individuals, the proportion of total 25(OH)D which is made up of free 25(OH)D, increases at low 25(OH)D levels.[69] This may have implications for studies investigating vitamin D deficient populations.

As discussed in detail below, UVB exposure and hence season is a major factor in vitamin D levels. It is important that the season of blood draw is taken into account when measuring vitamin D across populations over a period of time. Several adjustment methods have been utilised in

epidemiological studies, including making a correction of the 25(OH)D measure by season (summer, autumn, winter, spring) or modelling the seasonal variation in a sinusoid pattern.

2.3.4 Optimal vitamin D levels

Clinically, and in epidemiological studies, vitamin D is measured using an assay for 25(OH)D. This inactive metabolite has a half-life of approximately three weeks and is therefore relatively stable to daily diet and sunlight exposure variation. The active vitamin D metabolite 1,25(OH)D_2 has a half-life of approximately four hours and is therefore a poor long-term indicator of vitamin D levels.

It has been proposed that 25(OH)D levels in excess of 50 nmol/L are adequate for bone health[65] with this level being a minimum at the end of winter rather than a year-round average.[70] One group has suggested higher thresholds (>80 nmol/L) based on the levels of specific biomarkers that vitamin D effects, such as intact parathyroid hormone, calcium absorption and bone mineral density. These guidelines mainly relate to bone health however and there is little robust evidence to suggest levels which may be sufficient for the non-skeletal effects of vitamin D.

2.3.5 Factors that affect vitamin D levels

Many factors affect vitamin D levels. The main contributor to serum vitamin D levels is exposure to UVB radiation, although UVB radiation is not a risk factor for HNC (other than the external lip which in not included in this thesis) understanding the production of vitamin D in the body is important to several aspects of this thesis. One chapter of this thesis involves identifying genetic variants as proxy measures of vitamin D and applying them in causal analyses therefore the genetic factors that influence vitamin D are discussed. Additionally, the factors that are important to both vitamin D levels and HNC are summarised as these will be the main factors that may impact analyses investigating the association between vitamin D and HNC risk and prognosis. The impact could either be as a confounder, where the factor is a common cause of both decreased vitamin D and increase HNC incidence or an effect modifier, where vitamin D has an effect on HNC but it is dependent on the factor.

2.3.5.1 Exposure to ultraviolet B radiation

Given the initiation of vitamin D synthesis requires UVB photons to collide with 7-DHC, it is unsurprising that vitamin D levels are strongly influenced by UVB exposure. Factors that can affect this exposure can be external or internal to the individual.

External factors

The main external factors that control the ambient UVB are highly predictable as they affect the angle at which sunlight reaches the Earth's surface, or solar zenith angle (SZA). These include

latitude, season and time-of-day. Ambient UVB decreases as the SZA increases as this corresponds to an increased path distance through the atmosphere and an increased surface area of absorption. Small SZA, and therefore high UVB radiation, corresponds to low latitudes (closer to the equator), summer and when the sun is highest in the sky (noon). This factor creates a link between geographic location and vitamin D which can be important when considering studies of vitamin D conducted around the world.

Some external factors which are less predictable involve influences which absorb UVB radiation before they can reach the Earth's surface. More cloud cover, thicker ozone and higher pollution levels all absorb more UVB radiation and ultimately decrease conversion of 7-DHC to vitamin D.[60]

Internal factors

Internal factors relate to those pertinent to the individual with the major factor being skin type. This genetically determined influence affects the amount of vitamin D that can be synthesised for a given amount of UVB radiation. Melanin in the skin absorbs UVB radiation as a protective mechanism against UVB damage but also prevents the conversion of 7-DHC to pre-vitamin D_3. Evolutionarily, dark-skinned races have greater melanin due to originating at lower latitudes where UVB radiation is high. Pale skin, that maximises the use of available UVB radiation, was an evolutionary advantage at latitudes where risk of damage from UVB radiation is low. Migration of dark-skinned individuals to higher latitudes puts these populations at risk of sub-optimal vitamin D levels without the use of supplementation or targeted diets.[60]

Other internal factors that reduce the exposure of an individual to UVB radiation are skin coverage with clothing or sunscreen which absorb UVB radiation and reduce the amount available for vitamin D synthesis. [60] Cultural and behavioural patterns will also affect UVB exposure, in cultures where it is the norm to cover-up exposed skin there will be less surface area for UVB absorption. [60] Jobs and past times that involve being outside will increase the potential for UVB exposure whereas sedentary indoor jobs and reduced levels of physical activity will have the opposite effect. The recent increase in obesity has been partly blamed on an increase in sedentary occupations and more time spend indoors (watching television and being on the computer). There is the potential that these lifestyle changes are also having vitamin D lowering impacts. Good quality trend data for vitamin D in not available in the UK however there has been a dramatic increase in the number of tests for low vitamin D levels in the NHS.[71] This may suggest a higher incidence of low vitamin D levels but may also reflect a greater awareness amongst clinicians to the potential health consequences of low vitamin D.

2.3.5.2 Genetic factors

There have been many studies which have investigated and demonstrated a genetic component of vitamin D levels. These studies are of varied design and are summarised below:

Family and twin studies

Vitamin D levels have been shown to be heritable, but estimates vary dramatically. 25(OH)D heritability estimates in Caucasian populations have been reported between 0.22 and 0.86.[72–78] The variation in these estimates may be down to the changeable environmental factors such as season, Karohl e*t al.*[78] estimated the heritability of 25(OH)D to be 0.7 (95% CI: 0.31, 0.80) in winter and 0.00 (95% CI: 0.00, 0.35) in summer. There are very few studies estimating heritability in non-Caucasian populations but studies investigating Asian[79] and Hispanic[80] populations show similar estimates to Caucasian populations.

Linkage studies

Several studies have attempted to identify regions of the genome that are associated with 25(OH)D using linkage, none of these have identified regions with robust results that pass multiple testing thresholds.[74,77]

Candidate gene studies

Candidate gene studies look for associations between single nucleotide polymorphisms (SNP), single base-pair changes in DNA, with 25(OH)D levels. These studies do have limitations however:

1. Phenotype characterisation is often heterogeneous across studies which limits the ability to replicate findings
2. Sample sizes are often small
3. Multiple testing is not always adequately dealt with which can lead to type II errors.

The SNPs under investigation are selected using prior knowledge of the vitamin D related pathways (see 2.3.2 Vitamin D synthesis and metabolism). An overview of results is provided below.

CYP27B1

Cytochrome P450 family 27 subfamily B member 1 (*CYP27B1*) is found on chromosome 12 and encodes the 25(OH)D 1-alpha hydroxylase protein which catalyses the conversion of 25(OH)D to 1,25$(OH)_2$D. The C allele of rs10877012 has been associated with lower 25(OH)D in three studies[81–83] but it is unclear how this SNP exerts effects. Two other SNPs in this gene have also been associated with 25(OH)D in candidate gene studies.[75]

CYP2R1

Cytochrome P450 family 2 subfamily R member 1 (*CYP2R1*) encodes vitamin D 25-hydroxylase which converts vitamin D into 25(OH)D. This gene, found on chromosome 11 was first considered important in vitamin D metabolism when a mutation on exon 2 was identified to lead to vitamin D deficiency.[84] Later rs10741657,[85–87] rs12794714[86,87] and rs10766197[74,86] have been shown to have associations with 25(OH)D.

GC

The DBP, a group specific component is encoded by the *GC* gene on chromosome four and has been shown to be associated with 25(OH)D levels in gene level association studies.[88] Two SNPs have been consistently associated with 25(OH)D levels in numerous studies; rs7041 encodes a Glutamine to Aspartate change at codon 416[80,87–91] and rs4588 encodes a Threonine to Lysine amino acid change at codon 426.[80,87,89,90,92–94] A third SNP, rs2282679 in *GC*, is of interest as it shows consistent effects on 25(OH)D across American,[91] African American,[83] British[87] and Han Chinese populations.[90]

VDR

Vitamin D receptor protein is encoded by the Vitamin D receptor gene (*VDR*) on chromosome 12. This protein is a nuclear receptor transcription factor that binds $1,25(OH)_2D$ to exert its effects. rs2228570 (a mutation also referred to as *FokI*) increases the amino acid structure of this protein by three amino acids and influences its activity. This SNP has also been associated with increased 25(OH)D levels.[75,95] The C-G haplotype of rs7139166 and rs54516035 has been shown to increase the VDR protein activity in one study.[96]

CYP24A1

The cytochrome P450 family 24 subfamily A member 1 (*CYP24A1*) gene on chromosome 20, encodes an enzyme involved in the breakdown of the 25(OH)D and $1,25(OH)_2D$. rs17219315 in *CYP24A1* has been associated with 25(OH)D levels in a single study.[74]

Genome-wide association studies

Genome-wide association studies (GWAS) look across the genome in an unbiased manner to identify SNPs which have associations with the trait of interest, in this case 25(OH)D. Due to the number of SNPs tested they require large sample sizes to overcome the multiple testing burden. A more detailed description of GWAS is given in Chapter 3 and more recent findings from GWAS, including results from this thesis are discussed in Chapter 5.

The first 25(OH)D GWAS had limited coverage of the genome (<80,000 SNPs) and was conducted in a small sample, it is unsurprising then that no robust associations were identified.[97] Two larger GWAS

with discovery and replication sets have been more informative. Ahn *et al.* investigated 4,501 as a discovery and 2,221 in a replication set[98] whereas the largest GWAS (preceding this thesis) was a meta-GWAS of 15 cohorts totalling 33,996 across discovery and replications sets by the SUNLIGHT consortium (Study of Underlying Genetic Determinants of Vitamin D and Highly Related Traits). SNPs in *GC and CYP2R1* were identified as having statistical evidence of association beyond the multiple testing threshold in both studies, corroborating the candidate gene findings. A SNP near the 7-dehydrocholesterol reductase gene (*DHCR7*) was also identified in both studies. This gene is involved in the early stages of production of vitamin D in the skin from UV light. Additionally, the SUNLIGHT consortium confirmed the link with *CYP24A1* from a single candidate study[74] by identifying an association in the same locus. One limitation of this latter GWAS however was the authors did not adjust for ancestry informative principal components. This may have resulted in false positives as described in more detail in Chapter 3.

2.3.5.3 Age

Vitamin D levels have been shown to decrease with age.[99] This is likely to be partly due to changes in behaviour but levels of 7-DHC, the substrate for vitamin D synthesis, are also depleted in the skin of the elderly.[100]

2.3.5.4 Body mass index

It has been observed for some time that there is an inverse correlation between vitamin D and body mass index and other similar anthropometric traits.[101] More recently, this inverse association has been demonstrated for 25(OH)D with total body fat in women,[102] for 25(OH)D and $1,25(OH)_2D$ with BMI and total fat mass[103] and for 25(OH)D and $1,25(OH)_2D$ with BMI.[104]

Despite the consistent correlation being reported, the direction of causality has been difficult to determine. A Mendelian randomization study that assessed the association in both directions reported a robust causal vitamin D lowering effect of BMI but, if the reverse causal association exists, it is likely to be of a very small magnitude.[105]

2.3.5.5 Smoking

The association between vitamin D and smoking has been the subject of several studies, most of which report lower 25(OH)D levels in smokers compared to non-smokers in a range of populations.[106–109] In the most recent of these studies conducted in a Chinese cohort, there was a dose response relationship with increased number of cigarettes smoked per day being associated with progressively lower 25(OH)D.[108] Contrary to these reported associations, one study has found no evidence of association between smoking and 25(OH)D levels,[110] and in another study the

association was seen in males but not females.[111] One study examined the effect of the assay used to measure 25(OH)D, and found that some commonly used assays over-report 25(OH)D in smokers compared to non-smokers.[112] This may mean the magnitude of the association between smoking and 25(OH)D has been under-estimated. Despite the majority of studies demonstrating a vitamin D lowering effect of smoking even after correcting for likely confounders, the true causal link is not well understood.

2.3.5.6 Alcohol use

Given the correlation between alcohol and smoking, an association between vitamin D and alcohol would not be surprising. A systematic review reported that out of 49 papers with effect estimates of the association between alcohol use and vitamin D, 30.7% showed higher 25(OH)D levels with higher alcohol consumption, 36.7% showed the opposite direction of effect and 32.7% showed no evidence of an association.[113] This review also commented that over 70% of the studies were of cross-sectional design and therefore, even when associations were present, effect direction was difficult to determine. A further study which was not captured by the search terms of the above systematic review showed no evidence of an effect in a Greek population.[109]

2.3.6 Vitamin D supplementation

Guidelines from several sources exist which recommend vitamin D supplementation of the general population.[114] There is a range of suggested doses from 600 to 2000 International Units (IU) per day, with a maximum recommended dose of 4000 IU per day. The stated aims of the guidelines differ with some focussing on reducing skeletal problems, some on reducing vitamin D deficiency and others suggesting general health as the reason for supplementation. These recommendations are reflected in the global vitamin D market which is expected to be worth $2.5 billion by 2020 with an annual growth of approximately 11%.[115]

In the US increased vitamin D supplementation has been linked to a modest increase in vitamin D levels, as measured by 25(OH)D.[116] The increase was greatest in females, especially those known to use vitamin D supplements, and it is unlikely that this section of the population is of greatest risk of vitamin D deficiency. There is limited evidence of vitamin D levels increasing elsewhere across the globe. In contrast there are reports of higher incidences of vitamin D deficiency in the UK[117] and rickets in Australia[118] which may suggest the opposite trend.

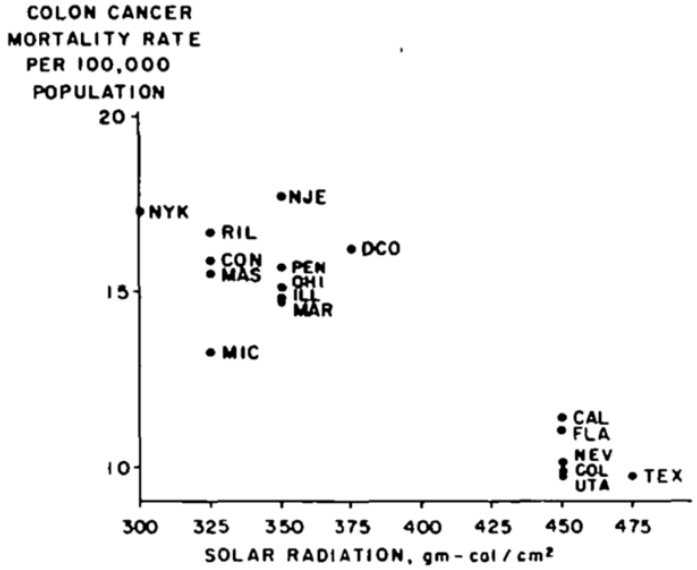
2.4 The association between vitamin D and head and neck cancer

2.4.1 Vitamin D and cancer

The link between vitamin D and cancer dates back nearly 80 years. An IARC review provides a detailed breakdown of the evidence assessing the association of 25(OH)D with the most common cancers.[9] The key findings are summarised below.

The first evidence linking vitamin D to cancer arose from ecological evidence of UVB exposure and cancer incidence and mortality. The hypothesis that low vitamin D increases cancer risk assumes that the greater the distance from the equator the lower the UVB exposure, and therefore the lower the population's vitamin D level. In 1941 Apperly looked at UVB exposure and cancer mortality across US states and Canada.[119] This work reported a positive correlation between index of radiation (a measure of UVB exposure) and skin cancer mortality, but an inverse association with all cancer mortality. Forty years later Garland and Garland reported the same inverse association for colorectal cancer across US states (Figure 2-8).[120]

Figure 2-8: The association of solar radiation with colon cancer mortality rate in US states from 1959 -1961. (from [120] by permission of Oxford University Press)

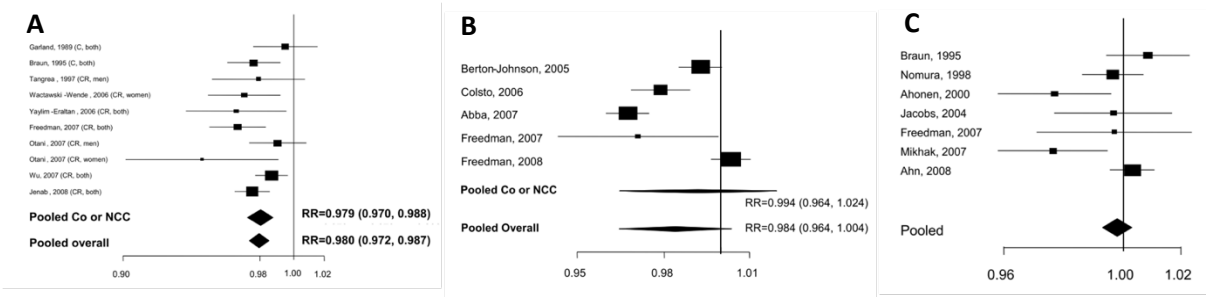


Since the 1980s approximately 20 studies have investigated the ecological link between UVB exposure and cancer with strong inverse associations identified for colorectal, prostate and breast cancers (see IARC review[9]). There is evidence for an association in the same direction between UVB exposure and non-Hodgkin's lymphoma, ovarian, cervical and uterine cancers but the evidence is weak. These studies are all however subject to the limitations of their study design including the ecological fallacy, confounding and misclassification which, unlike in individual level data, in ecological studies is more likely to bias estimates away from the null.

Participant level observational studies of sun exposure and cancer incidence also show evidence for a protective effect of increased UVB exposure on breast, colon and prostate cancers (see IARC review[9]). Studies of dietary vitamin D exposure of this type are less conclusive with weak evidence for a protective effect of dietary vitamin D on colorectal cancer but limited evidence for other cancers.[9]

Many observational studies have investigated the association of serum vitamin D levels, usually 25(OH)D, and cancer. The IARC review has meta-analysed these findings where possible and shows a strong protective effect of 25(OH)D on colorectal cancer incidence but less strong effects on breast and prostate cancers (Figure 2-9).

Figure 2-9: Meta-analysis of studies assessing association of 25(OH)D on (A) colorectal (B) breast and (B) prostate cancer incidence (Reprinted from[9] with permission of WHO).



2.4.2 Vitamin D and head and neck cancer

There are no published studies comparing the population trends of vitamin D levels to that of HNC incidence. However, as discussed in the previous sections on HNC and vitamin D, there is a steady increase in HNC incidence (despite decreasing smoking prevalence) and the potential for decreasing vitamin D levels, due to changes in lifestyle meaning individuals spend much less time outdoors. This provides a potential link between vitamin D and HNC which may be able to partly explain the recent changes in incidence. In the section below, the evidence linking these two factors is discussed.

2.4.2.1 Observational epidemiology for HNC risk

The association between vitamin D and cancers of the head and neck is less well studied than the association with colorectal, breast and prostate cancer. This is partly due to it being a relatively rare cancer, particularly when the sub-sites are considered independently, limiting power even in large cohort studies. Below, the epidemiological studies that have assessed the association of vitamin D and HNC are discussed. The study results are combined into a table but it should be made clear that the studies differ in terms of their design, vitamin D measure, vitamin D scale and summary effect

measure. For this reason, the studies should not be directly compared, the aim of the table is to give an overall representation of the observational evidence to date (Table 2-2).

The first interest in the association with HNC was an unexpected finding. Giovannucci *et al*. attempted to improve on the ecological studies that had preceded it, by using a prediction model of six 25(OH)D determinants, to estimate individual 25(OH)D levels in over 47,000 men in the Health Professionals Follow-Up Study, a prospective cohort study. The model, which explained 28% of the variability in 25(OH)D was trained in 1,095 men with the six determinants and serum 25(OH)D measured. The association between the predicted 25(OH)D measure and a variety of cancers was assessed, including 51 oral or pharyngeal cancer cases. Despite expecting to find associations for the digestive-system related cancers, the strongest effect was seen for the oral and pharyngeal cancers combined. After correcting for the common HNC risk factors of smoking and alcohol use, they identified a 70% (95% CI: 19% to 89%) reduction in risk of oral or pharyngeal cancer per 25 nmol/L increase in estimated vitamin D.[121]

A later case-control study of 804 oral and pharyngeal cancer cases assessed dietary intake, recruiting cancer cases at diagnosis and using hospital-based controls. Dietary questionnaire data was used to estimate vitamin D intake. This estimated vitamin D intake was then split into tertiles. For oral and pharyngeal cancer combined, there was a 24% decrease (95% CI: 6% to 40%) in the odds of developing cancer in the highest tertile of vitamin D intake compared to the lowest.[122] After stratifying by smoking and drinking the effect of vitamin D intake was only seen in the highest smoking and drinking strata. Estimates were stratified by HNC site, but no obvious differences were seen between oral cavity and pharynx cancers.

Subsequent studies have measured serum 25(OH)D concentration, a reliable biomarker of vitamin D from all sources rather than a predicted measure from diet, sunlight exposure or a combination of these and other determinants. A case control study investigating the effect of 25(OH)D concentration on the risk of oral cavity, pharynx and larynx cancers was nested in the Alpha-Tocopherol Beta-Carotene Cancer (ATBC) Study, a randomised, double blind, placebo-controlled primary cancer prevention trial in Finland. Cases and controls, all males smoking more than five cigarettes per day, were selected from the trial participants matched for age at randomization and date of baseline serum 25(OH)D measure. 25(OH)D was measured at the pre-randomization phase. In this study there was little evidence to support an association between 25(OH)D and the three cancer sites combined or individually, although estimates were imprecise with confidence intervals including relatively large effects in both directions.[123]

Afzal *et al*.[124] conducted a prospective study of 9,791 participants with baseline serum 25(OH)D measures, to investigate the association between serum 25(OH)D concentration and cancer. Although they do report results for all cancers combined, their particular focus was tobacco related cancers including head and neck cancer (n=122), lung cancer (n=507), bladder cancer (n=112), pancreatic cancer (n=109), stomach cancer (n=64), kidney cancer (n=55), liver cancer (n=55) and other tobacco-related cancers (n=96). The HR for all cancers, for a 50% reduction in 25(OH)D, was 1.06 (95% CI: 1.02 to 1.11). This effect was driven by the tobacco-related cancers only (HR=1.20 (95% CI: 1.13 to 1.28)) with very little heterogeneity seen between these cancers (I^2=0%, *P* for heterogeneity = 0.48). Additionally, a clear dose response relationship was seen across 25(OH)D strata for the tobacco-related cancers. Similar to the study by Giovannucci *et al.*,[121] the strongest association in the tobacco-related cancers was seen for HNC (HR=1.44 (95% CI: 1.19 to 1.73)).

In a Danish prospective cohort study of 12,204 participants with serum 25(OH)D measures at baseline the association between serum 25(OH)D and several cancers was assessed including 44 HNC cases. There was some weak evidence to support an association between serum 25(OH)D and HNC (HR per 10nmol/L increase in serum 25(OH)D = 0.88 (95% CI: 0.77 to 1.02)). However, this effect was almost completely attenuated after correcting for several confounders including smoking (HR = 0.97 (95% CI: 0.84 to 1.12)).[125]

The largest and most recent study examining the association between vitamin D and HNC was conducted in the European Prospective Investigation into Cancer and Nutrition (EPIC).[126] This nested case-control study included 250 HNCs and up to 940 controls, with baseline serum 25(OH)D measures. It found a strong protective effect of 25(OH)D, with a doubling of 25(OH)D associated with a 46% reduction in the odds of HNC (95% CI: 32% to 56%). This effect is very similar to the effect reported in the study by Afzal *et al*.[124] ($OR_{1/2}$=1.44) when the effect in the EPIC study is re-scaled for a halving of 25(OH)D ($OR_{1/2}$=1.45). Correction for alcohol consumption and tobacco use, including cotinine in current smokers, led to some attenuation of the effect but this was still of a clinically relevant magnitude (reduction of odds = 30% (95% CI: 44% to 12%)). When stratified by HNC site, the strongest fully adjusted estimate was seen for larynx and hypopharynx cancers (OR = 0.55 (95% CI: 0.39 to 0.78)), then oral cavity cancers (OR = 0.60 (95% CI: 0.42 to 0.87)) whereas there was little evidence for an effect in oropharynx cancers (OR = 0.92 (95% CI: 0.58 to 1.45)). Given the heterogeneity across HNC sites discussed above, it is not surprising that these different effect sizes exist, and they may reflect real differences in causal effects of vitamin D or different confounding structure across sites.

Table 2-2: Effect of vitamin D on HNC risk from observational studies.

First Author	Study Design	Vitamin D measure	Effect scale	Effect size	95% CI
Giovannucci	Cohort	Predicted 25(OH)D	25 nmol/L increase	RR = 0.3	(0.11,0.81)
Lipworth	Case control	Dietary intake	3rd vs 1st tertile	OR = 0.76	(0.60,0.94)
Arem	Case control	25(OH)D	4th vs 1st quartile	OR = 0.88	(0.56,1.39)
Afzal	Cohort	25(OH)D	Doubling	HR = 0.69	(0.58,0.84)
Skaaby	Cohort	25(OH)D	10 nmol/L increase	HR = 0.88	(0.77,1.02)
Fanidi	Case Cohort	25(OH)D	Doubling	OR = 0.54	(0.44,0.68)

These studies vary in terms of design, vitamin D measure, vitamin D scale and summary effect measure. They should not be directly compared but the figure should give a feel for the overall observational evidence to date.

Although these studies are difficult to compare directly, they all would be consistent with an increase in vitamin D reducing the risk of HNC (although some would also be consistent with a null or opposite association). Of these studies, three were of case-control design[122,123,126] and so would usually be subject to recall bias. However, the Fanidi *et al*.[126] and the Arem *et al.*[123] studies were nested within an RCT and a prospective cohort respectively. As a result, the measurement of the exposures can be considered as robust as those from a cohort study design (discussed below). The Lipworth *et al.*[122] study however, used a recall diet diary at the time of diagnosis to assess vitamin D levels for the previous two years. The recall of this information will certainly be subject to random error but may also be affected by recall bias as, the recent diagnosis of HNC may affect how a patient recalled their recent dietary intake, potentially over-estimating the effect of dietary vitamin D.

Three of the studies reported above are prospective in design with 25(OH)D measures at baseline. In terms of the measurement of exposures, the nested case-control (often referred to as case-cohort) studies can be considered alongside these. In most settings, prospectively measured 25(OH)D should be unbiased with respect to the outcome however this may not be the case here. HNC cancer has a very long latency, meaning its developing may be subclinical for many years or even decades before diagnosis. Even with the assumption that 25(OH)D casually effects HNC, it is unclear at what timepoint 25(OH)D levels are important in the development of HNC but it is likely to be years to decades rather than months to years from diagnosis. This has two implications on the results of the prospective cohort studies and the nested case-control studies. Firstly, despite serum 25(OH)D being a good measure of recent vitamin D levels, it is unlikely that the 25(OH)D measurement in these studies has been taken at the moist critical time and may not reflect the long-term 25(OH)D level of

the individual. This would result in reduced power to detect 25(OH)D effects on HNC. Secondly, in studies where follow-up is relatively short such as in the Fanidi *et al.* (6.3 years) and Arem *et al.* (~8 years) studies, there is a chance that the (as yet undetected) subclinical HNC is influencing the 25(OH)D levels. This would result in reverse causality and overestimate the effect of vitamin D on HNC risk. As the results from the studies above are not directly comparable, it is difficult to interpret what the effect of longer follow up has on the study results but the longest follow-up is 27 years in the Afzal *et al.* study which does report strong evidence of a protective effect of 25(OH)D on HNC risk.

A major limitation of both case-control and cohort studies is the inability to fully account for confounding. In both study designs, multivariable regression methods are used to correct for measured confounders but little can be done to account for unmeasured confounders or confounders which are measured with error or bias as is likely to be the case in retrospective studies. This concept is discussed further in the section below using tobacco use as an example.

2.4.2.2 Tobacco use and the association between vitamin D and HNC risk

Many of the observational studies reported above suggest that tobacco plays a role in the association between vitamin D and HNC risk. In all studies that report adjusted and unadjusted estimates, there is some attenuation of the effect after correcting for potential confounders, including smoking.[121,122,125,126] Additionally, where studies identified an association between vitamin D and HNC, the effect of vitamin D was seen predominantly in the current or former smokers.[122,124–126]

The study by Afzal *et al*. focussed on tobacco-related cancers, but the study was conducted in both ever and never smokers. As mentioned above, the effect of 25(OH)D on tobacco-related cancers was attenuated in the never smokers, although the estimate was imprecise, given the relatively small proportion (20%) of never smokers developing a tobacco-related cancer. Interestingly, within the smokers, there was minimal variation in the effect size across strata of current tobacco consumption (reported as grams/day) or pack-years. The authors reported that they consider tobacco use to be a confounder of the 25(OH)D – cancer association rather than an effect modifier however, the power to provide clear evidence of this interaction would have been limited in this study.

In the EPIC study,[126] the authors point out that residual confounding by tobacco exposure was identified as a concern. In sensitivity analyses a correction was made for additional indicators of tobacco exposure in the current and former smoker strata. These adjustments had little effect on the associations leading the authors to conclude that they did not feel residual confounding was able to explain the protective association identified in this study.

Smoking has been shown to reduce 25(OH)D levels (see section 2.3.5.5) and is the largest known risk factor for HNC, particularly for oral, laryngeal and non-HPV oropharyngeal cancers. There are two ways in which smoking may influence the effect of vitamin D on HNCs. Firstly, smoking is a known confounder of the association as shown by the attenuation of effects when adjustments are made for it. Residual confounding may partly explain any remaining effect of vitamin D on HNC. Secondly, smoking may act as an effect modifier of the vitamin D effect. In this scenario, the protective effect of vitamin D is causal but is only seen in smokers – having no effect in non-smokers. This theory is partly supported by several of the studies above which identify an effect of vitamin D only in smokers and previous smokers.

2.4.2.3 Randomised controlled trials

Double blinded randomised controlled trials (RCT) are the gold standard study design when assessing for causality between an intervention and an outcome. To date there have been no HNC specific RCTs to assess the effect of vitamin D supplementation. Four large RCTs which have cancer incidence as a primary or secondary outcome have been conducted. The first of those, the UK Trial for Prevention of Osteoporotic Fracture, assessed cancer incidence as a secondary outcome but found little evidence of a protective effect of vitamin D supplementation (age adjusted RR = 1.11 (95% CI: 0.86 to 1.42)). The Women's Health Initiative Trial assessed colorectal cancer incidence as a primary endpoint, finding no evidence of an association. They also reported a near null association with all cancer incidence (RR = 0.98 (95% CI: 0.85 to 1.09). The Nebraska Trial, much like the UK study, assessed osteoporotic fracture as a primary endpoint but also reported a protective effect of vitamin D on all-cancer incidence. The results of this trial have been much criticised[127–129] with claims of failed randomization, incorrect choice of statistical tests and not reporting key details that would be considered compulsory by the CONSORT guidelines. Re-analysis in the IARC report suggests the vitamin D effect would be near null using the correct methodology.[9] More recently the very large VITAL study randomised over 25,000 participants over the age of 50 years of age in a two-by-two factorial randomised double blind placebo controlled trial assessing vitamin D and omega 3 fatty acids. The primary outcome was primary cancer prevention and the 2000 IU per day vitamin D supplement was associated with a near null effect after the five year follow up (RR=0.96 (95% CI: 0.88 to 1.06)).

The VITAL trial was the largest of its type but even at this size we would expect only about 35 cases of HNC over the course of the trial. Assuming the incidence of HNC is approximately 20 per 100,000 people, even if the effect estimate is as large as that reported in the EPIC study above, a vitamin D trial would need approximately 300,000 participants in each arm to detect an effect. Given the VITAL

study cost approximately $22 million a trial assessing the effect of vitamin D and HNC using a RCT is not feasible.

2.4.2.4 Vitamin D and head and neck cancer prognosis

When considering epidemiological evidence there is a need to consider factors that affect the risk of cancer, often looked at in the general population, differently to those that affect the survival or prognosis of a cancer patient. This distinction is important and the factors that affect cancer risk are not necessarily the same as those that will influence the prognosis after diagnosis. One example is smoking and lung cancer; stopping smoking will have a large effect on an individual's risk of developing lung cancer but once lung cancer has developed, stopping smoking may only have a trivial effect on a patient's prognosis. In some study designs, such as the ecological studies mentioned above, that look at cancer mortality in the general population, the outcome is a combination of developing a cancer and then that cancer progressing, but this is not usually the case.

This thesis will focus predominantly on the effect of vitamin D on HNC incidence (risk), but where data are available it will assess the effect with cancer prognosis. The epidemiological evidence assessing the effect of vitamin D on prognosis is limited and is summarised below.

Meyer and colleagues first assessed the effect of 25(OH)D on HNC prognosis in 2011[130] is a study nested within a randomised trial for β-carotene or vitamin E. Serum 25(OH)D and vitamin D intake (from diet and supplements) was recorded at diagnosis and split into quartiles in the 540 participants. Associations between serum 25(OH)D or vitamin D intake quartile were assessed with HNC recurrence, second primary and overall mortality. There was no strong evidence for association of either exposure with the outcomes, although for overall mortality, the direction of effect was protective (lowest versus highest 25(OH)D quartile HR = 0.85 (95% CI: 0.57, 1.28)).

A study later that year investigating 88 consecutive newly diagnosed HNC patients found evidence of a protective effect of 25(OH)D levels at diagnosis with disease free survival (RR per 1ng/ml increase in 25(OH)D = 0.85 (95% CI: 0.75, 0.90)) and overall survival (RR = 0.89 (95% CI: 0.83, 0.97)).[131]

The most recent work assessing the association between 25(OH)D levels at diagnosis and HNC survival identified a similar protective effect.[132] In this study HNC patients were split into low (below the mean) and high (equal or higher than the mean) 25(OH)D groups. In the low vitamin D group 42.6% of patients died during follow up but in the high vitamin D group only 30.3% of patients died, equivalent to a risk ratio of 0.71.

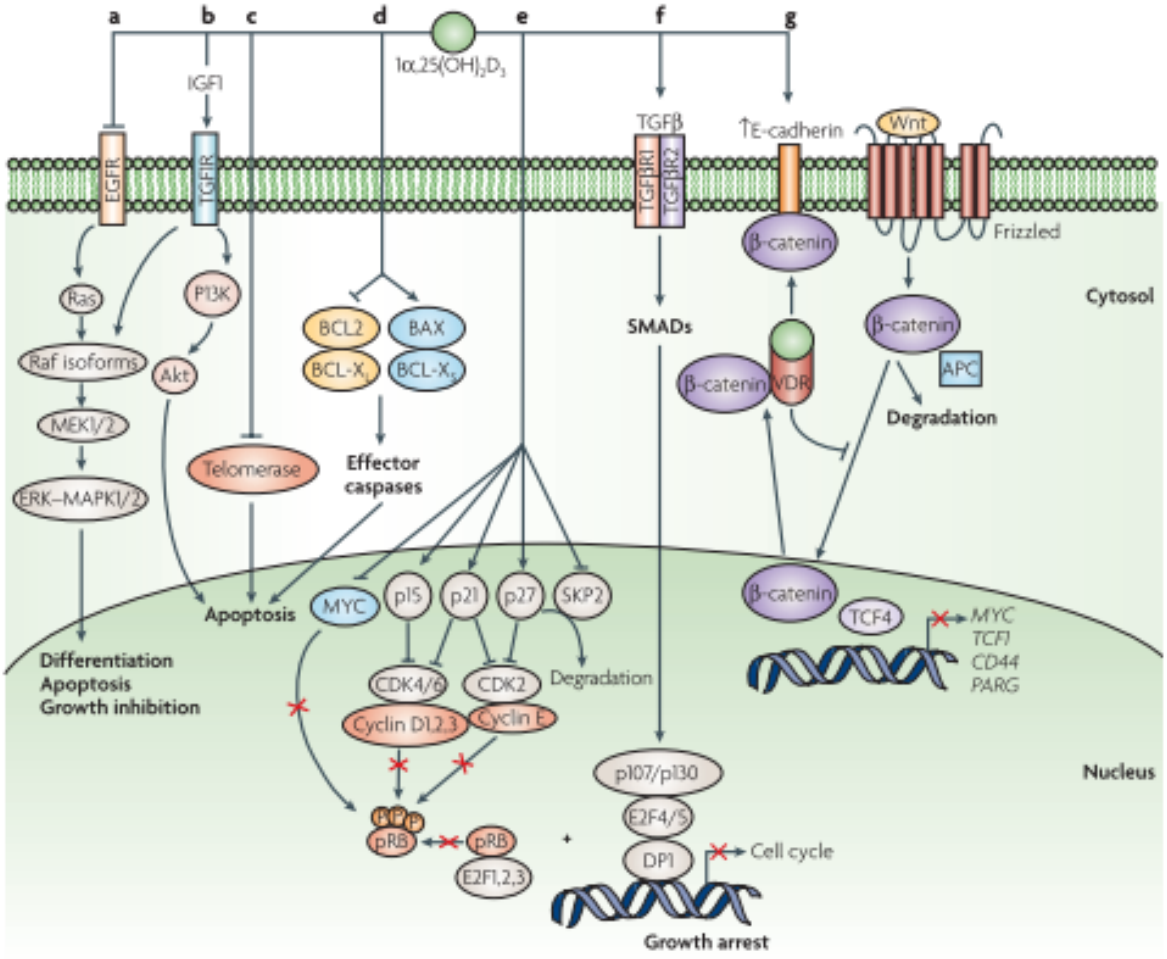
These studies seem to suggest an observational protective effect of vitamin D on HNC prognosis however they may all be affected by measuring vitamin D at diagnosis. It is conceivable that patients who have more advanced disease are more likely to remain inside and have poorer diets than those with less advanced disease, reducing their vitamin D levels. If this is the case, 25(OH)D may be a marker of cancer stage rather than a causal factor in patient prognosis. One study nested within the EPIC cohort study[126] recorded pre-diagnosis 25(OH)D so would not be affected by this issue. In this study of 250 HNC participants the hazard ratio (HR) for a doubling of 25(OH)D was 0.73 (95% CI: 0.53, 0.97). This association was not attenuated by adjusting for confounding or when only including participants who had 25(OH)D measures greater than five years before diagnosis. This suggests that cancer-related lowering of 25(OH)D is unlikely to be responsible for the association seen in the previously discussed studies. It does however seem as though stage at diagnosis has a role in the 25(OH)D – HNC prognosis association, in the 161 cases with stage data in this study, the association was attenuated after correction for stage.

2.4.3 Vitamin D cancer protective mechanisms

The clear distinction between the epidemiological study of risk and prognosis does not necessarily hold true for mechanistic studies of cancer. Cancer mechanisms are multi-stage processes which likely require errors to occur in multiple normal cellular processes. It can be difficult to distinguish if these process errors map to the concept of cancer risk, prognosis or both. They can however, be considered more closely related to cancer initiation (increased proliferation, errors in differentiation and reduced apoptosis) or more closely related to the prognosis of a cancer patient (factors that affect local spread and metastasis e.g. angiogenesis).

There is a large body of work investigating the mechanism by which vitamin D may exert protective effects on cancer. Much of this work focusses on the cancer sites that have also received the most focus in epidemiological studies, colorectal, breast and prostate cancers. This work has been reviewed by Deeb and collegues[133] and is summarised in Figure 2-10. Those processes that are of particular importance in HNC are described below.

Figure 2-10: Key cancer-related signalling pathways targeted by 1,25(OH)$_2$D (from [133] Reprinted by permission from Springer Nature)



2.4.3.1 Epidermal growth factor (Figure 2-10a)

Epidermal growth factor (EGFR) is a cell surface receptor which when activated by its ligand, transforming growth factor alpha (TGFα), inhibits cell differentiation, apoptosis and growth. The active form of vitamin D, 1,25(OH)$_2$D, has been shown to inhibit this pathway to exert its anti-cancer effects. It has been shown that EGFR mRNA is on average 29-fold higher in normal oral tissue of 91% of HNC patients compared to controls and TGFα mRNA is on average five-fold higher in 95% of normal tissue of HNC patients compared to controls, with differences even more pronounced in tumour tissue.[134]

2.4.3.2 Cyclin-dependent kinase inhibitor 1B and cyclin D (Figure 2-10e)

Cyclin-dependent kinase inhibitor 1B (p27) is an enzyme inhibitor which downregulates cell cycle progression at G_1 by inhibiting cyclin D dependent phosphorylation of retinoblastoma in the nucleus. 1,25(OH)$_2$D has been shown to activate p27 and therefore maintain control of cell cycle progression

but p27 is downregulated in oral SSCs suggesting it may have a role in these cancers.[135] This is further supported by studies showing cyclin D is amplified in 20-50% of oral SCC.[136]

2.4.3.3 Transforming growth factor beta (Figure 2-10f)

Transforming growth factor beta (TGFβ) acts via SMAD proteins to inhibit cell cycle progression via a different mechanism to cyclin D, a process which is activated by $1,25(OH)_2D$. TGFβ is thought to have several roles in HNC including a role in cancer initiation, as a potent tumour suppressor and later in the cancer mechanism by being involved in epithelial-mesenchymal transition and facilitating metastasis.[137]

2.5 Summary

HNC has an extremely high level of morbidity and mortality. Despite the main risk factors being well understood, approximately 30% of PAR cannot be explained by these factors. In addition, there has been a recent shift in characteristics and worrying increase in incidence of OC and non-HPV related OPC. This change is difficult to explain with the parallel reduction in smoking levels and stable levels of alcohol use and there is therefore a requirement to identify novel causal risk factors which may explain these trends.

Vitamin D has been linked to many disease states over recent decades but evidence of strong causal associations is difficult to establish. Vitamin D deficiency is easily modifiable and supplementation is becoming more common, despite this there are concerns that changes in working and lifestyle patterns mean populations have chronically low vitamin D levels.

There is strong evidence for a protective association of vitamin D on cancer incidence and prognosis from mechanistic and observational epidemiological studies. For HNC, the observational estimates suggest large reductions in risk, particularly for OC. Vitamin D levels, measured before or at diagnosis, have also been strongly associated with improved cancer prognosis. Despite this observational evidence, causality has not been established in either cancer incidence or prognosis.

Smoking is likely to be a strong confounder of the relationship with HNC incidence but well-designed prospective studies have shown clinically relevant effects even after robust correction for smoking. There is inconsistency across studies however and where strong evidence exists it tends to attenuate in non-smokers. Additionally, there is likely to be residual confounding of other imprecisely measured or unmeasured confounders, biasing the observational estimates away from the null.

RCTs are not a practical method to investigate HNC incidence due to the relatively rare disease in question and trials of vitamin D as an adjuvant in HNC patients would require stronger evidence

before clinical trials would be indicated. Alternative methods are required to assess the causal effect of vitamin D on HNC incidence and prognosis to inform whether future research and health resources should be used in this area. It is into this unoccupied niche which this thesis sits.

This thesis aims to identify the true causal effect of vitamin D on oral and oropharyngeal SCC risk by triangulating evidence from the most robust observational sources and Mendelian randomization. Secondly, the thesis aims to conduct preliminary studies to apply these methods to study the true causal association between vitamin D and oral and oropharyngeal SCC prognosis.

The next chapter

In the next chapter I introduce the concept of causality and present a series of models by which to appraise evidence for causal inference. The general methods and resources used in the rest of this thesis are also described.

Chapter 3 : Methodological Overview

3.1 Chapter objectives

In this chapter I introduce the concept of causality and issues which can prevent causal inference. I go on to describe the approaches that will be used in this thesis and how they can contribute to estimating a causal effect estimate. Finally, I describe the main resources used within the following results chapters.

3.2 Causality

3.2.1 What is causality

Causality in epidemiology is not a straightforward concept and there are various definitions that exist within the field that stem from many centuries of philosophical discussion. They span a range of concepts of causality[138] but typically relate to a "difference making" approach.[139] One such view of causality is the counterfactual view that, had the cause been "different", or more specifically absent, then the effect would also be absent, in at least some of those being studied. A second view is that of contrastive causation, in which the exposure is only a cause given a particular contrast e.g. moderate smoking is only a cause of lung cancer with reference to smoking abstinence but not with reference to heavy smoking. Within this view, the contrasts can be identified for both exposure and outcome. The third view, interventionist causation, relies on a cause being something we can imagine intervening on. This can be further split into thinking about causes which are possible to intervene on, e.g. a drug, or those that we can in principal intervene on but are not necessarily possible or feasible at the current time. An example of the latter is the *BRCA1* mutation which cannot currently be altered but is a very well-established cause of breast cancer. In the possible interventionist view of causality, the *BRCA1* mutation would not be considered causally related to breast cancer. This view of causality limits the type of causal questions which can be posed and limits the epidemiological tools available to investigate causality to only those that implement a possible intervention.[139] In this thesis, the less restrictive view of causality will be used in which anything we can imagine intervening on can be considered causal.

3.2.2 How do we investigate causality?

As discussed in Chapter 2, conducting a large RCT for 25(OH)D supplementation to assess the causal effect on HNC risk is possible but not necessarily feasible. In order to investigate causality in this question other strategies, other than an RCT are required. Sir Austin Bradford Hill proposed nine concepts that should be studied when deciding if any factor should be considered causal or merely the result of association. Hill suggested that these concepts should be applied to the research

question under study before deciding if the most likely interpretation is causality. The nine concepts are described below in the context of this thesis:

3.2.2.1 Strength of association

In this concept Hill argues that the larger the strength of the association the more likely it is to be true. In the context of HNC, the odds of developing laryngeal cancer in non-drinking heavy smokers (31-40 cigarettes per day) versus never smokers is very high (OR~18). In comparison the odds of developing oral cavity cancer in non-drinking heavy smokers is much more moderate (OR~3). It is inconceivable that a third factor could be so strongly interlinked with smoking that it was the real underlying cause in larynx cancer, without that third factor being easily detected. On the other hand, it is much more conceivable that a third factor could be related to smoking, for example poor oral hygiene, that might explain the smaller association with oral cavity cancer, without it being easily detected. It should be noted however that this concept does not work in reverse, as there are many instances in medicine in which factors that have small effects on an outcome are genuinely causally related.

3.2.2.2 Consistency

This concept relates to multiple sources of evidence all pointing to the same finding. The greater the variety of evidence, be that relating to situation or technique of the study, the lower the likelihood of there being a constant error that would explain a non-causal association in all situations. It is important to note that there may be circumstances where a constant error is in fact present in nearly all situations and if this is the case, there will likely be an alternative hypothesis to causality. In the context of this thesis, there is observational evidence for an effect of 25(OH)D on HNC risk from several prospective and retrospective sources and this begins to provide stronger evidence of a causal relationship. However, when smoking is considered across these differing studies the effect is often only seen in non-smokers. This may point to a consistent failure to appropriately account for smoking in each of the studies and therefore, the alternative hypothesis, that residual confounding explains the association across the differing study designs, is more likely. This concept is similar to that of Triangulation, which is discussed below.

3.2.2.3 Specificity

A specific effect of an exposure on a particular outcome may provide stronger evidence of a causal effect compared to that same exposure having an effect on many outcomes. The example Hill quotes relates to the causal association of a workplace exposure leading to death by a specific disease but, not being associated with other causes of death.[140] This would suggest that the exposure is causally related to the disease leading to death. Vitamin D has been linked to many

diseases over recent decades and so does not necessarily fulfil this criteria but this is also true for smoking. In these cases, where an exposure is related to many health outcomes, Hill suggests referring back to his first concept and assessing for specific (and large) effect sizes for a particular disease.

3.2.2.4 Temporality

The concept of temporality relates to identifying the order of exposure and outcome. Although this is often easily identified, it can be problematic in diseases which develop relatively slowly such as HNC. For vitamin D to be causally related to HNC, low levels must precede development of the cancer. The alternative hypothesis is that an early cancer may, unbeknown to the patient or researcher, influence diet and lifestyle and result in lower vitamin D levels, essentially the cancer becomes the cause for the low vitamin D (often referred to as reverse causality).

3.2.2.5 Biological gradient

This concept relates to an increase in the exposure corresponding to an increase in the outcome, often referred to as a dose-response relationship. An example is again provided by the effect of smoking on laryngeal cancer which, shows a clear dose-response relationship with both number of cigarettes smoked per day and the number of years smoked.[32] This gradient of effect is much more convincing of causality than an association which is stronger in light smokers than in heavy smokers.

3.2.2.6 Plausibility

This concept relates to there being an understanding of the biological mechanism underpinning the association. Hill suggested we should not hold this concept too highly, as it relies on biological knowledge at the time of study. There have been many, now established, causes of disease which did not seem plausible at the time of discovery.

3.2.2.7 Coherence

The concept of coherence is opposite to that of plausibility. Any causal association should not conflict with the known natural history and biology of the disease. For example, the incidence of laryngeal cancer does not conflict with the sex and geographical distribution of smoking, nor does it conflict with the understanding that tobacco smoke, a known carcinogen, contacts the laryngeal tissue during inhalation.

3.2.2.8 Experiment

In certain situations, the relationship under examination can be assessed using experimental evidence. In the case of this thesis, 25(OH)D supplementation is possible and an experimental study could form very strong evidence of causality if the study design is appropriate. Experiment does not

solely relate to RCTs as many other experimental methods exist and can contribute to assigning causality without being convincing on their own.

3.2.2.9 Analogy

This concept relies on there being similar causal associations which we can use to judge another. Analogy may be particularly useful when new evidence is found of a potential causative agent and/or when the consequences of ignoring a potentially causal relationship are very high. For example, if a drug similar to thalidomide was available it would be sensible for it not to be given to pregnant people. A HNC related example of this could be used when investigating explanations for the recent rise in non-HPV related oropharyngeal cancer in non-smokers. In the past, a similar increase in the incidence of oropharyngeal cancer occurred and it is now known to be an increase in HPV infection that was the cause. It would be unwise to ignore evidence of a new infective risk factor for oropharyngeal cancer, given what is known about the causal effect of HPV on this cancer.

Hill stated in his President's address to the Royal Society of Medicine, that these concepts do not form *"hard-and-fast rules of evidence that must be obeyed before we accept cause and effect"* [140] but instead provide a framework on which to base the study of causality.

3.2.2.10 Triangulation

One aspects of Hill's approach, the concept of consistency, is very similar to the practice of triangulation which is another useful framework when investigating causality. Triangulation is defined as:

> *"The practice of strengthening causal inferences by integrating results from several different approaches, where each approach has different (and assumed to be largely unrelated) key sources of potential bias."* [141]

Lawlor *et al.* set out a series of criteria that should be met for triangulation to be valid in observational epidemiology. These are: [141]

- Results from at least two, but ideally more, different approaches, with differing and unrelated key sources of potential biases, are compared.
- The different approaches address the same underlying causal question.
- Related to the above criterion, for each approach the duration and timing of exposure that it assesses is taken into account when comparing results.
- For each approach the key sources of bias are explicitly acknowledged when comparing results.

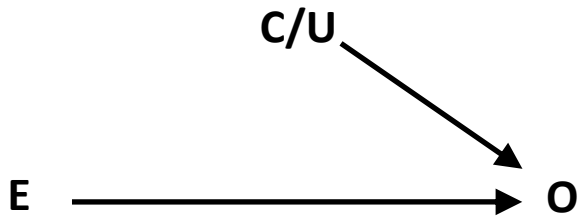
- For each approach the expected direction of all key sources of potential bias are made explicit where this is feasible, and ideally within the set of approaches being compared there are approaches with potential biases that are in opposite directions.

3.2.2.11 Causality in the context of this thesis

Triangulation and Hill criteria overlap but provide a sensible approach to assess the question of causality in the association between vitamin D and HNC risk (and prognosis). Potential explanations for the association between vitamin D and HNC risk are depicted in the directed acyclic graphs (DAGs) below. These DAGs represent the extreme situations to demonstrate the different potential relationships between vitamin D and HNC clearly and in reality more than one may be true at the same time. In terms of causal diagrams, the aim of this thesis is to identify which of these DAGs most likely represents the association between vitamin D (exposure - E) and HNC incidence (outcome - O). In these diagrams, measured covariables (C) and unmeasured covariables (U) are also included to demonstrate the possible situations.

Causal association

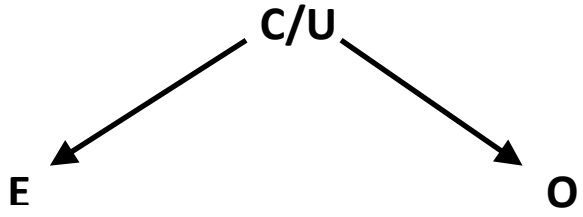
Figure 3-1: Directed acyclic graph representing a causal association between vitamin D and HNC incidence.



In the causal model vitamin D (E) is causally related to HNC incidence (O). Measured and unmeasured variables (C/U) also cause O. For the estimate of E-O to be unbiased there must be no effect of C/U on E. There is a special case of this model in which E is causally related to O, but only when a third variable (such as smoking) is present. This special case represents an interaction between E and C (C is acting as an effect modifer) and can be investigated by calculating E-O across levels of C.

Complete confounded association

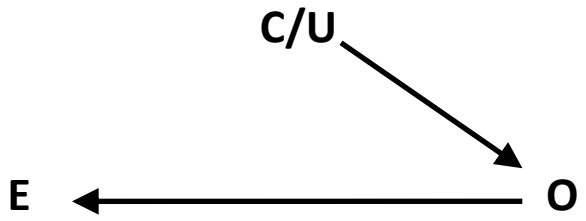
Figure 3-2: Directed acyclic graph representing a complete confounded association between vitamin D and HNC incidence.



In the complete confounded association vitamin D (E) does not cause HNC (O). The apparent association seen can be explained by a measured or unmeasured variable (O/U) being related to both E and O. If this variable is measured without error, conditioning the outcome by C should give a effect magnitude of zero. In cases when C is measured with error, or is unmeasured (U) the effect of E-O will be non-zero due to residual confounding. There may also be situations when there is partial confounding when a true causal relationship is biased by confounding, unless the study design or statistical approach is appropriate, this can be difficult to distinguah from residual confounding.

Reverse causal association

Figure 3-3: Directed acyclic graph representing a reverse causal association between vitamin D and HNC incidence.



In the reverse causal association, E does not cause O but O does cause E. This may be because the onset of the disease alters behaviour in some way. Assessing the association with proper understanding of temporality will usually allow the reverse causal to be distinguished from the causal.

In this thesis I will attempt to apply as many of the concepts of Hill and triangulation as possible to investigate the association between vitamin D and HNC. To accomplish and distinguish between the DAGs presented above, I will use a series of statistical approaches across a variety of study designs. These approaches and others used within the thesis are summarised below, I then provide summaries of the resources in which these approaches are applied.

3.3 Statistical approaches used in this thesis

3.3.1 Multivariable regression

Regression modelling is the process of fitting a line to best describe a relationship between two variables to give you the intercept and the slope of that line. In this thesis linear, logistic, conditional logistic and Cox-proportional hazards regression will be used to investigate continuous, binary, paired binary and time to event data respectfully (these are discussed in more detail below).

Using linear regression, the simplest form of regression, as an example, the intercept will be the value of the outcome when the exposure is zero and the slope (beta or regression coefficient) will equal the effect size for a one unit increase the exposure. In the case of a normally distributed continuous outcome and a binary exposure, this beta will be the difference between the means of the two levels of the exposure and the test for evidence against the null is equivalent to a t-test.

Multivariable regression allows us to include more than one exposure into the model and has two main advantages over calculating effect sizes numerically. Firstly, it allows the effect size of an exposure on an outcome to be estimated while taking into account other measured exposures which influence the outcome. This approach can be used to account for confounding in situations such as that detailed in Figure 3-2 above. In this thesis I will use this approach to adjust estimates of effect between vitamin D and HNC for bias by confounding. The second advantage, which does not apply for logistic and Cox regression, is that when other variables that explain variance in the outcome are included in the model, the standard error of the regression coefficient (effect size) will be lower and therefore improve the statistical power of the analysis.

3.3.1.1 Linear regression

As explained in the example above, linear regression is used to investigate the association between at least one exposure and a continuous outcome. In this thesis this approach will be used to assess the association of a range of variables with both serum and proxy measures of vitamin D. There are two main assumptions which should be met for the approach to be valid. Firstly, for any value of the exposure the outcome must be normally distributed. This is not always the case and an example of this is 25(OH)D which is right skewed. In these instances, transformation of the outcome is required so that it approximates normality. Secondly, the spread of the points around the regression line is assumed to be equal across the full length of the line. Linear regression is limited to estimating linear associations and if the exposure and outcome are not related in a linear fashion the effect estimate from this approach will be invalid.

3.3.1.2 Logistic regression

Logistic regression is an approach to investigate the effect of at least one exposure on a binary outcome. This approach is particularly important in this thesis to investigate factors which are related to getting HNC or not (a binary phenotype). Logistic regression allows a (generalised) linear model to be applied to a binary outcome by modelling the binary outcome on a log odds scale. When considering a binary exposure, the coefficient of a logistic regression is the log odds ratio for having HNC in the exposed group compared to the unexposed group and the intercept is the log odds of having HNC in the unexposed group (baseline odds). Commonly, the log odds ratio of being exposed is exponentiated to an odds ratio (OR) which would be comparable to calculate this value from a two-by-two table. This approach can be extended to a continuous exposure where the interpretation of the OR is the ratio of odds between an exposure of x+1 relative to an exposure of x. Much like linear regression the main advantage of logistic regression is to allow for the correction of confounding by measured covariables. This is achieved by including more than one exposure in the

model. Logistic regression requires observations to be independent of each other, there must be little or no co-linearity between the independent variables and the independent variables should be linearly related to the log odds.

3.3.1.3 Conditional logistic regression

Conditional logistic regression is a special case of logistic regression where there is the assumption that there is pairing of cases and controls. The pairing is achieved by matching values of specific confounders so that each individual with a matched set has the same (or very similar) value. In conditional logistic regression, confounding by the specific confounders is accounted for as the association between the exposure of interest and the outcome is only made within the matched sets. Adjustment for additional confounders, which have not been matched, can be achieved by including them in the regression model.

3.3.1.4 Cox proportional hazards regression

This is the most common approach when conducting regression analyses in time to event data or survival data. In this method the values of the exposures for an individual who has just had an event is compared to the exposures of all the other subjects who have not yet had an event. For this approach to be valid, the ratio of the hazards comparing different exposure groups must remain constant over time. The output (or coefficient) of the regression is the log hazard ratio but the baseline hazard is not estimated as it can change over time. Similarly, to the other regression approaches discussed above, the main advantage of Cox regression is that multiple exposure variables can be included in the model to account for confounding.

3.3.2 Stratification

Stratification is an alternative approach to correct for confounding. In this approach the association of interest, e.g. between vitamin D and HNC, is assessed within strata of the confounder of interest. Much like conditional logistic regression, the effect of the confounder is removed as there is no variation in the confounder within the strata. In order to achieve an overall effect estimate which accounts for the confounder, similar to that from a multivariable regression model, the strata specific effect estimates can be combined using a weighted average. A second advantage of stratified analyses is to identify if that variable is acting as an effect modifier. In this scenario, the effect of the exposure on the outcome will vary according to the level of the confounder. Effect modification is one of the explanations for the association between vitamin D and HNC risk only being identified in current or former smokers and stratification will be used to investigate this in this thesis.

3.3.3 Genome-wide association studies

Genome-wide association studies (GWAS) scan across the genome to identify genetic variation which is associated with a disease or trait of interest in a largely unbiased approach.[142] In this thesis I will use this method to identify genetic variation that robustly influences serum 25(OH)D levels which can then be used as genetic proxies for vitamin D. In a GWAS, a trait and genetic information from across the genome is measured in all individuals in the sample. As an example, consider a single nucleotide polymorphism (SNP) in which an adenine (A) or a cytosine (C) nucleotide base (referred to as alleles) exists at that genomic position in the population. In each individual, the SNP can either be homozygous for the reference base (AA, zero effect alleles), heterozygous (AC, one effect allele) or homozygous for the effect allele (CC, two effect alleles). This three-level variable becomes an ordered categorical exposure and is regressed against the trait of interest. This is repeated across all genetic sites. The output of a GWAS is an effect size and standard error for a one unit increase in the effect allele and a *P*-value against the null hypothesis of no effect. For GWAS estimates to be correct they assume that genotypes are randomly distributed across the sample population and that if the design in case-control in nature that the cases were drawn from the control population (in terms of genetics). Violation of these assumptions can invalidate findings and some examples are discussed below.

3.3.3.1 Measuring genetic information

Measuring genetic information is usually achieved using a genotyping array which uses a high throughput approach to genotype a set number of positions on the genome, approximately 500,000. These positions are strategically selected as they provide enough information to infer the remaining genetic information using imputation. Genetic imputation relies on linkage disequilibrium, where runs of genetic sites between breakpoints are inherited together. These breakpoints are identified from reference panels of whole genome sequenced individuals and then applied in a probabilistic approach to identify the most likely allele at sites not genotyped on the array. The genotyping stage is usually performed in batches as not all samples can be genotyped at the same time.

Alternatives to array-based genotyping exist including whole genome or exome sequencing. In these processes, there is much greater coverage of the genome compared to array-based genotyping and the process of imputation is not required. In these genotyping methods we can be surer of a particular allele at certain sites, particularly those that have low frequency in the reference population.

3.3.3.2 Complications in GWAS studies

Multiple testing burden

As GWAS studies use an unbiased, hypothesis free approach to discover genetic variation, they conduct a very high number of statistical tests creating a large multiple testing burden. To reduce false positives, a Bonferroni corrected *P*-value threshold is used by dividing the arbitrary 5% alpha by the effective number of independent tests. This effective number of tests is lower than the actual number of tests performed as it takes into account the correlation between genetic sites. It is typically set a $5x10^{-8}$ and is often referred to as the genome wide significance threshold.

Sample size

Common genetic variation predominantly results is small effects on the trait of interest. Variation with large effects are either selected against, if detrimental, or predominate if beneficial and therefore are uncommon. Given these small effect sizes and the high multiple testing burden, identifying genetic variation requires very large sample sizes. Very large population-based studies are starting to be developed, for example UK Biobank which has over 500,000 participants (described below), but these are relatively recent projects. The main stay of GWAS are large collaborations of many smaller studies which are combined by meta-analysis to provide adequate power for discovery, often termed meta-GWAS.

Sources of bias in GWAS studies

Information bias is not often a problem in GWAS studies as genotyping completeness and quality is very high. Furthermore, any small error is not often related to the trait in question. An exception to this is when the genotyping batch correlates in some way with the trait of interest. This could happen if samples are genotyped as they are collected and therefore any trait with seasonal variation, such as vitamin D, may be associated with any small genotyping errors between batches. The avoidance of this type of error is relatively straight forward with careful study design.

Selection bias can theoretically occur in case-control GWAS if the control population does not represent the population from which the cases are sampled or there is poor participation rates in the control sample. Despite this, very few instances of false positives have been reported due to this error and this is likely to be because the source of potential bias is not strongly associated with genotype.[143] Furthermore, with the move towards large meta-GWAS, selection bias within any single study bias becomes even smaller with respect to the overall pooled estimate.

Confounding bias in GWAS is most likely to come from population stratification. Genotypes are randomly allocated at gamete formation and segregation meaning that, in a population with

homogenous ancestry and the assumption of random mating, the association between genetic variant and trait of interest is unconfounded. However, in a non-homogenous population, differences in allele frequency which relate to ancestry may correlate with environmental or behavioural differences introducing bias. Vitamin D may be particularly affected by population stratification as 25(OH)D levels are strongly influenced by skin colour and latitude, both of which will be related to ancestry. For example, any genetic variants which are correlated with ancestry in a similar way to those that influence skin colour, but do not causally influence vitamin D may be identified in GWAS by error if ancestry is not dealt with. This type of bias can be addressed by including genetic principal components in the statistical model to account for the differences in allele frequency across sub-populations, or alternatively using a linear mixed model for the analysis. Further details of how other forms of latent structure in genetic studies can influence results of applied genetic analyses is discussed in work to which I contributed.[144]

3.3.4 Mendelian randomization

MR is an analytical approach to estimate causal effects between an exposure and an outcome.[145,146] It is a type of instrumental variable analysis as it uses a proxy measure of the exposure which, is unrelated to confounders and only influences the outcome through the exposure. In MR one or more genetic variants, robustly associated with the exposure of interest, are used as the instrumental variable. Genetic variants are randomly allocated at gamete formation and segregation and as a consequence are not associated with typical confounding factors at the population level. Furthermore, genetic variants are unaffected by behaviour, environment or disease. Given these two properties, the estimates of effect from MR are largely free from bias by confounding and reverse causality and therefore approximate the true causal effect estimate, as long as three crucial assumptions hold.

3.3.4.1 Assumptions

MR relies on certain assumptions for the effect estimate to be a valid estimate of the true causal estimate. These are:

A. **The relevance assumption:** The genetic variants used within the MR analysis are robustly associated with the exposure of interest.
B. **The independence assumption:** The genetic variants share no common cause with the outcome of interest, i.e. are not confounded.
C. **The exclusion restriction assumption:** The genetic variants only affect the outcome of interest through the exposure of interest, i.e. are not pleiotropic (see below for more detail on pleiotropy).

There are further assumptions of MR including monotonicity of genetic instruments, which assumes the genetic proxy influences every individual in the same direction, and no assortative mating, which assumes that individuals mate randomly with respect to the phenotype under study.

3.3.4.2 One sample MR

The first MR analyses required the exposure, genotype(s) related to that exposure and outcome to all be measured in the same study, now referred to as single-sample MR. These studies were limited in sample size due to the practicalities of having all these variables measured in large enough samples.

3.3.4.3 Two sample MR

Two sample MR has greatly improved the power of MR analyses. In this approach, the effect of the genetic variant on the exposure (G-E) is taken from one study and the effect of the genetic variant on the outcome (G-O) is taken from a separate study. These studies, usually meta-GWAS, are typically much larger than those used in the one sample MR analyses as they focus on either the exposure or the outcome and so only need genetic information and one other variable to be measured.

