Author:
Langdon, Ryan J

Title:
Genetic and epigenetic data as a tool to augment understanding of oropharyngeal cancer

Genetic and epigenetic data as a tool to augment understanding of oropharyngeal cancer

Ryan Langdon

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Bristol Medical School

MRC Integrative Epidemiology Unit

Department of Population Health Sciences

Bristol Medical School

University of Bristol

UK

November 2019

Words: 65,265

Abstract

Genetic and epigenetic data provide the opportunity to robustly appraise the causal effect of an exposure on an outcome of interest, improve understanding of risk and prognostic pathways, and predict the status of a risk factor, prognostic factor, or outcome. In the case of oropharyngeal cancer (OPC), genetic and epigenetic data have rarely been applied in these contexts.

Using large population-based OPC cohorts alongside bioinformatic, genetic and epigenetic resources, I have applied a series of methodologies which aim to improve understanding of the causal risk factor pathways associated with this disease, beyond the limited degree of inference afforded by conventional observational studies. To this end, throughout this thesis I have employed: enriched literature object mining, genome-wide association studies (GWAS), epigenome-wide association studies (EWAS), two-sample Mendelian randomization (MR), MR-phenome-wide association studies (MR-PheWAS), two-step MR and epigenetic prediction scores.

The OPC cohorts forming the core data resources in this thesis are the Head and Neck 5000 study (HN5000) and the head and neck cancer OncoArray study: HN5000 contains genetic, epigenetic and mortality data for 448 individuals with oropharyngeal cancer, whilst OncoArray contains genetic data on 2,641 cases and 6,585 controls.

Methods applied in this thesis have provided evidence for enrichment in literature of 4 risk factors for OPC, the association of 16 phenotypes with OPC incidence, novel whole-blood-based CpG sites associated with HPV, alcohol, smoking and oropharyngeal cancer survival, evidence for a causal effect of smoking-related methylation at the *SPEG* gene with OPC survival, and finally, evidence for the value of blood-based methylation signatures in predicting mortality in those with OPC. These findings highlight the merits of using genetic and epigenetic data to improve on conventional observational analyses, with the caveat that replication and triangulation of these findings from a range of methodological approaches is the optimal route to ensure their robustness and validity.

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:……

Contents

List of figures

List of tables

Research output

2020

1. Epigenetic prediction of complex traits and mortality in a cohort of individuals with oropharyngeal cancer. **Langdon, R. J.**, Beynon, R. A., Ingarfield, K., Marioni, R. E., MacCartney, D. M., Martin, R. M., Ness, A. R., Pawlita, M., Waterboer, T., Relton, C., Thomas, S. J. & Richmond, R. C., 22 Apr 2020, In : Clinical Epigenetics. 12, 14 p., 58 (2020).

Contribution: I generated and applied epigenetic scores for alcohol consumption, body mass index, educational attainment and smoking using MethylationEPIC data from the HN5000 clinical cohort. I then determined the predictive value of these scores (via Area Under the ROC curve) for 4-year excess mortality. In addition, I calculated the amount of phenotypic variance these scores explained in their respective phenotypes. All of the above are the focus of Chapter 6 of my thesis (Epigenetic prediction of complex traits in individuals with oropharyngeal cancer).

I wrote this article as a joint-first author with Rhona Beynon, seen in the reference above. I have used tables (Tables 6.2, 6.3, 6.5, 6.6, 6.7 and 6.8), and figures (Figures 6.2a and 6.2b) in Chapter 6 of my thesis which were created by Rhona as part of our shared paper. The nature of shared first-authorship meant that we were in active discussion and collaboration when refining methods, generating results and creating these tables and figures. Tables 6.2, 6.3 and 6.7 are descriptive tables of baseline measures in the Head and Neck 5000 data I used. The variables included in all of these were agreed between myself and Rhona. Figures 6.2a and 6.2b are used to provide a visual reference for how different phenotypes affect survival in my chapter results. Tables 6.5, 6.6 and 6.8 are results from Cox regression models where I supplied epigenetic scores, but regressions and tables were completed by Rhona. We jointly agreed on covariates to be used in these regressions. Tables 6.5, 6.6 and 6.8 provide hazard estimates of both directly-measured phenotypes (top 4 rows) and epigenetic scores (rest of each table, under “DNAm scores”) against mortality after 4 years. Whilst the above tables and figures are included and referred to, their absence would not fundamentally alter the narrative of Chapter 6 of my thesis; they are included purely to supplement my discussion of how well the epigenetic scores I have uniquely derived predict excess mortality in individuals with oropharyngeal cancer.

2019

2. Validation and characterization of a DNA methylation alcohol biomarker across the life course. Yousefi, P. D., Richmond, R. C., **Langdon, R**. J., Ness, A. R., Liu, C., Levy, D., Relton, C. L., Suderman, M. J. & Zuccolo, L. 23 Sep 2019, (Accepted/In press) In : Clinical Epigenetics.

Contribution: I generated an epigenetic predictor of alcohol consumption using MethylationEPIC data from the HN5000 clinical cohort study and results from a large EWAS of alcohol consumption by Liu et al. This predictor was compared against a healthy population as a biomarker for improved clinical or epidemiologic assessment of alcohol-related ill health.

3. Identifying epigenetic biomarkers of established prognostic factors and survival in a clinical cohort of individuals with oropharyngeal cancer
   Langdon, R., Richmond, R., Elliott, H. R., Dudding, T., Kazmi, N., Penfold, C., Ingarfield, K., Ho, K., Bretherick, A., Haley, C., Zeng, Y., Walker, R. M., Pawlita, M., Waterboer, T., Ring, S., Gaunt, T., Davey-Smith, G., Suderman, M., Thomas, S., Ness, A., Relton, C., 23 April 2020 (Under review) In: Clinical Epigenetics

Contribution: I wrote this paper, conducted all analyses (single-site EWAS, DMR analysis and MR) and derived a novel method of instrumenting differentially-methylated regions in an MR framework. The work featured in this paper forms a large portion of Chapter 5 in this thesis - DNA methylation as a mediator of OPC prognostic factors and mortality.

4. A phenome-wide Mendelian randomization study of pancreatic cancer using summary genetic data
   Langdon, R., Richmond, R., Hemani, G., Zheng, J., Wade, K., Carreras-Torres, R., Johansson, M., Brennan, P., Wootton, R., Munafo, M., Davey Smith, G., Relton, C., Vincent, E., Martin, R. & Haycock, P., 17 Jul 2019, In : Cancer Epidemiology, Biomarkers and Prevention.

Contribution: I wrote this paper and conducted all analyses within. This paper employs an MR-PheWAS framework, which I have used to investigate OPC genetic data and generate the results seen in Chapter 4 of my thesis (A phenome-wide mendelian randomization study of oropharyngeal cancer using summary genetic data). A peer-reviewed paper proved invaluable in determining statistical thresholds and sensitivity analyses, in addition to understanding how to accurately interpret results.

5. Treatment preference and recruitment to pediatric RCTs: A systematic review
   Beasant, L., Brigden, A., Parslow, R., Apperley, H., Keep, T., Northam, A., Wray, C., King, H., **Langdon, R.**, Mills, N., Young, B. & Crawley, E., 1 Jun 2019, In : Contemporary Clinical Trials Communications. 14, 100335.

Contribution: I was a second reviewer for this systematic review, which provided insight into the methodology, advantages and disadvantages of this style of paper. This insight proved useful when introducing literature mining and setting it in its wider context

2018

6. Mendelian randomization does not support serum calcium in prostate cancer risk
   Yarmolinsky, J., Berryman, K., **Langdon, R.**, Bonilla, C., the PRACTICAL Consortium, Davey Smith, G., Martin, R. & Lewis, S., 10 Oct 2018, In : Cancer Causes and Control.

Contribution: I conducted the MR analysis for this paper, conceptualised the idea along with Yarmolinsky and Berryman, and wrote the discussion section of the manuscript. This allowed me to develop an understanding of how to interpret MR results and discuss them in a wider epidemiological context.

7. DNA methylation derived systemic inflammation indices are associated with head and neck cancer development and survival
   Ambatipudi, S., **Langdon, R.**, Richmond, R. C., Suderman, M., Koestler, D. C., Kelsey, K. T., Kazmi, N., Penfold, C., Ho, K. M., McArdle, W., Ring, S. M., Pring, M., Waterboer, T., Pawlita, M., Gaunt, T. R., Davey Smith, G., Thomas, S., Ness, A. R. & Relton, C. L., 1 Oct 2018, In : Oral Oncology. 85, p.87-94

Contribution: I was co-first author on this paper with Ambatipudi. I developed DNA methylation predictors of systemic inflammation using cell sub-type information and epigenetic data from HN5000. I also contributed to the introduction, wrote the methods section of this manuscript and contributed to the discussion by appraising the validity of our findings.

8. Causal inference in cancer epidemiology: What is the role of mendelian randomization?
   Yarmolinsky, J., Wade, K. H., Richmond, R. C., **Langdon, R. J**., Bull, C. J., Tilling, K. M., Relton, C. L., Lewis, S. J., Smith, G. D. & Martin, R. M., 1 Sep 2018, In : Cancer Epidemiology, Biomarkers and Prevention. 27, 9, p.995-1010

Contribution: In this large review of Mendelian randomization in the context of cancer, I discussed the limitations of Mendelian randomization with respect to cancer latency and reverse causation. I also reviewed each section of this paper, which provided a much deeper understanding of the benefits and limitations of using Mendelian randomization in the context of cancer epidemiology.

9. Circulating selenium and prostate cancer risk: A Mendelian randomization analysis
   Yarmolinsky, J., Bonilla, C., Haycock, P., **Langdon, R.**, Lotta, L. A., Langenberg, C., Relton, C., Lewis, S., Evans, D., Davey Smith, G. & Martin, R., Sep 2018, In : Journal of the National Cancer Institute. 110, 9, p.1035-1038

Contribution: In addition to reviewing the paper, I conducted sensitivity analyses for this article and designed a figure comparing randomized controlled trials to Mendelian randomization. This figure is seen in Chapter 2.2.4 of this thesis – Mendelian randomization.

10. Influence of puberty timing on adiposity and cardiometabolic traits: A Mendelian randomisation study
    Bell, J., Carslake, D., Wade, K., Richmond, R., **Langdon, R.**, Vincent, E., Holmes, M., Timpson, N. & Davey Smith, G., 28 Aug 2018, In : PLoS Medicine. 15, 8, , e1002641.

Contribution: In addition to contributing generally to the manuscript, I provided guidance for the use of two-sample Mendelian randomization with summary data in this paper, in particular discussing instrument selection for an optimal instrumental variable for BMI.

11. DNA methylation as a marker for prenatal smoke exposure in adults
    Richmond, R., Suderman, M., **Langdon, R.**, Relton, C. & Davey Smith, G., 31 May 2018, In : International Journal of Epidemiology.

Contribution: I generated a DNA methylation predictor of smoking status for this paper using Illumina 450K data and smoking EWAS results from Joehanes et al. My description of DNA methylation score derivation is used in Chapter 6 of this thesis (6.2 – Epigenetic risk score generation).

12. The MR-Base platform supports systematic causal inference across the human phenome
    Hemani, G., Zheng, J., Elsworth, B., Wade, K., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., **Langdon, R.**, Tan, V., Yarmolinsky, J., Shihab, H., Timpson, N., Evans, D., Relton, C., Martin, R., Davey Smith, G., Gaunt, T. & Haycock, P., 30 May 2018, In : eLife. 7, , e34408.

Contribution: I worked closely with the first and last authors of this paper to generate the GWAS data and appraise the statistical functions used in the MR-Base resource. Accordingly, I have an excellent understanding of the advantages and limitations of a resource for systematic Mendelian randomization at this scale.

13. MELODI: Mining Enriched Literature Objects to Derive Intermediates
    Elsworth, B., Dawe, K., Vincent, E., **Langdon, R**., Lynch, B., Martin, R., Relton, C., Higgins, J. & Gaunt, T., Apr 2018, In : International Journal of Epidemiology. 47, 2, p.369–379

Contribution: I worked closely with Elsworth to appraise the various functions of MELODI and contributed the carnitine and pancreatic cancer exemplar found in the results of this paper. Other sections of my contribution to this paper can be seen in Chapters 2.2.1 and Chapter 3.1

2017

14. Application of Mendelian randomization: can we establish causal risk factors for type 2 diabetes in low-to-middle income countries?
    Langdon, R. & Wade, K., 1 Jan 2017, In : Revista Cuidarte. 8, 1, p.1391-1406

Contribution: Wade and I co-wrote this editorial article. I derived the directed acyclic graph figure for this paper which can be seen in Chapter 2.2.4 of this thesis. By investigating low-to-middle income countries in the context of Mendelian randomization, I gained a far greater understanding of the effect of population stratification on these analyses and in the wider context of GWAS.

Acknowledgements

There are few times in life you embark on something as consuming and exciting as a research PhD. I'd like to express my deepest gratitude to my supervisory team: Caroline Relton, Hannah Elliott, Rebecca Richmond and Steve Thomas. Without your guidance and care, completing this PhD would have been a much less enjoyable and fruitful experience. Thank you.

I'd like to thank Jessie Wilcox, Alex Creavin, Prianka Padmanathan, David Johnson and Kaitlin Wade for helping (sometimes pushing) me through the more difficult times and showing me true friendship and patience. Thank you all – I count myself extremely fortunate to know you.

I would like to acknowledge the work of Rhona Beynon in Chapter 6 of my thesis. As detailed in the Research Output section of my thesis, I have used Tables and Figures from a joint first-authored paper with her to supplement discussion of my own work regarding epigenetic prediction of complex traits in individuals with oropharyngeal cancer.

My sincerest gratitude goes out to the participants of the HNC OncoArray Study, HN5000, UK Biobank, and Generation Scotland for the data that was used in this thesis. Thank you for sharing your data and making this possible.

Finally, a huge thank you to the Integrative Epidemiology Unit at the University of Bristol and to the Integrative Cancer Epidemiology Programme for providing me with a truly special opportunity and a wonderful group of colleagues.

1.1. Introduction

This chapter introduces the cancer type and background relevant to this thesis. Firstly, oropharyngeal cancer is introduced by briefly describing its biology, epidemiology and clinical impact. Secondly, the current genetic and epigenetic understanding of oropharyngeal cancer (OPC) is outlined. Finally, the chapter is summarised. In addition to introducing the thesis, this section provides a justification of the work presented throughout it.

1.1.1. A brief introduction to oropharyngeal cancer biology

OPC is one of the aetiologically similar cancers of the mouth, pharynx, larynx, paranasal sinuses, nasal cavity and salivary glands collectively known as head and neck cancer (HNC). OPC is located in the pharynx – the hollow tube inside the neck that starts behind the nose and ends at the top of the oesophagus. The oropharynx is the central area of the pharynx (**Figure 1.1**), bordered by the nasopharynx (the top portion of the pharynx) and the hypopharynx (the bottom portion of the pharynx). It includes the soft palate, side and back walls of the throat and the back third of the tongue. The function of the pharynx is to ensure that air travels through the trachea, and that food and water travel to the oesophagus. The oropharynx accepts air from the nasopharynx, which passes through to the hypopharynx and laryngeal pharynx. The oropharynx also accepts food and water from the mouth, passing it to the oesophagus via peristaltic muscular contraction.

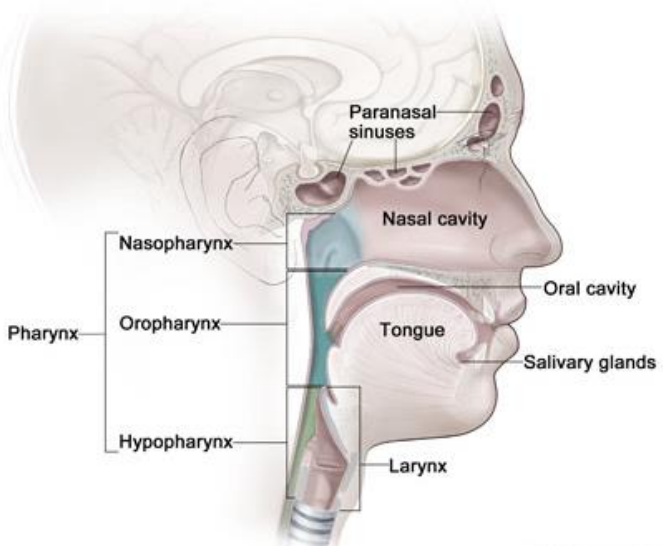


Figure 1.1 - Head and neck cancer anatomical sites

The majority of the oropharynx is lined by non-keratinizing stratified squamous epithelium (NKSSE) [1]. NKSSE consists of squamous (flattened) cells arranged in layers upon a basal membrane

(**Figure 1.2**) [2]. As it isn't keratinizing (i.e. it doesn't possess dead, keratin-transformed surface cells), the surface squamous cells in this type of epithelium lack rigidity, allowing damaged cells to be sloughed and replaced rapidly. Rapid sloughing of oropharyngeal epithelium is an important characteristic of the tissue as it provides a first-line defence from physical (e.g. mechanical shearing caused by the pharyngeal phase of swallowing [3]), chemical (e.g. gastric acid exposure via laryngopharyngeal reflux [4]), and biological (e.g. bacterial invasion of the pharyngeal mucosa [5]) damage.

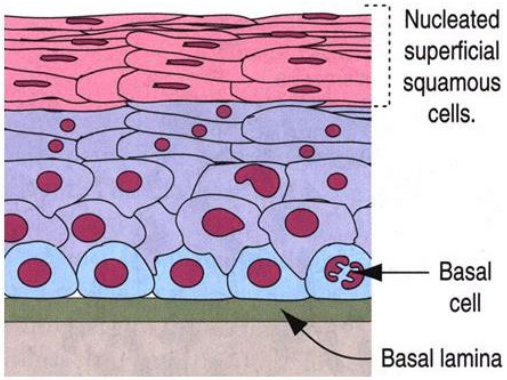


Figure 1.2 - Structure of non-keratinizing squamous stratified epithelium

The predominance of squamous cells over other cell types in the oropharynx result in approximately 95% of oropharyngeal tumours being oropharyngeal squamous cell carcinomas (OPSCCs) [6] . Less common neoplasms originating in the oropharynx include minor salivary gland tumours [7] and squamous carcinoma with lymphoid stroma (lymphoepithelioma) [8]. Additionally, malignant lymphomas can occur in Waldeyer's ring of the oropharynx [9]; a ringed arrangement of lymphoid tissue in the pharynx, surrounding the nasopharynx and oropharynx, with tonsillar tissue located above and below the soft palate and back of the mouth cavity.

1.1.2. Epidemiology

Incidence and prevalence of oropharyngeal cancer in the UK

Epidemiology typically quantifies a disease by considering its incidence (the number of new cases of a disease within a given time period) and prevalence (the proportion of individuals with a disease at a given time) within a defined population. When combined, HNCs are the seventh most common cancer type [10], accounting for an estimated 550,000 incident cases globally, per year, corresponding to ~3.9% of all cancer cases [11]. Of these incident cases, an estimated 100,500 are of

the oropharynx (~0.8% of all cancer cases; ~18% of all HNCs), with an age-standardised incidence rate (ASR: a weighted average of age-specific incidence rates, where the weights are the proportions of individuals in each age group) of 1.4 per 100,000 for both sexes combined (2.3 for men; 0.5 for women) [11].

The number of region-specific OPC diagnoses contributing to the 100,500 global diagnoses ranges substantially, from ~0.9% in Oceania to ~35% in south-central Asia. Northern and western Europe contribute ~13.2% of all OPC diagnoses; third highest, behind south-central Asia (35%; above) and northern America (15%) (**Figure 1.3** – inner chart). OPC diagnoses as a proportion of all HNC sub-types also show marked geographic variability, ranging from ~8.2% of all HNCs in northern Africa and western Asia, to ~34.2% of all HNCs in northern America. The proportion of OPC in HNC in northern and western Europe was second highest at ~29.5% (**Figure 1.3** - outer chart). Understandably, the regional heterogeneity of OPC incidence affects the heterogeneity of ASRs for this disease. Worldwide, the ASR for OPC is 1.4 per 100,000 (calculated using the world standard), but ranges from 0.5 in eastern/south-east Asia to as high as 2.9 in northern America. Rate disparities such as these indicate that a proportion of these cancers are preventable; if not entirely due to genetic differences between populations, ASR rate disparities are suggestive, in part, of a geographic and cultural heterogeneity in exposure to causal risk factors. Large regional differences in OPC prevalence provide a rationale for targeted investigation of risk factors for OPC, particularly where prevalence of this sub-type is proportionally high.

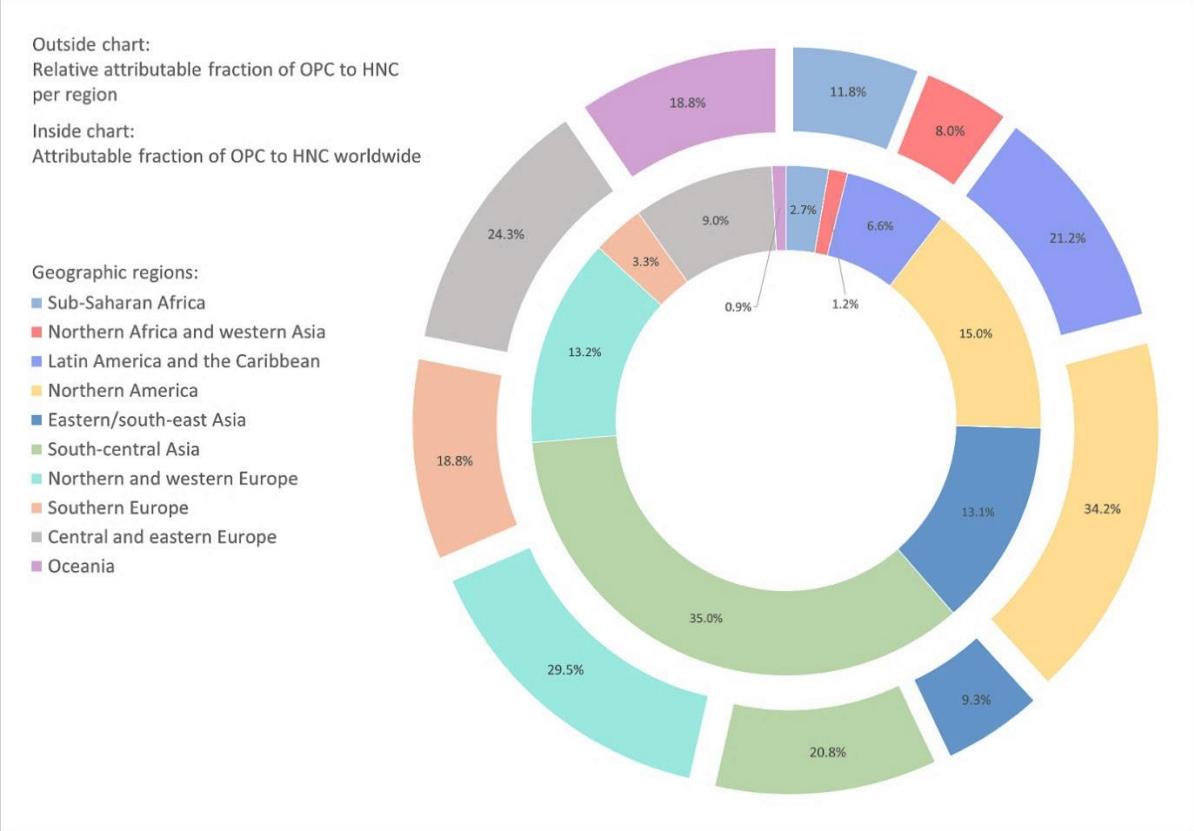


Figure 1.3 - Geographic variation in prevalence of OPC as a HNC sub-type

In 2011 in the UK, the ASR of HNC for both sexes was 15.9 per 100,000; a 30.3% increase from 12.2 per 100,000 in 2002 [12]. However, during this time, incidence of the OPSCC sub-type increased by 100.6%. Sex-specific incidence rates in the UK from 1995-2011 have been reported in epidemiological literature, showing that the ASR for OPSCC almost tripled in men (from 2.0 to 5.8), and more than doubled in women (0.8 to 1.7) during this period [12]. Additionally, the Oxford Cancer Intelligence Unit Profile of Head and Neck Cancers in England: Incidence, Mortality and Survival reports that the incidence of OPC more than doubled from 1990 to 2006, and nearly doubled again from 2006 to 2011 [13]. A notable increase in OPC incidence is also similar across Ireland, Scotland and Wales, respectively (**Figure 1.4**). It appears, therefore, that OPC is increasing at a rapid rate, establishing this disease as an increasing health problem in the UK.

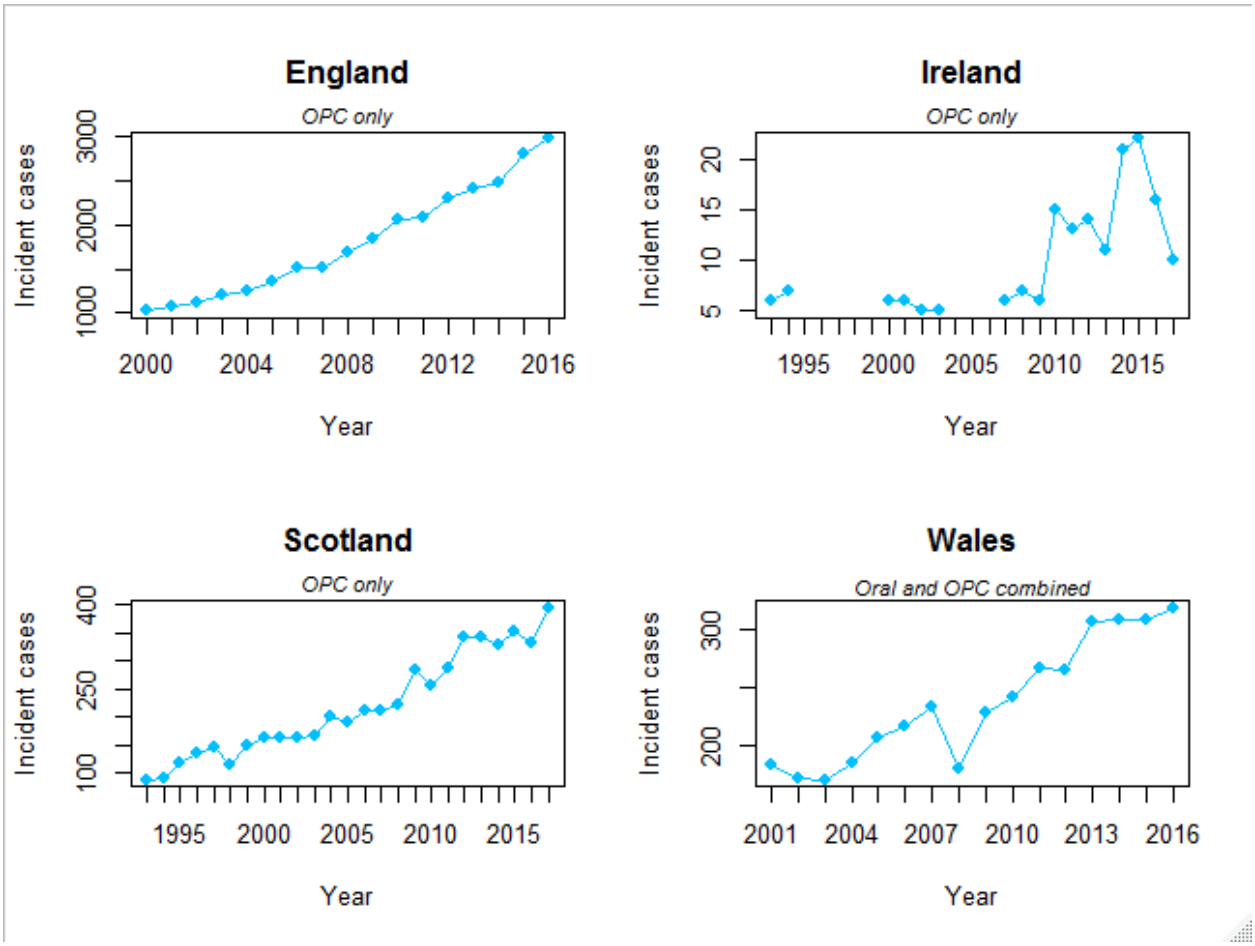


Figure 1.4 - Incident cases of OPC from 1993-2017 across England, Ireland, Scotland and Wales

Common OPC risk factors

Globally [11], and in the UK [14], the most prevalent risk factors in observational epidemiological literature for OPC are tobacco smoking [15, 16], alcohol consumption [17-19] and human papillomavirus (HPV) infection (notably the HPV16 subtype) [20-24]. Of these risk factors, tobacco smoking and alcohol consumption show synergistic effects on OPC [25]. Both smoking and HPV16 infection, and alcohol and HPV16 infection, however, do not currently appear to affect risk of OPC synergistically [26], despite both increasing risk of HPV [27, 28]. In 2015 in the UK, smoking, alcohol and HPV were estimated to attribute 88.4% of the population risk for pharyngeal cancer (tobacco 37.4%; alcohol 37.8%; infections 70.2%) [29]. A brief overview of each can be seen below.

<u>Alcohol consumption</u>

Consumption of alcohol varies significantly across the world. In 2016, the worldwide average (mean ± SD) alcohol consumption per capita was 6.4L ± 4.1L (litres of pure alcohol [to account for

differences in drink preference]; age 15+) (**Figure 1.5**) [30]. Moldova showed the highest alcohol consumption of 15.2L; the lowest alcohol consumption was 0L in Bangladesh, Kuwait, Libya, Mauritania and Somalia. In the European Union, which was the second highest alcohol consumption of any continent/grouping (Central Europe and the Baltics was highest at 12.2L), consumption averaged 11.3L [30]. The UK average in 2016 was 11.5L, establishing it as one of the largest consumers of alcohol in the world – 1.8x the global average – a prime target for primary prevention of OPC if established as a true causal risk factor.

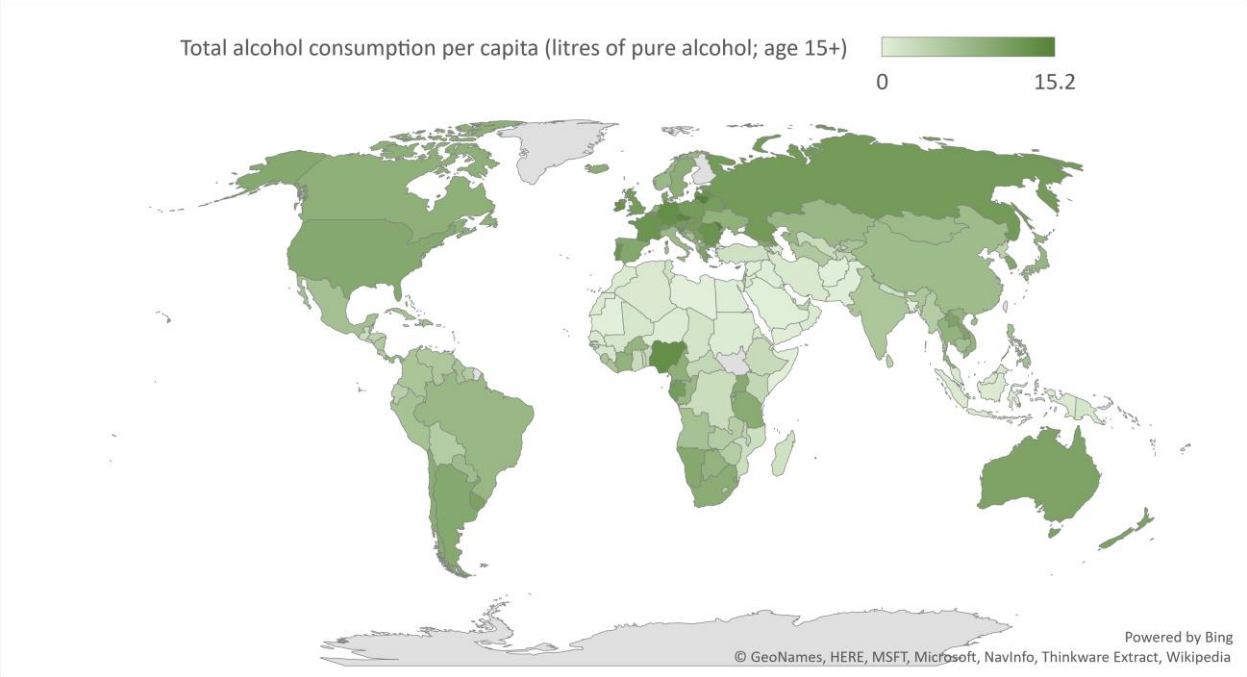


Figure 1.5 - Alcohol consumption across the world in 2016. Data are plotted from the World Bank - World Development Indicators

Observational evidence for the effect of general alcohol consumption on OPC

Epidemiological studies investigating alcohol consumption in relation to risk of OPC appear to show consistent evidence of elevated risk with increasing alcohol consumption [31]. In an analysis of data from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort study, the relative risk of oral-pharyngeal (combined oral cavity, nasopharyngeal, oropharyngeal and hypopharyngeal) cancer for over 60 grams (4 drinks) per day was calculated to be 9.2 (95% confidence interval [CI]: 2.8 to 30.9) [31, 32]. Other prospective cohort studies reported lower, albeit still notable risks of alcohol consumption on risk of oral-pharyngeal cancer. The American Cancer Society (ACS) prospective study report a risk ratio (RR) of 3.2 (95% CI: 1.7 to 6.1) for more than 4 drinks/day [33], and a prospective study of 10,000 Norwegian men report a RR of 3.9 (95% CI: 2.1 to 7.1) for

consumption of alcohol 4–7 times per week [34]. Prospective cohort studies are considered to have a high level of evidence in the hierarchy of evidence for epidemiological study design (**Figure 1.6**), due to their longitudinal nature and regular contemporaneous collection of results. Longitudinal observation in prospective studies allows a direction of effect to be established between an exposure and outcome (in this case, alcohol → OPC), whilst regular, contemporaneous collection of results minimises recall bias; a systematic error that occurs when study participants do not remember past experiences accurately or omit details (as generally found in case-control or cross-sectional studies). However, prospective cohort study design still presents with drawbacks. At baseline, it is not possible to measure and control for every factor which may affect OPC via alcohol consumption; thus it cannot be ruled out that the association of alcohol with OPC seen in the prospective studies above may be due to confounding. As such, association, not causation, can be asserted for the effect of alcohol on OPC in these studies. This is in contrast to the gold standard of epidemiological study design – the randomised controlled trial (RCT) - which uses random allocation of individuals to an intervention or control group at baseline to account for confounding. It should be noted, however, that an RCT investigating alcohol consumption on risk of OPC would be both unethical and infeasible, due to the relative rarity, latency and harmful nature of OPC (and indeed, other alcohol-related diseases) as an outcome. Therefore, RCT evidence of the effect of alcohol on OPC does not currently exist.

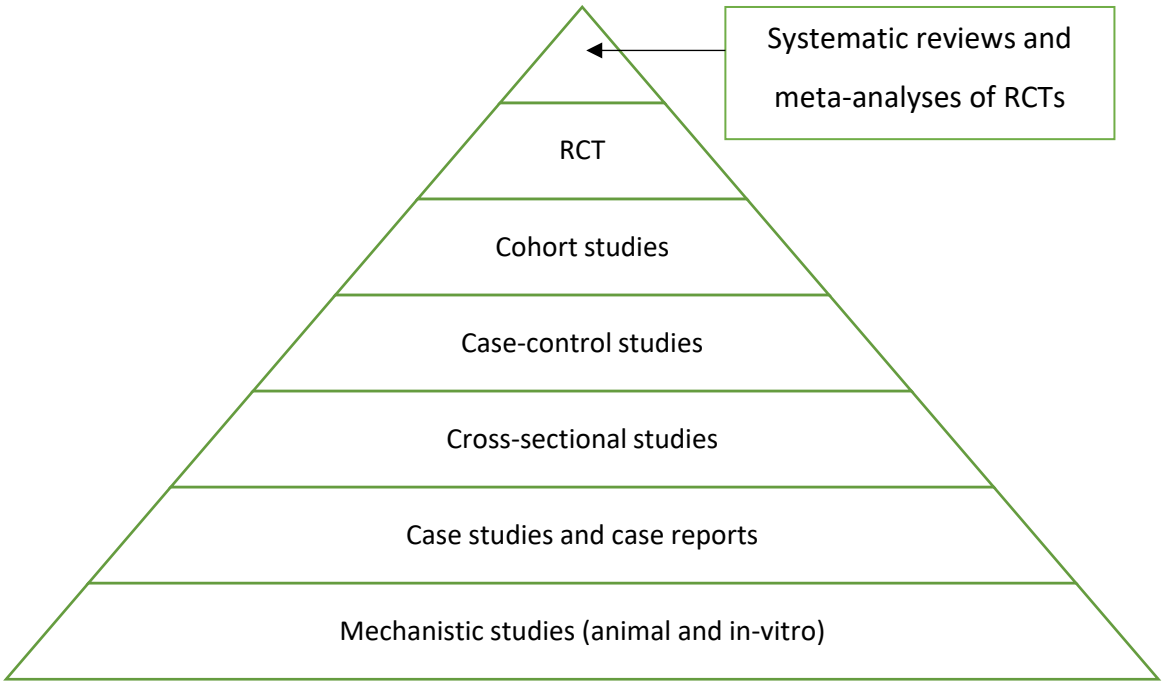


Figure 1.6 – Pyramid diagram detailing the hierarchy of evidence for epidemiological study design

Finally, there is observational evidence for the effect of alcohol on OPC in England. A case-control study of risk factors for oral cancer in newly diagnosed patients aged 45 years and younger found that males who drank >21 units of alcohol per week were over 8 times more likely to develop oral cancer or OPC (RR: 8.1; 95% CI: 1.6 to 40.1) [35]. However, the level of evidence from this study is lower than the prospective studies mentioned above. In addition to being unable to fully measure and control for confounding, case-control studies are not prospective, thus cannot establish a true direction of effect; it is uncertain whether alcohol affects risk of OPC, or those with OPC tend to drink more alcohol as a result. Case-control studies do have an advantage over prospective studies, however. Statistical power is increased by being able to match participants who already have OPC with healthy controls. This can afford a smaller sample size to achieve adequate power and, more generally, a viable method of investigating rarer outcomes with long latency periods (such as OPC). Statistical power is a pitfall of prospective cohort studies, as sample sizes need to be particularly large to have sufficient power to account for the rarity and latency of OPC (and other similar outcomes).

Observational evidence for the effect of type of alcohol consumed on OPC

Between 1890-2014, the population share of alcohol type consumed (beer, wine and spirits) in the UK changed significantly [36]. A sharp increase in the population share of wine drinking, from 4% in 1964 to 41% in 2014, affected a decrease in the population share of beer drinking from 81% in 1964 to 37% in 2014. Spirit drinking in the UK was 30% of the population share in 1890, but remained largely steady, at around 20% of the UK's total pure alcohol consumption from 1929-2014 (**Figure 1.7**). The notable increase in wine consumption is thought to be due to a combination of Big 6 brewers recognising the influence of women as consumers of alcohol, and also due to the ability of UK consumers to travel abroad for their holidays (particularly France, Italy and Spain), establishing wine as a commonplace beverage [37].

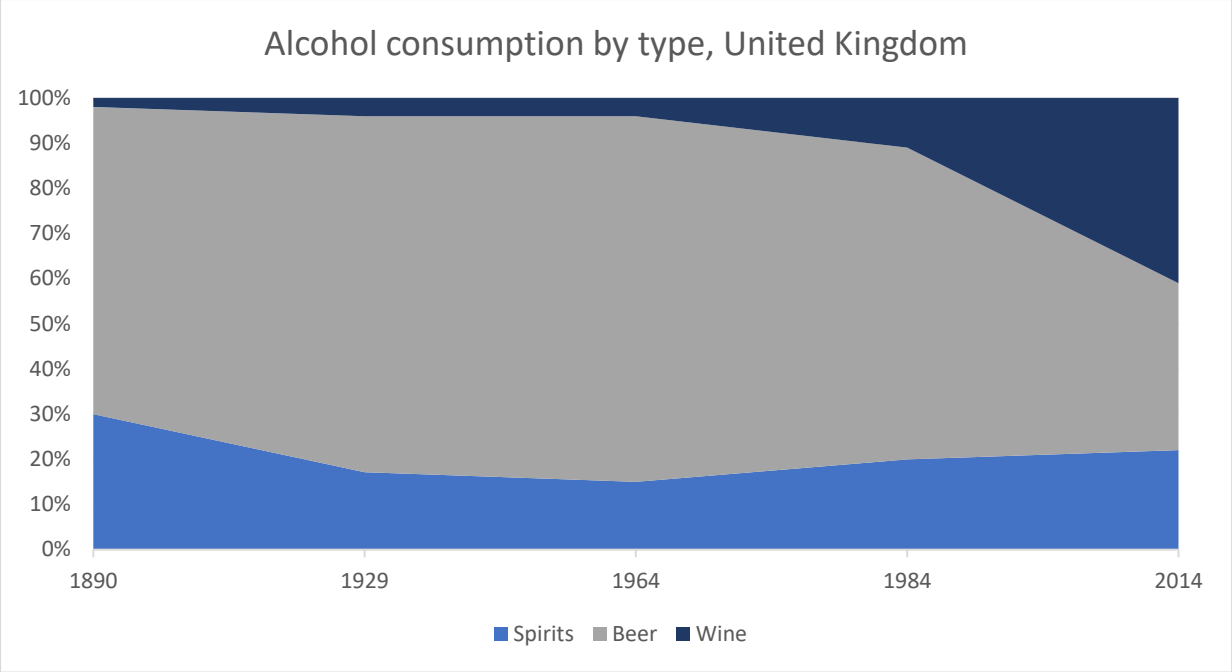


Figure 1.7 - Share of total alcohol consumption from wine, beer and spirits in the UK from 1890-2014, as a percentage of pure alcohol

Epidemiological evidence pertaining to type of alcohol consumed and risk of OPC show differences between risk of wine, beer and spirit drinking in relation to pharyngeal cancer (nasopharynx, oropharynx and hypopharynx cancers combined), yet no notable difference when solely investigating OPC. In a meta-analysis examining type of alcohol on HNC risk in Europeans (9,107 cases, 14,219 controls; N = 15 case-control studies), Purdue et al. found that, for moderate alcohol consumption (≤15 standardised drinks per week), wine and spirit consumption provided no conclusive evidence of association with pharyngeal cancer (wine OR: 1.4; 95% CI: 0.9 to 2.2 | spirit OR: 2.0; 95% CI: 0.9 to 4.6), as opposed to beer consumption, which showed a positive association (OR: 2.3; 95% CI: 1.7 to 3.1) [38]. By only analysing case-control studies, findings from this meta-analysis cannot be entirely free of unmeasured confounding factors, and the direction of effect cannot be ascertained. The authors discuss their wine consumption findings, providing evidence to suggest that confounding factors may have affected these results. In studies conducted in the United States and Northern Europe, wine has previously been associated with higher intake of a healthy diet, higher education and lower smoking levels compared to other alcohol types [39-41]. In seems, therefore, that wine consumption may be correlated with socio-economic position (SEP), which is itself an independent risk factor for OPC (see "**Socio-economic position**" below). However, it should be noted that findings from this study do not stratify by pharyngeal cancer sub-type, thus results may be biased by a different sub-type than OPC (i.e. nasopharyngeal or hypopharyngeal cancer) being susceptible to alcohol type consumed.

Turati et al. investigated type of alcohol consumed against risk of OPC uniquely, in Europeans. In a 2013 meta-analysis (N studies = 11), compared with non- or occasional drinking (combined), pooled RRs for type of alcohol consumed were found to be 2.12 (95% CI: 1.37 to 3.29) for wine-only, 2.43 (95% CI: 1.92 to 3.07) for beer-only and 2.30 (95% CI: 1.78 to 2.98) for spirit-only consumption [42]. No heterogeneity was found between types of beverages (P = 0.856), though interestingly, within the wine-only studies (N = 10), there was evidence of substantial heterogeneity (I^2: 76.9%; *P*: 0.000). Four studies of 11 showed null findings for the effect of this beverage type on OPC risk, indicating presence of a source of confounding. As with Purdue et al., Turati et al. cannot assert causality with their findings; confounding and reverse causation are key limitations of observational epidemiology and further causal inference methodologies are required to ascertain the true causal effect of type of alcohol consumed on risk of OPC.

Observational evidence for the effect of alcohol cessation on OPC

There is epidemiological evidence for an inverse association between cessation of alcohol consumption and risk of OPC. In a 2010 paper published by the International Head and Neck Cancer (INHANCE) consortium, length of alcohol cessation was investigated against risk of HNC in Europeans (N case-control studies = 13) [43]. Among subjects who drank one or more drinks per day, the risk of HNC seemed to decrease as years of drinking cessation increased. ORs for HNC risk were 0.99 (95% CI: 0.69 to 1.43) for >1–4 years cessation, 0.90 (95% CI: 0.62 to 1.30) for 5–9 years cessation, 0.94 (95% CI: 0.75 to 1.18) for 10–19 years and 0.60 (95% CI: 0.40 to 0.89) for ≥20 years cessation. However, when investigating OPC and hypopharyngeal cancer as a grouping, there was no evidence that alcohol cessation affected risk of these cancers (OR ≥20 years cessation: 0.74; 95% CI: 0.50 to 1.09). Given an initial evidence base for an inverse association between alcohol cessation and HNC, robust causal inference methods and investigation into an OPC-specific effect is needed to ascertain the true casual effect of this risk factor on OPC incidence.

Observational evidence for an independent effect of alcohol consumption on OPC

Alcohol consumption and smoking are strongly correlated in the general population [44] and thought to confer a synergistic (greater than multiplicative interaction) effect on HNC risk [25]. Given the strength of interaction effect conferred by these risk factors on OPC, it is important to distinguish the individual effect of alcohol consumption on OPC in addition to quantifying the joint effect of smoking and alcohol consumption. The same is true for the effect of smoking on OPC (see

Observational evidence for an independent effect of alcohol consumption on OPC below). By ascertaining these metrics, a more suitable, cost-effective intervention strategy can be generated for reduction of OPC; the sub-group which would benefit most (smokers vs drinkers) from intervention can be established. The effect of alcohol on oropharyngeal and hypopharyngeal cancer (combined) in American and European non-smokers (i.e. independently of smoking) has been assessed by the INHANCE Consortium. In a 2009 paper, including 3,899 cases and 15,751 controls, the authors found that 3 or more alcoholic drinks per day (vs no drinks per day) showed an elevated risk of these cancers in non-smokers (OR: 2.94; 95% CI: 1.73 to 5.02), though found no evidence for 1-2 drinks per day on pharyngeal cancer risk amongst non-smokers (OR: 1.26; 95% CI: 0.92 to 1.73) [25]. For oropharyngeal and hypopharyngeal cancer combined, the population-attributable fraction of alcohol (in 2009), independent of smoking, was 5.6 % (95% CI: 1.9% to 7.3%) [25].

Summary of alcohol as a risk factor for OPC

In summary, alcohol appears to be an independent risk factor for Europeans and Americans for HNC, oral cancer and pharyngeal cancer, with 3-4 drinks per day conferring a 3-fold increased risk of these cancers. However, there appears to be a paucity of OPC-specific publications with respect to this risk factor; most of the epidemiological evidence provided above was derived from articles which group OPC with either oral cavity of hypopharyngeal cancer, potentially biasing effect estimates. Using genetic and epigenetic data with robust causal inference methods such as Mendelian randomization (MR; see Chapter 2.2.4 – *Mendelian randomization*), an effect estimate specific to OPC for alcohol consumption could be derived to address this shortfall and augment current understanding of the risk of alcohol on OPC. Furthermore, using MR, the hypothesis that wine consumption specifically may be able to be appraised using MR to corroborate the doubling in OPC risk seen by Turati et al for wine consumption.

<u>Smoking</u>

Much like alcohol consumption, prevalence of tobacco consumption varies significantly across the world. The average worldwide prevalence of smoking in 2016 was 20.5% (**Figure 1.8**) [30]. Kiribati (a sovereign state in Micronesia) showed the largest smoking prevalence of 48.3% in 2016, with Montenegro and Greece second and third at 46.1% and 43.8%, respectively. Honduras showed the smallest smoking prevalence at 2.1%, followed by Ghana and Ethiopia at 4.0% and 4.4%, respectively. The European Union average was 28.2%; 7.7% above the global average and, according to the Tobacco

Control Scale (TCS; a quantitative scoring system based on expert opinions that evaluates policies on tobacco prices, smoke-free places, spending on information campaigns, bans on advertising, health warning labels and treatment for smoking cessation), possibly reflects a lack of meaningful change in policy implementation by the region (**Figure 1.9**). For Europe, TCS scores ranged between 40 and 60 (out of 100), showing substantial room for improvement in each of North, South, East and West Europe. In 2016, smoking prevalence was 23.1% in the UK. Unlike alcohol consumption, smoking prevalence in the UK is below the European average, reflecting being at the forefront of tobacco control policies and scientific research on tobacco control (#1 in the EU) with a TCS score of 81 in 2016 (https://www.europeancancerleagues.org/ecl-map/) [45, 46]. Such a push towards tobacco control has resulted in a steady decline in prevalence of smoking from 2000-2016 (**Figure 1.10**). Given a decrease in smoking in the UK over the past 16 years, in contrast to an increase in OPC incidence, it would appear that the causal landscape of OPC has shifted to include other risk factors (e.g. HPV infection) that aren't smoking.

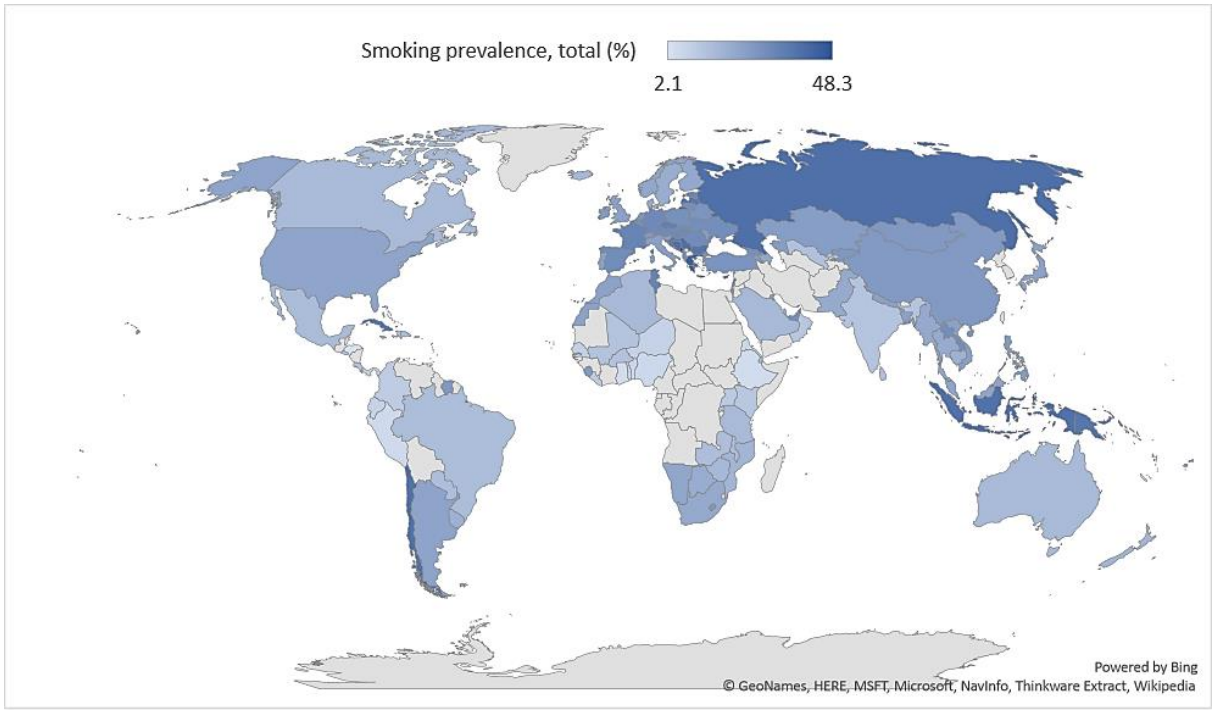


Figure 1.8 - Smoking prevalence across the world in 2016. Data are plotted from the World Bank - World Development Indicators

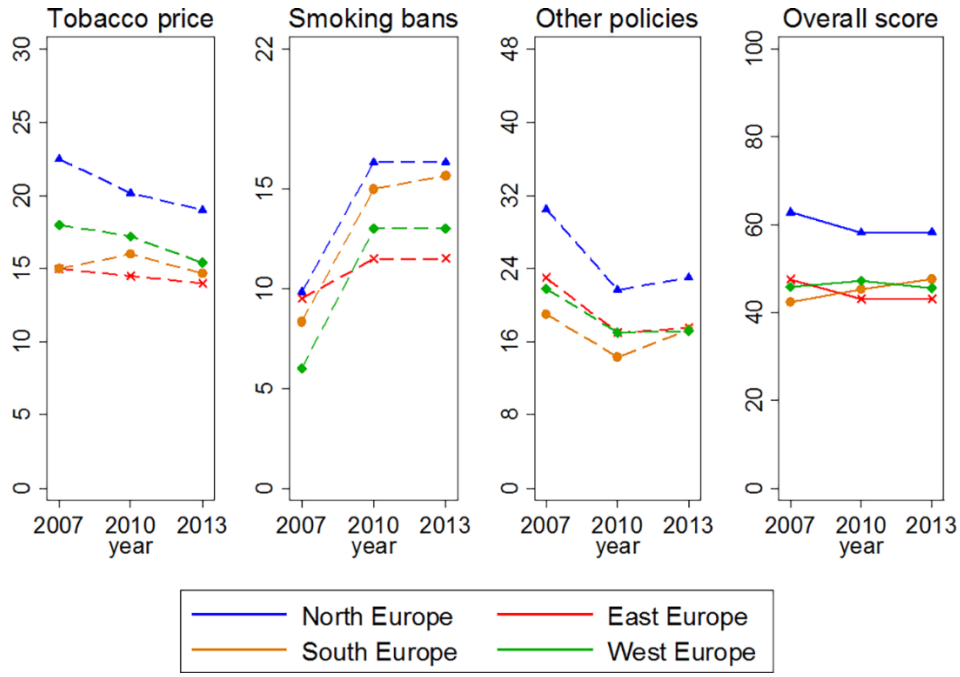


Figure 1.9 - Trends in implementation of tobacco control policies by European region (2007–2013). Plots obtained using the Tobacco Control Scale scoring system from Joossens & Raw. Markers indicate mean scores for the countries included in the study, except for Macedonia in East Europe (not available). Y-axes show theoretical ranges. "Other policies" is the combination of spending for public information campaigns, bans on advertising, health warning labels, and treatment for smoking cessation.

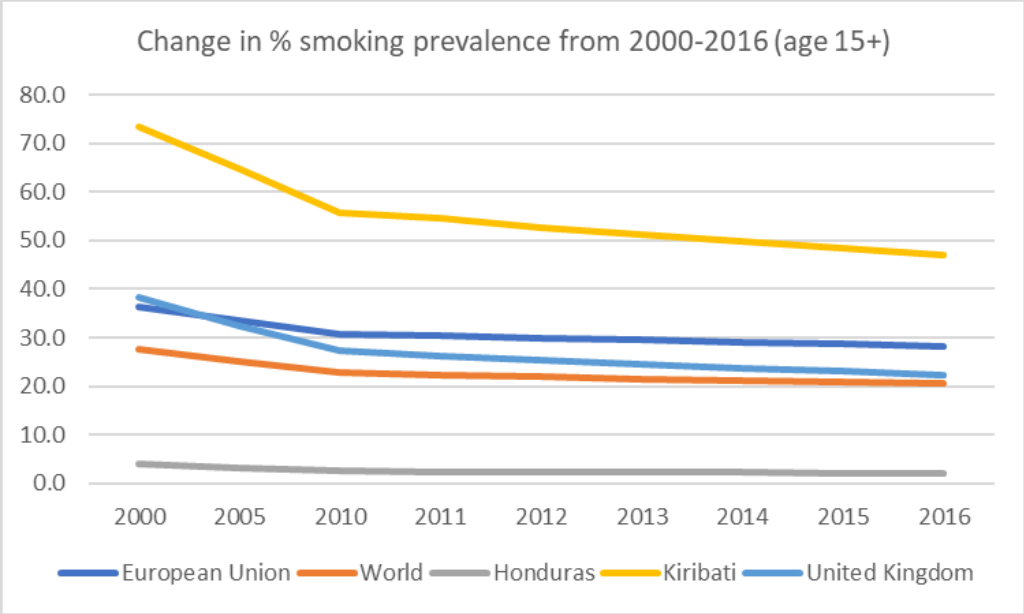


Figure 1.10 - Change in smoking prevalence in the UK from 2000-2016. Global average (World), highest prevalence (Kiribati) and lowest prevalence (Honduras) are included for comparison. Data are from The World Bank – World Development Indicators

Observational evidence for the effect of general tobacco consumption on OPC

Tobacco has been found to be associated with OPC in European populations by multiple epidemiological studies. The 2004 IARC monograph establishing the carcinogenic effect of tobacco smoke and involuntary smoking on cancer considered tobacco smoking to have sufficient epidemiological evidence to be causally related to risk of OPC [47]. More recent epidemiological evidence supports this assertion. Anantharaman et al. conducted a case-control analysis of smoking against OPC using data from the HNC case-control study within EPIC and the Alcohol-Related Cancers and Genetic Susceptibility in Europe (ARCAGE) study. With a sample size of 1,093 (274 cases and 819 controls), Anantharaman et al. observed greater than a 5-fold increase in risk of OPC (OR: 5.34; 95% CI: 3.89 to 7.33) when comparing current vs never smokers [48]. The increased risk of smoking on OPC appears to persist when examining the effect of ever vs never smoking (that is, with former smokers also included in the analysis). Wyss et al. report a tripling in risk of OPC in ever vs never smokers (OR: 3.01; 95% CI: 2.71 to 3.35 for their case-control study using INHANCE data, containing 3,828 cases and 11,277 controls [49]. The INHANCE consortium also report the effect of increasing cigarettes per day, revealing a consistently elevated risk estimate for every additional 10 cigarettes smoked (vs not smoking), observing a dose-response relationship (1-10 cigarettes/day OR: 2.55; 11-20 cigarettes/day OR: 2.15; 21-30 cigarettes/day OR: 3.86; 31 to 40 cigarettes/day OR: 4.82; >40 cigarettes/day OR: 3.10; P[trend] < 0.001) [19]. It should be noted for these studies that a case-control design is susceptible to the same biases as discussed for the alcohol case-control studies. Furthermore, there is a paucity of studies exclusively investigating the risk of smoking on OPC as a distinct sub-type. Despite existing epidemiological evidence, more studies with larger sample sizes (investigating OPC exclusively) and a longitudinal framework are required to provide greater causal inference, in an observational setting, between smoking and OPC.

Observational evidence for the effect of type of tobacco product on OPC

Wyss et al. studied the risk of cancer associated with several tobacco products in relation to HNC sub-types [49]. For combustible products, this revealed increased risks of OPC for cigars (OR: 2.31; 95% CI: 1.54 to 3.45; 543 cases and 6,979 controls) and pipes (OR: 1.65; 95% CI: 1.04 to 2.60; 553 cases and 7,145 controls) vs never cigarette smokers, in addition to (as seen above) an increased risk seen for ever cigarette smokers vs never smokers (OR: 3.01; 95% CI: 2.71 to 3.35; 3,828 cases and 11,277 controls). Interestingly, in a later publication by the same author, smokeless tobacco was compared against risk of OPC in the United States (US), with no evidence of increased risk of OPC seen

between smokeless tobacco users and never smokers (OR: 0.98; 95% CI: 0.57 to 1.68; 524 cases and 3,333 controls) , or smokeless tobacco users and ever smokers (OR: 0.94; 95% CI: 0.72 to 1.22; 1,847 cases and 5,041 controls) [50]. However, these risk estimates were derived in a population (US) where the average smokeless tobacco prevalence is ~4 times higher than the UK (US men: 6.5%, US women: 0.4% [51]; UK men: 1.6%, UK women: 0.5% [52]). Currently, no evidence exists for use of smokeless tobacco and increased risk of OPC in Europe [53]. A 2015 systematic review and meta-analysis examining the global burden of smokeless tobacco on disease found no evidence to suggest smokeless tobacco affected risk of pharyngeal cancer in Europeans (OR: 1.45; 95% CI: 0.84 to 3.01), albeit with data restricted to Scandinavian countries [54]. As such, it may not be viable to extrapolate either finding to the UK and further research is needed to ascertain the effect of smokeless tobacco in the UK population.

Observational evidence for the effect of smoking cessation on OPC

There is epidemiological evidence for an inverse association between cessation of smoking and risk of OPC. The same 2010 paper published by the INHANCE consortium mentioned above for alcohol cessation investigated the length of smoking cessation against risk of OPC in Europeans (N case-control studies = 13) [43]. Compared to current smokers who smoke ≥10 cigarettes a day, increased time since cessation of smoking was associated with decreased risk of OPC (grouped with hypopharyngeal cancer). Risk estimates were calculated as follows: OR >1-4 years cessation: 0.70 (95% CI: 0.53 to 0.94), OR 5-9 years cessation: 0.47 (95% CI 0.36 to 0.61), OR 10-19 years cessation: 0.34 (95% CI: 0.25 to 0.45), OR ≥20 years cessation: 0.23 (95% CI: 0.16 to 0.35); estimates showed a dose-response relationship (*P* for trend <0.01). Comparatively, never smokers were calculated as having an OR of 0.17 (95% CI: 0.11 to 0.27) vs current smokers, indicating that risk of OPC approaches that of never smokers at around 20 years smoking cessation.

Summary of smoking as a risk factor for OPC

In summary, smoking is reported in epidemiological literature as a significant risk factor for OPC. In European populations, smoking prevalence is still above average (28.2% vs the worldwide average of 20.5%), with no evidence of meaningful interventions being employed across the continent as a whole. Interestingly, smoking prevalence in the UK has steadily declined as a result of employing effective tobacco control measures, evidenced by the 2016 UK TCS score of 81/100; the highest in Europe. Nevertheless, the UK is still above the global average with respect to smoking prevalence, at

23.1%. Observational epidemiology consistently reports smoking as a significant risk factor for OPC, with current smokers at a ~3-fold risk of OPC compared to never smokers. Cessation of smoking has been seen to be associated with a marked decrease in risk of OPC, approaching the equivalent risk of never smokers past 20 years cessation. Current epidemiological research tends to combine OPC with other HNC sub-types, potentially biasing effect estimates for smoking on risk of OPC. Furthermore, evidence for association between smoking and OPC is derived from observational epidemiological literature, which has shown to be prone to confounding, reverse causation and other innate biases from using observational data. Methods such as MR could potentially overcome these shortcomings and provide an unconfounded, OPC-specific risk estimate (see Chapter 2). Furthermore, given the substantial estimated risk of smoking on OPC incidence, establishing potential causal intermediates between smoking and OPC may shed insight into potential therapeutic targets for this aetiological pathway.

HPV infection

HPV infection is recognised in epidemiological literature as one of the primary causes of OPC and is transmitted from person to person via oral, vaginal or anal sex [55]. Biologically, HPV is a DNA oncovirus, meaning its genome can integrate directly into the host genome (as opposed to RNA oncovirus genomes, which must be reverse-transcribed to DNA before integration) to affect tumorigenesis [56]. HPV is also epitheliotropic; it possesses an affinity for epithelium, making the oropharynx a common, local site of infection for this virus [57]. There are over 120 different HPV subtypes: "low-risk" types such as HPV6 and HPV11 are responsible for benign proliferation of epithelial tissue. Two "high-risk", oncogenic types, HPV16 and HPV18, are both well-established initiators of around 30% of worldwide OPSCC [58-60]. HPV16 in particular is thought to account for more than 90% of all HPV-positive OPC. The "high-risk" HPV sub-types are known to affect oncogenesis via two oncoproteins, HPV E6 and HPV E7, which inhibit p53 and pRB tumour suppressor pathways. Disruption of these two common cancer pathways greatly increases risk of mutation and oncogenesis [22, 61].

One of the most common methods for determining if an individual has a (high-risk) HPV infection is serological examination; that is, assessing presence of HPV antibodies (namely E6 and E7) in blood serum. Case–control studies of HPV serology in OPC are feasible in serum due to the relative ease and low patient burden of obtaining blood samples compared to fresh tumour biopsies. Detecting HPV in tissue samples (e.g. buccal rinses) is problematic due to a low yield of detectable HPV DNA and a lack of correlation with HPV DNA levels in biopsy samples [62]. However, serology is

not a perfect measure of HPV infection. As HPV infections are sexually transmitted diseases which are typically localised, it is difficult to distinguish between any combination of oral, cervical, anal and penile infections (all of which are possible HPV infection sites) based on serology alone. Furthermore, there are individual differences in the time between infection and development of an HPV E6- or E7-specific antibody (seroconversion) detectable in blood [63], with some individuals failing to seroconvert altogether [64]. Therefore, in studies examining HPV infection using seropositivity, HPV-seronegative subjects may still be infected with the virus, thus may bias results of observational analyses.

The attributable fraction of HPV to OPC varies dramatically depending on world region. A 2012 systematic review estimating this metric found fractions ranging from 13% in sub-Saharan Africa to 60% in Korea [60]. In 2008, the global population-attributable fraction of HPV for OPC was 25.6%, and 4.8% across all cancers [65]. In 2012, the global attributable fraction of HPV for OPC was estimated to be 30.8%. In the same year, HPV in North-West Europe showed a population attributable fraction of 42% [66]. The incidence of HNCs attributable to HPV can be seen in **Figure 1.11**. Generally, Western, more economically-developed countries have greater incidence of HPV-driven HNCs than their Eastern and less economically-developed contemporaries.

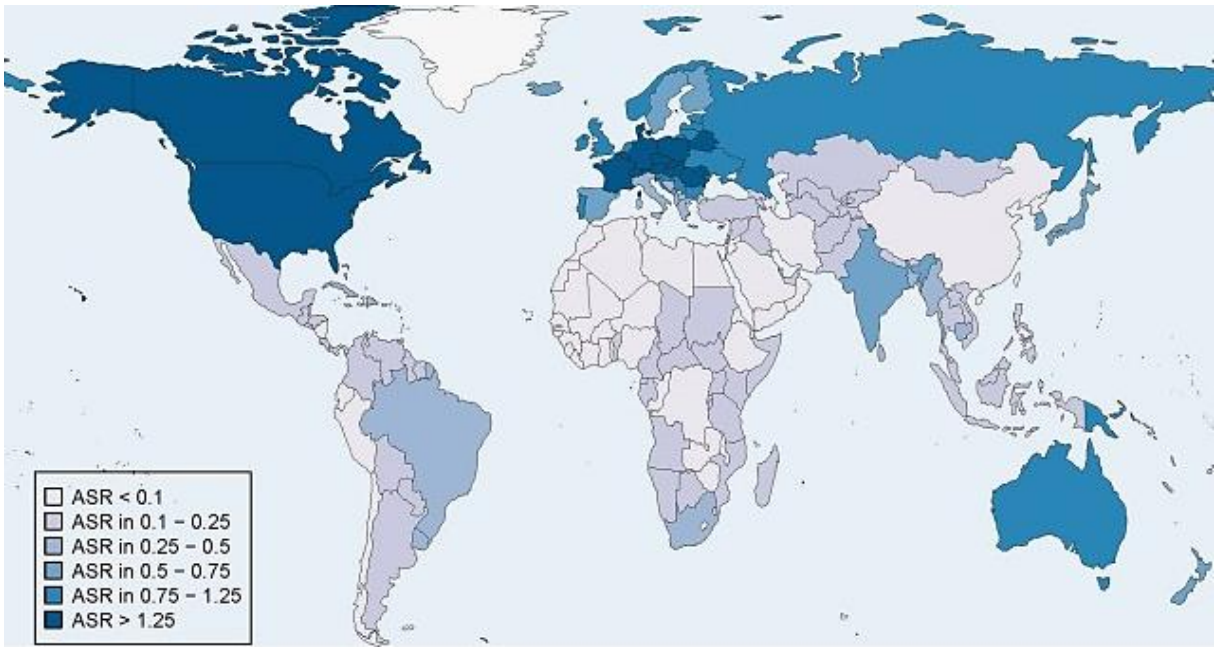


Figure 1.11 - Age-standardised incident rates for HPV-driven HNC worldwide in 2012. Figure adapted from Martel et al. [60]

Between 2002 and 2011, HPV prevalence amongst OPC in the UK was estimated to be 51.8% (95% CI: 49.3% to 54.4%) and remained unchanged throughout the decade [12]. Interestingly, during

this period, incidence of OPC (specifically, OPSCC) almost doubled (UK ASR for 2002: 2.1 [95% CI: 1.9 to 2.2]; UK ASR for 2011: 4.1 [95% CI: 4.0 to 4.3]). Although the total number of OPSCCs diagnosed within the United Kingdom from 2002 to 2011 nearly doubled, the proportion of HPV-positive cases remained at approximately 50%. Schache et al. argue that the rapidly increasing incidence of OPSCC in the United Kingdom cannot be solely attributable to the influence of HPV [12]. The stable, equal proportion of HPV-positive and HPV-negative cases provide a motivation to determine whether other risk factors are playing a role in OPC aetiology (particularly in the case of HPV-negative OPC).

Following evidence of the carcinogenic effect of HPV on cervical cancer, a national HPV immunisation program was introduced to women aged 16-18 the UK in September 2008 by the UK Department of Health [67]. Compared to 2010, prevalence of either HPV16 or HPV18 in 16 to 18-year-old women dramatically decreased by 2016 (postvaccination prevalence 2010: 8.2% [95% CI: 6.6% to 9.9%], postvaccination prevalence 2016: 1.6% [95% CI: 0.6% to 2.6%], $P_{trend} < 0.001$). These findings allude to the future emergence of a novel OPC population who are neither HPV-positive nor heavy or sustained smokers or drinkers. Currently, this population is restricted to women, as UK males are not part of the nation HPV immunisation program.