The analyses methods of one and two sample MR are different but the estimates are roughly equivalent. In the most common type of two sample MR analyses, only summary statistics of the genetic variant effects are required from the exposure and outcome study. There is a strong move towards these summary data being made freely available for the scientific community and so summary level two sample MR is an attractive analysis option. In situations where there is only one genetic variant being used as a genetic proxy, the causal estimate is calculated via the Wald ratio which is the ratio of G-O to G-E. Where more than one genetic proxy is available, these are combined into a single estimate. One approach to combine the effects of different genetic variants into an overall causal estimate is the inverse variance weighted method. This method, similar to the combining of studies in meta-analysis, takes into account the variance of each of the summary effect estimates. For two-sample MR to be valid, the studies from which the two sample are drawn must be homogenous (in terms of ancestry) and be non-overlapping. Violation of these assumptions can lead to false positives.

An alternative to only using summary level data, is to use the genetic variant effect on the exposure from a GWAS study but calculate the genetic variant effect on the outcome using individual level data. This approach has two main advantages, the first is that the effect of an exposure which is not measured in the study of interest can be estimated. For example, in Chapter 8, the effect of vitamin

D on HNC prognosis is estimated in the Head and Neck 5000 study which does not have serum 25(OH)D measured in the participants. The second advantage is the analyses can be stratified by certain covariables to assess their impact on the causal effect, something which cannot be achieved in summary level outcome data. The most common strategy for this type of MR is to generate a genetic risk score (GRS) which is a combination of all the genetic variants associated with the exposure of interest. This GRS can be unweighted or weighted by the effect size of the genetic variant on the exposure (from the GWAS study) and scaled to match that of the exposure of interest. This scaled and weighted GRS then acts as a genetically predicted exposure level which is unrelated to the typical confounding factors and can be used as the explanatory variable in traditional regression analyses.

One drawback of this approach compared to using summary level data only, is the imprecision of the genetic variant's effect on the exposure is ignored, potentially creating over-precise causal estimates and increasing the chances of false positive findings. To correct for this, a boot strap method can be utilised to adjust the standard error of the causal effect estimate.

3.3.4.4 Methodological issues in MR analyses

There are several issues which must be considered when designing MR analyses as they can cause violation of the MR assumptions. Each of the genetic variants used in the MR analysis should meet the three assumptions detailed above. If this is not the case then the causal estimate may be biased. Several issues which can lead to the MR assumptions being violated are discussed below before a section which discusses how the assumptions can be tested.

Weak instrument bias

Weak instruments are those genetic variants that are valid, in that they would pass all the MR assumptions in an infinite sample but have very small effects on the exposure of interest. This becomes a problem in finite samples where they can bias causal estimates. This bias occurs because of chance correlation between the genetic variants and confounders of the exposure outcome relationship. Where genetic variant-exposure associations are strong, any effect of the genetic variant on the outcome will be via the exposure. However, in the weak instrument situation, the effect of the variant on the outcome via the confounder starts to predominate and will therefore bias the association. In a one sample setting this bias is towards the observational (confounded) estimate whereas in the two-sample setting the bias is towards the null.[147]

Pleiotropy

Pleiotropy is where a genetic variant is associated with multiple risk factors. There are two types of pleiotropy, vertical and horizontal. Vertical pleiotropy is where a genetic variant is associated with two variables but they lie on the same biological pathway and therefore the effect of one is mediated by the other. This type of pleiotropy is not a concern in MR analyses. Horizontal pleiotropy is when a genetic variant is associated with more than one exposure but via different pathways. An example which could be if a genetic variant for smoking was actually related to risk taking behaviour. This variant, which increased the chances of smoking via increasing risky behaviour, would also likely increase alcohol use and unsafe sexual practices (both strong risk factors for HNC), but not through smoking. If this genetic variant was used in an MR analysis of smoking on HNC, the causal effect for that variant would be biased as it would account for smoking, alcohol use and unsafe sexual practices and likely overestimate the effect of smoking on HNC. In the context of this example, horizonal pleiotropy invalidates the exclusion restriction assumption but it can also invalidate the independence assumption if the genetic variant is pleiotropic with a confounder of the exposure outcome association.

Canalisation

This is the process of biological compensation for the role of a genetic variant so that the effect of that variant being absent or reduced is not seen in the phenotype.[148] Canalisation does not necessarily violate the MR assumptions but can influence the interpretation of a MR causal estimate. The aim of an MR study is to estimate the causal effect of an exposure, not the genetic variant, so if canalisation is occurring the genetic variant does not act as a very good proxy as it does not mirror what would happen if the level of the exposure was altered.

Linkage disequilibrium

Linkage disequilibrium is the situation where genetic variants are correlated because they are inherited together due to their proximity on the genome. Iit can be both useful and problematic in MR analyses. Although it is ideal to identify genetic variants that are causally related to the exposure of interest for use as instruments for an exposure, the inference is the same if a genetic variant which is highly correlated with the causal variant is used. Linkage disequilibrium can be problematic however when the variant used as an exposure proxy is correlated with another genetic variant which does not pass the MR assumptions.

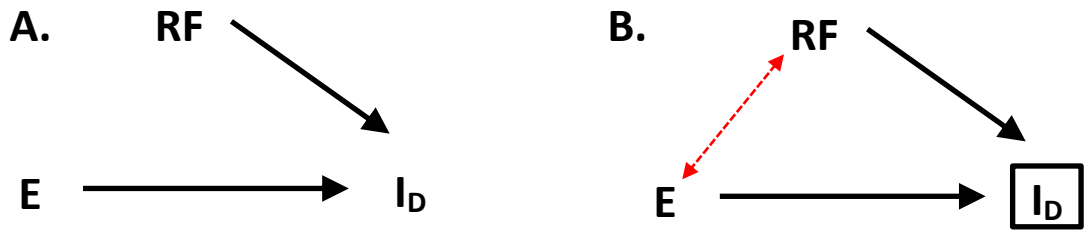
Population stratification

Population stratification in MR analyses is similar to that described above for GWAS. This can introduce confounding to an MR analysis and therefore bias the causal estimate.

Ascertainment effects (collider bias)

Ascertainment or collider bias is when artificial associations between two or more variables are induced because of the act of conditioning or selecting on a third variable. For collider bias to be an issue, the variable being conditioned on must be caused by at least two variables, i.e. the two causes 'collide' onto the third variable (Figure 3-4).

Figure 3-4: Ascertainment bias by selecting or conditioning on case status.

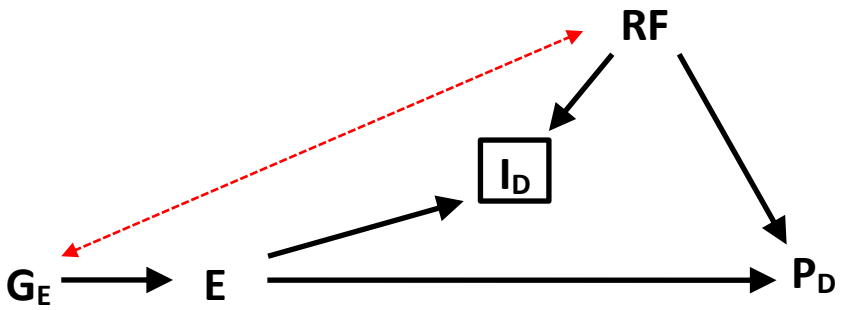


In A. the figure depicts a situation where we are investigating the association between an exposure (E) and a disease (I_D) where there are other causal risk factors for the disease (RF) but these are unrelated to E. In B. when we look only in the disease cases (indicated by the square around I_D), there is an association induced between E and RF (red dashed arrow).

MR relies on randomly assigned alleles to not be correlated with potential confounders of the exposure-outcome association under investigation. If the genetic variants associated with the exposure are associated with the chances of an individual being in the study then the distribution of the alleles for that genetic variant will not be the same as in the unselected population. This will induce spurious associations between the genetic variant and any other factors which are associated with being recruited into the study.

These induced associations create an alternative route from the genetic variant to the outcome and will therefore bias the causal estimate. This bias is often termed collider bias as it results from conditioning or selecting a variable on which two causal associations "collide" (Figure 3-5).

Figure 3-5: Collider bias in the context of studies assessing the causal association between an exposure and disease prognosis.



The genetic variant for the exposure (G_E) is causally related to the incidence of the disease (I_D) through the exposure (E). Other risk factors (RF) which also cause I_D are unrelated to G_E but also affect disease prognosis

(P_D). To investigate the effect of E on the prognosis of the disease (P_D), an MR analysis in conducted in cases only (indicated by the square around I_D). In the cases only G_E is non-randomly assigned and a spurious association is induced between G_E and RF (red dotted arrow). In the case only MR analysis, the causal effect of E – P_D is biased via the aternative route from G_E through other factors which affect disease risk and prognosis (RF).

3.3.4.5 Testing the MR assumptions

The relevance assumption

As described above the genetic variants used in MR analyses as proxy measures are often identified from GWAS using largely hypothesis-free and unbiased approaches. Furthermore, due to the GWAS design, a variant must have very strong statistical evidence for an association for it to pass the multiple testing threshold. The most common approach is to only accept genetic variants which pass the genome-wide statistical threshold which means that those variants which are selected as proxies are highly likely to be robustly associated with the exposure of interest. Some MR studies relax the threshold at which genetic variants are identified[149] which, will increasing the numbers of genetic variants used as a proxy. Some of these additional variants are likely to be false negatives (at the genome-wide threshold level) but there will also be an increased chance of identifying false positive associations which do not pass the relevance assumption. Regardless of whether the variant passes the relevance assumption, these variants are more likely to be weak instruments as there weaker statistical association most likely represents a smaller effect size on the exposure of interest. Statistical information should only be considered the first aspect of selecting genetic variants as proxy measures for MR analyses. Once genetic variants have been identified by statistical thresholds, it is important to assess the biological plausibility for a robust association with the exposure. This can be done using the set of Bradford-Hill criteria set out above, for example if a genetic variant has a strong, consistent and plausible association with the exposure it is more likely to be a valid genetic proxy than a genetic variant that does not show all these qualities.

The independence assumption

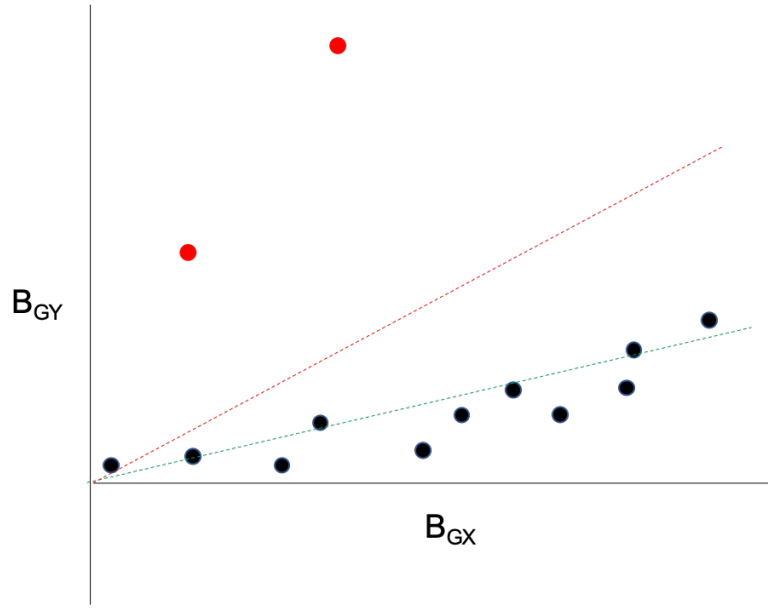
The assumption that the genetic variants used as a proxy share no common cause with the outcome can be tested in some MR studies. This process is similar to demonstrating the randomization in a RCT has been successful and so relies on having individual level data and measured confounders. If confounders are available in the study then associations between the genetic variant(s) and the potential confounders can be tested. In summary level, two-sample MR or when covariables are not available in the study of interest, testing of the genetic variant confounder association can be tested in a separate study or using summary level data. Neither of these strategies can test associations with unmeasured confounders. A further issue is that this method cannot distinguish between a

confounder and a variable which is on the causal pathway between exposure and outcome, a mediator. In the mediator scenario, you would expect an association of the genetic variant with the variable but this does not violate the assumption. It is best to assess these associations thinking about the biological processes in play to assess whether the genetic variant is violating this assumption.

The exclusion restriction assumption

This assumption cannot be easily tested and applying prior biological knowledge when choosing genetic variants is an important step as any other. One way to try to identify pleiotropic effects is to identify genetic variants which are outliers when comparing their effect on the exposure and outcome. Methods such as MR-Egger and Weighted Median use these methods to estimate pleiotropy robust MR estimates.[150,151]

Figure 3-6: Fictional example of how outlier genetic variants can influence the causal estimate.



In this fictional example the effect of the genetic variant on the exposure (B_{GX}) is plotted on the x-axis and the effect of the genetic variant on the outcome (B_{GY}) is plotted on the y-axis, the slope of the lines represents the causal estimate from MR. The dots represent the genetic variants with the two red dots representing pleiotropic varians which are therefore outliers. The outliers can influence the slope of the line as demonstrated by the difference between the green and red dashed lines. In this scenario removing the variants represented by the red dots would provide the true causal estimate (green dashed line).

If the MR assumptions are valid for the genetic proxy, the effect on the outcome will be proportional to the effect on the exposure. Plotting one against the other can help to visualise genetic variants which are not related in a way consistent with other genetic variants but, this is not very helpful with few variants. Pleiotropy can be an issue in all MR methods including one-sample and two-sample approaches combining all genetic variants into a single instrument. The difficulty with having a single

instrument as in these MR methods is there are no outliers to identify. Another way to visualise a single genetic variant which is overtly influencing a causal estimate is to repeat analyses sequentially removing a single genetic variant, referred to as a leave-one-out analysis. If a single variant is influencing the causal estimate substantially, this may be evidence that it is pleiotropic and this can be conducted regardless of the MR methods used.

There are several formal sensitivity analyses which can be conducted to assess the level of pleiotropy and estimate causal effect sizes which is not biased by the pleiotropy. One such method is the median estimator, which can provide reliable causal estimates when up to half of the genetic variants have pleiotropic effects.[150] A second is the MR-Egger method which is an adaptation of the Egger test for small sample bias in meta-analysis.[151] In this method, all instruments can be pleiotropic as long as the strength of the genetic variant exposure association is not correlated with the strength of the genetic variant pleiotropic association . Another benefit of the MR-Egger method is that is can provide a formal statistical test for the presence of unbalanced pleiotropy. Unfortunately, these sensitivity analyses, particularly MR-Egger lack power and utility when only a few genetic variants are available for the exposure of interest.

3.3.5 Meta-analysis

Meta-analysis is the combining of results from a number of studies. The idea of this approach is to combine effect estimates and precision of that estimate from multiple studies into a single pooled estimate of the effect. The method is designed to combine effects from RCTs with the assumption that each estimate is an unbiased estimate of the true population effect size and any difference between the estimates is due to sampling error, with that error being inversely proportional to the size of the study. The pooled estimated, a weighted average of the effect sizes from the individual studies, therefore represents the most precise estimate of the true causal effect in the population. There are two caveats to this interpretation. Firstly, it assumes that each of the RCTs included in the study was of good quality and therefore had no bias. If bias exists in some of the studies then this will bias the pooled estimate. Worse still, if the majority of the studies were conducted in a way that introduces the same bias then the pooled estimate will become a precise estimate of the biased effect. Secondly, meta-analysis assumes the studies being combined are representative of all the evidence available which have been identified using systematic searching approaches. Failure to identify studies or when studies reporting null or negative effects of a treatment are less likely to be reported, the results of the meta-analysis can be biased.

3.3.5.1 Meta-analysis of observational studies

Meta-analysis of observational studies is similar in principal to that of meta-analysis of RCTs but there are more opportunities for bias to be problematic. Observational estimates, as discussed already in this thesis, are likely to be biased by residual confounding, selection bias or other sources of systematic error. As discussed above, Hill's concept of consistency and the theory of triangulation rely on studies with different biases all pointing to the same result. When combining observational studies by meta-analysis however, it is highly likely that many of the studies will include similar biases resulting in a precise yet spurious pooled result. Meta-analysis of observational studies is still useful however as long as the analyst is aware that the process of meta-analysis will not mitigate the biases introduced within the individual studies. A more informative approach is to look at the heterogeneity between the individual estimates with an understanding of the magnitude of the biases that are potentially affecting them. This latter approach will be used in this thesis to combine observational estimates generated as part of this work with those from previous studies.

In this thesis, meta-analysis will also be used to combine estimates from multiple GWAS into a meta-GWAS study and estimates of causal effects from MR studies into a pooled MR-estimate. As explained above, both these study designs estimate an unbiased effect when their assumptions hold therefore, meta-analysis of effects from these studies is similar to that of RCTs.

3.3.5.2 Fixed versus random effects meta-analysis

There are two approaches to combining the individual study estimates into a pooled estimate. The first, fixed-effects meta-analysis assumes that the underlying effect being estimated by each study is the same and that any variation is due to sampling error. To assess this assumption, a test for heterogeneity can be conducted and broadly speaking if there is strong evidence for heterogeneity between the individual estimates the assumption is violated and the pooled estimate is likely to be over precise. The second approach, random effects meta-analysis, allows for possible heterogeneity between the individual studies by including the additional between-study variance in the weighting of the individual effects. A random effects meta-analysis estimate gives more weight to smaller studies than in fixed-effect meta-analysis leading to wider confidence intervals for the pooled estimate and larger *P*-values.[152]

3.4 Studies and participants

There are several study designs that are mentioned within this thesis, the main ones and their assumptions will be briefly discussed before the individual studies are described.

3.4.1 Cohort study

In this type of study the individuals are defined based on the basis of exposure to a suspected risk factor e.g. vitamin D. This can be either presence or absence or as in the case of vitamin D a continuous value. At the time of exposure status determination, all participants must be free from the disease in question. The "cohort" of individuals are then followed up over time to assess the development of the disease of interest in relation to the exposure level.

As participants are disease free when the exposure status is measured these studies are often considered robust to reverse causality however long lag times can violate this assumption. Cohort studies can be expensive as they require large numbers and long follow-up periods, but a major advantage is that many exposures and outcomes can be investigated in the same study. Examples where this is the case are UK Biobank and ALSPAC (described below). Cohort studies can be prospective or retrospective. In retrospective studies, both exposures and the outcomes have occurred when the study is initiated. In prospective cohort studies at least the outcomes are yet to happen.

Selection bias is often not considered to be a major problem in prospective cohort studies as the exposure is assessed before the occurrence of disease and the outcome can influence the classification or measurement of the exposure. Misclassification of exposure and outcome can be an issue in cohort studies. Non-differential misclassification, when there is some level of inaccuracy in the exposure or outcome status but this is similar across study groups, will tend to underestimate the effect towards the null but will not cause false positives. When classification error is related to exposure or disease status then directional bias in either direction can be expected. Similarly, loss to follow-up which is related to exposure or outcome status will introduce bias in either direction.

3.4.2 Case-control studies

Case control studies are designed by identifying individuals who are cases and controls of a particular disease. These individuals are then compared with respect to an exposure or multiple exposures of interest. This study design is good for studying rare diseases and with long latency periods.

The major issue with case-control studies is that both the exposure and the disease have already happened so selection bias can be an issue. It is important to select cases of a single disease entity so that when examining exposures the associations identified are specific to that disease. For case control studies it is also important that the controls represent the population from which the cases are drawn, if not then selection bias will occur. Information bias can also be an issue in case control

studies as information pertaining to exposures is often collected from participants after the onset of the disease and this may affect how they report their exposure status.

There are two variations of case control study which use an established cohort study. The advantage of these designs is that highly detailed information, which would be very expensive to collect for the whole cohort, can be collected for only those cases and controls included in the case-control like study. The first of these is the nested case-control study, in this study design, cases are identified from those followed up in a cohort study and then controls are selected from those still under observation in the cohort study. An alternative is the case-cohort study, in this design a sub-population is randomly selected from the cohort. This sub-population, plus any other cases from the whole cohort are then investigated in more detail. This latter design is slightly more powerful and has the advantage that more than one disease can be investigated in the same sub-population.

The following studies are used as the main sources of data in the thesis. An overview of each is given below in the order that they appear in the thesis, where chapter specific information is relevant this is in a specific methods section within each chapter.

3.4.3 UK Biobank

3.4.3.1 Study description

UK Biobank is a population-based health research resource consisting of approximately 500,000 people, aged between 38 and 73 years, who were recruited between the years 2006 and 2010 from across the UK.[153] Participants provided a range of information pertinent to adult and later life health outcomes via questionnaires, interviews, physical measurement and donating biological samples (data showcase available at www.ukbiobank.ac.uk).[154] UK Biobank received ethical approval from the North West Multi- centre Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

3.4.3.2 Genetic data

The UK Biobank genotype data (July 2017 release) contains 488,377 successfully genotyped samples. 49,979 individuals were genotyped using the UKBiLEVE array and 438,398 using the UK Biobank axiom array. Pre-imputation quality control (QC), phasing and imputation were completed.[155] In brief, prior to phasing, multiallelic variants or those with minor allele frequency ≤1% were removed. Phasing of genotype data was performed using a modified version of the SHAPEIT2 algorithm.[156] Genotype imputation was performed to a combined UK10K haplotype[157] and Haplotype reference consortium (HRC) reference panels using IMPUTE2 algorithms.[158] A further QC protocol was then

applied at the Wellcome Trust Centre for Human Genetics prior to release (http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf).

Individuals with sex-mismatch (derived by comparing genetic sex and reported sex) or individuals with sex-chromosome aneuploidy were excluded from the analysis (n=814). Estimated kinship coefficients using the KING toolset[159] identified 107,162 pairs of related individuals.[160] An in-house algorithm was then applied to this list and preferentially removed the individuals related to the greatest number of other individuals until no related pairs remain. These individuals were excluded (n=79,448). Additionally two individuals were removed due to them being relating to a very large number (>200) of individuals.[161] Principal components were generated by UK BIOBANK using flashPCA.

3.4.3.3 25 hydroxyvitamin D

Blood was drawn from consenting participants at the baseline clinic and transported to the UK Biobank central laboratory in Stockport, UK. The whole blood was aliquoted and frozen within an average time of 24 hours (+/- 2.4 hours). 25(OH)D was measured by immunoassay on a DiaSorin LIA LIASON XL with a detection range of 10 nmol/L < 25(OH)D < 375 nmol/L. Where samples were above 375 nmol/L and sufficient sample remained, a dilution was performed and the sample reanalysed. During the automated aliquoting process, an unintended dilution of the aliquots occurred however, approximately 92% of sample assay concentrations are affected by less than 1%, approximately 8% of assay concentrations are affected by less than 10% and very few are affected by more than this. Correction for this unintentional dilution has been made centrally by UK Biobank and more details can be found here: http://biobank.ndph.ox.ac.uk/showcase/docs/biomarker_issues.pdf.

3.4.4 Avon Longitudinal Study of Parents and Children

3.4.4.1 Study description

The ALSPAC longitudinal birth cohort recruited pregnant women living near Bristol, UK with an estimated delivery date between 1991 and 1992. The initial number of pregnancies enrolled was 14,541. Of these initial pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. As a result, when considering variables collected from the age of seven onwards there are data available for more than the 14,541 pregnancies mentioned above. The number of new pregnancies not in the initial sample is 913. The phases of enrolment are described in more detail in the cohort profile paper and its update.[162–164] The total sample size for

analyses using any data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these 14,901 were alive at 1 year of age. Follow up has included clinical assessment and questionnaires and is ongoing. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

3.4.4.2 Genetic data

ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, US. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness (>3%) and insufficient sample replication (IBD < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95% or evidence for violations of Hardy-Weinberg equilibrium (P < 5E-7) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD > 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 subjects and 500,527 SNPs passed these quality control filters.

ALSPAC mothers were genotyped using the Illumina human660W-quad array at Centre National de Génotypage (CNG) and genotypes were called with Illumina GenomeStudio. PLINK (v1.07) was used to carry out quality control measures on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium *P* value of less than 1.0e-06. Additionally, SNPs with a minor allele frequency of less than 1% were removed. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using a IBD estimate of more than 0.125 which is expected to correspond to roughly 12.5% alleles shared IBD or a relatedness at the first cousin level. Related subjects that passed all other quality control thresholds were retained during

subsequent phasing and imputation. 9,048 subjects and 526,688 SNPs passed these quality control filters.

SNP genotypes in common between the sample of mothers and sample of children were combined (477,482). SNPs with genotype missingness above 1% were removed due to poor quality (11,396 SNPs removed), a further 321 subjects were removed due to potential ID mismatches. This resulted in a dataset of 17,842 subjects containing 6,305 duos and 465,740 SNPs (112 were removed during liftover and 234 were out of HWE after combination). Haplotypes were estimated using ShapeIT (v2.r644) which utilises relatedness during phasing. The phased haplotypes were then imputed to the Haplotype Reference Consortium (HRC) panel of approximately 31,000 phased whole genomes. The HRC panel was phased using ShapeIt v2, and the imputation was performed using Impute V3.

Whole genome sequence (WGS) data are described within Appendix as it was sequenced alongside the Twins UK study.

3.4.4.3 25 Hydroxyvitamin D

After collection, blood samples were immediately spun, frozen, and stored at −80°C. Assays were performed after a maximum of 12 years in storage with no previous freeze-thaw cycles. The 25(OH)D_3 and 25(OH)D_2 concentrations were measured with HPLC tandem mass spectrometry using internal standard in a laboratory meeting the performance target set by the Vitamin D External Quality Assessment Scheme Advisory Panel for 25(OH)D assays. Interassay coefficients of variation for the assay were less than 10% across a working range of 2.5–624 nmol/L for both 25(OH)D_3 and 25(OH)D_2.[165] Total serum 25(OH)D concentrations were calculated as the sum of 25(OH)D_3 and 25(OH)D_2.

3.4.5 GAME ON

3.4.5.1 Study description

The study comprised 6,034 head and neck cancer cases and 6,585 controls from studies which were part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. Cases and controls from 10 different case-control studies were included, as well as from the European cohort study (EPIC) and cases from a UK case series (Head and Neck 5000). Details of the studies included have been described previously.[22] Informed consent was obtained for all participants and studies were approved by respective institutional review boards.

3.4.5.2 Genetic data

DNA was isolated from buccal cells or blood and was genotyped on the Illumina OncoArray by the Centre for Inherited disease Research (CIDR). The genotyping array was custom designed to be specific for the study of cancer by the OncoArray Consortium which was part of the Genetic Associations and Mechanisms in Oncology Network (GAME-ON). All cases and the majority of controls were genotyped together as part of the oral and pharynx OncoArray project. An additional 2,476 shared controls were genotyped as part of the Lung OncoArray project, also by CIDR. Genotype calls were made using the GenomeStudio software (Illumina). After single nucleotide polymorphism (SNP) filtering by call rate, 513,311 probes remained for analysis. Within each geographical area, principal components were generated using EIGENSTRAT[166]. Imputation of unknown genetic variation was performed on the Michigan Imputation Server[167] using the first release of the Haplotype Reference consortium panel.

3.4.6 Head and Neck 5000

3.4.6.1 Study description

Head and Neck 5000 (HN5000) is a large HNC clinical cohort, full details of the study methods and population are described elsewhere.[168] Briefly, 5,511 people with a new HNC diagnosis were recruited from 76 centres across the UK between April 2011 and December 2014. Recruitment rates to the study varied by centre, from around 20% to around 90% of eligible HNC cases. Individuals were recruited before they started treatment, unless their treatment was their diagnostic procedure. At baseline, 5,474 (99%) data capture forms and 4,099 (74%) health and lifestyle questionnaires were completed.[169] Baseline blood samples were obtained from 4,676 (85%) individuals. Full ethical approval was granted by The South West – Frenchay Regional Ethics Committee (ref: 10/H0107/57).

Baseline information was collected via data capture from clinical notes and questionnaires. Formal study-based follow-up was conducted with further data capture and questionnaires at 4 months, 12 months and 3 years. Data linkage with death registry sources provided details of survival with the last recorded data freeze being ???.

3.4.6.2 Genetic data

Blood samples were taken at recruitment and stored in EDTA tubes. Blood samples were spun at 3,500 rpm for 10 minutes and the buffy coat layer stored for future DNA extraction. DNA extraction was performed by LGC genomics (http://www.lgcgenomics.com/) using the Kleargene spin column extraction method (http://www.lgcgroup.com/products/dna-extraction-kits).

DNA samples from 1,079 participants were included in the genotyping of the GAME-ON study described above using the Illumina OncoArray. Imputation was completed using the HRC reference panel alongside all other GAME-ON samples.

The next chapter

In the next chapter I will assess the observational association between serum 25(OH)D and HNC incidence and prognosis in the UK Biobank study, using some of the methods described above. As far as is possible, I will attempt to mitigate the bias of confounding using multivariable regression and stratification. The results from this study will then be compared and combined with the estimates from previous observational evidence of these associations.

Chapter 4 : Revisiting the observational vitamin D HNC association

4.1 Chapter objectives

In the introduction, the body of observational evidence linking vitamin D to HNC risk and prognosis was presented. In Chapter 4 observational methods will be used to conduct a prospective study investigating this association. The results are then pooled with those from Fanidi *et al.* (2016)[126], the most recent observational study investigating this association. The aim of this chapter is to use newly available data in UK Biobank to re-examine the association between serum 25(OH)D and head and neck cancer risk and prognosis and to produce the most precise observational estimates of the association to date.

4.2 Introduction

In Chapter 2 epidemiological studies assessing the association between 25(OH)D and HNC were described. There is however heterogeneity in the design of these studies. Some of the early studies used predicted or dietary vitamin D levels whereas the more recent studies used serum 25(OH)D and there is inconsistency in how different sites within the head and neck region are combined. Given these differences, it is not surprising that there is also inconsistency in the results, with several studies reporting strong protective effects on HNC risk and prognosis and others providing no robust evidence for an effect of vitamin D on either.

The most recent study by Fanidi *et al.* (2016), is the largest and most robustly designed of the observational studies investigating the effect of vitamin D on head and neck cancer. As discussed, the prospective design and correction by known confounders likely provides the most valid estimate of the effect of 25(OH)D on HNC risk and prognosis to date. The results however, have not yet been replicated in a similarly designed and well powered study.

UK Biobank, a large prospective cohort study, provides an opportunity to replicate the findings of Fanidi *et al*. in an independent cohort. The breadth of data available in UK Biobank allows the analysis plan of Fanidi *et al.* to be closely replicated with consistency in vitamin D measurement and how HNC sites are defined. This will allow comparison between the studies and meta-analysis of the results. As set out in Hill's criteria for causation,[140] consistency between the study results and large effect magnitudes can add to the view that the association may be causal. Through meta-analysis of the results of the work in this chapter with the Fanidi *et al.* study, a more precise estimate of the association magnitude can be identified. It must be noted however, that any association identified in

this work could still be biased by unmeasured confounding and it is therefore important that the results are assessed for the impact of the major confounders such as smoking.

4.3 Methods

4.3.1 Overview

In this chapter UK Biobank data was analysed to mirror the analyses of the Fanidi *et al.*[126] study to assess the association of vitamin D with HNC risk and prognosis. In the analyses for HNC risk, to make the best use of the large numbers of participants in UK Biobank and to ensure analyses were consistent with those in the Fanidi *et al.* study,[126] two parallel study designs were used. Firstly, a matched case-control analysis was conducted where each case was matched to three controls. Secondly, a case-cohort design was used to ensure maximum power.

In the analyses assessing the effect of vitamin D with HNC prognosis, survival analyses were used in the cases only, using all-cause and HNC-specific mortality as end points. The results were meta-analysed to further refine the association estimates and the results are assessed for the effect of major confounders.

4.3.2 Studies

The study was conducted in UK Biobank, a population-based study of approximately 500,000 participants. A detailed description of this study and the 25(OH)D measures are given in Chapter 3.

4.3.3 Identification of incident cancer cases

Incident cancer cases were identified from UK cancer registry data which is linked to the UK Biobank resource (version: March 2019). The cancer cases were defined based on the ICD10 codes from the Fanidi *et al.* study.[126] As UK Biobank includes some cancer cases identified before the adoption of the ICD10 codes, equivalent ICD9 codes were also identified. These ICD9 and ICD10 codes were then allocated to the oral, oropharynx, larynx, hypopharynx and other HNC subsites as specified in Table 4-1.

The ICD codes were used to identify participants who had been diagnosed with a HNC in UK Biobank. For each diagnosis, the date of diagnosis, histology, behaviour (benign, carcinoma-in-situ, malignant, metastatic cancer) and age at diagnosis was extracted from the UK Biobank dataset. This process was repeated to identify participants diagnosed with any other cancer (except non-melanoma skin cancer).

Table 4-1: ICD codes for major head and neck cancer sites

Cancer site		ICD codes
Oral cavity cancer	ICD10	C003, C004, C005, C006, C009, C020, C021, C022, ,C028, C029, C030, C031, C039, C040, C041, C048, C049, C050, C058, C059, C060, C061, C062, C068, C069
	ICD9	1403, 1404, 1405, 1406, 1409, 1411, 1412, 1413, 1422, 1428, 1419, 1430, 1431, 1439, 1440, 1441, 1448, 1449, 1452, 1458, 1459, 1450, 1451, 1456, 1458, 1459
Oropharynx cancer	ICD10	C01, C024, C090, C091, C098, C099, C100, C101, C102,C103, C104, C108, C109, C140, C142, C051, C052
	ICD9	1410, 1416, 1461, 1462, 1468, 1460, 1463, 1464, 1466, 1467, 1468, 1469, 1490, 1491, 1453, 1454
Larynx cancer	ICD10	C320, C321, C322, C323, C328, C329
	ICD9	1610, 1611, 1612, 1813, 1618, 1619
Other HNC cancer	ICD10	C148, CD000, C110, C111, C118, C119
	ICD9	1498, 2300, 1470, 1471, 1478, 1479

Cases who had a history of another cancer (except non-melanoma skin cancer) were excluded. Due to the different aetiology of squamous cell carcinoma (SCC) compared with other non-SCC cancers, non-SCC cancers were also excluded.

4.3.4 Identification of cancer control participants

For each case, three controls were randomly selected from a risk set consisting of all cohort members alive and free of cancer (except non-melanoma skin cancer) at the time of diagnosis of the incident case. Matching criteria were sex, date of blood collection (± 1 month) and date of birth (± 1 year). Three matched controls were chosen as this allowed matching of all criteria. To select more matched controls per case would have meant relaxing the matching criteria and therefore providing less robust control for these potential confounders.

Additional unmatched controls were selected from the remaining UK Biobank participants who did not have a history of cancer (except non-melanoma skin cancer) and had serum 25(OH)D measured.

Participants with non-melanoma skin cancer were included as cases and controls for two reasons. Firstly, these cancers are common and removing them would reduce the power of the study (particularly by the removal of cases). Secondly, as these cancers are caused by sun exposure and the hypothesis in question in this study relates to vitamin D which is largely determined by sun

exposure. There is the potential for bias to occur if sun exposure if differential between the cases and controls. For example, if sun exposure is correlated with other HNC risk factors, removing participants with non-melanoma skin cancers may reduce the average serum vitamin D levels of the case more than the controls, biasing the results away from the null.

4.3.5 Follow-up of cancer cases

Cancer cases identified above were followed up in cancer and death registry data to identify the date of cancer recurrence or death.

Participants were considered to have had a cancer recurrence if they had a further diagnosis of any HNC site greater than three months after their initial HNC diagnosis. Although it is possible for a cancer to recur before this three-month period, it is very rare and more likely would relate to residual disease. Additionally, there are multiple occasions in the cancer registry data in UK Biobank where, the same participant has multiple records for the same cancer diagnosis – sometimes several weeks apart. Applying this three-month cut-off strategy reduces the chances of misclassifying a repeated cancer registry record as a cancer recurrence.

All-cause mortality was identified from death registry data, linked to UK Biobank, as death of a HNC case from any cause. HNC-specific cause of death was identified if an ICD10 code, relating to any of the HNC sites above, was mentioned in either the underlying (primary) or contributory (secondary) causes of death on the death certificate.

Survival time was calculated from time of diagnosis to the earlier of cancer recurrence or death from any cause. The survival analysis was censored at 14th February 2018 – the last recorded death in the UK Biobank data freeze used for this study.

4.3.6 25 hydroxyvitamin D

25(OH)D was measure from blood serum as described in Chapter 3.

4.3.7 Covariables

Covariables were chosen based on the prior knowledge of vitamin D related factors discussed in Chapter 2 and to match the analyses of the Fanidi *et al*. study.[126] Day of birth is not available in UK Biobank to protect the anonymity of the participants. Therefore, for each participant the date of birth was generated assuming everyone was born on the first day of the month. Self-reported ethnic background was recorded at the study assessment centres and grouped into White, Mixed, Asian, Black and Other. Smoking was recorded at baseline and grouped into never, former or current. For former smokers, these were further categorised into those who had quit 10 or more years ago and those who had quit within the last 10 years. Alcohol drinking status (‘never’, ‘former’ or ‘current’)

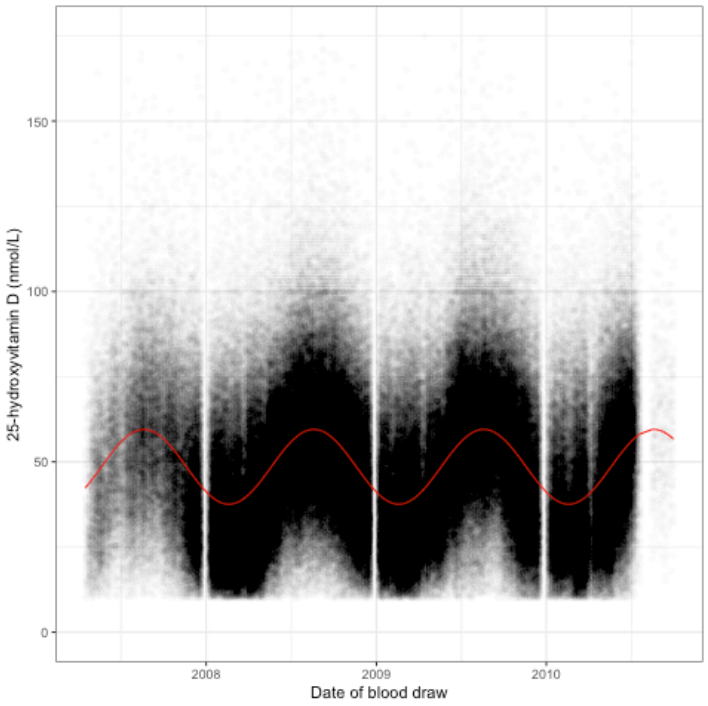
and frequency of drinking alcohol was also recorded. For the latter, the following categories were specified: 'daily or almost daily', 'three or four times a week', 'once or twice a week', 'one to three times a month', 'special occasions only' or 'never'. Highest education attainment was recorded and coded into the following categories: 'college or University degree', 'other professional qualifications e.g.: nursing', 'A levels/AS levels or equivalent', 'NVQ or HND or HNC or equivalent', 'O levels/GCSEs or equivalent' or 'CSEs or equivalent'. For participants who did not answer this question, those that reported leaving school before age 16 years were placed in a category of 'No formal qualifications' and the remaining participants were coded as 'missing' but included in analyses. Body mass index (BMI) was calculated from height and weight measured at the initial assessment centre attendance and categorised into three categories (<25 kg/m^2, 25-29.9 kg/m^2, >30 kg/m^2). For a subset of participants, a 24-hour food and beverage consumption questionnaire was used to estimate dietary vitamin D and calcium intake (excluding any supplements). Physical activity was recorded at initial study assessment centres as total Metabolic Equivalent Task (MET) minutes per week for all activity.

UK Biobank is a UK-wide study with participants recruited from across the country. As discussed, vitamin D is related to geographic location with more northerly latitudes receiving less UV light. It is therefore expected that there is a gradient of 25(OH)D concentrations from north to south. Participants home location in the south-north direction measured in meters from the most southerly point of the UK was used as a measure of latitude. As this continuous variable has a non-normal distribution it was split into quartiles to form a categorical variable. The association of this categorical variable with 25(OH)D was assessed using linear regression and with HNC case status using the chi-squared test.

4.3.8 Statistical analysis

To remove the variation in serum 25(OH)D attributable to season, the 25(OH)D measures were adjusted using the 'predict' function of the 'cosinor' package in R[170] to provide a predicted mean annual outcome value. Briefly, the package identifies an appropriate sinusoidal pattern to model the seasonal variation and then for each individual adds the residual from that model to the mean value of 25(OH)D. This method models the seasonal variation of the 25(OH)D concentration by providing smooth predictions without artificial discontinuities throughout the year (Figure 4-1) . Seasonally adjusted 25(OH)D values were log-2 transformed to normalise the distribution and allow easy interpretation of results – each unit increase in the exposure corresponds to a doubling of 25(OH)D.

Figure 4-1: Seasonal variation in unadjusted 25-hydroxyvitamin D levels.



The plot demonstrates the seasonal variation in 25(OH)D across the recruitment period, gaps in sampling likely represent the Christmas to New Year period. Individual 25(OH)D levels are indicated with the black shading, darker shades represent a greater number of individuals with similar values. The red line represents the predicted geometric mean concentration modelled using the cosinor package in r. This predicted regression line was used to adjust the raw values removing seasonal variation in 25(OH)D.

Within cases, matched and unmatched controls, categorical variables were described by frequency and percentage and continuous variables by median and interquartile range. Within the controls, associations between 25(OH)D and potential confounders was assessed using linear regression.

For each HNC site and all HNC sites combined, conditional logistic regression, conditioning on matched case-control quads, was used to calculate OR and 95% confidence intervals for the base 2 logarithm of seasonally adjusted 25(OH)D which, approximates the relative risk of doubling circulating 25(OH)D. To provide maximal power, matched and unmatched controls were combined in unconditional logistic regression models correcting for age at recruitment, sex and ethnicity. Combining the matched and un-matched controls could potentially reduce the representativeness of the control set however, in this study the matched controls were drawn randomly from those that were included in the unmatched analysis and should not be systematically different. This was assessed by comparing the baseline characteristics of the matched and unmatched controls. Additional adjustment was made for measured potential confounders, including smoking status,

alcohol consumption frequency, highest educational attainment (including missing values), BMI and latitude.

To assess whether a dose-response relationship exists between 25(OH)D and HNC risk, the effect was assessed across quintiles of 25(OH)D concentration, using the lowest 25(OH)D quintile as the reference. The assumption is that if the association is causal, the higher the quintile the stronger the effect on HNC should be. This was repeated for all and site specific HNCs.

To investigate the effect of confounding and/or interactions, the logistic regression models were run within strata of covariables. Where heterogeneity existed between strata, the presence of an interaction between vitamin D and the covariable was investigated by conducting logistic regression including and excluding an interaction term. A log likelihood test was used to check for evidence that the model including the interaction term was of better fit than the model without an interaction term and if this was the case, this provides evidence that the covariable is modifying the effect of vitamin D.

To provide the most precise estimate of the effect of 25(OH)D on HNC, the adjusted unconditional logistic regression estimate was meta-analysed with the estimates from Fanidi *et al*. (2016)[126] using a fixed effect model. Heterogeneity between the two studies was assessed using the I^2 statistic.

The association between pre-diagnostic 25(OH)D and prognosis in incident HNC cases was assessed. Due to the low number of cases, all HNC sites were considered together in the primary analysis. Recurrence of HNC or death from any cause was considered an event and time since diagnosis was used as the time scale. Participants were separated into low (<50 nmol/L) and high (>=50 nmol/L) 25(OH)D categories and survival distributions compared using Kaplan Meier plots and the log-rank test.

Using cox-proportional hazards regression, hazard ratios were estimated for a base 2 logarithm of seasonally adjusted 25(OH)D (i.e. a doubling of vitamin D) and adjustments for the same covariables as the unconditional logistic regression were made.

Sensitivity analyses

The OR and HR of a base 2 logarithm change in 25(OH)D was assessed stratified by the following variables: sex, ethnicity, smoking status, alcohol intake, educational level, age at diagnosis, time from blood draw to diagnosis and BMI.

To account for the effect of UK Biobank geographic location clustering may have had on the precision of estimates, the unconditional logistic regression was repeated using a measure of

participant's home location in the South to North direction as the cluster variable. Analyses were conducted using the "glm.cluster" function, in the 'miceadds' package in R,[171] which computed cluster robust standard errors.

In order to investigate if any effect of 25(OH)D on HNC prognosis was acting via effects on non-HNC causes of death, cox proportional hazards regression was repeated using HNC-specific death and compared with the estimates for all-cause mortality.

Deaths that occur early after HNC diagnosis (< 6 months) are likely to have a different cause than those that occur later on (> 6 months). In order to assess this the survival analysis was repeated only for those participants who had an event later than 6 months after diagnosis.

4.4 Results

4.4.1 Baseline characteristics

There were 502,543 participants enrolled into UK Biobank, 54,593 participants did not have valid seasonally adjusted vitamin D levels. After excluding participants with a previous diagnosis of any cancer except non-melanoma skin cancer (61,850) there were 386,100 participants included in the study. Overall there were 480 incident HNCs, of which 145 (30%) were oral cavity cancers, 220 (46%) were oropharyngeal cancers, 85 (18%) were laryngeal cancers and the remainder (6%) were other head and neck cancers including hypopharynx. 1,440 individually matched controls (three per case) and 384,180 additional unmatched controls were identified from UK Biobank (Figure 4-2). Greater than three quarters of the matched case-control sample were male compared to under half of the unmatched controls being male. Median age at recruitment was slightly lower in the unmatched control set (57 years, IQR: 49,63) compared to the matched cases and controls (59 years, IQR: 53, 64). Cases were more likely to be current smokers or previous smokers who have quit within the last 10 years, compared to matched and unmatched controls. There were a high proportion of cases from the two most northerly latitude quartiles whereas matched and un-matched controls were evenly spread across the quartiles (Table 4-2). In formal tests of association, there was no strong evidence for an association between latitude and the risk of HNC (χ^2 = 7.47, df = 3, P = 0.06) however this may reflect a lack for power.

Figure 4-2: Study profile

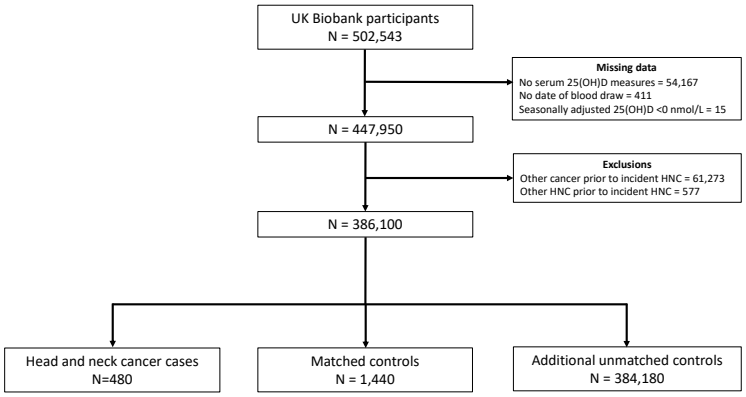


Table 4-2: Baseline characteristics of cases and controls

Summary Statistics	Cases	Matched Controls	Unmatched Controls
Ethnicity			
Asian	11 (2)	31 (2)	8,728 (2)
Black	0 (0)	16 (1)	6,392 (2)
Mixed	3 (1)	7 (0)	2,341 (1)
Other	5 (1)	7 (0)	3,595 (1)
White	460 (96)	1,370 (96)	361,308 (94)
Missing	*1 (0)*	*9 (1)*	*1,816 (0)*
Sex			
Male	370 (77)	1,110 (77)	180,888 (47)
Female	110 (23)	330 (23)	203,292 (53)
Education			
College or University degree	121 (25)	477 (33)	126,734 (33)
Other professional qualifications	0 (0)	0 (0)	0 (0)
A levels/AS levels or equivalent	49 (10)	153 (11)	43,277 (11)
O levels/GCSEs or equivalent	109 (23)	285 (20)	81,108 (21)
NVQ or HND or HNC or equivalent	0 (0)	0 (0)	0 (0)
CSEs or equivalent	22 (5)	66 (5)	21,625 (6)
No formal qualifications	122 (25)	249 (17)	62,291 (16)
Missing	*6 (1)*	*19 (1)*	*4,448 (1)*
Smoking Status			
Never	136 (28)	739 (51)	212,981 (56)
Previous	188 (39)	543 (38)	129,916 (34)
Years since quitting < 10	56 (38)	94 (24)	23,567 (27)
Years since quitting ≥ 10	91 (62)	303 (76)	62,299 (73)
Current	154 (32)	158 (11)	39,401 (10)
Missing	*2 (0)*	*0 (0)*	*1,882 (0)*
Alcohol Frequency			
Never	49 (10)	101 (7)	30,131 (8)
Special occasions only	35 (7)	131 (9)	43,539 (11)
One to three times a month	27 (6)	141 (10)	43,017 (11)
Once or twice a week	140 (29)	387 (27)	99,832 (26)
Three or four times a week	89 (19)	331 (23)	89,699 (23)
Daily or almost daily	140 (29)	349 (24)	77,101 (20)
Missing	*0 (0)*	*0 (0)*	*861 (0)*
BMI			
<25	154 (32)	416 (29)	128,010 (33)
25-29.9	194 (41)	673 (47)	162,225 (42)
≥30	130 (27)	351 (24)	92,446 (24)
Missing	*2 (0)*	*0 (0)*	*1,499 (0)*
Latitude (quartiles)			
1 (most southerly)	103 (22)	368 (26)	99,547 (26)
2	108 (23)	353 (25)	90,821 (24)
3	137 (29)	356 (25)	95,123 (25)
4 (most northerly)	128 (27)	355 (25)	94,737 (25)
Missing	*4 (1)*	*8 (1)*	*3,952 (1)*
Continuous variables, Median (Interquartile range)			
Age at recruitment	59 (53, 64)	59 (53, 64)	57 (49, 63)
Seasonally adj. 25-hydroxyvitamin (nmol/L)	43.10 (26.28, 60.85)	47.80 (32.30, 62.90)	46.70 (32.30, 62.30)
Physical activity METs (hours/week)	28.65 (9.95, 63.10)	28.75 (12.95, 56.87)	29.77 (13.55, 59.55)
Dietary variables			

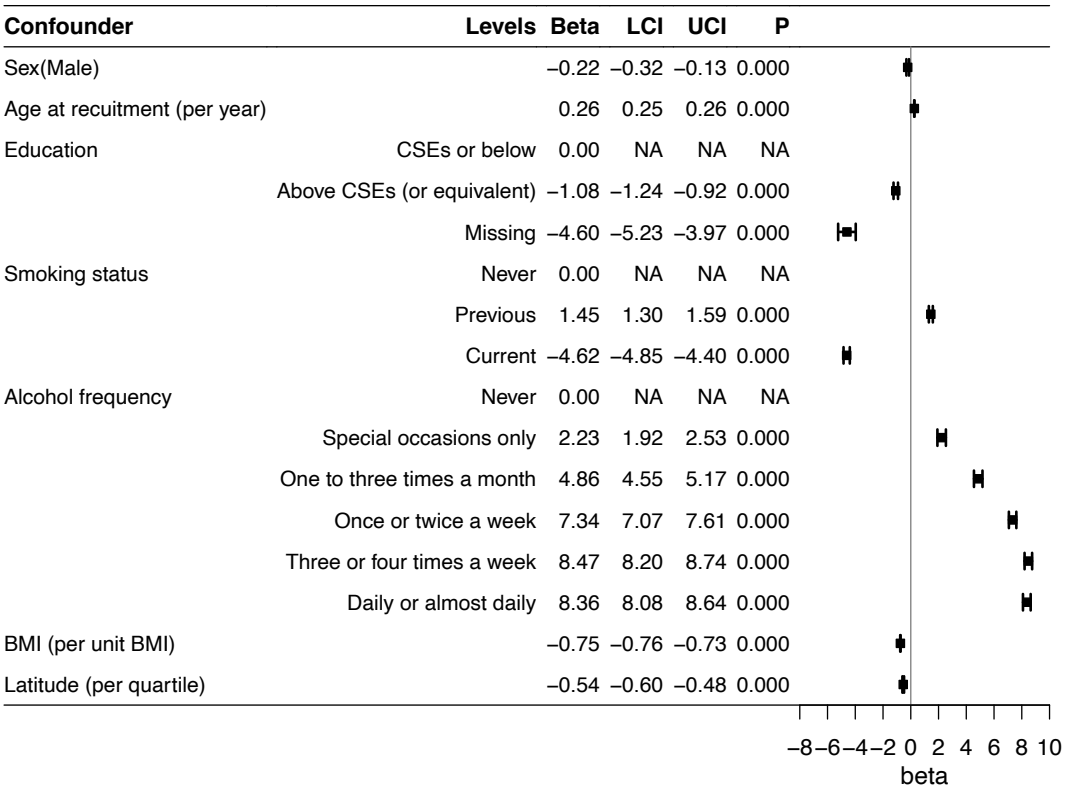
Vitamin D (ug/day)	1.58 (0.78, 2.96)	1.85 (0.93, 3.59)	1.73 (0.85, 3.31)
Calcium (mg/day)	872 (677, 1,153)	948 (730, 1,205)	920 (699, 1,188)
Case only variables			
Age at diagnosis	62.87 (57.79, 68.26)	NA (NA, NA)	NA (NA, NA)
Time from blood draw to diagnosis (years)	3.85 (2.10, 5.34)	NA (NA, NA)	NA (NA, NA)
Tumour site, No. (%)			
Oral	145 (30)	0 (0)	0 (NaN)
Oropharynx	220 (46)	0 (0)	0 (NaN)
Larynx	85 (18)	0 (0)	0 (NaN)
Hypopharynx	19 (4)	0 (0)	0 (NaN)
Other HNC	11 (2)	0 (0)	0 (NaN)
Overall	**480**	**1,440**	**384,180**

4.4.2 25(OH)D variation by baseline characteristics

25(OH)D showed strong statistical associations with sex, age at recruitment, educational level, smoking status, alcohol frequency, BMI and latitude (Figure 4-3). However, the associations with sex, age at recruitment (per year increase), education and BMI (per unit increase) are of relatively small magnitude. Current smokers had on average approximately 4.6 nmol/L lower 25(OH)D compared to never smokers but Previous smokers had slightly higher 25(OH)D levels. Alcohol frequency showed a positive association with 25(OH)D levels, with those who drink daily or almost daily having 25(OH)D levels approximately 8 nmol/L higher than never drinkers.

For the association with latitude, for each quartile further north the mean serum seasonally adjusted 25(OH)D reduced by 0.54 nmol/L (95% CI: 0.48, 0.60, P=1.89x10^{-72}). Although this is a small effect and the evidence for an association between HNC risk and latitude was weak, latitude was included as a covariate in the full adjusted regression models as it has the potential to confound the vitamin D HNC risk association.

Figure 4-3: Effects of potential confounders on circulating 25(OH)D.



Change in serum 25(OH)D across levels of potential confounders given in nmols/L. For age at recruitment the beta is per year increase, for BMI the beta is per unit increase in BMI, for Latitude the beta is quartile of the variable which goes from the most southerly to the most northerly part of the UK. LCI = lower confidence interval, UCI = upper confidence interval, P = P value. BMI = body mass index.

4.4.3 Association between 25(OH)D and HNC risk

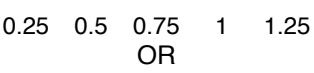
A unit increase in log2(25(OH)D), which corresponds to a doubling in 25(OH)D levels, was associated with a protective effect on all HNC sites combined. Unadjusted conditional logistic regression using three matched controls per case (OR = 0.69 (95% CI: 0.59 to 0.80)) and minimally adjusted unconditional logistic regression (OR = 0.63 (95% CI: 0.56 to 0.71)) showed relatively consistent magnitudes of effect, with the latter having slightly greater precision (Figure 4-4). As all results showed approximate consistency between conditional and unconditional analyses, only the unconditional estimates will be given in the text.

When HNC were considered by site, the largest effect of 25(OH)D was seen for laryngeal cancer (OR = 0.51 (95% CI: 0.39 to 0.65)) with the smallest effect for oropharyngeal cancer (OR = 0.75 (95% CI: 0.61 to 0.91)), the effect on oral cavity cancer was between these two effect sizes (OR = 0.61 (95% CI: 0.50 to 0.76)).

For HNC combined and when sites are considered separately, adjustment for education level, smoking status, alcohol frequency and BMI attenuated the magnitude of effect slightly in both conditional and unconditional logistic regression. In all sites combined and within each site, there was no obvious dose response relationship. The effect of 25(OH)D on HNC risk remained largely consistent across quintiles of 25(OH)D (Appendix 4.1).

Figure 4-4: Odds ratio for developing HNC per log2 increase in 25(OH)D using conditional and unconditional logistic regression.

Site	Regression	Model	OR	LCI	UCI	P
All	Conditional	Unadjusted	0.69	0.59	0.80	0.000
		Adjusted	0.78	0.66	0.92	0.003
	Unconditional	Unadjusted*	0.63	0.56	0.71	0.000
		Adjusted	0.71	0.63	0.81	0.000
Oral	Conditional	Unadjusted	0.64	0.49	0.84	0.001
		Adjusted	0.73	0.54	0.99	0.045
	Unconditional	Unadjusted*	0.61	0.50	0.76	0.000
		Adjusted	0.71	0.57	0.88	0.002
Oropharynx	Conditional	Unadjusted	0.81	0.65	1.02	0.068
		Adjusted	0.82	0.64	1.05	0.109
	Unconditional	Unadjusted*	0.75	0.61	0.91	0.004
		Adjusted	0.80	0.65	0.98	0.035
Larynx	Conditional	Unadjusted	0.60	0.42	0.86	0.005
		Adjusted	0.69	0.42	1.14	0.151
	Unconditional	Unadjusted*	0.51	0.39	0.65	0.000
		Adjusted	0.59	0.45	0.76	0.000

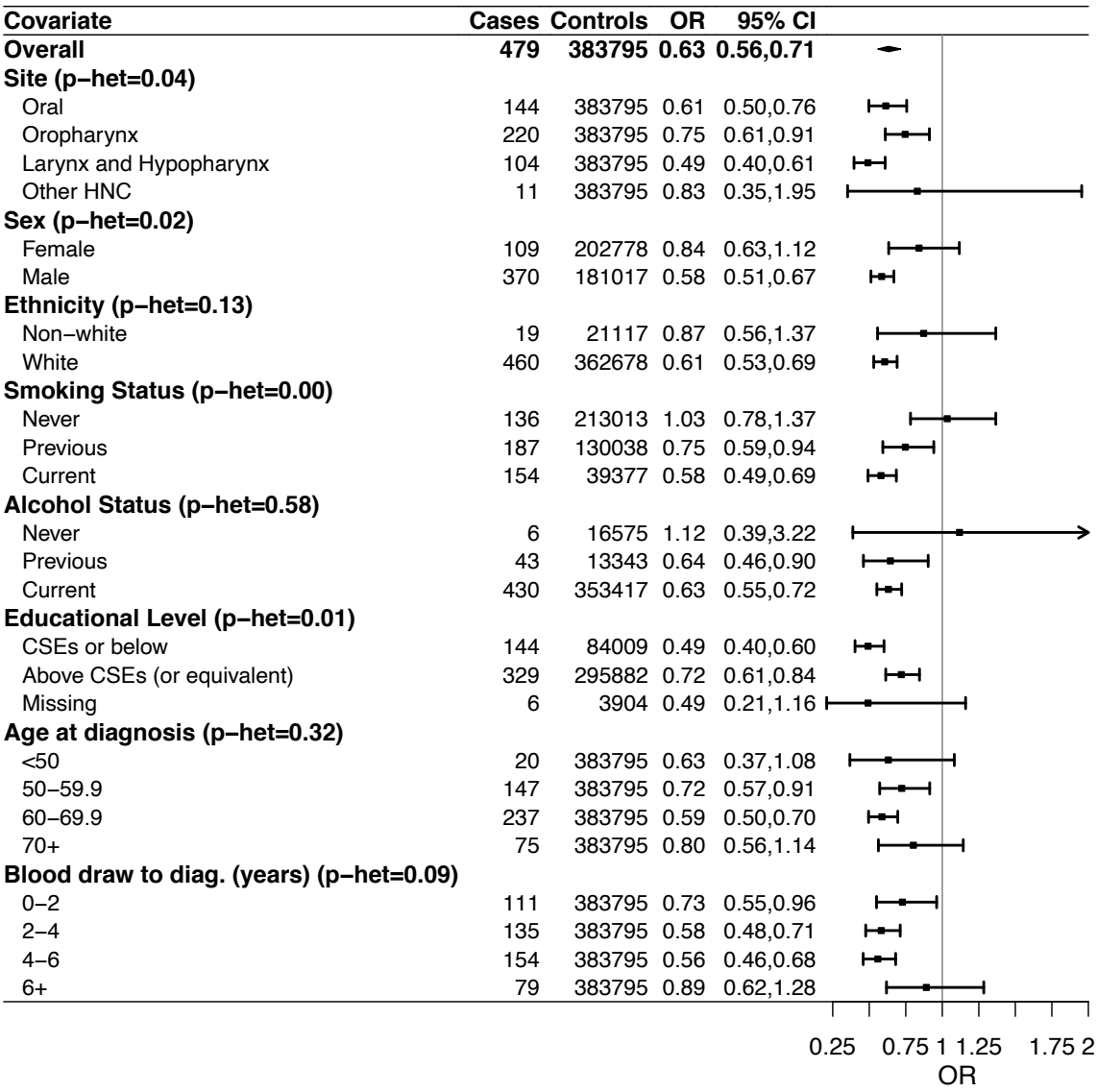


***Unconditional unadjusted model includes age, sex and ethnicity to match the conditional analysis as closely as possible.**

4.4.4 Stratification of results by confounders

When results are stratified by known confounders there was evidence of heterogeneity across strata of smoking status, sex and time from blood draw to diagnosis (Figure 4-5).

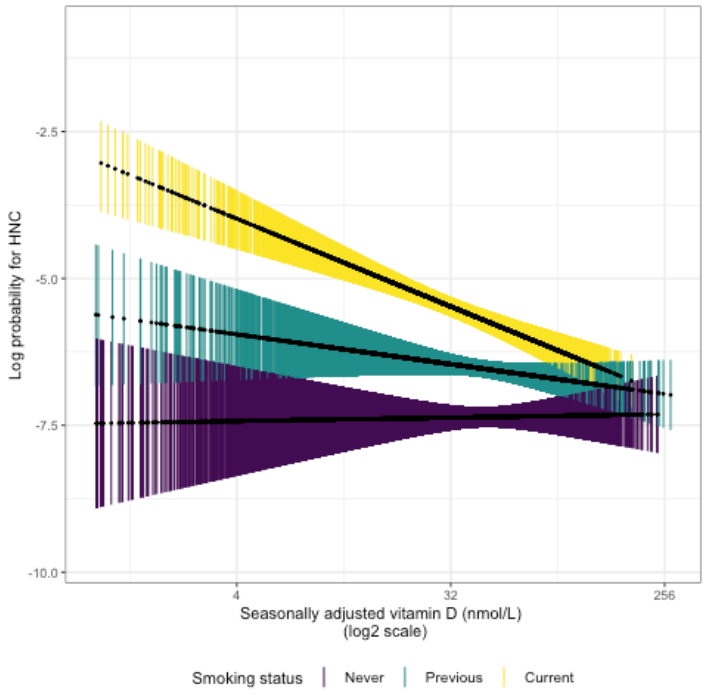
Figure 4-5: Odds ratio for developing HNC per log2 increase in 25(OH)D stratified by site and known confounders.



4.4.4.1 Smoking status

When stratified by smoking, there was strong evidence for heterogeneity between smoking status strata (I^2 = 84.2% (95% CI: 52.9, 94.7), P_{het} = 0.002). A strong protective effect of 25(OH)D was seen in current smokers (OR = 0.58 (95% CI: 0.49 to 0.69) with a smaller effect magnitude in previous smokers (OR = 0.75 (95% CI: 0.59 to 0.94) and no evidence of an effect in never smokers (OR = 1.03 (95% CI: 0.78 to 1.37). Logistic regression including and excluding an interaction term for 25(OH)D with smoking status was conducted and provided strong evidence that including an interaction term provides a better fit in the model (Log-likelihood ratio P = 0.002, χ^2 = 12.38, degrees of freedom = 2). When the interaction term is included, the model is in agreement with the stratified results suggesting no evidence for an effect of 25(OH)D on HNC risk in non-smokers but with a strong effect in current smokers, particularly at low 25(OH)D level (Figure 4-6).

Figure 4-6: The modelled probability of developing HNC within smoking status strata when an interaction between 25(OH)D and smoking status is included in the model.



Each point represents the predicted probability (on the log scale) for each individual given their 25(OH)D level. Vertical error bars represent the 95% confidence interval for the prediction. The model includes an interaction between 25(OH)D and smoking status. A line with horizontal slope suggests no effect of 25(OH)D and a decreasing slope suggests that the probability of HNC decreases as 25(OH)D increases.

4.4.4.2 Sex

Table 4-3: Stratification by sex and smoking status.

	Cases	Controls	OR	95% CI
Smoking Status: Female				
Never	43	122,653	1.35	0.82,2.22
Previous	47	62,166	0.82	0.52,1.27
Current	19	17,271	0.54	0.33,0.90
Smoking Status: Male				
Never	93	90,378	0.90	0.64,1.26
Previous	140	67,872	0.72	0.55,0.95
Current	135	22,106	0.59	0.49,0.70

When stratified by sex, there was some evidence that 25(OH)D had a smaller effect in females compared to males (OR = 0.84 (95% CI: 0.63 to 1.12) versus OR = 0.58 (95% CI: 0.51 to 0.67), I^2 = 80.6%, P_{het} = 0.023). The effect estimate in females was imprecise and there was some overlap of confidence intervals between the male and female estimates. When results are stratified by both smoking status and sex, effect estimates for current and previous smokers are similar in both sexes

but there is a large disparity in the imprecise estimates for non-smokers, between males and females (Table 4-3).

4.4.4.3 Time from blood draw to diagnosis

There was some very weak evidence of heterogeneity across strata of time from blood draw to diagnosis (Figure 4-5). There was not an obvious trend with increasing time from blood draw to diagnosis with the effect of 25(OH)D on HNC smaller at the extremes of time from blood draw.

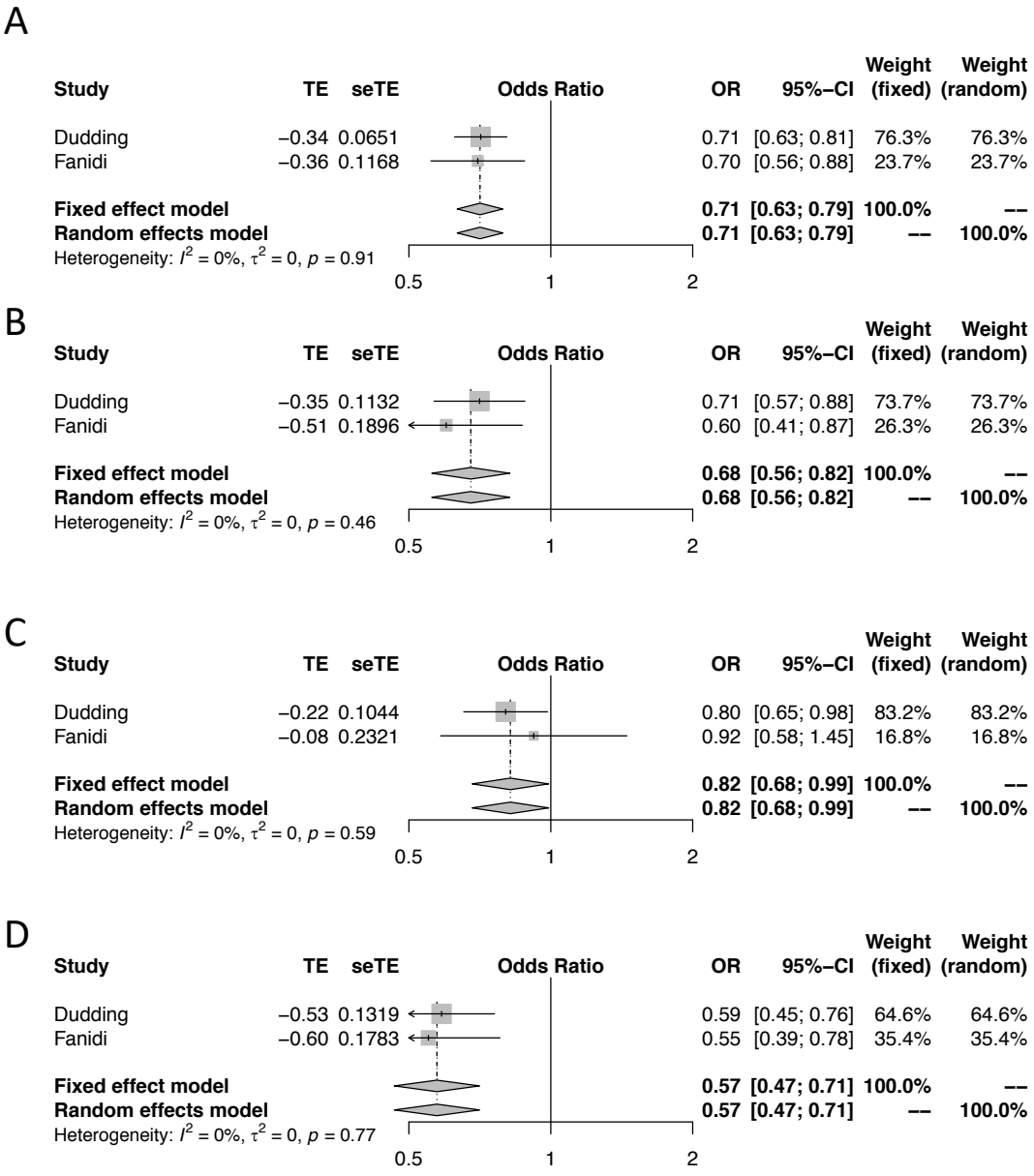
4.4.5 Sensitivity analyses

When the unconditional analyses were repeated generating standard errors robust to geographic clustering of UK Biobank individuals, there was little decrease in the precision of the estimates. For example, the log OR and standard error for the unadjusted unconditional analysis was -0.34 and 0.060 respectively and -0.34 and 0.061 respectively using the clustered analysis method. This small reduction in precision made no difference to the interpretation of the results.

4.4.6 Comparison and meta-analysis of effect sizes with previously published estimates

The estimates in this study are highly consistent with those from Fanidi *et al.* (2016),[126] within cancer site and when all sites are combined. There was no strong evidence of heterogeneity identified however, power to detect heterogeneity was low. In fixed effect meta-analyses, the magnitudes of effect for each cancer site remain similar to those in this study with a modest improvement in precision (Figure 4-7).

Figure 4-7: Forest plots showing the meta-analysis of results with Fanidi *et al.* (2016).



A = All HNC, B = Oral cavity cancer, C = Oropharyngeal cancer, D = Laryngeal cancer.

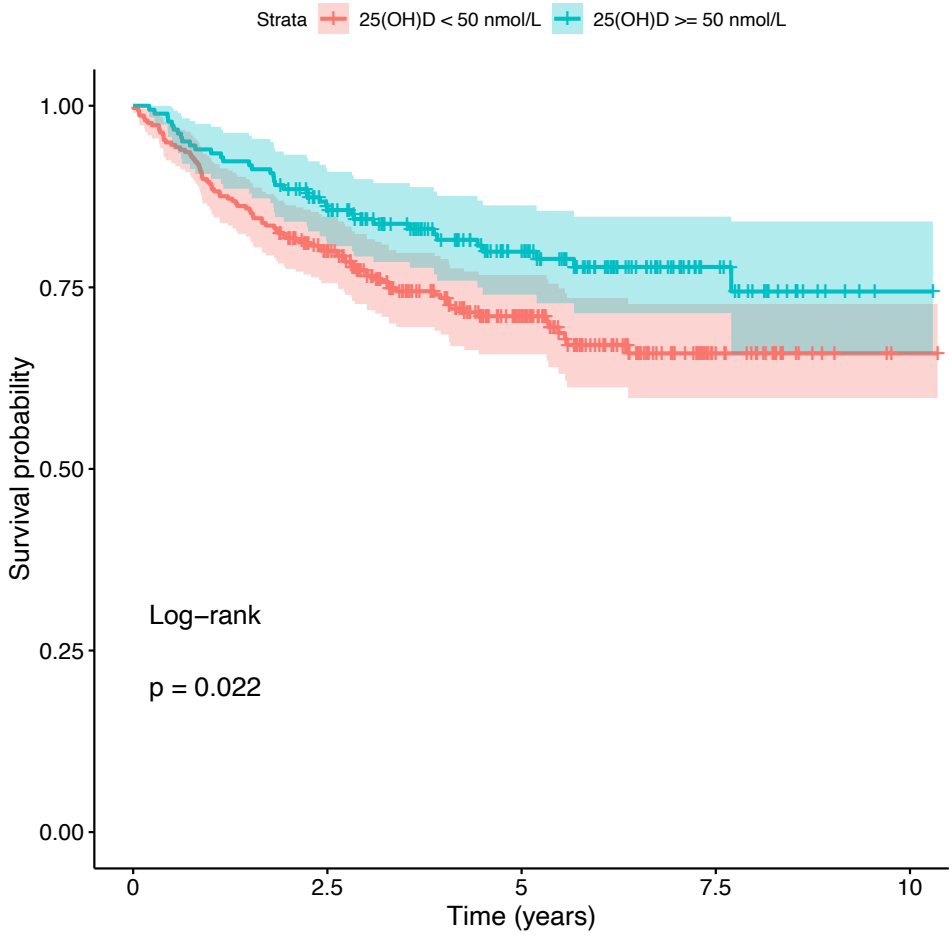
4.4.7 The association between 25(OH)D and HNC prognosis

The mean follow up time of the 480 incident HNCs identified in UK Biobank was 4.3 years (SD=2.3 years). There were 114 (23.8 %) deaths from any cause of which 61.4 % (n=70) were HNC specific deaths, there were nine (1.9%) HNC recurrences resulting in 123 events (deaths or recurrences) in survival analyses.

Of the 480 incident HNCs there were 297 participants in the low 25(OH)D group (mean 25(OH)D = 28.8 nmol/L and 183 participants in the high 25(OH)D group (mean 25(OH)D = 65.9 nmol/L). When

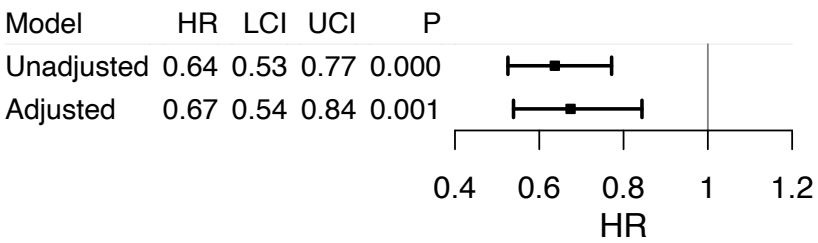
the disease-free survival distributions were compared, there was some evidence of poorer disease-free survival in the low 25(OH)D group (Figure 4-8).

Figure 4-8: Kaplan Meier plots of disease-free survival for low and high 25(OH)D.



In minimally adjusted cox-proportional hazards regression a log2(25(OH)D) increase was associated with a HR of 0.64 (95% CI: 0.53 to 0.77) with minimal attenuation after correcting for smoking, alcohol consumption and BMI (HR = 0.67 (95% CI: 0.54 to 0.84)) (Figure 4-9).

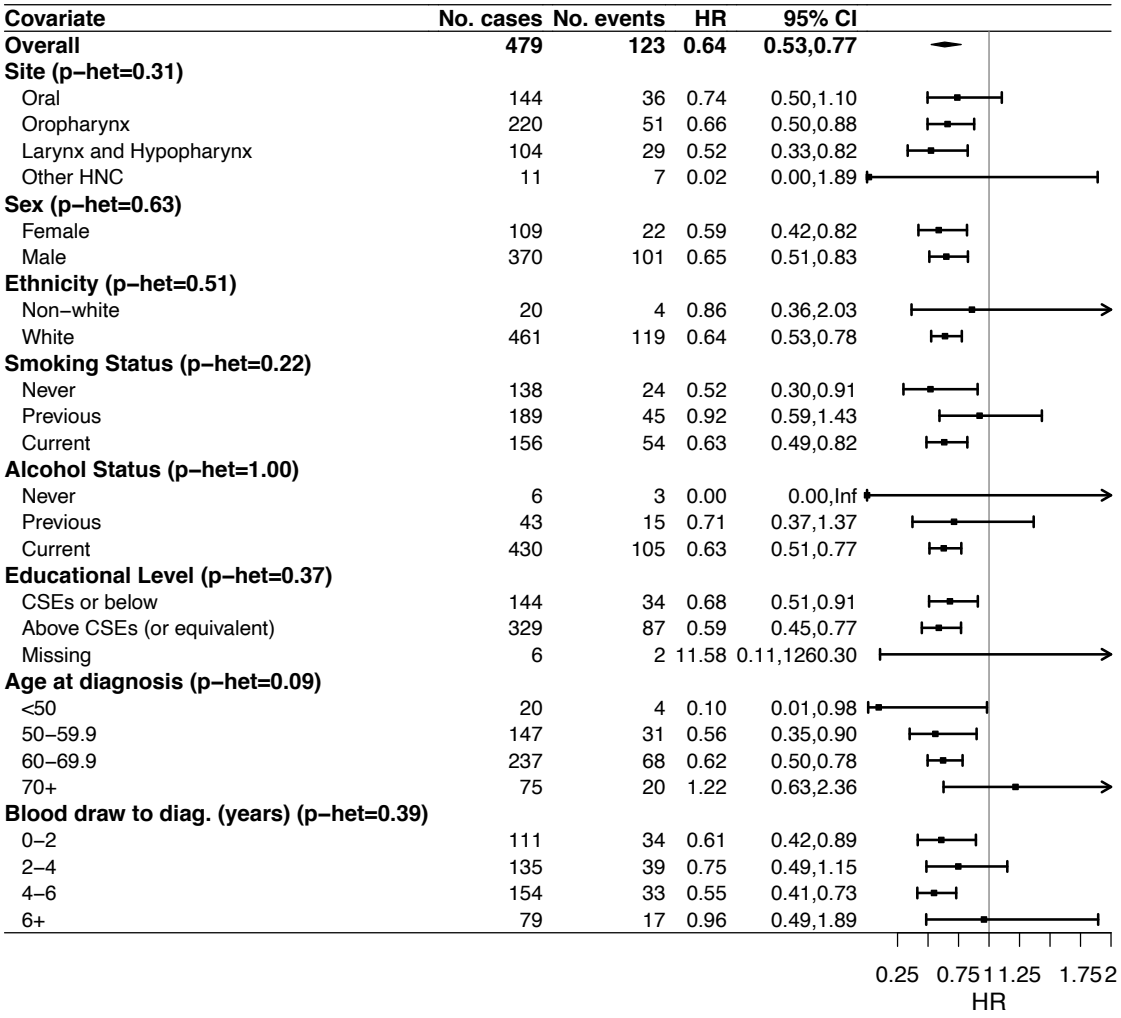
Figure 4-9: Hazard ratio for all-cause mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression.



4.4.8 Stratification of survival analyses

Due to the relatively small number of cases in UK Biobank, there was limited power to identify effects within specific sites or strata of potential confounders. The effect of 25(OH)D on prognosis was similar across the major HNC sites, with the greatest protective effect seen in laryngeal (and hypopharyngeal) cancers. The protective effect in oral cavity cancers was weakest with confidence intervals consistent with a large protective effect or a small detrimental effect on HNC disease free survival. There was very little evidence of heterogeneous effects across strata of potential confounders (Figure 4-10).

Figure 4-10: Hazard ratio for all-cause mortality or recurrence per log2 increase in 25(OH)D stratified by site and known confounders.



4.4.8.1 Sensitivity analyses

When HNC specific mortality only was considered, there was a small attenuation of effect size suggesting there may be a stronger association with non-HNC deaths than HNC related deaths

(Figure 4-11). In analyses removing those who died within six months of diagnosis (55 participants), there was minimal difference when compared to analyses including all deaths (Figure 4-12).

Figure 4-11: Hazard ratio for HNC specific mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression.

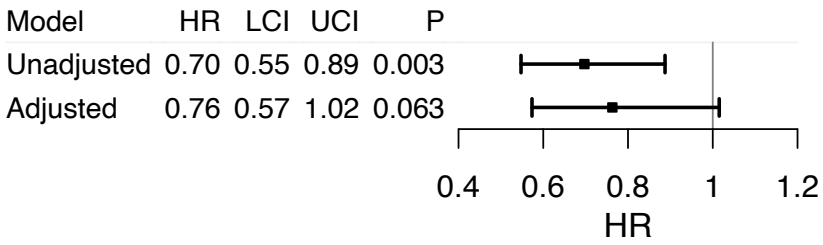
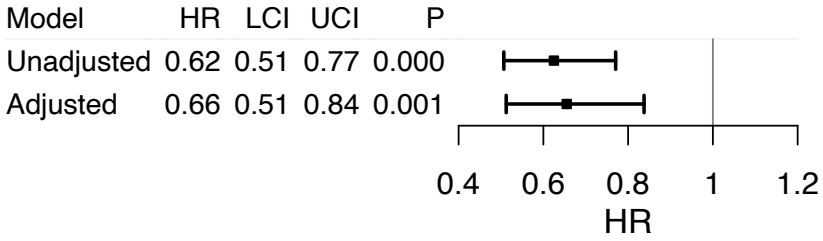


Figure 4-12: Hazard ratio for all-cause mortality or recurrence of HNC per log2 increase in 25(OH)D using Cox proportional hazards regression in participants who survived greater than six months post-diagnosis.



4.5 Discussion

In this chapter a protective effect of 25(OH)D on HNC risk is demonstrated in a prospective cohort study. The largest protective effect of 25(OH)D is seen for laryngeal and oral cavity cancers with weaker evidence for a small effect on oropharyngeal cancers. A protective effect is also seen on HNC prognosis, with higher 25(OH)D resulting in greater disease-free survival. The effects on both risk and prognosis are robust to correction by measured confounders and are similar to previous findings by Fanidi *et al*.[126]

4.5.1 Head and neck cancer risk

The effects of 25(OH)D on oral, laryngeal and oropharyngeal cancer risk identified here are consistent with several studies which have previously reported a protective effect of 25(OH)D on HNC cancers.[124–126] The effect magnitudes for laryngeal and oral cavity cancers are relatively large and clinically relevant, for example nearing a 50% reduction in risk for laryngeal cancer when 25(OH)D levels are doubled. The effect estimate for oropharyngeal cancer is smaller with confidence intervals crossing the null in adjusted conditional analyses. The effect estimates remain after correcting for known confounders.

As detailed in Chapter 2, there is a close relationship between tobacco use, vitamin D and HNC risk. Several previous studies that have identified protective effects of vitamin D on HNC risk, have commented that the effect was either stronger or only seen in smokers. This finding is replicated in this study, with the effect of 25(OH)D in only non-smokers having an OR of 1.0 but with very strong effects seen for smokers.

There are several possible explanations for the heterogeneity in effect size across smoking strata, these have been discussed in general in Chapter 2. The differing effect sizes within the smoking status strata may reflect the largely unbiased situation with smoking acting as an effect modifier on the 25(OH)D – HNC association. This suggests that 25(OH)D may be causally associated with a reduction in risk of HNC, but only in smokers. In the second scenario, the effect could be due to a combination of residual confounding and misclassification of smoking status. In this scenario, correcting for smoking status using current, previous or never smoking may not fully capture the risk attributable to smoking. Furthermore, this misclassification of smoking, and therefore the degree of residual confounding, may increase as smoking increases, resulting in biased estimates in smokers and unbiased estimates in never smokers. The combined and within smoking strata estimates of effect would therefore be biased away from the null (towards a protective effect of 25(OH)D in this case) with the true effect size seen in the non-smokers.

The results presented in this chapter could support smoking being either a confounder or an effect modifier, or even a combination of the two. Similar results were identified in the Fanidi *et al*. study where they had the advantage of cotinine measures as an unbiased measure of smoking status at blood draw. After correcting for cotinine levels, the researchers still saw heterogeneity across smoking strata and concluded that residual confounding was unlikely to fully explain this heterogeneity. Even with cotinine levels however, there is the potential for misclassification of the lifetime effect of smoking. Current smoking intensity (which is captured by cotinine) may not represent the overall risk of HNC attributable to smoking across the life-course and still leave the potential for residual confounding to bias estimates.

In this study the aim is to estimate a causal association, as discussed in Chapter 3 causality is likely to lead to a dose-response relationship between the exposure and outcome. When 25(OH)D was stratified into quintiles, no such dose-response relationship was evident in any of the HNC sites. One explanation is that the association is not causal and the confounding structure is the same across the quintiles of vitamin D. Despite this, the consistency of the results of this study with the Fanidi *et al*. study lends some support to a causal association. It is unlikely that the confounding structure across the two studies is identical and it is also unlikely that the results would by so similar by chance.

However, smoking does have the potential to confound both studies in a similar way and therefore residual confounding could explain the results in both studies. Using meta-analysis to combine these studies does not generate a pooled estimate that is necessarily closer to the true causal estimate. As discussed in Chapter 3, the pooled estimate will include the error of the individual studies and if these are systematically biasing the individual studies in one direction then the pooled estimate will also be biased.

4.5.2 Head and neck cancer prognosis

In this chapter, strong protective effects of pre-diagnosis 25(OH)D on cancer prognosis have been identified. Due to the limited power in stratified analyses it is difficult to comment on how the effect of 25(OH)D may differ across HNC sites or within strata of potential confounders.

This finding supports previous findings identifying a protective effect of pre-diagnosis 25(OH)D on HNC prognosis[126] and two studies that found a protective effect of 25(OH)D levels at diagnosis on HNC prognosis.[131,132] The latter two studies may have been biased by vitamin D levels dropping due to the ongoing but undiagnosed cancer process in the participants. However, in this work, pre-diagnostic 25(OH)D levels were used to mitigate this issue. Furthermore, sensitivity analyses removing participants that were diagnosed within six months of blood draw did not alter the results or their interpretation.

In this chapter we were able to identify participants that were thought to have died of HNC related deaths, as well as those that died from any cause. When analyses were restricted to those that died of HNC specific deaths, the effects of 25(OH)D were smaller and more imprecisely estimated.

Conducting analyses in a case only study has the potential to induce ascertainment or collider bias. This bias will be an issue if 25(OH)D is a cause for HNC, because, in those with HNC, any HNC risk factors will become correlated, potentially inducing spurious relationships. This topic is discussed further in Chapter 8 in the context of observational and MR analyses.

In this study routinely collected registry data was used to identify cancer cases and mortality data in the UK Biobank participants. As cancer and death registry data are collected as part of normal clinical care and not specifically for research purposes they have strengths and weaknesses. Registry data are largely complete and where data are missing it is most likely to be missing at random therefore, missing data will only lead to reduced precision rather than biased estimates. There are also likely to be errors in death registry data, particular relating to the cause of death.[172,173] These errors will have implications on cause-specific survival analyses but should be less influential when investigating all-cause mortality. Finally, registry data are not collected for the purposes of this

specific study and therefore does not have the detail that may be desired if you were to design a bespoke data collection. One example in this study is the lack of cancer staging data from the UK cancer registry. The impact of not having data on stage is discussed below.

Stage of cancer at diagnosis is one of the major determinants of prognosis and if vitamin D is a cause of cancer then it will likely be associated with stage. This relationship between vitamin D and stage can take one of two potential forms, each with its own impact on the survival analyses. Firstly, vitamin D may affect the biology of the cancer and therefore impact prognosis. In this scenario stage at diagnosis can be thought of as an intermediate on the causal pathway between vitamin D and prognosis. Therefore, correcting for stage will lead to estimating only the impact of vitamin D that does not act via its effect on cancer stage. Alternatively, vitamin D may be a good marker of cancer stage, rather than being a causal influence. In this scenario, cancer stage acts as a confounder and estimates that are not corrected for it, such as those in this study, will over-estimate the effect of vitamin D due to confounding by stage. It is unclear which of these scenarios or whether a combination of the two is the most likely.

4.5.3 Conclusions

Although it is not yet clear, there is growing evidence from this work and from Fanidi *et al*.[126] that 25(OH)D may be protective against HNC risk in smokers. If this is the case, there may be the potential for targeting high dose 25(OH)D interventions to smokers at high risk of developing HNC. However, given high profile failures of similar interventions in other cancers, for example vitamin E and selenium in prostate cancer,[174] it would be sensible to seek further evidence of causality before proceeding to clinical trials. Additionally, the protective effect of 25(OH)D on HNC prognosis provides a potential avenue for an adjuvant treatment alongside the primary cancer treatment modality. Arguably, evidence from cell-based and mechanistic studies supports the effect on prognosis more than the effect on risk. Several pre-clinical studies have assessed the effectiveness of potent 25(OH)D analogues on HNC cell lines[175–177] with some success but again further research to assess the true causal association of vitamin D on HNC prognosis is required.

The next chapter

The strong observational associations identified between vitamin D and HNC risk and prognosis, identified here and in previous similar studies, are likely to be affected by residual confounding. There is a need to assess the association using methods which are more robust to confounding such as MR. Effect estimates, from methods such as MR, that are consistent with the observational evidence will add credibility to a causal association existing between vitamin D and HNC risk and or

prognosis. In the next chapter, I conduct (as part of a small team) a study to identify new genetic variants which can proxy vitamin D in MR studies. After identifying these new variants, I combine them with previously identified variants and demonstrate that together these are a largely unconfounded measure of vitamin D and are therefore suitable to use in MR analyses.

Chapter 5 : Identifying vitamin D genetic variants

5.1 Chapter objectives

In the previous chapter it was shown that there is a growing body of evidence from well-designed prospective observational studies that vitamin D may be protective in HNC risk and prognosis. However, despite these well-designed studies, causality has not been established and the associations may be due to residual confounding. MR is a method that can provide causal inference from observation data and requires a robust genetic proxy for vitamin D. In this chapter a genome-wide association study (GWAS) of serum 25(OH)D is described with the aim to better understand the genetic underpinning of vitamin D biology and identify genetic variants which can be used as proxy measures for vitamin D in MR analyses. The findings from this study and similar published studies, are combined to select genetic variants which are robustly associated with 25(OH)D for use as proxy measures of vitamin D in MR analyses. In the following chapters, these proxies are assessed for suitability for use in MR analyses and then used in MR analyses to assess the causal association of vitamin D on HNC risk and prognosis.

The work presented in this chapter was conducted as part of consortium of which I was a major contributor to the study design, analysis planning, analysis and final manuscript. In terms of analysis, I prepared the phenotype and conducted the GWAS analysis of the whole genome sequenced and imputed ALSPAC datasets. For this thesis, I additionally conducted all analyses presented here which were conducted on the ALSPAC cohort such as the haplotype analyses. Other aspects of the analysis, for which I was involved in the planning but not the actual conduct, such as the meta-analysis of individual studies, is presented here for completeness.

5.2 Introduction

Instrumental variable analyses such as MR rely on proxy variables for the exposure of interest. In this thesis, genetic variants associated with vitamin D are ideal as they are unrelated to factors that may confound the association between vitamin D and HNC risk or prognosis. The more variability in the exposure that is explained by the genetic proxy the greater power there will be in MR analyses. It is therefore of benefit to identify novel genetic variants for vitamin D which would increase the proportion of variance explained by the vitamin D proxy measure.

As detailed in Chapter 2, approximately half of the variability in the concentration of the widely accepted vitamin D biomarker (25(OH)D), has been attributed to genetic factors in twin and family

studies.[72,77] Four common genetic variants (minor allele frequency (MAF) >5%) in loci near four genes known to be involved in cholesterol synthesis (*DHCR7*), hydroxylation (*CYP2R1*), vitamin D transport (*GC*) and catabolism (*CYP24A1*) are strongly associated with 25(OH)D levels, yet explain little of its heritability.[178] Low-frequency and rare genetic variants (defined as variants with a MAF of ≤5% and ≤1% respectively) have recently been found to have large effects on clinically relevant traits[179–181] providing an opportunity to better understand the biologic mechanisms influencing disease susceptibility in the general population.

The objective of this study was to detect genetic variants with as yet undiscovered effects on 25(OH)D levels through a large-scale meta-analysis. Unlike the previous GWAS of 25(OH)D,[178] this study attempted to also identify low-frequency and rare variants with large effects on 25(OH)D. This has only recently been made possible through whole-genome sequencing and the use of improved genotype imputation for low frequency and rare variants, with the recent availability of large whole genome sequencing reference panels.[157]

5.3 Methods

5.3.1 Overview

To achieve the objective of identifying novel genetic variants for vitamin D an association study of whole-genome sequence data and deeply imputed genome-wide genotypes was conducted to identify novel genetic determinants of serum 25(OH)D in 42,274 individuals. The study combined data from 17 studies by meta-analysis and also combined whole genome sequenced and imputed genetic data to identify variants which are rarer or have smaller effects than the previous studies.

5.3.2 Studies

All human studies were approved by each respective institutional or national ethics review committees, and all participants provided written informed consent. To investigate the role of rare and low-frequency genetic variation on 25(OH)D levels in individuals of European descent, whole genome sequencing (WGS) data at mean read depth of 6.7x in 2,619 subjects from two cohorts in the UK10K project[182] with available 25(OH)D phenotypes were used (Table 5-1). Imputation reference panels were also used to impute variants that were missing, or poorly captured, from previous GWAS in 39,655 subjects (Table 5-1 and Figure 5-1). The participating individuals were drawn from independent cohorts of individuals of European descent. Detailed description of each of the participating studies is provided in the Appendix (Appendix 5.1).

Figure 5-1: Schematic of the discovery single variant meta-analysis

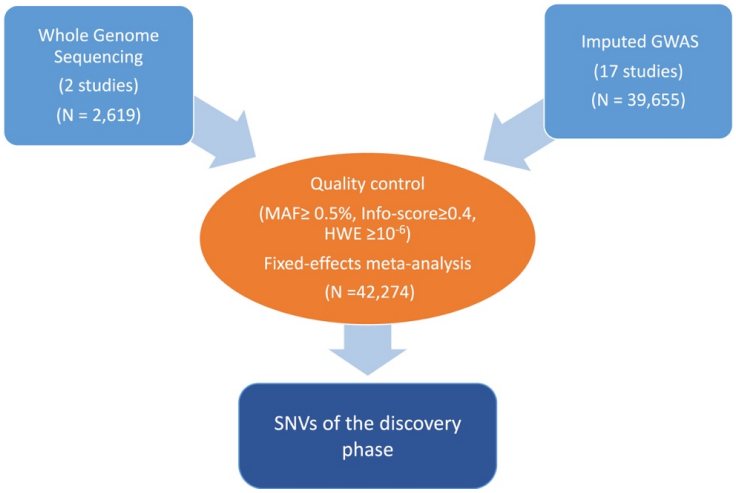


Table 5-1: Participating cohorts and number of DNA samples per cohort

Study Acronym*	Imputed	WGS
ALSPAC	3,679	1,606
TUK	1,919	1,013
Generation R	1,442	
BPROOF	2,514	
FHS	5,402	
MrOS	3,265	
RSI	3,320	
RSII	2,022	
RSIII	2,913	
CHS	1,792	
BMDCS	863	
MrOS GBG	945	
GOOD	921	
MrOS Malmo	893	
PIVUS	943	
ULSAM	1,095	
NEO	5,727	
TOTAL = 42,274	**39,655**	**2,619**

***For full names of the studies see Appendix (Appendix 5.1), WGS: Whole-Genome Sequenced**

5.3.3 Measurement of serum 25(OH)D

The methods applied to measure 25(OH)D levels differed among the participating cohorts (Appendix 5.1). The four methods used were tandem mass spectrometry (in BMDCS, MrOS and BPROOF), combined high-performance liquid chromatography with mass spectrometry (in ALSPAC, BPROOF, CHS, ULSAM, NEO, Generation R), chemiluminescence immunoassay (DiaSorin, Inc, Stillwater, MN) (in TUK, PIVUS, FHS, MrOS Malmo, MrOS GBG and GOOD) and an electrochemiluminescence

immunoassay (COBAS, Roche Diagnostics GmbH) (in RSI, RSII and RSIII). Detection limits for the different methods are provided in the Appendix (Appendix 5.1).

5.3.4 Whole-genome sequencing, genotyping and imputation

ALSPAC WGS and TUK WGS cohorts had been sequenced at an average read depth of 6.7x through the UK10K consortium (www.UK10K.org) using the Illumina HiSeq platform, and aligned to the GRCh37 human reference using Burrows-Wheeler Aligner (BWA)31.[183] Single-nucleotide variant (SNV) calls were completed using samtools/bcftools,[184] and VQSR[185] and GATK were used to recall these variants. The whole genome sequencing for the ALSPAC and TwinsUK cohorts has been described in detail in a previous publication.[180] Appendix 5.3 summarises the data generation method for sequencing-based cohorts.

Participating studies separately genotyped samples and imputed them to WGS-based reference panels. The most recent imputation panels, such as the UK10K and 1000Genomes Project (v3) combined panel, which in total contained 7,562 haplotypes from the UK10K Project and 2,184 haplotypes from the 1000 Genomes Project,[157] and the Haplotype Reference Consortium (HRC) panel, with 64,976 haplotypes,[186] enabled more accurate imputation of low frequency variants, when compared to the UK10K or the 1000Genomes reference panel alone.[157] Specifically, 11 out of the 17 participating cohorts were imputed to the UK10K and 1000 Genomes reference panel (total number of imputed individuals included in the meta-analysis N=25,589). Three of the participating cohorts were imputed using the HRC panel (total number of imputed individuals N=5,717). Finally, two cohorts were imputed to the 1000Genomes panel (N=7,536), and one cohort was imputed to the UK10K panel (N=863). (Appendix 5.1). Details on genotyping methods and imputation for the 17 participating cohorts are presented in Appendix 5.2. Info scores for the imputed SNVs per participating cohort are presented in Appendix 5.4.

5.3.5 Statistical analysis

A GWAS was conducted separately by each cohort using an additive genetic model for 25(OH)D levels. Because 25(OH)D concentrations were measured using different methods, log-transformed 25(OH)D levels were standardized to z-scores, after being adjusted for age, sex, BMI, and season of measurement. Specifically, the phenotype for each GWAS study was prepared according to the following steps:

1. 25(OH)D levels were log-transformed to ensure normality
2. Linear regression models were used to generate cohort-specific residuals of log transformed 25(OH)D levels adjusted for covariates (age, sex, BMI and season). Season was treated as a

non-ordinal categorical variable (summer: July to September, fall: October to December, winter: January to March, and spring: April to June)

3. The mean of log transformed 25(OH)D levels was added to the residuals to create the adjusted 25(OH)D phenotype.
4. The above phenotype was then normalized within each cohort (mean of zero with SD of one) to make the phenotype consistent across cohorts.
5. Finally, outliers beyond 5 standard deviations were removed from step 4.

For comparison purposes, 25(OH)D levels, adjusted for age, sex, BMI and season of measurement, were calculated in one cohort of the meta-analysis (ALSPAC WGS) in carriers and non-carriers of the lead SNV(s).

The software used by each cohort to perform a GWAS is listed in Appendix 5.1. Single variant tests were undertaken for variants with MAF>0.1%, using an additive effect of the minor allele at each variant in each cohort. The type of software employed for single variant testing for each cohort is shown in Appendix 5.1. Studies with related individuals used software that accounted for relatedness. The mean lambda was 1.015, cohort-specific genomic inflation factors (lambdas) are shown in Appendix 5.1.

The association results from all discovery cohorts were meta-analysed (N total = 42,274). This stage included validation of results file format, filtering files by the above QC criteria, comparison of trait distributions among different studies, identification of potential biases (large betas and/or standard errors, inconsistent effect allele frequencies, extreme lambdas). Meta-analysis quality control of the GWAS data included the following SNV-level exclusion criteria: i) Info score <0.4, ii) HWE P-value $<10^{-6}$ iii) Missingness >0.05, and iv) MAF <0.5%. Alignment of the SNVs across studies was done using the chromosome and position information for each variant according to genome build hg19. SNVs in the X chromosome were not included in the meta-analysis. Fixed–effects meta-analysis was performed using the software package GWAMA[187] adjusting for genomic control. Bi-allelic SNVs with MAF ≥ 0.5% were tested for association, declaring genome-wide statistical significance at $P \leq 1.2 \times 10^{-8}$ for variants present in more than one study. This stringent p-value threshold was set to adjust for all independent SNVs above the MAF threshold of 0.5%.[188]

Conditional analysis was undertaken for the four previously described lead vitamin D SNVs from the SUNLIGHT consortium using the GCTA package.[189] This method uses an approximate conditional analysis approach from summary-level statistics from the meta-analysis and linkage disequilibrium corrections between SNVs estimated from a reference sample. The UK10K individuals were used as the reference sample to calculate the linkage disequilibrium information of SNVs. The associated

regions flanking within 400kb of the top SNVs from SUNLIGHT were extracted and the conditional analyses were conducted within these regions. Conditional analyses of individual variants were conducted using GCTA v 0.93.9 using default parameters.

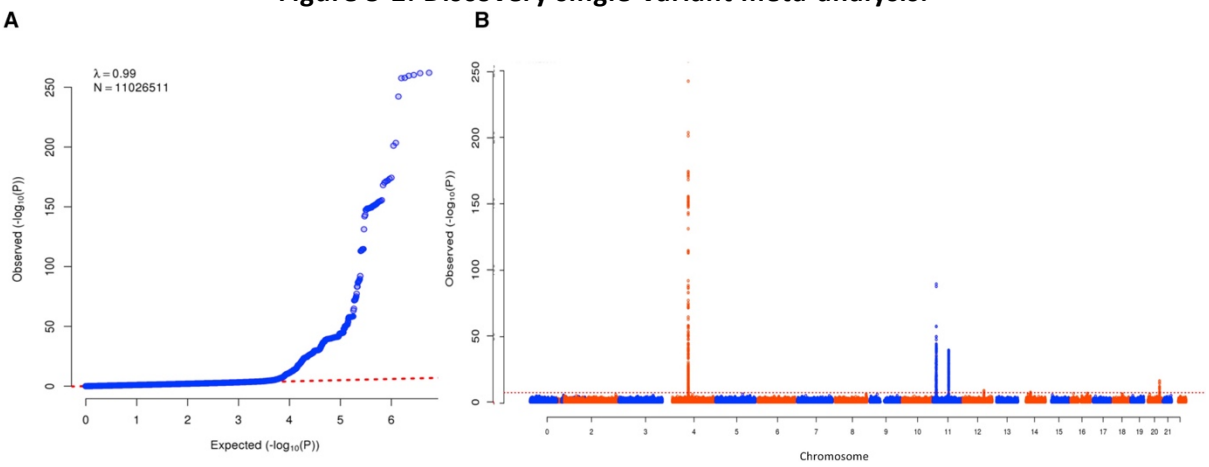
Haplotype block analyses were used for the candidate variants of interest by deriving phased haplotypes from 1,060 individuals from the ALSPAC WGS cohort using adjusted but untransformed 25(OH)D values.

5.4 Results

5.4.1 Low-frequency 25(OH)D-associated SNVs

After strict quality control, the genomic inflation factor for the meta-analysis of 19 GWAS studies was 0.99, suggesting lack of bias due to population stratification (Figure 5-2). Through meta-analysis of 11,026,511 sequenced and imputed variants from our discovery cohorts (Table 5-1), we identified a signal at the chromosome 11p.15.2 locus, harbouring variants associated with 25(OH)D levels (lead low-frequency SNV p.Asp120Asp [rs117913124(A)], MAF = 2.5%, allelic effect size = -0.43 standard deviations of the standardized log-transformed 25(OH)D levels [SD], P = 1.5×10^{-88}, Figure 5-3). The direction of effect was consistent across all discovery cohorts (Table 5-2 and Figure 5-3) and the mean imputation information score for the imputed studies was 0.97. This low-frequency synonymous coding variant is in exon 4 of the *CYP2R1* and is ~14 kb from the previously identified common *CYP2R1* variant, rs10741657 (r2 between these two SNVs= 0.03) (Figure 5-2). The rs117913124 SNV has not previously been reported to be associated with any vitamin D-related traits in humans.

Figure 5-2: Discovery single-variant meta-analysis.



A. Quantile-quantile plot for the single SNV meta-analysis. B. Manhattan plot of the meta-analysis. The plot depicts variants with MAF > 0.5% across the 22 autosomes against the –log10 p-value from the meta-analysis of 19 cohorts, which included 42,274 individuals.

Figure 5-3: Forest plot detailing the per allele effect of rs117913124 on standardised log25(OH)D by cohort.

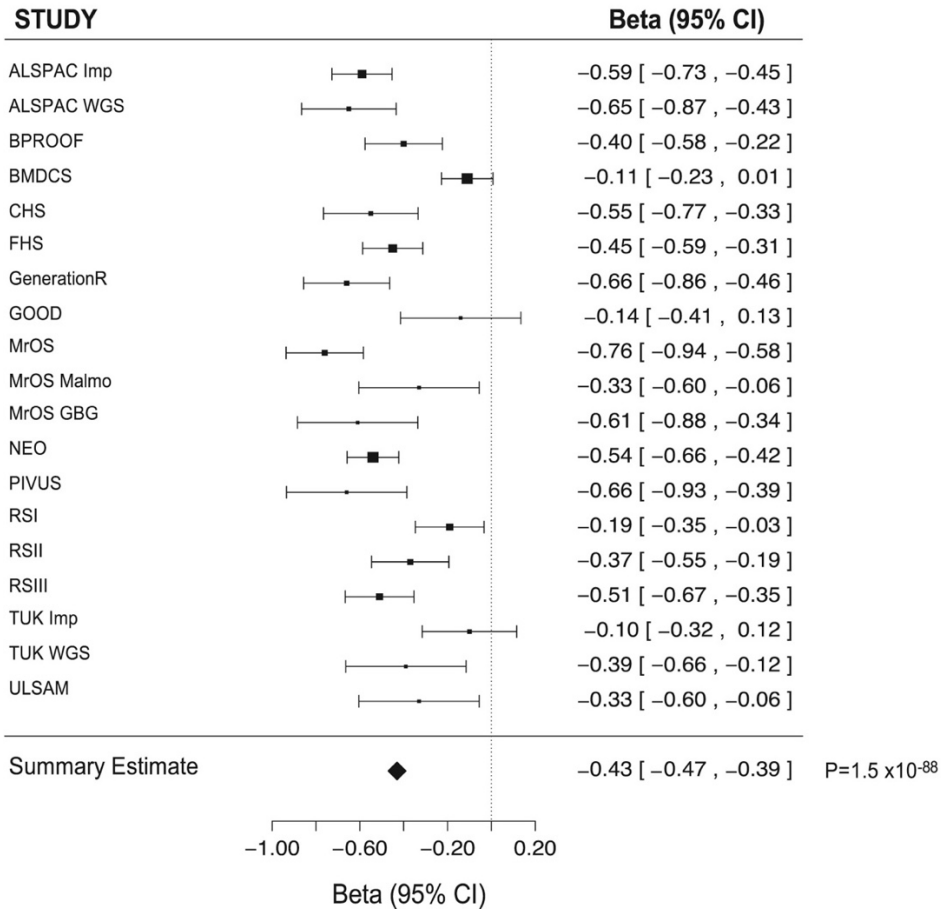


Table 5-2: Summary statistics results for the *CYP2R1* low-frequency variant, rs117913124, from 19 studies

STUDY	25OHD measurement method[#]	N	Effect Allele A[*] Frequency	Beta[$]	Standard Error	P-value	Information score
ALSPAC Imputed	MS	3675	0.028	-0.59	0.07	$3.43x10^{-18}$	0.99
ALSPAC WGS	MS	1606	0.028	-0.65	0.11	$8.23x10^{-10}$	NA
BPROOF	MS	2512	0.027	-0.4	0.09	$4.99x10^{-6}$	0.97
BMDCS	MS	863	0.019	-0.11	0.06	0.058	0.98
CHS	MS	1581	0.022	-0.55	0.11	$5.15x10^{-7}$	0.88
FHS	CLIA	5402	0.021	-0.45	0.07	$2.32x10^{-10}$	0.97
GenerationR	MS	1442	0.033	-0.66	0.1	$1.78x10^{-6}$	1
GOOD	CLIA	921	0.028	-0.14	0.14	0.31	0.96
MrOS	MS	3265	0.018	-0.76	0.09	$5.63x10^{-16}$	0.96
MrOS Malmo	CLIA	893	0.033	-0.33	0.14	0.016	0.94
MrOS GBG	CLIA	945	0.026	-0.61	0.14	$7.87x10^{-6}$	1
NEO	MS	5727	0.025	-0.54	0.06	$2.73x10^{-19}$	1
PIVUS	CLIA	943	0.028	-0.66	0.14	$2.56x10^{-6}$	0.99
RSI	ECLIA	3320	0.025	-0.19	0.08	0.019	0.98
RSII	ECLIA	2022	0.033	-0.37	0.09	$2.38x10^{-5}$	0.99
RSIII	ECLIA	2913	0.027	-0.51	0.08	$4.61x10^{-10}$	0.98
TUK Imputed	CLIA	1919	0.021	-0.1	0.11	0.35	0.98
TUK WGS	CLIA	1013	0.025	-0.39	0.14	0.006	NA
ULSAM	MS	1095	0.025	-0.33	0.14	0.02	1

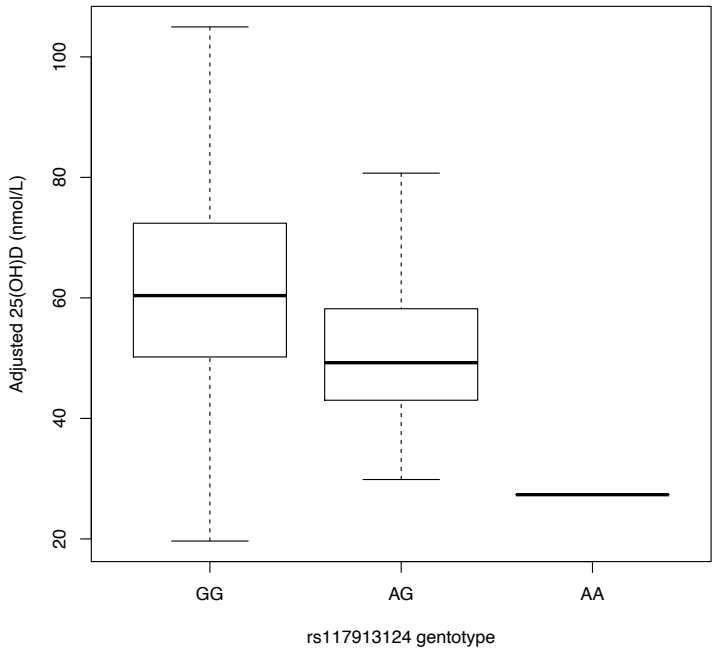
*** Effect allele is the 25OHD decreasing allele**

 # MS: mass spectrometry, CLIA: chemiluminescence immunoassay, ECLIA: electrochemiluminescence immunoassay

$ Betas represent changes in standard deviations of the standardized log-transformed 25OHD level

A comparison of the 25(OH)D levels, adjusted for age, sex, BMI and season of measurement, in non-carriers, heterozygote carriers and homozygous carriers of the A allele of rs117913124 in the ALSPAC WGS was performed. The average 25(OH)D levels, adjusted for age, sex, BMI and season of measurement were computed in 1,606 individuals from the ALSPAC WGS cohort, among which 1,516 were not carriers, 89 were heterozygote carriers of the A allele and 1 individual was a homozygous carrier of the A allele of rs117913124. A linear regression model with the adjusted 25(OH)D levels as the dependent variable and the dose of the "A" allele of rs117913124 (numeric factor 2, 1 or 0) as the independent variable demonstrated a 11.1 nmol/L decrease in the adjusted 25(OH)D levels per "A" allele. The mean adjusted 25(OH)D levels were 61.9 nmol/L in non-carriers vs 51.1 nmol/L in heterozygote carriers vs 27.3 nmol/L in the homozygous carrier (Figure 5-4).

Figure 5-4: Comparison of 25(OH)D levels in homozygous (AA), heterozygote (AG) and non-carriers (GG) of the rs117913124 variant in the ALSPAC WGS participants.



25(OH)D levels are adjusted for age, sex, BMI and season of measurement. Thick lines represent the median, boxes represent the interquartile range (IQR) and the whiskers indicate +/- 1.5*IQR from the edge of the IQR box.

Two-way conditional analysis between the *CYP2R1* common (rs10741657) and low-frequency (rs117913124) variants revealed that the two association signals are statistically independent. Specifically, after conditioning on rs10741657, rs117913124 remained strongly associated with 25(OH)D level (P_{cond} = 2.4x10^{-78}); after conditioning on rs11791324, the effect of rs10741657 on

25(OH)D level remained significant (P_{cond}= 4.0 x10^{-33} versus $P_{pre\text{-}cond}$= 8.8 X 10^{-45}) (Table 5-3). Furthermore, no other low frequency variant in the region remained significant when conditioning on rs117913124 (Table 5-3). To further disentangle the role of rs117913124 from rs10741657 on 25(OH)D levels, haplotype analysis based on WGS data from 1,606 individuals from the ALSPAC WGS cohort was conducted. It was found that the 25(OH)D decreasing allele A of rs117913124 was rarely transmitted in the same haplotype block with the 25(OH)D increasing allele A of the common *CYP2R1* variant rs10741657 (haplotype frequency = 0.1%). The 25(OH)D levels among carriers of the various haplotype blocks was compared. This showed evidence of a decrease in the 25(OH)D levels in carriers of the A allele of the rs117913124 compared to non-carriers independent of the presence of the effect allele G of the common *CYP2R1* variant (Table 5-4).

Table 5-3: Association results for genome-wide significant low-frequency variants from discovery 25(OH)D meta-analysis, before and after conditioning on the lead common *CYP2R1* SNP, rs10741657, and the lead low-frequency *CYP2R1* variant, rs117913124.

Genetic variant	Chr	Position	EA*	EAF	Candidate Gene	Beta[$]	P-value	Beta[$]	P -value	Beta[$]	P-value	N
								Conditional on rs10741657		Conditional on rs117913124		
rs117913124	11	14900931	A	0.025	*CYP2R1*	-0.43	1.5×10^{-88}	-0.39	2.4×10^{-78}	NA	NA	41336
rs116970203		14876718	A	0.025	*CYP2R1*	-0.43	2.2×10^{-90}	-0.40	3.3×10^{-80}	NA	NA	41138
rs117361591		14861957	T	0.014	*(nearest*	-0.44	9.1×10^{-51}	-0.40	2.2×10^{-44}	-0.05	0.017	38286
rs117621176		14861320	G	0.014	*gene: PDE3B)*	-0.44	8.7×10^{-51}	-0.40	2.1×10^{-44}	-0.05	0.016	38273
rs142830933		14838760	C	0.014		-0.44	1.4×10^{-48}	-0.40	1.7×10^{-42}	-0.05	0.03	37541
rs117672174		14746404	T	0.014		-0.43	2.8×10^{-45}	-0.39	2.9×10^{-39}	-0.04	0.062	37209

***Effect allele is the 25OHD decreasing allele**

[$] Betas represent changes in standard deviations of the standardized log-transformed 25OHD levels

Chr = chromosome, EA = effect allele, EAF = effecta allele frequency

Table 5-4: Effect of different haplotype combinations of the low frequency (rs117913124) and the common (rs10741657) *CYP2R1* variants on 25OHD levels.

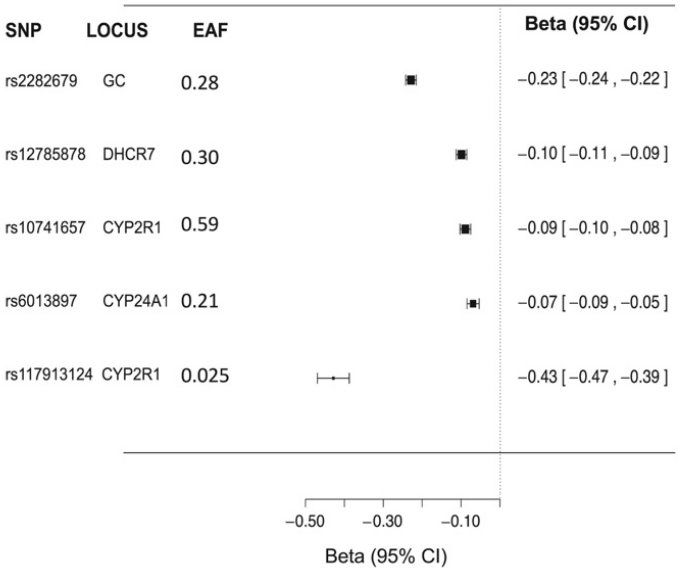
Haplotype		Beta[$]	P-value	N (%)
rare/common*	**rare/common***			**TOTAL = 1,606**
GA	GA	**REF** (Mean = 63.51 nmol/L)		254 (15.8)
AA	GG	-13.03	2×10^{-5}	37 (2.3)
AG	GG	-12.02	5.8×10^{-6}	52 (3.2)
AG	AG	-36.17	0.038	1 (0.1)
GA	GG	-0.51	0.688	718 (44.7)
GG	GG	-3.93	0.003	544 (33.9)

*** The first allele in each chromatid corresponds to the low-frequency variant rs117913124; the second allele corresponds to the common variant rs10741657. The 25(OH)D decreasing allele of the low-frequency variant is A, the 25OHD decreasing allele of the common variant is G.**

[$] Beta values represent changes in the untransformed 25OHD levels from the mean value in the GA GA haplotype.

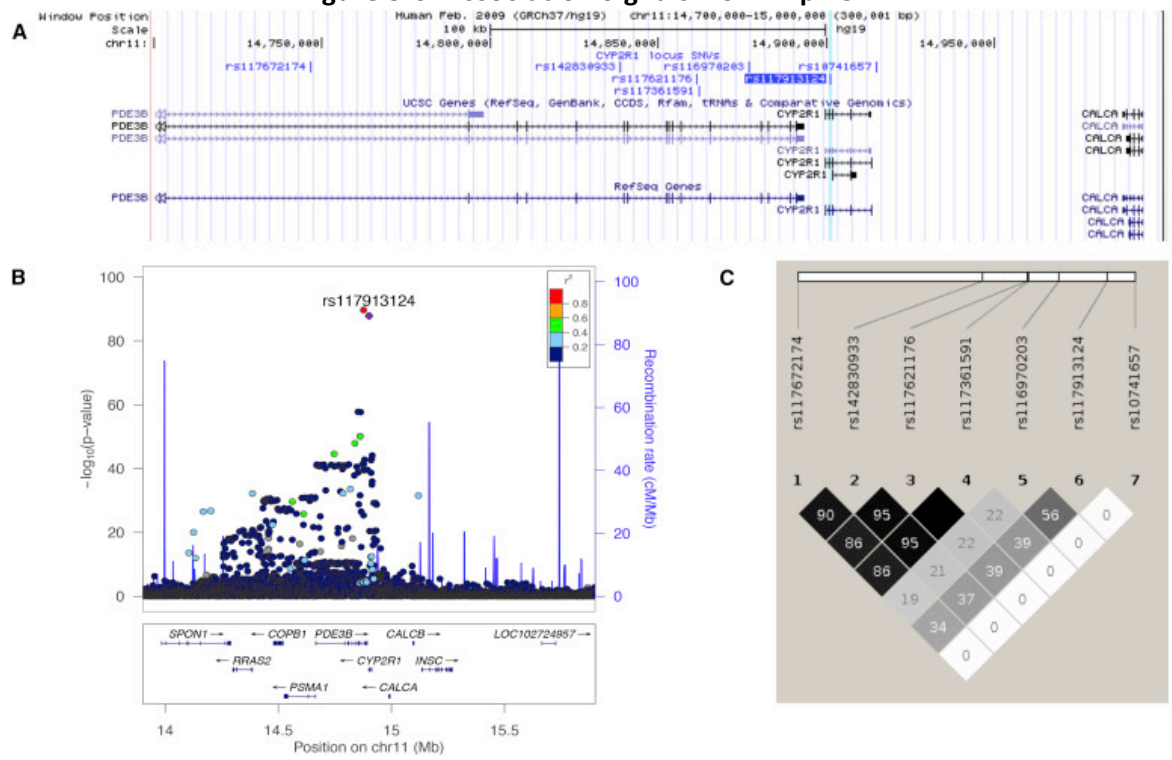
No other low-frequency or rare variants were identified in the three previously described vitamin D-related loci at *DHCR7*, *GC* and *CYP24A1*. The mean effect size of the four previously reported common genome-wide significant SNVs (MAF ≥ 5%) from the SUNLIGHT consortium was -0.13 SD and the largest effect size was -0.25 SD (for the *GC* variant) in this meta-analysis (Figure 5-5 and Appendix 5.5). The effect size of rs10741657(G), the known common *CYP2R1* variant, was -0.09 SD. Hence, the observed effect size of rs117913124 is 3-fold larger than the above-mentioned mean, 4-fold larger that of the common *CYP2R1* variant and almost twice that of the largest previously reported effect of the *GC* variant. Last, the percentage of the variance of the 25(OH)D phenotype explained by the low-frequency *CYP2R1* variant was more than double than the percentage of the variance explained by the *CYP2R1* common variant (0.9% vs 0.4%).

Figure 5-5: Forest-plot of the effect of the four common SUNLIGHT variants and of the *CYP2R1* low-frequency variant rs117913124 on standardised log-transformed 25(OH)D levels.



Eighteen genome-wide significant low-frequency and rare SNVs were also identified on the same chromosome 11 region as rs117914124. Six of these were located in the neighbouring *PDE3B* gene and had consistent effect direction in all 19 cohorts (Table 5-3, Figure 5-6B). Signals from these SNVs in *PDE3B* were independent of the common variant at *CYP2R1* (Table 5-3). Haplotype blocks with rs117913124 and SNVs at *PDE3B* were generated based on haplotype information from the 1,606 individuals from the ALSPAC WGS cohorts (Appendix 5.6) It was found that the 25(OH)D decreasing allele (A) of the rs117913124 was inherited with the 25(OH)D decreasing allele (A) of its perfect proxy rs116970203 (r^2=1) in all participants except one. Therefore, rs116970203 is not likely to have a distinct effect from rs117913124 on 25(OH)D levels. On the other hand, the 25(OH)D decreasing alleles of the remaining four low-frequency variants (all having a MAF of approximately 1.4%) were not always inherited in the same haplotype block as the rs117913124 and rs116970203 and were in moderate linkage disequilibrium with the rs117913124 (all r^2< 0.6, Figure 4B and Figure 4C). Each of the four alleles is in almost perfect linkage disequilibrium with the remaining three (all r^2 >0.96). This implied that these four SNVs might influence 25(OH)D levels independently of the rs117913124. Nevertheless, as mentioned above, when conditioning on the lead low-frequency *CYP2R1* SNV rs117913124, the P-values of the four *PDE3B* SNVs became non-significant and their betas decreased substantially (Table 5-3), demonstrating that they likely do not represent an independent signal at the chromosome 11 locus.

Figure 5-6: Association Signals from 11p.15.2



A. Snapshot from the UCSC genome-browser including the top low-frequency SNVs (see Table 5-2) and the lead common variant rs10741657 at the CYP2R1 locus. The position of rs117913124 is highlighted in light blue. B. Regional disequilibrium plot showing the rs117913124 (purple dot), its perfect proxy rs11670203 (red dot) and the other genome-wide significant SNVs in the same locus (blue and green dots). The plot depicts SNVs within 1 Mb of a locus' lead SNV (x-axis) and their associated meta-analysis p value (-log10). SNVs are color coded according to r^2 with the lead SNV (labelled, r2 calculated from UK10K whole genome sequencing dataset). Recombination rate (blue line), and the position of genes, their exons and the direction of transcription are also displayed (below plot). C. Linkage disequilibrium plot indicating the r^2 values between the SNVs of Table 5-2 (top low-frequency variants) and between these low-frequency SNVs and the lead common variant (rs107416570) at the same CYP2R1 locus (r^2 calculated from the 1000 Genomes dataset).

5.4.2 Common 25(OH)D-associated SNVs

Two additional novel loci associated with 25(OH)D levels (Table 5-5) were identified which passed the genome-wide significance threshold. Variants leading these associations were common and exerted a rather small effect on 25(OH)D: first, a variant in chromosome 12 (rs3819817[C], intronic to *HAL*), with a MAF of 45%, a beta of 0.04 and a P-value of 3.2×10^{-10}. Second, a variant in chromosome 14 (rs2277458[G], intronic to *GEMIN2*), with a MAF of 21%, a beta of -0.05 and a P-value of 6.0×10^{-9}. Both variants were present in all 19 studies and with consistent direction of effect. Neither the *HAL* nor the *GEMIN2* loci are previously known to be associated with 25(OH)D levels. Of note, neither variant was present in the HapMap imputation reference used in the SUNLIGHT study.

Table 5-5: Main novel findings of the GWAS meta-analysis

SNP	Chr	Candidate Gene	Effect allele	Effect allele frequency	Beta[$]	P-value	N
rs117913124	11	*CYP2R1*	A	0.025	-0.43	1.5×10^{-88}	41,336
rs3819817	12	*HAL*	C	0.45	0.04	3.2×10^{-10}	41,071
rs2277458	14	*GEMIN2*	G	0.21	-0.05	6.0×10^{-09}	39,746

$ Betas represent changes in standard deviations of the standardized log-transformed 25(OH)D levels, while controlling for age, sex, BMI and season of measurement

5.5 Discussion

5.5.1 GWAS meta-analysis

In this meta-analysis of genome-wide association studies for 25(OH)D levels in European populations, a low-frequency, synonymous coding genetic variant of large effect that strongly associates with 25(OH)D levels was identified. This variant has an effect size four-fold larger than that described for the common variant in the same gene (*CYP2R1*). The biologic plausibility of this finding is supported by the fact that the low-frequency variant is located in *CYP2R1*, the major hepatic 25-hydroxylase for vitamin D.[84] These findings are of clinical relevance since this study shos that approximately 5% of the general European population carry this variant in either the homozygous or heterozygous state.

This study was enabled by large imputation reference panels (UK10K/1000 Genomes and HRC), which offer at least 10-fold more European samples than the 1000 Genomes reference panel alone. Low-frequency variants of large effect on 25(OH)D were not identified in novel genes in Europeans. Although, variants with smaller effects in two loci not previously known to be associated with 25(OH)D were identified.

CYP2R1 encodes the enzyme responsible for 25-hydroxylation of vitamin D, and is one of the two main enzymes responsible for vitamin D hepatic metabolism.[84] Rare mutations in *CYP2R1* have already been described to cause rickets.[84,190]

Two common genetic variants of small effect size were also identified on chromosome 12 (*HAL* gene) and chromosome 14 (*GEMIN2* gene) that reached genome-wide level significance in the meta-analysis. Although there is no existing evidence of implication of *GEMIN2* in vitamin D related physiologic pathways, *HAL* is expressed in the skin and is involved in formation of urocanic acid, a "natural sunscreen".[191,192] Thus, this could constitute a plausible pathophysiologic mechanism implicating *HAL* in vitamin D synthesis in the skin. Additional functional follow-up of the signals in chromosomes 12 and 14 is needed to characterize the genes and/or mechanisms underlying these associations. Since the completion of this study a subsequent GWAS has been published which aimed to identify only common genetic variants associated with 25(OH)D.[193] This study of

approximately 80,000 individuals, also found strong evidence for 25(OH)D-associated variants on chromosomes 12 and 14. Although these two variants were not the same as those identified in this study they were in very similar loci. In European populations, the variant identified in this study on chromosome 12 (rs3819817) is in moderate LD (r^2 = 0.63) with the variant identified in the newer GWAS (rs10745742) and the variant identified in this this study on chromosome 14 (rs2277458) is in very high LD (r^2 = 0.98) with the variant identified in the newer study (rs8018720).

These findings may have clinical relevance for several reasons: First, individuals carrying at least one copy of the low-frequency *CYP2R1* variant have lowered levels of 25(OH)D by a clinically relevant degree. Secondly, these findings affect ~5% of individuals of European descent and may have utility in predicting individuals at risk of deficiency.

This study has its limitations. First, although part of the scope of this study was detection of low-frequency and rare variants, the two whole genome sequencing studies included in this study were of relatively low read depth (6.7x), and three studies were imputed to older imputation panels (1000Genomes and UK10K). These studies have a limited capacity to capture very rare variants, which might explain why such associations were not identified. Since the analysis is restricted to populations of European ancestry, no assumptions concerning the effect of rs117913124 in non-European populations can be made. Nonetheless, based on the 1000Genomes reference, this variant is rare in Africans (MAF = 0.3%) and has not been described in East Asians (MAF = 0%). Therefore, very large sample sizes of these populations will be required to describe with any certainty the effect of this variant on 25(OH)D level in these populations. In the absence of functional experiments showing the exact function of the rs117913124 on *CYP2R1* and given that this synonymous polymorphism does not affect protein sequence, it cannot be unequivocally shown that this low-frequency variant is causal, however, given that this is a coding variant in a well-documented 25(OH)D-associated gene, it seems most likely that it exerts its effect on *CYP2R1*. In this meta-GWAS the phenotype was adjusted for BMI in order to remove any of the variance of 25(OH)D explained by BMI. Although this may have slightly improved power to identify genetic variants associated with 25(OH)D, it could introduce biases when the summary results are used in downstream MR analyses. Specifically, by conditioning on BMI which could be considered a collider in the vitamin D – HNC association, collider bias may be introduced however, in reality the magnitude of this bias is likely to be very small. Finally, population stratification may have been an issue in some of the studies included in the meta-analysis which did not use a linear mixed model to account for population structure. Despite this potential limitations, the correction of the issue with ancestry informative principal components is unlikely to change the key findings dramatically as only genetic variants which were consistent across all 17 studies (included seven where population

stratification was corrected for) were considered true findings. In future GWAS efforts of vitamin D 25(OH)D will not be adjusted for BMI and will be conducted in a method which accounts for potential population structure.

5.5.2 Vitamin D genetic proxy

This GWAS demonstrates that vitamin D may have an oligogenic architecture. Compared to the previous GWAS[178], the Jiang *et al.*[193] study includes five times the number of participants in their discovery sample and the study in this chapter had twice the sample size and covered four times the number of genetic variants in the analysis. Despite this, there have been very few new genetic variants discovered which are strongly associated with vitamin D. Although this may partly relate to power, it may suggest that serum 25(OH)D biology is tightly controlled by only a few genes and pathways. This may have implications on the use of these variants as proxy measures of vitamin D. As mentioned, prior to this study there were four genetic variants which had been robustly associated with 25(OH)D and had been used in MR analyses. As a result of the low frequency and common variants identified in this study, and the similar findings for the two common variants in the Jiang *et al*. (2018) GWAS,[193] there are now seven genetic variants with robust associations with 25(OH)D which can be used as proxy measures of vitamin D. This increase in the number of genetic variants will increase power for MR analyses as a larger amount of variance in 25(OH)D will be explained by the seven variants than the original four. This increase in power is likely to be modest however, with the two novel common variants having only a small effects on the circulating 25(OH)D level. Despite the very large effects of rs117913124, it has a MAF of approximately 2.5%, and will therefore not explain a large amount of variance in 25(OH)D within the population.

The low frequency variant (rs117913124), does however have other research benefits due to its large effects. Studies in which individuals who are carriers of large effect variants, such as rs117913124, are recalled and undergo further testing or more specific phenotyping (either physically or *in-silico*) are referred to as recall-by-genotype studies and can be an efficient and powerful adjunct to GWAS. In an extension to the work in this chapter, the effect of this variant on multiple sclerosis (MS) risk was investigated. The hypothesis was that as rs117913124 exerts such large effects on 25(OH)D, and there is a causal association between 25(OH)D and MS risk, the 25(OH)D decreasing allele of the variant should predispose carriers to MS. In this work, strong evidence for this risk increasing association was identified with Individuals carrying one copy of this variant having an increased odds of multiple sclerosis (OR = 1.4 (95% CI = 1.19, 1.64); P = 2.63 $x10^{-5}$).[194]

As a continuation of this work, I am contributing to a new GWAS for 25(OH)D. This new study will combine the meta-analysis is this chapter with the results of a GWAS in UK Biobank. In total, this will include almost 450,000 participants and will therefore be adequately powered to detect genetic variants with smaller effect sizes and low frequency variants. This project is discussed further in the Future Studies section in Chapter 9.