Observational evidence for HPV as an OPC risk factor

In 2007, IARC concluded that there was sufficient evidence to report a carcinogenic role of HPV16 in tumours of the oropharynx, estimating that ~31% of OPSCC cases were attributable to HPV [68]. HPV has since been recognised as one of the predominant aetiological factors for OPC, with a growing body of research suggesting that HPV-positive OPC is a distinct entity from the "traditional" OPC caused by smoking and alcohol consumption [23, 69].

Few studies quantify the risk of HPV16 infection on incidence of OPC. One study by Anantharaman et al. observed huge odds ratios for the risk of OPC for HPV16 E6-positive individuals vs negative controls (OR: 147.31; 95% CI: 83.07 to 361.24). Additionally, this study investigated whether HPV16 infection and smoking exhibited a synergistic effect on OPC risk. HPV16 E6-positivity and smoking appeared to interact on an additive scale (synergy index [SI; a confidence interval estimation of interaction derived by Hosmer and Lemeshow [70]] for interaction 1.32; 95% CI: 0.51 to 3.45), finding similar results when HPV16 L1-positivity was examined (SI for interaction: 0.75; 95% CI: 0.51 to 1.12). These results indicate that smoking and HPV infection affect OPC independently.

As mentioned above, HPV16 E6 seropositivity is thought to be a marker of persistent HPV infection and HPV-driven OPC [71]. Seropositivity of another HPV protein, HPV16 L1, in the absence of HPV16 E6 seropositivity, is considered a marker of past infection [72-74]. In the same study as above, Anantharaman et al. found that even those with L1 seropositivity (i.e. a historic, perceivably cleared infection) were still at almost a 9-fold risk of OPC (OR: 8.96; 95% CI: 5.27 to 15.23) vs those that were L1 seronegative. D'Souza et al. estimated presence vs absence of the HPV16 L1 protein to confer an OR of 32.2 (95% CI: 14.6 to 71.3) for risk of OPC [23]. Furthermore, Mork et al. estimated the presence of HPV 16 L1 antibodies in pre-diagnostic serum samples to affect a 14.4-fold increased risk of OPC. The pre-diagnostic samples preceded OPC by over 10 years, providing evidence for a temporal association between HPV infection and OPC.

HPV-driven OPC as a unique biological entity

There is a growing body of evidence to suggest that HPV-driven OPC is a distinct biological and clinical entity compared to HPV-negative OPC (commonly associated with smoking and alcohol consumption). Firstly, HPV appears to affect younger populations [75]. In a retrospective analysis of stage III and IV OPC patients enrolled in a randomized controlled trial comparing radiotherapeutic methods, Ang et al. found the median age of a HPV-positive OPC patient to be 53.5 years (N = 206), compared to the median age of a HPV-negative OPC patient of 57 years (N = 117) [76]. Additionally, HPV-positive patients typically show less exposure to tobacco and alcohol. In a cohort study of 356 newly-diagnosed OPC patients by Dahlstrom et al., HPV-positive OPC patients showed markedly lower numbers of current smokers at diagnosis compared to those who were HPV-negative (19% vs. 51%; $P < 0.001$) and presented fewer individuals with a history of >10 pack years of smoking (37% vs. 76%; $P < 0.001$). A case-control study by Gillison et al. also found that, compared with subjects who neither smoked tobacco nor drank alcohol, those with heavy use of tobacco (≥20 pack-years) and alcohol had an increased risk of HPV-16-negative HNSCC (OR: 4.8; 95% CI: 1.8 to 12) but not of HPV-16-positive HNSCC (OR: 0.67; 95% CI: 0.29 to 1.9). Interestingly, HPV-driven OPC appears to confer a survival advantage over "traditional" OPC. Ang et al. found, in the same randomized controlled trial mentioned above, that HPV-positive patients had improved 3-year survival vs HPV-negative patients (82.4%, vs. 57.1%; $P<0.001$ by the log-rank test). Furthermore, after adjustment for age, race, tumor and nodal stage, tobacco exposure, and treatment assignment, HPV-positive OPC patients had under half the risk of mortality vs those that were HPV-negative (HR: 0.42; 95% CI: 0.27 to 0.66).

Summary of HPV as a risk factor for OPC

HPV is established in observational literature as one of the predominant risk factors for OPC. Given the huge ORs conferred by current and even historic HPV infection on OPC, in addition to the clear demographic and aetiological differences between those with HPV-driven OPC and those with "traditional" OPC, there appears to be strong observational evidence for a marked effect of HPV16 infection on risk of developing an aetiologically-distinct cancer. The paucity of risk estimates for current HPV16 infection (proxied by HPV16 E6 seropositivity), in conjunction with the abnormally large estimates seen in current observational literature, necessitate investigation into the establishing the true causal effect of this risk factor on OPC.

Socioeconomic position

In observational epidemiology, socioeconomic position (SEP) appears to be associated with a global increased risk of OPC independently of major established OPC risk factors (smoking, alcohol consumption and HPV infection). A systematic review assessing the impact of low SEP on the global risk of oral cancer (oral cavity cancer and OPC combined) (41 studies; 15,344 cases and 33,852 controls) found low SEP to be significantly associated with increased oral cancer risk in high- and lower-income countries across the world [77]. This association remained when adjusting for potential confounders, including: method of SEP measure, age, sex, global region, development index (as classified by the World Bank), time period and lifestyle factors (including smoking and alcohol consumption). Compared with individuals who were in high SEP strata, the pooled ORs for the risk of developing oral cancer were 1.85 (95% CI: 1.60 to 2.15; N studies: 37) for individuals with low educational attainment, 1.84 (95% CI: 1.47 to 2.31; N studies: 14) for those with low occupational social class and 2.41 (95% CI: 1.59 to 3.65; N studies: 5) for people with low income. As a systematic review, the above represents a high level of observational evidence for the association of SEP with OPC. However, findings from this research relate to a combination of OPC and oral cavity cancer, thus the exact effects of SEP measures on the risk of OPC as a unique HNC sub-type may differ from those reported.

At a regional level, a study investigating the socioeconomic factors associated with risk of upper aerodigestive tract cancer (cancer of the lip, tongue, major salivary glands, gums and adjacent oral cavity tissues, floor of the mouth, tonsils, oropharynx, nasopharynx, hypopharynx and other oral regions, nasal cavity, accessory sinuses, middle ear, and larynx) in Europe found educational attainment to be significantly associated with risk of oral cancer incidence [78]. Educational attainment levels were recorded in five categories: no education, primary, secondary,

further/technical and university. (Further and technical education is education beyond secondary level and includes further and technical colleges.) This was further grouped into three broad educational levels: primary (no education/primary); secondary and tertiary (further/technical/university). Compared to those with a tertiary education, primary education was associated with an 81% increase in risk of incident oral cancer (OR: 1.81; 95% CI: 1.46 to 2.23). Additionally, when compared to tertiary education in Europe, adults only possessing a primary education in the British Isles were associated with risk of upper aerodigestive tract cancer (oral cavity, pharynx [excluding nasopharynx], larynx and oesophagus) in men with an OR of 19.88 (95% CI: 2.55 to 154.94) and in women with an OR of 3.07 (95% CI: 0.61 to 15.42). Whilst this study adds to the evidence for SEP affecting risk of OPC (this time in Europe, via a large multi-centre study), as above, it does not examine SEP against OPC as a unique HNC sub-type.

In the UK, the association between socioeconomic factors and incidence of upper aerodigestive tract cancers has been assessed in the Scottish Longitudinal Study; a large-scale linkage study including: census data from 1991 onwards, vital events data (births, deaths, marriages), NHS Central Register data (gives information on migration into or out of Scotland), and education data (including Schools Census and SQA data) [79]. When examining the interrelationship between deprivation and educational attainment on risk of upper aerodigestive tract cancers, comparing individuals from affluent areas with a diploma/higher education to individuals from a deprived area with no education, men had a >2-fold increase in risk of upper aerodigestive tract cancer (OR: 2.10; 95% CI: 1.55 to 2.84) and women had a 64% increase in risk (OR: 1.64; 95% CI: 1.09 to 2.49). At a global European, and country-specific (Scotland) level, it appears that the risk of SEP on OPC has only been established as part of a broader combination of HNC sub-types. Accordingly, whilst these cancers may share some common risk factors, the risk of SEP on OPC specifically may differ from these estimates; they are not the same entities.

The evidence above provides a basis for investigation of macro-environment associated with low SEP against risk of OPC. Low educational attainment, restrictions in healthcare, poor hygiene, poor nutrition, type of profession and poor environment may all affect risk of OPC through complex societal interactions in synergy with risk behaviours (e.g. smoking, alcohol consumption, sexual behaviour) often shown by low SEP groups.

BMI

BMI appears to show moderate evidence of an inverse association with OPC in epidemiological literature, with lower BMI greatly increasing OPC risk. A 2003 case-control study of 304 cases and 304 controls by Nieto et al., lifetime BMI was investigated against risk of oral cancer and OPC (combined) in a European population (Spain) [80]. The authors found that both low weight (≤65kg) and low BMI (≤22kg/m^2) at diagnosis showed substantial increases in OPC risk (weight OR: 3.57; 95% CI: 2.32 to 5.51, BMI OR: 3.64; 95% CI: 2.27 to 5.82). Interestingly, these findings were also established 2 years prior to diagnosis, displaying equally large risk estimates (weight OR: 2.96; 95% CI: 1.93 to 4.53, BMI OR: 3.31; 95% CI: 2.04 to 5.39). Despite the presence of pre-diagnostic information, phenotype derivation was based on patient self-report questionnaires responses for patient height and weight. Accordingly, recall bias and measurement error may be a source of significant bias, decreasing the level of evidence these findings represent. Furthermore, these findings represent the effect of self-report BMI on a combination of oral cancer and OPC. As such, there is ambiguity with regard to the effect of weight and BMI on OPC specifically.

A 2015 paper by Maasland et al., using the Netherlands cohort study to investigate BMI on risks of sub-types of HNC, found that a BMI of <18.5 kg/m^2 at diagnosis (vs those with a BMI between 18.5 kg/m^2 and 25 kg/m^2) conferred a risk ratio for hypopharyngeal and OPC (combined) of 4.96 (95%CI: 1.34 to 18.33) [81]. BMI of greater than 25 kg/m^2 did not affect risk of these cancers. However, despite use of a population cohort study, BMI was again measured using self-report questionnaire, conferring the same risk of bias as Nieto et al. Furthermore, the authors combined OPC and hypopharyngeal cancer, preventing the quantification of an OPC-specific BMI risk estimate. Lastly, the number of cases in this study with a BMI <18.5 kg/m^2 was 3, with 44 cases as the 18 kg/m^2 to 25 kg/m^2 reference. The small number of cases and controls in this study greatly increases the likelihood of a type 1 error affecting the results, thus decreasing the confidence with which any findings can be established as true.

Finally, a study of BMI and risk of HNC in a pooled analysis of 17 case-control studies from the INHANCE consortium reports OPC risk estimates of 2.64 (95% CI: 2.05 to 3.39) for a BMI of ≤18.5 kg/m^2 (vs a BMI of >18.5 kg/m^2 to 25 kg/m^2; 215 cases, 358 controls) [82]. Interestingly, higher BMI categories show an inverse association with risk of OPC, including those with a BMI exceeding the obese index of 30 (BMI >25 kg/m^2 to 30 kg/m^2 OR: 0.49; 95% CI: 0.40 to 0.60; 845 cases and 5,714 controls, BMI >30 kg/m^2 OR: 0.42; 95% CI: 0.32 to 0.56; 302 cases and 2,211 controls). The same findings are also observed 2 to 5 years prior to diagnosis (BMI ≤18.5 kg/m^2 OR: 1.99; 95% CI: 1.03 to

3.84; 74 cases and 187 controls, BMI >25kg/m^2 to 30 kg/m^2 OR: 0.55; 95% CI: 0.44 to 0.70; 374 cases and 1,855 controls, BMI >30kg/m^2 OR: 0.46; 95% CI: 0.28 to 0.75; 139 cases and 867 controls). This study presents evidence for the effect of BMI on OPC as a distinct HNC sub-type and possesses a large sample size, increasing the reliability of evidence provided. However, it still relies on self-report questionnaire data to derive BMI, thus introduces recall and measurement biases to the risk estimate given. Finally, a key issue for all of the above observational studies is that reverse causality cannot be excluded. Whilst it is plausible that low BMI increases risk of OPC, it is equally plausible that OPC carcinogenesis and progression causes a reduction in BMI.

In summary, low BMI appears to be consistently associated with risk of OPC (or a combination of OPC and another HNC sub-type). However, the combination of OPC with other sub-types and small sample sizes in an observational design may lead to effect estimates which are biased by multiple factors, including recall bias and measurement error from self-report questionnaire use (affecting the precision of BMI measurement), reverse causation and a lack of precision (alluded to above). Accordingly, a directional, robust causal estimate for BMI on OPC risk has yet to be determined.

1.1.3. Clinical impact of OPC

Diagnosis of OPC

OPC often presents with nonspecific symptoms, resulting in a high number of late-stage diagnoses and highlighting the need for primary prevention of this disease. A common presentation of an OPC patient is a painless neck lump, with few other symptoms. Other possible symptoms may include a sore throat or tongue, otalgia (ear pain), pain and/or difficulty swallowing and/or a change in voice quality. The United Kingdom Multidisciplinary Guidelines for OPC [83] recommend the following four practices for assessment of the disease. **Box 1.1** describes the various techniques mentioned by these practices.

1. Cross-sectional imaging is required in all cases to complete assessment and staging
2. Magnetic resonance imaging (MRI) is recommended for primary site and computed tomography (CT) scan for neck and chest
3. Positron emission tomography (PET) combined with CT scanning (PET-CT) is recommended for the assessment of response after chemoradiotherapy, and has a role in assessing recurrence
4. Examination under anaesthetic is strongly recommended, but not mandatory

Box 1.1 - Description of techniques used in the UK Disciplinary Guidelines for management of OPC

Cross-sectional imaging methods employed in the diagnosis of OPC include MRI, CT, ultrasound and PET (typically combined with CT scanning). MRI employs strong magnetic fields and radio waves to create an image of human anatomy. Furthermore, MRI can be conducted with the inclusion of a contrast dye injection (gadolinium contrast media) to make tissue and blood vessels show up in greater detail. Consequently, MRI with contrast is considered optimal for staging the primary (origin) oropharyngeal tumour; the tumour and soft tissue spread from it can be visualised, allowing the full extent of the primary to be viewed.

CT scans combine a series of X-ray images to create cross-sectional images (sometimes 3-dimensional) of bones, blood vessels and tissue within the body. As such, CT scans are particularly useful in assessing the extent of nodal metastases and bony invasion; a vital component of determining cancer stage (see Staging, below). Distant metastases should be assessed by CT scanning of the chest and upper abdomen, to exclude metastatic disease to the lungs and liver.

Positron emission tomography (PET) involves injection of a radioactive sugar (a radiotracer) into the body, which is detected by a scanner in order to visualise metabolic processes. Cancers typically require more energy than healthy tissue, thus the cancer may be visualised by a detected concentration of radiotracer in the body. Fluoro-deoxy-glucose (FDG) is a radiotracer commonly used in oncological PET scans. Supported by the results of the UK PET-Neck randomised controlled trial (RCT) study, PET using FDG, combined with CT scanning (F-FDG PET–CT) is recommended for the assessment of an individual's response to treatment approximately three months post-chemoradiotherapy, particularly in patients with advanced nodal disease. F-FDG PET–CT scanning also has a role in the assessment of recurrent tumours and can detect recurrence at primary sites, neck nodes and/or distant metastases due to it allowing visualisation of contrasting (healthy vs cancer) metabolic rates.

Finally, ultrasound is a form of imaging which uses high-frequency sound waves to create an image of internal anatomy. Ultrasound can be used to guide a needle to conduct a histological biopsy (fine needle biopsy), or without a needle biopsy as a direct visualisation of an OPC tumour. According to the UK Multidisciplinary Guidelines for OPC, ultrasound should be carried out for all patients presenting with a neck lump and is an accurate method of staging nodal disease in experienced hands.

Staging of OPC

Pre-treatment staging for primary oropharyngeal tumours is based on the tumour–node–metastasis (TNM) classification (8th edition) [84], shown in **Box 1.2**. Each letter describes a different aspect of tumour biology: "T" describes the size of the tumor and any spread of cancer into nearby tissue, "N" describes spread of cancer to nearby lymph nodes and "M" describes metastasis. This system was created and is updated by the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC).

Box 1.2 – Details of TNM staging criteria for OPC

TNM staging for oropharyngeal squamous cell carcinoma
TX: Primary tumour cannot be assessed
T0: No evidence of primary tumour
Tis: Carcinoma in situ
T1: Tumour 2 cm or less in greatest dimension
T2: Tumour larger than 2 cm but 4 cm or less in greatest dimension
T3: Tumour larger than 4 cm in greatest dimension or extension to lingual surface of epiglottis
T4a: Tumour invades the larynx, deep/extrinsic muscle of tongue, medial pterygoid, hard palate or mandible
T4b: Tumour invades lateral pterygoid muscle, pterygoid plates, lateral nasopharynx, or skull base or encases carotid artery

Quality of life for OPC patients

Quality of life (QoL) in HNC is commonly assessed via self-report questionnaire [85-87]. Despite only being the 8th most common cancer worldwide, HNC development and treatment are known to cause a substantial deterioration in a patient's QoL [88]. Current OPC treatment modalities include external beam radiation therapy (EBRT: high-energy x-ray beams are targeted at the tumour from outside the body to destroy cancer cells), chemoradiation therapy (CRT: combined chemotherapy and radiation therapy to improve disease response to treatment [89]), altered-fractionation radiation therapy (AFRT: the use of varied, frequent, short doses of radiotherapy [90]), intensity-modulated radiation therapy (IMRT: a type of radiotherapy that closely conforms to the shape of a tumour to minimise radiation of healthy neighbouring tissue [91] - sometimes this modality is combined with chemotherapy), surgery or brachytherapy (BT: treatment which places sealed radioactive sources inside the patient to destroy tumour cells [92]). The above treatments impact QoL of OPC patients differentially, affecting swallowing, pain, speech, salivation, social and emotional factors [93]. Consequently, social and personal intimacy, nutrition, biological function and self-esteem (and depression) can be substantially affected, providing a strong rationale for development of prevention strategies for OPC.

1.1.4. OPC prognostication

Current prognostication for OPC in clinical practice is based on a combination of HPV status and TNM staging [94]. However, despite the prognostic value added by taking HPV status into consideration, subsets of patients continue to demonstrate outcomes discordant with their disease

stage [95]. The prognostic importance of molecular biomarkers and clinical features separate from TNM staging and HPV status have become increasingly apparent in epidemiological literature [96]. As such, there remains a need for accurate risk stratification that combines existing prognostic factors with novel molecular biomarkers and other personalized features. With more accurate prognostication in patients with OPC, patient treatment and management can be optimised and prioritised. Below, prognostic indicators of OPC will be explored.

Prognostic factors for OPC

Perhaps the most reported prognostic factor other than TNM stage for OPC in epidemiological literature is HPV. The substantial survival advantage seen by those with OPC who are HPV-positive compared with HPV−negative has led to the 8th edition TNM classification for these cancers to include p16^{INK4A} immunostaining as a proxy for HPV-positivity, resulting in a lower TNM stage for these tumours compared to previous editions [84]. Furthermore, clinical trials are beginning to investigate the effect of de-escalation of treatment for HPV-positive OPC. One proposal for de-escalation of treatment of this sub-group is that, to reduce toxicity, radiotherapy in conjunction with cisplatin may be able to be replaced with radiotherapy in conjunction with cetuximab. However, two recently published trials, De-ESCALaTE HPV (Determination of Epidermal growth factor receptor-inhibitor [cetuximab] versus Standard Chemotherapy [cisplatin] early And Late Toxicity Events in Human Papillomavirus-positive oropharyngeal squamous cell carcinoma) [97] and RTOG-1016 (Radiation Therapy With Cisplatin or Cetuximab in Treating Patients With Oropharyngeal Cancer) [98], have found that attempting to de-escalate treatment with cetuximab and radiotherapy vs cisplatin and radiotherapy unexpectedly causes a decrease in overall survival. Whilst these findings clearly highlight the need for caution in de-escalation trial design and operation, they also demonstrate the high level of evidence established for the improved prognostic value of HPV-positivity in OPC.

In 2015, Keck et al. discovered that HPV-positive HNC could be stratified depending on gene expression patterns [99]. Using gene expression-based consensus clustering, the authors found that HPV-positive HNC (including an OPC-only analysis) could be classified into 2 groups based on expression of *CD8A/B;* a marker of enrichment of cytotoxic T-cell infiltration. Zhang et al. were also able to stratify HPV-positive HNCs by gene expression, characterised by either by elevated immune response and mesenchymal differentiation (named HPV-IMU), or by elevated keratinocyte differentiation and oxidation-reduction process (named HPV-KRT) [100]. Expression analyses showed that HPV-KRT tumours showed a higher frequency of integrated HPV in genic regions and had higher levels of a spliced variant of HPV E6 (denoted E6* and associated with oxidative stress [101]) compared

to the non-spliced E6 variant (which is reported to downregulate a large number of genes involved in keratinocyte differentiation, and upregulate genes normally expressed in mesenchymal lineages [102]). Interestingly, HPV-IMU tumours were enriched for chromosome 16q losses compared to HPV-KRT tumours; shown to correlate with improved survival in OPC [103]. Moreover, HPV-IMU tumours displayed improved immune response compared to HPV-KRT, which has been shown to confer a survival advantage amongst those with HPV-positive cancer [104]. Given the divergence between the HPV-IMU and HPV-KRT subgroups, different treatment strategies may improve prognosis. Zhang et al. hypothesise that the HPV-IMU subgroup may benefit more from immunotherapies and treatment targeting epithelial-to-mesenchymal transition, whilst HPV-KRT may benefit from strategies to induce tissue-based inflammation. In light of unexpected de-escalation trial findings and a lack of personalised medicine for OPC, the above unique expression profiles offer important translational findings.

Other prognostic factors to have been shown to derive effective prognostic models in HPV-positive OPC, highlighting their prognostic importance. Ward et al. developed a prognostic model for HPV-positive tumours, consisting of a combination of tumour-infiltrating lymphocyte levels, heavy smoking, and T-stage [105]. Findings were validated by area under a receiver-operator-characteristic (ROC) curve (AUC: a measure between 0 and 1 of the sensitivity and specificity of a predictive model, where 1 is perfect prediction) and found to show good predictive value in an independent 'testing' cohort (AUC: 0·82). In agreement with current literature, Ward et al.'s findings suggest that immune response (in this instance shown by tumour-infiltrating lymphocyte levels) plays an important role in the improved survival seen in most HPV-positive patients and is relevant for the clinical evaluation of HPV-positive OPC. Notably, this model also features heavy smoking as a prognostic factor, highlighting the importance of this behaviour on OPC outcomes, independently of HPV infection.

In a prospective study to identify markers of response to therapy to prevent organ loss in advanced OPC, Kumar et al. report low *EGFR* expression and high p16 expression (highly correlated with HPV infection in HNC [106]) to be markers of good response to organ-sparing therapy (induction chemotherapy and chemoradiotherapy), overall survival and disease-specific survival [107]. Conversely, high *EGFR* expression, combined low p53 and high antiapoptotic protein Bcl-xL expression (associated with chemotherapy and radiation resistance [108], in addition to cisplatin resistance [109]), female sex and smoking were associated with a poor outcome. The authors found EGFR expression to be associated with current smoking (*P*: 0.04), female sex (*P*: 0.05), and lower HPV titre (*P*: 0.03). Given the findings of this study, cessation of smoking, stratification by *EGFR* expression,

determination of HPV status and stratification by combined p53/Bcl-xL expression are key considerations when determining the prognostic outcome, intensity of treatment and modality of treatment for OPC patients.

In summary, HPV status and smoking appear to be consistently, strongly associated with OPC prognosis. Interestingly, alcohol consumption did not appear as a major prognostic indicator in any of the models described above. Within HPV-positive tumours, there appears to be sub-categories of OPC which show gene expression profiles associated with improved immune response (HPV-IMU) or keratinocyte differentiation (HPV-KRT). These tumours have different prognostic outcomes and response to therapy; thus, care should be taken to discern the sub-categories of HPV-positive OPC when determining treatment strategies, particularly when considering de-escalation of treatment. Finally, expression of *EGFR*, p53 and Bcl-xL appear to also show both an interplay with HPV status and smoking, in addition to independent effects on survival and treatment resistance.

1.2. Genetics of oropharyngeal cancer

1.2.1. Introduction to genetics of OPC

The genetic contribution to OPC carcinogenesis describes an accumulation of genetic variations (via genetic mutation, amplification or translocation) in proto-oncogenes and tumor-suppressor genes, which cause genetic instability and cumulatively lead to OPC. An oncogene is created by mutation (typically a gain-of-function mutation) of a proto-oncogene, which alters the original function of the proto-oncogene to affect transformation of normal cells into cancer cells. Proto-oncogenes typically encode proteins such as growth factors and their receptors, signal-transduction proteins, transcription factors and cell cycle control proteins. Gain-of-function of these proteins in oncogenes results in rapid, abnormal cell growth seen in cancer. Conversely to proto-oncogenes, tumor-suppressor genes generally produce proteins which arrest and monitor cell proliferation; loss-of-function mutations in these genes promotes oncogenesis. Tumor-suppressor genes inhibit cell cycle progression at specific stages, inhibit general cell proliferation, stop cell cycle progression if DNA is damaged/abnormal, promote apoptosis and produce enzymes which aid in DNA repair. Loss-of-function of tumor-suppressor genes, coupled with oncogene activation, result in a severely dysregulated cell cycle, allowing rapid proliferation of cells containing abnormal DNA. Identifying factors which affect the genetic stability of a healthy oropharynx cell to promote oncogenesis could vastly increase understanding and prevention of these cancers. Three of the most

common genetic causes of tumorigenesis are chromosomal instability and copy number variation. These will be discussed below in the context of OPC.

1.2.2. Chromosomal instability

Chromosomal instability is a genetic variation, recognized as a hallmark of cancer [110], which affects chromosome structure and number in cells. Chromosomal instability is typically classified as either numerical or structural instability in solid tumors - structural instability involves gain or loss of part of a chromosome, whereas numerical instability involves gain or loss of entire chromosomes (known as aneuploidy) [111]. Although poorly understood, chromosomal instability is thought to result from defects in mitotic processes (**Figure 1.12**) [112].

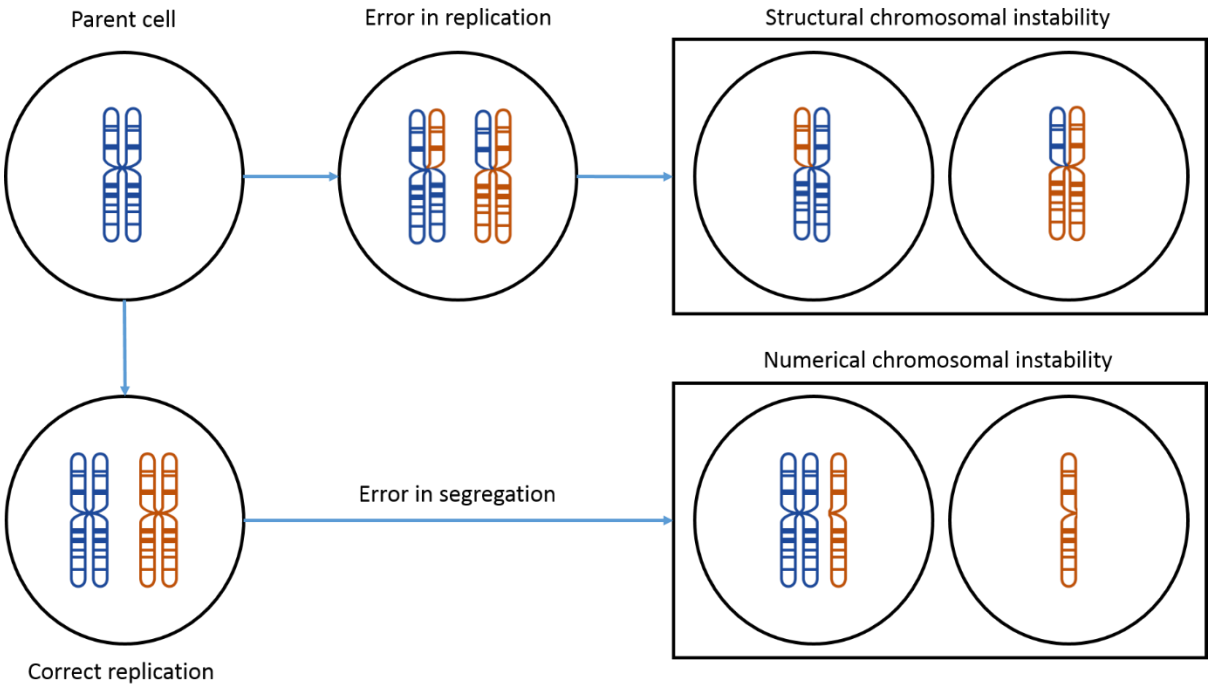


Figure 1.12 - The process of chromosomal instability via defects in mitotic processes. Incorrect replication is thought to cause structural instability and errors in segregation are thought to affect numerical instability.

One method of quantifying chromosomal instability is via fluorescence in-situ hybridization (FISH). FISH is a cytogenetic technique which implements fluorescent DNA probes to localize positions within a chromosome, through use of highly complementary sequences and a fluorescent microscope. A 2010 study by Sato et al. used FISH to investigate fine-needle aspiration (FNA – a procedure where a thin, hollow needle is used to retrieve cells) biopsy samples for chromosomal instability in oral squamous cell carcinoma (OSCC) [113]. This study investigated the degree of chromosomal instability

(CIN grade) in 77 OSCCs, choosing to examine chromosomes 7, 9 and 11. CIN grade was defined according to the proportion of cells within the tumor that had copies of chromosomal genetic information which differed to (greater or less than) the modal average [114]. A three-tier scale was used: CIN1 was defined as < 20% of the cells in the tumor differing from the modal average, CIN2 as ≧ 20% and < 40% of the cells, and CIN3 as ≧ 40% of the cells differing from the modal average. Using this method, Sato et al. established that CIN grade could "be useful in predicting of recurrence and poor prognosis in patients with oral SCCs". CIN3 was seen in 11.7% (9/77) OSCC tumors and showed a significant association with reduced disease-free 5-year survival (*P:* 0.008). Furthermore, CIN grade was associated with a poor outcome in disease-free 5-year survival (HR: 3.48; 95% CI: 1.10 to 11.1; P: 0.035) and overall 5-year survival (HR: 3.71; 95% CI: 1.06 to 13.0; P: 0.041) in Cox proportional-hazards analyses adjusting for age, sex, cellular differentiation and disease stage.

In the context of OPC, chromosomal instability has been investigated in relation to smoking and HPV16 infection. A 2019 study by Villepelet et al. investigated the effects of tobacco abuse on major chromosomal instability in HPV16-positive OPSCC using an array-based comparative genomic hybridization approach [115]. Examining 50 OPSCC patients, Villepelet et al. discovered that HPV-positive tumors had fewer genomic aberrations (P=0.008) and fewer breakpoints (P=0.048) than HPV-negative tumors. The authors confirmed an association between HPV-positive OPSCC and chromosomal losses at 11q. They also reported an association between HPV-negative OPSCC, losses at 3p and 9p and gains at 7q and 11q13. In the patients with OPSCC who were HPV-positive, the total number of chromosomal aberrations per tumor was significantly higher in the group of patients who were smokers (P=0.003). In support of these findings, recent *in-vitro* studies investigating the effect of HPV16 on chromosomal instability have found the two to be inversely associated, finding that high-risk HPV sub-types with a high viral immortalization capacity (HPV16/18/31/33/35) can immortalize a cancer cell much faster than other HPV types. Instead, it is thought that high-risk HPV with a lower viral immortalization capacity (HPV45/51/59/66/70) requires DNA to undergo many more aberrations to achieve the same immortalized state.

1.2.3. Copy number variation

As a tumor cell acquires genetic mutations, the number of copies of chromosomal regions or sections of genes can change (most commonly via deletion and duplication), known as copy number variation. Affected genes can include oncogenes, tumor suppressor genes and other genes associated with genomic instability. Few studies investigate these copy number variations specifically in the context of OPC. However, one study by Zagradišnik et al. investigated copy number variations which

were shared across oral cavity, hypopharyngeal cancer and OPC [116]. Gains of genomic regions from the long arm of chromosome 3 (3q), containing *PIK3CA* and *AGTR1* genes, were more frequent in cancer cases with lymph node involvement vs those without affected lymph nodes. In addition, cancer cases not treated with surgery were associated with gains of band 21 on the long arm of chromosome 7 (7q21) and gains of the entire long arm of chromosome 20 (20q) vs those cancer cases that were not [116]. Interestingly, Suda et al. found that copy number amplification of *PIK3CA* was associated with HNC relapse (which occurs in lymph nodes in most cases [117]) in those who did not have lymph node involvement [118]. This finding may partly explain the increase in copy number of *PIK3CA* seen by Zagradišnik et al.; *PIK3CA* copy number increase may be associated with relapse in individuals without lymph node involvement, causing the observed associated with lymph node involvement. In contrast to Zagradišnik et al. and Suda et al., Resteghini et al. found that an increase in copy number of *PIK3CA* genes was associated with favourable clinical outcome in HPV-negative OPC [119]. Despite being a known oncogene, the specific prognostic role of *PIK3CA* copy number in OPC is currently unclear.

1.2.4. Conclusion

OPC shows a higher incidence rate in developing countries due to increased tobacco and alcohol consumption. These risk behaviors are known to increase the rate of genetic mutation and could bring about carcinogenesis by disrupting the genomic stability of a healthy cell. The gene regions discovered by investigation of chromosomal instability and copy number variation could provide deeper insight into OPC etiology and may be useful as therapeutic targets for the disease. Greater insight is needed into the somatic mutational landscape of OPC as a unique HNC sub-type.

1.3. Epigenetics of oropharyngeal cancer

1.3.1. Introduction to epigenetics

The epigenetic contribution to cancer describes heritable changes in gene expression which occur independently of the primary DNA sequence to affect carcinogenesis [120]. Whereas genetic drivers of cancer directly affect the underlying sequence of bases in a DNA sequence, epigenetic changes regulate gene expression by affecting the access of cellular machinery to the DNA sequence. Two predominant types of epigenetic change are investigated with respect to cancer: DNA methylation and histone modification. DNA methylation involves the addition of a methyl (CH_3) group to the fifth carbon of cytosine (C) base, forming a 5-methylcytosine (5mC) molecule (**Figure 1.13**).