The next chapter

In this chapter several genetic variants were discovered to be associated with 25(OH)D permitting their use for genetic proxies of 25(OH)D. In the next chapter I combine these variants with those previously identified to generate a genetic risk score for 25(OH)D to use in MR studies. I demonstrate how this proxy measure has advantages over serum 25(OH)D for assessing causal relationships.

Chapter 6 : Assessing genetic variation as a potential vitamin D proxy

6.1 Chapter objectives

In Chapter 5 genetic variants that are strongly associated with vitamin D were identified. In this chapter the suitability for these genetic variants as proxy variables in MR analyses is investigated. Associations of the 25(OH)D genetic proxy variable with over 35,000 phenotypic variables are assessed and compared to the same associations with serum 25(OH)D. A comparison of the results of these studies demonstrates the favourable properties of the proxy variable for causal inference. The aim of this chapter is to demonstrate the properties of the 25(OH)D genetic proxy and how they allow complementary assessment of causal relationships, alongside the observational estimates from Chapter 4.

6.2 Introduction

As discussed in Chapter 3, MR utilises genetic variants as instrumental variables to provide estimates of association between exposure and outcome, which are, in the right conditions, less affected by reverse causality and confounding. Genetic proxies are not necessarily perfect instruments for use in instrumental variable analyses such as MR as they often lack power and may not be truly independent of other factors. They are however relatively easily identified and readily measured in large numbers of individuals and so lend themselves to these types of analyses. Furthermore, they are the closest biological examples get to true instrumental variables.

Phenome scans and PheWAS are hypothesis free approaches which are usually used to assess for associations between a variable or genetic variant with many potential outcomes. In this chapter, these hypothesis free approaches are repurposed to interrogate the MR assumptions where possible and assess the legitimacy of the genetic proxy over and above what can normally be achieved in a MR study. An alternative approach could be to select the phenotypes to include in the phenome scans as this would lower the multiple testing burden however, this approach would be heavily influenced by the variables selected and would not provide much more information to that in a standard MR analysis where measured confounders are tested for their association with the genetic instrument.

MR relies on the three assumptions previously discussed in Chapter 3:

A. The relevance assumption: The genetic variants used within the MR analysis are robustly associated with the exposure of interest

B. The independence assumption: The genetic variants share no common cause with the outcome of interest
C. The exclusion restriction assumption: The genetic variants only affect the outcome of interest through the exposure of interest.

The first assumption (A) can be readily tested. The identification of genetic proxies using GWAS, as in Chapter 5, relies on a hypothesis free approach and has been recognised as a technique able to yield reliable associations (through a combination of stringent thresholds to account for multiple testing and validation in multiple independent collections). In the case of 25(OH)D, the identified genetic variants are mostly found in or near genes which are vitamin D pathway related. Although this is no guarantee that the genetic variant is associated with these genes, it does lend further credence to their strong association with 25(OH)D. Despite the robust design of the GWAS experiment in Chapter 5, there remains the possibility that false positive genetic variants, which are not strongly associated with 25(OH)D, were identified. To further test the relevance assumption, the genetic variants can be assessed for their association with 25(OH)D in an independent study – if the association holds true then this provides strong evidence that their association with 25(OH)D is robust.

As mentioned previously, four of the genetic variants for 25(OH)D have been identified in genetic studies of 25(OH)D prior to this thesis[178] and the study in Chapter 5 provided further replication of these variants. Since the publication of the work in Chapter 5, a larger GWAS of 25(OH)D has been published by Jiang *et al.*[193] This more recent GWAS not only re-confirmed the association with the four original 25(OH)D variants but confirmed the association with the loci first identified in Chapter 5, on chromosomes 12 and 14, in both their discovery and replication sets.

The independence assumption (B) is at first thought easy to assess. Genetic variants which are allocated randomly at meiosis should have no common cause with a disease such as HNC. No action during an individual's life can affect their germline genetics and therefore it would seem that this assumption is not violated. There are however scenarios in which the assumption is violated, for example, where differing ancestry within a sample means that the allele frequencies of the 25(OH)D genetic variants are different across ethnic groups. If this coincides with a difference in disease prevalence across the same ethnic groups (due to cultural differences and not genetic effects) then the MR effect estimate will be biased. Some simple measures can be taken to reduce the confounding effect of ancestry including limiting the analysis to a particular ancestral group and/or including genetic principal components into the MR model to account for the effect of ancestry. This topic is beyond the scope of this thesis, but is discussed further in one of the papers to which I contributed to, published during this thesis.[144] Another potential source of bias which could violate

the independence assumption is ascertainment or collider bias. This is when a selection event introduces spurious associations between genetic variants and potential confounders. This topic is introduced in more detail in Chapter 3 and its implications investigated in Chapter 8.

It is challenging to show that the exclusion restriction assumption (C) has not been violated in an MR study. In the case of 25(OH)D, there are relatively few genetic variants and all are closely related to vitamin D pathway genes. This reduces the likelihood of any of those variants having effects on other pathways which may independently affect HNC, a situation referred to as horizontal pleiotropic effects. However, there is the possibility that the genetic variants have effects on other pathways or metabolites proximal to 25(OH)D in the pathway, which independently of 25(OH)D go on to effect HNC. If this is the case, the causal effect from MR will be biased.

In this chapter two parallel analyses are performed with the aim of systematically evaluating the causal effects of vitamin D and of understanding the suitability of the 25(OH)D genetic proxy for MR: (1) an observational study analysing the association of serum 25(OH)D with over 35,000 phenotypic variables and (2) an MR pheWAS analysis which estimates the causal effect of 25(OH)D on the same 35,000 measures using the genetic proxy of vitamin D identified in Chapter 5. These parallel studies will provide information on whether the 25(OH)D genetic proxy is suitable for use in MR. Firstly, the strength of the genetic proxy as a measure of 25(OH)D will be assessed in an independent study to evaluate whether the proxy passes the relevance assumption. Secondly, the comparison between the observational and MR PheWAS results will provide improved understanding of the confounded versus unconfounded effects of vitamin D. This is turn may provide further understanding of any disparity between observational and causal estimates of effect in this thesis. As mentioned above, assessing the exclusion restriction assumption is complicated. Theoretically, potentially pleiotropic phenotypes between vitamin D genetic variants and HNC could be identified however, in reality power to detect these pleiotropic pathways will be very low.

6.3 Methods

6.3.1 Overview

In this chapter, an observational phenome scan study and MR pheWAS study analysing the association of 25(OH)D with over 35,000 phenotypic variables was conducted. The first stage was to generate a 25(OH)D genetic risk score (GRS) by combining the effects of multiple genetic variants from those identified in this thesis and the more recent Jiang *et al.* GWAS.[193] This GRS was then assessed for associations with over 35,000 phenotypic variables in UK Biobank using a dedicated analysis program. For the observational phenome scan, the analyses were then repeated to assess

the same associations but using serum 25(OH)D instead of the 25(OH)D GRS. A series of positive and negative control look-ups were also conducted.

6.3.2 Study cohort

The study was conducted in UK Biobank, a population-based study of approximately 500,000 participants. A detailed description of this study, genotyping and the 25(OH)D measures are given in Chapter 3.

6.3.3 Generation of exposures

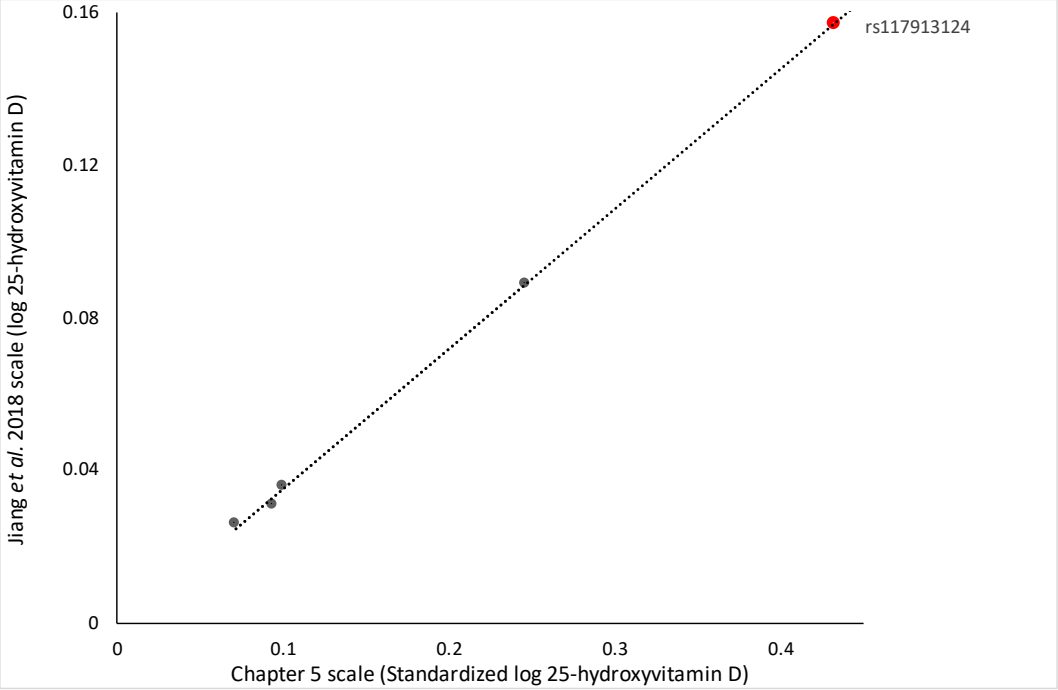
6.3.3.1 25-hydroxyvitamin D

To remove the variation in serum 25(OH)D attributable to season, the 25(OH)D measures were adjusted as described in Chapter 4.

6.3.3.2 Genetic risk score for vitamin D

A weighted genetic risk score (GRS) for vitamin D was generated by combining the genetic variants identified in Chapter 5 and the subsequent Jiang *et al.*[193] GWAS studies. Where possible effect estimates and standard errors were taken from the Jiang *et al.* GWAS as this has the largest sample size. The low frequency *CYP2R1* genetic variant identified in Chapter 5 was not included in the Jiang *et al.* summary results due to its low allele frequency.[193] The effect size and standard error of this variant was therefore scaled to match that of the variants from the Jiang *et al.* study. This was achieved by taking the four variants that were reported in both GWAS, matching their effect alleles and regressing the effect sizes from Jiang *et al.* on the effect sizes from Chapter 5 (Figure 6-1). The resultant linear equation was used to scale the effect of the low frequency *CYP2R1* variant.

Figure 6-1: Plot of effect size estimates of genetic variants included in both Jiang *et al.* and Chapter 5.



The effect estimates for the variants included in both Jiang *et al*. and Chapter 5 (rs3755967, rs10741657, rs12785878, rs172167070) are plotted against each other (•). The linear equation (y = 0.3662*Beta - 0.0012) was used to scale the effect estimate of the low frequency variant rs117913124 (•) in the Chapter 5 GWAS that was not reported in the Jiang *et al.* GWAS.

$$Beta_{Jiang\ predicted} = 0.3662 \cdot Beta_0 - 0.0012$$

$$Beta_{Jiang\ predicted} = 0.3662 \cdot -0.4323 - 0.0012$$

$$Beta_{Jiang\ predicted} = -0.160$$

The details of the seven genetic variants included in the GRS are shown in Table 6-1. The GRS was weighted by the effect size of the variants from the largest available GWAS (Jiang *et al.*) as this provides the most precise estimates. The weighting takes into account the variant's relative effect size and direction. The GRS is in log 25(OH)D units in nmol/L and is coded such that an increase in GRS corresponds to an increase in 25(OH)D.

$$GRS = Beta_{snp\ 1} \cdot Allele\ dose_{snp\ 1} + \cdots + Beta_{snp\ n} \cdot Allele\ dose_{snp\ n}$$

Table 6-1: Effect estimates from GWAS for the genetic variants included in the genetic proxy

Gene	RSID	chr:pos	Eff/Alt	EAF	Beta	SE	P
GC	rs3755967	4:72828262	T/C	0.28	-0.089	0.0023	4.74E–343
NADSYN1/ DHCR7	rs12785878	11:70845097	T/G	0.75	0.036	0.0022	3.80E–62
CYP2R1	rs10741657	11:14871454	A/G	0.40	0.031	0.0022	2.05E–46

CYP24A1	rs17216707	20:52165769	T/C	0.79	0.026	0.0027	8.14E–23
AMDHD1	rs10745742	12:94882660	T/C	0.40	0.017	0.0022	1.88E–14
SEC23A	rs8018720	14:38625936	C/G	0.82	-0.017	0.0029	4.72E–09
CYP2R1	rs117913124	11:14900931	A/G	0.025	-0.160	0.00645	1.50E–88

6.3.4 Outcome measures

The data included in these analyses (dataset 29244 in UK Biobank application 16729) includes 3,414 fields of the following types: integer, continuous, categorical (single) and categorical (multiple). These fields contain a wide range of phenotypes including phenotypes derived from health and lifestyle questionnaires, clinical assessments, blood and saliva assays (including serum 25(OH)D) and linkage to cancer and death registries and hospital episodes.

The following fields were removed: assessment centre; two fields described by UK Biobank as 'polymorphic', containing values with mixed data types; 26 genetic descriptor fields; one sex field; five age fields; 18 fields describing the assessment centre environment, 185 data processing indicators, and 21 categorical (single) fields with more than one value recorded per person. This resulted in 3,153 fields (591 integer, 1,475 continuous, 934 categorical (single) and 153 categorical (multiple)) which form the outcome dataset.

6.3.5 Covariates

When assessing the strength of the genetic proxy age, sex and the first 40 genetic principal components were included in the model to remove variability in 25(OH)D due to age and sex, and to account for potential population stratification.

For the observational analysis using serum 25(OH)D, age at attendance of UK Biobank assessment centre and sex were included as covariates. In the MR analysis using the GRS, age, sex and the first 40 genetic principal components were included in the model.

6.3.6 Statistical analysis

The relevance assumption was assessed by investigating the association between the GRS and seasonally adjusted 25(OH)D using linear regression to confirm the strength of the GRS. The effect of individual SNPs was also assessed against the seasonally adjusted 25(OH)D measure.

The PHESANT software (version 0.18), which uses an automated rule based method to assign a variable type and hence an analysis method, has been described in detail elsewhere.[195] Briefly, continuous or integer fields are usually assigned to the continuous type variable, unless there are only a few distinct values, in which case they will be assigned as an ordered categorical variable. Fields of the categorical (single) type, in which each participant has only a single record, are assigned

to either binary, unordered categorical or ordered categorical depending on the type of data the field refers to. Fields of the categorical (multiple) type, in which each participant can have multiple records, are converted to a series of binary variables – one for each record. For example, a participant may be asked whether they ate any of "white bread", "wholemeal bread" and "other bread products" the previous day. As a participant could have conceivably eaten all three, each bread type is converted to a binary variable denoting whether each participant had or had not eaten that bread type the previous day. In this (hypothetical) example, a single field would result in three binary variables, one for each bread type. This rule-based system attempts to provide the best available representation of all the phenotypes in UK Biobank.

The association of the seasonally adjusted 25(OH)D measure and the GRS was tested against all outcomes in the outcome dataset to demonstrate the difference in their associations with potential confounders. PHESANT estimates the association of the exposure (either the vitamin D GRS or the seasonally adjusted 25(OH)D measure) with each outcome variable in the outcome dataset. Outcome variables with continuous, binary, unordered categorical and ordered categorical data types are analysed using linear, logistic, ordered logistic and multinomial logistic regression respectively. To ensure normality of continuous variables each one is inverse normal rank transformed prior to running the regression model. Covariates are included in the model as described above.

The strength of evidence for associations was assessed, accounting for the number of statistical tests performed using the Bonferroni corrected p-value threshold. This was calculated by dividing 0.05 by the number of tests performed.

To estimate the power in each of the MR PheWAS statistical tests at the Bonferroni corrected *P*-value threshold, post-hoc power calculations were conducted based on the size of the observational estimate using the methods described by Brion *et al.* (2013).[196] Each calculation took into account the variance in 25(OH)D explained by the GRS, the sample size for the specific phenotype and the observational effect size. In these power calculations, for phenotypes analysed using linear or ordinal regression an estimate of the variance in the exposure and the outcome was taken into account, for phenotypes calculated using logistic regression the proportion of cases in the sample was taken into account.

6.3.6.1 Positive and negative control look-ups

For the MR PheWAS, positive and negative controls were looked up in the results. In these look-ups, phenotypes were selected that have been shown to be and not be associated with vitamin D in completely independent samples. The PheWAS results are then assessed to see they support these a

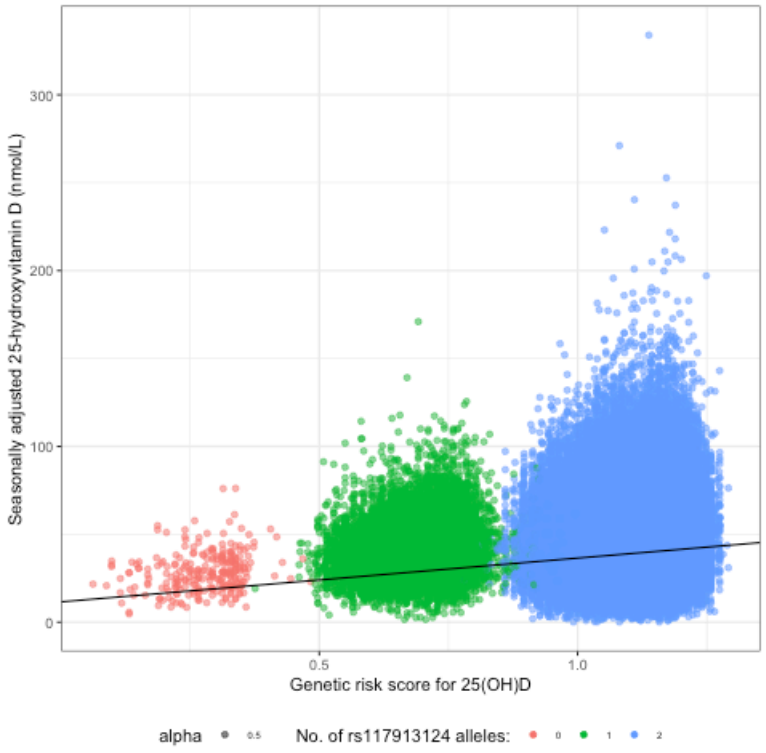
priori hypotheses. Multiple sclerosis was chosen as a positive control due to the strong evidence for a protective effect of 25(OH)D from previous MR studies.[197,198] Body mass index was selected as a negative control, again due to previous high quality MR studies suggesting there is no evidence of an effect of 25(OH)D on body mass index.[105]

6.4 Results

6.4.1 Assessing the GRS

The GRS was very strongly associated with the seasonally adjusted 25(OH)D measure, with a one unit increase in the GRS corresponding to a 24.9 nmol/L (95% CI: 24.3, 25.5) increase in 25(OH)D. This supports the genetic proxy does not violate the relevance assumption. The overall distribution of seasonally adjusted 25(OH)D with GRS is shown in Figure 6-2, the 25(OH)D level was strongly affected by the number of rs117913124 alleles due to the large effects of this variant. The GRS explained approximately 3.2% of the variance in seasonally adjusted 25(OH)D. The effect of each SNP on seasonally adjusted 25(OH)D was assessed (Table 6-2). All SNPs had effects in the same direction and of similar magnitude to those in the most recent 25(OH)D GWAS.

Figure 6-2: Distribution of seasonally adjusted 25(OH)D with genetic risk score.



Each dot represents an individual with their GRS plotted against their seasonally adjusted 25(OH)D level. Dot colour is determined by the number of rs117913124 alleles the individual carries. The groupings reflect the large effect size of the rs117913124 variant resulting in it having a large weighting in the GRS. The variation in number of dots in each grouping reflects that the variant is rare and so only few participants have both vitamin D lowering alleles.

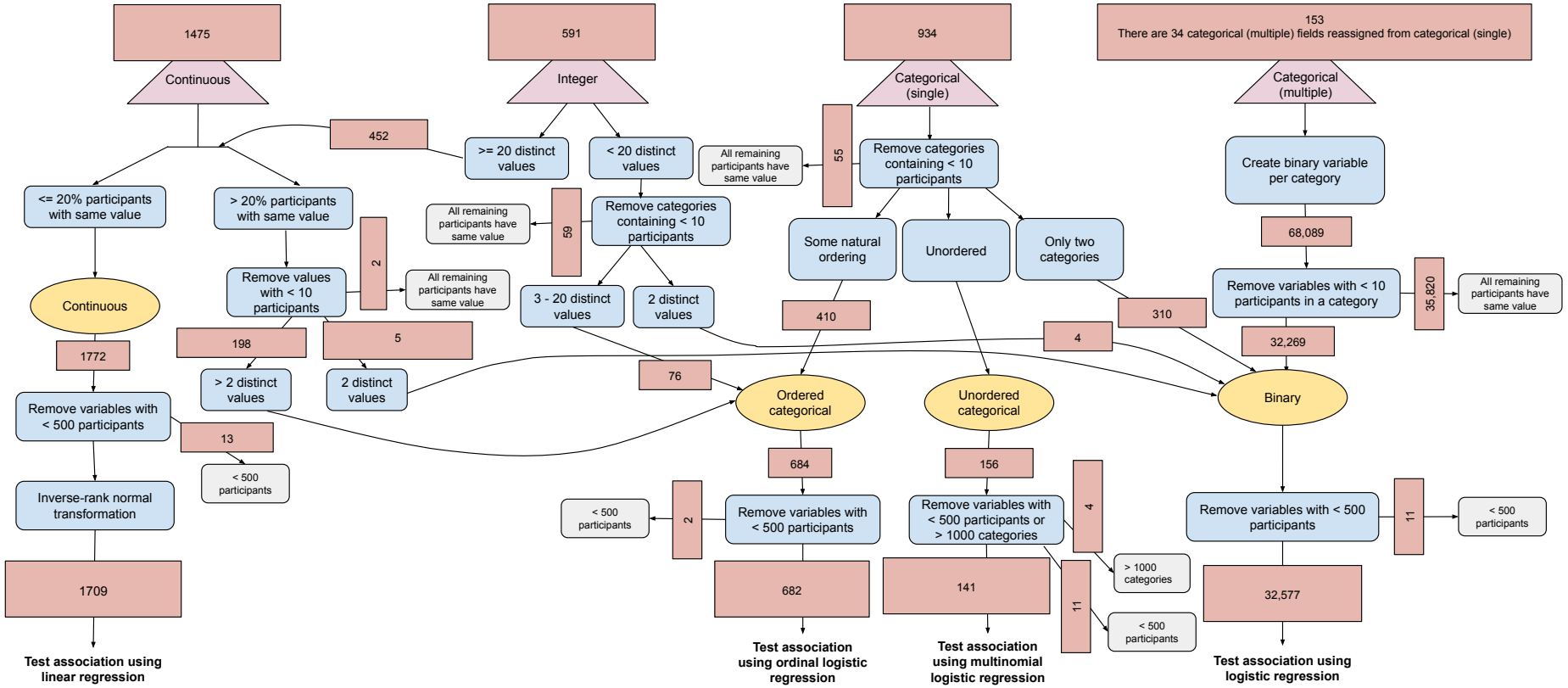
Table 6-2: Effect of SNPs on seasonally adjusted 25(OH)D in UK Biobank and Jiang *et al.* (2018).

		UK Biobank				Jiang *et al.*	
RSID	**Eff/Alt**	**EAF**	**Beta**	**SE**	**P**	**Beta**	**SE**
rs3755967	T/C	0.29	-0.084	0.0012	< 2e-16	-0.089	0.0023
rs12785878	T/G	0.76	0.051	0.0013	< 2e-16	0.036	0.0022
rs10741657	A/G	0.40	0.038	0.0011	< 2e-16	0.031	0.0022
rs17216707	T/C	0.81	0.017	0.0014	< 2e-16	0.026	0.0027
rs10745742	T/C	0.38	0.013	0.0011	< 2e-16	0.017	0.0022
rs8018720	C/G	0.82	-0.017	0.0014	< 2e-16	-0.017	0.0029
rs117913124	A/G	0.027	-0.172	0.0033	< 2e-16	-0.160	0.0065

6.4.2 Phenome scan analysis

The PHESANT data processing pipeline resulted in 35,109 phenotypes for testing in the observational and GRS analyses. A flow diagram detailing how the different data field types were analysed by PHESANT is given in Figure 6-3. The Bonferroni corrected p-value threshold accounting for the number of statistical tests was 1.42×10^{-06} (0.05/35,109).

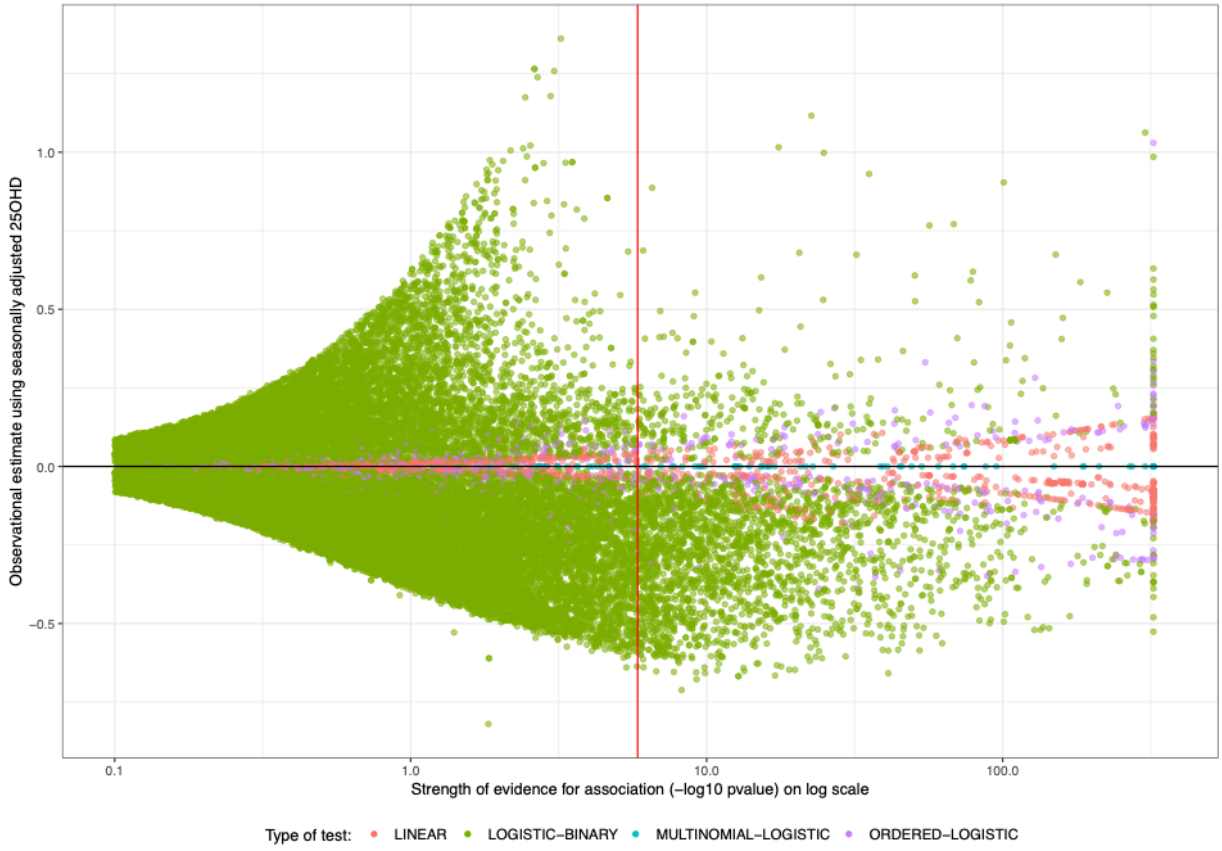
Figure 6-3: Flow diagram showing details of PHESANT analysis algorithm for both the seasonally adjusted 25(OH)D and exposures.



The flow diagram reflects the decision tree of PHESANT to allocate the four different variable field types, indicated by red triangles (Continuous, Integer, Categorical (single), Catigorical (multiple)) to the four different regression approaches (linear regresion, ordinal regression, multinomial logistic regression, logistic regression). The first part of the decision tree checks the field type, converts the field to a processable data type and allocates a regression method. The second part of the decision tree conducts quality checks on the variable. Red rectangles indicate numbers of phenotypes at each stage of the decision tree. Blue rectangles indicate steps. Grey rectangles indicate why phenotypes have been removed from analyses. As an example, take one of the 591 integer variable such as pulse rate (in beats per minute). The algorithm checks the number of distinct values (>= 20 for pulse rate) and whether more than 20% of participants have the same value (not true for pulse rate). This variable is then considered a continuous variable for the purposes of analyses. The variable is dropped if less than 500 participants have a value (not true for pulse rate). Finally the variable is inverse-rank normal transformed to deal with a non-normal distribution and then analyses by linear regression.

In the observational analyses 3,538 phenotypes passed the Bonferroni corrected p-value threshold. Figure 6-4 gives an overview of the results of the observational phenome scan and the top 50 results are provided in Appendix 6.1.

Figure 6-4: Overall graphical view of observational PheWAS results.

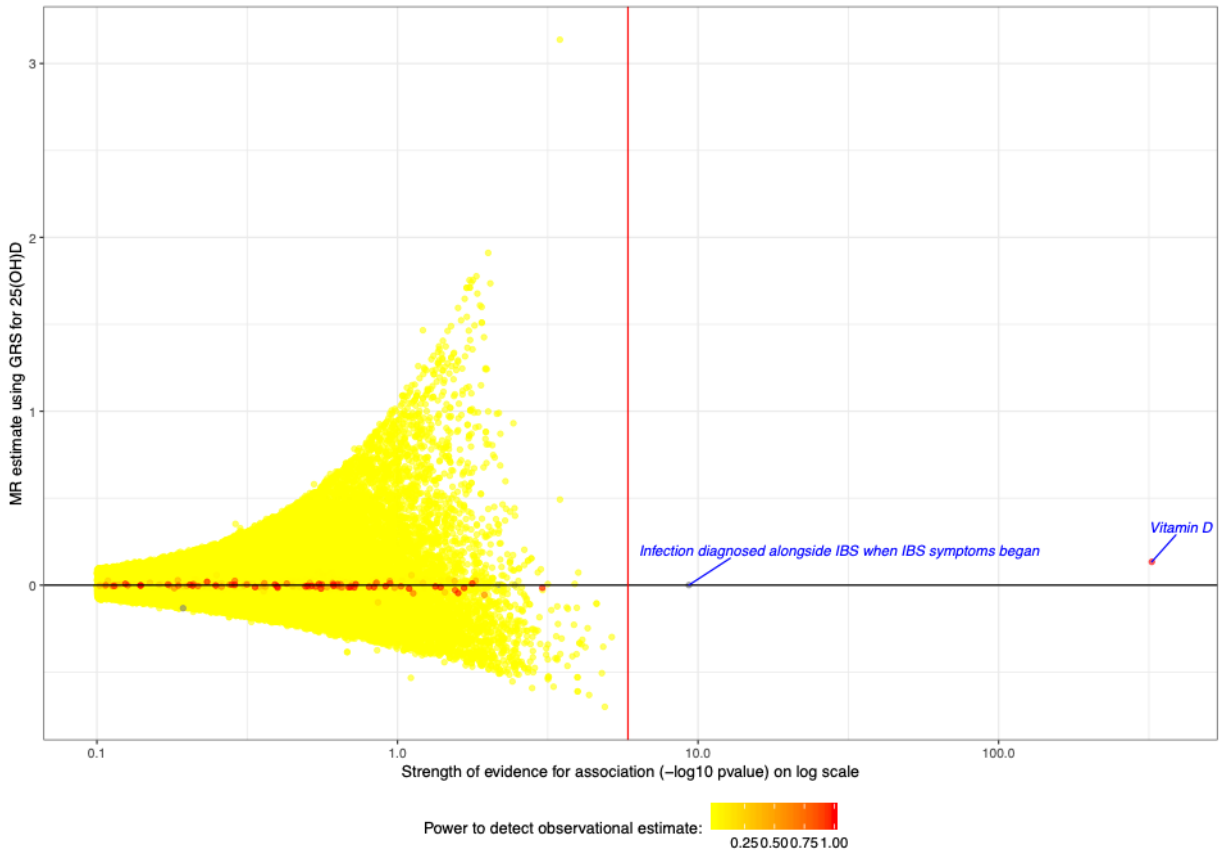


This plot gives an overall representation of the effect size and strength of evidence for all results in the observational phenome scan. It demonstrates that there are many phenotypes from all regression analyses that pass the multiple testing threshold for an association with serum 25(OH)D. Each dot represents a result from the PHESANT analysis with the colour corresponding to the type of statistical test used (see legend). The y-axis indicates the effect size of 25(OH)D on the phenotype. The x-axis indicates the strength of association with stronger evidence towards the right side of the plot. Multinomial-logistic results do not have effect sizes, only *P* values, so are plotted on the x-axis to allow the strength of evidence for the association to be visualised. The red line indicates the Bonferroni corrected threshold (*P* = 0.05/35,109 = 1.42x10^{-6}).

In the MR PheWAS analyses, there was greater than 80% power to detect the observational effect in 58 phenotypes at the Bonferroni corrected *P*-value threshold using the effect estimates from the observational analyses. However, in the MR PheWAS analyses, only two variables passed the Bonferroni threshold. An overview of the results of the MR PheWAS including an indication of the

power for each analysis are given in Figure 6-5. The strongest effect of the GRS was seen for the serum 25(OH)D measure (PHESANT standardised beta: 0.133 (95% CI: 0.130, 0.137) $P< 5x10^{-324}$, n=306,985). The only other variable to pass the multiple testing threshold was '21072' (UK Biobank description: 'Infection diagnosed alongside IBS [irritable bowel syndrome] when IBS symptoms began' (P =4.78x10^{-10}, n=769).

Figure 6-5: Overall graphical view of MR PheWAS results.



This plot gives an overall representation of the effect size and strength of evidence for all results in the MR PheWAS analyses. It demonstrates that there are very few phenotypes that pass the multiple testing threshold for an association with serum 25(OH)D. Although power is limited, even wehere there is sufficient power to detect effects, there is limited evidence for causal associations. Each dot represents a result from the PHESANT analysis with the colour corresponding to the power to detect the observational estimate magnitude, as detected in the observational analyses. The y-axis indicates the effect size of 25(OH)D on the phenotype. The x-axis indicates the strength of association with stronger evidence towards the right side of the plot. Multinomial-logistic results do not have effect sizes, only *P* values, so are plotted on the x-axis to allow the strength of evidence for the association to be visualised. The red line indicates the Bonferroni corrected threshold *P* = (0.05/35,109 = 1.42x10^{-6}). Phenotypes which pass this threshold are labelled with their UK Biobank phenotype description.

6.4.2.1 Comparison of estimates of major HNC risk factors between observational and MR PheWAS results.

In the observational PheWAS, consistent with what was expected, there was very strong evidence for a lowering effect of 25(OH)D on smoking and alcohol consumption (OR = 0.74 (95% CI: 0.74, 0.75), $P < 5x10^{-324}$; OR = 0.86 (95% CI: 0.85, 0.86), $P < 5x10^{-324}$ respectively) however in the MR PheWAS the estimates were on the null effect (OR = 1.00 (95% CI: 0.99, 1.01), *P*=0.62; OR = 1.00 (95% CI: 0.99, 1.00), *P*=0.42). There was good power to identify an association with smoking (power = 0.93) but less so with alcohol (power = 0.44).

6.4.2.2 Follow-up of potentially novel causal associations with 25(OH)D

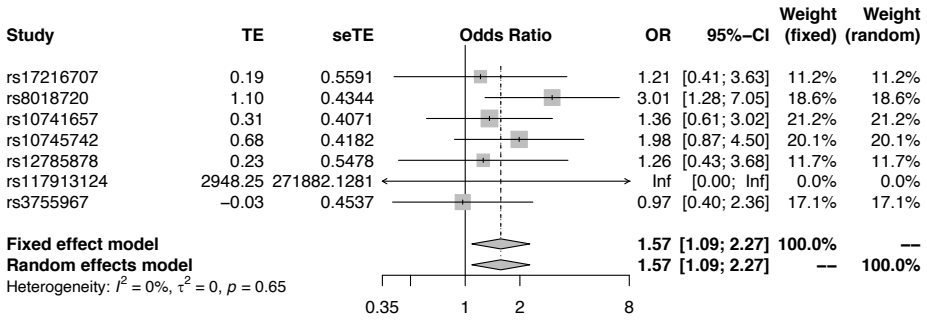
The only variable identified, in the MR PheWAS, as having a potential causal association with 25(OH)D, other than serum 25(OH)D, was a questionnaire item asking which type of 'Infection [was] diagnosed alongside IBS when IBS symptoms began' (field ID 21072). This was one of a series of UK Biobank questions around digestive health and this specific question was only asked to participants who reported having a diagnosis of IBS and also reported having an infection at the time of that diagnosis. The hypothesis that is being tested for this phenotype is does vitamin D cause an individual to get one infection over another infection alongside IBS. The analysis result for this phenotype came from a multinomial logistic regression analysis in PHESANT and therefore does not provide an overall effect estimate.

To inspect this further, the '21072' field was extracted from the dataset and analyses repeated manually. When analysing the results of the multinomial logistic regression, the beta for the shigella Infection category is greater than 2000 times the magnitude of the other categories. Of the 769 participants who reporting having an infection at the onset of IBS symptoms, only 12 had a shigella infection. Removing those 12 participants from the analyses removes the strong statistical evidence for effect (*P*-value for model after removing participants with Shigella = 0.71, *P*-value for model including participants with Shigella = $4.8 x10^{-10-}$).

Logistic regression analysis was conducted converting the shigella infection phenotype to a binary variable including only those that answered the question (coded 1 = shigella infection, n=12, 0 = answered the question but no shigella infection, n=757). This analysis confirmed the false positive association with the GRS with the results showing an unrealistic effect size and very wide confidence intervals (OR=27.3 (95% CI: 0.0, 35738.2), *P*=0.88). When this logistic regression analysis was repeated for each genetic variant in the GRS individually, the effect was driven entirely by the rs117913124 variant which had a very large but very imprecisely estimated effect on the binary

shigella variable (Figure 6-6). This extreme effect size is caused by all 12 shigella infection participants have the same number of risk alleles of the rs117913124 variant.

Figure 6-6: Effect of individual variants in the GRS on having a shigella infection at the time of IBS diagnosis.



When the individual genetic variant effects were meta-analysed the rs117913124 had a negligible effect on the pooled estimate due to its large standard error. The pooled effect therefore represents the effect estimate for the binary shigella variable when the rs117913124 variant has been removed (Figure 6-6).

6.4.2.3 Positive and negative control look-ups

The point estimate effect of the MR analysis on multiple sclerosis showed a protective effect of increased 25(OH)D in line with results from previous MR experiments (OR=0.66 (95% CI: 0.38, 1.15)). The analysis however was underpowered and confidence intervals crossed the null. The corresponding observational estimate showed strong evidence for a larger protective effect of vitamin D (OR=0.57 (95% CI: 0.49, 0.66)).

In the MR PheWAS, there was little evidence of an effect of 25(OH)D on body mass index which is in line with previous MR experiments (PHESANT standardised beta: -0.000 (95% CI: -0.004, 0.003)). This is in strong contrast to the observational analysis which showed very strong evidence for an inverse association between 25(OH)D and BMI (PHESANT standardised beta: -0.169 (95% CI: -0.173, -0.166)).

6.5 Discussion

The observational phenome scan identified over 3,500 phenotypes that were strongly associated with 25(OH)D after correcting for multiple testing. This was in contrast to there being no suggestive causal associations identified in the MR PheWAS analysis after discounting 25(OH)D itself and the false positive IBS infection variable. There was however a very strong association between the genetic proxy and 25(OH)D which provides further evidence that the proxy is robust and the relevance assumption for MR is not violated.

In general, the observational results demonstrate the highly confounded nature of vitamin D with an increase in its level being strongly associated with positive heath phenotypes such as lower BMI, less smoking and alcohol consumption as well many others. This is important as it may explain how confounding could explain observational associations between vitamin D and HNC, even after correcting for the most obvious confounders. For example, analyses looking at the risk of HNC have not routinely corrected for BMI but it is likely that this may confound the vitamin D-HNC risk association.

One aim of this chapter was to assess whether the vitamin D GRS, a combination of vitamin D genetic variants, was more robust to confounding that the serum 25(OH)D measure. When the results of the observational and MR PheWAS are compared it demonstrates the GRS is related to very few phenotypes, other than 25(OH)D itself. However, it must be considered that there was not power to detect effects for the majority of variables. Where power was high to detect effects in the MR PheWAS analyses, there were some clear demonstrations that the GRS is not related to potential confounders which would be strongly associated with the serum vitamin D measure. For example power was high to detect effects on sunbed use (PHESANT standardised beta: 0.00; 95% CI: -0.02, 0.02; *P*=1.0, power = 1) and vitamin D supplementation (PHESANT standardised beta: -0.01; 95% CI: -0.03, 0.00; *P*=0.2, power = 1) but the MR PheWAS estimates were very close to the null. This is important as these variables, although clearly important in serum vitamin D levels, are likely to be correlated with other behaviours which cause HNC and would therefore bias observational associations but, as they are not related to the GRS, will not bias MR estimates.

It is worth considering whether the stark difference between the MR PheWAS results and the observational phenome scan results are a result of chance. It could be that if repeated the observational and MR PheWAS associations would be much more similar, either with more associations identified in the MR PheWAS, less in the observational analyses or both. However, given the expectation that the GRS would be associated with less variables than the serum 25(OH)D measure matching the actual results, it is unlikely that this finding was only by chance. One way to confirm this would be to run a simulation study but that is beyond the scope of this thesis.

The results of this study can be used to look up whether the genetic proxy is independent of the known confounders in the vitamin D HNC association. For both smoking status and alcohol consumption frequency, there was strong evidence of serum 25(OH)D having observational associations with smoking and drinking. However, in the MR PheWAS results the causal effect estimates for these phenotypes were precisely estimated around the null (OR = 1). Therefore, in MR

analyses that assess the causal effect of vitamin D on HNC, the estimate will be minimally affected by any confounding by these two major HNC risk factors.

MR PheWAS has several limitations which must be considered when interpreting these results. Firstly, despite the PHESANT algorithm (depicted in Figure 6-3) including steps to quality control each variable under investigation, it cannot be as rigorous as would be expected if conducting a study into a single association. This can lead to problems, such as statistical assumptions being violated, which may lead to false positives. One example of this is the finding in this chapter relating to the type of infection coincident with an IBS diagnosis, where too few individuals in a strata of the variable led to a strong false positive association. This demonstrates the importance of following up any PheWAS result with complementary analyses to assess its validity. Secondly, there are a large number of variables which are likely to suffer different types of error. For example, self-reported data may be biased by recall whereas cancer registry data will not. Li *et al.*[199] show that, in a PheWAS of uric acid genetic variants, the effect size on gout depends on the source of the information. However, in their study, the uric acid genetic variants showed strong evidence for an association with both data sources (self-reported gout and hospital diagnosed gout). Finally, the main limitation of PheWAS is the large multiple testing burden – particularly when investigating such a large number of phenotypes as in this study. In this work the large number of statistical tests combined with a GRS which does not explain a large amount of variance in vitamin D meant there was limited power to detect effects for the majority of the variables assessed. However, power for the observational analyses were high and comparing the MR PheWAS and observational results can be informative when assessing whether a variable should be followed up with subsequent studies. In this study, the Bonferroni multiple testing threshold was used to provide an indication of where the null value might lie, this is likely to be over-stringent. The limited power in this study precludes the possibility of ruling out horizontal pleiotropy as even if the GRS has pleiotropic effects, these are likely to be small and therefore not pass the multiple testing burden.

6.5.1 Conclusions

In summary, this chapter demonstrates the highly confounded nature of serum vitamin D measures and highlights that in observational studies of vitamin D on HNC, there is likely to be substantial residual confounding by factors that may not traditionally be measured as potential confounders. This chapter also shows that the vitamin D GRS is largely unconfounded and therefore lends itself to be used in MR analyses assessing the effect of vitamin D on HNC risk and prognosis.

The next chapter

In this chapter I have demonstrated that the vitamin D genetic variants identified in Chapter 5 are a strong proxy of vitamin D and are largely unconfounded. In the following two chapters, I apply MR to assess the causal association of vitamin D on HNC risk and prognosis, using the GRS tested in this chapter.

Chapter 7 : Assessing the causal association between vitamin D and HNC

7.1 Chapter objectives

In Chapter 4 the observational association between 25(OH)D and HNC incidence was investigated and strong protective associations were identified however these estimates are likely biased by confounding. In this chapter, an MR analysis is conducted using the genetic proxy identified and tested in the previous chapter. The aim of this chapter is to estimate the causal association between vitamin D and HNC incidence using MR.

7.2 Introduction

An increase in 25(OH)D was associated with a large reduction in the odds of oral cavity cancer (OR per doubling of 25(OH)D = 0.71 (95% CI: 0.57, 0.88)) and a moderate reduction in odds of oropharyngeal cancer (OR per doubling of 25(OH)D = 0.80 (95% CI: 0.65, 0.98)), even after correction for measured confounders. These estimates were similar in magnitude to those in a previous prospective observational study conducted in EPIC[126] and together they represent the best estimate of the effect of 25(OH)D on HNC risk to date.

As detailed previously, observational estimates are often biased by residual confounding even after careful correction for measured confounders using multivariable regression or stratification. Residual confounding and reverse causality often bias away from the null meaning effects estimates from observational studies are larger than the true causal effect. Furthermore, reverse causality can also be an issue as the development of cancer is a slow process and even undiagnosed cancers may affect an individual's exposure to factors such as vitamin D.

The study described in Chapter 4, and the previous study by Fanidi *et al.*[126], are case-cohort in design and are therefore likely to be affected by the biases mentioned above, residual confounding and reverse causality. Bias by confounding may not have been fully accounted for in Chapter 4 during multivariable regression analyses because of unmeasured confounding factors or due to error in the measurement of the confounders accounted for. One example of this is confounding by smoking status which, is often recorded at the time of diagnosis and is commonly recorded as current smoker, ex-smoker or never smoker. It is unlikely that these broad categories fully capture the influence of smoking, as they do not take into account smoking intensity, duration or type of smoking or variability in any of these aspects. As a result of this confounder measurement error, the effect estimate of 25(OH)D on HNC may still be biased by smoking even after correcting for it.

Reverse causality may have also contributed to biased estimates in the observational studies, although this is less likely as 25(OH)D measures were taken prior to any HNC diagnosis and sensitivity analyses including only those with 25(OH)D measures five years prior to diagnosis show similar results.

As previously described in Chapter 3, MR reduces the impact of bias by confounding due to the randomised allocation of genetic alleles at gamete formation and segregation. This randomisation event results in confounders being approximately equal across exposure groups irrespective of whether they are perfectly measured, measured with error or unmeasured. Reverse causality is also not an issue in MR as the development of a HNC cannot alter germline genetics.

In this chapter, MR is utilised to provide an estimate of the effect of 25(OH)D on HNC risk which is minimally affected by confounding bias or reverse causality.

7.3 Methods

7.3.1 Overview

Genetic variants which proxy 25(OH)D levels were utilised in this chapter to conduct an MR study assessing the causal effect of 25(OH)D on oral and oropharyngeal cancer risk (laryngeal cancers were not available in the included studies and so were not investigated). As serum 25(OH)D, genetic and HNC outcome measures were not available in a large enough sample, a two-sample MR approach was adopted and modified to make the best of individual level HNC outcome data.

The genetic variants identified in the 25(OH)D GWAS in Chapter 5 were used as 25(OH)D proxies. The genotypes of these variants were identified in individuals of two studies with HNC cases and controls (GAME-ON and UK Biobank) and combined to make a relative 25(OH)D measure. To assess the assumption that this relative 25(OH)D measure was unconfounded, associations between the relative measure and available confounders was assessed in each of the studies.

The relative 25(OH)D measure was then included in regression analyses to provide an estimate of the causal effect of 25(OH)D on HNC risk. Finally, a bootstrap approach was used to adjust the standard errors of these effect estimates so that they took into account the imprecision of the genetic variant effects on 25(OH)D. Failure to conduct this last step could lead to false positive findings.

7.3.2 Proxy variant selection

Common genetic variants strongly related to serum 25(OH)D levels have been identified in GWAS[98,178,193] including the work described in Chapter 5. For this MR study, genetic variants that passed a genome wide association threshold in these GWAS studies ($p<5x10^{-8}$) were selected. Five of

these genetic proxies, including the low-frequency variant identified in Chapter 5, were located in or near 25(OH)D related genes: Group-Specific Component (*GC*), Cytochrome P450 family 2, subfamily R, polypeptide 1 (*CYP2R1*), 7-Dehydrocholesterol Reductase (*DHCR7*) and Cytochrome P450, family 24, polypeptide 1 (*CYP24A1*). The two common variants near *HAL* (rs3819817) and *GEMIN2* (rs2277458) identified in Chapter 5 were also included as strong proxies of 25(OH)D.

To provide the most precise estimate of the causal association of 25(OH)D on HNC, effect estimates for SNP associations with serum 25(OH)D concentrations were taken from the largest 25(OH)D GWAS available (Jiang *et al.* (2018), published shortly after the study in Chapter 5). Where the variants identified in Chapter 5 were not available in the Jiang *et al.* summary results, the effect size from Chapter 5 was scaled to match the Jiang *et al.* paper, as in the pheWAS study reported in Chapter 6.

To be valid proxies of 25(OH)D in MR analyses, the genetic variants must be independent of each other. To assess independence, variants on the same chromosome were assessed for linkage disequilibrium (LD) using SNAP (http://archive.broadinstitute.org/mpg/snap/ldsearchpw.php) in Europeans from the 1000 Genomes (Pilot 1) project.

In total, seven variants were identified for the 25(OH)D proxy from GWAS (Table 7-1). rs4588 in *GC*, rs10741657 in *CYP2R1* and rs6013897 in *CYP24A1* have been used as genetic proxies for 25(OH)D in MR studies before. rs4423214 near *DHCR7* is in perfect LD with rs12785878, identified in the original SUNLIGHT consortium 25(OH)D GWAS.[178] Finally, rs116970203 near *CYP2R1,* rs3819817 near *HAL* and rs2277458 near *GEMIN2* were identified in Chapter 5 with the latter two (or variants highly correlated with them) being replicated in the Jiang *et al.* GWAS.

For each individual the seven 25(OH)D variants were combined into a genetic risk score weighted by the beta coefficients derived from GWAS. To ease clinical interpretation and to allow easy comparison with observational estimates, this score was converted from a per allele scale to a log2 25(OH)D scale. Therefore, each unit increase in this variable (hereafter referred to as relative 25(OH)D) represented a doubling of 25(OH)D.

Table 7-1: 25-Hydroxyvitamin D genetic variant details

							Vitamin D GWAS			GAME ON		UK Biobank	
RSID	Chromosome	Position	Gene	Effect Allele	Other Allele	EAF	Beta log2 25(OH)D	SE	P value	Rsq (minimac3)	EAF	Info (snptest)	EAF
rs4588	4	72618323	*GC*	G	T	0.717	0.089	0.0023	1.68E-263	Genotyped	0.722	Genotyped	0.712
rs116970203	11	14876718	*PDE3B*	G	A	0.975	0.157	0.0064	2.29E-90	0.98	0.978	0.99	0.973
rs4423214	11	71173254	*DHCR7*	T	C	0.697	0.036	0.0023	1.39E-40	0.99	0.718	1.00	0.761
rs10741657	11	14914878	*CYP2R1*	A	G	0.415	0.031	0.0022	8.76E-45	Genotyped	0.383	Genotyped	0.398
rs3819817	12	96378771	*HAL*	C	T	0.451	0.014	0.0024	1.88E-14	1.00	0.539	0.997	0.422
rs2277458	14	39583481	*GEMIN2*	A	G	0.214	0.017	0.0031	4.72E-09	Genotyped	0.825	0.998	0.824
rs6013897	20	52742479	*CYP24A1*	T	A	0.791	0.021	0.0024	9.06E-16	0.74	0.781	1.00	0.805

EAF = effect allele frequency, Rsq = R squared, SE = standard error

7.3.3 Studies

7.3.3.1 GAME-ON

An overall description of the GAME-ON study including the genetic data are provided in Chapter 3. For this study cancer cases of interest comprised the following ICD codes: oral cavity (C02.0–C02.9, C03.0–C03.9, C04.0–C04.9, C05.0–C06.9) and oropharynx (C01.9, C02.4, C09.0–C10.9). DNA extraction, genotyping, quality control and imputation has been described previously.[22] The study population included participants from Europe, North America and South America. To reduce the effect of heterogeneity across these regions only participants with >70% CEU ancestry were included in analyses (n=11,117).

7.3.3.2 UK Biobank

An overall description of the UK Biobank study including the genetic data are provided in Chapter 3. Age on 31 December 2010 (coinciding with the approximate end of baseline data collection) was used as the age variable and not age at cancer diagnosis. Prevalent and incident oral and oropharyngeal cancers (ICD10 codes matched to GAME-ON) were identified from linked cancer registry data. To reduce selection bias in allocating controls, all remaining participants, after removing those participants with other HNC, were used as controls. A full description of the UK Biobank study design, participants and quality control (QC) methods have been described in detail previously.[154] UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382). [161] In UK Biobank analyses were restricted to individuals of white British ancestry who self-report as “White British” and who have very similar ancestral backgrounds according to the principal component analysis, as described by Bycroft.[160] The full data release contains the cohort of successfully genotyped samples (n=488,377).

7.3.4 DNA extraction genotyping and imputation

DNA extraction, genotyping, imputation and the generation of genetic principal components for GAME-ON and UK Biobank is described in Chapter 3.

7.3.5 Assessment of potential confounders

One of the principals of MR is that as alleles are assigned randomly during gamete formation and segregation. The genetic proxy is therefore not associated with factors that typically confound the association in observational analyses. To test this assumption, associations between relative 25(OH)D and measured potential confounders were investigated. Sex, smoking status (never, ever, current) and alcohol use (never, ever) were recorded by questionnaire in both cohorts. In GAME-ON only, sub-study and study country were also considered potential confounders. The associations of

relative 25(OH)D with strata of these potential confounders was examined in boxplots of range and means +/- one standard deviation. Tests of associations were made using linear regression with categorical variables as factor variables, the overall *P*-value for the model was shown on the boxplots. Where evidence of an association exists, there was the potential that the relative 25(OH)D measure is associated with confounders, violating the independence assumption of MR.

7.3.6 Statistical analysis

The relative 25(OH)D measure was used in logistic regression models to estimate a causal OR for a unit increase in log2 25(OH)D. To correct for within region population structure, region specific principal components were added to the logistic regression models.

The standard error of this estimate was corrected to account for the imprecision of the effect of the genetic variant on 25(OH)D using a bootstrap approach in the 'boot' package in R.[200] Five hundred imputed relative 25(OH)D values were generated from a sampling distribution based on the beta coefficient and standard error from the 25(OH)D GWAS study. This set of newly imputed relative 25(OH)D values were used in logistic regression models, with cancer case status as the outcome, to generate a distribution of causal estimates. The standard deviation of this distribution was added to the standard error of the original causal OR to give a corrected standard error.

1. **STAGE 1: Create estimate of vitamin D (referred to as relative vitamin D) in each participant using Beta from GWAS and dosage of variant**

$$\alpha_{1\ldots j} = Beta\ from\ GWAS\ for\ variant_{1\ldots j}$$
$$Z_i = \ Dosage\ for\ variant_{1\ldots j}\ in\ individual$$
$$\hat{X}_i = estimate\ of\ unit\ change\ in \log 2\ 25(OH)D\ \ using\ variant_{1\ldots j}$$
$$\hat{X}_i = \ \alpha_1 Z_{1i} + \cdots + \alpha_j Z_{ji}$$

2. **STAGE 2: Estimate causal odds ratio of relative vitamin D on OC/OPC risk as the coefficient of the below logistic regression**

$$p_i = probability\ of\ HNC\ in\ idividual$$
$$logit(p_i) = \ \beta_0 + \psi \hat{X}_i$$

3. **CORRECTED STANDARD ERRORS: The SE from stage 2 does not consider the error in $\alpha_{1\ldots j}$, a bootstrap method was used to generate corrected SE for the second stage regression.**

$$B = number\ of\ bootstraps\ (500)$$
$$1)\ \ for\ b = 1\ to\ B;\ Generate\ \hat{X}_i^*$$

$$\hat{X}_i^{*b} = \alpha_1^{*b} Z_{1i} + \cdots + \alpha_j^{*b} Z_{ji}\,, \qquad \alpha_j^{*b} \sim (\alpha_j, se(\alpha_j)^2)$$

$$2)\ \ Generate\ distribution\ of\ bootstraped\ causal\ OR\ (\psi^*)$$

$$logit(p_i) = \beta_0 + \psi^* \hat{X}_i^*$$

$$3)\ \ Corrected\ se(\psi) \approx se(\psi) + se(\psi^*)$$

7.3.6.1 GAME-ON

MR analyses were performed within each geographic region of the GAME-ON consortium accounting for age, sex and the first 15 population specific principal components. As there were likely to be differences between the causal estimates across regions, a random-effects meta-analysis was used to combine GAME-ON causal estimates using the R package '*meta*'.[201] Heterogeneity between study populations was assessed using I^2.[202] Meta analyses were repeated excluding geographical regions that were outliers to assess for their influence on the overall estimate.

7.3.6.2 Replication in UK Biobank and meta-analysis

MR analyses were repeated in UK Biobank. GAME-ON and UK Biobank estimates were compared and then combined using meta-analysis. This employed a random-effects model and was conducted using the 'meta' package in R.[201] Heterogeneity between studies was assessed using I^2.[202]

7.3.6.3 Stratification by smoking status

In the observational analyses in Chapter 4, the effect of 25(OH)D on HNC risk differed depending on whether the participants smoked or not, with the effect only seen in current or former smokers. To assess whether the heterogeneity in effect size across smoking strata reflects true heterogeneity in the causal effect, MR analyses were repeated within each smoking strata in the GAME-ON study.

7.3.6.4 Sensitivity analyses

In order to assess whether any single genetic variant was driving a causal estimate, which could be driven by events such as horizontal pleiotropy which invalidate the exclusion restriction MR assumptions**,** a leave-one out method was applied. In this approach, the analysis is sequentially repeated, each time removing one variant from the genetic proxy used to derive relative 25(OH)D. To demonstrate this further, where a variant overtly influenced an association, each possible genetic proxy from combinations of the seven 25(OH)D variants, was used to estimate relative 25(OH)D and thereafter a causal OR. These OR were plotted as histograms highlighting those that include the highly influential variant to see if there was an obvious grouping of these estimates.

A UK Biobank nested focus study (UKBiLEVE),[203] investigating lung disease, preferentially selected and genotyped heavy and non-smokers using a different array to the remainder of the participants. 49,979 individuals were genotyped using the UKBiLEVE array and 438,398 using the UK Biobank axiom array. UK Biobank analyses were repeated removing UKBiLEVE participants.

7.4 Results

7.4.1 Study participants

In the GAME-ON study, after removing participants with <70% CEU ancestry, there were 5,133 cases (oral n=2,700, oropharyngeal n=2,433) and 5,984 controls). Participants from Europe and North America represented approximately 90% of cases and controls. In UK Biobank, after removing related and non-White British participants, there were 337,108 eligible participants, consisting of 585 cases (oral n=294, oropharyngeal n=291) and 336,523 controls. In GAME-ON the cases and controls were well matched for age whereas in UK Biobank, the cases were on average older than the controls. In both GAME-ON and UK Biobank there were more males and current smokers in the cases than there were in the controls (Table 7-2).

Table 7-2: GAME-ON and UK Biobank participant summaries

	GAME-ON			UK Biobank		
	All n (%) N=11117	Case n (%) N=5133	Control n (%) N=5984	All n (%) N=337108	Case n (%) N=585	Control n (%) N=336523
Site						
Oral		2700 (52.6)			294 (50.3)	
Oropharyngeal		2433 (47.4)			291 (49.7)	
Age						
=<50	2217 (19.9%)	1071 (20.9%)	1146 (19.2%)	65349 (19.4%)	51 (8.72%)	65298 (19.4%)
50-<60	3443 (31.0%)	1696 (33.0%)	1747 (29.2%)	94102 (27.9%)	180 (30.8%)	93922 (27.9%)
60-<70	3348 (30.1%)	1504 (29.3%)	1844 (30.8%)	153829 (45.6%)	298 (50.9%)	153531 (45.6%)
>=70	2108 (19.0%)	861 (16.8%)	1247 (20.8%)	23828 (7.07%)	56 (9.57%)	23772 (7.06%)
Missing	1 (0.01%)	1 (0.02%)	0 (0.00%)			
Sex						
Male	7680 (69.1%)	3798 (74.0%)	3882 (64.9%)	155747 (46.2%)	387 (66.2%)	155360 (46.2%)
Female	3437 (30.9%)	1335 (26.0%)	2102 (35.1%)	181361 (53.8%)	198 (33.8%)	181163 (53.8%)
Smoking status						
Never	3302 (29.7%)	1002 (19.5%)	2300 (38.4%)	184081 (54.6%)	208 (35.6%)	183873 (54.6%)
Previous	3655 (32.9%)	1590 (31.0%)	2065 (34.5%)	118478 (35.1%)	267 (45.6%)	118211 (35.1%)
Current	3300 (29.7%)	2019 (39.3%)	1281 (21.4%)	33364 (9.90%)	107 (18.3%)	33257 (9.88%)
Missing	860 (7.74%)	522 (10.2%)	338 (5.65%)	1185 (0.35%)	3 (0.51%)	1182 (0.35%)
Alcohol use						
Never	1825 (16.4%)	767 (14.9%)	1058 (17.7%)	10406 (3.09%)	10 (1.71%)	10396 (3.09%)
Ever	8422 (75.8%)	4030 (78.5%)	4392 (73.4%)	326467 (96.8%)	575 (98.3%)	325892 (96.8%)
Missing	870 (7.83%)	336 (6.55%)	534 (8.92%)	235 (0.07%)	0 (0.00%)	235 (0.07%)
Geographic region						
Europe	5251 (47.2%)	2323 (45.3%)	2928 (48.9%)			
North America	4583 (41.2%)	2254 (43.9%)	2329 (38.9%)			
South America	1283 (11.5%)	556 (10.8%)	727 (12.1%)			

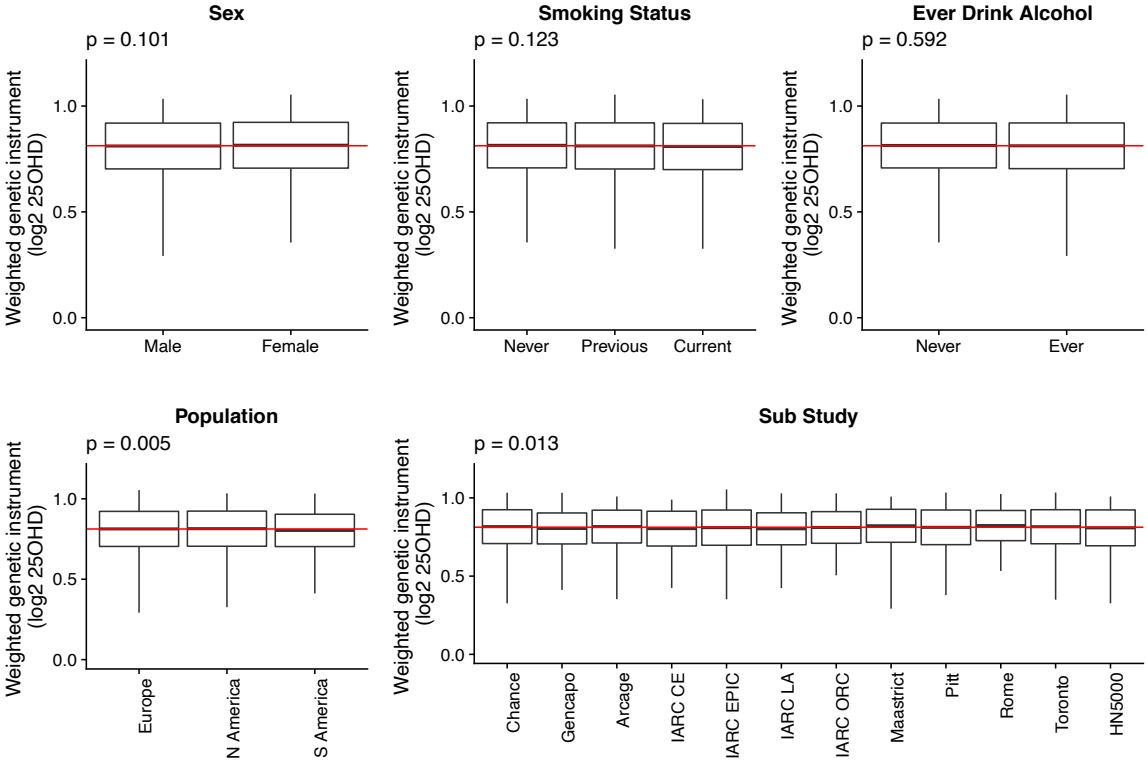
7.4.2 SNP Selection

Of the seven SNPs included in the analyses, rs4588 and rs10741657 were directly genotyped in both the GAME-ON and UK Biobank studies, rs2277458 was genotyped in the GAME-ON consortium . All other SNPs were imputed and the imputation quality metric (Minimac3 r^2 for GAME-ON; SNPTEST info score for UK Biobank) was greater than 0.98, except rs6013897 in the GAME-ON study which had an r^2 of 0.74 (Table 7-1) . No two SNPs on the same chromosome had an LD r^2 greater than 0.01.

7.4.3 Assessment of confounders

In the GAME-ON study there was no strong evidence of association of the 25(OH)D relative measure with sex, smoking status or drinking status after taking multiple testing into account. However, there was evidence for a difference in the 25(OH)D genetic proxy across regions and sub-studies in GAME-ON (Figure 7-1).