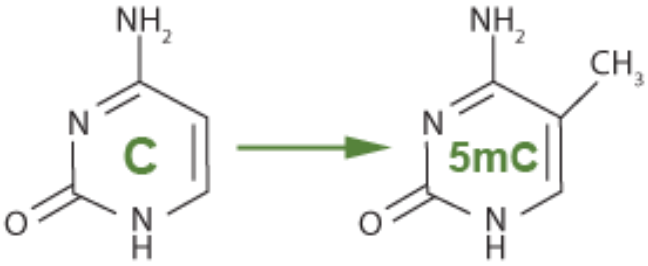


Figure 1.13 - The process of DNA methylation, as the addition of a methyl group to the fifth carbon of a cytosine base

This reversible process is catalysed by enzymes known as DNA methyltransferases (DNMTs), predominantly occurring (>98% in somatic cells) between cytosine and guanine bases, naturally separated by a single phosphate molecule in the DNA sequence. As such, specific sites of methylation in the genome are referred to as "CpG" dinucleotides. CpG sites typically play a central role in the silencing of gene expression, though in less common situations, methylation can also increase gene expression.

It is largely accepted that a variety of environment and lifestyle factors (diet, behaviour, stress, physical activity, working habits, alcohol and tobacco consumption, etc.) can have a negative impact on health, and contribute to the development of a large array of diseases (including that of cancer). Synergistically with other epigenetic mechanisms, DNA methylation allows cells (and by extension, organisms) to adapt to these factors at a speed that mutational mechanisms cannot. Thus far, DNA methylation has been associated with numerous cellular processes, including transcriptional repression, X chromosome inactivation, cell differentiation, genomic imprinting, alteration of chromatin structure and transposon inactivation. If dysregulated, the adaptation DNA methylation confers can be disrupted, potentially mediating the development of diseases such as cancer. DNA methylation has been the most measured epigenetic mark because of its obviously fundamental biological interest, its mitotic stability, the availability of methods for its quantification (globally or in targeted regions), its stability during DNA extraction and purification procedures, and its durability in archival biological materials.

The second type of epigenetic modification commonly investigated is histone modification. Histones are proteins which affect the compaction of DNA, which provide structural support and affect how available DNA is for transcriptional activation. A histone "core" consists of 8 histone proteins (2 H2A, 2 H2B, 2 H3 and 2 H4 proteins), around which DNA is wrapped. DNA which is wrapped around 8

histone cores (each of 8 histone proteins) is known as a nucleosome and is considered a basic unit of chromatin. Histone proteins have an outward-facing "tail" of amino acid residues, which are positively charged. The positive charge of histone protein amino acid tails interacts with the negatively charged phosphate groups in DNA to form tightly compacted chromatin. To relax the compaction of DNA, a process known as histone acetylation modifies the charge of histone protein amino acid tails. Facilitated by enzymes known as histone acetyltransferases, acetyl groups are transferred from available acetyl coenzyme-A molecules to NH_3^+ groups of lysine amino acids of histone protein tails, neutralising the overall positive charge. This reduces the degree to which histone cores are bound to DNA, relaxing a nucleosome, thus making it more available for transcriptional machinery to access. The process can be reversed, known as histone deacetylation, via facilitation by enzymes known as histone deacetylases.

1.3.2. Example of blood-based methylation changes in OPC

Tumour suppressor activity has been shown to be associated with changes in DNA methylation in OPC. The galanin gene codes for a neuropeptide which regulates anterior pituitary hormone secretion and acts as neurotransmitter [121]. GALR1 and GALR2, two receptors for GAL, are members of the G protein-coupled receptor (GPCR) superfamily. Galanin and these receptors regulate cell growth by inhibiting extracellular signal-regulated kinase (ERK) 1/2, upregulating cyclin-dependent kinase inhibitors p27 and p57, and decreasing expression of cyclin D1 [122]. These cellular processes work to arrest the cell cycle. Hypermethylation of these genes causes them to lose their tumour suppressor activity. Misawa et al. found that, in a study of HNC methylation patterns, 13 out of 20 OPC cases (65%) had hypermethylation of at least one of the above genes [123].

1.3.3. Example of saliva-based methylation changes in OPC

Saliva is an emerging diagnostic medium for OPC which may prove a lucrative resource for examination of shed tumour DNA, in addition to being a non-invasive collection technique. Furthermore, saliva may prove to be a tissue proxy for the tonsillar crypt DNA, which are difficult to extract DNA samples from. Genes observed to be hypermethylated in saliva samples in OPC in wider literature are listed in **Table 1.1**.

Table 1.1 - List of genes found hypermethylated in saliva samples and their function

Gene	Gene product	Function
CCNA1	Cyclin A1	Cell cycle regulation
DAPK1	Death-associated protein kinase 1	Apoptosis and autophagy
ERCC1	Excision repair cross-complementation group 1	Repair of DNA damage induced by ultraviolet light or cisplatin
TIMP3	Metalloproteinase inhibitor 3	Inhibitor of the matrix metalloproteinases

1.4. Summary, research gaps and aims of this thesis

OPC is an increasingly prevalent disease in the UK, affecting greater proportions of younger individuals. Epidemiological literature surrounding risk factors for OPC incidence provide observational estimates, with most studies combining OPC with other HNC sub-types. It should also be noted that a vast amount of epidemiological literature has been published investigating the same three risk factors. Observational epidemiological literature is prone to bias from confounding and reverse causation; coupled with few studies investigating OPC exclusively, it is difficult to ascertain the true causal effect of reported risk factors on OPC. Genetic epidemiology can circumvent the common pitfalls of observational epidemiology by proxying these phenotypes with genetic variants, in an approach known as Mendelian randomization (MR; see Chapter 2). Furthermore, few epigenetic studies of OPC exist (indeed, a paucity in relation to OPC prognosis/mortality) – those that do display low sample sizes with a lack of replication of findings. Larger sample sizes, repeated findings and causal inference methods are necessary to elucidate and appraise legitimate causal pathways, potentially improving our knowledge of risk factor and prognostic pathways in OPC. The aims of this thesis are therefore as follows:

1) Retrieve existing risk factors from observational epidemiological literature, via literature mining, to form an evidence base for downstream analyses and identify novel areas of investigation
2) In a hypothesis-free approach, appraise risk factors against OPC incidence in a robust causal inference framework (Mendelian randomization) using genetic data. Risk factors currently reported in observational literature will be prioritised

3) Use epigenetic data from whole-blood to investigate whether DNA methylation is associated with OPC risk factors, and if so, whether DNA methylation can be used to interrogate biological pathways associated with these risk factors. Furthermore, ascertain whether DNA methylation is associated with OPC mortality, and if so, whether shared methylation patterns with risk factors can be used to appraise causal pathways between the two
4) Exploit the utility of whole-blood methylation as a biosocial archive to derive epigenetic predictors of risk factors and mortality, to ascertain whether epigenetic predictors of the above have add value to risk stratification

CHAPTER 2. INTRODUCTION TO METHODOLOGY

2.1. Introduction

This chapter will introduce the basic methodological principles which underpin the thesis, from which more applied approaches are derived and applied in later results chapters. Core methods are described, then classified according to how they supplement the thesis workflow, in **Figure 2.1**. Additionally, the datasets utilised throughout this thesis are introduced and described, detailing which methods they are applicable to.

We are currently in an era where a vast amount of genetic and epigenetic data are publicly available, as well as a multitude of bioinformatic techniques to curate them. These data can be used (in both a hypothesis-generating, and subsequent hypothesis-driven manner) to improve elucidation and understanding of the causal pathways of OPC incidence and progression. Using a combination of hypothesis-generating methodology, including genome-wide and epigenome-wide association studies (GWAS and EWAS, respectively), enriched literature object mining and MR analyses (hypothesis-free [124], two-sample [125], and two-step MR [126]), robust causal pathways associated with OPC incidence and progression may be established and appraised. This chapter highlights how the employed methods and data resources combine to form a theoretical workflow which aims to augment understanding of OPC incidence and mortality.

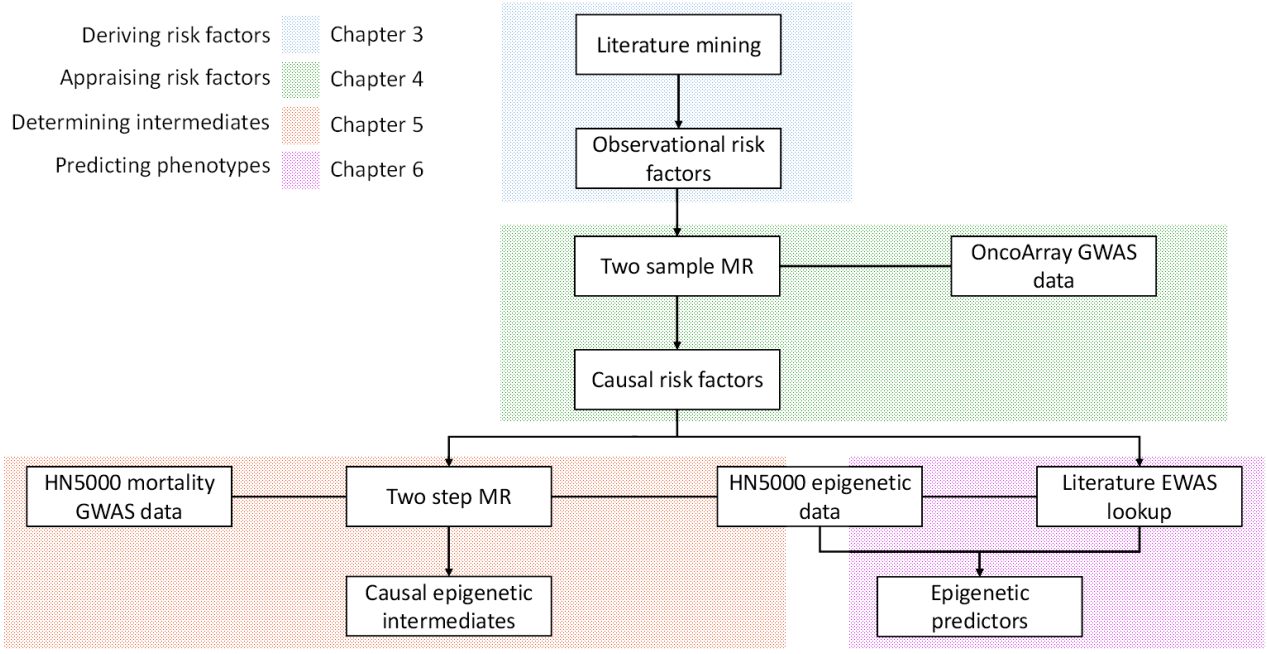


Figure 2.1 - Methodological workflow for results presented in this thesis

2.2. Methodological workflow

2.2.1. Literature mining

Published scientific literature contains a wealth of information from multiple specialist fields. Given that, at its core, science is a cumulative endeavour, possessing a solid foundation of existing knowledge to build upon and compare findings to is fundamentally important to the progression of any scientific research. Review of scientific literature allows the researcher to: identify what information already exists, identify patterns or trends in data, combine existing findings to answer a specific research question, generate novel hypotheses and identify areas where a paucity of research exists. Templier and Paré outline 6 steps necessary to write a typical review article [127]:

1. Formulating the research question(s) and objective(s)
2. Searching the extant literature
3. Screening for inclusion
4. Assessing the quality of primary studies
5. Extracting data
6. Analysing data

Despite the generic steps taken to write a review article, methodologies differ widely depending on the research question, aims and scope of the review. Nine review types comprise the vast majority of published literature reviews [128], overviewed in **Table 2.1**. These are then briefly described in **Table 2.2** with examples relevant to epidemiological literature.

Table 2.1- Nine of the most common scientific review types, the goals they seek to achieve and the methods they employ to achieve them

Overarching goal	Theoretical review types	Scope of questions	Search strategy	Nature of primary sources	Explicit study selection	Quality appraisal	Methods for synthesizing/analyzing findings
Summarization of prior knowledge	Narrative review	Broad	Usually selective	Conceptual and empirical	No	No	Narrative summary
	Descriptive review	Broad	Representative	Empirical	Yes	No	Content analysis/frequency analysis
	Scoping review	Broad	Comprehensive	Conceptual and empirical	Yes	Not essential	Content or thematic analysis
Data aggregation or integration	Meta-analysis	Narrow	Comprehensive	Empirical (quantitative)	Yes	Yes	Statistical methods (meta-analytic techniques)
	Qualitative systematic review	Narrow	Comprehensive	Empirical (quantitative)	Yes	Yes	Narrative synthesis
	Umbrella review	Narrow	Comprehensive	Systematic reviews	Yes	Yes	Narrative synthesis
Explanation building	Theoretical review	Broad	Comprehensive	Conceptual and empirical	Yes	No	Content analysis or interpretive methods
	Realist review	Narrow	Iterative and purposive	Conceptual and empirical	Yes	Yes	Mixed-methods approach
Critical assessment of extant literature	Critical review	Broad	Selective or representative	Conceptual and empirical	Yes or no	Not essential	Content analysis or critical interpretive methods

Table 2.2 - Examples and rationales for each of the most common types of epidemiological review

Review type	Why is it conducted?	Example
Narrative review	To identify what has been written on a subject	Skogen and Øverland conducted a narrative review to identify information pertaining to the fetal origins of adult disease [129]
Descriptive review	To examine data for patterns/trends, against pre-existing research questions, hypotheses or focuses	Harborne et al. conducted a descriptive review of the evidence for the efficacy of metformin in polycystic ovary syndrome, to improve clinicians' knowledge of the available published clinical evidence [130]
Scoping review	To indicate the volume and nature of literature related to a particular subject	Takahashi et al. conducted a scoping review of the volume and quality of literature describing risks and supports to competence for occupational therapists, pharmacists, physical therapists and physicians [131]
(Quantitative systematic review and) Meta-analysis	To aggregate and appraise quantitative data in the form of standard effect measures (e.g. odds ratios, prevalence etc.) from two or more functionally similar studies, taking into account the relative sample size of each study	Zhou et al. conducted a meta-analysis of spider mite sensitivity, to estimate the global prevalence of allergies to spider mites, using data from 23 studies [132]
Qualitative systematic review	To search, identify, select, appraise, and abstract data from literature using narrative and more subjective (rather than statistical, above) methods to bring together the findings of included studies	Beasant et al. conducted a qualitative systematic review to determine treatment preference and recruitment in paediatric randomised controlled trials [133]
Umbrella review	To integrate information from multiple systematic reviews (qualitative or quantitative) into one accessible and usable document to address a narrow research question	Veronese et al. conducted an umbrella review to collate systematic review information on whether chocolate consumption is associated with health outcomes [134]
Theoretical review	To develop a conceptual framework or model with a set of research propositions or hypotheses	Grover et al. published a review protocol. They will evaluate the methodological and clinical aspects of Mendelian randomization studies using neurodegenerative disorders as outcome. They intend to develop an in-

Review type	Why is it conducted?	Example
		depth understanding of what can be done in future for the derivation of true causal risk factors [135]
Realist review	To inform, enhance, extend or supplement conventional systematic reviews by explaining conflicting evidence about complex interventions applied in diverse contexts, typically to inform policy decision making	Mogre et al. conducted a realist review of educational interventions (what sort of educational interventions work, how, for whom, and in what circumstances) to improve the delivery of nutrition care by doctors and future doctors [136]
Critical review	To critically analyse the existing literature related to a broad topic to reveal weaknesses, contradictions, controversies, or inconsistencies	Chang et al. conducted a critical review of the various epidemiological studies investigating Agent Orange or 2,3,7,8-tetrachlorodibenzo-*p*-dioxin and lymphoid malignancies [137]

In the context of appraising risk factors and intermediates for OPC, various stages in the 6-point list for conducting literature reviews are a source of potential subjectivity and a large time burden. Firstly, deriving a search strategy to systematically appraise all epidemiological risk factors and intermediates for OPC (above simply screening all OPC literature) would introduce subjective bias based on what a researcher deemed necessary to include as a potential risk factor. Secondly, the sheer volume of literature available for review introduces a substantial time burden when searching and screening manually. Search strategies are often refined by inefficient methodologies such as arbitrary criteria, filtering by journal impact factor, social media influence and word of mouth [138]. In addition to the introduction of bias to a literature review, "filtering" in this way also greatly reduces the number of hypotheses that can be examined. Automated, systematic text-mining may be a review approach that can augment discovery and appraisal of mechanisms of disease alongside current literature review approaches.

Semantic MEDLINE Literature Annotation Objects

Rather than extracting information from raw literature text, pre-calculated literature annotation objects can be implemented. The Semantic MEDLINE Database (SemMedDB) is a computationally-derived repository of semantic predications which utilises the Unified Medical Language System (UMLS). When scientific literature are parsed by SemMedDB, a "subject-PREDICATE-object" annotation "triple" is generated from the title and abstract [139, 140]. The "subject" and "object" are

concepts from the UMLS Metathesaurus (a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships) [141, 142], whereas the predicate is a relation term from the UMLS Semantic Network (a set of broad subject categories and relationships between them) linking them together. For example, the sentence "We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalaemia" will produce the following four triples:

- Hemofiltration-TREATS-Patients
- Digoxin overdose-PROCESS_OF-Patients
- Hyperkalaemia-COMPLICATES-Digoxin overdose
- Hemofiltration-TREATS(INFER)-Digoxin overdose

Mining Enriched Literature Objects to Derive Intermediates (MELODI)

MELODI is a hypothesis-free application that can be used to derive mechanistic pathways using SemMedDB annotation [138]. It utilises a graph database to find enriched relationships between two “search sets” of articles, with the entire MEDLINE database as a preloaded repository. Data analysis of this nature is well-suited to graph databases, as the presence of graphs (relationships) between individual articles, or “nodes”, allow for efficient retrieval of related data using a single operation. First, for a given search set (a set of articles that represent a concept, such as “smoking”, or “head and neck cancer”), the enriched elements are identified. Identification of enriched elements is based on the number of times a SemMedDB triple has been annotated within the articles in the search set, compared to the frequency of this triple in every article in the MEDLINE database. The enrichment is restricted to concepts which are present less than 150,000 times in SemMedDB; to remove generic terms that would otherwise introduce unnecessary noise. Additionally, a Fisher's exact test is performed for each element and corrected for multiple testing using Bonferroni correction [143]. The Fishers exact test determines if there are non-random associations between two categorical variables, using the following formulae:

$$(1) \quad N = \sum_i R_i = \sum_j C_j$$

$$(2) \quad P_{cutoff} = \frac{(R_1!\, R_2! \cdots R_m!)(C_1!\, C_2! \cdots C_n!)}{N! \prod_{i,j} a_{i\,j}!}$$

Let there exist two variables, X and Y, with m and n observed states, respectively. Form an m x n matrix in which the entries $a_{i\,j}$ represent the number of observations in which $x = i$ and $y = j$. Calculate the row and column sum R_i and C_j,

respectively, and the total sum of the matrix using (1). Next, calculate the conditional probability of getting the actual matrix given the derived row and column sums, given by (2). Finally, find all possible matrices of nonnegative integers consistent with the row and column sums R_i and C_j. For each one, calculate the associated conditional probability using (2), where the sum of these probabilities must be 1.

A Bonferroni correction for multiple testing can be calculated as follows:

$$\alpha' = \alpha / k$$

Assuming k independent significance tests at the α level, the probability of no significant differences in all tests is simply the product of the individual probabilities: $(1 - \alpha)^k$. For example, with α = 0.05 and k = 10 we get $p = 0.95^{10} = 0.60$. For 10 tests, there is a 40% chance that one is significant, despite each individual test only being at a 5% level (α = 0.05). In order to guarantee that the overall significance test is still at the intended α level after multiple tests, we have to adapt the significance level α' of the individual test by dividing the original alpha, α, by the number of tests, k.

Once enriched elements are identified for two search sets, overlap elements are identified. The use of SemMedDB triples allows for a high-resolution analysis which provides a direction from one search set to another, allowing for multi-step relationships, traversing from one search set to another via overlapping concepts.

Systematic retrieval of epidemiological risk factors

Implementing the enrichment stage of MELODI's comparison process (as above), it is possible to retrieve subject-PREDICATE-object triples that are enriched in a specific search set, prior to comparing them against another search set. This allows the user to parse risk factors that arise in literature, related to a search term of interest (e.g. OPC) more times than expected by chance. By filtering SemMedDB predicates to those that may describe a potentially modifiable risk factor (AFFECTS, ASSOCIATED_WITH, CAUSES, and PREDISPOSES), a scan of epidemiological observational risk factors for a specific search term can be performed, with enriched 'risk factors' for that concept determined. Additionally, two search sets can be compared against each other using MELODI; overlapping predicates between them indicate potential intermediates (which may also be independent risk factors) for greater resolution of potential causal pathways. Both approaches are of particular use in genetic epidemiology as, if a genetic proxy (see "Genome-wide Association Studies below) exists for a novel modifiable risk factor or intermediate, they can be validated using robust causal inference methods such as Mendelian randomization (also below).

2.2.2.Genome-wide association studies

A brief introduction to genome-wide association studies

A genome-wide association study (GWAS) involves comparing the genetic data of many individuals to discern potential genetic variations between them based on a phenotype of interest. GWAS most commonly investigate single-nucleotide polymorphisms (SNPs) as the genetic variation of interest. SNPs are differences at single positions in an individual's DNA sequence which occur roughly every 1000 nucleotides, resulting in around 4-5 million SNPs in the human genome of over 3 billion nucleotides. A 1999 paper by Halushka et al. estimated that 50% of SNPs occur in noncoding DNA, whereas 25% affect missense mutations in coding DNA, and a further 25% affect "silent" mutations in coding DNA [144]. SNPs causing "silent" mutations are named so because they don't change the amino acid they encode and are thought not to impact an individual's health. SNPs which do cause a change in an amino acid, however, may impact an individual's health. In a given population, SNP differences can be observed between people which influence (but are not limited to) transcriptional and translational efficacy of cell machinery, protein structure and protein function. As such, SNPs can contribute towards the biological function of an individual, phenotypic expression of tens of thousands of traits, metabolism of drugs, and the susceptibility of an individual to disease states. For example, SNPs have been associated with nicotine addiction [145], self-reported risk-taking [146], heterochromia [147], susceptibility to infection [148] and OPC [149]. With their high frequency in the genome and scope for impact on the health and biology of an individual, SNPs are an important source of genetic variation to investigate.

Since the first study applying GWAS methodology was published in 2007, the frequency of published GWAS has increased exponentially. A well-established repository of published GWAS data, the NHGRI-EBI GWAS Catalog, has grown ~40-fold since its launch in 2008 with 139 GWAS studies, to 5687 studies in 2018 [150]. Furthermore, the complexity of methodology, sample size of published GWAS and number of traits per published GWAS has increased over time, generating robust SNP-trait associations for a huge number of phenotypes (**Figure 2.4**) (GWAS Catalog SNP-trait associations in 2018: 71,673). Accordingly, GWAS data represent a resource of vast scope and scale for investigation of phenotypes and disease.

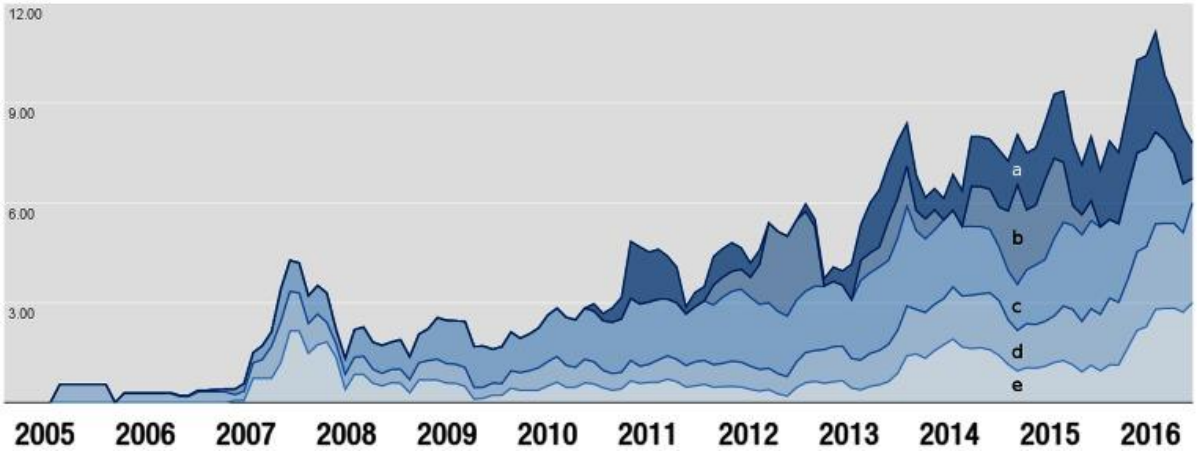


*Figure 2.2 - The increasing complexity of GWAS over time. Adapted from MacArthur et al. 2017 [151]: Increasing complexity of GWAS studies over time (**A**) number of SNP-by-environment interaction studies, (**B**) number of SNP-by-SNP interaction publications, (**C**) number of traits per publication, (**D**) number of ancestry categories each GWAS publication analyzed and (**E**) number of GWAS analyses per publication. Values were normalized to provide equal weighting to each category.*

GWAS Design

Brief introduction to linear, logistic and survival GWAS formulae

Depending on the phenotype being assessed in a GWAS, different statistical models are employed to accurately determine and interpret the magnitude of effect an associated genetic variant confers [152]. For GWAS of continuous phenotypes (e.g. height or alcohol in units per week) a linear regression model is commonly used. Linear regression models assume that a continuous phenotype is normally-distributed within a population and typically calculated as:

$$y = \beta_0 + \beta_G G_i + covariate_{1-j} + \varepsilon$$

Where G_i *is the genotype of the i-th person,* β_G *is the quantitative increase in phenotype per additional reference allele,* β_0 *is the intercept term, covariates 1-j are included in the model (i.e. age, sex, ethnicity etc.) and ε denotes model residuals*

The beta value from this model interpreted as the quantitative change in phenotype (e.g. for smoking, the change in units per week) per additional reference SNP.

For GWAS of binary phenotypes (e.g. ever vs never smoking, or HPV16E6 seropositivity vs HPV16E6 seronegativity) a logistic regression model is typically employed. Logistic regressions generally assume that a phenotype of interest is Bernoullli-distributed; that is, it has two possible outcomes where the probabilities of either occurring sum to 1, but neither can have a probability of 1 or 0 of occurring [153]. It should be noted that whilst the *probability* of the outcomes cannot equal 0

or 1, the *state* of the outcome is typically denoted as being either 0 or 1 (i.e. for ever vs never smoking, "1" would denote an ever smoker and "0" would denote a never smoker). This is depicted in statistical notation as the random variable Y_i (i.e. flipping a coin) having a probability of π (i.e. for flipping a fair coin, π = 0.5) of occurring for each i-th individual:

$$Y_i \sim Bernoulli(\pi_i)$$

If this is the case, a "logit link" function is used to link the probability of a phenotypic state occurring in a Bernoulli distribution to a linear function of the predictors (in the case of a GWAS, how much a trait increases per additional copy of a SNP) [154]:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = x_i'\beta$$

Where $\mathrm{x_i}$ *is a vector of covariates and* β *is a vector of regression coefficients.*

With this function linking π_i to a linear function, a logistic regression is calculated using:

$$x_i'\beta = \beta_0 + \beta_G G_i + covariate_{1-j} + \varepsilon$$

Where $\mathrm{G_i}$ *is the genotype of the i-th person,* β_G *is the log(odds ratio) per additional reference allele,* β_0 *is the intercept term, covariates 1-j are included in the model (i.e. age, sex, ethnicity etc.) and ε denotes model residuals*

The beta value from this model is interpreted as the change in the log(odds ratio) of an individual being a "case" (e.g. smoker) vs a "control" (e.g. non-smoker), per copy of a given reference SNP.

Finally, for time-to-event phenotypes (e.g. 5-year mortality), a Cox proportional-hazards GWAS model is regularly used. One of the key assumptions for Cox proportional-hazards models, known as the proportional-hazards assumption, is that over time the model covariates have a constant, multiplicative effect on the hazard function. The hazard function models the probability of an event (e.g. death in the case of 5-year mortality) occurring at a given point in time. Therefore, the proportional hazards assumption ensures that covariates don't interact with time to cause a disproportionate change in occurrence of an event of interest over time. A Cox proportional-hazards model is calculated as follows:

$$h_i(t) = h_0(t)\exp\left(x_i'\beta\right)$$

Where $\mathrm{h_i}(t)$ *is the hazard function for i-th individual,* $\mathrm{h_0}(t)$*is the baseline hazard when time = 0,* $\mathrm{x_i}$ *is a vector of model covariates and* β *is a vector of regression coefficients*

Beta values from Cox proportional-hazards GWAS models are interpreted as the change in log(hazard ratio) per additional reference SNP on the probability of an event occurring.

Confounding and GWAS precision

When conducting GWAS, it is important to be able to determine whether a genetic variant has an independent, direct association with a phenotype of interest, rather than a spurious association caused by confounding factors. Perhaps the largest confounding factor in GWAS is population structure [155], which generates systematic differences between individuals based on their ancestry. In a 1994 paper by Lander and Schork, a useful exemplar is given to illustrate this issue in the case of genetic association studies [156]:

> *"In a mixed population, any trait present at a higher frequency in an ethnic group will show positive association with any allele that also happens to be more common in that group. To give a light-hearted example, suppose that a would-be geneticist set out to study the "trait" of ability to eat with chopsticks in the San Francisco population by performing an association study with the HLA complex. The allele HLA-Al would turn out to be positively associated with ability to use chopsticks not because immunological determinants play any role in manual dexterity, but simply because the allele HLA-Al is more common among Asians than Caucasians."*

Standard GWAS regression models assume that the source population is not related, and therefore that each variable is identically and independently distributed [157]. However, in modern GWAS datasets of up to hundreds of thousands of individuals, a certain degree of ancestral relatedness is inevitable [158]. Whilst the chopstick association above may be a light-hearted example, confounding of health-related phenotypes by population structure presents a serious issue for researchers when attempting to understand and improve disease aetiology. Several methods have been developed to address confounding by population stratification, including (more recently) the use of mixed models [159] and principal components analysis (PCA) [160]. Mixed model approaches include fixed-effects phenotypes (phenotypes constant over time across individuals) such as SNPs, age and sex in their statistical model, but also model heritable variation as a random effect in the form of a kinship matrix of pairwise genotypic similarity between study individuals. By including a kinship matrix of this design, phenotypic differences which are a result of population stratification can be accounted for in a GWAS regression model [161]. PCA, in the context of a GWAS, aims to determine independent linear combinations of SNPs that account for the greatest variation in the source genetic

data. If genetic data is converted to a high-dimensional construct (a hypersphere) containing X dimensions (where X is the number of SNPs), the first principle component is an estimation of the direction through the hypersphere that explains the greatest amount of variance in the genetic data. The second principle component is the direction through the hypersphere that will create the second-greatest spread of data, with the constraint that it is uncorrelated (orthogonal) to the previous principle component [162]. All consequent principal components will iteratively explain the largest variance in the genetic data, given that they are orthogonal to the other principal components. This process creates an axis of "directions" (the principal components) with fewer dimensions than the original data but with the same variance, which the genetic data is projected onto. For PCA in GWAS, the underlying assumption is that most genetic variation is due to population structure. Therefore, identifying and adjusting for the top principal components from a PCA analysis should correspondingly ensure adjustment for population structure.

Another source of bias in GWAS studies can result from either a lack of adjustment, or problematic adjustment for covariates in a GWAS regression model. If not accounted for correctly, a scenario can occur where an association between a SNP and a particular phenotype (e.g. smoking) can cause an apparent association between the same SNP and another closely-associated phenotype (e.g. alcohol - smokers are more likely to drink alcohol, causing SNPs associated with smoking to appear in a GWAS of alcohol consumption, if not accounted for) [163]. This "contamination" can bias GWAS results by affecting SNP beta values (falsely inflating or attenuating them) or simply creating spurious SNP-phenotype associations. Known confounding can be adjusted for by including the confounder as a covariate in the GWAS regression model. Unknown confounding can largely be accounted for by adjusting GWAS regression models for principal components as detailed above. However, care should be taken in choice of covariates when adjusting for known confounding. If a covariate in a GWAS model is heritable (i.e. a proportion of the variation in the trait is attributable to genetic factors, rather than demographic factors), then a SNP can potentially be associated with both the phenotype of interest and the covariate [164]. In this scenario, given the number of pleiotropic genes associated with many complex traits, the effect of a SNP on a phenotype of interest is likely not independent, and instead biased towards the covariate [165]. Without sound knowledge of the covariate's pathophysiology, it is difficult to ascertain what the independent effect of a SNP on a phenotype of interest is likely to be. Therefore, care should be taken to understand how a SNP-trait association was derived (e.g. how a GWAS has been conducted) and how to interpret the effect estimate of the SNP on a trait (e.g. what other traits it may be associated with, and the resultant direction of bias, if any).

Multiple-testing burden

GWAS involve investigating the effect of millions of SNPs against a phenotype of interest. Therefore, due to the number of tests, the probability of spurious false-positive associations (type 1 error) being discovered is increased. To account for the multitude of statistical tests, a p-value correction is applied to GWAS results. In an attempt to quantify a significance threshold for GWAS, Dudbridge and Gusnato estimated the threshold for GWAS in Caucasian populations to be $P < 5x10^{-8}$ [166]. Accordingly, this p-value threshold has become a standard for most GWAS investigating common genetic variants.

Other than the aforementioned threshold, two of the most common corrections for determining an adequate p-value threshold for GWAS include Bonferroni correction and false discovery rate (FDR) [167]. The Bonferroni correction attempts to minimise the likelihood of a single false-positive finding by adjusting the p-value threshold for a single test of 0.05 by the overall number of tests (0.05/N, where N is the overall number of tests – see 2.2.1 - Mining Enriched Literature Objects to Derive Intermediates (MELODI)) [168]. In the case of a GWAS, this number of tests is equal to the number of SNPs (per person) in the genetic data. However, many SNPs are correlated due to linkage disequilibrium (LD: see "Linkage Disequilibrium" below) and are therefore not entirely independent. Consequently, the Bonferroni correction is likely too conservative a correction for GWAS, elevating the likelihood of false-negative findings (type 2 error) for the sake of minimising type 1 error. FDR correction adjusts the p-value for a single test by assuming that a fixed proportion of findings are false-positives [169]. FDR is calculated as follows:

$$\mathrm{FDR} = \mathbb{E}(\mathrm{V/R} \mid \mathrm{R} > 0)\, \mathrm{P}(\mathrm{R} > 0)$$

Where V = number of Type I errors and R = number of rejected hypotheses

Like a Bonferroni correction, FDR assumes that SNPs are independent, thus making the estimation of the fixed proportion of findings expected to be false-positive difficult. However, despite their limitations, these methods, alongside the Dudbridge and Gusnato $P < 5x10^{-8}$ threshold, are currently the most commonly accepted corrections for use in GWAS.

Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random distribution of alleles within a population. Linkage equilibrium would occur if all genetic variants were distributed randomly, as a result of continuous breaking apart and recombination of chromosomes. However, contiguous stretches of

DNA sequence still exist in human populations, allowing certain SNPs to be inherited with others. In GWAS, LD is measured using the r-squared statistic (r^2). The r^2 statistic is a quantification of correlation, with a high r^2 value between two SNPs describing a relationship where one SNP contains most of the information of the other. A benefit of this phenomenon is that ~80% of common, genome-wide SNPs in European populations can be captured using a subset of around 500,000-1,000,000 "tag" SNPs, usually incorporated onto a microarray (see 2.3.2 – Introduction to the OncoArray platform). However, due to the number of correlated SNPs and high complexity of the genome, when using GWAS results for downstream analysis (such as MR; see 2.2.4) it is difficult to ascertain the degree of bias which arises from non-independent instruments. To address this issue, a statistical approach known as LD clumping can be used to minimise the presence of correlated variants. LD pruning uses p-values of GWAS results to order the SNPs by their association with a phenotype of interest. The most-associated SNP is selected, and SNPs at a specified r^2 value (typically around 0.1 – 0.01) in a "window" around it (commonly 10,000 kilo-bases [kb]) are removed. The process is repeated iteratively, selecting the subsequent most-associated SNP that has not yet been removed, to result in a dataset containing SNPs which are largely independent of each other, associated with a phenotype of interest.

Genetic imputation

In order to comprehensively investigate the genome for novel variants associated with a trait of interest in a GWAS, larger sample sizes, next-generation DNA sequencing techniques or SNP arrays with denser coverage are required. As this isn't always financially or practically feasible (for example, due to lack of next-generation sequencing equipment or a ceiling on the coverage currently provided by SNP arrays), a technique known as genotype imputation can be used as an alternative to predict untyped genetic information, using a reference panel. To provide a reliable reference panel, international consortia with national whole-genome sequencing projects have been established [170]. Shared blocks of genetic variants which are inherited together, per chromosome, known as haplotypes, exist between individuals with a recent common ancestor. It is from haplotypes that genetic imputation is possible. To perform genetic imputation, the data to undergo genotype imputation are compared against multiple reference haplotype sequences which contain many more genetic markers. Stretches of shared haplotype between the original samples and the reference panel are then identified and genetic variants are assigned to missing genotypes in the original samples where possible. Multiple computational tools exist to facilitate genetic imputation, which have been grouped by Li et al. as either computationally intensive tools, such as IMPUTE [171], MACH [172] and fastPHASE/BIMBAM [173, 174] that use every available genotype when attempting to infer each

missing allele, or more computationally efficient tools such as PLINK [175], TUNA [176], WHAP [177] and BEAGLE [178], which impute genotypes based on a window of nearby markers [179]. Both classifications of programs are able to provide an estimate of the quality of imputation known as an INFO score. INFO score ranges from 0 to 1, where 1 constitutes no uncertainty in the imputed genotype, and 0 constitutes a situation akin to guessing the genotype based on population allele frequencies. An INFO score of 0.8 is commonly accepted as a sufficiently accurate estimate of genotype. INFO scores are calculated per locus, l, using the following formula:

$$\mathrm{INFO}_l = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{v_{il}}{w_l},$$

Where v_{il} is the variance of person i's genotype distribution at locus l, and w_l denotes the variance of the genotype distribution under Hardy-Weinberg equilibrium.

Scott et al. were the first authors to publish an investigation into the performance of genotype imputation in GWAS [180]. The authors genotyped around 300,000 SNPs in a type 2 diabetes case-control study with 1,161 cases and 1,174 controls, before imputing over 2 million SNPs to allow them to compare results with two other type 2 diabetes GWAS that had used different genotyping platforms. Scott et al. compared imputed genotypes to actual genotypes for 510 SNPs not present on their GWAS panel, finding an imputed to "actual" genotype concordance rate of 98.5%. Given the huge coverage increase and high concordance between imputed and measured SNPs, genetic imputation is an extremely valuable technique for GWAS and genetic epidemiology more generally.

2.2.3. Epigenome-wide association studies

A brief introduction to epigenome-wide association studies

As mentioned previously (see 1.3.1. Introduction to epigenetics of OPC), the epigenome is known to regulate gene expression whilst being independent of the underlying genetic architecture. The epigenome is also modifiable, in that the degree of epigenetic regulation of gene expression is affected by internal (biological – e.g. lung function [181]) and external (lifestyle – e.g. smoking [182]) factors such as immune response, alcohol consumption or cigarette smoking. By being both modifiable and a regulator of gene expression, the epigenome can lie on and potentially mediate the causal pathway between a factor and an outcome of interest, such as smoking on cancer.

The modifiable changes to the epigenome are partly elastic; if an internal or external factor causes a change in epigenetic regulation of gene expression, and that factor no longer occurs, some epigenetic modifications will begin to return to their inherited "normal" level [183]. Others, however, will persist as a marker of historical exposure, particularly where an individual has had chronic exposure to a factor. Ultimately, in addition to being a potential factor-disease mediator, the epigenome can be thought of as a "biosocial archive", capturing the exposure history of many biological and behavioural factors important to epidemiological research [184].

Similar to a GWAS with respect to SNPs, an epigenome-wide association study (EWAS) is an epidemiological method which compares epigenetic marks (typically DNA methylation) between large populations of individuals to find epigenetic variation associated with a particular phenotype. Epigenetic information (typically extracted from blood samples) is compared against a phenotype using regression analyses pertinent to linear, binary or time-to-event phenotypes (see "*Brief introduction to linear, logistic and survival GWAS formulae"*). Epigenetic marks associated with a phenotype of interest can provide valuable insight into potential biological pathways which modify them, affect downstream biological pathways, or highlight specific gene regions for further investigation.

In addition to finding individual epigenetic marks associated with a phenotype of interest, another common EWAS examines the association between differentially-methylated regions (DMRs) and a phenotype. DMRs are stretches of correlated epigenetic marks which commonly occur around gene regions and may show increased biological plausibility compared to single CpG sites [185]. DMR regression analyses are conducted much in the same way as single-site EWAS, albeit adjust for autocorrelation between neighbouring CpG sites to give a regional (across multiple CpG sites) rather than individual CpG p-value of association [186].

Whole blood, saliva and tumour tissue DNA extraction

DNA can be extracted from a variety of biological tissues for downstream generation of methylation data for use in EWAS and other epigenetic analyses. Among the most common sources of DNA methylation are saliva, whole blood and formalin-fixed, paraffin-embedded (FFPE) tumour tissue blocks. In this thesis, whole blood is used as the tissue of choice due to its minimally-invasive, high-yield and reliable response to exposures without contamination. The comparative characteristics of common tissue sources of DNA have been summarised in Table 2.3, below:

Table 2.3 - Comparative characteristics from commonly-extracted DNA sources for use with DNA methylation arrays

Characteristic	Methylation source		
	Saliva	**Whole blood**	**Tumour tissue block**
Collection method	Non-invasive collection by buccal swab or mouth rinse	Blood draw; minimally invasive	Biopsy, typically formalin-fixed; invasive, may not be possible in live patients
Ease of collection	Self-collection possible/no need for medically trained personnel	Venepuncture training necessary	Specialist/surgical collection of specimens
DNA stability	Stable storage at room temperature for up to 5 years [187]	Stable storage for years on a Whatman FTA Card; stable for 24 hours at ambient temperature or 4C in an EDTA tube; over a year at -80C or -20C with preservative [188]	No notable difference between blocks stored over 11–12 years, 5–7 years, or 1–2 years in comparison to current year blocks [189]
Yield	Mean 24µg Range 0.2–52µg [190]	Mean 210µg Range 58–577µg [190]	Mean 57µg Range 4.3-141.6µg (Average of Nanodrop results) [191]
Heterogeneity	Large (also varies with choice of cell reference) [192]	Large (prior to deconvolution) [193]	Large (dependent on origin of sample) [194]
Acute/chronic exposure indicator	Both	Both	Both
Used in biomarker discovery in literature?	Yes	Yes	Yes
Contamination	Variable bacterial contamination [195]	Negligible [196]	Negligible [197]
Cell reference for decomposition	Yes	Yes	Yes
Applicability to OPC	May prove lucrative in understanding HPV-associated methylation patterns due to proximity	Reliable, robust archive of biosocial exposure with substantial literature evidence.	Opportunity to assess OPC-specific changes rather than more global exposure history. Useful

Characteristic	Methylation source		
	Saliva	Whole blood	Tumour tissue block
	to tonsillar crypts. May also prove useful for acute tobacco and alcohol-related epigenetic changes. Comparative lack of reference studies to other DNAm sources and unable to determine site-specific changes	Useful for determining exposure history, with many previous blood-based EWAS of smoking and alcohol consumption. Unable to determine site-specific epigenetic changes	for determining expression differences between cases and controls, in addition to low vs high grade tumours. Potential high prognostic value.

Genetic associations with DNA methylation

With the advent of genome-wide association studies (GWAS) it has become increasingly clear that much of the genetic variation linked to disease acts not through altering protein coding genes but via gene regulatory pathways. Genetic variants identified in these studies explain only a small proportion of disease variability with effects that confer only small increase in risk. The 'missing' or unexplained heritability observed in many GWAS has been suggested to be partially due to the effects of the epigenome, which form a second "layer" of heritability through gene control. This leaves open the possibility for an epigenetic component to almost all human complex diseases that are at least partially heritable. Recent studies have demonstrated that a sizable proportion of GWAS SNPs exert their influence on disease through their effect on DNA methylation [198]. Furthermore, it has been demonstrated that DNA methylation is associated with extensive genetic variation, such that a high proportion of methylation variable CpG sites are associated with one or more SNPs. Over 50% of CpG sites are estimated to have a genetic component, such that variance in methylation at any given site includes both genetic and environmental contributions. Methylation quantitative trait loci (mQTLs) are sequence variants that associate with DNA methylation and are categorised based on their proximity to CpGs as either in-cis (usually defined as <1Mb or nearby on the chromosome) or in-trans (>1Mb or on different chromosomes) [199, 200]. Cis-mQTLs are generally of large effect whereas trans-mQTLs are more polygenic and have only small effects on methylation variation at any given CpG site. The genetic architecture of DNA methylation is only beginning to emerge but forms an important component of the interpretation of epigenetic variation.

DNA methylation as a predictor of phenotype

Transient and permanent changes in DNA methylation can be used to construct DNA methylation "scores" to predict phenotypes. If an EWAS discovers CpG sites associated with a trait of interest, the beta values from the EWAS at those sites can be used as weights to derive a weighted score of the trait in an independent population. Both smoking and alcohol have shown high concordance between methylation score and directly-measured phenotype [201, 202]. The derivation of a simple weighted score can obtained by multiplying the methylation value at a given CpG by the effect size from an EWAS, and then summing the values:

$$b_1cpg_1 + b_2cpg_2 + \ldots + b_ncpg_n$$

Where "cpg" is the normalized methylation value from a BeadChip or other methylation measurement platform and "b" is the effect size from an EWAS of a trait of interest

2.2.4. Mendelian randomization

Genetic variants as instrumental variables

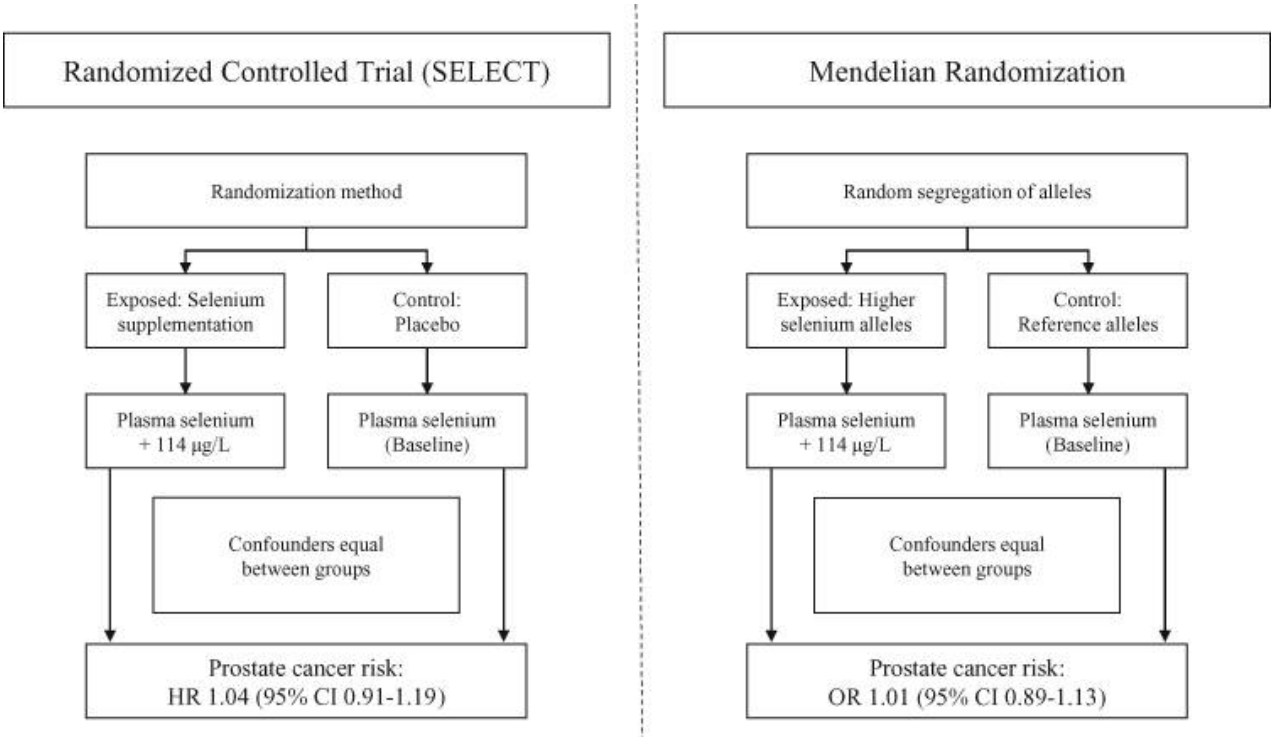
Gregor Mendel, known as "the father of modern genetics", was a monk born in Austria in 1822. Mendel discovered the basic principles of heredity through experiments conducted in his monastery's garden. His experiments showed that the inheritance of certain traits in pea plants followed specific patterns, upon which the foundation of the laws of modern genetic inheritance were developed; the first two of which are pertinent to Genetic Epidemiology:

1. The Law of Segregation of genes states that every trait within an individual contains two alleles, and that these alleles separate during meiosis such that each gamete contains only one of these alleles. As such, an offspring will receive a pair of alleles for a trait by inheriting homologous chromosomes; one allele for each trait from each parent.
2. The Law of Independent Assortment states that alleles for separate traits are passed independently of one another from parents to offspring. This means that the biological selection of an allele for one trait occurs completely independently of an allele for another.

As mentioned previously (see Genome-Wide Association Studies), genetic variants can proxy for a certain phenotype by being strongly associated with it. Building on the Laws of Segregation and Independent Assortment, Genetic Epidemiology can implement genetic variants as instrumental variables to assess the causality of an exposure on an outcome. Given the random nature of their

inheritance within a target population, these proxies are analogous to arms of a randomised control trial (RCT), thus are largely independent of confounding factors (**Figure 2.5**) [203].

Figure 2.3 – Schematic comparison of a randomized controlled trial (RCT; Selenium and Vitamin E Cancer Prevention Trial [SELECT]) to a Mendelian randomization analysis.



Taken from Yarmolinsky et al. 2018 [204]: In an RCT, individuals are randomly allocated to an intervention or control group (In SELECT, 200 µg/d selenium [114 µg/L increase in blood selenium] or placebo). If the trial is adequately sized, random assignment should ensure that intervention and control groups are comparable in all respects (eg, approximately equal distribution of potential confounding factors) except for the intervention being tested. In an intention-to-treat analysis, any observed differences in outcomes between intervention and control groups can then be attributed to the trial arm to which they were allocated. In a Mendelian randomization (MR) analysis, alleles that influence levels of a trait of interest are randomly allocated at conception. (In MR, the additive effects of selenium-raising alleles on 11 single nucleotide polymorphisms were scaled to mirror a 114 µg/L increase in blood selenium.) Groups defined by genotype should be comparable in all respects (eg, distribution of both genetic and environmental confounding factors) except for their exposure to a trait of interest. Any observed differences in outcomes between groups defined by genotype can then be attributed to differences in lifelong exposure to the trait of interest under study. Mendelian randomization is an application of the technique of instrumental variable (IV) analysis. In order for a genetic variant (or a multi-allelic instrument) to be used as an IV, three key assumptions must be met: 1) the instrument must be reliably associated with the exposure of interest, 2) the instrument should be independent of other factors affecting the outcome (confounders), and 3) the instrument should only affect the outcome through the exposure of interest (known as the exclusion restriction criterion). CI = confidence interval; HR = hazard ratio; SELECT = Selenium and Vitamin E Cancer Prevention Trial.

Furthermore, an individual's genotype is determined at conception; therefore, genetic variants are not modified by the later development of a disease or health outcome, removing the complication of reverse causation. Finally, the increasing accuracy of genotyping arrays produces very little measurement error when examining the effect of a genetic variant on a phenotype. When used in instrumental variable analyses such as Mendelian randomization (MR), a genetic variant proxying for a phenotype is free of the limitations that would otherwise weaken causal inference in observational studies [205].

Methodological assumptions of Mendelian randomization

MR relies on a number of key assumptions to reliably derive the causal effect of a genetic variant on an outcome of interest, shown in **Figure 2.6**. Provided these assumptions are met, genetic variants can be used as instrumental variables to provide a true causal effect estimate between an exposure (e.g. smoking) and an outcome (e.g. OPC).

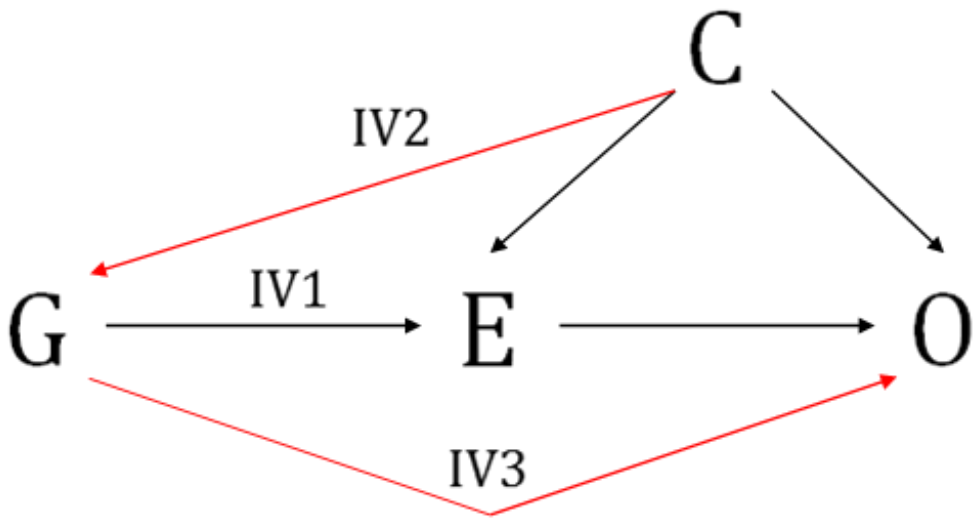


Figure 2.4 - Directed acyclic graph (DAG) of the theory and key assumptions of Mendelian randomization. A genetic variant (or variants, G) can be used as instrumental variables for an exposure of interest (E) to assess the causal association between E and the outcome of interest (O) given that the following three assumptions hold: (IV1) G must be robustly associated with E; (IV2) G must not be associated with any measured or unmeasured confounding variable (C); and (IV3) there must be no independent association between G and O, given E and C.

Applications of Mendelian randomization

Two-sample Mendelian Randomization

Two-sample MR involves using two different (non-overlapping) study samples to estimate the instrument-risk factor and instrument-outcome associations to estimate a causal effect of the risk factor on the outcome [125, 206, 207]. This can be useful when the risk factor or outcome, or both, are expensive to measure. It also provides an opportunity to substantially increase the statistical power, by incorporating data from multiple sources, including large consortia. Furthermore, MR is able to utilise summary-level instrument-exposure and instrument-outcome association results (typically, per-allele regression coefficients and standard errors from GWAS) to obtain causal effect estimates. The use of summary-level data prevents potential identification of study participants and

improves the potential for data sharing, consequently improving the scope of MR analyses which can be conducted.

Phenome-wide Mendelian Randomization

An extension of the two-sample MR approach is phenome-wide MR analysis [124, 208]. Given the availability of vast amounts of genetic data (pertaining to numerous exposures and outcomes) and the efficacy of two-sample MR as a statistical method, phenome-wide MR can be an effective method to test a large number of exposure-outcome associations; the results of which can be used to prioritise and inform downstream hypotheses and analyses. The two-sample MR framework allows for analyses of this scale, in that it can be used to test both the association of a genetically instrumentable exposure across all potential outcomes, or conversely test the association of all instrumentable exposures for a given outcome (with available genetic data).

This phenome-wide approach may provide novel insights into disease aetiology that may not have been captured using previous hypothesis-driven approaches. However, it is important to replicate any putative findings from the phenome-wide search in an independent data set. As large GWAS and consortia are becoming more prevalent, the number of studies and exposure-cancer associations that can be analysed using a two-sample MR framework will increase in quantity and power over time.

Two Step Mendelian Randomization

As mentioned in 2.2.3 – Genetics of epigenetics, methylation at specific loci can be instrumented by mQTLs. Building on two-sample MR, two-step MR is an extension of the MR framework which allows for appraisal of causality of molecular intermediates [126]. Firstly, the association between an exposure and intermediate is established using a traditional MR approach (e.g. a genetic IV for smoking is regressed against methylation levels at the *AHRR* loci). Next, in a second step, an IV proxying the intermediate is regressed against the outcome of interest (e.g. an mQTL proxying methylation at *AHRR* is regressed against incidence of OPC). The IVs for each step should come from independent samples.

2.3. Data and resources

The analyses in this thesis predominately utilise genetic, epigenetic and phenotype data from Head and Neck 5000 (HN5000) and the OncoArray Consortium. Below, each resource is briefly introduced and described. Any chapter-specific resources are described in the relevant results chapter.

2.3.1. Head and Neck 5000 (HN5000) clinical cohort study

Between April 2011 and December 2014, 5511 individuals with HNC were recruited from 76 centres across the UK [209]. All people with a new diagnosis of HNC were eligible to join the study and were recruited before or within a month of their cancer treatment commencing. Individuals with cancers of the pharynx, mouth, larynx, salivary glands and thyroid were included, while those with lymphoma, tumours of the skin or a recurrence of a previous head and neck cancer were excluded from the study. There were 119 exclusions between recruitment and our data release (v2.3) for the following reasons: withdrawn by study/ineligible (n = 72), patient choice withdrawal (n = 12), and not HNC (n = 35). Participants with OPC were selected from the wider pool of individuals (post-exclusion) in HN5000 (N: 5392) based on an ICD-10 coding (pathological where available, clinical if otherwise) of oropharynx (CO1, CO5, CO9, C10.0-2, C10.3, C10.8 and C10.9; N: 1909/5392), availability of OncoChip genotype data generated previously (N: 1034/1909), baseline questionnaire and data capture information (see below), and the availability of blood samples taken at baseline (prior to treatment; N: 448/1034).

N.B. HN5000 OncoArray data is a subset of GWAS data from people of European descent with HNC and matched controls were obtained from the OncoArray Consortium GWAS of oral cavity and pharyngeal cancer. Therefore, information on how this data was derived will be described in section 2.3.2 – OncoArray Consortium - oral cavity and pharyngeal cancer GWAS.

Local research nurses obtained informed consent from individuals, which included agreement to collect, store and use biological samples; obtain samples of stored tissue; carry out genetic analyses and collect clinical information from hospital notes and mortality data through record linkage. Ethics approval for this study was granted by the National Research Ethics Committee (South West Frenchay Ethics Committee, reference 10/H0107/57, 5th November 2010) and approved by the research and development departments from participating NHS Trusts.

Baseline data collection

Participants completed a series of three self-administered questionnaires at baseline enquiring about: 1) social and economic circumstances, overall health and lifestyle behaviours; 2) physical and psychological health, well-being and quality of life; and 3) past sexual history and behaviours. Information on diagnosis, treatment and co-morbidity was recorded on a short data capture form using questions based on a national audit. Diagnoses were coded using the International Classification of Diseases (ICD) version 10 and clinical staging of the tumour was derived based on the American Head and Neck Society TNM staging. Research nurses collected a blood sample from all consenting participants. These were then sent to the study centre laboratory at ambient temperature for processing. The blood samples were centrifuged at 3500 rpm for 10 minutes and the buffy coat layer used for DNA extraction. Any additional samples from the same participant were frozen and stored at -80°C.

Assessment of tobacco, alcohol and HPV infection

Detailed information on tobacco and alcohol history was obtained at baseline via the self-administered questionnaire. Participants were asked about their current smoking and drinking status and their use of tobacco and alcohol products prior to receiving their HNC diagnosis. Among smokers, information on age at smoking initiation and number of years of smoking was obtained. The questionnaire differentiated between use of cigarettes, hand-rolled cigarettes, cigars and smokeless tobacco, whereby a cigar was considered equivalent to four cigarettes. From this information, participants were dichotomised into ever and never smokers. Ever smokers were defined as those who smoked at the equivalent of at least 1 tobacco product a day per year, or ≥100 cigarettes in their lifetime. Never smokers were those who reported not smoking in any of the questions answered.

Respondents were asked to report their average weekly alcohol consumption of a range of beverage types (wine, spirits, and beer/larger/cider) before they were diagnosed with cancer. From these measures, we derived an average intake of alcohol consumption in units per week.

HPV serologic testing (HPV16 E6, E7, E1, E2, E4, and L1) was conducted at the German Cancer Research Center (DKFZ, Heidelberg, Germany) using glutathione S-transferase multiplex. Median fluorescence intensity (MFI) values were dichotomized to indicate HPV16 E6 seropositivity using a cut-off of ≥ 1000 MFI. E6 seropositivity is known to be a marker of with a high sensitivity and specificity for HPV16-driven oropharyngeal cancer.

Study follow-up and survival

Regular updates were received from the NHS Central Register (NHSCR) and the NHS Information Centre (NHSIC) notifying on subsequent cancer registrations and survival among cohort members in the Head and Neck 5000 study. Recruitment for the study finished in December 2014 and follow-up information on survival status was obtained on 30th September 2017, resulting in at least 2.75 years of follow-up for all participants (median: 3.1 years; range: 2.75 to 4.9 years: inter-quartile range: 1.1 years).

DNA methylation

Introduction to the Illumina MethylationEPIC BeadChip

The Illumina Infinium MethylationEPIC BeadChip (EPIC) is a bead-based array which can quantify CpG methylation levels at over 850,000 positions across the genome. In 2008, Illumina introduced their first BeadChip; the Human Methylation 27K (HM27) BeadChip, capable of quantifying methylation at 27,000 CpG sites across the genome. In 2011, Illumina released their Human Methylation 450K (HM450) BeadChip, drastically improving genome-wide methylation coverage to interrogate methylation at 450,000 CpG sites. The EPIC array was released in 2016 and is currently the largest commercially available methylation BeadChip, almost doubling the coverage of its 450K predecessor. Recent studies using platforms such as whole-genome bisulfite sequencing (WGBS; the entire genome methylation levels at a single-base resolution) have demonstrated that DNA methylation at regulatory enhancers can determine transcription and phenotypic variation, through modulation of transcription factor binding. Therefore, accurate quantification of DNA methylation at more regulatory regions is essential for our understanding of the role of DNA methylation in human development and disease. To this end, the EPIC BeadChip targets enhancer regions, possessing 90% of the existing coverage of the HM540, but with more than 350,000 CpGs at regions identified as enhancers by FANTOM5 and the ENCODE project. The EPIC array is therefore an extremely useful tool to further our understanding of DNA methylation mechanisms in human disease, particularly the DNA methylation landscape of distal regulatory elements.

Data generation

Following extraction, DNA from whole-blood samples were bisulphite-converted using the Zymo EZ DNA Methylation^{TM} kit (Zymo, Irvine, CA, USA). Genome-wide methylation data were

generated using the Infinium MethylationEPIC BeadChips (EPIC array) (Illumina, USA) according to the manufacturer protocol. The arrays were scanned using an Illumina iScan (version 2.3) by Bristol Bioresource Laboratories.

Pre-processing and quality assurance

Raw data files (IDAT files) were pre-processed using the R package *meffil* (https://github.com/perishky/meffil/) [210] to perform quality control (QC) and normalization [211]. From the initial 448 samples available, 8 samples did not pass QC; 2 samples with incorrect sex prediction based on autosomal DNA methylation, 3 samples with sex detection outliers, 1 sample with an outlier in predicted median methylated vs unmethylated signal, and 2 duplicate samples. An additional 32 individuals were subsequently removed from the analysis owing to pathological re-classification, leaving 408 participants with DNAm data available. During QC, probe intensities were dye-bias and background corrected using the 'noob' method developed by Triche et al [212]. A total of 3674 probes were excluded, leaving 863,289 CpGs with which to perform analyses - 2704 probes were removed due to a high proportion of high detection p-values (>10% of samples with a detection p-value > 0.1) and 970 CpGs had low bead numbers in a high proportion of samples (<3 beads in >10% samples). Following QC, functional normalization (originally developed by Fortin et al. [213]) was performed using the *Meffil* R package, which exploits control probes to separate biological variation from technical variation. Data were normalized using 6 control probe principal components derived from technical probes. During the normalization process, probe intensity quantiles were normalized between samples by fitting linear models to these 6 derived principal components. The resulting quantile residuals for each QC object were retained as a set of normalized quantiles and used in a second normalization step, where the raw probe intensities for each sample were adjusted to conform to its own set of normalized quantiles. After the second step had been completed for each sample, the resulting normalized DNAm data subsets were merged into a single dataset for analysis.

Post-normalization, estimation of blood cell proportions, per sample, were estimated via the Houseman cellular composition prediction algorithm [214]. Reinius et al. 2012 [215] was used as a cell type reference to estimate proportions of neutrophils, natural killer cells, B cells, eosinophils, CD4T cells, CD8T cells and monocytes.

2.3.2. OncoArray Consortium – oral cavity and pharyngeal cancer GWAS

Introduction to the OncoArray platform

The development of chip-based microarray technology has made GWAS increasingly viable and frequent by allowing for standardized, cost-effective assay of millions of SNPs. A genotype microarray

is a collection of thousands (usually hundreds of thousands) of wells, containing oligonucleotide (short nucleic acid polymers) probes which bind to unique SNPs considered to have biological relevance in the genome. In this thesis, large volumes of genetic data (1,034 individuals [case only] with OPC in the HN5000 study; 2,641 OPC cases, 6,585 controls in the OncoArray consortium HNC study – HN5000 is nested within the OncoArray HNC study) are used. All of these samples have been genotyped using the OncoArray BeadChip (OncoChip) microarray. Currently, the OncoChip array is perhaps the most cost-effective microarray for Cancer Epidemiology studies, interrogating almost 500,000 expert-selected SNPs, notably including >200,000 cancer-specific genetic variants [216]. As mentioned above, OncoChip is a BeadChip; it contains wells filled with 3µm silicon beads, covered with oligonucleotide probes (rather than oligonucleotides bound directly to the chip surface), which provide a large surface area with a high density of probes for processed DNA to bind to.

Oral cavity and pharyngeal cancer GWAS

In 2016, Lessuer et al. published GWAS results using genetic data from the OncoArray Consortium oral cavity and pharyngeal cancer GWAS [149]. This study examined 12,619 individuals (6,034 cases, 6,585 controls) from Europe, North America and South America. Cancer cases comprised the following ICD codes: oral cavity (C02.0-C02.9, C03.0-C03.9, C04.0-C04.9, C05.0-C06.9) oropharynx (C01.9, C02.4, C09.0-C10.9), hypopharynx (C13.0-C13.9), overlapping (C14 and combination of other sites) and 25 oral or pharyngeal cases with unknown ICD code (other). Samples were originally genotyped using aforementioned Illumina OncoArray platform (see 2.3.1 – OncoArray genetic data), designed for cancer studies by the OncoArray Consortium, part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. The majority of samples were genotyped as part of the oral and pharynx cancer OncoArray, with the exception of 2,476 shared controls (1,453 from the European cohort study and 1,023 from the Toronto study) that were genotyped at the Center for Infectious Disease Research (Seattle, Washington, United States), but as part of the Lung OncoArray. Genotype calls were made by the Dartmouth team in GenomeStudio software (Illumina, Inc.) using a standardized cluster file for OncoArray studies. A total of 2,641 cases and 6,585 controls were examined in relation to OPC as a specific HNC sub-type.

Pre-processing

Initial quality control steps and analyses were performed at the International Agency for the Research of Cancer (IARC), Lyon. After removing duplicates, related samples, samples with sex discrepancy and population outliers, genotype imputation was performed using the Michigan

Imputation Server. Genotypes were pre-phased (i.e. their haplotypes were inferred) using SHAPEIT v2 and imputed with Minimach v3 using the Haplotype Reference Consortium panel. After imputation, SNPs with an imputation quality (R^2) lower than 0.7 were removed from the datasets.

Statistical analysis

Effect estimates for oral cavity and pharyngeal cancer risk were obtained after adjusting for age, sex and significant principal components for population stratification using R software (R version 3.3.1). Results were calculated for geographic region of HNC (overall oral cancer and pharynx cancer, site-specific oral cancer and oropharyngeal cancer) were then combined using a fixed-effects inverse-variance approach implemented in PLINK.

CHAPTER 3. SYSTEMATIC RETRIEVAL OF OROPHARYNGEAL CANCER RISK FACTORS ENRICHED IN EPIDEMIOLOGICAL LITERATURE

3.1. Introduction

In current scientific literature, epidemiological studies with a focus on oropharyngeal cancer (OPC) are uncommon and display marked homogeneity with respect to risk factors of interest [21, 48, 83, 217-219]. However, OPC incidence is increasing globally, with affected demographics shifting to include younger individuals [11]. Populations affected by this cancer are changing, in part an artefact of a shifting causal landscape for OPC risk factors. Therefore, further research is required to determine how best to prevent and detect OPC in the first instance, and to reduce the elevated mortality and morbidity risks associated with OPC in those with late-stage cancer.

Prior to using genetic and epigenetic data to augment understanding of OPC risk factors and prognostic factors, it is important to know which risk factors exist with good evidence of association with OPC and which risk factors exist with more limited evidence. Notably, the latter group may be important in terms of effect on OPC but may simply be understudied or under-reported in current literature. Once known, an understanding of the mechanisms that relate identified risk factors to health outcomes is crucial for discovery of potential drug targets and disease biomarkers and prevents duplication of effort (i.e. the attempt of investigating an already-established hypothesis). Finally, identifying the mechanistic pathway from a given risk factor to OPC allows consideration of potentially modifiable intermediates, thus offering the potential to identify new biomarkers and treatments to reduce risk and mortality of the disease.

In this chapter, a novel application of the MELODI platform (www.melodi.biocompute.org) is employed (see below) to identify epidemiologically-relevant factors associated with OPC in current literature. In a systematic framework, putative risk factors found to be enriched in literature for OPC are discovered. These risk factors are then recursively investigated against OPC using MELODI's standard framework [138] in an attempt to derive intermediates in addition to risk factors. Any derived intermediates could be important therapeutic targets to reduce risk or mortality of OPC. Of note, intermediates discovered this way could also be independent risk factors and would require downstream investigation in a causal inference framework.