Figure 7-1: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders for GAMEON.

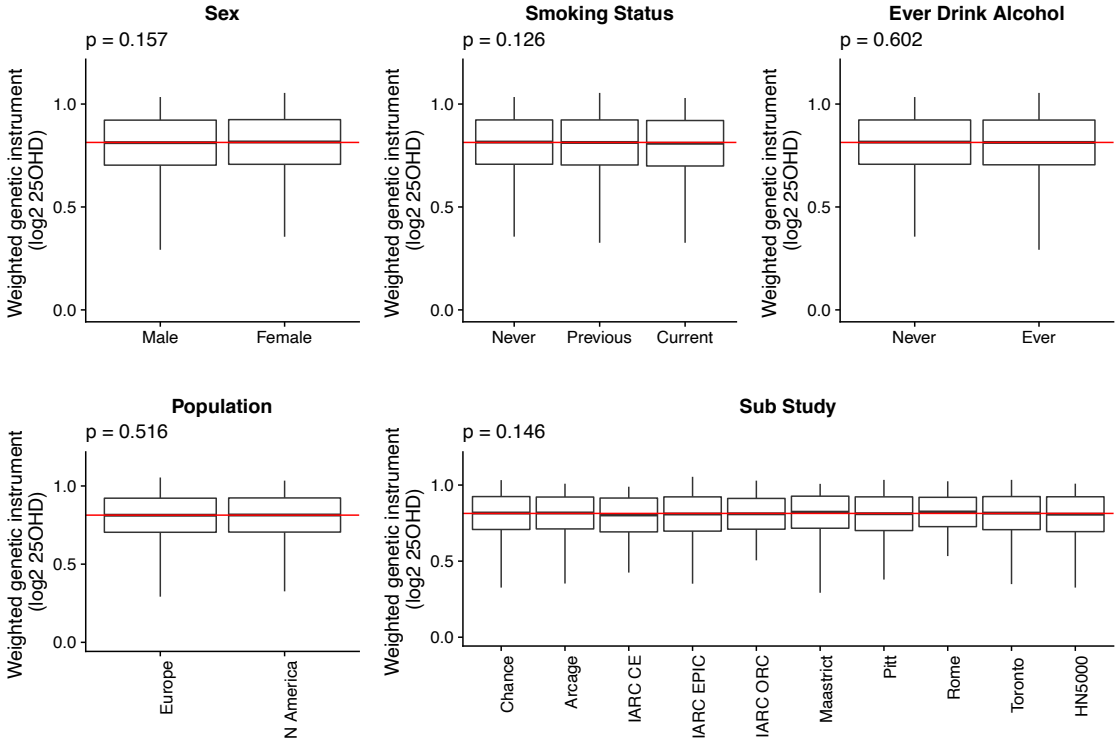


Overall mean relative 25-hydroxyvitamin D for those included in plot indicated by red line. 25OHD = 25 hydroxyvitamin D, p=p value for linear model

The region showing deviation from the others was South America and any differences across confounders were no longer seen when the South American region was excluded from analyses (Figure 7-2). Within the South American region alone there was no strong evidence for associations

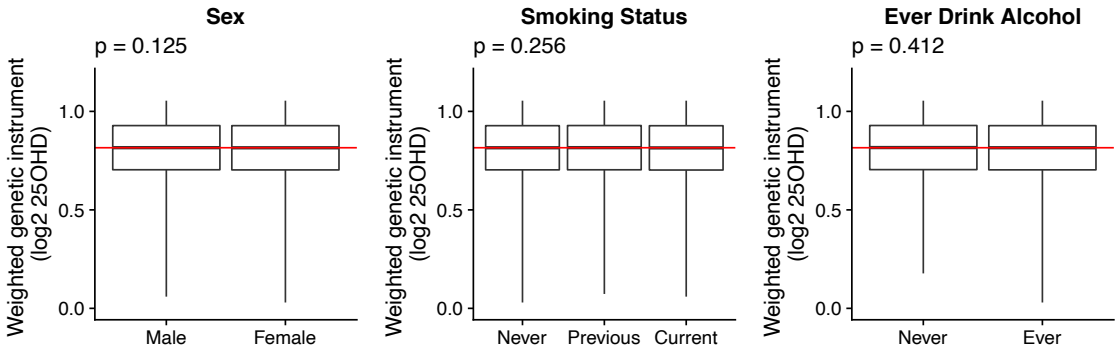
between the 25(OH)D genetic proxy and confounders (Appendix 7.1). In UK Biobank there was no evidence for a difference in relative 25(OH)D across strata of sex, smoking status or drinking status (Figure 7-3). These findings provide some evidence that the independence assumption is not violated by the relative 25(OH)D measure across the European and North American regions or within the South American region. However, there is some evidence that allele frequencies may differ in South America compared to North American and Europe, although these differences are small. This finding supports the decision to conduct the main analyses within each region and meta-analyse the effect estimates.

Figure 7-2: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders for GAMEON with the South American studies removed.



Overall mean relative 25-hydroxyvitamin D for those included in plot indicated by red line. 25OHD = 25 hydroxyvitamin D, p=p value for linear model

Figure 7-3: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders UK Biobank.



Overall mean relative 25-hydroxyvitamin D for those included in plot indicated by red line. 25OHD = 25 hydroxyvitamin D, p=*p* value for linear model

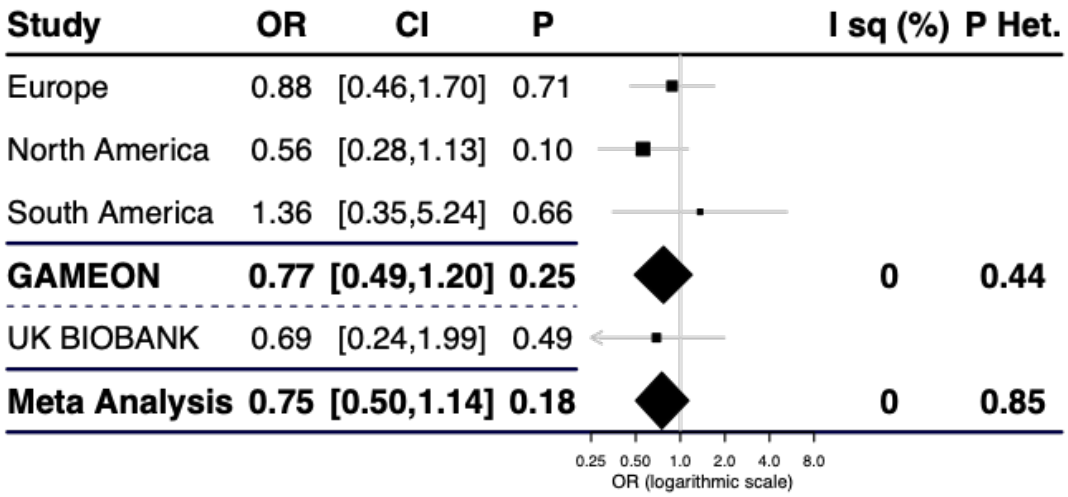
7.4.4 Causal association between 25(OH)D and oral and oropharyngeal cancer

In GAME-ON, for oral and oropharyngeal cancer there was little evidence for a causal association with relative 25(OH)D when the three geographic regions were meta-analysed, OR for a doubling of 25(OH)D = 0.77 (95% CI: 0.49, 1.20), *P*=0.25 and OR = 1.65 (95% CI: 0.55, 5.00), *P*=0.37 respectively. For oral sites, estimates were homogenous across geographic regions (I^2=0% (95% CI: 0, 87), P_{het}=0.44) but due to only having three estimates to meta-analyse, power to detect heterogeneity between regions was low. Despite this, there was evidence of heterogeneity between regions for the oropharyngeal sites (I^2=75% (95% CI: 17, 92), P_{het}=0.02). This was also reflected when all sites were examined together (I^2=58% (95% CI: 0, 88), P_{het}=0.09) and was most likely driven by the strong risk increasing association with oropharyngeal cancer identified in the South American region (OR = 13.81 (95% CI: 2.29, 83.09), *P*<0.01) compared to null associations in both European (OR = 0.98 (95% CI: 0.47, 2.03), *P*=0.95) and North American regions (OR = 0.90 (95% CI: 0.47, 1.84), *P*=0.76) (Figure 7-4).

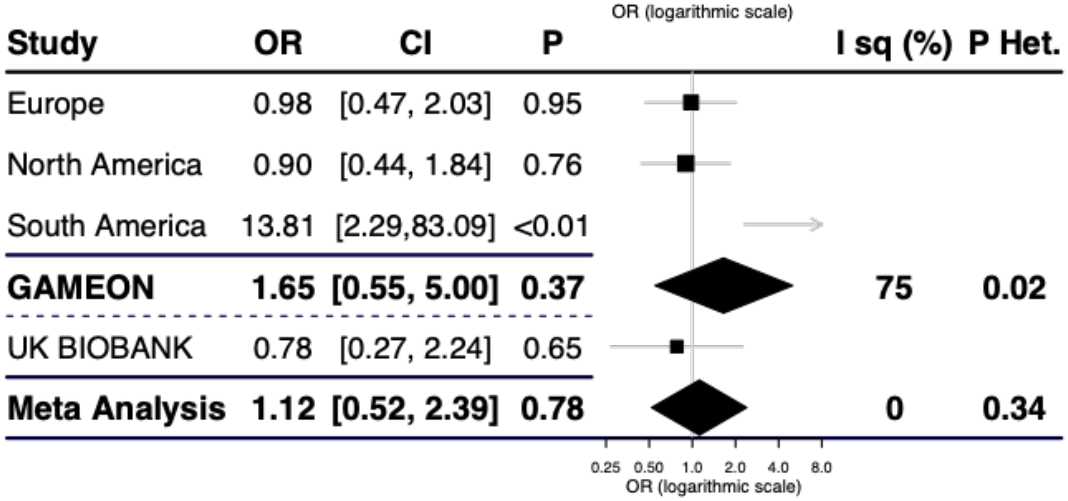
When replicated in UK Biobank, estimates for oral cavity cancer (OR=0.69 (95% CI: 0.24, 1.99), *P*=0.49) were similar to those from the GAME-ON study. For oropharyngeal sites (OR=0.78 (95% CI: 0.27, 2.47), *P*=0.65) point estimates differed between GAME-ON and UK Biobank although confidence intervals overlapped. When the two studies were meta-analysed using random-effects there was no convincing evidence of a causal association with either cancer or when all sites were analysed (Figure 7-4).

Figure 7-4: Odds ratio (95% CI) for developing cancer for D doubling in 25-hydroxyvitamin D for (A) oral cavity cancer, (B) oropharyngeal cancer and (C) all sites combined in Mendelian randomization analysis.

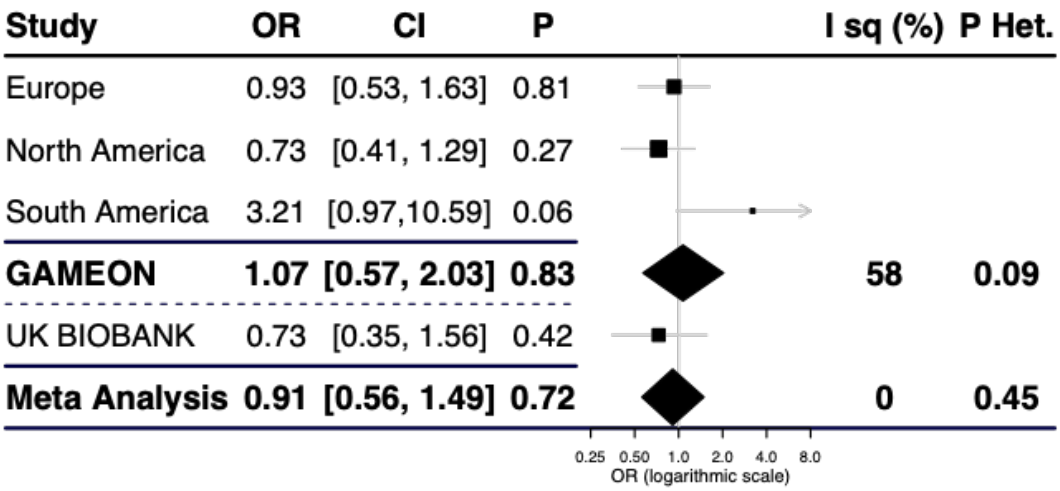
A



B



C

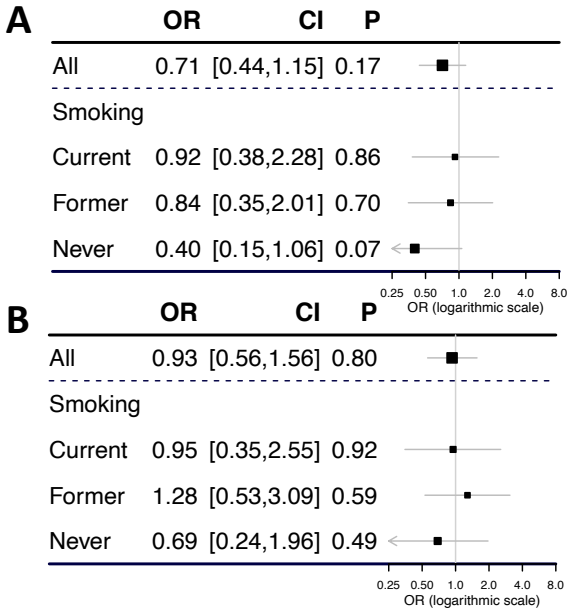


7.4.5 Stratification by smoking status

When stratified by smoking status, there was no clear pattern of effect across strata of smoking status. For both oral and oropharyngeal sites, the strongest protective effect of 25(OH)D was seen in

never smokers and the estimate for current smokers was on or near the null (Figure 7-5). This is the opposite of the observational findings which see any protective effect of 25(OH)D on HNCs driven by the current smoking strata (Chapter 4).

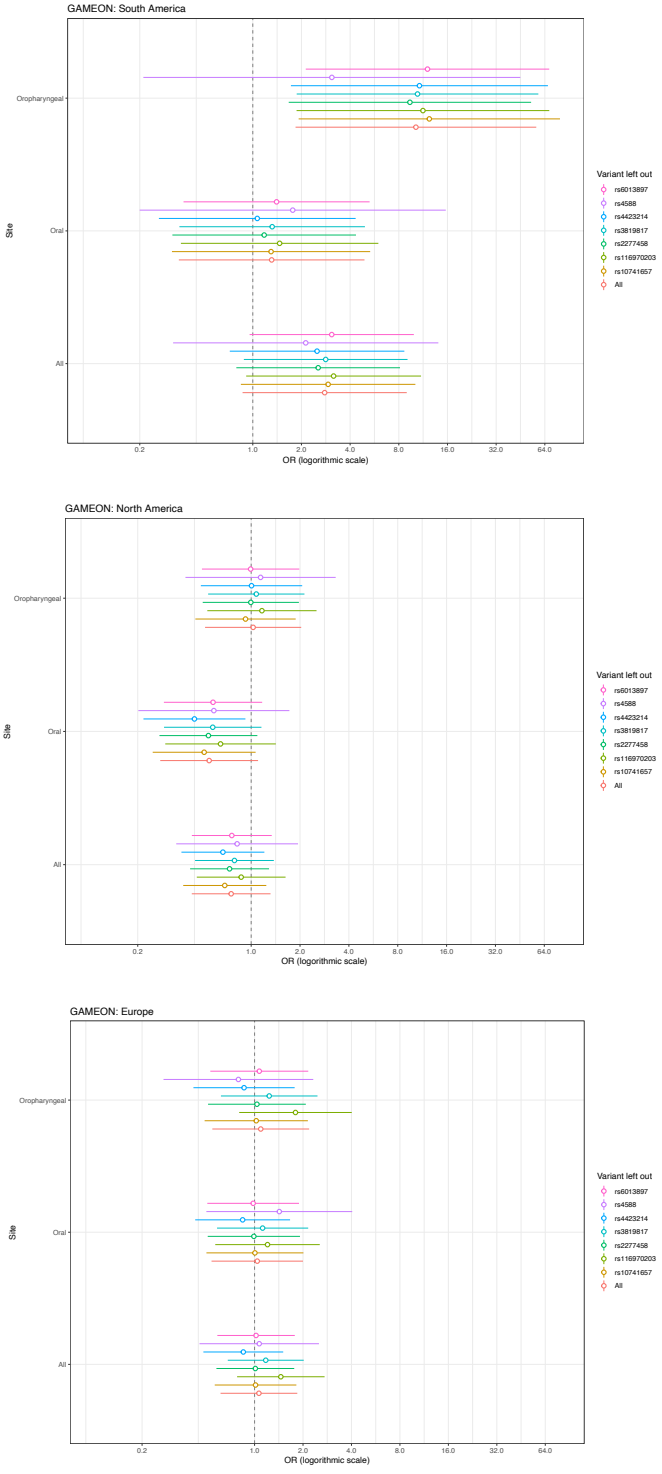
Figure 7-5: Odds ratio (95% CI) for developing cancer for doubling in 25-hydroxyvitamin D stratified by smoking in (A) oral cavity cancer and (B) oropharyngeal cancer.



7.4.6 Sensitivity analyses

In leave one out analyses, sequentially omitting each of the seven variants provided similar causal estimates for any site within the European and North American regions of GAME-ON. This is demonstrated in Figure 7-6 A and B, where the effect estimates differ very little when each of the genetic variants are left out of the analysis. However, in the South American region of the GAME-ON study, the risk increasing association between 25(OH)D and oropharyngeal cancer was most strongly influenced by the rs4588 variant. This is demonstrated clearly in Figure 7-6 C, where the effect estimate not including rs4588, is substantially closer to the null compared to the estimates leaving out the other genetic variants.

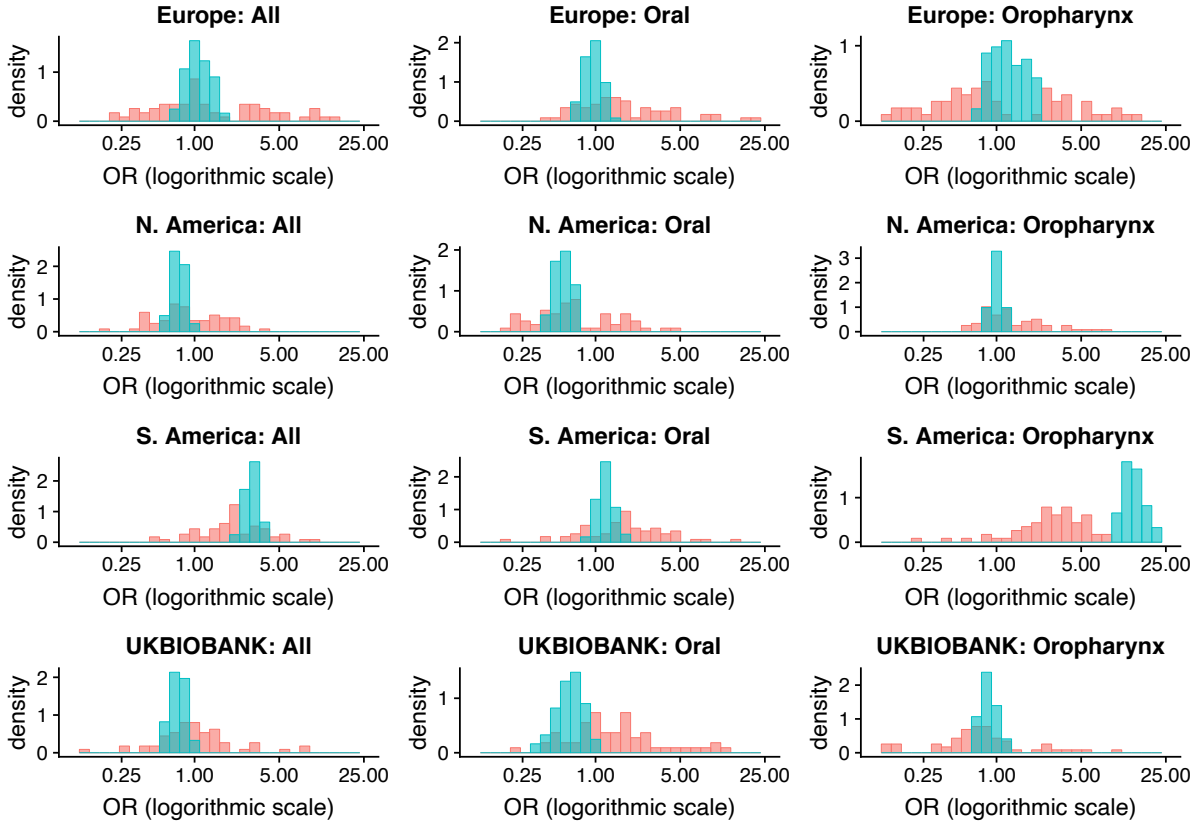
Figure 7-6: Leave-one-out plots for the three GAME-ON regions.



To investigate the influence of the rs4588 variant further, analyses were repeated using the relative 25(OH)D measure consisting of all possible combinations of the 25(OH)D variants with one variant, then two variants then three variants, all the way to seven variants. In the South American region, for oropharyngeal cancer, whenever the rs4588 variant was included, there was a much stronger odds increasing effect of 25(OH)D compared to when it was not included in the relative 25(OH)D measure. However, even with the rs4588 variant removed there was still an increased risk

association for oropharyngeal cancer in the South American region (Figure 7-7). Removing the South American region from the oropharyngeal analyses reduced heterogeneity between the GAME-ON regions with the effect estimate suggesting little evidence of a causal effect of 25(OH)D on oropharyngeal cancer (OR = 0.93 (95% CI: 0.56, 1.56), P=0.80; I^2 = 0, P_{het}=0.87). When combined with the UK Biobank estimate, the pooled estimate made little difference to the interpretation (OR = 0.90 (95% CI: 0.57, 1.43), P=0.67; I^2 = 0, P_{het}=0.77).

Figure 7-7: Histograms of OR for MR analyses repeated for each possible combination of 25-hydroxyvitamin D genetic proxies.



Proxies including rs4588 variant (•), proxies not including rs4588 variant (•).

7.5 Discussion

This study set out to investigate the causal association between 25(OH)D and oral and oropharyngeal cancer risk. Given the lack of 25(OH)D measures in large scale oral and oropharyngeal cancer studies a two-sample MR approach was adopted[204] that allowed genetic variants to be used as largely unconfounded measures of 25(OH)D. Strong evidence of a causal association was not identified with oral or oropharyngeal cancer. The point estimate for the association of 25(OH)D with oral cavity cancer (OR = 0.75 (95% CI: 0.50, 1.14)) was smaller but had some consistency with the observational estimates identified in Chapter 4 (OR = 0.71 (95% CI: 0.57, 0.88)) however, confidence

intervals were wide in these MR analyses and include the null. The MR estimate for oropharyngeal cancer form the meta-analysis of GAME-ON and UK Biobank (OR = 1.12 (95% CI: 0.52, 2.39)), was in the opposite direction to that of the observational results in Chapter 4 (OR = 0.80 (95% CI: 0.65, 0.98), however there was clear heterogeneity across the regions of the GAME-ON study. When removing the estimate from the South American participants the pooled GAME-ON, UK Biobank MR estimate was imprecisely estimated but, similar to the oral cavity cancer result, smaller but potentially consistent with the observational estimate.

The lack of strong evidence for a causal association between 25(OH)D and cancer risk mirrors a recently published MR study of colorectal, breast, prostate, ovarian, lung and pancreatic cancer, and neuroblastoma that did not support a causal role of 25(OH)D cancer risk.[205] Similar to the inference from this work on oral and oropharyngeal cancer, the authors of the previous study state that although they are unable to rule out clinically relevant effects of small magnitude, their study combined with previous literature, "provide evidence that population-wide screening for vitamin D deficiency and subsequent widespread vitamin D supplementation should not currently be recommended as a strategy for primary cancer prevention". Recent RCTs, including the large VITAL study, examining the effect of vitamin D on cancer incidence do not show strong evidence of a protective effect and support the findings from MR studies.[206,207] Contrary to this however, Ong *et al.* (2016)[208] did demonstrate a protective role for 25(OH)D in ovarian cancer in a well powered MR study but the more recent pan-cancer MR study seems to refute this. Despite the body of evidence that does not support a protective effect of 25(OH)D on cancer incidence there is some evidence to support the hypothesis that vitamin D has an effect on cancer prognosis[54,126] and this is supported by the observational results in Chapter 4. Recently a large meta-analysis of RCTs for vitamin D supplementation on all-cancer mortality demonstrated a protective effect with supplementation (RR = 0.84 (95% CI: 0.74, 0.95)).[209] Given the lack of RCT and MR evidence supporting a strong role for vitamin D on cancer risk, it may be that the effect on mortality seen in this recent meta-analysis is driven by the effect of 25(OH)D on cancer prognosis. The role of 25(OH)D on HNC prognosis is examined in Chapter 8.

The genetic relative 25(OH)D measure used in this study specifically proxies average total (free and bound) circulating 25(OH)D. This measure does not necessarily predict the concentrations of free 25(OH)D available at the tissues or concentrations of free or bound $1,25(OH)_2D$ in circulation or at the tissue level. Potential anti-cancer effects of vitamin D are via free $1,25(OH)_2D$ interacting with the vitamin D receptor within tissues resulting in reduced angiogenesis, metastasis, cell invasion, inflammation, and proliferation, as well as upregulation of apoptosis.[54] Levels of $1,25(OH)_2D$ are assumed to be correlated with circulating 25(OH)D, but this is not necessarily the case. Although it is

unlikely, the genetic proxy may not be a valid measure for 25(OH)D and this is discussed further in Chapter 9. However, as the genetic variants used here as genetic proxies for 25(OH)D have robust associations with 25(OH)D and are mainly in genes known to affect this metabolite they are likely to be valid for the purposes of instrumenting serum 25(OH)D in MR.

As highlighted previously, in order to be valid the genetic measure must proxy circulating 25(OH)D without affecting cancer risk through other causal pathways (violation of this assumption is referred to as horizontal pleiotropy). The SNPs used in this proxy are in genes with known effects on vitamin D pathways and have been consistently associated with circulating levels of 25(OH)D in GWA[98,178] and MR studies,[105] reducing the chance of horizontal pleiotropy. Sensitivity analyses such as MR Egger[6] and the weighted median method[150] can detect or correct for the presence of directional pleiotropy but have low power with few SNPs comprising the genetic proxy, as in this case. Here, causal estimates in the European and North American populations are largely consistent across all five SNPs, providing some evidence against the presence of strong pleiotropy that could bias findings.

In the South American population, some evidence for a large risk increasing association was identified between 25(OH)D and oropharyngeal cancer. Use of individual level data in this study allowed scrutiny of the genetic proxy-confounder independence assumption and showed that the 25(OH)D proxy generally held up to this assumption however there was some evidence the genetic proxy was associated with region and sub-study, driven by the South American region. As mentioned above it is unlikely then that pleiotropy can explain the large risk increased effect seen for oropharyngeal cancer in the South American region. There is also little evidence that the genetic proxy for 25(OH)D is associated with potential confounders, as demonstrated in this work and in the PheWAS in Chapter 6. However, as the risk increasing association seen between 25(OH)D and oropharyngeal cancer in the South American region, is inconsistent with observational effect directions and MR estimates from other geographic regions, it is unlikely to be truly causal: it most likely represents structure within this population that is by chance associated with oropharyngeal cancer risk. To assess the impact of this potentially biased estimate on the overall causal OR, the analyses were repeated with the South American individuals removed, resulting in more consistent estimates. Future studies could firstly use GWAS results for 25(OH)D which are specific to the South American population which would reduce problems with the genetic instrument potentially not being relevant to this population. Furthermore increasing the numbers of participants from the South American regions would improve power to produce a more precise estimate of the causal effect and therefore help to clarify whether these interesting finding are real or not.

In observational studies (including work from this thesis), the HNC risk decreasing effects of 25(OH)D are seen to be strongest in current smokers, with no evidence of an effect seen in never smokers. Despite being underpowered to show clear differences across smoking strata, the trend of effect sizes in this study does not match the observational analyses and shows no clear pattern. This result does not support the theory that smoking and vitamin D interact to make vitamin D protective but only in smokers.

The genetic 25(OH)D relative measure is a good proxy for circulating total 25(OH)D, the same metabolite that is used in supplementation and is measured in the observational studies**.** Despite the improved proxy identified in Chapter 5 which has been shown to be robustly associated with 25(OH)D, the genetic variants used here only explain a relatively small amount of variance in 25(OH)D. This means that causal effects of small magnitude cannot be ruled out. A new GWAS, currently being conducted, has identified a large number of new genetic variants that are associated with 25(OH)D. These variants will improve the proxy 25(OH)D measure beyond that of this thesis and increase power for future MR analyses. Furthermore, a larger case-control HNC study with genetic data will also help to provide more precise causal estimates from MR and, a collaborative effort is underway to achieve this. These new initiatives, to which I am contributing, are both discussed in the Future Work section in Chapter 9.

Using MR, it cannot be determined whether 25(OH)D has an effect on cancer risk in individuals below a certain 25(OH)D concentration threshold. Where inference is likely to be useful clinically is in relation to supplementation. Supplementation increases total circulating 25(OH)D and is presumed to subsequently increase free $1,25(OH)_2D$ availability at the tissues. The effect sizes in this chapter reflect a doubling in 25(OH)D to mirror those in previous observational studies. In the general population (with normal 25(OH)D levels), this would reflect a very large increase in 25(OH)D, likely to be larger than would be achieved by vitamin D supplementation. The 25(OH)D difference in this study is genetically determined and will have a relative effect on total 25(OH)D across the whole life course. If this long-term difference in relative 25(OH)D does not produce a detectable and clinically useful effect in this study, any 25(OH)D causal effect is unlikely to be of a magnitude that would warrant long-term supplementation in the general population.

7.6 Conclusion

This chapter does not support the large size of the observational association between 25(OH)D and oral cavity cancer risk and is consistent with evidence that a causal, clinically relevant protective effect of 25(OH)D on oropharyngeal cancer risk is unlikely.

The next chapter

As detailed above there is a growing body of evidence from observational and randomised studies that 25(OH)D has an effect on cancer mortality, even if not on cancer risk. The effect of 25(OH)D on oral and oropharyngeal cancer prognosis was not assessed in this chapter due to the lack of prognosis data in the GAME-ON contributing studies. In the next chapter, a study assessing the causal effect of 25(OH)D on HNC prognosis is conducted on the currently available data.

Chapter 8 : Assessing the causal association of vitamin D with HNC progression

8.1 Chapter objectives

In Chapter 4 the observational association between 25(OH)D and HNC prognosis was investigated and strong protective associations were identified however these estimates are likely biased by confounding and other biases specific to case only analyses. In this chapter, an MR analysis is conducted using the genetic proxy identified and tested in Chapter 5 and 6. The aim of this chapter is to estimate the causal association between vitamin D and HNC prognosis using MR.

8.2 Introduction

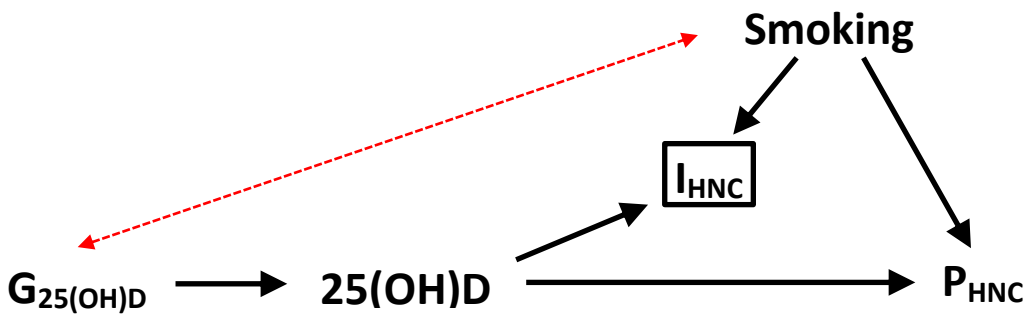
As reported in the summary of previous observational evidence in Chapter 2 and demonstrated in the study described in Chapter 4, there is a growing body of observational evidence that higher 25(OH)D confers a survival advantage in HNC patients. This protective effect is robust to correction by confounding and when 25(OH)D measures are recorded before the onset of cancer (for example five years prior to diagnosis). Despite these results, these estimates from observational studies are likely to be affected by the issues of residual confounding that were described above for the association with HNC incidence. MR can be used to investigate the association between 25(OH)D and HNC prognosis but statistical and logistical issues exist which make this less straight forward.

One major concern in case only analyses, including MR, is that when selecting on case status an ascertainment or selection event occurs which can introduce bias into any subsequent analyses. This type of bias, referred to as ascertainment or collider bias, is explained in more detail in Chapter 3. In the context of this chapter, any factors which cause an individual to develop HNC will become correlated in case only analyses of HNC. Therefore, if vitamin D does cause HNC then, in the case only setting, there will be induced correlations between vitamin D and any other cause of HNC. As will be discussed in Chapter 9, it is not clear whether there is a causal association between vitamin D and HNC risk and so the implication of collider bias is unknown.

In this example, vitamin D is assumed to be causally related to HNC incidence and smoking is used as a common cause of HNC. In an observational study, we would expect there to be associations between vitamin D and factors such as smoking, even in the non-ascertained (case-control or cohort) study but, in the ascertained (case-only) study these associations may be altered in magnitude and or direction. The collider bias from this induced or altered association with smoking can be unpredictable but can be partially mitigated by correcting for smoking in the analyses.

However, much like when correcting for confounding, there can be residual bias if there is error in the measurement of smoking. One of the major benefits of MR analyses over observational studies is the genetic proxies are uncorrelated with the potential confounders of the exposure-outcome relationship. However, in the case only setting, genotypes which proxy causal risk factors will become correlated, inducing associations between the genetic proxies, of vitamin D in this case, and other risk factors such a smoking. In an MR study of vitamin D on HNC prognosis, this creates a pathway between the genetic variants for vitamin D and HNC prognosis, which is not via vitamin D, violating the independence assumption of MR (Figure 8-1).

Figure 8-1: Collider bias in the context of an MR study assessing the causal association between 25(OH)D and HNC prognosis.



The genetic variant for 25(OH)D ($G_{25(OH)D}$) is causally related to the incidence of HNC (I_{HNC}) through 25(OH)D. Smoking also causes I_{HNC}, is unrelated to $G_{25(OH)D}$ but also affects HNC prognosis (P_{HNC}). To investigate the effect of 25(OH)D on the prognosis of the disease (P_{HNC}), an MR analysis in conducted in cases only (indicated by the square around I_{HNC}). In the cases only $G_{25(OH)D}$ is non-randomly assigned and a spurious association is induced between $G_{25(OH)D}$ and smoking (red dotted arrow). In the case only MR analysis, the causal effect of 25(OH)D – $P_{25(OH)D}$ is biased via the aternative route from $G_{25(OH)D}$ through smoking.

One problem in all studies which assess the process of disease, is what measure to use to as the prognosis outcome. The most commonly used measure is survival and this will be used here. It is routinely recorded from death registry data and so available on all participants. One method is to generate a binary variable of survived versus not survived at a set timepoint. The approach has the advantage of being more amenable to current MR methods but there is a large amount of information loss as the amount of time in the study is not taken into account. In this study time to death will be used as it makes the best use of available information but MR methods using this type of data are less developed.

There are very few instances in the literature of MR analyses being applied to survival data. One study used MR to assess the effect of BMI on all-cause mortality in UK Biobank,[210] using a genetic risk score (GRS) in an instrumental variable ratio method. In this study, the authors calculated the effect of the GRS on the exposure (BMI) as the denominator of the ratio and the GRS was used in a cox-proportional hazards model to calculate the effect of the GRS on all-cause mortality, as the

numerator of the ratio. The causal hazards ratio (HR) was calculated as the exponential of the ratio of the GRS effect on mortality to the GRS effect on BMI. Confidence intervals were calculated using a Taylor series expansion.

A more recent paper assessed the effect of BMI on breast cancer survival.[211] In this study two MR methods were used to assess the causal association, both estimating a HR. The first method used a weighted GRS in a Cox proportional hazards model, similar to that used in the previous chapter for HNC risk. The second method used a two-sample approach in which GWAS summary statistics for BMI and breast cancer progression were utilised in a method identical to that which would be used to generate an OR.

The main logistical challenge is that there are limited sample sizes available for well powered studies due to the difficulty in getting long-term HNC follow-up data in large enough samples. Genetic data and 25(OH)D measures are not both available in HNC specific studies with HNC follow data, and due to the relatively rare incidence of HNC, there are few cases in large population-based studies which do have 25(OH)D and genetic data. These issues make single sample MR impractical. Furthermore, studies with HNC follow up data and genetics are limited and there are no published GWAS of HNC prognosis which, would allow for a two-sample MR approach using summary level data. Given the relatively small proportion of variance in 25(OH)D explained by the 25(OH)D genetic variants and the small samples with genetics and HNC follow-up data available, a well powered MR study is not possible at this time.

Despite the issue of power, there are advantages to conducting an MR study looking at the effect of 25(OH)D on HNC prognosis. A study of this type will provide preliminary estimates of effect to inform sample size calculations for well powered MR studies and allow testing of the methodologies discussed above which have not been formally compared.

In this chapter several MR methods will be utilised to study the effect of 25(OH)D on HNC prognosis using available data in Head and Neck 5000 and UK Biobank. The reader is pointed to the Future Work section of Chapter 9, for a description of how this study will be developed in the future.

8.3 Methods

8.3.1 Overview

Genetic variants which proxy 25(OH)D levels were utilised in this chapter to conduct an MR study assessing the causal effect of 25(OH)D on oral and oropharyngeal cancer prognosis in two case only studies using three MR methods. All methods estimate the causal effect as a HR, two using individual level data and one using summary level data.

As in the previous chapter, genetic variants identified in the 25(OH)D GWAS in Chapter 5 were used as 25(OH)D proxies. The genotypes of these variants were identified in individuals of two studies with follow up of HNC cases (HN5000 and UK Biobank) and combined to make a 25(OH)D genetic risk score (GRS). To assess the assumption that this 25(OH)D GRS was unconfounded in the ascertained (case only) setting, the association between the GRS and available confounders was assessed in UK Biobank in both the whole cohort and the ascertained, HNC case only, participants.

The first MR method was conducted in both HN5000 and UK Biobank as serum 25(OH)D levels were not required. The 25(OH)D GRS was scaled to match 25(OH)D on the log2 scale then used in a Cox-proportional hazards survival regression model to estimate a HR for a log2 increase in 25(OH)D.

The second method applied the ratio method discussed above to assess the effect of BMI on all-cause mortality in UK Biobank. As serum 25(OH)D levels were required this was only conducted in UK Biobank, and not HN5000.

In the final method, a two sample MR analysis was used to estimate a causal HR using summary level data. Summary level data for 25(OH)D was used from published GWAS and summary level data for HNC prognosis was used from a GWAS of HNC survival conducted in UK Biobank and HN5000 for the purposes of this chapter.

8.3.2 Studies

8.3.2.1 Head and neck 5000

HN5000 is a large collection of HNC cases with detailed phenotyping and follow-up. An overall description of the HN5000 study is provided in Chapter 3.

8.3.2.2 UK Biobank

UK Biobank is a population-based study of approximately 500,000 participants. A detailed description of this study, genotyping and the 25(OH)D measures is given in Chapter 3.

8.3.3 DNA extraction genotyping and imputation

DNA extraction, genotyping and imputation for UK Biobank and HN5000 is described in Chapter 3.

8.3.4 Follow-up of HNC cases

Cancer cases were followed up in both studies using death registry data to identify the date of death. All-cause mortality was identified from death registry data as a death of a participant from any cause. Due to the limited sample size (UK Biobank) and data-availability (HN5000), HNC-specific cause of death was not assessed in this study.

Survival time was calculated from date of diagnosis to date of death. Those participants alive at the end of follow up were censored. Censoring was on 31st March 2018 in HN5000 (last data capture from death registry) and 14th February 2018 in UK Biobank (last recorded death in the UK Biobank data freeze used for this study).

8.3.5 Vitamin D genetic variants for MR analysis

The seven genetic variants used in Chapter 7 were used as a genetic proxy for 25(OH)D in MR analyses. As in Chapter 7, for each individual the seven 25(OH)D variants were combined into a GRS weighted by the beta coefficients of the genetic variant's effect on 25(OH)D levels, derived from GWAS.

For the assessment of confounders and the first MR method the GRS was converted from a per allele scale to a log2 25(OH)D scale. Therefore, each unit increase in this variable represented a doubling of 25(OH)D. For the second MR method, the GRS was used on the per allele scale.

For the summary level data the same seven genetic variants were identified from the largest published GWAS of 25(OH)D.[193] Effect estimates and standard errors were scaled to match the log 2 scale. As the low frequency variant in frequency variant in *CYP2R1*, identified in Chapter 5, was not available in this GWAS, the effect estimate identified in this thesis was used and scaled to match that of the other six variants, as done in previous chapters.

8.3.6 GWAS of HNC prognosis to generate summary level outcome data

A GWAS was conducted for HNC survival in HN5000 and UK Biobank using the SurvivalGWAS_Sv program.[212] Full details of the method are described in Appendix 8.1. Briefly, a Cox proportional hazards regression model was run for each genetic variant. The number of risk alleles at each variant was used as the exposure and time to event data for all HNC sites combined (as detailed above) was used as the outcome, age and sex were used as covariates in the regression model. Results from the two studies were meta-analyses using a fixed effect in GWAMA.[187] Summary effect estimates (log HR) and standard errors were extracted for the seven 25(OH)D genetic variants for the pooled results and the individual studies.

8.3.7 Statistical analysis

8.3.7.1 Assessment of GRS in UK Biobank

The strength of the GRS as a proxy measure for 25(OH)D was assessed using the F-statistic from a linear regression model of the GRS on seasonally adjusted 25(OH)D. The F-statistic gives a measure of the strength of an instrumental variable in MR analyses. As discussed in Chapter 3, where an instrumental variable is weak the analyses can be affected by weak instrument bias. This is where

chance associations between the instrumental variable and confounders outweigh the direct effect of the instrumental variable on the outcome and can bias the causal estimate. The R^2 was used to assess the proportion of variance in 25(OH)D which is explained by the GRS. This was performed in the HNC case only sample and compared to the same metrics in the whole UK Biobank dataset. A scatter plot was used to demonstrate the correlation between individual level log2 25(OH)D and the GRS with the linear relationship indicated.

8.3.7.2 Assessment of confounding in case only setting

To investigate the potential impact selecting only a set of HNC cases has on further analyses, the association between vitamin D and a series of variables was assessed in both the full UK Biobank cohort and the selected HNC cases. Smoking status and self-reported number of sexual partners (a proxy measure of HPV status) were selected as they are strong risk factors for HNC and have been shown to influence survival.[213,214] The association between these variables and both 25(OH)D and the GRS was assessed using a linear regression model. The magnitude and direction of effects was compared between the full and selected study samples.

8.3.7.3 MR method 1: GRS analysis in HN5000 and UK Biobank

In both studies the GRS was dichotomised at the median value and Kaplan Meier survival plots were generated to graphically represent the survival in "high" and "low" GRS individuals. This was repeated for the 25(OH)D levels in UK Biobank, again using the median 25(OH)D level as the cut point.

In both HN5000 and UK Biobank, the effect of the GRS on survival in the HNC cases was assessed using a cox-proportional hazards model using the 'survival' package in R.[215] Age at diagnosis and sex were included as the only covariables in the regression models as the GRS had been shown to be unrelated to potential confounders. In UK Biobank, including genetic principal components in the model made minimal impact on the effect estimates and were not included in further analyses for either cohort. In UK Biobank where 25(OH)D was available, the observational estimate in those included in the GRS analysis was calculated using the same regression model.

The estimates from HN5000 and UK Biobank, generated using the GRS, were meta-analysed using fixed effects in the 'meta' R package.[201]

8.3.7.4 MR method 2: Two-stage ratio method in UK Biobank

The second method involved two stages and was only conducted in UK Biobank as serum 25(OH)D levels were required. In this method, a GRS was generated for 25(OH)D scaled to a per allele effect. In the first stage, a beta and standard error for the effect of the GRS on 25(OH)D were estimated

from a linear regression model. In the second stage the beta and standard error for the effect of the GRS on HNC survival were estimated from a Cox proportional hazards model. The causal HR was generated by dividing the beta from the second stage by the first stage. Standard errors for the causal estimate were estimated using a Taylor series expansion which takes into account the imprecision of both stages of the method.

8.3.7.5 MR Method 3: Two-sample MR using summary statistics

The R package 'TwoSampleMR'[216] was used to conduct the MR analyses with option "2" used to harmonise the data. With this option, positive strand alleles were inferred using allele frequencies for palindromic variants. An inverse variant weighted method was used to provide the causal estimate which was compared to the observational and MR estimates from the first method. The MR analysis was conducted for the pooled HN5000 and UK Biobank GWAS summary results and then the studies individually. This was to allow direct comparison between the different MR methods in the same studies.

8.3.7.6 Sensitivity analyses

To assess whether any one genetic variant is overtly influencing the causal estimate a leave one out sensitivity analysis was performed for both the GRS and two-sample MR methods. This repeats the primary analysis sequentially removing one of the seven genetic variants. If any of the resulting estimates differ substantially from the others, this may provide evidence that the genetic variant is pleiotropic.

8.4 Results

8.4.1 Study participants

In the HN5000 study, there were 1,020 cases (oral n=395, oropharyngeal n=577, larynx = 6, other HNC=41) with follow-up and genetic data. Median follow up was 4.5 years (range: 11 days, to 6.4 years) and there were 335 (32.8%) deaths from any cause during this period. This corresponds to a follow period of 3,918.7 person-years. In UK Biobank, there were 457 cases (oral n=136, oropharyngeal n=212, larynx = 83, other HNC=27) with follow up and genetic data. Median follow up was 4.4 years (range: 1 day to 10.3 years) and there were 111 (24.3) deaths from any cause during this period. This corresponds to a follow period of 2,035.2 person-years.

8.4.2 SNP selection

Of the seven SNPs included in the analyses, rs117913124, rs10741657 and rs12785878 were directly genotyped in HN5000 and rs10741657, rs12785878 and rs8018720 were directly genotyped in UK Biobank. All other SNPs were imputed and the imputation quality metric (Minimac3 r^2 for HN5000;

SNPTEST info score for UK Biobank) was greater than 0.96, except rs6013897 which in the HN5000 study had an r^2 of 0.75 (Table 8-1). No two SNPs on the same chromosome had an LD r^2 greater than 0.01.

Table 8-1: 25-Hydroxyvitamin D genetic variant details

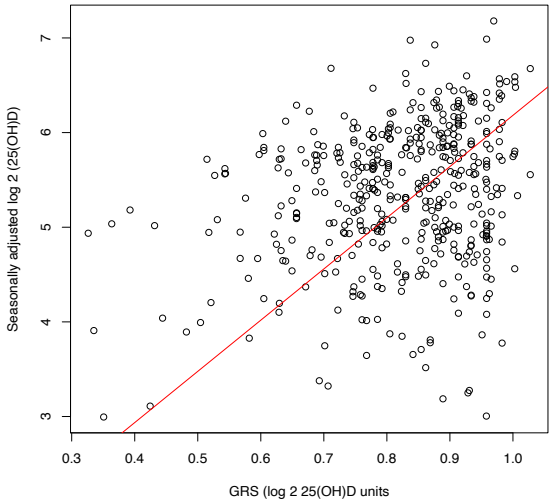
						Vitamin D GWAS				HN5000		UK Biobank	
RSID	Chromosome	Position	Gene	Effect Allele	Other Allele	EAF	Beta	SE	P value	Rsq (minimac3)	EAF	Info (snptest)	EAF
rs3755967	4	72609398	GC	C	T	0.713	0.128	0.0033	4.74E-343	1.0	0.702	1.0	0.713
rs117913124	11	14900931	CYP2R1	G	A	0.973	0.227	0.0093	2.29E-88	Genotyped	0.965	0.99	0.973
rs10741657	11	14914878	CYP2R1	A	G	0.398	0.045	0.0032	2.05E-46	Genotyped	0.393	Genotyped	0.398
rs12785878	11	71167449	DHCR7	T	G	0.758	0.052	0.0032	3.80E-62	Genotyped	0.777	Genotyped	0.758
rs10745742	12	96358529	AMDHD1	T	C	0.384	0.025	0.0017	1.88E-14	0.98	0.363	1.00	0.384
rs8018720	14	39556185	SEC23A	G	C	0.178	0.025	0.0042	4.72E-09	0.99	0.187	Genotyped	0.178
rs17216707	20	52732362	CYP24A1	T	C	0.812	0.038	0.0039	8.14E-23	0.75	0.817	0.96	0.812

Beta = per allele effect on 25(OH)D on log2 scale, EAF = effect allele frequency, Rsq = R squared, SE = standard error

8.4.3 Assessment of GRS

The strength of the GRS, as a proxy measure of 25(OH)D, was assessed in UK Biobank. The F-statistic from the linear model of the GRS on seasonally adjusted 25(OH)D was 28.2, in the sample included in survival analyses with the GRS explained approximately 5.8% of the variance in 25(OH)D. Figure 8-2 demonstrates the relationship between the GRS and actual 25(OH)D levels in the analysed samples. In the full UK Biobank sample, the F statistic was much higher (F=1,453) due to the larger sample size (n= 421,686) but the variance in 25(OH)D explained by the GRS in the full UK Biobank sample was lower (3.3%).

Figure 8-2: Scatter plot demonstrating the relationship between GRS and 25(OH)D levels.



8.4.4 Assessment of confounding in case only setting

To assess how selecting only cases may influence further analyses, the association between vitamin D and smoking, number of sexual partners and mean corpuscular haemoglobin concentration was compared in the case-only and full UK Biobank samples. When using the serum 25(OH)D measure, there was strong evidence for lower 25(OH)D in former and current smokers compared to never smokers in the full UK Biobank sample and in the case only sample. However, the effect of smoking on 25(OH)D was of a larger magnitude in the case only sample compared to the full UK Biobank sample. When assessing the association between the GRS and smoking there was little evidence that the GRS differed across smoking strata in either the case-only or full UK Biobank samples, with differences in effect likely only reflecting the lower precision in the case-only sample (Table 8-2).

Table 8-2: Association of 25(OH)D and smoking in full and selected UK Biobank samples.

		Full UK Biobank sample			Case only sample		
	Smoking status	Beta	95% CI	P	Beta	95% CI	P
Serum 25(OH)D	Never	REF			REF		
	Former	-0.138	-0.142, -0.133	$<2x10^{-16}$	-0.401	-0.519, -0.283	$8.31x10^{-11}$
	Current	-0.094	-0.098, 0.091	$<2x10^{-16}$	-0.131	-0.239, -0.023	0.02
25(OH)D GRS	Never	REF			REF		
	Former	-0.001	-0.001, 0.000	0.22	-0.011	-0.032, 0.010	0.31
	Current	-0.001	-0.001, 0.000	0.09	0.006	-0.013, 0.026	0.52

The association of 25(OH)D with number of sexual partners was weaker than the association with smoking. There was no obvious alteration in effect size or direction between the full and case only samples. There was strong evidence that the GRS was not associated with number of sexual partners in the full sample however, in the case only study there was some weak evidence that the GRS was higher in individuals who had more than two sexual partners compared to less than two. Due to the limited sample size and multiple tests used here, this result could be a false positive and should be viewed with caution (Table 8-3).

Table 8-3: Association of 25(OH)D and number of sexual partners in full and selected UK Biobank samples.

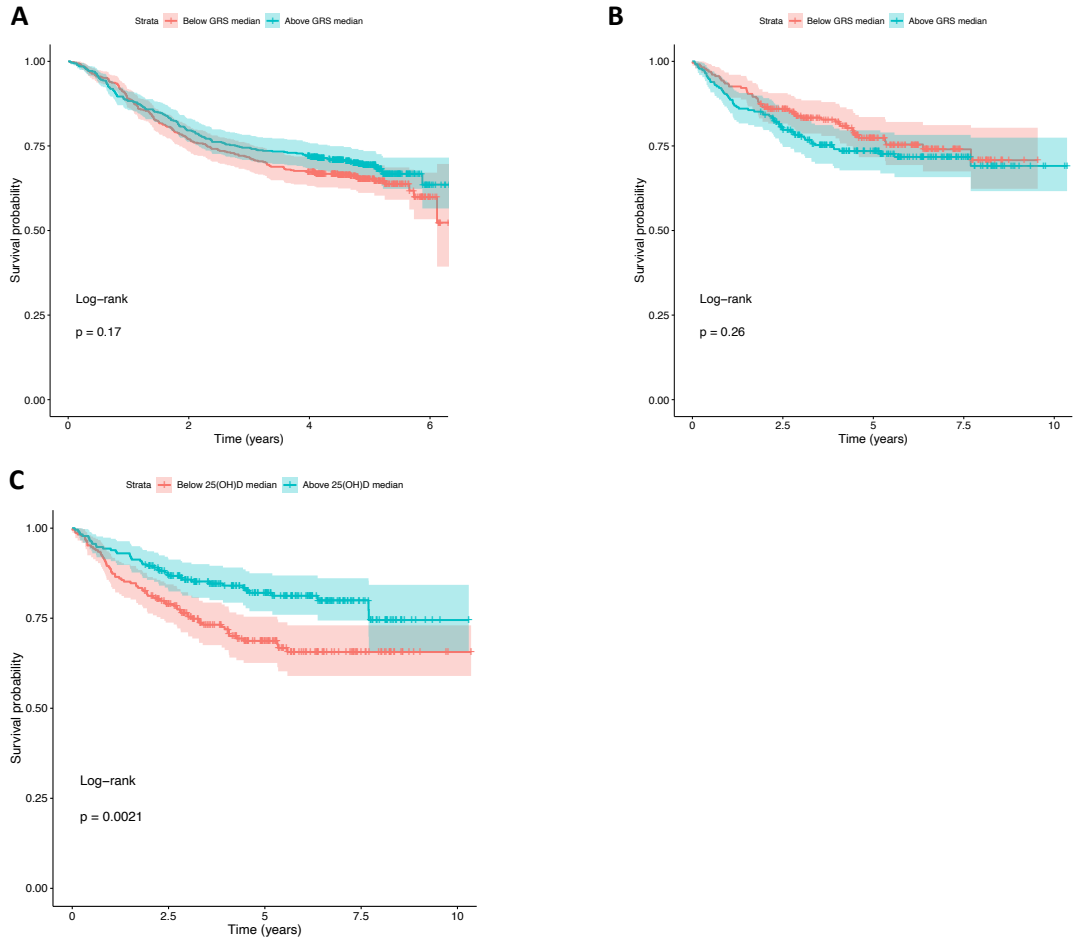
		Full UK Biobank sample			Case only sample		
	No. of sexual partners	Beta	95% CI	P	Beta	95% CI	P
Serum 25(OH)D	0-1	REF			REF		
	2-4	-0.015	-0.018, -0.112	$<2x10^{-16}$	-0.040	-0.163, 0.083	0.327
	5+	0.003	-0.000, 0.006	0.091	0.046	-0.093, 0.185	0.521
25(OH)D GRS	0-1	REF			REF		
	2-4	-0.000	-0.000, 0.001	0.496	0.024	0.002, 0.046	0.025
	5+	-0.000	-0.000, 0.001	0.439	0.014	-0.225, 0.523	0.244

8.4.5 Survival analyses

8.4.5.1 MR method 1: GRS analysis in HN5000 and UK Biobank

The GRS was dichotamised into a high-25(OH)D and a low-25(OH)D GRS in each study. In UK Biobank, where 25(OH)D measures were available, the high-25(OH)D GRS group had a mean 25(OH)D level of 49.7 nmol/L and the low-25(OH)D GRS group has a mean 25(OH)D of 42.1 nmol/L. In these analyses, there was little evidence from graphical representations that either group has a survival benefit over the follow up period in either UK Biobank or HN5000.

Figure 8-3: Kaplan Meier curves for high- and low-25(OH)D genetic risk scores in (A) HN5000 and (B) UK Biobank and (C) high- and low-25(OH)D levels in UK Biobank.



High-25(OH)D GRS = GRS above the median value in each cohort, Low-25(OH)D GRS = GRS below the median value in each cohort. Shaded area represent the 95% confidence interval of line. Vertical lines represent censoring. P = P-value for the log-rank test assessing whether there is a difference in survival between the two groups.

In survival analyses using cox-proportional hazards regression, the UK Biobank effect estimate was very imprecisely estimated. The point estimate corresponded to a large hazard increasing effect for a doubling 25(OH)D with the 95% confidence interval ranging from a large protective, to a large

hazard increasing effect (HR=2.14 (95% CI: 0.46, 9.97)). The HN5000 estimate was more precise but confidence intervals still spanned the null effect (HR=0.64 (95% CI: 0.26, 1.58)), the point estimate was similar to that of the UK Biobank observational analysis (HR=0.64 (95% CI: 0.50, 0.81). When meta-analysed together, the combined estimate was HR=0.87 (95% CI: 0.40, 1.91) (Figure 8-4).

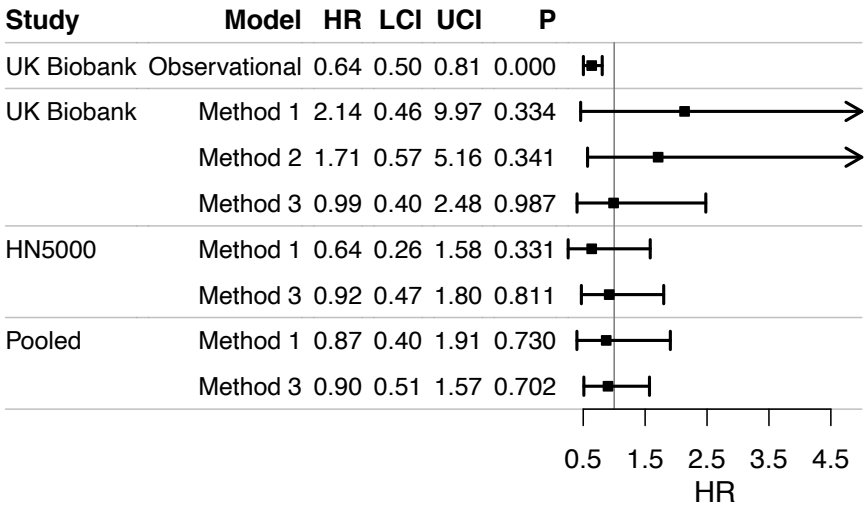
8.4.5.2 MR method 2: Two-stage ratio method in UK Biobank

Using the ratio based method the causal estimate for a log2 increase in 25(OH)D was in the same direction but of a smaller magnitude and slightly more precise than method 1 (HR= 1.71 (95% CI: 0.57, 5.16)) (Figure 8-4).

8.4.5.3 MR method 3: Two-sample MR using summary statistics

In the two-sample MR method, the effect estimates for both studies combined was HR = 0.90 (95% CI: 0.51, 1.57) which is similar to the pooled estimate from method 1. However, when the analyses were run within the individual studies the direction of effect in UK Biobank was not consistent with the previous two estimates (Figure 8-4).

Figure 8-4: Hazard ratio for death from any cause for a doubling in 25(OH)D from observational and three MR methods.

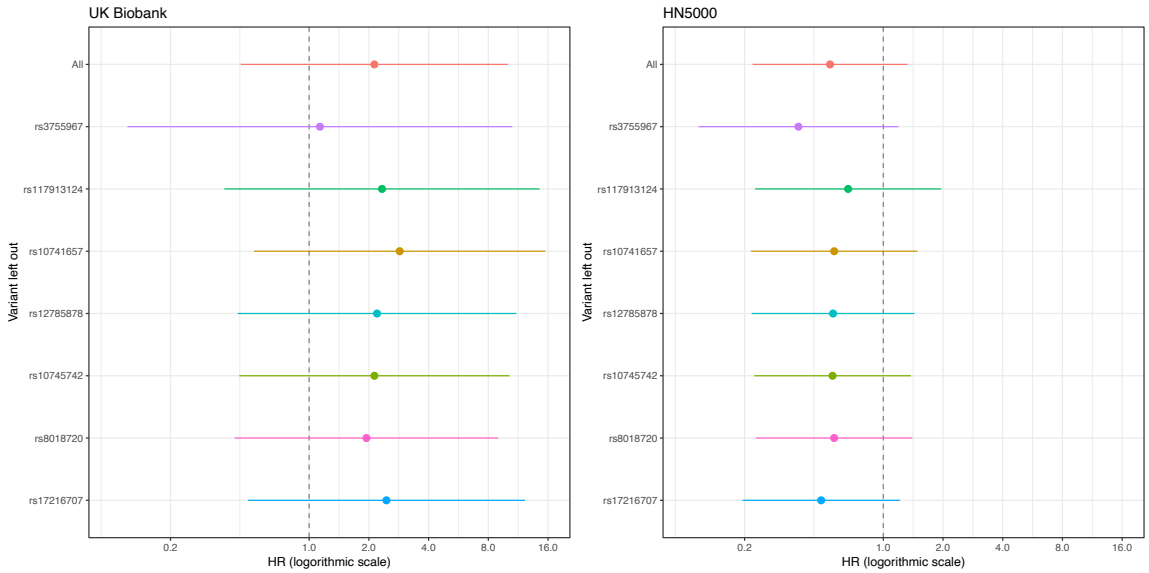


8.4.6 Sensitivity analyses

8.4.6.1 Leave-one-out

In both Methods 1 and 3, a leave-one-out sensitivity analysis was performed to see if any one variant was overtly affecting the causal estimate. In both methods, the removal of each genetic variant in turn did not substantially change the effect estimate, suggesting that no one variant was driving the effect in a particular direction. The results of this sensitivity analysis are presented for Method 1 (Figure 8-5).

Figure 8-5: Leave-one-out sensitivity analysis plots for method 1.



Each line represents a hazard ratio and 95% confidence interval. The top line represents the overall estimate with all genetic variants included. The esitmates below are the those which exclude the variant labelled on the left of the plot.

8.5 Discussion

This study demonstrates three MR methods to investigate the effect of 25(OH)D on HNC prognosis. These MR methods have several advantages over a serum measure of 25(OH)D. Firstly, these approaches are less affected by confounding factors than serum 25(OH)D, meaning that causality can be inferred. Secondly, two of the methods allow the association to be investigated in studies that do not have serum 25(OH)D measures, such as HN5000. Furthermore, if genome-wide genetic data are available in studies such as HN5000, these methods can be used to investigate other factors of interest that can be effectively proxied by genetic variants. This ultimately provides a cost-effective means of investigating a large range of factors in comparison to measuring them all individually.

The causal estimates from all three methods and across the two studies were imprecisely estimated. The point estimate for two studies showed opposite effect directions for Method 1, the GRS based method. These differing effect sizes and directions may represent the random spread of effects around the true estimate due to error and it is difficult to say with any certainty where the true causal estimate lies without improved power. The most precise estimate was from Method 3, the two-sample summary statistic method. However, when using the summary data from UK Biobank and HN5000 combined, the estimate is still consistent with a moderate survival advantage, no effect or indeed a survival disadvantage. When compared to estimates from observational studies, including the work in Chapter 4, the pooled MR effect estimates from Methods 1 and 3 suggest the

observational estimates are an overestimation but again it is difficult to say with any certainty as the MR effects sizes are imprecisely estimated. A well powered study is required to identify the exact causal effect and how this compares with the precise but likely-confounded observational estimates.

As discussed in previous chapters, the genetic variants used as a proxy measure of 25(OH)D in this study are likely to be a valid measure of circulating 25(OH)D levels. This was demonstrated by the relatively large F-statistic for the GRS in UK Biobank, despite the small sample size. The amount of variance in 25(OH)D explained in the HNC cases (5.8%) was interesting. This proportion of variance explained was almost double the value than in the full UK Biobank cohort. Although this may reflect a chance finding it could be that those with HNC are different, with regard to vitamin D, to the general population. For example, they may acquire less 25(OH)D from sunlight and their diet, meaning that the impact of the genetic variants for vitamin D in this group is higher. In Chapter 4, Table 4-2 shows that those participants who went on to get HNC had lower dietary 25(OH)D which may support this hypothesis.

The MR analyses in this study are conducted in a selected, case only scenario. The use of UK Biobank means some impacts of this selection can be interrogated. To do this, the associations between 25(OH)D and smoking and 25(OH)D and number of sexual partners were used as examples. In the case-only setting, the association between smoking and 25(OH)D was substantially exaggerated in terms of effect size which, may have implications on observational analyses: Any bias by residual confounding in the case-only analysis of 25(OH)D on HNC survival may be increased by the exaggerated association between smoking and 25(OH)D. As the GRS for 25(OH)D is uncorrelated with smoking there should be no exaggeration of confounding when a GRS is used in a case only context, however this is not always the case. Using a GRS (or other MR related method), would be problematic if the factor of interest (25(OH)D here) is causally related to HNC risk. In this scenario, when selecting on case (and therefore the 25(OH)D genetic variants) the 25(OH)D variants would become correlated with other HNC risk factors and would then introduce bias into the causal estimates – this is the issue previously referred to as ascertainment or collider bias.

In the assessment of the association of 25(OH)D and the 25(OH)D GRS with number of sexual partners, there was weak evidence for an association between the GRS and number of sexual partners in the case only scenario. Although this likely represents a chance finding, it could theoretically be the result of collider bias. If this is the case it could invalidate the independence assumption of MR. However, this study does demonstrate that the GRS for 25(OH)D is not associated with smoking in either the full UK Biobank or the case-only datasets. This has two important implications. Firstly, it suggests the estimates are unbiased with respect to confounding

by smoking and secondly it provides some evidence that 25(OH)D is not causally related to HNC risk as if it were an association between the 25(OH)D GRS and smoking would be induced in the case only setting.

Statistical methods to conduct MR in time to event analyses are much less established than those for continuous or binary variables in case-control or cohort designs. In this work three methods which have been reported in the literature were used. The first method used a GRS to estimate the causal effect of a unit increase in the GRS (scaled to a log2 increase in 25(OH)D). This method does not take into account the imprecision of the effect of the genetic variant on 25(OH)D and so may produce confidence intervals which are artificially narrowed. The second method, the ratio method, requires serum 25(OH)D in addition to the genetic variants and survival data. This method produces similar but slightly more precise estimates than Method 1. This increased precision likely comes from the GRS explaining a greater amount of variance in serum 25(OH)D in the case only scenario than in population-based studies. The final method which uses summary level data produced the most precise estimates and these were consistently closer to the null than the estimates from the other methods. This summary level method has the disadvantage that if only summary level data are available, no assessment of the independence assumption is possible. Summary level MR has attracted much attention over the past few years and many statistical approaches have been generated which can assess the exclusion restriction assumption such as MR-Egger.[150] However, these have little power to detect pleiotropy with low numbers of genetic variants and their utility is unknown in time to event MR contexts.

The outcome used in these analyses was all-cause mortality. This was due to limited data in HN5000 and limited numbers in UK Biobank. In future studies it would be interesting to see if estimates of HNC specific death differ from all-cause death.

Finally, the estimates generated in this study can be used to plan future MR studies of 25(OH)D on HNC prognosis. The best available causal estimate for the effect of 25(OH)D on HNC prognosis, provided by the meta-analysis of HN5000 and UK Biobank, is closer to the null than observational estimates. This implies that future studies should be powered to detect effects smaller than that of the observational estimate if estimate precision is to be clinically useful.

8.6 Conclusion

MR is likely to provide a useful method for assessing causality in situations where time to event data are available and in case-only scenarios however, refinement of methods is required. This study

demonstrates the utility of using population-based studies to nest case-only analyse within, allowing certain methodological issues such as collider bias, to be assessed.

The next chapter

The final chapter of this thesis will summarise the main findings of this work and discuss the important strengths and weaknesses of the body of work. Finally, plans for future work that hopes to address some of the issues highlighted will be provided.

Chapter 9 : Discussion

In this thesis, I describe my work attempting to estimate the true causal effect of vitamin D on the risk and prognosis of head and neck cancer. I did this by generating a precise estimate of the effect of vitamin D on HNC risk and prognosis from observational epidemiology. I then identified a robust genetic proxy for vitamin D which was minimally affected by confounding and reverse causality and used this genetic proxy in MR studies to estimate the causal effect of vitamin D on HNC risk and prognosis.

In this discussion I summarise the evidence gained as to whether there is a true causal link between vitamin D and HNC risk and prognosis by synthesising evidence from various chapters as well as previous studies. I use the Bradford Hill framework of causation,[140] detailed in Chapter 3, to assess this body of evidence. From this evidence I provide my conclusions on the association of vitamin D and HNC risk and prognosis and discuss these conclusions in the context of its clinical application. I discuss the strengths and limitations of the work in this thesis and how they may help support or reduce confidence in the main conclusions. Finally, I outline current work and planned research that will help to address some of the limitations raised.

9.1 Synthesis of findings

In this section I will assess the evidence presented in this thesis in the context of previously published research using the Bradford-Hill criteria which are described in detail in Chapter 3. A summary of findings from this thesis is presented in Table 9-1. As I have explained in the introductory chapters, cancer risk and prognosis should be considered separately and I will use that strategy here. It is also important to consider the heterogeneity between HNC sites as it could be that vitamin D has an effect which is dependent on the anatomical site of the cancer. In this section, I will where possible, summarise the evidence for oral and oropharyngeal cancers separately. The evidence for laryngeal cancer is limited to the observational chapter and so will not be discussed in detail. For HNC prognosis, stratification by site was not possible in the MR study therefore the association between vitamin D and all HNC combined will be discussed.

Table 9-1: Summary of results.

	Method	Oral cavity	Oropharynx	All HNC
RISK	**CHAPTER 4: Observational** (unconditional minimally adjusted)	0.61 (0.50, 0.76)	0.75 (0.61, 0.91)	0.63 (0.56, 0.71)
	CHAPTER 7: Mendelian randomization (meta-analysis: GAME-ON, UK BIOBANK)	0.75 (0.50, 1.10)	1.12 (0.52, 2.39)	0.91 (0.56, 1.49)
PROGNOSIS	**CHAPTER 4: Observational** (unadjusted)	0.74 (0.50, 1.10)	0.66 (0.50, 0.88)	0.64 (0.53, 0.77)
	CHAPTER 8: Mendelian randomization (method 3: 2-sample)	-	-	0.90 (0.51, 1.57)

9.1.1 Vitamin D and HNC risk

9.1.1.1 Strength of evidence

Strength of evidence is the first of the Bradford-Hill criteria that I will use to assess causality in the association between vitamin D and HNCs. For oral cavity cancers, the effect size estimated from the observational study in this thesis was moderate, with a doubling of vitamin D relating to a 41% reduction in odds (95% CI: 24%, 50%) from the minimally adjusted model. This estimate was slightly smaller in magnitude than previous studies and attenuated slightly after correcting for known confounders.[126] Although the MR estimate was imprecise, the point estimate was smaller again but would also have been consistent with no effect or one more extreme than the observational (OR = 0.75 (95% CI: 0.50, 1.10). The oropharyngeal effects were consistently smaller than the oral cavity cancer estimates in both observational and MR studies. The observational study in Chapter 4 estimated a modest protective effect for oropharyngeal cancer (0.75 (95% CI: 0.61, 0.91)), and this effect was closer to the null when meta-analysed with the Fanidi *et al.* study.[126] In the MR analysis the effect estimate was OR = 1.12 (95% CI: 0.52, 2.39). The South American region suffered potential population stratification issues which led to the risk increasing effect in this region being unreliable. After excluding this region the effect size was even closer to the null for oropharyngeal cancer (OR = 0.93 (95% CI: 0.56, 1.56)). To put these effect sizes in context, the effect of not smoking compared to smoking, in the same UK Biobank sample used in the observational study in Chapter 4, reduces the odds of HNC by approximately 94%. In terms of causality, it is conceivable that residual confounding factors could account for modest or small observational effects, particularly when considering the large effect size of known confounders such as smoking.

9.1.1.2 Consistency and specificity of evidence

As discussed in Chapter 3, the concept of consistency is similar to the triangulation approach to assessing causality. Despite the effect sizes being fairly modest, there is consistency in effect sizes particularly between the observational study in Chapter 4 and the Fanidi *et al* paper.[126] Furthermore, if you include older studies that report effects for the same change in vitamin D,[124] the effect size for all HNC sites was again consistent. It is unlikely that the confounding structures in these studies are identical and so consistency across these studies lends some support to a causal association. When including the MR studies, there is again some consistency across the sites – there is a consistent pattern in both observational and MR analyses that the effect in oropharyngeal cancer risk is smaller than in oral cavity cancer risk. This latter point relates also to specificity, as it appears that the effect of vitamin D on oral cavity cancer risk is greater than the effect on oropharyngeal cancer but the

difference in effect size is small. It is known that many of the oropharyngeal cancers will be HPV driven cancers and one explanation for the difference in vitamin D effect magnitude is that HPV cancers have a different association with vitamin D than non-HPV cancers. Although it is likely to be true that vitamin D has a different association with HPV driven cancer, this does not imply that associations with vitamin D in non-HPV driven cancers are causal. It is just as likely that the confounding structures of these cancers are different. For example, HPV related cancer is more common in non-smokers than non-HPV related cancers and, as we have seen in the observational results in Chapter 4 and several previous studies, [121,122,125,126] the effect of vitamin D on HNC risk in non-smokers is close to the null. It may be that there is less residual confounding by smoking in HPV-oropharyngeal cancers as less are driven by smoking. Unfortunately, there was not data on HPV related cancer in either the observational or MR studies in this thesis and so this could not be investigated further.

9.1.1.3 Temporality

Temporality is important when considering causality of vitamin D and cancer. The diagnosis of a HNC will almost certainly change patients' lifestyle which may reduce time outside, alter diet and ultimately affect vitamin D levels. In this thesis, the design of both the observational and MR studies adequately deal with the concept of temporality. In the observational study, only incident cancers, after measurement of serum 25(OH)D, were included in the analysis. Furthermore, when assessing the effect of vitamin D on HNC stratified by time since blood draw there was no strong evidence of an attenuation of effect as the time since blood draw increased, as would be expected if there was a reverse causal effect of cancer on vitamin D. In the MR analyses, genetic proxies were used as measures of vitamin D and these genetic variants are unchanged by the process of cancer and are therefore unaffected by temporality.

9.1.1.4 Dose-response relationship

In the observational analyses, a dose-response relationship was investigated by assessing the association of vitamin D with HNC risk in quintiles of 25(OH)D. There was no evidence of a dose response relationship, with effect sizes consistent across all 25(OH)D quintiles. If the association between vitamin D and HNC was causal, the most likely intervention would be to increase serum 25(OH)D with supplementation. The lack of a dose-response however, suggests that those with higher 25(OH)D are not at a lower risk and this does not support a causal relationship. MR analyses do not lend themselves to the investigation of dose-response relationships and so cannot provide further insight.

9.1.1.5 Plausibility and coherence

The plausibility and coherence of the association between vitamin D and HNC risk was not investigated in this thesis. A summary of the evidence supporting the biological explanation for the association between vitamin D and HNC is provided in the introductory chapters. It is worth noting that there are likely to be many unknown aspects of the biological relationship and as these are discovered they may support or refute the association of vitamin D with HNC risk.

9.1.1.6 Experimental evidence and analogy

Experimental evidence can provide strong evidence of causality. In the introduction I have discussed why a randomised controlled trial specific for vitamin D in HNC would be problematic. It would be extremely expensive requiring large numbers and long follow up periods. However, MR can be considered a quasi-experimental method as long as its assumptions hold. The MR study of vitamin D on HNC risk in this thesis suggests that the effect estimates of vitamin D are likely smaller than those from the observational studies but more precise MR estimates are required to provide clarity. The risk of more common cancers such as colorectal cancers have been investigated in large RCTs of vitamin D supplementation with all results finding no strong evidence of a protective effect of supplementation.[206,217] Given that the observational evidence for the association between vitamin D and colorectal cancer is considered much stronger than for HNCs,[9] considering the concept of analogy an RCT of vitamin D supplementation for HNC risk would be unlikely to support a protective association.

9.1.1.7 Triangulation of evidence

Applying the Bradford-Hill criteria to the evidence presented in this thesis, it is not completely clear whether the association between vitamin D and HNC risk is causal. A triangulation approach to assessing causality involves assessing the strength and direction of the particular biases from each set of evidence and using this to help clarify if an association could be causal. In this thesis, there are two main sources of evidence, the prospective observational evidence and the MR evidence.

As mentioned, neither method is likely to be affected by reverse causality. However, the observational study conducted in Chapter 4 and those that have gone before are likely to be affected by confounding. This is clearly demonstrated in Chapter 4, where effect estimates adjusted for confounders attenuate towards the null compared to the unadjusted effect estimates. This suggests the direction of the confounding for the variables corrected for in the regression models (ethnicity, smoking status, alcohol consumption frequency, highest educational attainment, BMI and latitude) is away from the null in a protective direction. Therefore, any residual confounding by these variables, not explained by their measurement, is likely to be acting in the same direction. The

magnitude of the residual confounding is unknown but there are several indications which may provide an approximation. Smoking is likely to be one of the main confounders in this association so, like in previous chapters, it will be used in these examples.

Firstly, the Fanidi *et al*. study,[126] recognised residual confounding by smoking as a potential issue. They used urinary cotinine, an objective measure of current smoking exposure in their regression analysis. In these analyses there was a very modest attenuation in effect size (OR = 0.54 unadjusted, OR = 0.55 cotinine adjusted) suggesting smoking at the time of 25(OH)D measurement was not a strong confounder of the vitamin D – HNC risk association. However, it is important to note that cotinine is only an objective measure of recent smoking (less than one week[218]) and so would not account for smoking history before this point. Future studies that wish to remove the effect of long-term smoking should consider alternative measures of smoking such as epigenetic profiles, which have been demonstrated to predict previous smoking over longer periods.[219]

Secondly, if smoking is to be considered the main confounder of the vitamin D – HNC risk association and there is limited mis-classification by self-reported current smoking (as indicated by the cotinine adjustment in the first point) then there should not be residual confounding by smoking status. However, it is unlikely that simply smoking or not smoking captures the true confounding effect of smoking and that the intensity and duration of smoking will play a large part. If this is true, there will be disparity across smoking strata as to the magnitude of the residual confounding effect. In true never-smokers there will be no error in the measurement of intensity or duration as both will be zero. In current smokers there will be a vast range of smoking intensities and durations and this will not be captured by the “current smoker” status and thus may lead to greater residual confounding in smokers than non-smokers. When considering an estimate of the causal effect from observational studies, the most robust estimate will be one with the least error in the measurement of the confounders and in the scenario above this would be the non-smoking strata. As highlighted in the introductory chapters[121,122,125,126] and shown in Chapter 4, there is a consistent near-null effect of vitamin D on HNC risk in non-smokers which can be considered the best estimate of the causal effect from the observational work presented here. An alternative explanation to this is that vitamin D interacts with smoking to provide a protective effect but only in smokers. There is some weak evidence to support this in the observational study in Chapter 4 and the Fanidi *et al.* study[126] but even with these relatively large samples there is limited power to distinguish an interaction from residual confounding.

Finally, one of the main advantages of MR is that the effect estimate from this approach is largely unaffected by confounding. In this thesis, the genetic proxy used in the MR analyses has been

demonstrated to be unrelated to smoking in both the hypothesis free MR PheWAS analyses in Chapter 6 and in the MR study in Chapter 7, whereas the serum 25(OH)D measure was strongly associated with smoking status in both the observational study in Chapter 4 and the phenome wide scan in Chapter 6. The MR estimate therefore provides the best estimate of the unconfounded effect however, due to the issues of power the estimate is imprecise. Because of this, it is not completely clear where the unconfounded effect estimate lies but the comparison of effect estimates from this thesis suggests some degree of residual confounding as the MR estimates are consistently smaller in magnitude than the adjusted observational estimates.

The issue of confounding is not limited to smoking. The phenome scans in Chapter 6 demonstrate that the serum 25(OH)D measure is strongly associated with a wide range of factors which may be causally related to HNC risk – either directly or by covariation with other causal factors. If this is the case, even if smoking is adequately corrected for, the observational effects will be confounded by these other, unmeasured, factors. Again, as demonstrated in the MR PheWAS, the genetic proxy of vitamin D was unrelated to these potential confounders and so the MR effect estimate is robust to them.

Considering these examples, it is likely that the magnitude of residual confounding could be as large as the full observational effect size (as suggested by the observational estimate in non-smokers) but may be smaller than this (as demonstrated by the consistently smaller but imprecisely estimated MR effect sizes).

When synthesising this evidence, I feel it is unlikely that low vitamin D can explain some of the recent patterns in HNC incidence or the unexplained population attributable risk discussed in Chapter 2. It seems doubtful that vitamin D supplementation provides genuine potential to reduce the risk of HNCs in the general public. It is still unclear whether vitamin D does have a causal association with HNC risk but if it does, the effect is likely to be small and potentially limited to specific groups, such as smokers or specific cancer types, such as oral or non-HPV oropharyngeal cancers. Before an expensive population-wide intervention could be considered, robust evidence from intervention studies would be required but the work of this thesis suggests we do not have the required threshold of evidence for a RCT to be considered. A more sensible approach would be to conduct a well powered MR study to provide more precise minimally confounded estimates, before even considering a trial. I am involved with several projects that will allow a subsequent MR study with greater power and these projects are described in the future work section below.

9.1.2 Vitamin D and HNC prognosis

9.1.2.1 Strength and consistency of evidence

The observational effect of vitamin D on HNC prognosis is approximately a 33% reduction in hazards of death. Given the high mortality of HNCs, this relative effect corresponds to a clinically important outcome that would have clear benefit for patients if real. In observational studies, when assessing the effect on HNC-specific deaths, there is some attenuation of the effect size but the majority of the effect is driven by HNC-related deaths. However, MR estimates suggest this effect size is smaller than that reported in the observational work but these MR estimates are very imprecise due to the relatively small sample sizes. There is consistency of effect size between the observational study in Chapter 4 and the Fanidi *et al.*[126] study but these are not consistent with the MR estimates which are smaller in size and, in the case of the UK Biobank MR estimate, in the opposite direction to the observational estimates.

9.1.2.2 Temporality

The temporality of the effect of vitamin D on HNC prognosis is important to consider. Several studies discussed in the introduction used 25(OH)D measures at diagnosis and there is the potential that the process of cancer will be affecting 25(OH)D levels in these studies. In this thesis, I use pre-diagnosis 25(OH)D levels to avoid this issue in the observational analyses. Furthermore, in stratifying by time from blood draw to diagnosis, I was able to assess if undiagnosed cancer was potentially having an effect on 25(OH)D levels. In these stratified analyses, the effect of vitamin D on HNC prognosis completely attenuated in those with vitamin D measures more than six years before diagnosis, which may be evidence that HNC is lowering 25(OH)D levels. However, in this greater than six-year strata, there are few participants and very few deaths limiting power to reliably detect an effect. Additionally, if there is a true causal relationship between 25(OH)D and HNC prognosis, a single measure of 25(OH)D greater than six years before diagnosis is unlikely to represent a valid measure of 25(OH)D during the HNC cancer process and this would also explain the null result in this stratum. Similar to HNC risk, the vitamin D proxy measure used in MR analyses is unaffected by the cancer process and estimates from MR should be robust to it.

9.1.2.3 Dose-response relationship

There is limited evidence from this thesis that there is a dose response relationship between vitamin D and HNC prognosis. The Fanidi *et al.*[126] study did model the effect of increasing 25(OH)D on the hazards of all-cause mortality and this demonstrated a U-shaped association with both very low and very high levels of 25(OH)D being associated with greater hazards of dying. This U-shaped

association was not investigated in this thesis but is not necessarily supported by current biological understanding.

As in the case of HNC risk, it is unclear from the work in this thesis whether a causal relationship does exist between vitamin D and HNC prognosis partly due to limited power from MR analyses. However, unlike for HNC risk, there does appear to be potential clinical utility if a causal association exists. For that reason, it is important that this topic is revisited in better powered observational studies and MR studies with approaches that are specific to the application on MR in case only data. Some of these approaches which I am involved in are discussed below.

9.2 Strengths and limitations

9.2.1 Power

In this thesis there was generally sufficient power for observational approaches, particularly in the assessment of the association with HNC risk. In these approaches, stratification by certain key factors was possible allowing heterogeneity between smoking strata to be identified. Power was moderate in observational studies of prognosis meaning there was limited inference from results stratified by site or potential confounders.

In the GWAS analyses, there was good power to detect common genetic variants with moderate to large associations with 25(OH)D and two novel associations were identified which were later replicated in a larger GWAS. One of the aims of the GWAS was to identify lower frequency variants with larger effects and one such variant was identified which had a large effect on 25(OH)D. Other variants of similar frequency may exist but are likely to have smaller effects. There was limited power to identify these variants.

Despite the discoveries in Chapter 5, there are few genetic variants which can be used as reliable genetic proxies of 25(OH)D which has two implications in MR analyses. Firstly, the genetic proxy explains a small amount of the variance in 25(OH)D limiting the power for MR. Secondly, with few variants many of the MR approaches which have been designed to be robust to pleiotropy are underpowered and would be uninformative. The power in MR studies also relies heavily on the size of the outcome sample and very large HNC datasets are currently unavailable. Work I am contributing to which will hopefully address both these power issues is described in the Future Work section below.

9.2.2 Head and neck cancer data

As mentioned in the section above there is limited HNC data available for approaches such as MR which required very large sample sizes. Collaborative efforts such as the International Head and

Neck Cancer Epidemiology Consortium (INHANCE), a collaboration of over 40 studies with approximately 40,000 cases and a similar number of controls, do provide an opportunity to conduct large observational studies but there is limited recording of phenotypes such as vitamin D across these collections. If genetic data are available on samples of this size, MR will allow a range of potential risk and prognostic factors to be investigated even if they are not measured across these studies. However, many of the studies contributing to the INHANCE consortium are relatively old and genotyping in newer collections such as HN5000 should be encouraged.

9.2.3 Vitamin D genetic proxy

In this thesis MR is used to estimate the causal effect of vitamin D on HNC risk and prognosis. For the MR estimates to be valid, the genetic proxy for vitamin D must not violate the three core MR assumptions. I will discuss each of these three assumptions below in the context of this thesis highlighting where limitations or potential violation exist and how future work may be able to overcome these.

9.2.3.1 The relevance assumption

The relevance assumption relates to the genetic proxy being robustly associated with the exposure of interest, here vitamin D or more specifically serum 25(OH)D levels. Under this assumption there are two aspects that should be considered, the strength of the association between the proxy and the exposure and the appropriateness of the proxy for that exposure.

In this thesis I have demonstrated that the seven genetic variants used as proxy measures of vitamin D are strongly and consistently associated with 25(OH)D. This is demonstrated in Chapter 5 where in the GWAS, the four previously identified 25(OH)D genetic variants showed strong associations with 25(OH)D and the three novel variants passed the genome-wide threshold in the meta-analysis and were consistent in the direction of effect across the 19 studies. The relevance assumption was also tested in Chapter 6 and Chapter 8, where the genetic variants were combined into a GRS and shown to have very strong associations with 25(OH)D in an independent sample, UK Biobank. The strength of a genetic proxy is often discussed in terms of the F-statistic which depends on both the percentage of variance in the exposure explained and the sample size where the association is being tested. The smallest sample where the GRS was assessed for its strength to proxy 25(OH)D was in the prognosis MR analyses in Chapter 8. Even in this relatively small sample the genetic proxy had an F-statistic of 28.2 and would seem strong enough to exclude weak instrument bias. Despite this strength, the genetic proxy only accounted for between 3% and 5% of the variance in 25(OH)D. Although this is large for the relatively low number of genetic variants used in the GRS, this means that power for MR analyses is low as discussed above.

The second important point to consider is the appropriateness of the genetic proxy for investigating the association between 25(OH)D and HNC risk and prognosis. The previously identified genetic variants, and the low-frequency variant identified in Chapter 5, which were used as proxies in this thesis, are in or near genes that are close to vitamin D related biology. Although it must be considered that being physically close to a gene is no guarantee that the variant affects that gene, it does support the relevance assumption as there is biological plausibility for how these variants could influence serum 25(OH)D. The two novel common variants identified in Chapter 5 have less clear links with vitamin D biology but were successfully replicated in a subsequent GWAS[193] and at least one of these may be involved in the utilisation of UV light in the skin which again demonstrates a biologically plausible mechanism for influencing serum 25(OH)D.

Despite the GRS being a strong proxy of serum 25(OH)D levels it must be considered whether this is appropriate for investigating the effect of vitamin D in HNC risk and prognosis. As highlighted in the introduction, serum 25(OH)D assays measure total 25(OH)D, a combination of free 25(OH)D and bound (to vitamin D binding protein (DBP) and albumin) 25(OH)D. Additionally, the GRS is a proxy for 25(OH)D and not the active metabolite, $1,25(OH)_2D$. It is not completely clear whether a proxy of the total inactive form is suitable when investigating the effect of vitamin D on cancer, as it has been demonstrated that the anti-cancer effects are via free-$1,25(OH)_2D$ acting at the tissues.[133] Despite these points, the most common form of vitamin D intervention is by supplementation with 25(OH)D, which will act to increase total 25(OH)D in the first instance. Therefore, in order to estimate the effects of this intervention using MR, it makes sense to have a genetic proxy which has the same influence. However, there is one genetic proxy used in this thesis which may be problematic even when trying to mirror 25(OH)D supplementation, this is discussed below.

In the kidney, 25(OH)D is transported into the cell using a megalin-mediated transport system, bound to DBP. It is likely that the effect of vitamin D on cancer is extra-renal and, outside of the kidney, 25(OH)D is likely to enter the cell in its free state and therefore there is controversy within the field as to whether the total or free level is the most important.[68] Although there is strong evidence to suggest free and total levels are highly correlated, it may be that the amount of DBP and its affinity to bind free 25(OH)D, dictates the availability of free 25(OH)D at peripheral sites such as the oral cavity and oropharynx. If this is the case an increase in DBP or its affinity for 25(OH)D will increase total 25(OH)D but may limit how much free 25(OH)D is available to provide the potential anti-cancer actions at the tissue level.

This issue is important because the genetic variant in *GC* (rs3755967) used as a proxy in this thesis has been linked with altered levels of DBP and binding affinity for 25(OH)D[178] and this may be

problematic. When *GC* variants are included in the genetic proxy which increase DBP concentration or its affinity for 25(OH)D, they will be robustly associated with increased total 25(OH)D as more 25(OH)D will be bound and therefore captured by the assay. However, this may also result in reduced free 25(OH)D at the peripheral tissues as more is in the bound form, particularly when 25(OH)D levels are low.[69] If so, the effect of the *GC* variant on free vitamin D may be in the opposite direction to that of total 25(OH)D and therefore *GC* may not be a suitable variant for the purposes of investigating the effect of 25(OH)D on cancer.

To assess the potential impact of this issue, leave-one-out approaches can be informative. Across the MR studies in this thesis, leaving the *GC* variant out does not impact the MR estimates in a consistent and systematic way. This indicates above *GC* issues are not influencing the causal estimates. Finally, megalin-mediated transport in extra-renal tissues such as the breast[220] and prostate[221] has been discovered so there is the potential that this transport process is also occurring in the oral and oropharyngeal mucosal cells which reduce the impact of this concern with the *GC* variant.

9.2.3.2 The independence assumption

The independence assumption relates to the genetic proxy being unrelated to potential confounders of the exposure – outcome association. In situations where this assumption is violated, the effect estimate can be biased in a similar way to a confounded observational effect estimate.

In this thesis, I have demonstrated that the vitamin D genetic proxy is unrelated to the measured confounders in the two MR chapters (Chapter 7 and Chapter 8) but have gone further than this by conducting the MR PheWAS in Chapter 6. In Chapter 6, I showed that compared to the serum 25(OH)D measure, the vitamin D GRS is unrelated to other factors in the UK Biobank study including phenotypes that would be the most likely confounders of the vitamin D – HNC risk and prognosis association such as smoking and alcohol use.

In Chapter 7, there was some evidence that the GRS was associated with geographic region and this may be a result of population stratification which was not accounted for by the addition of genetic principal components in the regression model. This population stratification could lead to a violation of the independence assumption and to account for this, analyses were repeated removing individuals from this geographic region in sensitivity analyses.

The independence assumption relies on genetic variants being randomly allocated at gamete formation and segregation in the population of interest. Where an ascertained population is under investigation this may not hold and the independence assumption may be violated. In Chapter 8, a

case only sample of HNC patients was selected. In this selected sample, the vitamin D genetic variants will not be randomly distributed with respect to confounders, if vitamin D (and hence those genetic variants) causally influence whether an individual develops HNC. As detailed above, it is not completely clear whether vitamin D is a cause of HNC so the implication of ascertainment or collider bias is unknown in the MR analyses assessing the association of vitamin D on HNC prognosis. In Chapter 8, I was able to compare GRS – confounder associations in both the unselected and selected samples to assess whether they differed. The strategy of comparing the effect of genetic variants in selected and un-selected populations is likely to be informative in future case-only MR analyses by allowing the identification of genetic variants that may lead to collider bias. These approaches may also be able to be extended so that they can correct for any identified bias.

9.2.3.3 The exclusion restriction assumption

The exclusion restriction assumption relates to the genetic proxy acting on the outcome only through the exposure of interest and not via alternative pathways. As mentioned this is hard to assess particularly when there are few genetic variants included as proxy measures.

As mentioned above, the majority of the genetic variants are near or within genes that are involved in vitamin D related pathways but this is no guarantee that they are acting on that gene or not acting on other (horizontally) pleiotropic pathways. For example, the *DHCR7* genetic variant used in this thesis (rs12785878) codes for the protein responsible for converting 7-dehydrocholesterol into cholesterol, therefore removing it from the vitamin D synthesis pathway. It might be expected that this variant also has roles in other cholesterol related pathways and therefore may be pleiotropic. However, in this example this is unlikely to be the case as it has been shown that the *DHCR7* variant is not related to cholesterol levels in at least two well conducted studies.[222,223] This example demonstrates how difficult it would be to rule out all potential pleiotropic pathways. In reality, we will only look for evidence of pleiotropy in biologically plausible pathways but this relies on adequate knowledge of such pathways. Furthermore, if potentially pleiotropic pathways are selected we also need the availability of correct data and adequate power to detect if a potential pleiotropic association is real.

MR sensitivity approaches which estimate causal effects which are robust to pleiotropy have been developed and provide a more realistic way of identifying the presence of pleiotropy.[150,151] Unfortunately, these methods have little power when only a small number of genetic variants are available as proxies for the exposure of interest. As discussed below, future effects to identify more genetic proxies of vitamin D will increase the number of genetic variants allowing these approaches to be used to assess the presence of pleiotropy.

9.3 Next steps

Despite this thesis adding to the body of evidence assessing the association between vitamin D and HNC, the MR aspects of the thesis would benefit from improved genetic proxies for vitamin D and larger genotyped HNC resources. During the latter part of this thesis, I have been working to establish these to further the investigation of the vitamin D – HNC association. Below I briefly describe these projects.

9.3.1 GWAS of 25(OH)D in UK Biobank

I am currently contributing to a meta-analysis of the GWAS in Chapter 5 with a new GWAS analysis of 25(OH)D in UK Biobank. This work which has a sample size of over 400,000 participants has potentially identified over 100 novel 25(OH)D related variants. This work will have two impacts on future MR studies assessing the effect of vitamin D on HNC. Firstly, these novel variants can be included in the genetic proxy of vitamin D and will increase the proportion of variance explained directly improving power for MR analyses. Secondly, a larger number of variants will allow the use of pleiotropy robust MR approaches.

9.3.2 Genotyped HNC resources

I have been involved in two projects which aim to increase the size of HNC case – control datasets with genome-wide genetic coverage. In the first of these I have obtained funding to complete genotyping of over 2,500 HNC (oral, oropharyngeal and laryngeal cancers) cases from HN5000 and to perform technical validation checks so that these cases can be linked to over 10,000 population-based controls. In the second project I have worked closely with collaborators at the International Agency for Research on Cancer (IARC) to obtain funding to genotype a further 5,351 HNC cases and 4,351 controls.

These two projects are run in parallel and when combined with the previously genotyped HNC datasets (described as GAME-ON in this thesis) there will be over 14,000 cases and 26,000 controls. This more than doubles the sample size which will be available for MR analyses of HNC risk and will also include a large number of laryngeal cancer samples, allowing MR analyses of this cancer site for the first time.

MR analyses of HNC prognosis will also benefit from the above projects. All of the HNC cases that are newly genotyped and some of the samples in the GAME-ON genotyping round have high quality follow-up data. In total there will be at least 8,000 HNC cases with genetic and follow up data.

The genotyped HNC resources discussed here not only provide the opportunity to conduct MR analyses assessing the effect of vitamin D on HNC risk and prognosis but also any other exposure which has valid genetic proxies.

9.4 Concluding remarks

In summary, this thesis cannot rule out small effects of 25(OH)D on HNC risk but suggests that effects as large as those reported in observational studies are unlikely. Given this, it would not be a sensible strategy to pursue a RCT of 25(OH)D supplementation unless more precise MR estimates suggest clinically relevant causal effects. The effect of 25(OH)D prognosis may be more clinically important, as even small effects will have clinical benefit but, larger samples and more established MR approaches are required before causality can be established using the methods in this thesis.

The genetic proxy for 25(OH)D has been improved by the work of this thesis but questions remain about how appropriate a genetic proxy for 25(OH)D may be in studies of cancer. Future work will help to improve power in both studies of HNC risk and prognosis by further improving the genetic proxy for vitamin D and increasing the size of genotyped HNC resources.

References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).

2. Klussmann, J. P. Head and Neck Cancer - New Insights into a Heterogeneous Disease. *Oncol. Res. Treat.* **40**, 318–319 (2017).

3. Gillison, M. L. *et al.* Distinct Risk Factor Profiles for Human Papillomavirus Type 16–Positive and Human Papillomavirus Type 16–Negative Head and Neck Cancers. *JNCI J. Natl. Cancer Inst.* **100**, 407–420 (2008).

4. Maier, H., Dietz, A., Gewelke, U., Heller, W. D. & Weidauer, H. Tobacco and alcohol and the risk of head and neck cancer. *Clin. Investig.* **70**, 320–327 (1992).

5. Hashibe, M. *et al.* Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Cancer Epidemiol. Biomarkers Prev.* **18**, 541–50 (2009).

6. Louie, K. S., Mehanna, H. & Sasieni, P. Trends in head and neck cancers in England from 1995 to 2011 and projections up to 2025. *Oral Oncol.* **51**, 341–348 (2015).

7. Holick, M. F. Vitamin D and bone health. *J. Nutr.* **126**, 1159S–64S (1996).

8. Nagpal, J. K. & Das, B. R. Oral cancer: Reviewing the present understanding of its molecular mechanism and exploring the future directions for its effective management. *Oral Oncol.* **39**, 213–221 (2003).

9. International Agency for Research on Cancer. *Vitamin D and Cancer. IARC Working Group Reports Vol.5, Lyon* (2008).

10. Gandini, S. *et al.* Meta-analysis of observational studies of serum 25-hydroxyvitamin D levels and colorectal, breast and prostate cancer and colorectal adenoma. *Int. J. Cancer* **128**, 1414–1424 (2011).

11. Galbiatti, A. L. S. *et al.* Head and neck cancer: causes, prevention and treatment. *Braz. J. Otorhinolaryngol.* **79**, 239–247 (2013).

12. Montero, P. H. & Patel, S. G. Cancer of the oral cavity. *Surg. Oncol. Clin. N. Am.* **24**, 491–508 (2015).

13. Muller, P., Belot, A., Morris, M. & Rachet, B. *Net survival and the probability of cancer death from rare cancers, London School of Hygiene and Tropical Medicine, London*. (2016).

14. Carpén, T. *et al.* Presenting symptoms and clinical findings in HPV-positive and HPV-negative oropharyngeal cancer patients. *Acta Otolaryngol.* **138**, 513–518 (2018).

15. Mehanna, H. *et al.* Oropharyngeal cancer: United Kingdom National Multidisciplinary Guidelines. *J. Laryngol. Otol.* **130**, S90–S96 (2016).

16. Gillison, M. L., Chaturvedi, A. K., Anderson, W. F. & Fakhry, C. Epidemiology of Human

Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma. *J. Clin. Oncol.* **33**, 3235–42 (2015).

17. Schache, A. G. *et al.* HPV-related oropharynx cancer in the United Kingdom: An evolution in the understanding of disease etiology. *Cancer Res.* **76**, 6598–6606 (2016).

18. Nutting, C. M., Robinson, M. & Birchall, M. Survival from laryngeal cancer in England and Wales up to 2001. *Br. J. Cancer* **99**, S38–S39 (2008).

19. Cattaruzza, M. S., Maisonneuve, P. & Boyle, P. Epidemiology of Laryngeal Cancer. *Eur J Cancer* **3**, 293–305 (1996).

20. Mehanna, H. *et al.* PET-CT Surveillance versus Neck Dissection in Advanced Head and Neck Cancer. *N. Engl. J. Med.* **374**, 1444–1454 (2016).

21. Cadoni, G. *et al.* A review of genetic epidemiology of head and neck cancer related to polymorphisms in metabolic genes, cell cycle control and alcohol metabolism. *Acta Otorhinolaryngol. Ital.* **32**, 1–11 (2012).

22. Lesseur, C. *et al.* Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat. Genet.* **48**, 1544–1550 (2016).

23. McKay, J. D. *et al.* A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet.* **7**, e1001333 (2011).

24. Boccia, S. *et al.* Aldehyde Dehydrogenase 2 and Head and Neck Cancer: A Meta-analysis Implementing a Mendelian Randomization Approach. *Cancer Epidemiol Biomarkers Prev* **18**, 248–54 (2009).

25. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).

26. Sarode, G., Sarode, S. C., Tupkari, J. & Patil, S. Is oral squamous cell carcinoma unique in terms of intra- and inter-tumoral heterogeneity? *Transl. Res. Oral Oncol.* **2**, 2057178X1770357 (2017).

27. Anantharaman, D. *et al.* Geographic heterogeneity in the prevalence of human papillomavirus in head and neck cancer. *Int. J. Cancer* **140**, 1968–1975 (2017).

28. Freedman, N. D., Schatzkin, A., Leitzmann, M. F., Hollenbeck, A. R. & Abnet, C. C. Alcohol and head and neck cancer risk in a prospective study. *Br. J. Cancer* **96**, 1469–74 (2007).

29. Chaturvedi, A. K. *et al.* Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. *J. Clin. Oncol.* **29**, 4294–4301 (2011).

30. International Agency for Research on Cancer. *IARC monographs on the evaluation of carcinogenic risks to humans, volume 100 E: Tobacco Smoking, Lyon*. (2012).

31. Di Credico, G. *et al.* Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: A bivariate spline model approach. *Oral Oncol.* **94**, 47–57 (2019).

32. Wyss, A. *et al.* Cigarette, Cigar, and Pipe Smoking and the Risk of Head and Neck Cancers:

Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium. *Am. J. Epidemiol.* **178**, 679–690 (2013).

33. International Agency for Research on Cancer. *IARC monographs on the evaluation of carcinogenic risks to humans, volume 44: Alcohol Drinking, Lyon*. (1988).

34. International Agency for Research on Cancer. *IARC monographs on the evaluation of carcinogenic risks to humans, volume 96: Alcohol consumption and ethyl carbamatem, Lyon*. (2010).

35. Lubin, J. H. *et al.* Systematic Reviews and Meta-and Pooled Analyses Body Mass Index, Cigarette Smoking, and Alcohol Consumption and Cancers of the Oral Cavity, Pharynx, and Larynx: Modeling Odds Ratios in Pooled Case-Control Data. *Am J Epidemiol* **171**, 1250–1261 (2010).

36. Purdue, M. P. *et al.* Type of Alcoholic Beverage and Risk of Head and Neck Cancer—A Pooled Analysis Within the INHANCE Consortium. *Am. J. Epidemiol.* **169**, 132–142 (2009).

37. Gillison, M. L. & Lowy, D. R. A causal role for human papillomavirus in head and neck cancer. *Lancet* **363**, 1488–1489 (2004).

38. Kreimer, A. R., Clifford, G. M., Boyle, P. & Franceschi, S. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: A systemic review. *Cancer Epidemiol. Biomarkers Prev.* **14**, 467–475 (2005).

39. Hobbs, C. G. L. *et al.* Human papillomavirus and head and neck cancer: a systematic review and meta-analysis. *Clin. Otolaryngol.* **31**, 259–266 (2006).

40. Conway, D. I. *et al.* Components of socioeconomic risk associated with head and neck cancer: A population-based case–control study in Scotland. *Br. J. Oral Maxillofac. Surg.* **48**, 11–17 (2010).

41. Conway, D. I. *et al.* Estimating and explaining the effect of education and income on head and neck cancer risk: INHANCE consortium pooled analysis of 31 case-control studies from 27 countries. *Int. J. Cancer* **136**, 1125–1139 (2015).

42. Stanford-Moore, G. *et al.* Interaction between known risk factors for head and neck cancer and socioeconomic status: the Carolina Head and Neck Cancer Study. *Cancer Causes Control* **29**, 863–873 (2018).

43. Winn, D., Lee, Y.-C., Hashibe, M. & Boffetta, P. The INHANCE consortium: toward a better understanding of the causes and mechanisms of head and neck cancer. *Oral Dis.* **21**, 685–693 (2015).

44. World Health Organization. *WHO global report on trends in tobacco smoking 2000-2025 - First edition*. (2018).

45. World Health Organization. *WHO global status report on alcohol and health*. (2011).

46. Office for national statistics (ONS). *Adult smoking habits in the UK: 2017*. (2018).

47. Thomas, S. J., Penfold, C. M., Waylen, A. & Ness, A. R. The changing aetiology of head and neck squamous cell cancer: A tale of three cancers? *Clinical Otolaryngology* **43**, 999–1003 (2018).

48. D'Souza, G. & Dempsey, A. The role of HPV in head and neck cancer and review of the HPV vaccine. *Prev. Med. (Baltim).* **53**, S5–S11 (2011).

49. Hammarstedt, L. *et al.* Human papillomavirus as a risk factor for the increase in incidence of tonsillar cancer. *Int. J. Cancer* **119**, 2620–2623 (2006).

50. Auluck, A. *et al.* Trends in oropharyngeal and oral cavity cancer incidence of Human Papillomavirus (HPV)-related and HPV-unrelated sites in a multicultural population: The British Columbia experience. *Cancer* **116**, 2635–2644 (2010).

51. Underwood, J. L. & DeLuca, H. F. Vitamin D is not directly necessary for bone growth and mineralization. *Am. J. Physiol.* **246**, E493-8 (1984).

52. Wang, T. J. Vitamin D and Cardiovascular Disease. *Annu. Rev. Med.* **67**, 261–272 (2016).

53. Agmon-Levin, N., Theodor, E., Segal, R. M. & Shoenfeld, Y. Vitamin D in Systemic and Organ-Specific Autoimmune Diseases. *Clin. Rev. Allergy Immunol.* **45**, 256–266 (2013).

54. Swami, S. *et al.* The role of vitamin D in reducing cancer risk and progression. *Nat. Rev. Cancer* **14**, 342–357 (2014).

55. Mellanby, E. An Experimental Investigation on rickets. *Lancet* **193**, 407–412 (1919).

56. McCollum, E. V, Simmonds, N., Becker, J. E. & Shipley, P. G. An experimental demonstration of the existence of a vitamin which promotes calcium deposition. *J. Biol. Chem.* **53**, 293–312 (1922).

57. Askew, F. A., Bourdillon, R. B., Bruce, H. M., Jenkins, R. G. C. & Webster, T. A. The Distillation of Vitamin D. *Proc. R. Soc. London. Ser. B, Contain. Pap. a Biol. Character* **107**, 76–90 (1930).

58. Deluca, H. F. History of the discovery of vitamin D and its active metabolites. *Bonekey Rep.* **3**, 479 (2014).

59. Esvelt, R. P., Schnoes, H. K. & DeLuca, H. F. Vitamin D3 from rat skins irradiated in vitro with ultraviolet light. *Arch. Biochem. Biophys.* **188**, 282–286 (1978).

60. Webb, A. R. Who, what, where and when-influences on cutaneous vitamin D synthesis. *Prog. Biophys. Mol. Biol.* **92**, 17–25 (2006).

61. Rosenstreich, S. J., Rich, C. & Volwiler, W. Deposition in and release of vitamin D3 from body fat: evidence for a storage site in the rat. *J. Clin. Invest.* **50**, 679–87 (1971).

62. Zerwekh, J. E. Blood biomarkers of vitamin D status. *Am J Clin Nutr* **87**, 1087–1091 (2008).

63. Howie, A. J. *et al.* Extrarenal Expression of 25-Hydroxyvitamin D 3 -1α-Hydroxylase 1. *J. Clin. Endocrinol. Metab.* **86**, 888–894 (2014).

64. Bikle, D. D. Vitamin D metabolism, mechanism of action, and clinical applications. *Chem. Biol.*

21, 319–29 (2014).

65. IOM (Institute of Medicine). *Dietary Reference Intakes for Calcium and Vitamin D*. (The National Academies Press, 2011).

66. Holick, M. F. Vitamin D status: measurement, interpretation, and clinical application. *Ann. Epidemiol.* **19**, 73–8 (2009).

67. Denburg, M. R. *et al.* Comparison of Two ELISA Methods and Mass Spectrometry for Measurement of Vitamin D-Binding Protein: Implications for the Assessment of Bioavailable Vitamin D Concentrations Across Genotypes. *J. Bone Miner. Res.* **31**, 1128–1136 (2016).

68. Tsuprykov, O. *et al.* Why should we measure free 25(OH) vitamin D? *J. Steroid Biochem. Mol. Biol.* **180**, 87–104 (2018).

69. Oleröd, G., Hultén, L. M., Hammarsten, O. & Klingberg, E. The variation in free 25-hydroxy vitamin D and vitamin D-binding protein with season and vitamin D status. *Endocr. Connect.* **6**, 111–120 (2017).

70. Nowson, C. A. *et al.* Vitamin D and health in adults in Australia and New Zealand: a position statement. *Med. J. Aust.* **196**, 686–687 (2012).

71. Crowe, F. L. *et al.* Trends in the incidence of testing for vitamin D deficiency in primary care in the UK: a retrospective analysis of The Health Improvement Network (THIN), 2005-2015. *BMJ Open* **9**, e028355 (2019).

72. Livshits, G., Karasik, D. & Seibel, M. J. Statistical genetic analysis of plasma levels of vitamin D: Familial study. *Ann. Hum. Genet.* **63**, 429–439 (1999).

73. Hunter, D. *et al.* Genetic Contribution to Bone Metabolism, Calcium Excretion, and Vitamin D and Parathyroid Hormone Regulation. *J. Bone Miner. Res.* **16**, 371–378 (2001).

74. Wjst, M. *et al.* Asthma families show transmission disequilibrium of gene variants in the vitamin D metabolism and signalling pathway. *Respir. Res.* **7**, 60 (2006).

75. Orton, S.-M. *et al.* Evidence for genetic regulation of vitamin D status in twins with multiple sclerosis. *Am. J. Clin. Nutr.* **88**, 441–447 (2008).

76. Snellman, G. *et al.* Seasonal Genetic Influence on Serum 25-Hydroxyvitamin D Levels: A Twin Study. *PLoS One* **4**, e7747 (2009).

77. Shea, M. K. *et al.* Genetic and non-genetic correlates of vitamins K and D. *Eur. J. Clin. Nutr.* **63**, 458–464 (2009).

78. Karohl, C. *et al.* Heritability and seasonal variability of vitamin D concentrations in male twins. *Am. J. Clin. Nutr.* **92**, 1393–1398 (2010).

79. Arguelles, L. M. *et al.* Heritability and Environmental Factors Affecting Vitamin D Status in Rural Chinese Adolescent Twins. *J. Clin. Endocrinol. Metab.* **94**, 3273–3281 (2009).

80. Engelman, C. D. *et al.* Genetic and Environmental Determinants of 25-Hydroxyvitamin D and 1,25-Dihydroxyvitamin D Levels in Hispanic and African Americans. *J. Clin. Endocrinol. Metab.*

93, 3381–3388 (2008).

81. Ramos-Lopez, E. *et al.* Gestational diabetes mellitus and vitamin D deficiency: genetic contribution of CYP27B1 and CYP2R1 polymorphisms. *Diabetes, Obes. Metab.* **10**, 683–685 (2008).

82. Hyppönen, E., Berry, D. J., Wjst, M. & Power, C. Serum 25-hydroxyvitamin D and IgE - a significant but nonlinear relationship. *Allergy* **64**, 613–620 (2009).

83. Signorello, L. B. *et al.* Common Variation in Vitamin D Pathway Genes Predicts Circulating 25-Hydroxyvitamin D Levels among African Americans. *PLoS One* **6**, e28623 (2011).

84. Cheng, J. B., Levine, M. A., Bell, N. H., Mangelsdorf, D. J. & Russell, D. W. Genetic evidence that the human CYP2R1 enzyme is a key vitamin D 25-hydroxylase. *Proc. Natl. Acad. Sci.* **101**, 7711–7715 (2004).

85. Ramos-Lopez, E., Brück, P., Jansen, T., Herwig, J. & Badenhoop, K. CYP2R1 (vitamin D 25-hydroxylase) gene is associated with susceptibility to type 1 diabetes and vitamin D levels in Germans. *Diabetes. Metab. Res. Rev.* **23**, 631–636 (2007).

86. Bu, F.-X. *et al.* Comprehensive association analysis of nine candidate genes with serum 25-hydroxy vitamin D levels among healthy Caucasian subjects. *Hum. Genet.* **128**, 549–556 (2010).

87. Cooper, J. D. *et al.* Inherited Variation in Vitamin D Genes Is Associated With Predisposition to Autoimmune Disease Type 1 Diabetes. *Diabetes* **60**, 1624–1631 (2011).

88. Hibler, E. A., Hu, C., Jurutka, P. W., Martinez, M. E. & Jacobs, E. T. Polymorphic Variation in the GC and CASR Genes and Associations with Vitamin D Metabolite Concentration and Metachronous Colorectal Neoplasia. *Cancer Epidemiol. Biomarkers Prev.* **21**, 368–375 (2012).

89. Janssens, W. *et al.* Vitamin D deficiency is highly prevalent in COPD and correlates with variants in the vitamin D-binding gene. *Thorax* **65**, 215–220 (2010).

90. Lu, L. *et al.* Associations between common variants in GC and DHCR7/NADSYN1 and vitamin D concentration in Chinese Hans. *Hum. Genet.* **131**, 505–512 (2012).

91. Ahn, J. *et al.* Vitamin D-related genes, serum vitamin D concentrations and prostate cancer risk. *Carcinogenesis* **30**, 769–776 (2009).

92. Kurylowicz, A., Ramos-Lopez, E., Bednarczuk, T. & Badenhoop, K. Vitamin D-Binding Protein (DBP) Gene Polymorphism is Associated with Graves' Disease and the Vitamin D Status in a Polish Population Study. *Exp. Clin. Endocrinol. Diabetes* **114**, 329–335 (2006).

93. Fu, L. *et al.* Functional T436K SNP in DBP is a predictor of serum 25(OH)D: A replicate study in a young healthy adult population. *Clin. Biochem.* **41**, 1267 (2008).

94. Sinotte, M., Diorio, C., Bérubé, S., Pollak, M. & Brisson, J. Genetic polymorphisms of the vitamin D binding protein and plasma concentrations of 25-hydroxyvitamin D in premenopausal women. *Am. J. Clin. Nutr.* **89**, 634–640 (2009).

95. Smolders, J., Damoiseaux, J., Menheere, P., Tervaert, J. W. C. & Hupperts, R. Fok-I vitamin D receptor gene polymorphism (rs10735810) and vitamin D metabolism in multiple sclerosis. *J. Neuroimmunol.* **207**, 117–121 (2009).

96. d'Alésio, A. *et al.* Two single-nucleotide polymorphisms in the human vitamin D receptor promoter change protein–DNA complex formation and are associated with height and vitamin D status in adolescent girls. *Hum. Mol. Genet.* **14**, 3539–3548 (2005).

97. Benjamin, E. J. *et al.* Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med. Genet.* **8**, S11 (2007).

98. Ahn, J. *et al.* Genome-wide association study of circulating vitamin D levels. *Hum. Mol. Genet.* **19**, 2739–2745 (2010).

99. Gallagher, J. C. Vitamin D and Aging. *Endocrinol. Metab. Clin. North Am.* **42**, 319–332 (2013).

100. MacLaughlin, J. & Holick, M. F. Aging decreases the capacity of human skin to produce vitamin D3. *J. Clin. Invest.* **76**, 1536–1538 (1985).

101. Liel, Y., Ulmer, E., Shary, J., Hollis, B. W. & Bell, N. H. Low circulating vitamin D in obesity. *Calcif. Tissue Int.* **43**, 199–201 (1988).

102. Arunabh, S., Pollack, S., Yeh, J. & Aloia, J. F. Body fat content and 25-hydroxyvitamin D levels in healthy women. *J. Clin. Endocrinol. Metab.* **88**, 157–161 (2003).

103. Parikh, S. J. *et al.* The relationship between obesity and serum 1,25-dihydroxy vitamin D concentrations in healthy adults. *J. Clin. Endocrinol. Metab.* **89**, 1196–1199 (2004).

104. Lagunova, Z., Porojnicu, A. C., Lindberg, F., Hexeberg, S. & Moan, J. The dependency of vitamin D status on body mass index, gender, age and season. *Anticancer Res.* **29**, 3713–20 (2009).

105. Vimaleswaran, K. S. *et al.* Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. *PLoS Med.* **10**, (2013).

106. Brot, C., Jøorgensen, N. R., Sørensen, O. H., Jorgensen, N. R. & Sorensen, O. H. The influence of smoking on vitamin D status and calcium metabolism. *Eur. J. Clin. Nutr.* **53**, 920–926 (1999).

107. Cutillas-Marco, E., Fuertes-Prosper, A., Grant, W. B. & Morales-Suárez-Varela, M. Vitamin D deficiency in South Europe: effect of smoking and aging. *Photodermatol. Photoimmunol. Photomed.* **28**, 159–161 (2012).

108. Jiang, C. Q. *et al.* Smoking and serum vitamin D in older Chinese people: cross-sectional analysis based on the Guangzhou Biobank Cohort Study. *BMJ Open* **6**, e010946 (2016).

109. Kassi, E. N. *et al.* Smoking is a significant determinant of low serum vitamin D in young and middle-aged healthy males. *Hormones* **14**, 241–250 (2015).

110. Lange, N. E., Sparrow, D., Vokonas, P. & Litonjua, A. A. Vitamin D Deficiency, Smoking, and Lung Function in the Normative Aging Study. *Am. J. Respir. Crit. Care Med.* **186**, 616–621

(2012).

111. Shinkov, A. *et al.* Winter 25-hydroxyvitamin D levels in young urban adults are affected by smoking, body mass index and educational level. *Eur. J. Clin. Nutr.* **69**, 355–360 (2015).

112. Grimnes, G. *et al.* Effect of smoking on the serum levels of 25-hydroxyvitamin D depends on the assay employed. *Eur. J. Endocrinol.* **163**, 339–348 (2010).

113. Tardelli, V. S., Lago, M. P. P. do, Silveira, D. X. da & Fidalgo, T. M. Vitamin D and alcohol: A review of the current literature. *Psychiatry Res.* **248**, 83–86 (2017).

114. Pludowski, P. *et al.* Vitamin D supplementation guidelines. *J. Steroid Biochem. Mol. Biol.* **175**, 125–135 (2018).

115. www.marketsandmarkets.com. Vitamin D Market by Analog, Application, Region - 2020. (2018). Available at: https://www.marketsandmarkets.com/Market-Reports/vitamin-d-market-22034298.html. (Accessed: 22nd February 2019)

116. Schleicher, R. L. *et al.* The vitamin D status of the US population from 1988 to 2010 using standardized serum concentrations of 25-hydroxyvitamin D shows recent modest increases. *Am. J. Clin. Nutr.* **104**, 454–61 (2016).

117. Basatemur, E., Horsfall, L., Marston, L., Rait, G. & Sutcliffe, A. Trends in the Diagnosis of Vitamin D Deficiency. *Pediatrics* **139**, e20162748 (2017).

118. Robinson, P. D. *et al.* The re-emerging burden of rickets: a decade of experience from Sydney. *Arch. Dis. Child.* **91**, 564–8 (2006).

119. Apperly, F. L. The Relation of Solar Radiation to Cancer Mortality in North America. *Cancer Res.* **1**, 191–195 (1941).

120. Garland, C. F. & Garland, F. C. Do sunlight and vitamin D reduce the likelihood of colon cancer? *Int. J. Epidemiol.* **9**, 227–31 (1980).

121. Giovannucci, E. *et al.* Prospective study of predictors of vitamin D status and cancer incidence and mortality in men. *J. Natl. Cancer Inst.* **98**, 451–459 (2006).

122. Lipworth, L. *et al.* Dietary vitamin D and cancers of the oral cavity and esophagus. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **20**, 1576–1581 (2009).

123. Arem, H. *et al.* Serum 25-hydroxyvitamin D and risk of oropharynx and larynx cancers in finnish men. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1178–1184 (2011).

124. Afzal, S., Bojesen, S. E. & Nordestgaard, B. G. Low plasma 25-hydroxyvitamin D and risk of tobacco-related cancer. *Clin. Chem.* **59**, 771–780 (2013).

125. Skaaby, T. *et al.* Prospective population-based study of the association between serum 25-hydroxyvitamin-D levels and the incidence of specific types of cancer. *Cancer Epidemiol. Biomarkers Prev.* **23**, 1220–1229 (2014).

126. Fanidi, A. *et al.* Circulating vitamin D in relation to cancer incidence and survival of the head and neck and oesophagus in the EPIC cohort. *Sci. Rep.* **6**, 36017 (2016).

127. Sood, M. M. & Sood, A. R. Dietary vitamin D and decreases in cancer rates: Canada as the national experiment. *Am. J. Clin. Nutr.* **86**, 1549–1549 (2007).

128. Ojha, R. P., Felini, M. J. & Fischbach, L. A. Vitamin D for cancer prevention: valid assertion or premature anointment? *Am. J. Clin. Nutr.* **86**, 1804–1805 (2007).

129. Schabas, R. Artifact in the control group undermines the conclusions of a vitamin D and cancer study. *Am. J. Clin. Nutr.* **87**, 792–792 (2008).

130. Meyer, F. F. *et al.* Dietary vitamin D intake and serum 25-hydroxyvitamin D level in relation to disease outcomes in head and neck cancer patients. *Int. J. Cancer* **128**, 1741–1746 (2011).

131. Gugatschka, M. *et al.* Vitamin D status is associated with disease-free survival and overall survival time in patients with squamous cell carcinoma of the upper aerodigestive tract. *Eur. Arch. Oto-Rhino-Laryngology* **268**, 1201–1204 (2011).

132. Bochen, F. *et al.* Vitamin D deficiency in head and neck cancer patients – prevalence, prognostic value and impact on immune function. *Oncoimmunology* **7**, e1476817 (2018).

133. Deeb, K. K., Trump, D. L. & Johnson, C. S. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat. Rev. Cancer* **7**, 684–700 (2007).

134. Rubin, D. & Levij, I. S. Suppression by vitamins D2 and D3 of hamster cheek pouch carcinoma induced with 9,10 dimethyl 1,2 benzanthracene. *Path.Microbiol.* **39**, 446–460 (1973).

135. Choi, H.-R. *et al.* Differential expressions of cyclin-dependent kinase inhibitors (p27 and p21) and their relation to p53 and Ki-67 in oral squamous tumorigenesis. *Int. J. Oncol.* **22**, 409–414 (2003).

136. Niméus, E. *et al.* Amplification of the cyclin D1 gene is associated with tumour subsite, DNA non-diploidy and high S-phase fraction in squamous cell carcinoma of the head and neck. *Oral Oncol.* **40**, 624–629 (2004).

137. Prime, S. S., Davies, M., Pring, M. & Paterson, I. C. The role of TGF-β in epithelial malignancy and its relevance to the pathogenesis of oral cancer (part II). *Critical Reviews in Oral Biology and Medicine* **15**, 337–347 (2004).

138. Beebee, H., Hitchcock, C. & Menzies, P. *The Oxford Handbook of Causation*. *The Oxford Handbook of Causation* (2010). doi:10.1093/oxfordhb/9780199279739.001.0001

139. Vandenbroucke, J. P., Broadbent, A. & Pearce, N. Causality and causal inference in epidemiology: The need for a pluralistic approach. *Int. J. Epidemiol.* **45**, 1776–1786 (2016).

140. Bradford Hill, A. The environment and disease: Association or causation. *Proc. R. Soc. Med.* **58**, 295–300 (1965).

141. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, dyw314 (2017).

142. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–7 (1996).

143. Hunter, D. J. Commentary: Lessons From Genome-wide Association Studies for Epidemiology. *Source Epidemiol.* **23**, 363–367 (2012).

144. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, (2019).

145. 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. *Stat. Med.* **27**, 1133–63 (2008).

146. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

147. Burgess, S. & Thompson, S. G. *Mendelian randomization: Methods for using genetic variants in causal estimation*. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation* (CRC press, 2015). doi:10.1201/b18084

148. Debat, V. & David, P. Mapping phenotypes: canalization, plasticity and developmental stability. *Trends Ecol. Evol.* **16**, 555–561 (2001).

149. van den Broek, N. *et al.* Causal associations between body mass index and mental health: a Mendelian randomisation study. *J. Epidemiol. Community Health* **72**, 708–710 (2018).

150. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

151. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

152. Kirkwood, B. & Sterne, J. *Essential Medical Statistics*. *Blackwell Publishing company* (Blackwell Publishing Ltd, 2003).

153. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank Data: Come and Get It. *Sci. Transl. Med.* **6**, 224ed4-224ed4 (2014).

154. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).

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

156. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).

157. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).

158. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3 Genes, Genomes, Genet.* **1**, 457–470 (2011).

159. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

160. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017). doi:10.1101/166298

161. Mitchell, R., Hemani, G., Dudding, T. & Paternoster, L. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 1. (2017). doi:10.5523/bris.3074krb6t2frj29yh2b03x3wxj

162. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).

163. Northstone, K. *et al.* The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. *Wellcome Open Res.* **4**, 51 (2019).

164. Boyd, A. *et al.* Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children. *Int. J. Epidemiol.* **42**, 111–127 (2013).

165. Gilbert, R. *et al.* Associations of vitamin D pathway genes with circulating 25-hydroxyvitamin-D, 1,25-dihydroxyvitamin-D, and prostate cancer: a nested case???control study. *Cancer Causes Control* **26**, 205–218 (2015).

166. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

167. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

168. Ness, A. R. *et al.* Establishing a large prospective clinical cohort in people with head and neck cancer as a biomedical resource: head and neck 5000. *BMC Cancer* **14**, 973 (2014).

169. Ness, A. R. *et al.* Recruitment, response rates and characteristics of 5511 people enrolled in a prospective clinical cohort study: head and neck 5000. *Clin. Otolaryngol.* **41**, 804–809 (2016).

170. Sachs, M. cosinor: Tools for estimating and predicting the cosinor model. R package version 1.1. (2014).

171. Robitzsch, A., Grund, S. & Henke, T. miceadds: Some additional multiple imputation functions, especially for mice. R package version 3.1-37. (2019).

172. McGivern, L., Shulman, L., Carney, J. K., Shapiro, S. & Bundock, E. Death Certification Errors and the Effect on Mortality Statistics. *Public Health Rep.* **132**, 669–675 (2017).

173. Lee, P. N. Comparison of autopsy, clinical and death certificate diagnosis with particular reference to lung cancer. A review of the published data. *APMIS. Suppl.* **45**, 1–42 (1994).

174. Nicastro, H. L. & Dunn, B. K. Selenium and prostate cancer prevention: insights from the selenium and vitamin E cancer prevention trial (SELECT). *Nutrients* **5**, 1122–48 (2013).

175. Yang, S.-W. *et al.* MART-10, a newly synthesized vitamin D analog, represses metastatic potential of head and neck squamous carcinoma cells. *Drug Des. Devel. Ther.* **10**, 1995–2002 (2016).

176. Chiang, K.-C. *et al.* MART-10, a novel vitamin D analog, inhibits head and neck squamous carcinoma cells growth through cell cycle arrest at G0/G1 with upregulation of p21 and p27 and downregulation of telomerase. *J. Steroid Biochem. Mol. Biol.* **138**, 427–434 (2013).

177. Shintani, T. *et al.* Eldecalcitol (ED-71), an analog of 1alpha,25-dihydroxyvitamin D3 as a potential anti-cancer agent for oral squamous cell carcinomas. *J. Steroid Biochem. Mol. Biol.* **164**, 79–84 (2016).

178. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: A genome-wide association study. *Lancet* **376**, 180–188 (2010).

179. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).

180. Zheng, H. F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).

181. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science (80-.).* **305**, 869–872 (2004).

182. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).

183. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

184. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

185. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).

186. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

187. Mägi, R. & Morris, A. P. GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, (2010).

188. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38**, 281–290 (2014).

189. Yang, J., Lee, H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

190. Casella, S. J., Reiner, B. J., Chen, T. C., Holick, M. F. & Harrison, H. E. A possible genetic defect in 25-hydroxylation as a cause of rickets. *J. Pediatr.* **124**, 929–932 (1994).

191. Barresi, C. *et al.* Increased sensitivity of histidinemic mice to UVB radiation suggests a crucial role of endogenous urocanic acid in photoprotection. *J. Invest. Dermatol.* **131**, 188–194 (2011).

192. Suchi, M. *et al.* Molecular cloning and structural characterization of the human histidase gene (HAL). *Genomics* **29**, 98–104 (1995).

193. Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun.* **9**, 260 (2018).

194. Manousaki, D. *et al.* Low-Frequency Synonymous Coding Variation in CYP2R1 Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis. *Am. J. Hum. Genet.* **101**, 227–238 (2017).

195. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Smith, G. D. & Tilling, K. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).

196. Brion, M.-J. J. A., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).

197. Mokry, L. E. *et al.* Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Med.* **12**, e1001866 (2015).

198. Rhead, B. *et al.* Mendelian randomization shows a causal effect of low vitamin D on multiple sclerosis risk. *Neurol. Genet.* **2**, e97 (2016).

199. Li, X. *et al.* MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank. *Ann. Rheum. Dis.* **77**, 1039–1047 (2018).

200. Canty, A. & Ripley, B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20. (2017).

201. Schwarzer, G. meta: An R package for meta-analysis. *R News* **7**, 40–45 (2007).

202. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ Br. Med. J.* **327**, 557–560 (2003).

203. Wain, L. V *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).

204. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–52 (2015).

205. Dimitrakopoulou, V. I. *et al.* Circulating vitamin D concentration and risk of seven cancers: Mendelian randomisation study. *BMJ* **359**, j4761 (2017).

206. Lappe, J. *et al.* Effect of Vitamin D and Calcium Supplementation on Cancer Incidence in Older Women. *JAMA* **317**, 1234 (2017).

207. Manson, J. E. *et al.* Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. *N. Engl. J. Med.* **380**, 33–44 (2019).

208. Ong, J. S. *et al.* Association of vitamin D levels and risk of ovarian cancer: A Mendelian randomization study. *Int. J. Epidemiol.* **45**, 1619–1630 (2016).

209. Zhang, Y. *et al.* Association between vitamin D supplementation and mortality: systematic review and meta-analysis. *BMJ* **366**, l4673 (2019).

210. Wade, K. H., Carslake, D., Sattar, N., Davey Smith, G. & Timpson, N. J. BMI and Mortality in UK Biobank: Revised Estimates Using Mendelian Randomization. *Obesity* **26**, 1796–1806 (2018).

211. Guo, Q. *et al.* Body mass index and breast cancer survival: a Mendelian randomization analysis. *Int. J. Epidemiol.* **46**, 1814–1822 (2017).

212. Syed, H., Jorgensen, A. L. & Morris, A. P. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics* **17**, 907–915 (2016).

213. Beynon, R. A. *et al.* Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer. *Int. J. Cancer* **143**, 1114–1127 (2018).