MELODI is an online literature-mining tool which employs the Semantic Medline Database (SemMedDB) [140] of predications to infer association between an exposure and outcome in scientific literature [138]. The use of SemMedDB predications allows for directionality to be established between exposure and outcome, and for overlapping mechanisms to come from independent articles (i.e. one article details a high-calorie diet being associated with increased pancreatic function, and

another shows increased pancreatic function to be associated with type 2 diabetes. The link between high-calorie diet and type 2 diabetes could be overlooked by a typical PubMed search, but in MELODI would show a pathway from high-calorie diet to type 2 diabetes with increased pancreatic function as an intermediate). MELODI also removes background "noise" by performing an enrichment step of the "concepts" (see Methods) within a particular set of articles, aiding the reliability of any exposure-outcome pathways discovered. Finally, MELODI contains a Neo4j graph database of all available literature in PubMed. Unlike traditional databases which store data in rows, columns, and tables, Neo4j has a flexible structure which saves the relationships that connect data. With Neo4j, each data record, or "node", contains directional links to all the nodes it is connected to. For example, if Bob is friends with Jane, a node called "Bob" would point directly to a node called "Jane" through a stored relationship. This network of nodes connected by relationships is a graph. Most other databases, including more recent node SQL types, don't save relationship data directly; they can create connections by searching a separate data structure called an index, but this process has to be repeated to find each connection, which is a time-intensive computational step. Accordingly, traditional databases tend to be inherently slower than Neo4j for relationship-intensive queries, particularly with large amounts of data such as PubMed literature. Neo4j avoids repeated index lookups because its native storage layer is a connected graph. The MELODI Neo4j graph contains data from every PubMed abstract, which has been systematically parsed for SemMedDB predicates. This allows the entire PubMed database (and its search engine) to be implemented in the creation of custom article "search sets". A "search set" is the information and relationships shared by articles relating to a search term of interest.

Here, a novel application of the MELODI tool for the hypothesis-free identification of epidemiological risk factors for OPC present in scientific literature was used: if a search set containing PubMed articles pertaining to OPC (or indeed, any outcome) is compared against *itself* (rather than a unique second search set), MELODI provides a results output which can identify enriched associations (and therefore potential risk factors) associated with it. Typically, MELODI is used to compare information across two search sets to discover intermediates. The output consists of a table of subject-PREDICATE-object pathways, attempting to find enriched concepts in the search set. This set of results helps to identify risk factors (as mentioned above) which are both directly associated with OPC (intra-publication: a risk factor has been directly discovered by the author of the paper i.e. smoking causes OPC. Smoking is therefore a risk factor for OPC), and 1 step removed from the outcome but potentially affecting it via an intermediate (inter-publication: two authors have discovered associations which

may form a causal pathway e.g. one author discovers that oral sex leads to HPV infection, another discovers that HPV infection leads to OPC. Oral sex may therefore be a risk factor for OPC).

3.2. Methods

3.2.1. Risk factor retrieval

Within MELODI, SemMedDB data is available for all PubMed articles, in both "triple" and "concept" form, for those published on or before 30th June 2018. SemMedDB "concepts" are indexing articles similar to MeSH terms [220], derived from the Unified Medical Language System (UMLS) Metathesaurus [141] and describing the content of a journal article. SemMedDB "triples" are a pair of "concepts" (as before) linked by a "PREDICATE" derived from the UMLS Semantic Network [221]. These "triples" have been obtained for PubMed literature through use of the UMLS-based program, SemRep [222], and downloaded into MELODI's database via the SemMedDB repository. Rindflesch and Fiszman [222], who developed SemRep, use the following example to show how their program extracts SemMedDB "triples":

The sentence in (1) extracts the predications in (2):

1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.

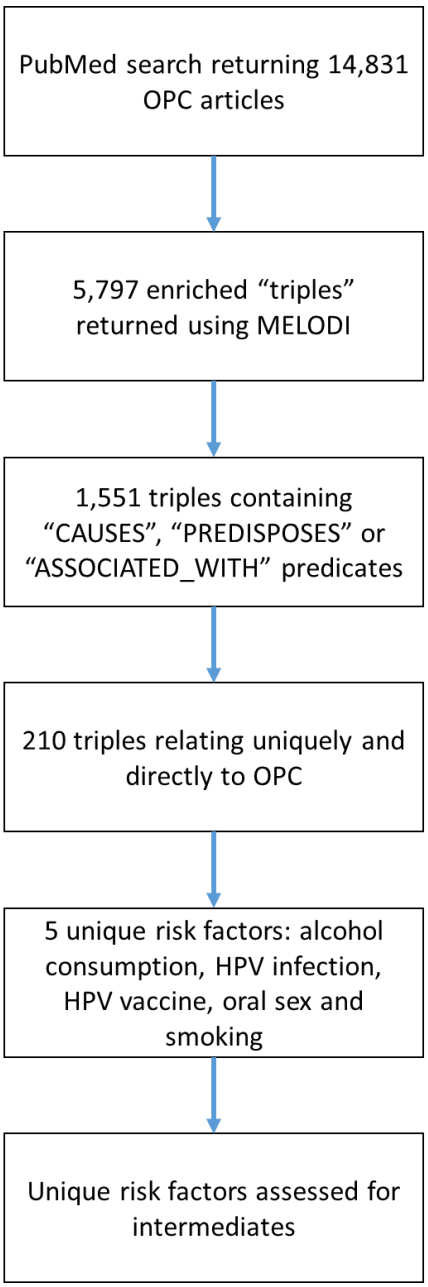
2. Hemofiltration-TREATS-Patients
   Digoxin overdose-PROCESS_OF-Patients
   Hyperkalemia-COMPLICATES-Digoxin overdose
   Hemofiltration-TREATS(INFER)-Digoxin overdose

As evidenced above, SemRep can parse sentences from journal article abstracts, returning refined predications which can be used to generate hypotheses. In the risk factor retrieval stage of this chapter, SemMedDB "triples" (rather than "concepts", of which there would be tens of thousands, with less precise associations due to a lack of directionality) were retrieved using the MELODI platform, in relation to OPC. MELODI returns predications that are present in PubMed literature, between two search sets of articles, greater than expected by chance. However, if a search set pertaining to OPC is compared against *itself*, the MELODI program will return a list of overlapping,

enriched SemMedDB “triples” from within the single set of articles. As of 25/10/2018, a PubMed search set of 14,831 articles pertaining to OPC - “oropharynx cancer OR oropharyngeal cancer OR OPC OR OPSCC OR oropharyngeal squamous cell carcinoma OR oropharynx squamous cell carcinoma” - was assessed for enriched predications. From this search set, 5,797 overlapping “triples” (referred to herein as a pathway) were retrieved.

From all UMLS Sematic Network predicates available in the OPC search set, the predicates of ’CAUSES’, ‘PREDISPOSES’ and ‘ASSOCIATED_WITH’ described potential evidence of a causal epidemiological relationship between retrieved risk factors and OPC, whereas others (e.g. LOCATION_OF, IS_A, PART_OF) did not. These 3 predicates were therefore used to filter the resultant 5,797 potential pathways to 1,551 – if any of the overlapping “triples” contained one of these predicates, it was kept and others removed. Next, the search set was refined to those with a ‘CAUSES’, ‘PREDISPOSES’ or ‘ASSOCIATED_WITH’ predicate prior to one of the only two concepts in the UMLS Metathesaurus which uniquely describes oropharyngeal cancer (either PREDICATE-”malignant neoplasm of oropharynx”, or PREDICATE-“oropharyngeal squamous cell carcinoma”), producing a final set of ‘risk factor – PREDICATE – OPC’ associations (**Figure 3.1**).

Figure 3.1 - Flowchart of MELODI risk factor retrieval process



Intermediate retrieval

After risk factors were obtained for OPC, MELODI was implemented according to its original function; to mine enriched literature objects to derive intermediates. By comparing a search set of each retrieved risk factor against the OPC search set, respectively, intermediates for risk factor-OPC associations may be discovered, where available. Search strategies for risk factors (see Results) were generated using current systematic review literature as reference. HPV vaccine and HPV infection were aggregated in a search term that more broadly encapsulated both terms. The following four search strategies were employed:

Alcohol

Search strategy: alcohol* OR alcoholic beverage OR alcohol consumption OR alcohol drinking OR alcohol use OR alcohol intake OR alcoholism OR alcohol abuse OR ethanol* OR ethanol concentration

Returning a search set of 495,565 articles

Reference: Simou et al. 2018 [223]

Smoking

Search strategy: cigarette OR smoking OR smoke OR tobacco OR snus OR snuff OR "environmental tobacco smoke" OR "passive smoking" OR "smoking cessation" OR "betel nut" OR bidi OR pipe OR cigar

Returning a search set of 321,171 articles

Reference: Aune et al. 2018 [224]

HPV infection

Search strategy: papillomavirus infections/epidemiology[MeSH Terms] OR papillomavirus infections/etiology[MeSH Terms] OR papillomavirus infections/transmission[MeSH Terms] OR "papillomaviridae"[MeSH Terms] OR HPV[Text Word] OR papillomavir*[Text Word] OR papilloma virus*[Text Word]

Returning a search set of 47,399 articles

Reference: Tam et al. 2018 [225]

Oral sex

Search strategy: “sexual behaviour” OR sexual partner* OR sexuality OR “oral sex” OR orogenital* OR “sexual activity”

Returning a search set of 151,688 articles

Reference: Marston and King 2006 [226]

As mentioned previously, MELODI contains information on both SemMedDB “concept” and “triple” predications. In the case of retrieving intermediates, “concepts” were examined. In contrast to the risk factor retrieval stage, “concepts” are more appropriate for intermediate retrieval as they are less frequently reported in epidemiological literature and may be as easily inferred by SemRep in journal abstracts. Enriched concepts were restricted to those of interest to molecular epidemiology; those that were modifiable, those that were molecular intermediates, or those that were able to be proxied by genetic loci to appraise causality (**Table 3.1**). Following identification of these enriched concepts, potential intermediates were manually screened for evidence of mediation of OPC risk.

Table 3.1 - Inclusion and exclusion terms for SemMedDB concepts of interest to molecular epidemiology

Included concepts	**Excluded concepts**			
Amino acid, peptide or protein	Anatomical abnormality	Human	Fish	Professional or Occupational Group
Bacteria	Body Location or Region	Individual Behaviour	Finding	Qualitative concept
Biologically active substance	Body Part, Organ, or Organ Component	Injury or Poisoning	Health care related organization	Research Activity
Carbohydrate	Body Space or Junction	Intellectual product	Organism Function	Research device
Fungus	Cell component	Laboratory Procedure	Organism	Tissue
Gene	Cell	Mental process	Organization	Therapeutic or Preventive Procedure
Hormone	Clinical attribute	Mental or Behavioural Dysfunction	Organ or Tissue Function	
Lipid	Daily or Recreational Activity	Neoplastic Process	Pathologic Function	

Neuroreactive substance or biogenic amine	Disease or Syndrome	Organism Attribute	Phenomenon or process
Pharmacologic substance	Element, Ion, or Isotope	Embryonic structure	Plant
Virus	Health care activity	Family Group	Population Group

3.3. Results

Two hundred and ten "triples" were found to be enriched in the OPC search set, from which 5 unique risk factors for OPC were identified: alcohol consumption (# triples = 17), HPV infection (# triples = 99), HPV vaccination (# triples = 77), oral sex (# triples = 1) and smoking (# triples = 16). For each respective risk factor, over 30 SemMedDB concepts were found to be enriched, including pharmacological, gene region and molecular intermediates. See below for details of enriched concepts per risk factor.

HPV

HPV infection was responsible for the majority of enriched literature associations (n = 99/210). All triples returned from this literature mining approach described HPV as causing ("CAUSES") OPC; there were no triples described HPV predisposing ("PREDISPOSES") or being associated ("ASSOCIATED_WITH") with the disease. The HPV vaccine was also uniquely associated with OPC, present in 77 triples; 56 triples were "ASSOCIATED_WITH" OPC, 21 contained the "CAUSES" predicate.

From the retrieval of enriched literature intermediates between HPV infection (a search was constructed encompassing both; see methods) and OPC, there were 35 potential intermediates: 10 amino acid, peptide, or protein, 5 biologically active substances, 1 carbohydrate, 1 fungus, 5 gene regions, 3 hormones, 7 pharmacologic substances and 3 viruses (see **Table 3.2**).

Table 3.2 - Potential intermediate factors between HPV and OPC

Category	Intermediate	# Articles in HPV search set	# Articles in OPC search set	Overlap
Amino acid, peptide, or protein	Cyclin D1|CCND1	42	13	11
	Epidermal Growth Factor Receptor	84	38	22
	Interferon-beta|EREG|ESR1	21	3	0
	Interferons	196	2	1
	Interleukin-1 beta	27	9	0

	Interleukin-4	39	4	0
	Recombinant Interferon-gamma|CALR	7	3	0
	Retinoblastoma Protein|RB1	285	4	21
	Superoxide Dismutase	12	5	0
	Tumour Necrosis Factor-alpha	64	17	1
Biologically active	Cholesterol	3	6	0
substance	Nitric Oxide	85	19	0
	Sodium Chloride	38	24	0
	Triglycerides	2	3	0
	Zinc	20	4	0
Carbohydrate	Ascorbic acid	13	2	0
Fungus	Saccharomyces cerevisiae	83	10	0
Gene	*CDKN2A*	626	34	201
	EGFR	42	12	10
	MMP8	8	6	4
	SERPINB3	113	51	49
	TP53	1425	67	69
Hormone	Estradiol	40	1	0
	Estrogens	81	4	0
	Hydrocortisone	6	3	0
Pharmacologic	Cetuximab	38	58	30
substance	Cyclosporine	33	12	0
	Dexamethasone	28	14	0
	Fluconazole	1	60	0
	Hydrogen Peroxide	28	7	0
	Indomethacin	20	3	0
	Morphine	2	4	0
Virus	Hepatitis C virus	49	3	1
	Herpesvirus 4, Human	191	57	15
	Human Papillomavirus	13350	901	901

Alcohol

Seventeen triples were returned pertaining to alcohol consumption and OPC risk from this analysis (17/88 triples). All 17 described alcohol consumption predisposing ("PREDISPOSES") OPC, rather than alcohol consumption causing OPC or alcohol consumption being associated with OPC.

From the retrieval of enriched literature intermediates between alcohol consumption and OPC, there were 74 potential intermediates: 22 amino acid, peptide, or protein, 1 bacteria, 9 biologically active substances, 1 carbohydrate, 2 fungi, 6 gene regions, 8 hormones, 1 neuroreactive substance or biogenic amine, 20 pharmacologic substances and 4 viruses (see **Table 3.3**).

Table 3.3 - Potential intermediate factors between alcohol consumption and OPC

Category	Intermediate	# Articles in alcohol search set	# Articles in OPC search set	Overlap
Amino acid, peptide, or protein	Cyclin D1	80	18	6
	Epidermal Growth Factor Receptor	107	54	6
	Interferons	292	3	0
	Interleukin-1 beta	744	9	0
	Interleukin-4	234	3	1
	Leptin	356	4	0
	NF-kappa B	587	6	0
	Recombinant Interferon-gamma|CALR	15	3	0
	Retinoblastoma Protein 1	20	21	6
	Superoxide Dismutase	1273	4	1
	Tumor Necrosis Factor-alpha	946	17	1
Biologically active substance	Cholesterol	1891	5	1
	Nitric Oxide	1590	19	0
	Sodium Chloride	2837	23	1
	Zinc	700	3	1
Carbohydrate	Fluorodeoxyglucose	5	17	0
Fungus	Candida albicans	256	45	0
	Saccharomyces cerevisiae	5136	10	0
Gene	CA2	963	22	0
	CDKN2A	94	212	23
	EGFR	17	21	1
	IGHE	116	1	0
	TNFRSF6B	20	22	0
	TP53	479	116	20
Hormone	Adrenal Cortex Hormones	852	12	1
	Estradiol	637	1	0
	Estrogens	763	4	0

	Glucocorticoids	392	3	0
	Hydrocortisone	545	2	1
	Insulin	970	3	0
	Progesterone	373	6	0
	Testosterone	784	1	0
Neuroreactive substance or biogenic amine	Norepinephrine	969	11	0
Pharmacologic substance	Amifostine	5	11	0
	Angiotensin-Converting Enzyme Inhibitors	184	1	0
	Argipressin	339	109	1
	Aripiprazole	53	25	0
	Bleomycin	82	48	1
	Carboplatin	18	71	1
	Cetuximab	7	85	3
	Cilostazol	12	44	1
	Cisplatin	222	258	6
	Cyclosporine	253	11	1
	Dexamethasone	439	14	0
	Docetaxel	64	42	1
	Fluconazole	59	60	0
	Fluorouracil	119	118	2
	Hydrogen peroxide	1495	7	0
	Iron	1237	4	0
	Inamrinone	7	11	0
	Morphine	799	4	0
	Paclitaxel	244	51	1
	Phosphodiesterase Inhibitors	63	23	0
Virus	Hepatitis C virus	2079	3	1
	Herpesvirus 4, Human	42	68	4
	Human papillomavirus	199	780	134
	Papillomavirus	6	15	0

Smoking

There were 16 triples pertaining to tobacco smoking and OPC risk in this analysis. 14 triples described tobacco predisposing ("PREDISPOSES") OPC and 2 triples described tobacco being associated with ("ASSOCIATED_WITH") OPC. There were no triples describing tobacco causing ("CAUSES") OPC.

From the retrieval of enriched literature intermediates between smoking and OPC, there were 74 potential intermediates: 17 amino acid, peptide, or protein, 1 bacteria, 6 biologically active substances, 1 carbohydrate, 2 fungi, 5 gene regions, 8 hormones, 2 neuroreactive substance or biogenic amine, 19 pharmacologic substances and 3 viruses (see **Table 3.4**).

Table 3.4 - Potential intermediate factors between smoking and OPC

Category	Intermediate	# Articles in smoking search set	# Articles in OPC search set	Overlap
Amino acid, peptide, or protein	Angiotensin II	72	6	0
	ATP8A2	163	15	0
	Collagen	283	6	0
	Gamma-Aminobutyric Acid	116	4	0
	Interleukin-1 beta	438	9	0
	Interleukin-4	147	2	2
	Leptin|LEP	254	4	0
	Myelin Basic Proteins	8	36	0
	NF-kappa B	246	6	0
	Phosphoric diester hydrolase	18	31	0
	Recombinant Interferon-gamma|CALR	11	3	0
	Retinoblastoma Protein|RB1	25	20	7
	Somatotropin	52	1	0
	Substance P	113	1	0
	Superoxide Dismutase	433	4	1
	Tumor Necrosis Factor-alpha	481	16	2
	Vasopressin Receptor	2	37	0
Biologically active substance	Calcium	702	9	1
	Nitric Oxide	829	17	2
	Phospholipids	213	3	0
	Sodium Chloride	721	23	1

	Triglycerides	1156	1	2
	Zinc	320	3	1
Bacteria	Neisseria meningitidis	9	43	0
Carbohydrate	Ascorbic Acid	774	2	0
Fungus	Candida albicans	24	45	0
	Saccharomyces cerevisiae	700	10	0
Gene	CA2	163	22	0
	EGFR	221	18	4
	SERPINB3	69	81	19
	TNFRSF6B	29	22	0
	TP53	939	103	33
Hormone	Adrenal Cortex Hormones	906	13	0
	Estradiol	268	1	0
	Estrogens	658	3	1
	Glucocorticoids	211	3	0
	Hydrocortisone	247	2	1
	Insulin	368	3	0
	Progesterone	129	6	0
	Testosterone	398	1	0
Lipid	Lipopolysaccharides	355	19	0
Neuroreactive	Dopamine	483	17	0
substance or	Norepinephrine	185	11	0
biogenic amine				
Pharmacologic	Amifostine	4	11	0
substance	Argipressin	68	111	0
	Aripiprazole	19	25	0
	Bleomycin	92	48	1
	Carboplatin	40	71	1
	Cetuximab	28	76	12
	Cisplatin	163	246	18
	Clotrimazole	1	12	0
	Cyclosporine	64	12	0
	Dexamethasone	143	14	0
	Docetaxel	33	40	3
	Fluconazole	8	60	0
	Fluorouracil	38	115	5
	Hydrogen Peroxide	466	7	0

	Indomethacin	57	3	0
	Iron	406	4	0
	Morphine	86	4	0
	Nystatin	2	11	0
	Paclitaxel	45	48	4
Virus	Hepatitis C virus	274	2	2
	Herpesvirus 4, human	68	67	5
	Human papillomavirus	687	704	210
	Human papillomavirus 16	130	113	21

Oral sex

Oral sex had a single SemMedDB triple found to be associated with OPC risk. This triple showed oral sex to predispose OPC risk, rather than having a "CAUSES" or "ASSOCIATED_WITH" predicate linking it to the disease.

From the retrieval of enriched literature intermediates between oral sex and OPC, there were 74 potential intermediates: 19 amino acid, peptide, or protein, 9 biologically active substances, 1 carbohydrate, 2 fungi, 5 gene regions, 8 hormones, 1 lipid, 2 neuroreactive substance or biogenic amine, 13 pharmacologic substances and 6 viruses (see **Table 3.5**).

Table 3.5 - Potential intermediate factors between oral sex and OPC

Category	**Intermediate**	**# Articles in oral sex search set**	**# Articles in OPC search set**	**Overlap**
Amino acid, peptide, or protein	Adiponectin	3	1	0
	Amino Acids	58	13	0
	Angiotensin II	15	6	0
	ATP8A2	13	15	0
	Collagen	16	6	0
	Epidermal Growth Factor Receptor	3	60	0
	Gamma-Aminobutyric Acid	88	4	0
	Human papillomavirus antibody	6	6	1
	Interferons	55	3	0
	Interleukin-1 beta	21	9	0
	Interleukin-4	9	4	0
	Leptin	41	4	0

	NF-kappa B	3	6	0
	Phosphoric diester hydrolase	10	31	0
	Retinoblastoma Protein|RB1	1	27	0
	Somatotropin	31	1	0
	Substance P	24	1	0
	Superoxide Dismutase	10	5	0
	Tumour Necrosis Factor-alpha	21	18	0
Biologically active	Calcium	31	10	0
substance	Cholesterol	55	6	0
	Fatty Acids	21	4	0
	Nitric Oxide	117	19	0
	Phospholipids	2	3	0
	Plasmids	14	10	0
	Sodium Chloride	186	24	0
	Triglycerides	28	3	0
	Zinc	16	4	0
Carbohydrate	Ascorbic Acid	15	2	0
Fungus	Candida albicans	13	45	0
	Saccharomyces cerevisiae	36	10	0
Gene	CA2	8	22	0
	CDKN2A	8	233	2
	EGFR	2	21	1
	IGHE	4	1	0
	TP53	11	136	0
Hormone	Adrenal Cortex Hormones	48	13	0
	Estradiol	984	1	0
	Estrogens	1249	4	0
	Glucocorticoids	95	3	0
	Hydrocortisone	151	3	0
	Insulin	37	9	0
	Progesterone	925	6	0
	Testosterone	2005	1	0
Lipid	Lipopolysaccharides	51	19	0
Neuroreactive	Dopamine	370	17	0
substance or	Norepinephrine	168	11	0
biogenic amine				

Pharmacologic substance	Angiotensin-Converting Enzyme Inhibitors	12	1	0
	Argipressin	13	15	0
	Bleomycin	7	49	0
	Cetuximab	1	88	0
	Cisplatin	9	264	0
	Cyclosporine	9	12	0
	Dexamethasone	30	14	0
	Fluconazole	8	60	0
	Fluorouracil	13	120	0
	Hydrogen Peroxide	3	7	0
	Indomethacin	25	3	0
	Iron	9	4	0
	Morphine	75	4	0
Virus	Hepatitis C virus	731	4	0
	Herpesvirus 4, human	39	72	0
	Human Papillomavirus	1567	823	91
	Human papillomavirus 16	158	123	11
	Human papillomavirus 18	37	8	0
	Human papillomavirus 6	46	11	0

3.4. Discussion

From this systematic retrieval of OPC risk factors and intermediates between OPC and risk factors, 5 phenotypes appear to be highly enriched in current PubMed literature: alcohol consumption, tobacco use, HPV infection, the HPV vaccine and sexual behaviour (namely oral sex). Intermediate SemMedDB concepts relating to pharmacological substances, specific gene regions and select molecular intermediates (e.g. hormones, amino acids) were also explored to augment the results from MELODI.

Quantifying HPV risk in observational literature

Although many journal articles quantify the prevalence of HPV within populations of individuals with OPC, very few attempt to quantify the risk of developing OPC from HPV infection. The largest of the few studies that does estimate risk of HPV infection on OPC incidence examines OPC cases and matched controls in the European Prospective Investigation into Cancer and Nutrition

cohort [227]. The study uses seropositivity of HPV16 E6 oncoproteins as a marker of current oncogenic infection. Of 135 OPCs, 88 showed HPV16 E6 seropositivity vs 47 showing HPV16 E6 seronegativity, with 1,599 controls (9 HPV16 E6 seropositive, 1590 HPV16 E6 seronegative). The odds ratio of having OPC given HPV16 E6 a seropositive status was 274 (95% CI: 110 to 681); a huge risk increase vs seronegative controls. Such evidence appears to support the predications obtained for HPV on OPC risk.

Quantifying alcohol risk in observational literature

The effect of alcohol consumption on OPC risk has been investigated by many observational epidemiological studies. In a large meta-analysis of 52 alcohol-OPC (combined as oral cavity and pharyngeal cancer) studies [17], alcohol consumption was split into frequency categories. Light (RR: 1.13; 95% CI: 1.00 to 1.26), moderate (RR: 1.83 95% CI: 1.62 to 2.07) and heavy (RR: 5.13; 95% CI: 4.31 to 6.10) consumption of alcohol all conferred an increased risk of these cancers vs never drinking. Furthermore, a study examining alcohol consumption in cancers of the pharynx found similar evidence [228]. Versus non/occasional drinking, light (OR: 1.39; 95% CI: 1.02 to 1.89), moderate (OR: 2.87; 95% CI: 1.91 to 4.30) and heavy drinking (OR: 5.70; 95% CI: 3.61 to 9.02) showed significant increase in risk of OPC. As with HPV, the effect of alcohol consumption on OPC risk from the studies above show consistent evidence supporting the SemMedDB triples obtained from this analysis – alcohol consumption is positively associated with OPC risk.

Quantifying smoking risk in observational literature

As with alcohol consumption, smoking has been investigated with respect to OPC risk in large, multi-centre meta-analyses of observational studies. In large meta-analyses of pharyngeal cancer, both current vs never smoking and former vs never smoking show marked effects on cancer risk [229] (current vs never RR: 6.76, 95% CI: 2.86 to 15.98, N studies = 7; former vs never smoking RR: 2.28, 95% CI: 0.95 to 5.50, N studies = 3). Similarly, current smoking was associated with cancer risk (OR: 5.83; 95% CI: 4.50 to 7.54) for oral cavity and oropharyngeal cancer combined in the Alcohol-Related Cancers and Genetic Susceptibility in Europe (ARCGSE) project; a multicentre case-control study in 10 European countries pertaining to over 4000 individuals [15]. Finally, a prospective study conducted by the International Head and Neck Cancer Epidemiology (INHANCE) consortium from 1981-2007, with over 3,800 cases and 18,000 controls [49] found ever vs never smoking in OPC to be associated with a threefold increase in cancer risk (OR: 3.01; 95% CI: 2.71 to 3.35). Smoking findings from observational

epidemiological studies appear to provide good evidence for the SemMedDB predications relating to an association between smoking and OPC risk.

Quantifying oral sex risk in observational literature

In the only meta-analysis of oral sex and oral cancer, six case-control studies and one cross-sectional study, relating to 5,553 individuals, indicated that there was no significant association between oral sex and risk of oral cancer (OR: 1.15; 95% CI 0.86 to 1.54; P: 0.33) [230]. The authors of this meta-analysis suggest that oral sex is a "risk behaviour" that could lead to HPV infection, which in turn could cause OPC, rather than that oral sex is an independent risk factor for oral cancer.

Example oral sex hypothesis network

The most associated intermediate (by overlap in PubMed literature) between OPC and oral sex is HPV. Sexual behaviour is an established risk factor for HPV-related OPC [231], with lifetime number of oral sex partners the factor most strongly associated with OPC [232]. When compared with patients with non-HPV-related OPC squamous cell carcinoma, patients with HPV-related OPSCC tend to be younger (aged <60 years) and do not have a history of sustained smoking and drinking [233]. Additionally, a higher percentage of them are men (in most regions), and report more oral sex partners and a higher socioeconomic status [231]. It is entirely plausible and generally accepted that differences in sexual behaviour could explain some of the differences in risk-attributable fractions observed across regions and across decades for HPV-positive OPC; demographics with more oral sex partners result in an increased number of high-risk HPV infections [234], resulting in a higher frequency of HPV-driven OPC [21].

Common intermediates between risk factors

From these results, it is clear that research needs to be undertaken to disentangle the causal pathways of the risk factors returned by MELODI. Between alcohol consumption, smoking, HPV infection and oral sex, there is considerable correlation. The risk factors retrieved by MELODI could represent an over-arching "risk-taking" phenotype which is difficult to disentangle from its constituent factors. This interpretation is also supported by the discovery of dopamine as an intermediate for each of alcohol consumption, smoking and HPV infection; dopamine agonists have been shown to lead to impulse control disorders and pathological gambling [235, 236]. It may be that smoking, alcohol and HPV infection affect dopamine levels, which in turn affect risk of OPC; elevated dopamine is commonly

reported in response to smoking [237] and alcohol consumption [238], but not for HPV infection. An alternative explanation that cannot be discounted in this analysis is that dopamine levels affect risk of smoking, alcohol consumption and sexual behaviour (leading to HPV infection), which in turn affect risk of OPC, highlighting a limitation of SemMedDB "concepts" vs "triples" – whilst use of "concepts" allows for recovery of a greater number of potential intermediates, directionality cannot be established between risk factor and outcome.

In observational literature, it has long been established that smoking and alcohol are highly correlated [25]. Alcohol and sexual activity are also correlated [239] – with a greater number of sexual partners, oral sex and propensity for HPV infection are generally more frequent in heavier/more frequent drinkers [231]. The shifting prevalence of HPV-driven OPC affecting younger individuals (<45 years old) seen in literature likely reflects the change in sexual (and social) behaviour in those demographics, more than it reflects a fundamental change in how sustained alcohol consumption, HPV infection or sustained smoking cause OPC. A key pitfall of observational epidemiological studies is the inability to isolate the effect of a single phenotype on disease risk. For each risk factor "triple" seen in the results section of this chapter, the other 4 are always considered intermediates by multiple studies. Confounding, reverse causation and many other epidemiological biases cannot be accounted for in observational study designs, adding to the necessity to establish true causal risk factor-OPC estimates. Mendelian randomization (MR) is a method that could overcome these limitations; by genetically-proxying phenotypes in a causal inference framework, direction of effect is established (genetic information is fixed at conception, thus reverse causation is not possible) and confounding is typically overcome (genetic information is inherited at random during meiosis, thus population distribution of genotypes should be naturally "randomized").

Furthermore, 200 of the 210 intermediates were shared between at least two of the risk factors retrieved by MELODI and, when examining the literature mentioning smoking, alcohol consumption and oral sex indexing concepts, the vast majority of articles returned investigate all three concepts as risk factors. Solely using the approach detailed in this chapter, it is unclear whether smoking, alcohol and HPV do indeed describe a singular "risk-taking" phenotype, where each risk factor mediates the others. However, more likely is that these "mediations" are an artefact of a saturation of published articles examining the three risk factors on OPC risk simultaneously, with SemRep parsing the abstracts of these articles to describe independent risk factors as cofactors for OPC risk. In the context of this cancer, the paucity of PubMed literature investigating other potential

risk factors/hypotheses (rather than highlighting combinations of established risk factors in different populations), prevents this data-driven approach from generating more hypotheses.

Exemplar data-driven hypothesis network using enriched predications

As an exemplar, overlapping intermediates have been combined (where possible) into a hypothesis "network", sharing common biological pathways. It should be noted, *all* intermediates discovered via MELODI will be used as an evidence base for analyses in latter chapters in this thesis.

The amino acid, peptide or protein showing the largest degree of overlap between HPV and OPC in PubMed literature was the epidermal growth factor receptor (EGFR) protein. *In vitro* literature suggests that HPV16 oncoproteins (an artefact of HPV16 infection, transmitted through oral sex) regulate the translocation of β-catenin via the activation of EGFR [240]. There is a correlation seen between HPV-positive OPC, levels of E5 and E6 oncoprotein and decreased membrane EGFR. Furthermore, lower membrane EGFR has been shown to be significantly associated with disease-free and overall survival. However, whilst lower membrane EFGR is associated with survival, activation of this receptor can cause translocation of β-catenin from the membrane to the nucleus of a cell, which is strongly associated with lymph node metastases [240]. It is therefore hypothesised that HPV-positive OPC, whilst associated with better survival than HPV-negative OPC, promotes early lymph node (and potentially distant) metastatic disease via activation of EGFR and subsequent translocation of β-catenin, through its E5 and E6 oncoproteins [241]. These findings also establish β-catenin as a distinct biomarker for HPV-positive OPC. Contrarywise, smoking and alcohol have been reported to increase expression of membrane EGFR [241-243], potentially explaining part of the decreased survival in more "traditional" OPC (caused by sustained tobacco and alcohol consumption) compared to HPV-driven OPC [11].

TP53 shared the most "gene" indexing articles between HPV and OPC search sets. TP53 is a well-established oncogene in a pan-cancer context with many associated biological pathways [244]. In the case of HPV-positive OPC, one of the many roles TP53 plays is to dysregulate the induction of hBD3 expression in human oral epithelial cells and oral cancer cell lines via HPV16 E6 when compared to E6 from non-oncogenic HPV types [245]. The hBD3 peptide is an epithelial cell-derived antimicrobial and immunoregulatory peptide that, under normal conditions, acts to defend mucosal surfaces from microbial challenges [246]. Clinically confirmed HPV-positive head and neck cancers overexpress hBD3, which appears to be induced by activated EGFR via TP53. HPV-positive cancers appear to recruit

and activate tumour-associated macrophages in the tumour microenvironment through hBD3, contributing to cancer progression [245]. As such, hBD3 may be a therapeutic target for OPC.

The pharmacologic substance with the most overlap between HPV and OPC search sets in this analysis was cetuximab – an EGFR monoclonal antibody used to prevent overexpression of EGFR in HNC patients in conjunction with radiotherapy. Interestingly, cetuximab seems to show inferior efficacy in HPV-positive OPC compared to HPV-negative OPC, potentially an artefact of decreased expression of membrane (due to translocation to the nucleus; see above) EGFR seen in those with HPV. As proof of principle of these connected hypotheses, recently, cisplatin has been advocated above cetuximab due to marked overall (77·9% [95% CI: 73·4 to 82·5] in the cetuximab group versus 84·6% [95% CI: 80·6 to 88·6] in the cisplatin group) and progression-free survival (5-year progression-free survival: cetuximab 67·3% [95% CI: 62·4 to 72·2] *vs* cisplatin 78·4% [95% CI: 73·8 to 83·0]) seen in a trial of 849 individuals with HPV-positive OPC, comparing the two chemotherapeutic agents [97, 98].