214. Ang, K. K. *et al.* Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *N. Engl. J. Med.* **363**, 24–35 (2010).

215. Therneau, T. M. A Package for Survival Analysis in S. R package version 2.38. (2015).

216. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).

217. Trivedi, D. P., Doll, R. & Khaw, K. T. Effect of four monthly oral vitamin D3 (cholecalciferol) supplementation on fractures and mortality in men and women living in the community: randomised double blind controlled trial. *BMJ* **326**, 469 (2003).

218. Benowitz, N. L. *et al.* Biochemical verification of tobacco use and cessation. *Nicotine Tob. Res.* **4**, 149–159 (2002).

219. Shenker, N. S. *et al.* DNA Methylation as a Long-term Biomarker of Exposure to Tobacco Smoke. *Source Epidemiol.* **24**, 712–716 (2013).

220. Rowling, M. J., Kemmis, C. M., Taffany, D. A. & Welsh, J. Megalin-Mediated Endocytosis of Vitamin D Binding Protein Correlates with 25-Hydroxycholecalciferol Actions in Human Mammary Cells. *J. Nutr.* **136**, 2754–2759 (2006).

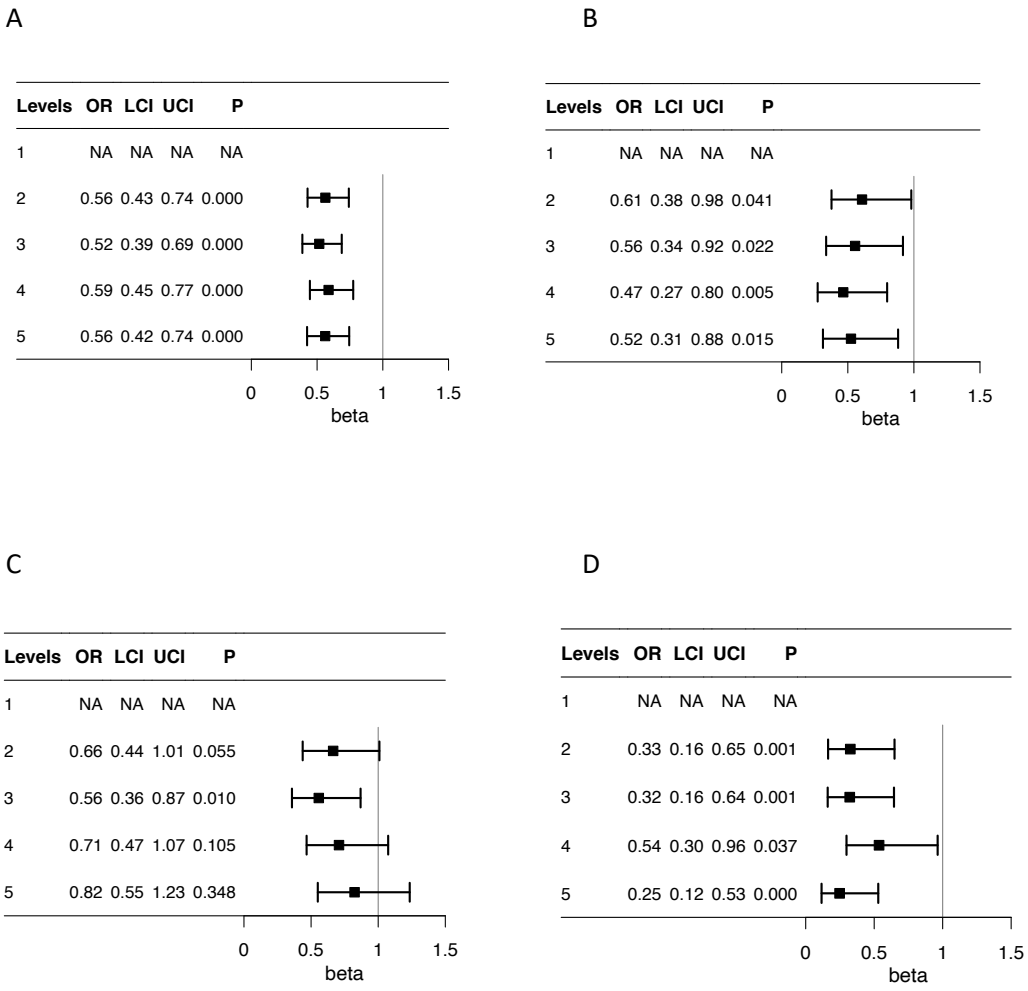
221. Ternes, S. B. & Rowling, M. J. Vitamin D Transport Proteins Megalin and Disabled-2 Are Expressed in Prostate and Colon Epithelial Cells and Are Induced and Activated by All- *Trans* -Retinoic Acid. *Nutr. Cancer* **65**, 900–907 (2013).

222. Berry, D. J., Vimaleswaran, K. S., Whittaker, J. C., Hingorani, A. D. & Hyppönen, E. Evaluation of genetic markers as instruments for Mendelian randomization studies on vitamin D. *PLoS One* **7**, 1–10 (2012).

223. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

Appendix

Appendix 4.1: Odds ratio for HNC across quintiles of 25 hydroxyvitamin D.



A = All HNC sites, B = Oral cavity cancer, C = Oropharyngeal cancer, D = Laryngeal cancer. Levels = quintiles of 25(OH)D from lowest (1) to highest (5). Odds ratio for each quintile compared to the 1st quintile.

Appendix 5.1: Detailed Description of the Participating Cohorts

Study Acronym	Imputation panel	Association testing software	Lamda	Number of SNVs analysed	N phenotyped	25OHD measurement technique	25OHD mean (SD)	Age mean(SD)	BMI mean(SD)	sex (males/females)	Detection limits for 25OHD measurement
ALSPAC Imp	HRC	SNPtest v2.5.0	1.0133	8247588	3679	HPLC/Tandem mass spec	64.2 (23.5)	9.8 (1.3)	17.6 (2.8)	1909/1770	2.5 to 624 nmol/L
ALSPAC WGS	-	SNPtest v2.5.0	0.999	9886094	1606	HPLC/Tandem mass spec	61.8 (19.9)	10.1 (0.8)	17.6 (2.9)	789/817	2.5 to 624 nmol/L
BPROOF	UK10K/1000G	GEMMA	0.9991	9857390	2514	SLC-MS/Ms/tandem mass spec	56.2(24.6)	73.5(6.5)	27.1 (4.0)	1271/1243	The inter-assay coefficient of variation was 9% at 25 nmol/L and 6% at 62.4 nmol/L
MROS	UK10K/1000G	GEMMA	1.0111	10012764	3266	Mass spec	62.1 (20.8)	74.1(5.8)	27.4(3.8)	3266/0	1.6nmol/L
TUK Imputed	UK10K/1000G	GEMMA	1.0076	9614367	1920	chemiluminescence immunoassay (DIASORIN)	76.5(42.4)	49.2(13.8)	25.5(4.8)	1744/176	4 to 10 nmol/L
TUK WGS	-	GEMMA	0.9983	9505928	953	chemiluminescence immunoassay (DIASORIN)	76.9 (41.1)	48.1 (11.7)	25.3 (4.8)	953/0	4 to 10 nmol/L
CHS	1000G	Custom R program	0.9951	8569308	1792	HPLC/Tandem mass spec	66.6 (27.0)	74.1(4.7)	26.4(4.5)	1250/542	2.0 to 200 nmol/L
PIVUS	HRC	SNPTEST	1.0018	8830580	943	chemiluminescence immunoassay (DIASORIN)	57.6 (19.7)	70.1 (0.2)	27.0(4.4)	471/472	10 nmol/L
ULSAM	HRC	SNPTEST	0.9949	8879176	1095	HPLC-APSI-MS	68.7 (19.1)	71.0 (0.6)	26.3(3.4)	1095/0	1 to 4 nmol/L
NEO	1000G	PROBABEL	1.0146	8454808	5744	LC-MS/MS	68.9 (23.6)	56.0 (5.9)	30.0 (4.8)	2757/2987	1.6nmol/L
MROS Malmo	UK10K/1000G	PLINK	1.0079	10292745	893	chemiluminescence immunoassay (DIASORIN)	66.6 (24.3)	75.6 (3.2)	26.5 (3.6)	893/0	10 nmol/L
MROS GBG	UK10K/1000G	PLINK	1.015	10311274	947	chemiluminescence immunoassay DIASORIN)	66.7 (18.8)	75.2 (3.2)	26.2 (3.5)	947/0	10 nmol/L
GOOD	UK10K/1000G	PLINK	1.0042	10336932	921	chemiluminescence immunoassay (DIASORIN)	65.2 (21.7)	18.9 (0.6)	22.4 (3.2)	921/0	10 nmol/L
BMDCS	UK10K	GEMMA	1.0032	9537533	863	Tandem mass spec	49.3 (15.4)	14.9 (4.7)	21.1 (4.1)	420/443	1.6-285 nmol/L
FHS	UK10K/1000G	R Kinship package	1.1917	9396749	5402	chemiluminescence immunoassay (DIASORIN)	78.4 (37.1)	47.4 (13.1)	27.0 (5.3)	2539/2863	4 to 10 nmol/L
Generation R	UK10K/1000G	SNPTEST	1.0222	9555270	1442	LC-MS/MS	76.5 (26.4)	6.11(0.4)	15.9(1.5)	731/711	4 nmol/L
RSI	UK10K/1000G	SNPTEST	1.0117	9540518	3320	electrochemiluminescence immunoassay (COBAS, Roche Diagnostics GmbH)	49.4 (25.3)	73.4 (7.6)	26.9(4.0)	1390/1930	3 nmol/L
RSII	UK10K/1000G	SNPTEST	0.9936	9554142	2022	electrochemiluminescence immunoassay (COBAS, Roche Diagnostics GmbH)	62.4(28.3)	64.7(8.0)	27.3(4.0)	919/1103	3 nmol/L
RSIII	UK10K/1000G	SNPTEST	1.0037	9558464	2913	electrochemiluminescence immunoassay (COBAS, Roche Diagnostics GmbH)	61.9(27.6)	57.1(6.8)	27.7(4.6)	1277/1636	3 nmol/L

Appendix 5.2: Participating Cohorts: study designs and locations

Short name	Full name	Study design	Study Type	Country of Origin	Ethnicity
ALSPAC	The Avon Longitudinal Study of Parents and Children	Cohort	Population-based, family-based	United Kingdom	North-western European
TUK	TwinsUK	Cohort	Population-based, family-based	United Kingdom	North-western European
FHS	Framingham Heart Study	Cohort	Population-based, family-based	United States of America	European American
MrOS	Osteoporotic Fractures in Men USA	Cohort	General population	United States of America	Non-Hispanic white
BPROOF	B-Vitamins for the PRvention Of Osteoporotic Fractures	Cohort	Population-based, family-based		
MrOS GBG	Osteoporotic Fractures in Men Gothenburg	Cohort	Population-based	Sweden	North-western European
MrOS MALMO	Osteoporotic Fractures in Men Malmo	Cohort	Population-based	Sweden	North-western European
GOOD	Gothenburg Osteoporosis and Obesity Determinants	Cohort	Population-based	Sweden	North-western European
CHS	Cardiovascular Health Study	Cohort	Population-based	United States	European American
RSI	Rotterdam Study-I	Cohort	Population-based, family-based	The Netherlands	North-western European
RSII	Rotterdam Study-II	Cohort	Population-based, family-based	The Netherlands	North-western European
RSIII	Rotterdam Study-III	Cohort	Population-based, family-based	The Netherlands	North-western European
PIVUS	Prospective Investigation of the Vasculature in Upssala Seniors	Cohort	Population-based sample from the general population	Sweden	North-western European
ULSAM	Upssala Longitudinal Study of Adult Men	Cohort	Population-based sample from the general population	Sweden	North-western European
Generation R	The Generation R Study	Cohort	Population-based, family-based	The Netherlands	Admixed
BMDCS	Bone Mineral Density in Childhood Study	Cohort	population-based	United States of America	European American
NEO	Netherlands Epidemiology of Obesity	Cohort	Population-based, oversampling of individuals with overweight	The Netherlands	European American

Further details about the participating studies can be found in Supplementary material for the published paper (https://www.cell.com/ajhg/fulltext/S0002-9297(17)30277-X)

Appendix 5.3: Metrics for Sequencing-Based Cohorts

Cohort	Target	Platform	Capture	Sample size	Mean depth	Monomorphic/ Singletons?	Variant QC	Calling method	Additional variant QC filters pre-association
TWINSUK	Genome	Illumina HiSeq 2000	No capture, WGS.	1934	6x	Singletons included	VQSR (with standard GATK parameters), only variants with INFO=PASS were included	samtools and bcftools	Subsequent to conversion to binary GEN format, variants were filtered by removing variants with INFO < 0.4 and HWE p-value < 1e-6.
ALSPAC	Genome	Illumina HiSeq 2000	No capture, WGS.	1976	6x	Singletons included	VQSR (with standard GATK parameters), only variants with INFO=PASS were included	samtools and bcftools	Subsequent to conversion to binary GEN format, variants were filtered by removing variants with INFO < 0.4 and HWE p-value < 1e-6.

Appendix 5.4: Info Scores for Imputed SNVs per Participating Cohort

			Average Info score according to MAF			Number of SNVs according to MAF			
Cohort	**Imputation reference panel**	**Average Info Score**	**0.5%-1%**	**1%-5%**	**>5%**	**0.5%-1%**	**1%-5%**	**>5%**	**total**
ALSPAC Imputed	HRC	0.953898	0.86779	0.929003	0.983156	1039739	2240923	4966926	8247588
BMDCS	UK10K	0.9337	0.813959	0.8921	0.97537	1170457	2508752	5864874	9544083
BPROOF	UK10K/1000G	0.903492	0.761488	0.851148	0.957815	1269564	2683536	5904456	9857556
CHS	1000G	0.861307	0.695624	0.764028	0.912802	636475	1989577	5806416	8432468
FHS	UK10K/1000G	0.901056	0.761261	0.856422	0.952144	1308922	2515245	5779103	9603270
GenerationR	UK10K/1000G	0.991865	0.99749	0.993785	0.979851	1065474	2130392	2963012	6158878
GOOD	UK10K/1000G	0.963843	0.855494	0.930359	1.00184	1365394	2699130	6272408	10336932
MrOS	UK10K/1000G	0.906883	0.777368	0.843646	0.958579	1057189	2608734	5839794	9505717
MROS MALMO	UK10K/1000G	0.969319	0.861881	0.93872	1.00542	1338140	2691611	6262995	10292746
MROS GBG	UK10K/1000G	0.973695	0.871793	0.945207	1.00796	1350188	2700557	6260530	10311275
NEO	1000G	0.860829	0.67615	0.778314	0.924774	900228	2163238	5391343	8454809
PIVUS	HRC	0.963494	0.883933	0.94073	0.989421	1096722	2326148	5407710	8830580
RSI	UK10K/1000G	0.992253	0.997564	0.99386	0.980092	1106406	2065041	2828682	6000129
RSII	UK10K/1000G	0.991956	0.997524	0.993753	0.979589	1103454	2099961	2883101	6086516
RSII	UK10K/1000G	0.992352	0.997571	0.993931	0.980243	1092398	2042174	2826194	5960766
TUK Imputed	UK10K/1000G	0.936431	0.830858	0.89187	0.973654	923412	2599834	5731415	9254661
USLAM	HRC	0.976243	0.913023	0.963587	0.994818	1122991	2339719	5416466	8879176

MAF = minor allele frequency

Appendix 5.5: Results for the Lead SUNLIGHT SNVs in the Discovery Meta-analysis

SNV	**Effect Allele***	**Effect Allele* Frequency**	**Gene**	**SUNLIGHT (N=16,125)**	**SUNLIGHT_Seq meta (N=42,274)**	
				P-value	beta	P-value
rs2282679	G	0.28	*GC*	1.9×10^{-109}	-0.25	2.6×10^{-258}
rs12785878	G	0.3	*DHCR7*	2.1×10^{-27}	-0.1	3.2×10^{-40}
rs10741657	G	0.58	*CYP2R1*	3.3×10^{-20}	-0.09	8.8×10^{-45}
rs6013897	A	0.21	*CYP24A1*	6.0×10^{-10}	-0.07	9.1×10^{-16}

***effect allele is the 25OHD decreasing allele**

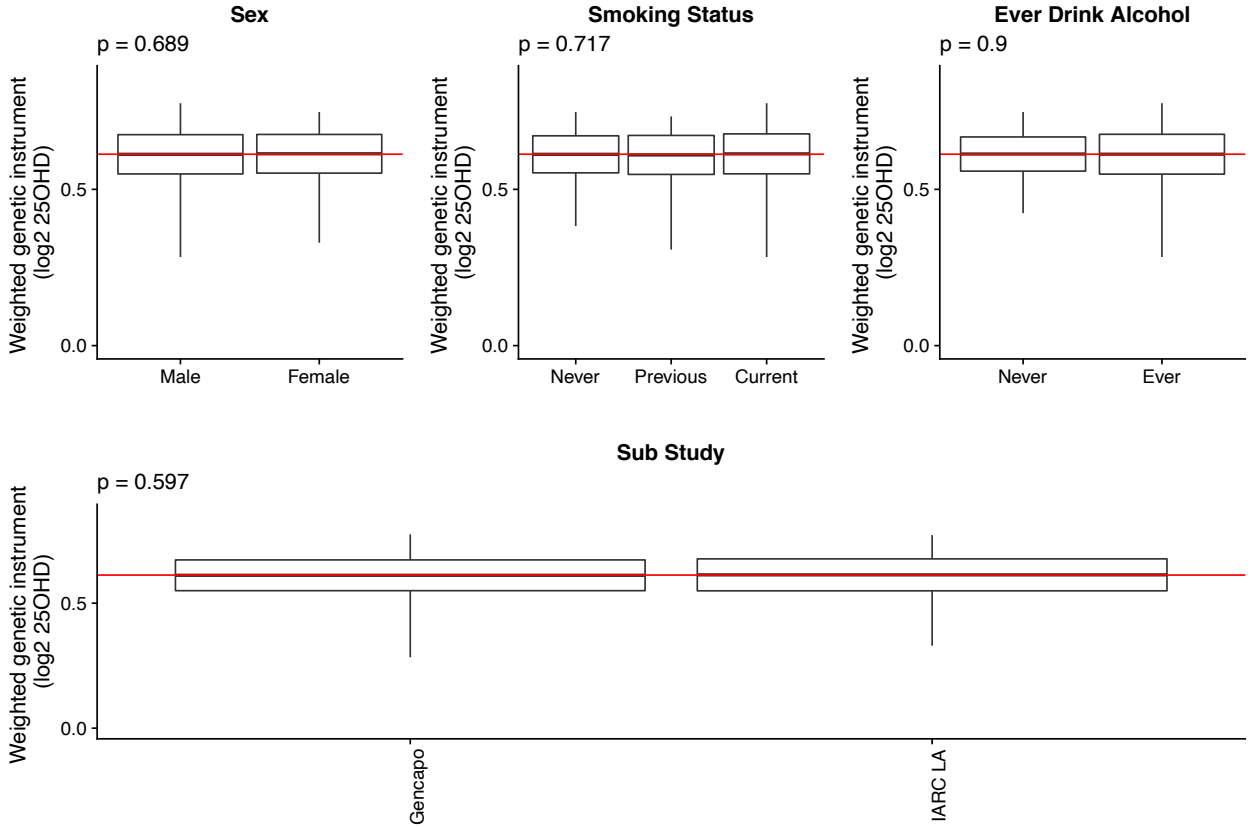
Appendix 5.6: Haplotype Counts in the ALSPAC WGS Study for the Six SNVs of the Same Chromosome 11 Locus.

N of chromatids			**Gene/SNV**					
			PDE3B	PDE3B	PDE3B	PDE3B	PDE3B	CYP2R1
all	**chromatid2**	**chromatid1**	**rs117672174**	**rs142830933**	**rs117621176**	**rs117361591**	**rs116970203**	**rs117913124**
7341	1516	1605	C	A	A	G	G	G
111	51	0	C	A	A	G	**A**	**A**
75	37	1	**T**	**C**	**G**	**T**	**A**	**A**
14	1	0	C	**C**	**G**	**T**	**A**	**A**
1	1	0	**T**	**C**	**G**	**T**	G	**A**

Appendix 6.1: Top 50 results (by p-value) for observational 25(OH)D phenome scan

varName	varType	n	beta	lower	upper	P	resType	description
48	CONTINUOUS	306489	-0.175754407	-0.178803462	-0.172705353	0	LINEAR	Waist circumference
49	CONTINUOUS	306451	-0.157643996	-0.161148576	-0.154139415	0	LINEAR	Hip circumference
670-1	CAT-SIN	279740/306249	-999	-999	-999	0	MULTINOMIAL-LOGISTIC	Type of accommodation lived in
680-1	CAT-SIN	165316/303268	-999	-999	-999	0	MULTINOMIAL-LOGISTIC	Own or rent accommodation lived in
728	CAT-SIN	305191	0.256794482	0.249967627	0.263621338	0	ORDERED-LOGISTIC	Number of vehicles in household
864	INTEGER	302066	0.135211317	0.128596412	0.141826222	0	ORDERED-LOGISTIC	Number of days/week walked 10+ minutes
884	INTEGER	292622	0.188947452	0.182424402	0.195470502	0	ORDERED-LOGISTIC	Number of days/week of moderate physical activity 10+ minutes
904	INTEGER	292796	0.307301383	0.300464504	0.314138263	0	ORDERED-LOGISTIC	Number of days/week of vigorous physical activity 10+ minutes
924	CAT-SIN	305360	0.227605588	0.220426912	0.234784263	0	ORDERED-LOGISTIC	Usual walking pace
102	INTEGER	289255	-0.150934017	-0.154536185	-0.147331849	0	LINEAR	Pulse rate, automated reading
189	CONTINUOUS	306624	-0.122308471	-0.125827624	-0.118789318	0	LINEAR	Townsend deprivation index at recruitment
914	INTEGER	168444	0.100654627	0.095594106	0.105715148	0	LINEAR	Duration of vigorous activity
1418-2	CAT-SIN	200299/306835	-999	-999	-999	0	MULTINOMIAL-LOGISTIC	Milk type used
1428-3	CAT-SIN	164383/306685	-999	-999	-999	0	MULTINOMIAL-LOGISTIC	Spread type
1050	INTEGER	279370	0.272977265	0.265555694	0.280398836	0	ORDERED-LOGISTIC	Time spend outdoors in summer
1090	INTEGER	204779	0.16487138	0.15666761	0.17307515	0	ORDERED-LOGISTIC	Time spent driving
1239	CAT-SIN	306826	-0.296045976	-0.307381124	-0.284710829	0	ORDERED-LOGISTIC	Current tobacco smoking
1329	CAT-SIN	305626	0.192251789	0.185643682	0.198859897	0	ORDERED-LOGISTIC	Oily fish intake
1558	CAT-SIN	306770	-0.156142573	-0.162560472	-0.149724674	0	ORDERED-LOGISTIC	Alcohol intake frequency.
1717	CAT-SIN	303242	0.231474636	0.223436374	0.239512898	0	ORDERED-LOGISTIC	Skin colour
1727	CAT-SIN	301139	-0.287464016	-0.294138657	-0.280789374	0	ORDERED-LOGISTIC	Ease of skin tanning
2050	CAT-SIN	294320	-0.174462601	-0.182984364	-0.165940838	0	ORDERED-LOGISTIC	Frequency of depressed mood in last 2 weeks
2060	CAT-SIN	297464	-0.19738052	-0.206215313	-0.188545726	0	ORDERED-LOGISTIC	Frequency of unenthusiasm / disinterest in last 2 weeks
2080	CAT-SIN	298262	-0.213030766	-0.220017353	-0.206044179	0	ORDERED-LOGISTIC	Frequency of tiredness / lethargy in last 2 weeks
2188	CAT-SIN	202852/97316(300168)	-0.193801949	-0.201567342	-0.18604043	0	LOGISTIC-BINARY	Long-standing illness, disability or infirmity
2316	CAT-SIN	239762/61889(301651)	-0.177903734	-0.186605653	-0.16920025	0	LOGISTIC-BINARY	Wheeze or whistling in the chest in last year
2443	CAT-SIN	291189/15096(306285)	-0.308404292	-0.323702047	-0.293074295	0	LOGISTIC-BINARY	Diabetes diagnosed by doctor
2178	CAT-SIN	305934	-0.294071875	-0.301144621	-0.286999129	0	ORDERED-LOGISTIC	Overall health rating
2267	CAT-SIN	305323	0.159430799	0.15272301	0.166138588	0	ORDERED-LOGISTIC	Use of sun/uv protection
2277	INTEGER	289968	1.029646469	1.007372112	1.051920826	0	ORDERED-LOGISTIC	Frequency of solarium/sunlamp use
3062	CONTINUOUS	279762	0.070531398	0.067841152	0.073221643	0	LINEAR	Forced vital capacity (FVC)
3063	CONTINUOUS	279762	0.069737647	0.066985902	0.072489392	0	LINEAR	Forced expiratory volume in 1-second (FEV1)
3064	INTEGER	279762	0.065110813	0.06209445	0.068127176	0	LINEAR	Peak expiratory flow (PEF)
3143	CONTINUOUS	177345	-0.103598869	-0.107618772	-0.099578966	0	LINEAR	Ankle spacing width
3637	CAT-SIN	147724	0.211638156	0.201613089	0.221663223	0	ORDERED-LOGISTIC	Frequency of other exercises in last 4 weeks
4194	INTEGER	99605	-0.145886783	-0.152122018	-0.139651547	0	LINEAR	Pulse rate
4717	CAT-SIN	88851/10215(99066)	-0.388752912	-0.408579652	-0.368929006	0	LOGISTIC-BINARY	Shortness of breath walking on level ground
4548	CAT-SIN	100348	-0.26728787	-0.27946725	-0.255108491	0	ORDERED-LOGISTIC	Health satisfaction
6141#1	CAT-MUL	77190/228009(305199)	0.263728018	0.2555961	0.271864664	0	LOGISTIC-BINARY	How are people in household related to participant: Husband, wife or partner
6142#2	CAT-MUL	192885/113233(306118)	0.216318398	0.20534117	0.227305726	0	LOGISTIC-BINARY	Current employment status: Retired
6142#4	CAT-MUL	295010/11108(306118)	-0.370747378	-0.387704492	-0.353748459	0	LOGISTIC-BINARY	Current employment status: Unable to work because of sickness or disability
6143#3	CAT-MUL	130255/33034(163289)	-0.228728321	-0.240614883	-0.216845675	0	LOGISTIC-BINARY	Transport type for commuting to job workplace: Public transport
6145#6	CAT-MUL	271610/33858(305468)	-0.280661302	-0.291602496	-0.269715927	0	LOGISTIC-BINARY	Illness, injury, bereavement, stress in last 2 years: Financial difficulties
6146#100	CAT-MUL	17170/287882(305052)	0.320633608	0.306385723	0.334859128	0	LOGISTIC-BINARY	Attendance/disability/mobility allowance
6146#2	CAT-MUL	292206/12846(305052)	-0.34646434	-0.362409564	-0.330483965	0	LOGISTIC-BINARY	Attendance/disability/mobility allowance: Disability living allowance
6155#100	CAT-MUL	95200/210619(305819)	-0.415549718	-0.424054816	-0.40705948	0	LOGISTIC-BINARY	Vitamin and mineral supplements
6155#7	CAT-MUL	240196/65623(305819)	0.478724694	0.468930364	0.488538765	0	LOGISTIC-BINARY	Vitamin and mineral supplements: Multivitamins +/- minerals

Appendix 7.1: Mean (+/- standard deviation) and range of relative 25-hydroxyvitamin D across strata of potential confounders for South America Region only in GAMEON.



Appendix 8.1: Methods for HNC prognosis GWAS

Studies

Head and neck 5000

HN5000 is a large collection of HNC cases with detailed phenotyping and follow-up. An overall description of the HN5000 study is provided in Chapter 3.

UK Biobank

UK Biobank is a population-based study of approximately 500,000 participants. A detailed description of UK Biobank study is provided in Chapter 3.

DNA extraction genotyping and imputation

DNA extraction, genotyping and imputation for UK Biobank and HN5000 is described in Chapter 3.

Follow-up of HNC cases

Cancer cases were followed up in both studies using death registry data to identify the date of death. All-cause mortality was identified from death registry data as a death of a participant from any cause. Due to the limited sample size (UK Biobank) and data-availability (HN5000), HNC-specific cause of death was not assessed in this study.

Survival time was calculated from date of diagnosis to date of death. Those participants alive at the end of follow up were censored. Censoring was on 31st March 2018 in HN5000 (last data capture from death registry) and 14th February 2018 in UK Biobank (last recorded death in the UK Biobank data freeze used for this study).

Statistical analysis

GWAS analysis

A GWAS was conducted for HNC survival in HN5000 and UK Biobank separately using the SurvivalGWAS_Sv program.[212] This program applies a Cox proportional hazards regression model at each genetic variant. The number of risk alleles at each variant was used as the exposure and time to event data for all HNC sites combined was used as the outcome, age and sex were used as covariates in the regression model. A large number of genetic variants in both studies have minor allele frequencies less than 1%. Due to the relatively small sample sizes available these low frequency alleles would not be present in enough cases to allow meaningful analyses and were removed prior to running the GWAS. Genetic variants with imputation quality scores less than 0.3 were also removed prior to running the GWAS analysis.

Appendix

Meta-analysis

Results from the two studies were meta-analyses using a fixed effect in GWAMA.[187] This approach includes automated quality control steps which identify strand problems, incorrect alleles, nonsensical effect sizes and duplicate variants.

Appendix

Appendix papers:

Paper 1: Manousaki, D., Dudding, T. *et al.* (2017) 'Low-Frequency Synonymous Coding Variation in CYP2R1 Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis', *American Journal of Human Genetics*, 101(2), pp. 227–238. **(Chapter 5)**

Paper 2: Dudding, T. *et al.* (2018) 'Assessing the causal association between 25-hydroxyvitamin D and the risk of oral and oropharyngeal cancer using Mendelian randomization.', *International journal of cancer*, 143(5), pp. 1029–1036. **(Chapter 7)**

Low-Frequency Synonymous Coding Variation in *CYP2R1* Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis

Despoina Manousaki,[1,2,49] Tom Dudding,[3,49] Simon Haworth,[3,49] Yi-Hsiang Hsu,[4,5,6,49] Ching-Ti Liu,[7,49] Carolina Medina-Gómez,[8,9,10,49] Trudy Voortman,[9,10,49] Nathalie van der Velde,[8,11,49] Håkan Melhus,[12,49] Cassianne Robinson-Cohen,[13,49] Diana L. Cousminer,[14,15,49] Maria Nethander,[16,17,49] Liesbeth Vandenput,[16,49] Raymond Noordam,[18,49] Vincenzo Forgetta,[1,2] Celia M.T. Greenwood,[1,2,19,20] Mary L. Biggs,[21] Bruce M. Psaty,[22,23] Jerome I. Rotter,[24,25] Babette S. Zemel,[26,27] Jonathan A. Mitchell,[26,27] Bruce Taylor,[28] Mattias Lorentzon,[16,29,30]

(Author list continued on next page)

Vitamin D insufficiency is common, correctable, and influenced by genetic factors, and it has been associated with risk of several diseases. We sought to identify low-frequency genetic variants that strongly increase the risk of vitamin D insufficiency and tested their effect on risk of multiple sclerosis, a disease influenced by low vitamin D concentrations. We used whole-genome sequencing data from 2,619 individuals through the UK10K program and deep-imputation data from 39,655 individuals genotyped genome-wide. Meta-analysis of the summary statistics from 19 cohorts identified in *CYP2R1* the low-frequency (minor allele frequency = 2.5%) synonymous coding variant g.14900931G>A (p.Asp120Asp) (rs117913124[A]), which conferred a large effect on 25-hydroxyvitamin D (25OHD) levels (−0.43 SD of standardized natural log-transformed 25OHD per A allele; p value = 1.5×10^{-88}). The effect on 25OHD was four times larger and independent of the effect of a previously described common variant near *CYP2R1*. By analyzing 8,711 individuals, we showed that heterozygote carriers of this low-frequency variant have an increased risk of vitamin D insufficiency (odds ratio [OR] = 2.2, 95% confidence interval [CI] = 1.78–2.78, p = 1.26×10^{-12}). Individuals carrying one copy of this variant also had increased odds of multiple sclerosis (OR = 1.4, 95% CI = 1.19–1.64, p = 2.63×10^{-5}) in a sample of 5,927 case and 5,599 control subjects. In conclusion, we describe a low-frequency *CYP2R1* coding variant that exerts the largest effect upon 25OHD levels identified to date in the general European population and implicates vitamin D in the etiology of multiple sclerosis.

Introduction

Vitamin D insufficiency affects approximately 40% of the general population in developed countries.[1] This could have important public health consequences, given that vitamin D insufficiency has been associated with musculoskeletal consequences and several common diseases, such as multiple sclerosis (MIM: 126200), type 1 diabetes (MIM: 222100), type 2 diabetes (MIM: 125853), and several cancers.[2] Further, repletion of vitamin D status can be achieved safely and inexpensively. Thus, understanding the determinants of vitamin D insufficiency, and their effects, can provide a better understanding of the role of vitamin D in disease susceptibility with potentially important public health benefits.

Approximately half of the variability in the concentration of the widely accepted biomarker for vitamin D status, 25-hydroxyvitamin D (25OHD), has been attributed to genetic factors in twin and family studies.[3,4] Four common (minor allele frequency [MAF] > 5%) genetic variants in loci near four genes known to be involved in cholesterol synthesis (*DHCR7* [MIM: 602858]), hydroxylation (*CYP2R1* [MIM: 608713]), vitamin D transport (*GC* [MIM: 139200]), and catabolism (*CYP24A1* [MIM: 126065]) are strongly associated with 25OHD levels yet explain little of its heritability.[5] Low-frequency and rare genetic variants (defined as those with a MAF ≤ 5% and 1%, respectively) have recently been found to have large effects on clinically relevant traits,[6–8] providing an opportunity to better understand the

[1]Department of Human Genetics, McGill University, Montreal, QC H3A 1B1, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada; [3]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK; [4]Institute for Aging Research, Hebrew SeniorLife, Boston, MA 02131, USA; [5]Harvard Medical School, Boston, MA 02115, USA; [6]Broad Institute of MIT and Harvard, Boston, MA 02142, USA; [7]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA; [8]Department of Internal Medicine, Erasmus Medical Center, Rotterdam 3015 GE, the Netherlands; [9]Generation R Study Group, Erasmus Medical Center, Rotterdam 3015 GE, the Netherlands; [10]Department of Epidemiology, Erasmus Medical Center, Rotterdam 3015 GE, the Netherlands; [11]Section of Geriatrics, Department of Internal Medicine, Academic Medical Center, Amsterdam 1105 AZ, the Netherlands; [12]Department of Medical Sciences, Uppsala University, Uppsala 751 85, Sweden; [13]Kidney Research Institute, Division of Nephrology, University of Washington, Seattle, WA 98195, USA; [14]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [15]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [16]Centre for Bone and Arthritis Research, Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg 40530, Sweden; [17]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg,

(Affiliations continued on next page)

© 2017 American Society of Human Genetics.

Magnus Karlsson,[31,32] Vincent V.W. Jaddoe,[9,10] Henning Tiemeier,[9,10,33] Natalia Campos-Obando,[8] Oscar H. Franco,[10] Andre G. Utterlinden,[8,9,10] Linda Broer,[8] Natasja M. van Schoor,[34] Annelies C. Ham,[8] M. Arfan Ikram,[10,35] David Karasik,[4] Renée de Mutsert,[36] Frits R. Rosendaal,[36] Martin den Heijer,[37] Thomas J. Wang,[38] Lars Lind,[12,50] Eric S. Orwoll,[39,40,50] Dennis O. Mook-Kanamori,[36,41,50] Karl Michaëlsson,[42,50] Bryan Kestenbaum,[13,50] Claes Ohlsson,[16,50] Dan Mellström,[16,29,50] Lisette C.P.G.M. de Groot,[43,50] Struan F.A. Grant,[14,26,44,50] Douglas P. Kiel,[4,5,6,45,50] M. Carola Zillikens,[8,50] Fernando Rivadeneira,[8,9,10,50] Stephen Sawcer,[46,50] Nicholas J. Timpson,[3,50] and J. Brent Richards[1,2,47,48,50,*]

biologic mechanisms influencing disease susceptibility in the general population.

Therefore, the principal objective of the present study was to detect low-frequency and rare variants with large effects on 25OHD levels through a large-scale meta-analysis and describe their biological and clinical relevance. Similar to an earlier genome-wide association study (GWAS) examining common (MAF $\geq$ 5%) genetic variation by the SUNLIGHT consortium,[5] we sought to increase understanding of the genetic etiology of vitamin D variation within the general population; however, our current study focused on genetic variation with a MAF $<$ 5%. This has only recently been made possible through whole-genome sequencing (WGS) and the use of improved genotype imputation for low-frequency and rare variants with the recent availability of large WGS reference panels.[9] The second objective of this study was to better understand whether low-frequency genetic variants with large effects on 25OHD could predict a higher risk of vitamin D insufficiency in their carriers and whether vitamin D intake through diet might interact with such genetic factors to prevent, or magnify, vitamin D insufficiency. Finally, we sought to understand whether these genetic determinants of 25OHD levels are implicated in multiple sclerosis, a disease influenced by low 25OHD levels.[10]

To do so, we first undertook an association study of WGS data and deeply imputed genome-wide genotypes to identify novel genetic determinants of vitamin D in 42,274 individuals. We next tested if these genetic variants conferred a higher risk of vitamin D insufficiency in 8,711 subjects and whether this insufficiency showed effect modification by dietary intake. Last we assessed their effect on multiple sclerosis in a separate sample of 5,927 case and 5,599 control subjects.

Material and Methods

Cohorts

All human studies were approved by each respective institutional or national ethics review committee, and all participants provided written informed consent. To investigate the role of rare and low-frequency genetic variation on 25OHD levels in individuals of European descent, we used WGS data at a mean read depth of 6.7× in 2,619 subjects from two cohorts with available 25OHD phenotypes in the UK10K project[11] (Table 1). We also used imputation reference panels to impute variants that were missing, or poorly captured, from previous GWASs of 39,655 subjects (Table 1 and Figure 1). The participating individuals were drawn from independent cohorts of individuals of European descent. A detailed description of each of the participating studies is provided in Table S1.

25OHD Measurements

The methods applied for measuring 25OHD levels differed among the participating cohorts (Tables S1 and S6). The four methods

Gothenburg 41390, Sweden; [18]Section of Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [19]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC H3A 1A2, Canada; [20]Department of Oncology, McGill University, Montreal, QC H4A 3T2, Canada; [21]Cardiovascular Health Research Unit, Departments of Medicine and Biostatistics, University of Washington, Seattle, WA 98101, USA; [22]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA 98101, USA; [23]Kaiser Permanente Washington Health Research Unit, Seattle, WA 98101, USA; [24]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute, Torrance, CA 90502, USA; [25]Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [26]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [27]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [28]Menzies Institute for Medical Research University of Tasmania, Locked Bag 23, Hobart, Tasmania 7000, Australia; [29]Geriatric Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, 43180 Mölndal, Sweden; [30]Geriatric Medicine, Sahlgrenska University Hospital, 43180 Mölndal, Sweden; [31]Clinical and Molecular Osteoporosis Research Unit, Department of Clinical Sciences, Lund University, 22241 Malmö, Sweden; [32]Department of Orthopaedics, Skåne University Hospital, 22241 Malmö, Sweden; [33]Department of Child and Adolescent Psychiatry/Psychology, Erasmus Medical Center, Rotterdam 3015 GE, the Netherlands; [34]Department of Epidemiology and Biostatistics and EMGO Institute of Health and Care Research, VU University Medical Center, Amsterdam 1081 HV, the Netherlands; [35]Department of Radiology and Nuclear Medicine, Erasmus Medical Center, Rotterdam 3015 GE, the Netherlands; [36]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [37]Department of Endocrinology, VU University Medical Center, Amsterdam 1081 HV, the Netherlands; [38]Division of Cardiovascular Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [39]Bone and Mineral Unit, Oregon Health & Science University, Portland, OR 97239, USA; [40]Department of Medicine, Oregon Health & Science University, Portland, OR 97239, USA; [41]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; [42]Department of Surgical Sciences, Uppsala University, 75105 Uppsala, Sweden; [43]Division of Human Nutrition, Wageningen University, Wageningen 6708 WE, the Netherlands; [44]Division of Endocrinology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [45]Beth Israel Deaconess Medical Center, Boston, MA 02215, USA; [46]Department of Clinical Neurosciences, University of Cambridge, Box 165, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0QQ, UK; [47]Department of Twin Research and Genetic Epidemiology, King's College London, London WC2R 2LS, UK; [48]Department of Medicine, McGill University, Montreal, QC H3G 1Y6, Canada
[49]These authors contributed equally to this work
[50]These authors contributed equally to this work
*Correspondence: brent.richards@mcgill.ca
http://dx.doi.org/10.1016/j.ajhg.2017.06.014.

Table 1. Participating Cohorts and Number of DNA Samples per Cohort

Study	Imputed	Whole-Genome Sequenced
ALSPAC	3,679	1,606
TUK	1,919	1,013
Generation R	1,442	–
BPROOF	2,514	–
FHS	5,402	–
MrOS	3,265	–
RSI	3,320	–
RSII	2,022	–
RSIII	2,913	–
CHS	1,792	–
BMDCS	863	–
MrOS GBG	945	–
GOOD	921	–
MrOS Malmo	893	–
PIVUS	943	–
ULSAM	1,095	–
NEO	5,727	–
Total	**39,655**	**2,619**

used were tandem mass spectrometry (in the Bone Mineral Density in Childhood Study [BMDCS], Osteoporotic Fractures in Men USA [MrOS], and B-Vitamins for the Prevention of Osteoporotic Fractures [BPROOF]), combined high-performance liquid chromatography and mass spectrometry (in the Avon Longitudinal Study of Parents and Children [ALSPAC], BPROOF, Cardiovascular Health Study [CHS], Upssala Longitudinal Study of Adult Men [ULSAM], Netherlands Epidemiology of Obesity [NEO], and Generation R Study [Generation R]), chemiluminescence immunoassay (DiaSorin) (in TwinsUK [TUK], the Prospective Investigation of the Vasculature in Upssala Seniors [PIVUS], the Framingham Heart Study [FHS], Osteoporotic Fractures in Men Malmo [MrOS Malmo], Osteoporotic Fractures in Men Gothenburg [MrOS GBG], and Gothenburg Osteoporosis and Obesity Determinants [GOOD]), and electrochemiluminescence immunoassay (COBAS, Roche Diagnostics) (in Rotterdam Studies I [RSI], II [RSII], and III [RSIII]). Detection limits for the different methods are provided in Table S6.

WGS, Genotyping, and Imputation

ALSPAC WGS and TUK WGS cohorts had been sequenced at an average read depth of 6.7× through the UK10K consortium on the Illumina HiSeq platform and aligned to the GRCh37 human reference sequence with Burrows-Wheeler Aligner 31.[12] Single-nucleotide variant (SNV) calls were completed with SAMtools/BCFtools,[13] and VQSR[14] and GATK were used to recall these variants. WGS for the ALSPAC and TUK cohorts has been described in detail in a previous publication from our group.[7] Table S8 summarizes the data-generation method for sequencing-based cohorts.

Participating studies separately genotyped samples and imputed them to WGS-based reference panels. The most recent imputation panels, such as the UK10K and 1000 Genomes Project (v.3) combined panel (7,562 haplotypes from the UK10K project and 2,184 haplotypes from the 1000 Genomes Project[9]) and the Haplotype Reference Consortium (HRC) panel (64,976 haplotypes[15]), enabled more accurate imputation of low-frequency variants than the UK10K or 1000 Genomes reference panel alone.[9] Specifically, 11 of the 17 participating cohorts were imputed to the combined UK10K and 1000 Genomes reference panel (total number of imputed individuals included in the meta-analysis = 25,589). Three of the participating cohorts were imputed with the HRC panel (n = 5,717). Finally, two cohorts were imputed to the 1000 Genomes panel (n = 7,536), and one cohort was imputed to the UK10K panel (n = 863) (Table S1). Details on genotyping methods and imputation for the 17 participating cohorts are presented in Table S6. Info scores for the imputed SNVs per participating cohort are presented in Table S7. To assess the quality of imputation, we tested the non-reference discordance rate for the low-frequency genome-wide-significant SNVs and found this to be 0% (Table S9).

Association Testing for 25OHD Levels and Meta-analysis

We conducted a GWAS separately for each cohort by using an additive genetic model for 25OHD levels. Because 25OHD concentrations were measured by different methods, log-transformed 25OHD levels were standardized to Z scores after adjustment for age, sex, BMI, and season of measurement. Specifically, the phenotype for each GWAS was prepared according to the following steps: (1) We log transformed 25OHD levels to ensure normality. (2) We used linear regression models to generate cohort-specific residuals of log-transformed 25OHD levels adjusted for covariates (age, sex, BMI, and season). Season was treated as a non-ordinal categorical variable (summer: July to September; fall: October to December; winter: January to March; and spring: April to June). (3) We added the mean of log-transformed 25OHD levels to the residuals to create the adjusted 25OHD phenotype. (4) We then normalized the above phenotype within each cohort (mean of 0 with 1 SD) to make the phenotype consistent across cohorts, given that our consortium has measured 25OHD levels in different cohorts by different methods. (5) Finally, we removed outliers beyond 5 SD from step 4.

For comparison purposes, we computed the average 25OHD levels, adjusted for age, sex, BMI, and season of measurement, in one cohort of our meta-analysis (TUK WGS) in carriers and non-carriers of the lead SNV(s).

The software used for each cohort's GWAS is listed in Table S1. We performed single-variant tests for variants with MAF > 0.1% by using an additive effect of the minor allele at each variant in each cohort. The type of software employed for single-variant testing for each cohort is shown in Table S1. Studies with related individuals used software that accounted for relatedness. Cohort-specific genomic inflation factors (lambda values) are also shown in Table S1 (the mean lambda value was 1.015).

We then meta-analyzed association results from all discovery cohorts (n = 42,274). This stage included validation of the results file format, filtering files by the above quality-control (QC) criteria, comparison of trait distributions among different studies, and identification of potential biases (large beta values and/or standard errors, inconsistent effect allele frequencies, and/or extreme lambda values). Meta-analysis QC of the GWAS data included the following SNV-level exclusion criteria: (1) information score < 0.4, (2) Hardy-Weinberg equilibrium (HWE) p value < 10^{-6}, (3) missingness > 0.05, and (4) MAF < 0.5%.

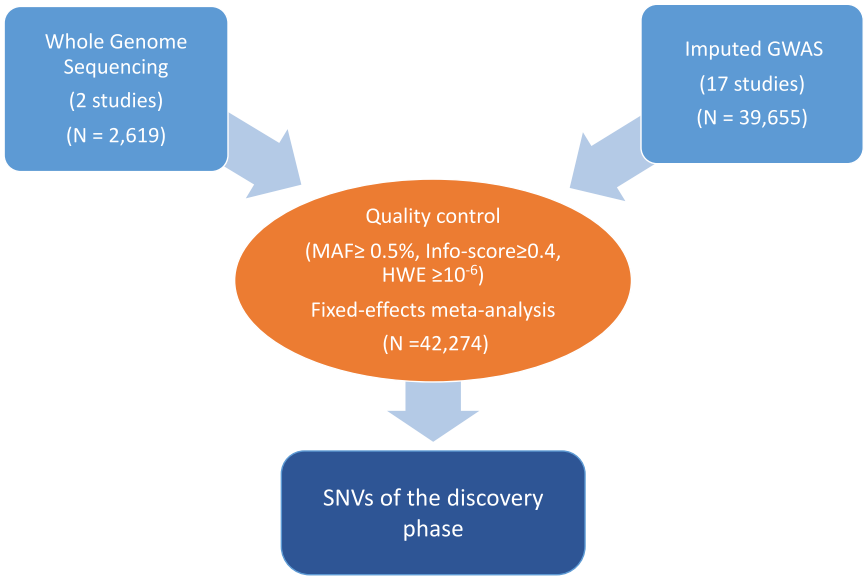


Figure 1. Schematic of the Discovery Single-Variant Meta-analysis

SNV alignment across studies was done with the chromosome and position information for each variant according to genome build hg19 (UCSC Genome Browser). SNVs in the X chromosome were not included in the meta-analysis. Fixed-effects meta-analysis was performed with the software package GWAMA[16] with adjustment for genomic control. We tested bi-allelic SNVs with MAF $\geq$ 0.5% for association and declared genome-wide statistical significance at $p \leq 1.2 \times 10^{-8}$ for variants present in more than one study. This stringent p value threshold was set to adjust for all independent SNVs above the MAF threshold of 0.5%.[17]

Conditional analysis was undertaken for the four previously described lead vitamin D SNVs from the SUNLIGHT consortium with the Genome-wide Complex Trait Analysis (GCTA) package.[18] This method uses an approximate conditional-analysis approach from summary-level statistics from the meta-analysis and inter-SNV linkage-disequilibrium corrections estimated from a reference sample. We used UK10K individuals as the reference sample to calculate the linkage disequilibrium of SNVs. The associated regions flanking within 400 kb of the top SNVs from SUNLIGHT were extracted, and the conditional analyses were conducted within these regions. Conditional analyses of individual variants presented in Tables 2 and S5 were conducted with GCTA v.0.93.9 and default parameters.

We used analyses of haplotype blocks for the candidate variants of interest by deriving phased haplotypes from 1,013 individuals from the TUK WGS cohort with a custom R package.

Effects on Vitamin D Insufficiency

To investigate the effect of genome-wide-significant SNVs on vitamin D insufficiency (defined as 25OHD levels below 50 nmol/L), we used data from four cohorts: TUK imputed, TUK WGS, BPROOF, and MrOS (n = 8,711). We performed logistic regression of this binary phenotype against the SNVs by adjusting for the following covariates: age, sex, BMI, and season of measurement. Meta-analysis of cohort-level summary statistics was performed in R[19] with the epitools[20] and metafor[21] packages.

Interaction Analysis with Vitamin D Intake

We analyzed interactions between our candidate SNV(s) and vitamin D dietary intake (continuous and tertiles) in 9,224 individuals from five of the cohorts (FHS, PIVUS, ULSAM, BPROOF, and RSIII) participating in our discovery phase. A detailed description of the method for capturing vitamin D intake in each of the participating cohorts appears in Table S6. Linear regression was conducted in each of these studies under an additive genetic model. The following variables and co-variables were included in the model: log-transformed serum 25OHD as the dependent variable; SNV genotype (coded as 0, 1, or 2) as an independent variable; SNV (genotype) × dietary vitamin D intake (continuous or tertiles) as an interaction term; and age, sex, BMI, season of 25OHD measurement, dietary vitamin D intake (continuous or tertiles), supplemented vitamin D (yes or no), and total energy intake as covariates. The results from the five studies were meta-analyzed by a fixed-effects model with the metafor tool of the R statistical package.

Effects on Multiple Sclerosis

We tested the effect of the genome-wide-significant SNVs on the risk of multiple sclerosis in 5,927 case and 5,599 control samples by assuming an additive genetic model. Control samples were obtained from the UK Biobank[22] by random selection of participants without multiple sclerosis. Case samples were obtained from the UK Biobank,[22] previously published multiple sclerosis GWASs,[23,24] and newly genotyped UK subjects. Before genotype imputation of the genotyped case samples, we applied numerous QC criteria to ensure unbiased genotype calls between cohorts. These included retaining only SNVs with a MAF > 1% and excluding SNVs or samples with high missingness.[25] Further, samples were assessed for population stratification with EIGENSTRAT,[26,27] and outliers were removed. Genotype data were then imputed by the Sanger Imputation Service[15] with the combined UK10K and 1000 Genomes Phase 3 reference panels,[9,28] the same reference panel used for the UK Biobank control samples. Genotype data were phased with EAGLE2[29] and imputed with PBWT.[30] SNPTEST[31] was used for association testing on the combined case-control dataset, which included testing the additive effect of each allele on multiple sclerosis status and using the top ten principal components from EIGENSTRAT[26,27] to adjust for population stratification and batch effects.

Results

GWAS

After strict QC, the genomic inflation factor for the meta-analysis of 19 GWASs was 0.99, suggesting a lack of bias due to population stratification (Figure 2). Through meta-analysis of 11,026,511 sequenced and imputed variants from our discovery cohorts (Table 1), we identified a signal at the chromosomal locus 11p.15.2, which harbors variants associated with 25OHD levels (lead low-frequency

Table 2. Association Results for Genome-wide-Significant Low-Frequency Variants from Discovery 25OHD Meta-analysis before and after Conditioning on the Lead Common *CYP2R1* SNP, rs10741657, and the Lead Low-Frequency *CYP2R1* Variant, rs117913124

SNV	Chr	Position	EA[a]	EAF[b]	Candidate Gene	Function	Beta[c]	p Value	Conditional on rs10741657		Conditional on rs117913124		n
									Beta[c]	p Value	Beta[c]	p Value	
rs117913124	11	14900931	A	0.025	*CYP2R1*	exon 4 (synonymous codon)	−0.43	1.5×10^{-88}	−0.39	2.4×10^{-78}	NA	NA	41,336
rs116970203	11	14876718	A	0.025	*CYP2R1*[d]	intron 11 variant	−0.43	2.2×10^{-90}	−0.40	3.3×10^{-80}	NA	NA	41,138
rs117361591	11	14861957	T	0.014	*CYP2R1*[d]	intron 11 variant	−0.44	9.1×10^{-51}	−0.40	2.2×10^{-44}	−0.05	0.017	38,286
rs117621176	11	14861320	G	0.014	*CYP2R1*[d]	intron 11 variant	−0.44	8.7×10^{-51}	−0.40	2.1×10^{-44}	−0.05	0.016	38,273
rs142830933	11	14838760	C	0.014	*CYP2R1*[d]	intron 5 variant	−0.44	1.4×10^{-48}	−0.40	1.7×10^{-42}	−0.05	0.03	37,541
rs117672174	11	14746404	T	0.014	*CYP2R1*[d]	intron 1 variant	−0.43	2.8×10^{-45}	−0.39	2.9×10^{-39}	−0.04	0.062	37,209

Abbreviations are as follows: Chr, chromosome; EA, effect allele; EAF, effect allele frequency; NA, not applicable; SNV, single-nucleotide variant.
[a]Effect allele is the 25OHD decreasing allele.
[b]Effect allele frequency.
[c]Beta values represent changes in standard deviations of the standardized log-transformed 25OHD levels.
[d]Nearest gene: *PDE3B*.

SNV g.14900931G>A [p.Asp120Asp] [rs117913124(A)] [GenBank: NC_000011.9]; MAF = 2.5%, allelic effect size = −0.43 SD of the standardized log-transformed 25OHD levels [SD], $p = 1.5 \times 10^{-88}$; Figure 3 and Table 2). The direction of effect was consistent across all discovery cohorts (Table 3 and Figure 3A), and the mean imputation information score for the imputed studies was 0.97. This low-frequency synonymous coding variant is in exon 4 of *CYP2R1* and is ~14 kb from the previously identified common *CYP2R1* variant rs10741657 (r^2 between these two SNVs = 0.03) (Figure 4). To our knowledge, rs117913124 has not previously been associated with any vitamin-D-related traits in humans.

Figure S1 shows a comparison of the average 25OHD levels, adjusted for age, sex, BMI, and season of measurement, in non-carriers and heterozygous carriers of the A allele of rs117913124 in the TUK WGS cohort. The average 25OHD levels, adjusted for age, sex, BMI, and season of measurement were computed in 542 individuals from the TUK WGS cohort, among which 510 were not carriers and 32 were heterozygous carriers of the A allele of rs117913124 (no homozygous carriers were present in this cohort). After removing outliers (adjusted 25OHD levels ±3 SD from the mean), we included in our analysis 449 non-carriers and 30 heterozygous carriers (for a total of 479 individuals). A linear-regression model with the adjusted 25OHD levels as the dependent variable and the dose of the A allele of rs117913124 (numeric factor 1 or 0) as the independent variable demonstrated an 8.3 nmol/L decrease in the adjusted 25OHD levels per A allele. The mean adjusted 25OHD levels were 64.3 nmol/L in non-carriers and 56.0 nmol/L in heterozygous carriers.

Two-way conditional analysis between the *CYP2R1* common (rs10741657) and low-frequency (rs117913124) variants revealed that the two association signals are largely independent. Specifically, after conditioning on rs10741657, rs117913124 remained strongly associated with 25OHD levels ($p_{cond} = 2.4 \times 10^{-78}$); after conditioning on rs11791324, the effect of rs10741657 on 25OHD levels remained significant ($p_{cond} = 4.0 \times 10^{-33}$ versus $p_{pre\text{-}cond} = 8.8 \times 10^{-45}$; Tables 2 and S5). Further, no other low-frequency variant in the region remained significant after conditioning on rs117913124 (Table 2). To further disentangle the role of rs117913124 from that of rs10741657 on 25OHD levels, we undertook a haplotype analysis based on WGS data from 3,781 individuals from the TUK WGS and ALSPAC WGS cohorts. We found that the 25OHD decreasing A allele of rs117913124 was always transmitted in the same haplotype block with the 25OHD decreasing G allele of the common *CYP2R1* variant rs10741657. By using 25OHD data from the TUK WGS cohort, we compared the 25OHD levels among carriers of the various haplotype blocks. We observed lower levels of 25OHD in carriers of the A allele of rs117913124 than in non-carriers, independently of the presence of the effect allele G of the common *CYP2R1* variant (Table 4).

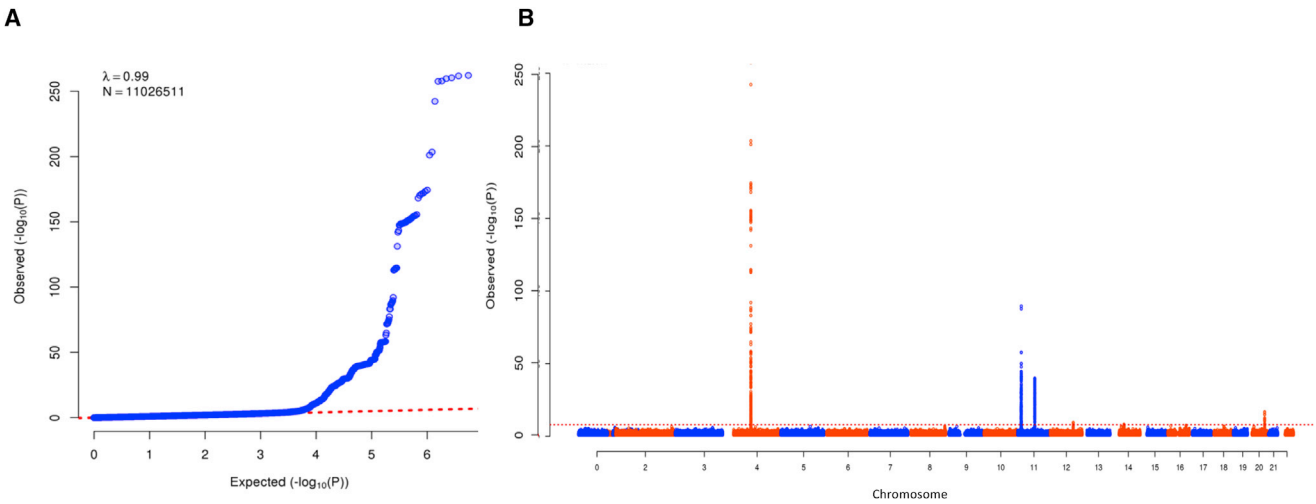


Figure 2. Discovery Single-Variant Meta-analysis
(A) Quantile-quantile plot for the single SNV meta-analysis.
(B) Manhattan plot of the meta-analysis depicts variants with MAF > 0.5% across the 22 autosomes against the $-\log_{10}$ p value from the meta-analysis of 19 cohorts, which included 42,274 individuals.

No other low-frequency or rare variants were identified in the three previously described vitamin-D-related loci at *DHCR7*, *GC*, and *CYP24A1*. The mean effect size of the four previously reported common (MAF ≥ 5%) genome-wide-significant SNVs from the SUNLIGHT consortium was −0.13 SD, and the largest effect size was −0.25 SD (for the *GC* variant) in our meta-analysis (Table S3 and Figure 3B). The effect size of rs10741657(G), the known common *CYP2R1* variant, was −0.09 SD. Hence, the observed effect size of rs117913124 is 3-fold larger than the above mean, 4-fold larger than that of the common *CYP2R1* variant, and almost twice that of the largest previously reported effect of the *GC* variant. Last, the percentage of the 25OHD phenotype variance explained by the low-frequency *CYP2R1* variant (0.9%) was more than double the percentage of the variance explained by the *CYP2R1* common variant (0.4%).

We also identified 18 genome-wide-significant low-frequency and rare SNVs on the same chromosome 11 region as rs117914124 in the neighboring *PDE3B* (MIM: 602047)

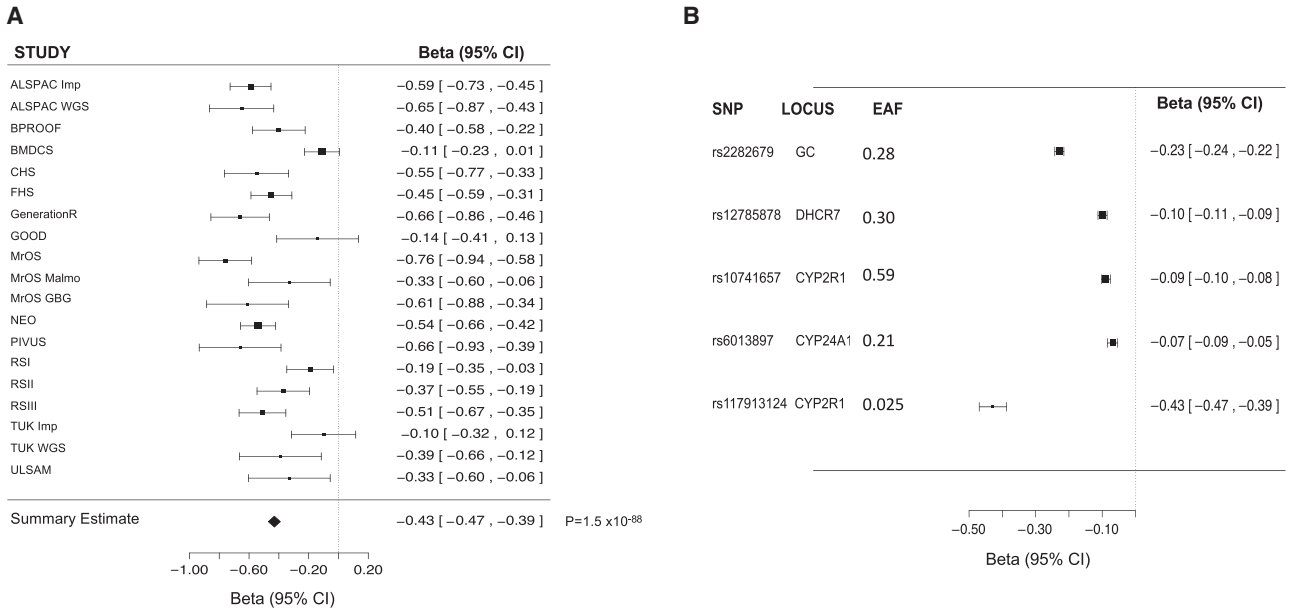


Figure 3. Forest Plot by Cohort for rs117913124 and Forest Plot for rs117913124 and the Previously Described Common 25OHD-Related Variants from Discovery Meta-analysis
(A) Forest plot of estimates from all 19 studies for the low-frequency *CYP2R1* variant rs117913124.
(B) Forest plot of the effect of the four common SUNLIGHT variants and the *CYP2R1* low-frequency variant rs117913124 on log-transformed 25OHD levels.
Squares represent beta values in the 19 studies, and bars around the squares represent 95% confidence intervals (CIs).

Table 3. Summary Statistics for the *CYP2R1* Low-Frequency Variant, rs117913124, from 19 Studies

Study	25OHD Measurement Method	n	EAF (A Allele[a])	Beta[b]	Standard Error	p Value	Information Score
ALSPAC imputed	MS	3,675	0.028	−0.59	0.07	3.43×10^{-18}	0.99
ALSPAC WGS	MS	1,606	0.028	−0.65	0.11	8.23×10^{-10}	NA
BPROOF	MS	2,512	0.027	−0.4	0.09	4.99×10^{-6}	0.97
BMDCS	MS	863	0.019	−0.11	0.06	0.058	0.98
CHS	MS	1,581	0.022	−0.55	0.11	5.15×10^{-7}	0.88
FHS	CLIA	5,402	0.021	−0.45	0.07	2.32×10^{-10}	0.97
GenerationR	MS	1,442	0.033	−0.66	0.1	1.78×10^{-6}	1
GOOD	CLIA	921	0.028	−0.14	0.14	0.31	0.96
MrOS	MS	3,265	0.018	−0.76	0.09	5.63×10^{-16}	0.96
MrOS Malmo	CLIA	893	0.033	−0.33	0.14	0.016	0.94
MrOS GBG	CLIA	945	0.026	−0.61	0.14	7.87×10^{-6}	1
NEO	MS	5,727	0.025	−0.54	0.06	2.73×10^{-19}	1
PIVUS	CLIA	943	0.028	−0.66	0.14	2.56×10^{-6}	0.99
RSI	ECLIA	3,320	0.025	−0.19	0.08	0.019	0.98
RSII	ECLIA	2,022	0.033	−0.37	0.09	2.38×10^{-5}	0.99
RSIII	ECLIA	2,913	0.027	−0.51	0.08	4.61×10^{-10}	0.98
TUK imputed	CLIA	1,919	0.021	−0.1	0.11	0.35	0.98
TUK WGS	CLIA	1,013	0.025	−0.39	0.14	0.006	NA
ULSAM	MS	1,095	0.025	−0.33	0.14	0.02	1

Abbreviations are as follows: CLIA, chemiluminescence immunoassay; EAF, effect allele frequency; ECLIA, electrochemiluminescence immunoassay; MS, mass spectrometry; NA, not applicable; 25OHD, 25-hydroxyvitamin D.
[a]Effect allele is the 25OHD decreasing allele.
[b]Beta values represent changes in standard deviations of the standardized log-transformed 25OHD levels.

(Tables 2 and S4 and Figure 4B). Signals from these SNVs in *PDE3B* were independent of the common variant at *CYP2R1* (Table 2). We then created haplotype blocks with rs117913124 and SNVs at *PDE3B* on the basis of haplotype information from the 3,781 individuals from the TUK WGS and ALSPAC WGS cohorts (Table S2). We found that the 25OHD decreasing allele (A) of rs117913124 was always inherited with the 25OHD decreasing allele (A) of its perfect proxy rs116970203 ($r^2 = 1$). Therefore, rs116970203 is not likely to have a distinct effect from that of rs117913124 on 25OHD levels. On the other hand, the 25OHD decreasing alleles of the remaining four low-frequency variants (all with a MAF of approximately 1.4%) were not always inherited in the same haplotype block as rs117913124 and rs116970203 and were in moderate linkage disequilibrium with rs117913124 (all $r^2 < 0.6$; Figures 4B and 4C). Each of the four alleles was in almost perfect linkage disequilibrium with the remaining three (all $r^2 > 0.96$). This implies that these four SNVs might influence 25OHD levels independently of rs117913124. Nevertheless, as mentioned above, after conditioning on the lead low-frequency *CYP2R1* SNV rs117913124, the p values of the four *PDE3B* SNVs became non-significant and their beta values decreased substantially (Table 2), demonstrating that they probably do not represent an independent signal at the chromosome 11 locus.

rs117913124 and Risk of Vitamin D Insufficiency

To further investigate the clinical significance of the low-frequency *CYP2R1* variant rs117913124, we tested its effect on a binary outcome for vitamin D insufficiency (defined as 25OHD levels < 50 nmol/L) in 8,711 individuals from four studies (TUK WGS, TUK IMP, BPROOF, and MrOS). rs117913124 was strongly associated with an increased risk of vitamin D insufficiency (odds ratio [OR] = 2.20, 95% confidence interval [CI] = 1.8–2.8, p = 1.2×10^{-12}) (Figure 5) after control for relevant covariates as described in the Material and Methods.

Common 25OHD-Associated SNVs

We report two additional loci associated with 25OHD levels (Table 5). Variants leading these associations were common and exerted a rather small effect on 25OHD: (1) a variant in chromosome 12 (rs3819817[C], intronic to *HAL* [MIM: 609457]) with a MAF of 45%, a beta value of 0.04, and a p value of 3.2×10^{-10}; and (2) a variant in chromosome 14 (rs2277458[G], intronic to *GEMIN2* [MIM: 602595]) with a MAF of 21%, a beta value of −0.05, and

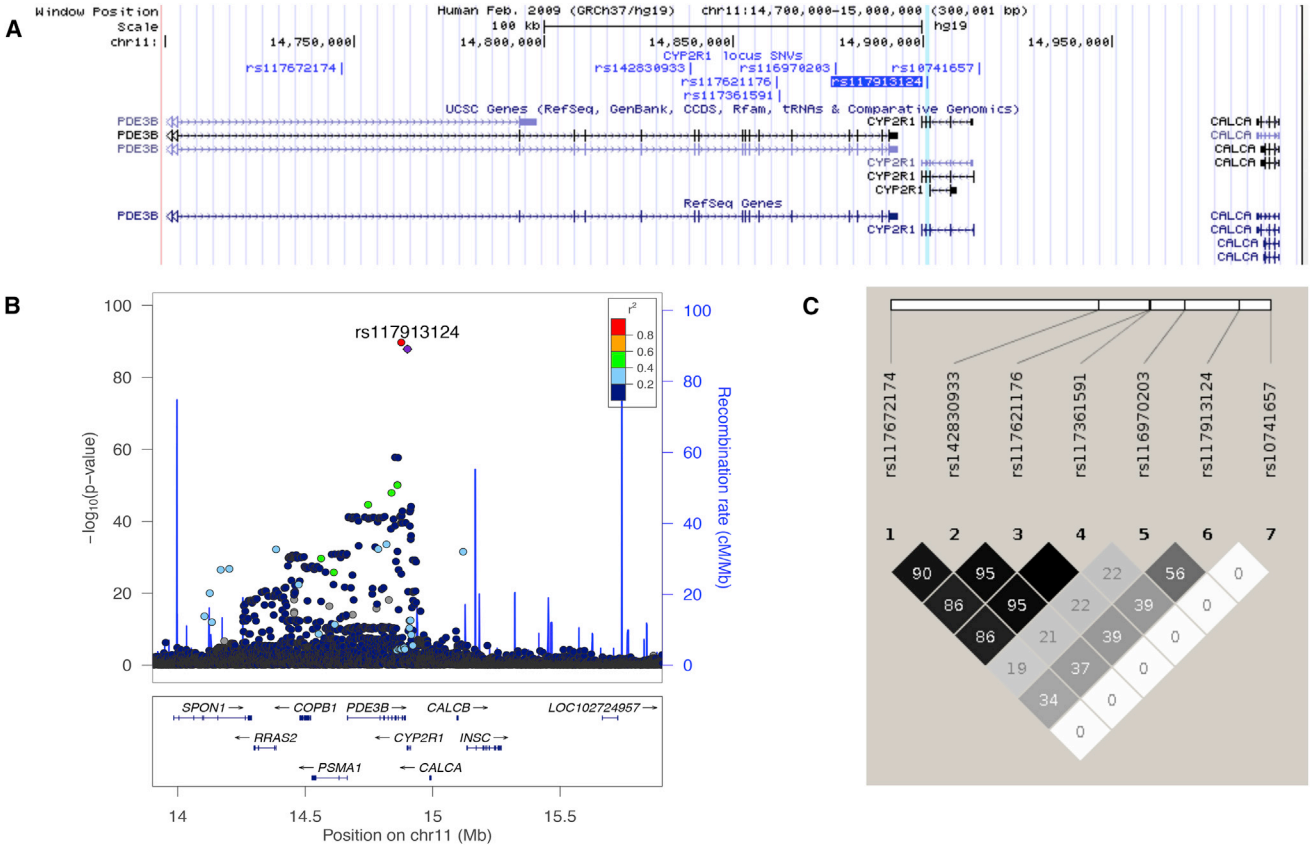


Figure 4. Association Signals from 11p.15.2
(A) UCSC Genome Browser snapshot including the top low-frequency SNVs (see Table 2) and the lead common variant rs10741657 in *CYP2R1*. The position of rs117913124 is highlighted in light blue.
(B) Regional disequilibrium plot showing rs117913124 (purple dot), its perfect proxy rs11670203 (red dot), and the other genome-wide-significant SNVs in the same locus (blue and green dots). The plot depicts SNVs within 1 Mb of a locus's lead SNV (x axis) and their associated meta-analysis p value ($-\log_{10}$) (see Table S10 for more details). SNVs are color coded according to r^2 with the lead SNV (labeled; r^2 was calculated from the UK10K WGS dataset). The recombination rate (blue line), position of genes and their exons, and direction of transcription are also displayed (below plot).
(C) Linkage-disequilibrium plot indicating the r^2 values between the SNVs of Table 2 (top low-frequency variants) and between these low-frequency SNVs and the lead common variant (rs107416570) at the same *CYP2R1* locus (r^2 calculated from the 1000 Genomes dataset).

a p value of 6.0×10^{-9}. Both variants were present in all 19 studies, and the direction of the effect was the same among the 19 studies (Figure 6). Neither the *HAL* nor the *GEMIN2* locus is previously known to be associated with 25OHD levels. Of note, neither variant was present in the HapMap imputation reference used in the SUNLIGHT study.

Interaction Analysis

CYP2R1 encodes the enzyme responsible for 25-hydroxylation of vitamin D in the liver,[32] a necessary step in the conversion of dietary vitamin D and vitamin D oral supplements to the active metabolite, 1,25 dihydroxy-vitamin D. Therefore, we hypothesized that, in contrast with non-carriers, individuals heterozygous or homozygous for rs117913124 in *CYP2R1* would not show a response in their 25OHD levels to vitamin D intake. In other words, we expected carriers of the effect allele of rs117913124 to have steadily lower 25OHD levels, independently of their vitamin D intake. To investigate this hypothesis, we tested the presence of interaction between rs117913124 and vitamin D dietary intake (continuous values and tertiles) on 25OHD levels in 9,224 individuals from five studies (Figure S2). We found no interaction between rs117913124 and dietary vitamin D intake (beta value = −0.0002 and interaction p value = 0.41 for continuous vitamin D intake; beta value = 0.012 and p value = 0.60 for tertiles of vitamin D intake). Given that the two common 25OHD-associated SNVs are located in genes (*HAL* and *GEMIN2*) with no known role in the processing of dietary vitamin D, we found no biological rationale for undertaking a gene-diet interaction analysis for these variants.

25OHD-Assosiated Variants and Risk of Multiple Sclerosis

We tested whether the *CYP2R1* low-frequency variant rs117913124 and the common variants rsrs3819817 and

Table 4. Effect of Different Haplotype Combinations of the Low-Frequency (rs117913124) and Common (rs10741657) *CYP2R1* Variants on 25OHD Levels

Haplotype[a]	Beta[b]	p Value	n
GA GA	−0.02	0.79	156
AG GA	−0.49	0.02	23
AG GG	−0.3	0.13	27
GA GG	0.01	0.87	477
GG GG	0.05	0.58	330

Results are based on individuals from the TUK WGS cohort.
[a]The first allele in each chromatid corresponds to the low-frequency variant rs117913124; the second allele corresponds to the common variant rs10741657. The two AG blocks contain the 25OHD decreasing allele (A) of the low-frequency variant, which is always inherited with the 25OHD decreasing allele (G) of the common variant.
[b]Beta values represent changes in standard deviations of the standardized log-transformed 25OHD levels.

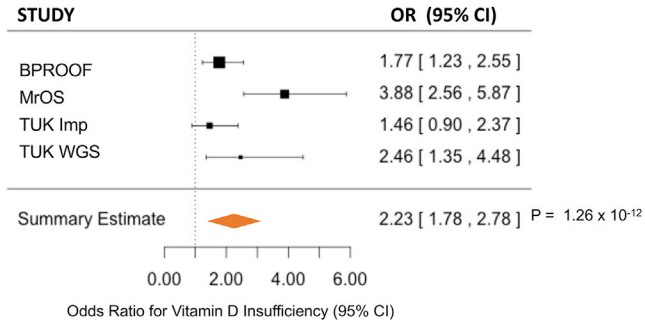


Figure 5. Effect of rs117913124 on Vitamin D Insufficiency
Forest plot of the effect of the low-frequency *CYP2R1* variant rs117913124 on vitamin D insufficiency in four studies. Squares represent odds ratios for vitamin D insufficiency in the four studies, and bars represent 95% CIs.

rs2277458 in *HAL* and *GEMIN2*, respectively, influence the risk of multiple sclerosis. In 5,927 multiple sclerosis samples and 5,599 control samples, we found that the 25OHD decreasing allele at rs117913124[A] was associated with increased odds of multiple sclerosis (OR = 1.40; 95% CI = 1.19–1.64; p value = 2.6×10^{-5}). By way of comparison, the OR of multiple sclerosis for the common *CYP2R1* variant was 1.03 (95% CI = 0.97–1.08; p value = 0.03) in the same study and has previously been reported to be 1.05 (95% CI = 1.02–1.09; p value 0.004) in a separate study.[33] Thus, the effect per allele of rs117913124 on multiple sclerosis was 12.4-fold larger than that attributed to the already known common variant at *CYP2R1*. With regard to the two common SNVs, the 25OHD decreasing allele (T) at the *HAL* variant rs3819817 was not clearly associated with risk of multiple sclerosis; however, there was a trend in the expected direction: OR = 1.05 (95% CI = 1.00–1.11; p value = 0.07). We found no association between the 25OHD decreasing allele (G) at the *GEMIN2* variant rs2277458 and risk of multiple sclerosis: OR = 1.03 (95% CI = 0.96–1.11; p value = 0.34).

Discussion

In the largest GWAS meta-analysis of 25OHD levels in European populations to date, we have identified a low-frequency, synonymous coding genetic variant that has a large effect and strongly associates with 25OHD levels. This variant has an effect size 4-fold larger than that described for the common variant in the same gene (*CYP2R1*) and is associated with a 2-fold increase in risk of vitamin D insufficiency and a 40% increase in the odds of developing multiple sclerosis. The biological plausibility of these findings is supported by the fact that the low-frequency variant is located in *CYP2R1*, encoding the major hepatic 25-hydroxylase for vitamin D.[32] These findings are of clinical relevance given that 5% of the general European population carries this variant in either the homozygous or heterozygous state, and it is associated with a clinically relevant increase in the risk of multiple sclerosis.

Our study was enabled by large imputation reference panels (UK10K-1000 Genomes and HRC) that offer at least 10-fold more European samples than the 1000 Genomes reference panel alone. We did not identify genome-wide-significant variants with a large effect on 25OHD in novel genes in Europeans, although we did find variants with smaller effects in two loci not previously known to be associated with 25OHD. We also identified in a known vitamin-D-related gene low-frequency variants with much larger effects than those of the previously described common variants.

CYP2R1 encodes the enzyme that is responsible for 25-hydroxylation of vitamin D and is one of the two main enzymes responsible for vitamin D hepatic metabolism[32] (Figure 7). Rare mutations in *CYP2R1* have already been described to cause rickets (MIM: 27744).[32,34] Given the important role of *CYP2R1* in the conversion of dietary vitamin D and vitamin D oral supplements to the active form of vitamin D, we hypothesized that carriers of the low-frequency *CYP2R1* variant might respond poorly to vitamin D replacement therapy. We tested this hypothesis by undertaking an interaction analysis between the *CYP2R1* low-frequency variant and dietary vitamin D intake, which showed no clear interaction. However, we note that studies of gene-environment interactions are generally underpowered, measurement error in dietary data is common, and this interaction was further limited by time differences between assessment of dietary intake and measurement of 25OHD levels. Therefore, whether this genetic variant influences 25OHD response to vitamin D administration requires further study.

Although the aim of the present study was to describe variants of low MAF and large effect on 25OHD, we report two common chromosome 12 (*HAL*) and 14 (*GEMIN2*) variants that have a small effect size and reached genome-wide significance in our

Table 5. Main Findings of the GWAS Meta-analysis

SNP	Chr	Candidate Gene	EA	EAF	Beta[a]	p Value	n
rs117913124	11	*CYP2R1*	A	0.025	−0.43	1.5×10^{-88}	41,336
rs3819817	12	*HAL*	C	0.45	0.04	3.2×10^{-10}	41,071
rs2277458	14	*GEMIN2*	G	0.21	−0.05	6.0×10^{-9}	39,746

Abbreviations are as follows: Chr, chromosome; EA, effect allele; EAF, effect allele frequency; SNP, single-nucleotide polymorphism.
[a]Beta values represent changes in standard deviations of the standardized log-transformed 25OHD levels while controlling for age, sex, BMI, and season of measurement.

meta-analysis. Although no existing evidence implicates *GEMIN2* in vitamin-D-related physiological pathways, *HAL* is expressed in the skin and is involved in the formation of urocanic acid, a "natural sunscreen."[35,36] Thus, this could constitute a plausible pathophysiologic mechanism implicating *HAL* in vitamin D synthesis in the skin. Additional functional follow-up of the signals in chromosomes 12 and 14 is needed to characterize the genes and/or mechanisms underlying these associations.

Our findings could have clinical relevance for several reasons. First, individuals carrying at least one copy of the low-frequency *CYP2R1* variant have lower levels of 25OHD by a clinically relevant degree. Specifically, the risk of vitamin D insufficiency is doubled in these individuals. Second, their risk of multiple sclerosis is also increased in accordance with previous evidence supporting a causal role for vitamin D in the risk of multiple sclerosis.[10] Third, these findings affect ~5% of individuals of European descent. Fourth and finally, rs117913124 could be used along with the previously identified common vitamin-D-related variants as an additional genetic predictor of low 25OHD levels in Mendelian randomization studies investigating the causal role of low vitamin D levels in human disease.

Our study also has its limitations. First, although the scope of our study was detection of low-frequency and rare variants, we opted to include in our meta-analysis two WGS studies with a relatively low read depth of 6.7×, as well as three studies imputed to older imputation panels (1000 Genomes and UK10K). These studies have a limited capacity to capture very rare variants, which might explain why we failed to identify such associations. In addition to the limitations arising from the time difference between assessment of dietary vitamin D intake and 25OHD measurements, the analysis of the gene-diet interaction, as mentioned above, might have lacked statistical power. Because our analysis was restricted to populations of European ancestry, we cannot make any assumptions concerning the effect of

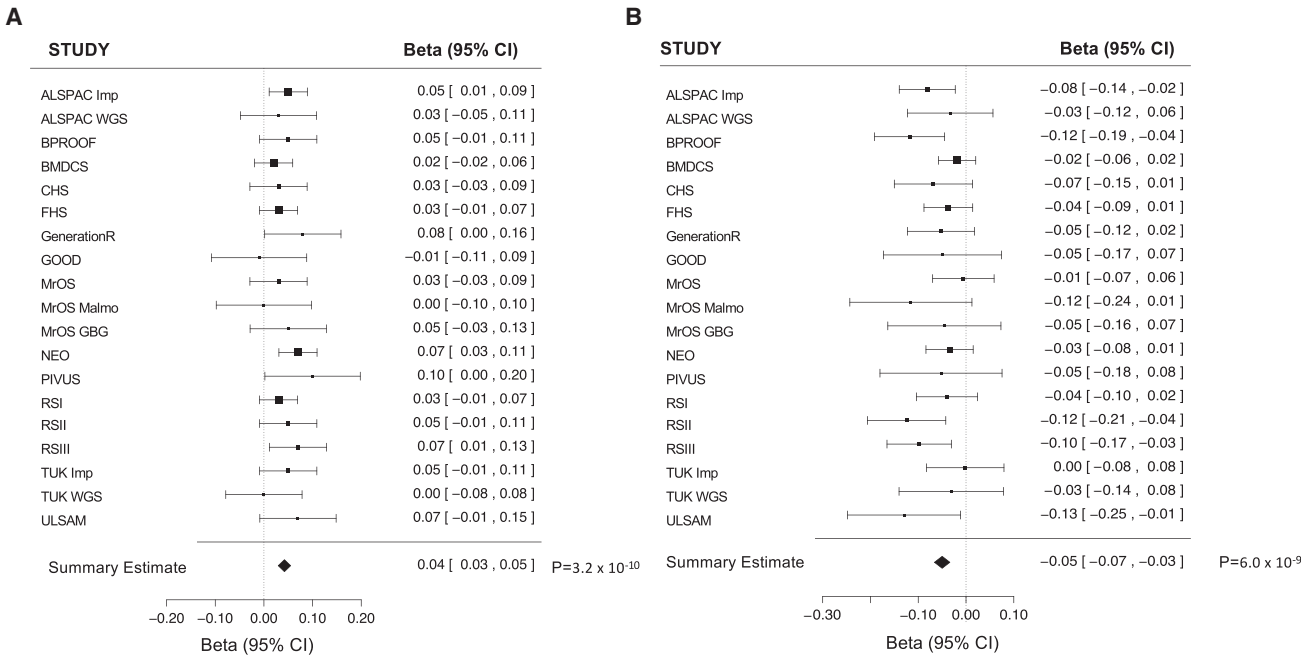


Figure 6. Association Signals from Chromosomes 12 and 14
Forest plots with (A) estimates for the chromosome 12 common variant rs3819817 and (B) estimates for the chromosome 14 common variant rs2277458 from all 19 studies of the meta-analysis where both variants were present. Squares represent beta values in the 19 studies, and bars around the squares represent 95% CIs.

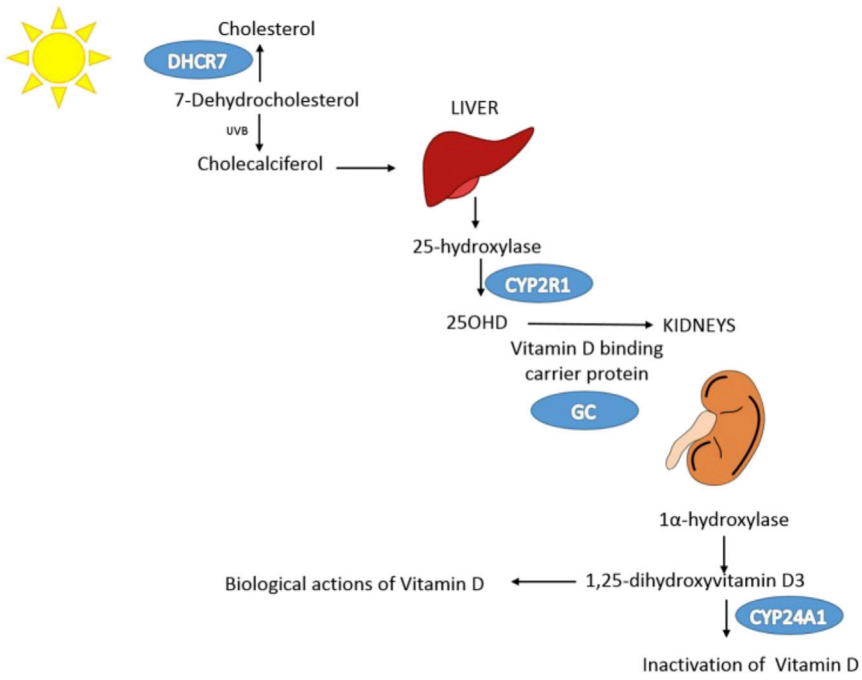


Figure 7. Schematic of the Vitamin D Metabolic Pathway
UVB, ultraviolet B rays.

rs117913124 in non-European populations. Nonetheless, according to the 1000 Genomes reference, this variant is rare in Africans (MAF = 0.3%) and has not been described in East Asians (MAF = 0%). Therefore, describing with any certainty the effect of this variant on 25OHD levels in these populations will require large sample sizes of these populations. Finally, in the absence of functional experiments showing the exact function of rs117913124 in *CYP2R1* and given that this synonymous polymorphism does not affect protein sequence, we cannot unequivocally confirm that this low-frequency variant is causal; however, given that this is a coding variant in a well-documented 25OHD-associated gene, it seems likely that it exerts its effect on *CYP2R1*.

In conclusion, our findings demonstrate the utility of WGS-based discovery and deep imputation for enabling the characterization of genetic associations, offering an improved understanding of the pathophysiology of vitamin D, providing an enriched set of genetic predictors of 25OHD levels for future study, and enabling the identification of groups at increased risk for vitamin D insufficiency and multiple sclerosis.

Accession Numbers

The GWAS summary statistics reported in this paper have been deposited in the Genome-wide Repository of Associations between SNPs and Phenotypes (GRASP).

Supplemental Data

Supplemental Data include 2 figures, 10 tables, and Supplemental Acknowledgments and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.06.014.

Web Resources

GCTA, http://cnsgenomics.com/software/gcta/
GenBank, https://www.ncbi.nlm.nih.gov/genbank/
GRASP: Genome-wide Repository of Associations between SNPs and Phenotypes, https://grasp.nhlbi.nih.gov/Overview.aspx
GWAMA, http://www.geenivaramu.ee/en/tools/gwama
OMIM, http://www.omim.org
UCSC Genome Browser, https://genome.ucsc.edu/
UK10K, http://www.uk10k.org
VQSLOD, http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

Received: February 20, 2017
Accepted: June 29, 2017
Published: July 27, 2017

References

1. Forrest, K.Y., and Stuhldreher, W.L. (2011). Prevalence and correlates of vitamin D deficiency in US adults. Nutr. Res. *31*, 48–54.
2. Rosen, C.J., Adams, J.S., Bikle, D.D., Black, D.M., Demay, M.B., Manson, J.E., Murad, M.H., and Kovacs, C.S. (2012). The nonskeletal effects of vitamin D: an Endocrine Society scientific statement. Endocr. Rev. *33*, 456–492.
3. Shea, M.K., Benjamin, E.J., Dupuis, J., Massaro, J.M., Jacques, P.F., D'Agostino, R.B., Sr., Ordovas, J.M., O'Donnell, C.J., Dawson-Hughes, B., Vasan, R.S., and Booth, S.L. (2009). Genetic and non-genetic correlates of vitamins K and D. Eur. J. Clin. Nutr. *63*, 458–464.
4. Livshits, G., Karasik, D., and Seibel, M.J. (1999). Statistical genetic analysis of plasma levels of vitamin D: familial study. Ann. Hum. Genet. *63*, 429–439.
5. Wang, T.J., Zhang, F., Richards, J.B., Kestenbaum, B., van Meurs, J.B., Berry, D., Kiel, D.P., Streeten, E.A., Ohlsson, C., Koller, D.L., et al. (2010). Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet *376*, 180–188.

6. Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. *47*, 1272–1281.
7. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al.; AOGC Consortium; and UK10K Consortium (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature *526*, 112–117.
8. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.
9. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F., et al.; UK10K Consortium (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. *6*, 8111.
10. Mokry, L.E., Ross, S., Ahmad, O.S., Forgetta, V., Smith, G.D., Goltzman, D., Leong, A., Greenwood, C.M., Thanassoulis, G., and Richards, J.B. (2015). Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. PLoS Med. *12*, e1001866.
11. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
12. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.
13. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
14. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
15. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
16. Mägi, R., and Morris, A.P. (2010). GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics *11*, 288.
17. Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., Greenwood, C.M.; and UK10K Consortium (2014). Estimating genome-wide significance for whole-genome sequencing studies. Genet. Epidemiol. *38*, 281–290.
18. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
19. Team, R.C. (2013). R: A language and environment for statistical computing (R Foundation for Statistical Computing).
20. Aragon, T.J., Wollschlaeger, D., and Omidpanah, A. (2017). epitools: Epidemiology Tools. https://cran.r-project.org/package=epitools.
21. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. J. Stat. Softw. *36*, 1–48.
22. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.
23. Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., Ivinson, A.J., et al.; International Multiple Sclerosis Genetics Consortium (2007). Risk alleles for multiple sclerosis identified by a genomewide study. N. Engl. J. Med. *357*, 851–862.
24. Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat. Genet. *41*, 824–828.
25. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.
26. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
27. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
28. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
29. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448.
30. Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics *30*, 1266–1272.
31. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.
32. Cheng, J.B., Levine, M.A., Bell, N.H., Mangelsdorf, D.J., and Russell, D.W. (2004). Genetic evidence that the human CYP2R1 enzyme is a key vitamin D 25-hydroxylase. Proc. Natl. Acad. Sci. USA *101*, 7711–7715.
33. Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A., et al.; International Multiple Sclerosis Genetics Consortium (IMSGC); Wellcome Trust Case Control Consortium 2 (WTCCC2); and International IBD Genetics Consortium (IIBDGC) (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet. *45*, 1353–1360.
34. Casella, S.J., Reiner, B.J., Chen, T.C., Holick, M.F., and Harrison, H.E. (1994). A possible genetic defect in 25-hydroxylation as a cause of rickets. J. Pediatr. *124*, 929–932.
35. Barresi, C., Stremnitzer, C., Mlitz, V., Kezic, S., Kammeyer, A., Ghannadan, M., Posa-Markaryan, K., Selden, C., Tschachler, E., and Eckhart, L. (2011). Increased sensitivity of histidinemic mice to UVB radiation suggests a crucial role of endogenous urocanic acid in photoprotection. J. Invest. Dermatol. *131*, 188–194.
36. Suchi, M., Sano, H., Mizuno, H., and Wada, Y. (1995). Molecular cloning and structural characterization of the human histidase gene (HAL). Genomics *29*, 98–104.

IJC
International Journal of Cancer

Assessing the causal association between 25-hydroxyvitamin D and the risk of oral and oropharyngeal cancer using Mendelian randomization

Tom Dudding [1,2,3], Mattias Johansson[4], Steven J. Thomas[3], Paul Brennan[4], Richard M. Martin[1,2,5] and Nicholas J. Timpson[1,2,5]

[1] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[2] Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, United Kingdom
[3] Bristol Dental School, University of Bristol, Bristol, United Kingdom
[4] International Agency for Research on Cancer, Section of Genetics–Genetic Epidemiology Group, Lyon, France
[5] National Institute for Health Research (NIHR) Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, United Kingdom

Circulating 25-hydroxyvitamin D (25OHD) is an appealing potential intervention for cancer risk and has been associated with oral and oropharyngeal cancer risk but evidence is inconsistent. The availability of genetic variants, uncorrelated with known confounders, but predictive of 25OHD and genetic data in a large oral and oropharyngeal cancer collaboration aids causal inference when assessing this association. A total of 5,133 oral and oropharyngeal cancer cases and 5,984 controls with genetic data were included in the study. Participants were based in Europe, North America and South America and were part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. Five genetic variants reliably associated with circulating 25OHD were used to create a relative genetic measure of 25OHD. In the absence of measured 25OHD, two-sample Mendelian randomization using individual level outcome data were used to estimate causal odds ratios (OR) for cancer case status per standard deviation increase in log25OHD. Analyses were replicated in an independent population-based cohort (UK Biobank). In the GAME-ON study, there was little evidence of a causal association between circulating 25OHD and oral cancer (OR = 0.86 [0.68;1.09], $p = 0.22$), oropharyngeal cancer (OR = 1.28 [0.72;2.26], $p = 0.40$) or when sites were combined (OR = 1.01 [0.74;1.40], $p = 0.93$). Replication in UK Biobank and pooled estimates produced similar results. Our study suggests that a clinically relevant protective effect of 25OHD on oral and oropharyngeal cancer risk is unlikely and supplementation of the general population with 25OHD is unlikely to be beneficial in preventing these cancers.

Introduction

Each year, there are approximately 300,000 new oral cavity and 230,000 new oropharyngeal cancers worldwide.[1] Squamous cell cancers of the head and neck (HNC) are heterogeneous and oral and oropharyngeal cancers have differing aetiology. Alcohol and tobacco use are the major risk factors for oral and oropharyngeal cancers,[2,3] explaining approximately 65–70% of the population attributable risk.[4] Infection with human papillomavirus (HPV) has emerged as an important risk factor, particularly in oropharyngeal cancers.[5] The disease burden of these cancers remains high despite reductions in the prevalence of tobacco and alcohol use. Oropharyngeal cancer incidence continues to increase in the UK despite no increase in HPV infection in the past 10 years;[6] therefore, identification of other modifiable risk factors remains important.

Key words: oral cancer, oropharyngeal cancer, 25-hydroxyvitamin D, Mendelian randomization

Abbreviations: $1,25OH_2D$: 1,25 di-hydroxyvitamin D; 25OHD: 25-hydroxyvitamin D; CI: confidence interval; CEU ancestry: northern Europeans from Utah; GAME-ON: Genetic Associations and Mechanisms in Oncology; GWAS: Genome Wide Association Study; HNC: head and neck cancer; MAF: minor allele frequency; MR: Mendelian randomization; OR: odds ratio; QC: quality control; RCT: randomized controlled trial; SNP: single nucleotide polymorphism

Additional Supporting Information may be found in the online version of this article.

Conflict of interest: All the authors declare that they have no potential conflicts of interest.

Grant sponsor: Wellcome Trust; **Grant number:** 201237/Z/16/Z

DOI: 10.1002/ijc.31377

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

History: Received 22 Jan 2018; Accepted 2 Mar 2018; Online 14 Mar 2018

Correspondence to: Nicholas J. Timpson, MRC Integrative Epidemiology Unit (IEU), Bristol Medical School, Population Health Sciences, Oakfield House, Oakfield Grove, University of Bristol, Bristol BS8 2BN, United Kingdom, Tel.: +44-117-331-0131, E-mail: n.j.timpson@bristol.ac.uk

What's new?

Evidence for a role of vitamin D in head and neck cancers is contradictory, and randomized controlled trials are not well suited to investigate the potential effect of 25-hydroxyvitamin D (25OHD) supplementation. For the first time, this study used Mendelian randomization with genetic variants associated with 25OHD and uncorrelated with known confounders to derive the causal effect of 25OHD on oral cavity and oropharyngeal cancer risk. The findings suggest that a clinically-relevant protective effect of 25OHD on oral and oropharyngeal cancer risk is unlikely, making supplementation of the general population with 25OHD unlikely to be beneficial in preventing these cancers.

Vitamin D is predominantly synthesised by exposure to ultra-violet light. The inactive metabolite (25-hydroxyvitamin D [25OHD]) is the most commonly clinically measured and supplemented form and is correlated with the active form, 1,25-dihydroxyvitamin D ($1,25OH_2D$). An endocrine feedback mechanism controls activation of 25OHD to $1,25OH_2D$ in the kidneys which is important for calcium and phosphate balance.[7] Evidence from mechanistic studies suggests a protective role of $1,25OH_2D$ on cancer incidence and progression,[8,9] and this may act by the same endocrine mechanism or via paracrine effects at the cancer site. It is not clear whether higher circulating 25OHD is itself protective against cancer or whether other factors such as the ratio of 25OHD to $1,25OH_2D$ at the cancer site add further complexity (Tan V et al., Submitted for publication).

Evidence for a role of vitamin D in HNCs is contradictory. In an observational study of tobacco-related tumours, participants with 50% lower circulating 25OHD were more likely to develop HNC after accounting for available potential confounders (Hazard ratio = 1.44 [95% confidence interval (CI): 1.19, 1.73]).[10] A recent prospective European cohort study also found strong evidence for an inverse association between circulating 25OHD on HNC risk (odds ratio [OR] per doubling of circulating 25OHD concentration = 0.69 [95% CI: 0.56, 0.87], $p = 9\times10^{-4}$).[11] However, Arem *et al.*[12] and Skaaby *et al.*[13] found no convincing evidence of a protective effect of higher 25OHD levels on HNC.

Well-designed randomized studies of 25OHD supplementation are required to identify whether there is a true causal link with cancer incidence. However, randomized controlled trials (RCTs) are not well suited to investigate the effect of 25OHD supplementation on rare diseases. Long follow-up duration, non-adherence to intervention or control regime and an unknown optimal supplementation dose limit such trials[14] meaning they are prohibitively expensive and lack power to detect relatively small but potentially clinically important effects.

In the context of this lack of evidence in the literature, Mendelian randomization (MR) can help assess whether there is a causal association between 25OHD and oral and oropharyngeal cancer risk by utilising genetic data, even in studies where 25OHD itself has not been measured. MR uses genetic variants known to be reliably associated with a risk factor of interest (*e.g.,* 25OHD) to derive estimates of the causal effect of that risk factor on health outcomes (*e.g.,* oral cavity and oropharyngeal cancer risk).[15–17] Other studies have used these methods to provide evidence for causal associations of 25OHD with health outcomes including other cancers[18] and multiple sclerosis.[19]

Our study aims to assess the causal association between vitamin D and oral and oropharyngeal cancer risk and estimate the size of any effect using MR.

Methods

Participants and genotyping

GAME-ON. The study comprised 6,034 HNC cases and 6,585 controls from studies which were part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. Cases and controls from 10 different case–control studies were included, as well as from the European cohort study (EPIC) and cases from a UK case series (Head and Neck 5000 [HN5000]). Details of the studies included have been described previously.[20] Informed consent was obtained for all participants and studies were approved by respective institutional review boards. For our study, cancer cases of interest comprised the following ICD codes: oral cavity (C02.0–C02.9, C03.0–C03.9, C04.0–C04.9 and C05.0–C06.9) and oropharynx (C01.9, C02.4 and C09.0–C10.9). DNA extraction, genotyping, quality control and imputation has been described previously.[20] The study population included participants from Europe, North America and South America. To reduce the effect of heterogeneity across these regions, only participants with >70% CEU ancestry were included in analyses ($n = 11,117$). These consisted of 5,133 cases (oral $n = 2,700$ and oropharyngeal $n = 2,433$) and 5,984 controls (Table 1).

SNP selection and validation

Common genetic variants have been identified in Genome Wide Association Studies (GWAS)[21,22] of 25OHD, only variants that passed a genome wide association threshold ($p < 5 \times 10^{-8}$) and had been replicated were selected. These genetic proxies or instruments are located in or near four 25OHD related genes: Group-specific component (*GC*), cytochrome P450 family 2, subfamily R, polypeptide 1 (*CYP2R1*), 7-dehydrocholesterol reductase (*DHCR7*) and cytochrome P450, family 24, polypeptide 1 (*CYP24A1*). A recent GWAS of 25OHD, a meta-analysis of 19 studies totalling 42,274 individuals with European ancestry, identified a low frequency variant in *CYP2R1* (rs117913124), independent of the common variant in the same gene.[23] Effect estimates for all

Table 1. GAME-ON participant summaries

	All, *n* (%) *N* = 11,117	Case, *n* (%) *N* = 5,133	Control, *n* (%) *N* = 5,984
Site			
Oral		2,700 (52.6)	
Oropharyngeal		2,433 (47.4)	
Age (years)			
≤50	2,217 (19.9)	1,071 (20.9)	1,146 (19.2)
50–<60	3,443 (31.0)	1,696 (33.0)	1,747 (29.2)
60–<70	3,348 (30.1)	1,504 (29.3)	1,844 (30.8)
≥70	2,108 (19.0)	861 (16.8)	1,247 (20.8)
Missing	1 (0.01)	1 (0.02)	0 (0.00)
Sex			
Male	7,680 (69.1)	3,798 (74.0)	3,882 (64.9)
Female	3,437 (30.9)	1,335 (26.0)	2,102 (35.1)
Smoking status			
Never	3,302 (29.7)	1,002 (19.5)	2,300 (38.4)
Previous	3,655 (32.9)	1,590 (31.0)	2,065 (34.5)
Current	3,300 (29.7)	2,019 (39.3)	1,281 (21.4)
Missing	860 (7.74)	522 (10.2)	338 (5.65)
Alcohol use			
Never	1,825 (16.4)	767 (14.9)	1,058 (17.7)
Ever	8,422 (75.8)	4,030 (78.5)	4,392 (73.4)
Missing	870 (7.83)	336 (6.55)	534 (8.92)
Geographic region			
Europe	5,251 (47.2)	2,323 (45.3)	2,928 (48.9)
North America	4,583 (41.2)	2,254 (43.9)	2,329 (38.9)
South America	1,283 (11.5)	556 (10.8)	727 (12.1)

single nucleotide polymorphism (SNP) associations with serum 25OHD concentrations were taken from this GWAS. To assess SNP independence, SNPs on the same chromosome were assessed for linkage disequilibrium (LD) using SNAP (http://archive.broadinstitute.org/mpg/snap/ldsearchpw.php) in Europeans from the 1,000 Genomes (Pilot 1) project. Five SNPs were identified for 25OHD from GWAS (Table 2). rs4588 in *GC*, rs10741657 in *CYP2R1* and rs6013897 in *CYP24A1* have been used in genetic instruments for 25OHD in MR studies before. rs4423214 near *DHCR7* is in perfect LD with rs12785878, identified in the original SUNLIGHT consortium 25OHD GWAS.[22] Finally, rs116970203 near *CYP2R1* is a low frequency variant identified in the most recent GWAS of 25OHD and has been shown to act independently of rs10741657.[23]

In both studies, to correct for within region population structure, region specific principal components were added to the logistic regression models. Effect allele frequencies of the five 25OHD genetic variants were calculated within the three GAME-ON regions and plotted in pie charts superimposed over a world map.

For each individual, the five 25OHD variants were combined into a genetic risk score. To ease clinical interpretation, this score was converted from a per allele scale to the standardized log 25OHD scale by weighting by the beta coefficients derived from GWAS. Therefore, each unit increase in this variable (hereafter referred to as relative 25OHD) represented a standard deviation (SD) increase in log 25OHD.

Assessment of potential confounders

One of the principals of MR is that as alleles are assigned randomly during gamete formation and segregation, the genetic instrument is not associated with factors that typically confound the observational association. To test this assumption, associations between the 25OHD instrument and all available potential confounders in GAME-ON were investigated. Sex, smoking status (never, ever and current) and alcohol use (never and ever) were recorded by questionnaire in both cohorts. GAME-ON sub-study and study country were also considered potential confounders.

Associations of the 25OHD instrument with strata of potential confounders was examined in boxplots of range and means ± 1 SD. Tests of associations were made using linear regression with categorical variables as factor variables, the overall *p* values for the model is shown on the boxplots.

Replication dataset

UK Biobank is a population-based health research resource consisting of approximately 500,000 people, aged between 38 years and 73 years, who were recruited between the years 2006 and 2010 from across the UK.[24] Age on 31 December 2010 (coinciding with the approximate end of baseline data

Table 2. 25OHD genetic variant details

						Vitamin D GWAS				GAME ON	
RSID	Chromosome	Position	Gene	Effect Allele	Other Allele	EAF	Beta	SE	*p* value	Rsq (minimac3)	EAF
rs4588	4	72618323	*GC*	G	T	0.717	0.2469	0.0070	1.68E-263	Genotyped	0.722
rs116970203	11	14876718	*PDE3B*	G	A	0.975	0.4323	0.0209	2.29E-90	0.98	0.978
rs4423214	11	71173254	*DHCR7*	T	C	0.697	0.0998	0.0073	1.39E-40	0.99	0.718
rs10741657	11	14914878	*CYP2R1*	A	G	0.415	0.0938	0.0065	8.76E-45	Genotyped	0.383
rs6013897	20	52742479	*CYP24A1*	T	A	0.791	0.0658	0.0080	9.06E-16	0.77	0.781

Abbeviations: EAF: effect allele frequency; Rsq: R squared; SE: standard error.

collection) was used as the age variable and not age at cancer diagnosis. Prevalent and incident oral and oropharyngeal cancers (ICD10 codes matched to GAME-ON) were identified from linked cancer registry data. To reduce selection bias in allocating controls, all remaining participants, after removing those participants with other HNC, were used as controls. A full description of the study design, participants and quality control (QC) methods have been described in detail previously.[25] UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).[26] In UK Biobank, analyses were restricted to individuals of white British ancestry who self-report as "White British" and who have very similar ancestral backgrounds according to the principal component analysis, as described by Bycroft.[27] The full data release contains the cohort of successfully genotyped samples ($n = 488{,}377$). Any impact of genotyping array was investigated and did not meaningfully affect results (Supporting Information text). Pre-imputation QC, phasing and imputation are described elsewhere.[27] Individuals with sex-mismatch or sex-chromosome aneuploidy ($n = 814$) and related individuals ($n = 79{,}448$) were excluded from the analysis (Supporting Information text). After removing related and non-White British participants, there were 337,108 eligible participants, consisting of 585 cases (oral $n = 294$, oropharyngeal $n = 291$) and 336,523 controls (Supporting Information Table 1).

Statistical analyses

MR analyses. In a method analogous to using a genetic risk score,[28] relative 25OHD was used in logistic regression models to estimate a causal OR for a SD increase in log 25OHD. The standard error of this estimate was corrected to account for the imprecision of the beta coefficients used to generate the relative 25OHD estimates using a bootstrap technique (Supporting Information text).

GAME-ON. MR analyses were performed within each geographic region of the GAME-ON consortium accounting for age, sex and the first 15 population specific principal components. As there were likely to be differences between the causal estimates across regions, a random-effects meta-analysis was used to combine GAME-ON causal estimates using the R package '*meta*'.[29] Heterogeneity between study populations was assessed using I^2.[30] Meta analyses were repeated excluding geographical regions that were outliers to assess for their influence on the overall estimate.

Replication in UK biobank and Meta-analysis. MR analyses were repeated in UK Biobank. GAME-ON and UK Biobank estimates were compared and then combined using meta-analysis. This used a random-effects model and was conducted using the 'meta' package in R.[29] Heterogeneity between studies was assessed using I^2.[30]

Sensitivity analyses. To assess whether any single SNP was driving a causal estimate, which could be driven by events such as horizontal pleiotropy which invalidate the MR assumptions, a leave-one out method was applied. This repeats the analysis sequentially removing one SNP from the genetic instrument used to derive relative 25OHD. To further demonstrate this, where a SNP overtly influences an association, each possible genetic instrument from combinations of the five 25OHD SNPs was used to estimate relative 25OHD and thereafter a causal OR. These OR were plotted as histograms highlighting those that include the highly influential SNP to see whether there was an obvious grouping of these estimates.

Power. The 25OHD genetic variants have been reported to explain 3–5% of the variance in 25OHD.[21,22,31] Given this, in the GAME-ON study for oral ($n = 2{,}700$ cases, 5984 controls) or oropharyngeal sites ($n = 2{,}433$ cases, 5984 controls) alone, there was adequate power to detect an OR of 0.67 to 0.74 per SD increase in log 25OHD with power of 0.8 and an alpha of 0.05. For all sites combined ($n = 5{,}133$ cases, 5984 controls), the study was powered to detect an OR of between 0.74 and 0.79 (http://cnsgenomics.com/shiny/mRnd/).[32]

Results

Of the five SNPs included in the analyses, rs4588 and rs10741657 were directly genotyped in both the GAME-ON and UK Biobank studies. All other SNPs were imputed and the imputation quality metric (Minimac3 r^2 for GAME-ON; SNPTEST info score for UK Biobank) was >0.98, except rs6013897 in the GAME-ON study which had an r^2 of 0.77 (Table 2; Supporting Information Table 2). No two SNPs on the same chromosome had an LD r^2 >0.01. Allele frequencies were similar across all three regions (Supporting Information Fig. 1).

In the GAME-ON study, there was no strong evidence of association of the 25OHD genetic instrument with sex, smoking status or drinking status after taking multiple testing into account. There was evidence for a difference in the 25OHD genetic instrument across regions and sub-studies (Fig. 1). The region showing deviation from the others was South America and any differences across confounders were no longer seen when the South American region was excluded from analyses (Supporting Information Fig. 2*a*). Within the South American region alone, there was no strong evidence for associations between the 25OHD genetic instrument and confounders (Supporting Information Fig. 2*a*). In UK Biobank, there was no evidence for a difference in relative 25OHD across strata of sex, smoking status or drinking status (Supporting Information Fig. 3).

In GAME-ON, for oral and oropharyngeal cancer, there was little evidence for a causal link with relative 25OHD when the three geographic regions were meta-analysed OR = 0.86 [0.68, 1.09], $p = 0.22$ and OR = 1.28 [0.72; 2.26], $p = 0.40$, respectively. For oral sites, estimates were homogenous (I^2=0% [0, 87], p_{het}=0.44) but due to only having three estimates to meta-analyse (one for each geographic region), power to detect heterogeneity between regions was low.

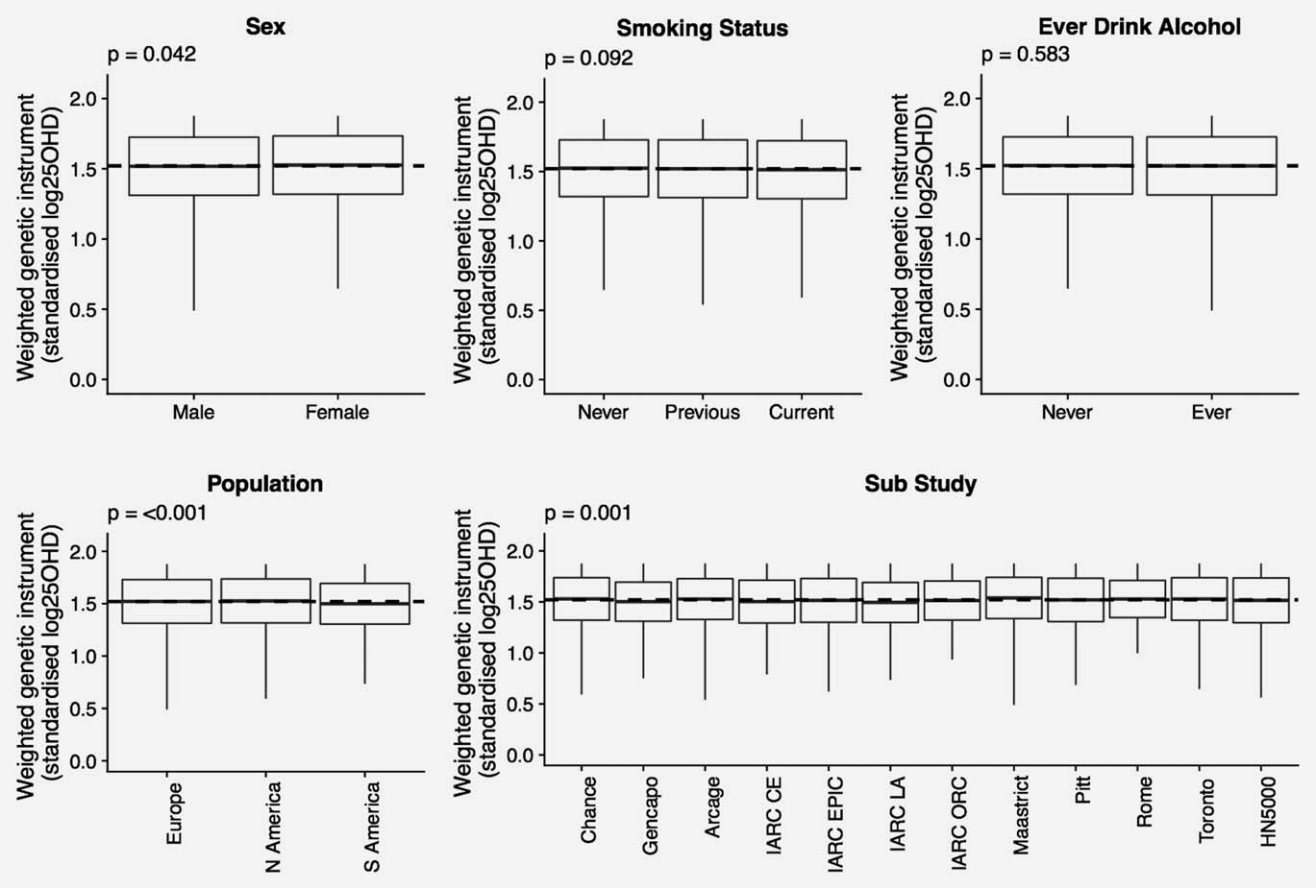


Figure 1. Mean (± SD) and range of 25OHD weighted genetic instrument across strata of potential confounders for GAME-ON. Overall mean 25OHD weighted genetic instrument for those included in plot indicated by dashed line. Abbreviation: *p*: *p* values for linear model.

Despite this, there was evidence of heterogeneity between regions for the oropharyngeal sites (I^2=74% [16, 92], p_{het}=0.02). This was also reflected in all sites (I^2=55% [0, 87], p_{het}=0.10) and was most likely driven by the strong risk increasing association with oropharyngeal cancer identified in the South American region (OR = 3.91 [1.64; 9.29], $p < 0.01$) compared to null associations in both European (OR = 0.98 [0.67; 1.44], $p = 0.94$) and North American regions (OR = 0.92 [0.64; 1.34], *p*0.68) (Fig. 2).

When replicated in UK Biobank, estimates for oral cancer (OR = 0.86 [0.50; 1.51], $p = 0.61$) were very similar with those from the GAME-ON study. For oropharyngeal sites (OR = 0.85 [0.49; 1.47], $p = 0.56$), point estimates differed between GAME-ON and UK Biobank although CIs overlapped. When the two studies were meta-analysed using random-effects, there was no convincing evidence of a causal association with either cancer or when all sites were analysed (Supporting Information Fig. 4).

Sensitivity analyses

In leave one out analyses, sequentially omitting each of the five SNPs provided similar causal estimates for any site within the European and North American regions of GAME-ON. In the South American region of the GAME-ON study, the risk increasing association between 25OHD and oropharyngeal cancer was most strongly influenced by the rs4588 variant (Supporting Information Fig. 5). In analyses using an instrument with all possible combinations of the 25OHD SNPs, even with the rs4588 variant removed there was still a risk increasing association for oropharyngeal cancer in the South American region (Supporting Information Fig. 6). Removing the South American region from the oropharyngeal analyses reduced heterogeneity between the GAME-ON studies (OR = 0.95 [0.73, 1.25], $p = 0.73$; $I^2 = 0$, p_{het}=0.82) (Supporting Information Fig. 7).

Discussion

Our study set out to investigate the causal association between 25OHD and oral and oropharyngeal cancer risk. Given the lack of 25OHD measures in large scale oral and oropharyngeal cancer studies, a two-sample MR approach,[33] that allowed genetic variants to be used as largely unconfounded measures of 25OHD, was used. No strong evidence of a causal association was identified with oral or oropharyngeal cancer. Given power calculations predict an OR smaller than 0.74 could be detected, if present any potential true

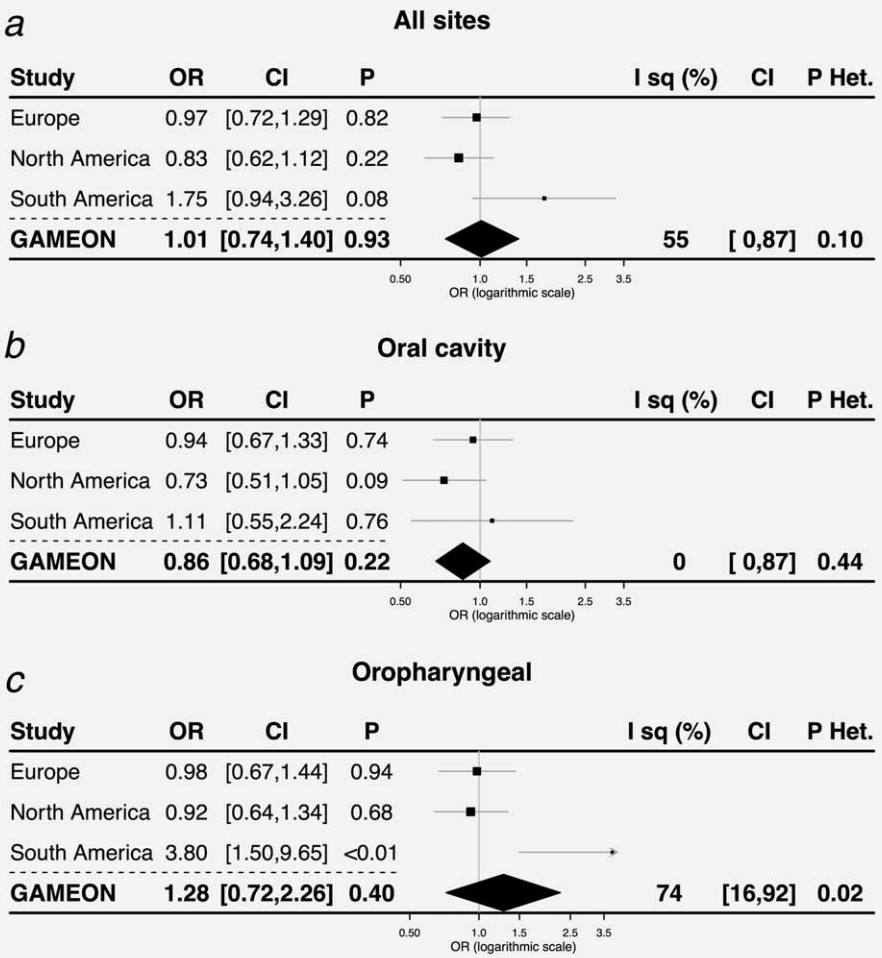


Figure 2. OR (95% CI) for developing cancer for a SD increase in 25OHD for all sites (*a*), oral sites (*b*) and oropharyngeal sites (*c*) in Mendelian randomization analysis. Abbreviations: *p*: *p* values; *p* Het: *p* values for heterogeneity.

effect is likely to be smaller than this and thus threaten the clinical relevance of the association.

These findings mirror a recently published MR study of colorectal, breast, prostate, ovarian, lung and pancreatic cancer, and neuroblastoma that did not support a causal role of 25OHD in the risk of these cancers.[34] Similar to the inference from this work on oral and pharyngeal cancer, the authors of the previous study state that although they are unable to rule out clinically relevant effects of small magnitude, their study combined with previous literature, 'provide evidence that population-wide screening for vitamin D deficiency and subsequent widespread vitamin D supplementation should not currently be recommended as a strategy for primary cancer prevention'. Recent RCTs examining the effect of vitamin D on cancer incidence do not show strong evidence of a protective effect, although these studies are limited to older females.[35,36] Contrary to this, Ong *et al.*[18] did demonstrate a protective role for 25OHD in ovarian cancer in a well powered MR study. Despite the body of evidence that does not support a protective effect of 25OHD on cancer incidence, there is evidence to support the hypothesis that vitamin D has an effect on cancer progression.[8,11]

The genetic instrument used in our study specifically proxies average total (free and bound) circulating 25OHD and does not necessarily predict the concentrations of free 25OHD available at the tissues or concentrations of free or bound $1,25OH_2D$ in circulation or at the tissue level. Potential anti-cancer effects of vitamin D are via free $1,25OH_2D$ interacting with the vitamin D receptor within tissues resulting in reduced angiogenesis, metastasis, cell invasion, inflammation, and proliferation as well as upregulation of apoptosis.[8] Levels of $1,25OH_2D$ are assumed to be correlated with circulating 25OHD, but this is not necessarily the case. Although less likely than a null association between 25OHD and oral and oropharyngeal cancer risk, the instrument being invalid for 25OHD could explain the disparity between the results of our study and previous findings showing a protective effect of 25OHD. However, as the genetic variants used here as instruments for 25OHD have robust associations with 25OHD and are in genes known to

effect this metabolite they are likely to be valid for the purposes of MR.

To be valid, the genetic instrument must proxy circulating 25OHD without affecting cancer risk through other casual pathways (violation of this assumption is referred to as horizontal pleiotropy). The SNPs used in this instrument are in genes with known effects on vitamin D pathways and have been consistently associated with circulating levels of 25OHD in GWA[21,22] and MR studies,[37] reducing the chance of horizontal pleiotropy. Sensitivity analyses such as MR Egger[38] and the weighted median method[39] can detect or correct for the presence of directional pleiotropy but have low power with few SNPs comprising the genetic instrument, as in this case. Here, causal estimates in the European and North American populations are largely consistent across all five SNPs, providing some evidence against the presence of strong pleiotropy that could bias findings.

In the South American population, a risk increasing association was identified between 25OHD and oropharyngeal cancer. Use of individual level data in our study allowed scrutiny of the genetic instrument-confounder independence assumption and showed that the 25OHD instrument generally held up to this assumption; however, there was some evidence the genetic instrument was associated with region and sub-study, driven by the South American region. Given the above points, it is unlikely then that pleiotropy or confounding can explain the large risk increasing association seen for oropharyngeal cancer in the South American region. However, as this risk increasing association is inconsistent with observational effect directions and MR estimates from other geographic regions, it is unlikely to be truly causal: it most likely represents structure within this population that is by chance associated with oropharyngeal cancer risk. To assess the impact of this potentially biased estimate on the overall causal OR, the analyses were repeated with the South American individuals removed, resulting in more consistent estimates.

Conclusions

Our study does not support the observational association between 25OHD and oral cancer risk and is consistent with evidence that a causal, clinically relevant protective effect of 25OHD on oropharyngeal cancer risk is unlikely. The effect of 25OHD on oral and oropharyngeal cancer progression was not assessed here and requires further investigation.

The genetic instrument for 25OHD is a good proxy for circulating total 25OHD, the same metabolite that is used in supplementation and is measured in the observational studies. Despite being robustly associated with 25OHD, the genetic variants used here only explain a relatively small amount of variance in 25OHD. This means that causal effects of small magnitude cannot be ruled out. Furthermore, it cannot be determined whether 25OHD has an effect on cancer risk in individuals below a certain 25OHD concentration threshold. Where inference is likely to be useful clinically is in relation to supplementation. Supplementation increases total circulating 25OHD and is presumed to subsequently increase free $1,25OH_2D$ availability at the tissues. The SD 25OHD increase in our study will have a relative effect on total 25OHD across the whole life course. If this long-term difference in relative 25OHD does not produce a detectable effect in our study, any 25OHD causal effect is unlikely to be of a magnitude that would warrant long-term supplementation in the general population from a particular age.

Acknowledgements

Quality Control filtering of the UK Biobank data were conducted by R.M.M, G. Hemani, T.D. and L. Paternoster as described in the published protocol (doi: doi.org/10.5523/bris.3074krb6t2frj29yh2b03x3wxj). This research is supported by the National Institute for Health Research (NIHR) Bristol Biomedical Research Centre which is funded by the National Institute for Health Research and is a partnership between University Hospitals Bristol NHS Trust and the University of Bristol; by the MRC/University of Bristol Integrative Epidemiology Unit (IEU) supported by the MRC and the University of Bristol (MC_UU_12013/1-9). RMM, NJT, PB are supported by a Cancer Research UK Programme Grant, the Integrative Cancer Epidemiology Programme (C18281/A19169). TD is supported by the Wellcome Trust (201237/Z/16/Z). NJT is a Wellcome Trust Investigator (202802/Z/16/Z), is a programme lead in the MRC Integrative Epidemiology Unit (MC_UU_12013/3) and works within the University of Bristol NIHR Biomedical Research Centre (BRC).

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136:E359–86.
2. Maier H, Dietz A, Gewelke U, et al. Tobacco and alcohol and the risk of head and neck cancer. *Clin Investig* 1992;70–70:320–7:
3. Freedman ND, Schatzkin A, Leitzmann MF, et al. Alcohol and head and neck cancer risk in a prospective study. *Br J Cancer* 2007;96:1469–74.
4. Hashibe M, Brennan P, Chuang S-C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Cancer Epidemiol Biomarkers Prev* 2009;18:541–50.
5. Chaturvedi AK, Engels EA, Pfeiffer RM, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *JCO.* 2011; 29:4294–301.
6. Schache AG, Powell NG, Cuschieri KS, et al. HPV-related oropharynx cancer in the United Kingdom: an evolution in the understanding of disease etiology. *Cancer Res* 2016;76:6598–606.
7. Holick MF. Vitamin D and bone health. *J Nutr* 1996;126:1159S–64S.
8. Feldman D, Krishnan AV, Swami S, et al. The role of vitamin D in reducing cancer risk and progression. *Nat Rev Cancer* 2014;14:342–57.
9. Deeb KK, Trump DL, Johnson CS. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. *Nat Rev Cancer* 2007;7:684–700.
10. Afzal S, Bojesen SE, Nordestgaard BG. Low plasma 25-hydroxyvitamin D and risk of tobacco-related cancer. *Clin Chem* 2013;59:771
11. Fanidi A, Muller DC, Midttun Ø, et al. Circulating vitamin D in relation to cancer incidence and survival of the head and neck and oesophagus in the EPIC cohort. *Sci Rep* 2016;6:36017
12. Arem H, Weinstein SJ, Horst RL, et al. Serum 25-hydroxyvitamin D and risk of oropharynx and larynx cancers in finnish men. *Cancer Epidemiol Biomarkers Prev* 2011;20:1178–84.

13. Skaaby T, Husemoen LLN, Thuesen BH, et al. Prospective population-based study of the association between serum 25-hydroxyvitamin-D levels and the incidence of specific types of cancer. *Cancer Epidemiol Biomarkers Prev* 2014; 23:1220–9.
14. IARC. Vitamin D and Cancer. IARC, Lyon France, 2008.
15. Lawlor DA, Harbord RM, Sterne JAC, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statist Med* 2008;27:1133–63.
16. Davey Smith G, Ebrahim S. Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32:1–22.
17. Hingorani A, Humphries S. Nature's randomised trials. *Lancet* 2005;366:1906–8.
18. Ong J-S, Cuellar-Partida G, Lu Y, t al. Association of vitamin D levels and risk of ovarian cancer: a Mendelian randomization study. *Int J Epidemiol.* 2016;45:1619–30.
19. Mokry LE, Ross S, Ahmad OS, et al. Vitamin D and risk of multiple sclerosis: a Mendelian Randomization Study. *PLoS Med* 2015;12:e1001866
20. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* 2016;48:1544–50.
21. Ahn J, Yu K, Stolzenberg-Solomon R, et al. Genome-wide association study of circulating vitamin D levels. *Hum Mol Genet* 2010;19: 2739–45.
22. Wang TJ, Zhang F, Richards JB, et al. Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* 2010;376:180–8.
23. Manousaki D, Dudding T, Haworth S, et al. Low-frequency synonymous coding variation in CYP2R1 has large effects on vitamin D levels and risk of multiple sclerosis. *Am J Hum Genet* 2017;29:98–104.
24. Allen NE, Sudlow C, Peakman T, et al. UK biobank data: come and get it. *Sci Transl Med* 2014; 6:224ed4
25. Collins R. What makes UK Biobank special? *Lancet* 2012;379:1173–4.
26. Mitchell R, Hemani G, Dudding T, et al. UK biobank genetic data: MRC-IEU quality control, Version 1. University of Bristol, Bristol, UK, 2017.
27. Bycroft C, Freeman C, Petkova D, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 2017;166298
28. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol* 2013;42:1134–44.
29. Schwarzer G. meta: an R package for meta-analysis. *R News* 2007;7:40–5.
30. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ Br Med J* 2003;327:557–60.
31. Hiraki LT, Major JM, Chen C, et al. Exploring the genetic architecture of circulating 25-hydroxyvitamin D. *Genet Epidemiol* 2013;37:92–8.
32. Brion MJA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* 2013;42: 1497–501.
33. Burgess S, Scott RA, Timpson NJ, EPIC InterAct Consortium E-I, et al. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* 2015;30:543–52.
34. Dimitrakopoulou VI, Tsilidis KK, Haycock PC, et al. Circulating vitamin D concentration and risk of seven cancers: Mendelian randomisation study. *BMJ* 2017;359:j4761
35. Hill KB, Chadwick BL, Freeman R, et al. Adult Dental Health Survey 2009: relationships between dental attendance patterns, oral health behaviour and the current barriers to dental care. *Br Dent J* 2013;214:25–32.
36. Lappe J, Watson P, Travers-Gustafson D, et al. Effect of vitamin D and calcium supplementation on cancer incidence in older women. *JAMA* 2017; 317:1234
37. Vimaleswaran KS, Berry DJ, Lu C, et al. Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts. *PLoS Med* 2013;10: e1001383.
38. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;44:512–25.
39. Bowden J, Davey Smith G, Haycock PC, et al. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* 2016;40:304–14.