Strengths

The method outlined in this chapter is a rapid, systematic approach to obtain risk factors relating to an outcome of interest. In the context of epidemiology, when attempting to generate novel hypotheses or relationships, advanced data mining methodology is becoming increasingly important. Systematic, automated approaches offer enormous potential to assist in identification of existing evidence and prioritization of mechanisms to investigate, examining vast amounts of publication data using pre-calculated literature objects, such as SemMed predications. The scale and size of high-throughput data mining techniques simply cannot be matched by manual effort.

A key strength of using MELODI is that it can appraise intermediates between a risk factor and OPC. Results obtained from this approach provide the number of individual studies relating to a certain indexing concept for each of two search sets, whilst also providing an estimation of shared literature (i.e. an article describing both concepts simultaneously, greater than chance) between concepts. Shared literature represents existing evidence of an intermediate between a risk factor and OPC, whilst no shared literature could indicate a novel pathway. From the combined results across all risk factors, 71 intermediates showed evidence of association with OPC from published literature, with 139 intermediates showing no previous evidence. By simultaneously estimating the degree of evidence for existing risk factor pathways and deriving novel hypotheses for appraisal, MELODI is less confined by published literature as traditional methods of risk factor ascertainment (for example, a

systematic review); traditional methods typically cannot link a risk factor and outcome which are a step removed from each other on a causal pathway, whereas MELODI can.

Limitations

The resolution of SemMedDB triples and the inclusion of a PREDICATE between concepts allows for directionality (proposed by the author of an article in an abstract) to be established between potential risk factor/intermediate and OPC when one is inferred. However, the efficacy of retrieval of SemMedDB concepts and triples from literature is largely dependent on how an abstract has been parsed by the SemRep program. As mentioned previously, SemRep infers predications from PubMed article abstracts. As such, an article would have to explicitly state that a risk factor was associated with OPC risk in its abstract for a triple to be inferred (e.g. "mouthwash was associated with increased risk of oropharyngeal squamous cell carcinoma in our analysis"). Combined with MELODI filtering results to remove background noise, risk factors and intermediates actually associated with OPC risk, but rare, unique, or phrased in a unique way, would likely be filtered out of the results of a search set comparison. This is a key limitation of this method; it is evidenced by a small number of epidemiologically-relevant triples returned for a small number of already well-established risk factors, in the case of OPC. Despite the lack of discovery of novel modifiable risk factors, those that have been retrieved by MELODI explain a significantly high proportion of the variance explained of OPC risk, validating this method of literature mining.

In observational literature, many articles investigating risk factors relating to OPC tend to combine HNC subtypes. Whilst HNCs are proximal to each other, they are not always aetiologically similar. For example, Epstein-Barr virus is associated with nasopharyngeal cancer, whilst unassociated with other HNC cancer sub-types. Equally, HPV16 seems to have a much more significant association with OPC than any other HNC sub-type [247]. Combining HNCs raises two key issues for the method presented in this chapter; the first is that SemRep may not infer the association of a risk factor with OPC if it is grouped ambiguously, the second is that SemRep may infer an association that is not genuine for the same reason. Whilst MELODI filters out background noise and "triples" were filtered to only include those relating to OPC uniquely, it cannot be ruled out that some OPC risk predications were either not genuine, or simply missed.

An important limitation of any literature-based tool is that the published literature may be a biased subset, or a biased over-representation, of research that has been undertaken. A large

proportion of negative findings are never published, and groups often publish many related papers with similar ideas discussed in the abstract. In addition, the algorithms used to produce the SemMedDB data and the humans used to assign the MeSH terms may introduce unconscious bias. Using more flexible agnostic methods such as those mentioned above would enable the use of other publicly available data sets, alleviating some of the bias associated with published literature. Even so, MELODI is always going to give a biased representation of what is really known about a topic since it is based on published literature. However, the alternative to a computational approach is manual curation, which is impossible at this scale and potentially prone to much greater bias. As long as the caveats and limitations are understood, then the output of this kind of approach can still be valuable and provide reliable hypotheses.

A key limitation of this analysis is that it was restricted to PubMed literature prior to 30th June 2018. Whilst PubMed at this timepoint contains over 27.8 million articles (www.melodi.biocompute.org.uk), it does not encapsulate the entirety of scientific literature. Other biomedical literature repositories may contain information regarding OPC risk, such as Scopus, which contains over 69 million records [248]. However, whilst Scopus is a larger database, PubMed (including MEDLINE) is the default repository for biomedical articles and is currently the only database parsed for SemMedDB data [140], making it the only database available for use with this method.

3.5. Conclusion

In summary, MELODI provides a means of systematically interrogating vast amounts of literature data at a speed manual curation simply cannot match. Automated, systematic retrieval of risk factors and intermediates for downstream investigation, using pre-calculated literature objects, is extremely effective for rapid hypothesis generation. For the purposes of this thesis, gene and protein intermediate findings are of particular interest, as they may be able to be incorporated (using quantitative trait loci) into a Mendelian randomization framework to assess robust causality. Additionally, the list of risk factors and intermediates provide a robust evidence base and comparison for future chapters, where data-driven approaches will be incorporated to agnostically retrieve intermediates. It should be noted that this approach is not without weaknesses. Specifically, it is reliant on correct parsing of literature abstracts by SemRep, and the MELODI enrichment step will filter out any unique OPC risk factors if they are present in literature “less than chance”. In this respect, MELODI is overly conservative. However, for the purposes of deriving a robust evidence base from which to compare future analyses, it is a suitable method of risk factor and intermediate retrieval from literature.

CHAPTER 4. A PHENOME-WIDE MENDELIAN RANDOMIZATION STUDY OF OROPHARYNGEAL CANCER USING SUMMARY GENETIC DATA

4.1. Introduction

In this chapter, 17,449 phenotypes proxied by 647,283 genetic variants are examined in relation to oropharyngeal cancer risk in an MR-PheWAS framework. Exposure data is combined into 4 broad groups: metabolites, immune cell traits, UK Biobank traits (see 4.2.1 – Phenotype grouping for an explanation of trait origins) and non-Biobank traits. Sensitivity analyses are conducted to ensure that any key assumptions of MR methodology are not violated. Finally, results surpassing a stringent FDR correction for multiple testing are then explored in current literature to ascertain biological plausibility. A relaxed p-value threshold is given to exposures pertaining to exposures (smoking, alcohol consumption, oral sex and HPV infection) presented in Chapter 3.

Marked geographic and temporal differences in distribution and pattern of OPC indicates that a large proportion of these cancers are potentially avoidable. However, the nonspecific nature of OPC symptoms results in a high frequency of late-stage diagnoses and underscores the need for an effective screening programme with robust risk stratification. Tobacco smoking, alcohol consumption, and viral infection are among the major risk factors for OPC, with tobacco smoking and alcohol consumption reported as having synergistic effects on this disease. However, these reports are based on observational epidemiological studies, which are prone to unmeasured or residual confounding and reverse causation, precluding robust causal inference. Furthermore, conventional epidemiological studies often test a narrow set of hypotheses using prior subject knowledge, typically based on other observational studies. Whilst essential, hypothesis-driven approaches can constrict a field of research, and preoccupation with previously-hypothesised risk factors for a given disease can introduce publication bias [249], preventing both the identification of novel risk factors and de-prioritization of non-causal associations.

Mendelian randomization (MR) is a well-established type of instrumental variable (IV) analysis that addresses some of the shortcomings of conventional observational studies by using genetic anchors to appraise the causal relevance of exposures on disease [250-254]. It is an increasingly recognised and powerful tool for identifying causes of a broad spectrum of outcomes, including cancer [255, 256]. Traditional MR studies focus on hypothesized exposure-outcome combinations, often using phenotype data collected on all participants in a single sample. Two-sample MR, using summary-level data from published genome-wide association studies (GWASs), greatly extends the scope of the approach [125], allowing causal appraisal of hypothesized exposure-outcome associations using gene-exposure and gene-disease associations collected in separate studies [257-261]. Further, the two-sample MR method can be extended to appraise causality in a hypothesis-free manner, appraising 1-

to-many, many-to-1 or many-to-many exposure-outcome combinations, in an approach known as a MR phenome-wide association study (MR-PheWAS) [262, 263].

Here, MR-PheWAS was used to screen the phenome for potential causes of OPC. Our aims were twofold: to identify potentially novel causes of OPC that may not have been captured using previous epidemiological approaches, and to prioritise hypotheses identified in current literature.

4.2. Methods

4.2.1. Data preparation

OPC data

GWAS data from people of European descent with OPC and matched controls were obtained from the International Head and Neck Cancer Epidemiology Consortium (INHANCE; 12 studies; 6034 cases, 6585 controls). Cancer cases comprised the following ICD codes: oral cavity (C02.0-C02.9, C03.0-C03.9, C04.0-C04.9, C05.0-C06.9) oropharynx (C01.9, C02.4, C09.0-C10.9), hypopharynx (C13.0-C13.9), overlapping (C14 and combination of other sites) and 25 oral or pharyngeal cases with unknown ICD code (other). The samples were originally genotyped using Illumina OncoArray, designed for cancer studies by the OncoArray Consortium, part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. The majority of samples were genotyped as part of the oral and pharynx cancer OncoArray, with the exception of 2,476 shared controls (1,453 from the European cohort study and 1,023 from the Toronto study) that were genotyped at the Center for Infectious Disease Research (Seattle, Washington, United States), but as part of the Lung OncoArray. Genotype calls were made by the Dartmouth team in GenomeStudio software (Illumina, Inc.) using a standardized cluster file for OncoArray studies.

Initial quality control steps and analyses were performed at the International Agency for the Research of Cancer (IARC), Lyon. After removing duplicates, related samples, samples with sex discrepancy and population outliers, genotype imputation was performed using the Michigan Imputation Server [264]. Genotypes were pre-phased (i.e. their haplotypes were inferred) using SHAPEIT v2 [265] and imputed with Minimach v3 [266] using the Haplotype Reference Consortium panel [267]. After imputation, SNPs with an imputation quality (R^2) lower than 0.7 were removed from the datasets. Effect estimates for OPC risk were obtained after adjusting for age, sex and significant principal components for population stratification using R software (R version 3.3.1). Results were

calculated, at IARC, for site of OPC (overall oral cancer and pharynx cancer, site-specific oral cancer and oropharyngeal cancer) were then combined using a fixed-effects inverse-variance approach implemented in PLINK [268].

Genetic instruments for phenotypes

Two-sample MR was conducted using the TwoSampleMR R package [269], in R version 3.5.1. Genetic data on cognitive, anthropometric, metabolic, immune and behavioural phenotypes were obtained from the MR-Base database of harmonised GWAS summary data (**Figure 4.1**). Those phenotypes possessing robust genetic proxies (defined as $P<5e^{-8}$) with which to conduct MR analyses were considered for further analysis (N=21,158). Duplicate (N=477) and non-European studies (N=16) were excluded from the analysis at this stage, leaving 21,074 potential phenotypes for analysis. Genetic instruments for each phenotype were single-nucleotide polymorphisms (SNPs) independently associated with the phenotype of interest after linkage disequilibrium (LD) clumping (radius = 10,000kb; $r^2 = 0.001$). For each identified SNP, the reported effect size was expressed as a one standard deviation (SD) increase in the level of the phenotype per risk allele, along with the standard error (SE). In the case of a binary phenotype (e.g. presence or absence of coronary heart disease), the reported effect size was expressed as a log odds ratio (OR). For studies in which the genetic effects were not originally reported in SD units of the phenotype, these were recalibrated according to the mean and SD reported in the original study and, if appropriate, weighted for sample size across the different studies contributing to the meta-analysed GWAS for a phenotype. For each genetic variant associated with the identified phenotypes, effect-estimates and SEs were then extracted from the summary genetic data for OPC. Phenotypes without genetic variants at $P< 5x10^{-8}$ (N = 2,540) were excluded from analysis. To harmonise the data, effect alleles in the OPC summary data were coded to reflect the phenotype increasing allele, using allele frequencies to resolve strand ambiguities for palindromic SNPs (A/T or C/G). Those phenotypes that did not have genetic variants in the OPC GWAS were excluded (N = 669; not imputed in the dataset), resulting in a final list of 17,449 phenotypes on which to perform MR analyses.

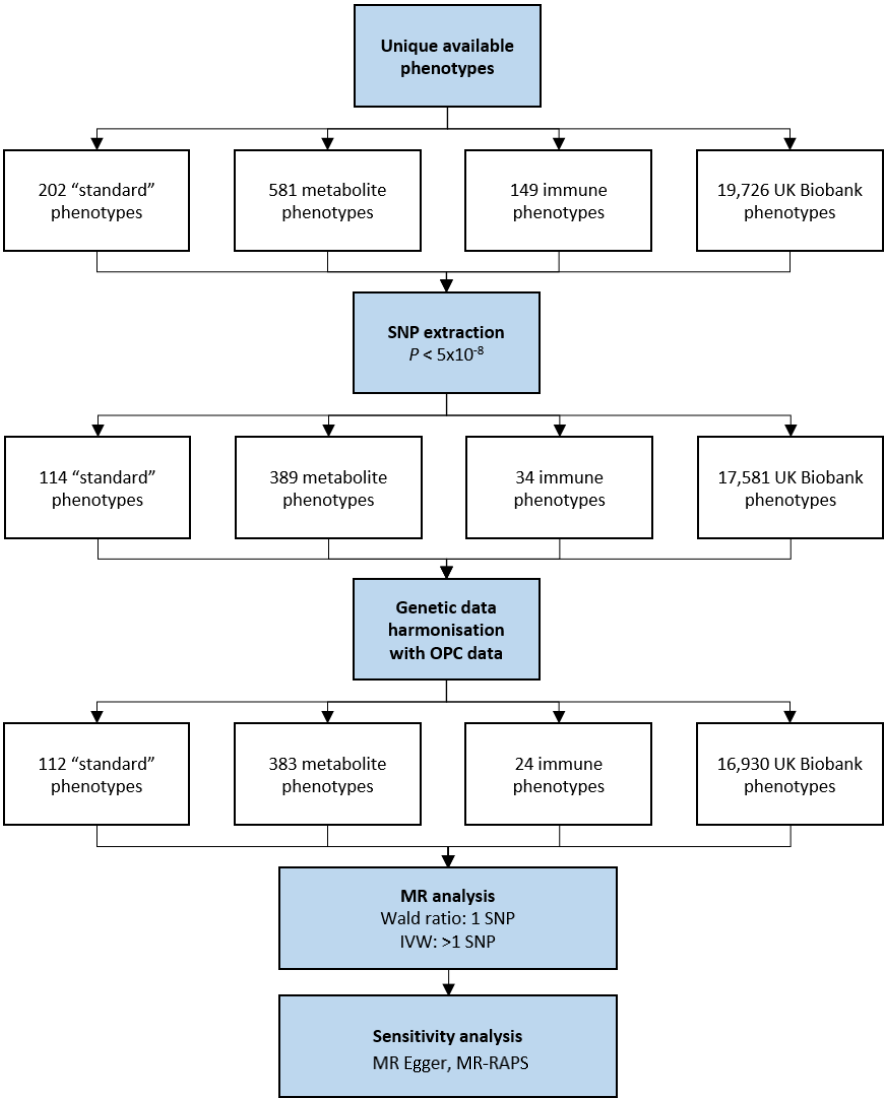


Figure 4.1 - Flowchart detailing phenotype extraction process for MR-PheWAS. Phenotypes were extracted from MR-Base using R v3.5.1 in August 2019. IVW: Inverse variance weighted; RAPS: Robust Adjusted Pleiotropy Score

Phenotype grouping

Phenotypes were grouped into 4 categories: metabolites, immune traits, UK Biobank traits, and traits not defined by the other categories (defined as "standard", non-UK Biobank, immune or metabolite traits). Metabolic traits included those from Shin et al. [270], Kettunen et al. [271], Lemaitre et al. [272], Guan et al. [273], Wu et al.[274], Mozaffarian et al. [275], Dastani et al. [276], Paterson et al. [277] and Kilpelainen et al. [278], corresponding to 383 phenotypes. Metabolites were grouped due to their vast numbers, highly correlated nature and the rising significance of the metabolome's effect on disease in literature. Immune traits consisted of phenotypes from Roederer et al. [279], who examined the genetic contribution to different proportions of immune cell subtypes. Immune traits (N = 24) were grouped separately due to their highly correlated nature and due to the

large proportion of attributable risk of OPC given to HPV infection. UK Biobank traits consisted of 16,930 phenotype instrumental variables (IVs) generated by the Integrative Epidemiology Unit at the University of Bristol, UK and by Neale Laboratories in Boston, Massachusetts. Both institutes conducted GWAS using pipelines incorporating the PHEnome Scan Analysis Tool (PHESANT). PHESANT was created to systematically perform PheWAS in UK Biobank, automating the coding of UK Biobank variables and testing their association with genetic information. Where possible, all phenotypes were converted into normally-distributed quantitative or binary categorical variables and GWAS models were adjusted for principal components, sex, age, age^2, *sex*age and *sex*age^2. Binary traits were regressed using BOLT-LMM; a linear mixed-model approach which accounts for relatedness and population stratification. BOLT-LMM outputs betas as an absolute risk difference, therefore to allow for estimation of log odds ratios, beta values from UK Biobank binary phenotype GWAS were converted prior to MR using the following formulae:

$$\mu = \frac{n_{case}}{(n_{case} + n_{control})}$$

$$\log(OR) = \frac{\beta_{BOLT}}{\mu(1-\mu)}$$

Where β_{BOLT} is the BOLT-LMM output beta

$$\text{se}(\textbf{log}OR) = \frac{se_{BOLT}}{\mu(1-\mu)}$$

Where SE_{BOLT} is the BOLT-LMM standard error

Non-Biobank traits consisted of 112 phenotypes. This grouping of phenotypes contained neurological, anthropometric and disease traits, including common modifiable epidemiological phenotypes such as body mass index, smoking and educational attainment.

Finally, phenotypes pertaining to alcohol consumption, smoking, sexual activity and HPV infection (**see Box 4.1**) were investigated in a separate analysis due to existing observational epidemiological evidence of association with OPC from MELODI (see Chapter 3). For these phenotypes, a multiple-testing correction was not necessary (see 1.2.2); they were treated as individual analyses. Other phenotypes were corrected using

Box 4.1 - Phenotypes used to proxy risk factors identified by MELODI as associated with OPC

The following phenotypes were matched with risk factors found to be enriched in current epidemiological literature by MELODI. From each broad category of phenotypes in this analysis, the keywords “smok*”, “ciga*”, “alc*”, “drink*”, “beer”, “wine”, “spirit*”, “sex*”, “intercourse”, “HPV” and “papillomavirus” were used to systematically retrieve potential phenotype matches. Where similar phenotypes were retrieved, the IV with the largest sample size or largest number of SNPs was used, resulting in 13 phenotypes being extracted from the UK Biobank dataframe of exposures.

Phenotype	**#SNPs**	**Matching phenotype**
Age first had sexual intercourse	189	Sexual activity
Average weekly beer plus cider intake	20	Alcohol consumption
Average weekly champagne plus white wine intake	4	Alcohol consumption
Average weekly red wine intake	17	Alcohol consumption
Average weekly spirits intake	4	Alcohol consumption
Current tobacco smoking	34	Smoking
Ever smoked	77	Smoking
Lifetime number of sexual partners	61	Sexual activity
Nondependent abuse of alcohol	64	Alcohol consumption
Nondependent abuse of tobacco	15	Smoking
Pack years of smoking	10	Smoking
Papillomavirus as the cause of diseases classified to other chapters	3	HPV infection
Past tobacco smoking	92	Smoking

4.2.2. Statistical analysis

Mendelian randomization analyses

We used fixed-effects inverse-variance weighted (IVW) [280, 281] MR analyses when the number of SNPs available to instrument a phenotype was greater than 1. IVW is an established, reliable MR analysis method when using summary genetic data with phenotype instruments containing multiple SNPs [125]. For phenotypes instrumented by a single SNP, we derived Wald ratio effect estimates [125, 282]. Results were expressed as odds ratios (ORs) with a corresponding 95%

confidence interval (CI) per 1 standard deviation (SD) increase in continuous traits (e.g. height), and as ORs with 95% CI per a doubling in odds for binary traits (e.g. type 2 diabetes).

Multiple testing correction

To account for the large multiple testing burden accrued by the MR-PheWAS, analysis p-values were adjusted via the Benjamini-Hochberg FDR correction [169]. An FDR-adjusted p-value of 0.05 was used to determine results with strong statistical evidence to support an association between a phenotype and OPC.

Sensitivity analyses

MR-Egger regression [283] was used as a sensitivity analysis to detect bias due to horizontal pleiotropy in the causal estimates. Horizontal pleiotropy is where a genetic variant affects the outcome via a different biological pathway from the phenotype under investigation and is a violation of a key assumption of MR. MR-Egger regression performs a weighted linear regression of the SNP-disease and SNP-phenotype associations, the intercept of which is not constrained to the origin and can therefore be used to detect and estimate the magnitude of horizontal pleiotropy [283]. Due to the lack of constraint to the origin, deviation from the origin in an MR-Egger regression may suggest the effect of the SNP is operating via a separate pathway. MR-Egger regression relies on the existence of at least three SNPs to estimate a linear relationship. Additionally, we used MR Robust Adjusted Pleiotropy Score (MR-RAPS) to detect the presence of the many weak instruments bias.

4.3. Results

All phenotype-OPC associations below an FDR-corrected p-value of 0.05 can be seen in **Table 4.1**. Results are described in more detail, per grouping, below.

Metabolite trait grouping

Using a conservative FDR p-value correction for multiple testing, of the 383 metabolite traits analysed, none of our results showed a corrected p-value of association below 0.05. All results from this analysis can be seen in **Figure 4.2**.

Table 4.1 - Phenotype associations with OPC displaying sufficient evidence of association below an FDR-corrected p-value of 0.05. 95% confidence intervals (95% CI) and p-values are shown for each phenotype, in addition to the number of SNPs used in the IV, the variance explained by the IV in the phenotype of interest and the broad phenotype grouping the exposure belongs to in this analysis. OR: Inverse-variance weighted odds ratio for the effect of the exposure on incidence of OPC. Units are standardised - continuous traits are in standard deviation units; binary traits are in log odds units.

Phenotype	OR	95% CI	FDR P-value	Variance explained	#SNPs	Grouping
Alcohol consumed (binary)	10.4	4.8 to 22.6	1.42×10^{-5}	1.24%	1	UK Biobank
Former alcohol drinker (binary)	11.7	5.2 to 26.6	1.42×10^{-5}	1.23%	1	UK Biobank
Alcohol drinker status: never (binary)	5.0×10^{-21}	8.5×10^{-28} to 3.0×10^{-14}	1.42×10^{-5}	1.03%	1	UK Biobank
Red wine intake (units/wk)	178.1	31.6 to 1002.8	1.42×10^{-5}	2.35%	1	UK Biobank
Alcohol (units/wk)	154.6	28.8 to 830.6	1.42×10^{-5}	1.67%	1	UK Biobank
Non-cancer illness code, self-reported: bronchiectasis (binary)	0.55	0.45 to 0.68	7.54×10^{-5}	2.00%	2	UK Biobank
Ischemic stroke (binary)	5.21	2.47 to 11.0	1.76×10^{-3}	2.93%	1	Non-Biobank
Treatment/medication code: flecainide (binary)	1.29	1.16 to 1.43	4.10×10^{-3}	2.00%	2	UK Biobank
Treatment/medication code: sotalol (binary)	1.55	1.27 to 1.89	0.027	1.00%	1	UK Biobank
Qualifications: College or University degree (binary)	0.72	0.61 to 0.83	0.027	1.29%	244	UK Biobank
PCT where patients GP was registered: SWALE PCT (binary)	1.00	1.002 to 1.007	0.027	2.11%	114	UK Biobank
Age first had sexual intercourse (years)	0.44	0.30 to 0.64	0.037	2.07%	189	UK Biobank
Operative procedures - main OPCS: P09.1 Biopsy of lesion of vulva (binary)	1.72	1.33 to 2.22	0.040	1.00%	1	UK Biobank
Age at first live birth (years)	0.23	0.11 to 0.46	0.040	2.02%	34	UK Biobank
Diagnoses - main ICD10: D64.9 Anaemia, unspecified (binary)	2.77	1.71 to 4.48	0.041	1.01%	1	UK Biobank
Weight (kg)	0.42	0.25 to 0.70	0.046	18.03%	8	Non-Biobank

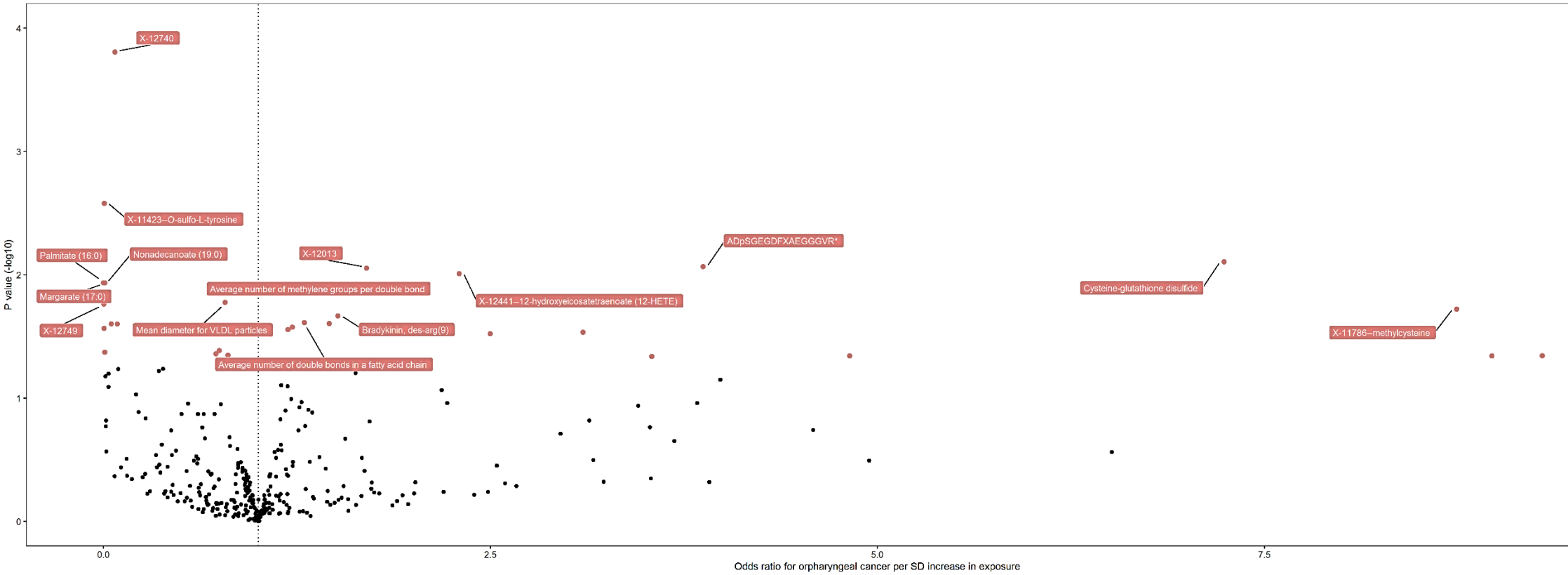


Figure 4.2 - Volcano plot showing the odds ratio derived from MR analyses of 383 metabolic phenotypes against incident OPC across the x-axis and a corresponding MR analysis p-value (-log10 scale) on the y-axis. Units are standardised - continuous traits are in are in standard deviation units, whereas binary traits are in log odds units. Small red points denote analyses with an unadjusted p-value < 0.05. Large red points denote analyses with a Bonferroni-adjusted p-value < 0.05

Immune trait grouping

None of the 24 immune traits investigated showed any evidence of association with OPC below our FDR correction. Results of all immune trait phenotype-OPC associations can be seen in **Figure 4.3**.

UK Biobank trait grouping

Fourteen of the 16,930 UK Biobank traits analysed showed evidence of association with OPC incidence below the FDR multiple testing threshold. Of these findings, 8 analyses were proxied by a single SNP, hence sensitivity analyses could not be conducted. The other 6 analyses were conducted using an IVW regression and allowed for 1 or more sensitivity analyses to be performed. Of the phenotypes below $P_{FDR} = 0.05$, alcohol consumption appeared to show dramatic ORs for OPC incidence (alcohol consumed OR = 10.4; former alcohol drinker OR = 11.7; alcohol drinker status: never OR = 5.0×10^{-21}; red wine intake OR = 178.1; alcohol OR = 154.6). However, none of these findings allowed for sensitivity analyses due to single genetic proxies per phenotype. Age at first sexual intercourse and age at first live birth both showed inverse associations with OPC incidence (age first had sexual intercourse OR = 0.44; age at first live birth OR = 0.23). Both of these phenotypes possessed >1 genetic proxy, and both had consistent directions of effect between IVW, MR Egger and MR-RAPS analyses (age first had sexual intercourse IVW OR = 0.44; Egger OR = 0.60; RAPS OR = 0.42 | age at first live birth IVW OR = 0.23; Egger OR = 0.11; RAPS OR = 0.22). The magnitude of effect differed for both these phenotypes for the MR Egger findings, though low statistical power for Egger may explain these discrepancies.

Non-Biobank trait grouping

Of the non-Biobank phenotypes, 2 traits showed evidence of association at an FDR-corrected p-value < 0.05. Ischemic stroke showed a strong positive association with OPC (OR: 5.22; 95% CI: 2.47 to 11.05; P_{FDR}: 1.8×10^{-3}). This phenotype was instrumented by a single SNP, hence corresponds to a Wald ratio estimate. Weight showed a negative association with OPC (OR: 0.42; 95% CI: 0.25 to 0.70; P_{FDR}: 0.046) Weight was instrumented by 8 SNPs, thereby was estimated using the IVW method. Additionally, high grade serious ovarian cancer showed a positive association with OPC at a suggestive FDR-corrected p-value < 0.1 (OR: 1.24; 95% CI: 1.08 to 1.43; P_{FDR}: 0.09). All results for this grouping are shown in **Figure 4.4**.

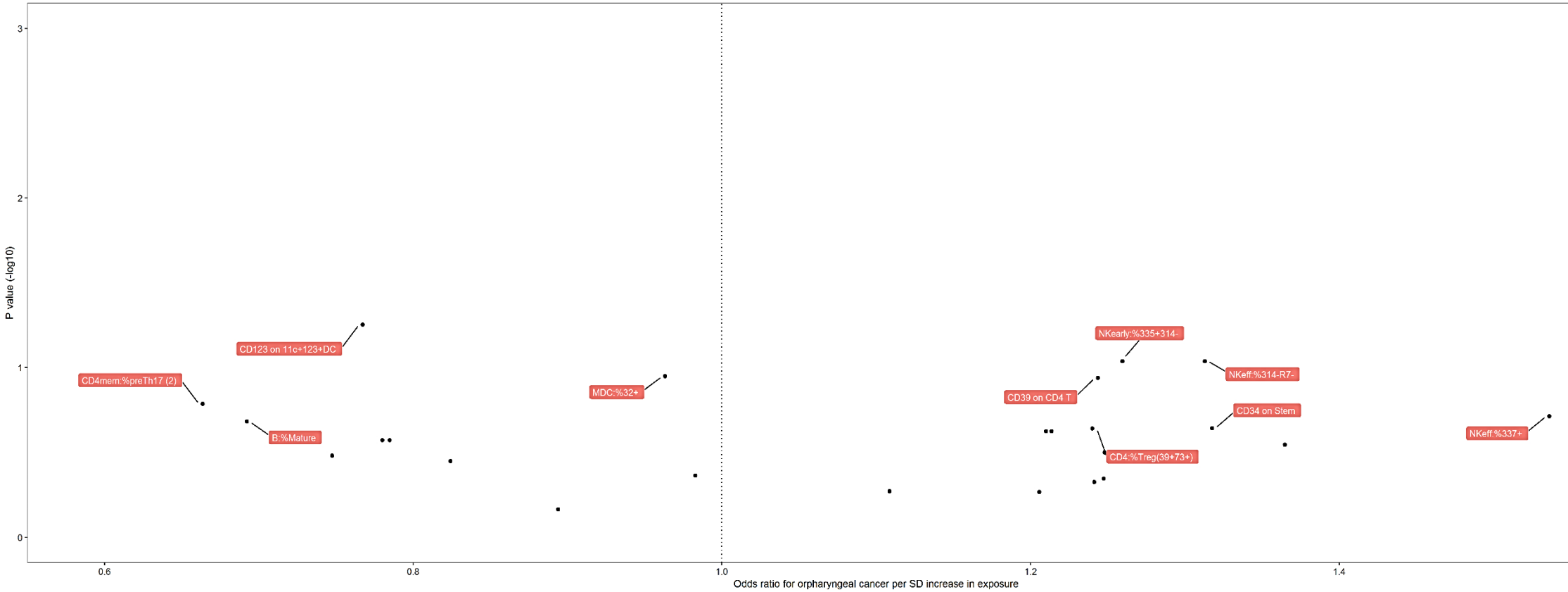


Figure 4.3 - Volcano plot showing the odds ratio derived from MR analyses of 24 immune phenotypes against incident OPC across the x-axis and a corresponding MR analysis p-value (-log10 scale) on the y-axis. Units are standardised - continuous traits are in are in standard deviation units, whereas binary traits are in log odds units. Small red points denote analyses with an unadjusted p-value < 0.05. Large red points denote analyses with a Bonferroni-adjusted p-value < 0.05

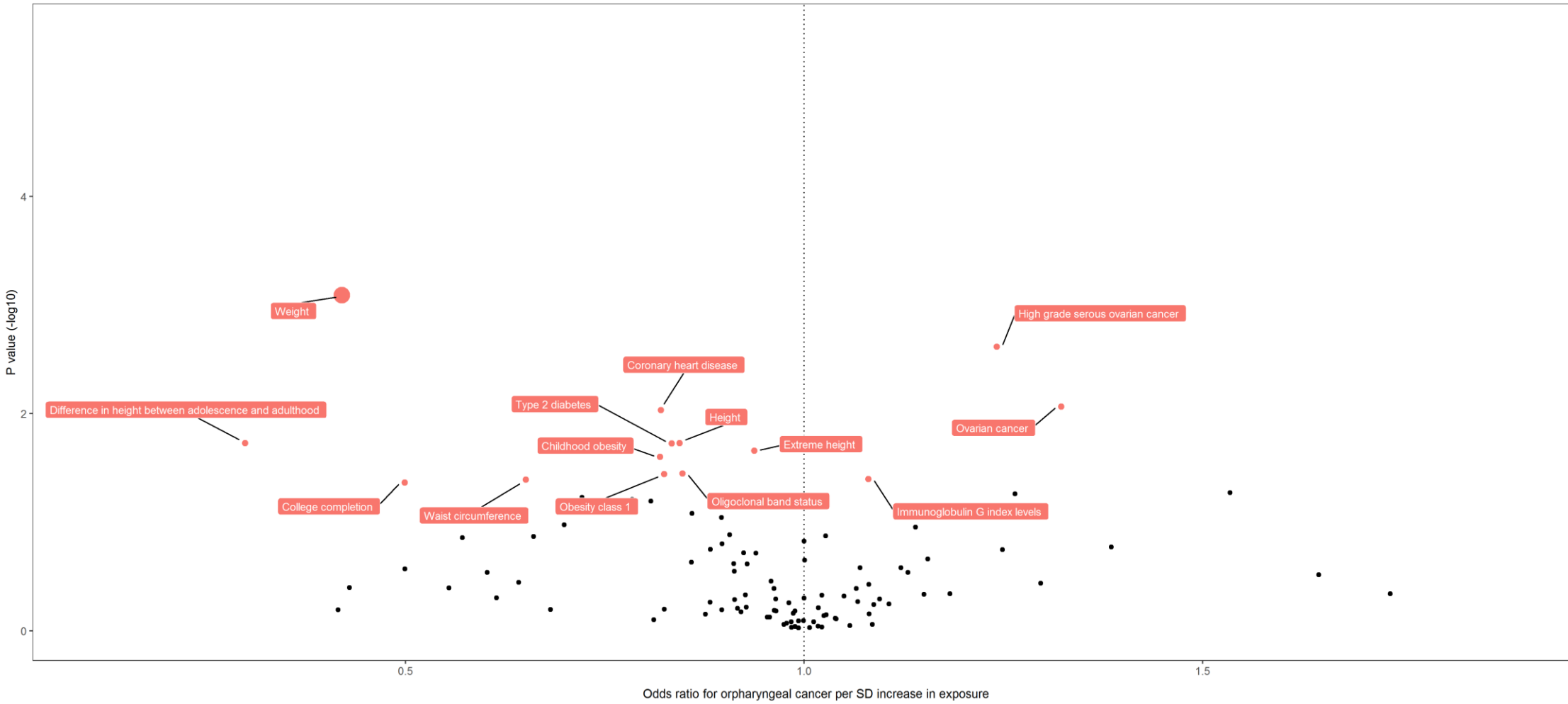


Figure 4.4 - Volcano plot showing the odds ratio derived from MR analyses of 112 "standard" (non-UK Biobank) phenotypes against incident OPC across the x-axis and a corresponding MR analysis p-value (-log10 scale) on the y-axis. Units are standardised - continuous traits are in are in standard deviation units, whereas binary traits are in log odds units. Small red points denote analyses with an unadjusted p-value < 0.05. Large red points denote analyses with a Bonferroni-adjusted p-value < 0.05

Phenotypes possessing observational epidemiological evidence (MELODI)

Of the 13 phenotypes matched to MELODI risk factors for OPC, 5 showed an FDR adjusted p-value of association below 0.05 (**Table 4.2**).

Table 4.2 - Association between MELODI-derived risk factors and OPC risk. OR: odds ratio, CI: confidence interval

Phenotype	OR	95% CI	P-value	#SNPs
Age first had sexual intercourse	0.44	0.30 to 0.64	2.4×10^{-5}	189
Current tobacco smoking	11.1	2.35 to 52.3	2.4×10^{-3}	34
Average weekly beer plus cider intake	21.5	2.87 to 161.8	2.8×10^{-3}	20
Ever smoked	1.46	1.06 to 2.00	0.02	77
Lifetime number of sexual partners	2.31	1.03 to 5.12	0.04	61
Past tobacco smoking	0.71	0.46 to 1.08	0.11	92
Nondependent abuse of alcohol	1.00	0.99 to 1.01	0.21	64
Average weekly red wine intake	2.22	0.24 to 20.9	0.48	17
Papillomavirus as the cause of diseases classified to other chapters	1.03	0.94 to 1.12	0.54	3
Average weekly spirits intake	0.21	0.0 to 38.4	0.55	4
Pack years of smoking	0.89	0.36 to 2.17	0.79	10
Nondependent abuse of tobacco	1.00	0.98 to 1.02	0.87	15
Average weekly champagne plus white wine intake	0.80	0.01 to 44.3	0.91	4

4.4. Discussion

We undertook an MR-PheWAS of the association of 17,449 phenotypes with OPC, across 4 broad groupings, pertaining to cognitive, anthropometric, clinical, metabolic, immune and behavioural phenotypes. We provide evidence that 16 of the 17,449 phenotypes we tested were associated with OPC, pertaining to alcohol consumption, sexual behaviour, disease, medication use, weight and educational attainment.

The inverse association seen between weight and risk of OPC is concurrent with findings from conventional observational studies investigating body fatness on the disease [80, 81, 219, 284, 285]. Across these studies, ORs were consistently above 2.0 for any measure of lower body fatness against risk of OPC, including weight and BMI quartiles, even amongst never-smokers and never-drinkers

[284]. In this PheWAS, elevated weight is associated with a ~2.3-fold decrease in OPC risk. Results did not change substantially in sensitivity analyses that made allowance for violations of MR assumptions, thus are compatible with a true inverse association between weight and risk of OPC. Findings for other body fatness phenotypes almost exclusively show the same inverse direction of association with OPC although do not reach the FDR adjusted p-value threshold used for filtering. In a study assessing weight loss at time of diagnosis in HNC patients, it was found that 20% of all HNC patients suffered from weight loss of either >5% in a month or >10% in 6 months [286]. Other studies examining weight loss report prevalence varying from 31% to 57% [287-289]. Weight loss has been linked to presence of typical HNC symptoms such as dysphagia/passage difficulties, loss of taste/aversion and loss of appetite, appearing to be particularly frequently observed in hypopharyngeal, oropharyngeal and supraglottic laryngeal cancer [286]. Thus, our finding of weight being inversely associated with OPC may be interpreted as "lack of weight loss" being inversely associated with OPC. Alternatively, it has been hypothesised that HNC patients may suffer from the anorexia-cachexia syndrome [290]; where elevated cytokine production in those with HNC [291, 292] cause a loss of taste and appetite. The inverse association of weight with OPC may also proxy for a lack of anorexia-cachexia syndrome; further investigation into this phenomenon is required to disentangle these phenotypes.

Dramatic findings were observed between genetically-proxied alcohol phenotypes and risk of OPC. ORs for these phenotypes ranged from 10.4 to 178.1 for alcohol intake traits, indicating that, as observational literature suggests, alcohol consumption is a significant risk factor for OPC. However, given a distinct lack of sensitivity analyses for these phenotypes, the robustness of the MR analyses is unclear and definitive causal inference cannot be made as a result. The top 5 phenotype-OPC associations were: alcohol consumed, former alcohol drinker, alcohol drinker status: never, red wine intake and alcohol. These phenotypes were all proxied by the same single SNP, rs1229984, evidenced by the exact same p-value for each of their associations. This SNP encodes a form of the alcohol dehydrogenase, *ADH1B*, gene that significantly reduces the clearance rate of alcohol from the liver [293], providing biological plausibility for the discovery of rs1229984 from multiple alcohol-related UK Biobank GWAS. However, we can only state that this locus displays a profound effect on OPC incidence; we cannot assert true causality of any of these phenotypes without more well-defined IVs to proxy them, and without additional sensitivity analyses. Amongst phenotypes possessing observational evidence of association with OPC from MELODI, average weekly beer plus cider intake showed a markedly elevated OR for risk of OPC (OR: 21.5; 95% CI: 2.87 to 161.8). This phenotype was proxied by 20 SNPs, thus sensitivity analyses be conducted alongside the IVW MR regression. Despite very large confidence intervals from the MR Egger analysis, overlap between IVW, MR Egger (OR: 3.05

$x10^4$; 95% CI: 89.1 to 1.05 $x10^7$) and MR-RAPS (OR: 14.4; 95% CI: 3.40 to 61.2) confidence intervals resulted in a lack of evidence for horizontal pleiotropy or weak instrument bias, respectively.

Educational attainment (Qualifications: College of University degree) was found to be one of the more robust results in this analysis. The phenotype was proxied by 244 genetic variants, showing little evidence of bias via multiple weak instruments (IVW OR: 0.72; 95% CI: 0.61 to 0.83 | MR-RAPS OR: 0.71; 95% CI: 0.61 to 0.83) and little evidence of horizontal pleiotropy (MR Egger OR: 0.92; 95% CI: 0.33 to 1.14). Therefore, there is sufficient evidence in this analysis to suggest a robust causal association between genetically-proxied educational attainment and OPC risk. One potential explanation for this finding could be that educational attainment typically describes a large amount of variance in socio-economic position (SEP); a proposed driver of OPC risk in observational epidemiological literature and compound measure of risk factor exposure for many diseases. Amongst other behaviours, elevated SEP has been shown to correspond to a decreased proportion of heavy smoking and alcohol consumption in European population [294]; both well-established as causal risk factors for OPC in observational epidemiological literature.

Findings pertaining to sexual behaviour agree with current observational literature. Increasing age at first sexual intercourse showed an inverse association with OPC. On the contrary, increasing lifetime number of sexual partners showed a positive association with OPC. In a cancer case/healthy control study, Schwartz et al. found a significant increase in OPC risk among men (OR: 3.4; 95% CI: 1.5 to 7.5) if their age at first regular intercourse (defined as ≥ 3 times per month) was <18 years old [295]. The authors also found that the risk of oral cancer in males increased if the number of opposite-sex partners was ≥15 (vs a single partner) (OR: 2.3; 95% CI: 1.1 to 5.0). Findings were adjusted for age, smoking and alcohol consumption. Additionally, D'Souza et al. found that being ≤17 years at first sexual intercourse showed a significant increase in the risk of HPV16-positive OPC (OR: 2.1; 95% CI: 1.1 to 3.6) after adjustment for gender, tobacco use, alcohol use, dentition and toothbrushing and family history of HNC [23]. This study also found that a high lifetime number of vaginal sex partners (≥26 vs 0–5) was associated with a significantly greater risk of OPC among men (OR: 3.1; 95% CI: 1.5 to 6.5). Dahlstrom et al. found that patients with OPC were significantly more likely than patients with other head and neck cancers to have had ≥10 lifetime sexual partners (OR: 39.2; 95% CI 8.2 to 187.3), with a trend of increasing risk of OPC with an increasing number of lifetime sexual partners ($P < 0.01$) after adjustment for age, sex, ethnicity, smoking and income [296]. Finally, a pooled analysis study by Heck et al. found that having 2 or ≥6 (vs 1) lifetime sexual partners significantly increased the risk of

cancer of the oropharynx (OR: 1.63; 95% CI: 1.22 to 2.18, and OR: 1.25; 95% CI: 1.01 to 1.54, respectively) [297].

The summary statistics used in this analysis were not stratified by sex. Therefore, the effect estimates seen between ovarian cancer and OPC, and vulvar lesions and OPC are potentially biased towards the null. By not being able to restrict these analyses to women, the true causal effect of these phenotypes on OPC are likely diluted by the presence of male genetic information. Observationally, in support of the positive association of ovarian cancer with OPC seen in this analysis, complications of women treated for ovarian cancer with pegylated liposomal doxorubicin (PLD) include oral squamous cell carcinoma and leucoplakia [298]. Liposomes have been shown to accumulate in skin and mucous membranes and release doxorubicin and its metabolites over time. The prolonged exposure to doxorubicin is presumed to be the cause of an increased rate of secondary oral malignancies. However, it should be noted that numbers of individuals exhibiting these symptoms are rare. Additionally, the findings from this MR-PheWAS should not be interpreted as a direct causal effect of ovarian cancer on OPC, or a direct effect of vulvar lesions on OPC. More likely, it is shared genetic architecture translating in the same direction to affect the risk between a genetic predisposition to both these phenotypes and OPC, respectively.

Mechanistic evidence for findings

Flecainide and sotalol are both drugs used to suppress abnormally high heart rates and arrhythmias. Arrhythmias are common in those with alcoholism, thus both findings may be proxying alcoholism if enough of the population from UK Biobank taking these medications were alcoholics. Cardiovascular disease is also significantly associated in those with HNC as a direct result of the patient demographics for this cancer; older individuals, smokers, sustained high alcohol consumption. The flecainide and sotalol findings may also be proxying a population of individuals more likely to exhibit risk behaviours for OPC. Sotalol has been shown in observational literature to decrease prostate cancer risk, though to the best of our knowledge, has not been examined with respect to OPC or HNC risk. No evidence in current epidemiological literature appears to investigate the effect of flecainide on OPC or (more generally) HNC.

Strengths

The association of a multitude of phenotypes with an uncommon cancer type were appraised in a hypothesis-free manner. This approach features a two-sample MR design that utilises summary-

level data; a particularly valuable method when the outcome of interest is rare (as in the case of OPC), or when the capacity to investigate phenotypes in single studies is limited. For example, given limited power and sample size due to the cost of metabolomic platforms, many metabolites would unlikely have been investigated in relation to OPC risk in observational studies. However, since genetic instruments for a multitude of metabolites have been obtained in previous studies with large sample sizes [270, 271], the two-sample MR framework allows these data to be harnessed to appraise the causal effect of the metabolome on health and disease.

Few other MR studies exist for the same OPC outcome as this analysis, and none utilise an MR-PheWAS framework. Pastorino et al. [299] examined the effect of adult height on combined HNC in a sample of ~5,000 Europeans. However, this MR analysis did not stratify their outcome by HNC subtype, thus the increase in risk observed between height and HNC (OR: 1.14; 95% CI: 0.99 to 1.32) may reflect a heterogeneity between subtypes, with one particular subtype biasing the true direction or magnitude of effect of another subtype. In this MR-PheWAS analysis, height showed an OR of height against OPC of over double that reported by Pastorino (OR: 2.33; 95% CI: 2.02 to 2.68), suggesting the presence of subtype heterogeneity in a combined HNC sample in their paper. The negation of potential HNC subtype heterogeneity in this analysis is a key strength as it reduces sources of bias. Kachuri et al. [300] also investigated HNC in an MR framework, this time against leukocyte telomere length. They found an inverse effect of telomere length on HNC (OR: 0.90; 95% CI: 0.70 to 1.05) using a novel IV for this phenotype. The present analysis could not investigate the effect of telomere length due to not possessing an IV for it, therefore could not assess the effect of telomere length on the OPC subtype rather than HNC combined.

This analysis serves to illustrate a method that may shed light on phenotype-disease associations otherwise untested in the literature, whilst also appraising the association of phenotypes purportedly associated with disease in observational epidemiological literature using an MR framework. That some of the phenotypes identified in this study have already been highlighted in the literature strengthens the validity of the MR-PheWAS approach in identifying potentially causal phenotype-disease associations. MR-PheWAS can be extended to other outcomes to help prioritise potentially modifiable phenotypes for investigation in further observational, animal, MR, and cell studies. Having prior evidence for an association between a subset of a large panel of phenotypes (as in this study) and an outcome of interest will allow for refinement of subsequent hypotheses.

As more GWAS are published, the number of hypotheses available for investigation using this method will steadily increase. Additionally, continued publication of GWAS of molecular intermediates such as those of the proteome, epigenome and transcriptome will allow for more detailed hypotheses of causal mechanisms to be generated, revealing intermediate pathways through which a phenotype could lead to disease.

Limitations

One limitation of the approach applied here is that not all possible phenotypes have genetic instruments or have not yet been curated in MR-Base (as seen for telomere length above). Therefore, some potentially associated phenotypes (e.g. HPV infection, oral sex) with OPC could not be appraised. Additionally, multiple phenotypes could only be proxied by a single SNP, negating the utilisation of sensitivity analyses and consequently restricting the evidence of causal association to suggestive at best. Conducting MR analysis with genetic IVs of ≥3 SNPs would allow a greater degree of certainty to be placed on the causality of a phenotype-OPC association, as it would increase the reliability of MR-Egger and MR-RAPS.

Due to the multiple testing burden of this analysis, there was potential for false-negative findings (i.e. true causal effects that we dismissed because they did not surpass our statistical threshold for further evaluation). To remain conservative in such a broad approach, only phenotypes that surpassed a strict FDR correction in the main analysis are presented. On the other hand, the MR approach may identify false positive findings, particularly if there is a horizontal pleiotropic effect of a genetic instrument on the outcome, which was evident for some of the phenotypes identified here.

A limitation of the approach applied here is a significant risk of false negative findings due to a lack of granularity afforded by using summary genetic data. As summary genetic data reflects the average genetic association across all individuals with OPC in the OncoArray study, this will include both those who have an HPV-driven OPC and those who do not. However, HPV-driven and non-HPV-driven OPC are recognised as distinct clinical entities and may have different factors predisposing them, respectively. Accordingly, with the outcome genetic data reflecting the average effect between the two entities, the results shown in this chapter may be biased towards the null, potentially missing important sub-type-specific (HPV-driven vs not) causal effects. It appears from the results that alcohol, smoking, body weight and sexual activity display a marked effect on OPC generally and are either not subject to, or able to overcome, any bias towards the null. However, any factors which do not have such a large effect on either OPC may be missed, despite potentially being an important factor in OPC

aetiology. This is particularly the case for HPV-driven OPC, where perhaps the only known predisposing factor for the disease is sexual history. Given that exposure to the high-risk HPV 16 and 18 sub-types is near-universal and incidence of HPV-driven OPC is currently increasing, knowledge of predisposing factors to the disease may prevent a large proportion of cases. This is notably important whilst the UK (and much of Europe) waits for the current time lag of vaccination programs to affect HPV-driven OPC incidence. An MR-PheWAS of HPV-driven OPC could be conducted similarly to how one has been conducted in this chapter – summary GWAS data would be extracted for a large number of exposures, then applied in a robust MR framework to summary genetic outcome data for HPV-stratified OPC incidence. The caveat of obtaining HPV-driven OPC summary genetic data is that a GWAS would require careful consideration of a control group; sexual history, geographic location and cultural experiences are all key factors for a cancer that is predisposed by sexual history. For younger individuals, matched controls based on e.g. university attendance and sexual history may optimal; for older individuals, matching based on long-term friendship groups (where possible) may best capture and match sexual behaviour, sex, age and "traditional" risk factor exposure (smoking, alcohol consumption and, to a degree, body weight).

4.5. Conclusion

Results from this MR-PheWAS of OPC show clear trends around phenotype groupings. Weight appears to be robustly associated with OPC risk. This finding is perhaps an artefact of not *losing* weight; weight loss appears to consistently preclude OPC in epidemiological literature and is also one of the nonspecific risk factors for the disease highlighted in Chapter 1. Alcohol consumption phenotypes show extreme effect sizes on OPC incidence, often showing ORs above 10, even with IVs possessing 20 SNPs (as is the case with red wine consumption). Whilst robust causality cannot be definitively stated with the alcohol IVs instrumented with a single SNP, rs1229984, associated with the *ALDB1* gene, provides biological plausibility for their association with the alcohol phenotype. Alcohol consumption affects OPC with a consistently large direction of effect in the results of this chapter. Sexual behaviour, determined by MELODI to be observationally associated with OPC, appeared to show corresponding directions of effect to those reported in epidemiological literature; the more sexual partners one has, the higher the risk of OPC; the older one is at first age of intercourse, the lower the risk of OPC. Unfortunately, HPV16 or HPV18 infection could not be instrumented in this analysis. Current tobacco smoking and having ever smoked were also associated with increased OPC risk, in agreement with current observational literature. For the findings here which suggest robust

causal association, investigation of biological pathways and molecular intermediates (such as DNA methylation) may provide more insight into the mechanisms through which OPC is caused by them.

CHAPTER 5. DNA METHYLATION AS A MEDIATOR OF OPC PROGNOSTIC FACTORS AND MORTALITY

5.1. Introduction

Observationally, smoking, alcohol and HPV16 infection are reported to affect both incidence and prognosis of OPC in classic epidemiological literature [19, 26, 217, 301]. Smoking and, to a lesser extent, heavy drinking at time of diagnosis are associated with increased incidence and poor prognosis of OPC [217, 302, 303]. Interestingly, HPV16 infection, while being a risk factor for OPC incidence, has been associated with improved prognosis for the disease [76, 304-306].

Given that the vast majority of OPC is thought to be associated with prolonged exposure to one or more of smoking, alcohol consumption and HPV16 infection [14], insight into prognostic pathways related to these phenotypes will be of notable value for risk stratification and intervention strategies for the disease. Methylation, due to its proximity to the genome and utility as an exposure indicator, can be used to highlight gene expression pathways between prognostic factor exposure and mortality. Additionally, by investigating the causality of methylation as an intermediate between prognostic factor exposure and mortality in those with OPC, evidence for novel, disease-specific, potentially druggable therapeutic targets between each exposure and mortality may be discovered.

Whilst methylation associated with smoking and alcohol have both been extensively investigated in epidemiological literature in healthy populations [182, 307], in the context of OPC, only methylation associated with HPV infection has been investigated [308]. Methylation associated with mortality in the context of OPC has not currently been investigated in epidemiological literature.

In this chapter, whole-blood-based genome-wide DNA methylation in HN5000 is examined in relation to the main prognostic factors reported in epidemiological literature for OPC, and in relation to overall OPC mortality. DNA methylation patterns associated with the prognostic factors of smoking (ever vs never), alcohol (units per week) and HPV16 infection (proxied by HPV16 E6 seropositivity) are identified through EWAS and DMR analysis, in addition to those associated with ~3-year mortality (dead vs alive at a median 3.1 years [range: 2.75 to 4.9 years; inter-quartile range: 1.1 years] post-follow-up from participant recruitment to study). Shared DNA methylation patterns between prognostic factors and mortality are then investigated. Finally, causal analyses are conducted within a novel MR framework to establish whether these overlapping methylation patterns between prognostic factors and mortality serve as causal intermediates.

5.2. Methods

5.2.1. Study population

Participants for this analysis were selected from a wider pool of individuals (post-exclusion) in HN5000 [209, 309] (N:5392), based on an ICD-10 coding (pathological where available, clinical if otherwise) of oropharynx (CO1, CO5, CO9, C10.0-2, C10.3, C10.8 and C10.9; N:1909/5392), availability of OncoChip genotype data [149] (N:1034/1909), baseline questionnaire and data capture information, and the availability of Illumina MethylationEPIC BeadChip data from blood taken at baseline, prior to treatment; N:448/1034. The HN5000 cohort structure and enrolment is described in detail in Chapter 2 section 1.5.

Phenotype definitions

Smoking, alcohol consumption and HPV16 infection

Detailed information on tobacco and alcohol history were obtained at baseline via a self-administered questionnaire. Participants were asked about their current smoking and drinking status and their use of tobacco and alcohol products prior to receiving their HNC diagnosis. Among smokers, information on age at smoking initiation and number of years of smoking was obtained. The questionnaire differentiated between use of cigarettes, hand-rolled cigarettes, cigars and smokeless tobacco, whereby a cigar was considered equivalent to four cigarettes. From this information, participants were dichotomised into ever and never smokers. Ever smokers were defined as those who smoked at the equivalent of at least 1 tobacco product a day per year, or ≥100 cigarettes in their lifetime. Never smokers were those who reported not smoking in any of the questions answered. Respondents were also asked to report their average weekly alcohol consumption of a range of beverage types (wine, spirits, and beer/larger/cider) before they were diagnosed with cancer. From these measures, an average intake of alcohol consumption in units per week was derived; weekly units were estimated from volumes of each beverage type reported by respondent and divided by 7.

HPV serologic testing (HPV16 E6, E7, E1, E2, E4, and L1) was conducted at the German Cancer Research Center (DKFZ, Heidelberg, Germany) using a glutathione S-transferase multiplex [310]. Median fluorescence intensity (MFI) values were dichotomized to indicate HPV16 E6 seropositivity using a cut-off of ≥ 1000 MFI [227]. E6 seropositivity is known to be a marker of current HPV16 infection and has a high sensitivity and specificity for HPV16-driven oropharyngeal cancer [311].

Mortality

Regular updates were received from the NHS Central Register (NHSCR) and the NHS Information Centre (NHSIC) notifying on subsequent cancer registrations and mortality among cohort members in HN5000. Recruitment for the study finished in December 2014 and follow-up information on mortality status was obtained on 30th September 2017, resulting in at least 2.75 years of follow-up for all participants (median: 3.1 years; range: 2.75 to 4.9 years: inter-quartile range: 1.1 years).

Other variables

Surrogate variables were generated from the HN5000 DNA methylation data using the *SVA* R package [312] and attempt to generate variables to include in a model which explain the largest amount of variance, similar to the PCA analyses described in Chapter 2.2.2 "Confounding and GWAS precision".

DNA methylation data

Illumina MethylationEPIC BeadChip-derived whole-blood DNA methylation data for 448 individuals from HN5000 was available for this analysis as part of the work conducted in the Integrative Cancer Epidemiology Programme at the University of Bristol. Details of HN5000 methylation data extraction and quality control can be found in Chapter 2.3.1.

Methylation quantitative trait loci

Methylation quantitative trait loci were generated using MethylationEPIC BeadChip data from a quality-controlled subset of individuals (N: 5101) from the Generation Scotland: Scottish Family Health Study [313, 314].

Mortality GWAS data

Prior to MR analysis of methylation against mortality being conducted (see section 1.2.2 "Statistical analyses" below), instrumental variables (IVs) were derived, proxying CpG sites jointly identified with prognostic factors and mortality, consisting of mQTLs. Methylation quantitative trait loci used to instrument intermediate CpG sites were identified in the HN5000 OncoChip data. Analysis of the association of these sites with mortality was conducted using the SurvivalGWAS_SV program in a Linux shell script to run Cox proportional-hazards survival analyses with an additive dosage model for each mQTL SNP [315]. Death from any cause was used as the failure variable and time to death (or

censoring) in days as the time variable. Age at cancer diagnosis and sex were used as covariables in the model. For each SNP the log-hazard ratio (and standard error) per minor allele was reported.

5.2.2. Statistical analyses

Single-site epigenome-wide association analyses

Epigenome wide association study (EWAS) analysis was conducted to identify associations between DNA methylation and 1) alcohol consumption 2) smoking status and 3) HPV16E6 seropositivity. EWAS were conducted in *meffil* [210] using R (version 3.4.1), using a linear regression model of DNA methylation regressed on the prognostic factors, adjusting for age, sex, surrogate variables obtained by SVA and the other prognostic factors (e.g. for alcohol intake, adjusting for smoking and HPV16E6 seropositivity; for smoking, alcohol intake and HPV16 E6 seropositivity).

Of the 443 individuals who passed QC (see Chapter 2.3.1 – Head and Neck 5000: DNA Methylation), the number of individuals with complete phenotype data for alcohol intake, smoking status and HPV16E6 seropositivity with which to conduct an EWAS was 408 as of the 2018, version 2.3 release of HN5000 data. All samples possessed information on mortality status.

EWASs for mortality from recruitment (last participant recruited December 2014) – September 2017 (or time of censoring; whichever occurred first) was conducted using Cox proportional-hazards models using code adapted from the meffil R package. Two models were assessed: Model 1, adjusting for age, sex and surrogate variables obtained by SVA, and Model 2, adjusting for age, sex, surrogate variables obtained by SVA, HPV16 E6 seropositivity, smoking status and alcohol intake. Model 1 was run to assess overlap with prognostic factors by not adjusting for them; Model 2, by adjusting for prognostic factors, would provide mortality-specific hits independent of them. Death from any cause was used as the failure variable and time to death (or censoring) in days as the time variable.

Due to the large number of tests conducted in the EWAS, a Bonferroni correction was employed to derive a conservative p-value threshold of 5.7×10^{-8} (0.05/862491 independent tests), determining those sites showing strong evidence of association with each prognostic factor of interest or mortality, respectively. The alpha value calculated for the Illumina 450K array was also employed as a p-value threshold of 2.4×10^{-7} for suggestive evidence of association [316].

Differentially-methylated region analyses

Following each EWAS, DMR analysis was conducted, using the *dmrff* R package [186]. This analysis identified regions (> 1 CpG site per region) enriched for low P-values (P<0.05), corrected for dependencies between other CpG sites in the DMR and adjusted for multiple testing.

Generation Scotland methylation quantitative trait loci

DNA methylation can be influenced by genetic sequence variations, such that individual genotypes at a given locus may result in different patterns of DNA methylation due to allele-specific methylation [198, 317, 318]. Such sites, called methylation quantitative trait loci (mQTLs), can influence the methylation pattern across an extended genomic region [317], and can be used as a proxy for methylation levels in a Mendelian randomization (MR) framework [126].

To generate mQTLs, methylation data from a quality-controlled subset of individuals (N: 5101) from the Generation Scotland: Scottish Family Health Study who had undergone EPIC array DNA methylation profiling, described previously [314], were used. Following measurement of DNA methylation, normalization was performed using the R package *minfi* [213], producing M-values [319] for downstream analysis. Briefly, linear mixed modelling was used to remove potential effects from technical factors, adjusting for both fixed and random effects. Fixed effects included: the top 50 principal components of control probe intensities (explaining 99% of variation in control probe intensities) [320], clinic centre for blood draw appointment, processing batch, year of clinic visit, and sentrix position (position of the sample on EPIC array slide). Random effects included: blood draw appointment date and sentrix ID (EPIC array slide). The model converged successfully for 712,595 sites. Outliers from this normalisation with residualized-M-values more than five interquartile ranges from the nearest quartile were removed [321].

A GKFSC model [322, 323] was then fitted to derive mQTLs from the normalised data, including 5 matrices as random effects, and other covariates as fixed-effects. The matrices were: G (a genomic relationship matrix), K (a kinship relationship matrix) [324, 325], F (an environmental matrix representing nuclear-family-member relationships), S (an environmental matrix representing full-sibling relationships) and C (an environmental matrix representing couple relationships) [322, 323]. Covariates (as fixed effects) included: age, age^2, gender, estimated cell counts, season of clinic visit, appointment time of the day and appointment day of the week. The model successfully converged in 638,737 CpG sites.

Generation of instrumental variables for DMRs

Prior to MR analysis being conducted (see "Mendelian randomization" below), instrumental variables (IVs) were generated, proxying CpG sites identified concurrently in analyses of each prognostic factor and mortality. Where possible, DMRs ($P < 0.05$) were identified for each prognostic factor, and DMRs located in the survival analysis which spanned the same regions (Model 1 – unadjusted for prognostic factors; $P < 0.05$). CpG sites present in both DMRs were retained.

Next, using the summary genetic data for mQTLs from Generation Scotland, all mQTLs proxying any CpG site per DMR grouping (MAF >0.05; $P < 5x10^{-8}$) were extracted. From this list, instruments were generated by LD pruning iteratively; first taking all mQTLs associated with the sentinel CpG (defined as the CpG in each DMR with the lowest p-value) and LD pruning with an r2 of 0.01. Then, the second most-associated CpG in the DMR was identified; all mQTLs associated with this CpG which were not associated with the previous, more-associated CpG were extracted. The remaining mQTLs were clumped and combined with mQTLs proxying the sentinel CpG. This process was repeated for each CpG within a DMR. Clumping and mQTL extraction were conducted using R 3.4.1, with the TwoSampleMR R package [269].

In order to account for mQTL proxies influencing methylation at multiple CpG sites, a meta-analysis of mQTL-CpG effects was conducted. Per DMR, the *metafor* R package [326] was used to meta-analyse each mQTL effect (beta) on methylation levels at each CpG, using a restricted maximum likelihood (REML) model, adjusting for pairwise correlation between the CpG sites proxied by each instrument. From this, the mQTL effect on average methylation levels across the DMR could be obtained.

Mendelian randomization

Following identification of shared methylation patterns between prognostic factors and OPC survival, two-sample MR was utilised in an attempt to ascertain whether methylation was a true causal intermediate, or rather concurrently associated with a prognostic factor and survival. In the first sample, mQTL-DMR effect estimates (βGP) from Generation Scotland were used; in the second sample, mQTL-survival estimates (βGD) from HN5000. For each mQTL, the log HR per unit (β) increase in DNA methylation was calculated by the formula βGD/βGP (Wald ratio). Standard errors were approximated by the delta method. Where multiple mQTLs were available for one DMR, these were combined in a fixed effects meta-analysis after weighting each ratio estimate by the inverse variance of their associations with the outcome (IVW approach). In order to account for correlation between

mQTLs, genetic correlation was adjusted for. An online platform, LDMatrix [327], was used to generate a genetic correlation matrix (using the 1000 Genomes reference standard) of mQTLs, which was included as a covariate in each MR regression analysis [328]. Wald ratios were calculated for CpGs proxied by a single mQTL and IVW MR estimates were calculated when multiple mQTLs were available to proxy a CpG.

Sensitivity analyses

In addition to the main analysis detailed above, multivariable MR Egger [329, 330] analysis was conducted as an assessment of IV heterogeneity. Sensitivity MR analyses were conducted by calculating the log HR per unit increase in DNA methylation for the sentinel CpG within each DMR analysed. Other prognostic factors for mortality include stage and comorbidity. Mortality EWAS were conducted with these covariates included and found the effect size remained largely unaffected by the addition of stage and comorbidity. Therefore, the EWAS of mortality was conducted without stage and comorbidity as covariates. Finally, where possible, multivariable MR Egger analysis was conducted on a subset of independent SNPs for each DMR as a sensitivity analysis for using multivariable MR Egger with correlated SNPs in the main analysis.

5.3. Results

5.3.1. Sample characteristics

Baseline characteristics of samples with epigenetic data, compared to all HNC and OPC samples in HN5000 are shown in **Table 1**. Notably, the proportion of those with OPC under the age of 60 is higher than those with other sub-types of HNC and the degree to which those with OPC differ to other HNC sub-types with respect to HPV16 E6 seropositivity (an established biomarker of current HPV16 infection) is substantial. **Table 5.1** shows that the demographics of those who were selected to have their methylation patterns typed were sufficiently representative of others with the OPC sub-type in HN5000 with respect to exposure to prognostic factors, albeit not necessarily representative of HNC as a combined entity.

Table 5.1 - Comparison of patient demographics in OPC samples selected for methylation data extraction, all samples in HN5000 identified as OPC, and all samples in HN5000

Variable	OPC in HN5000 with methylation data and complete phenotype data (N=408)	OPC in HN5000 (N=1,909)	All HN5000 (all sub-types) (N=5,392)
ICD group (% oropharynx)	100	100	35.4
Sex (% female)	27.0	21.9	27.2
Age (% <60)	58.4	52.4	42.7
Smoking (% never smoked)	27.1	28.0	24.6
Alcohol (% non-drinker)	25.9	26.6	28.4
HPV16 E6 (% negative)	33.3	32.3	72.0
Survival (% died, prior to 30/09/2017)	26.2	24.2	28.0

5.3.2. Epigenome-wide association analyses

Epigenome-wide association study of smoking

The single-site EWAS of ever vs never smokers revealed 52 CpG site associations annotated to 27 unique gene regions, or loci ($P<5.7x10^{-8}$, Bonferroni-adjusted $p < 0.05$ for 862,491 tests) (**Figure 5.1**). CpG site cg05575921, which annotates to the *AHRR* gene region, was most strongly associated ($P< 1.48x10^{-40}$) and also showed the largest effect size of -29.5% difference in methylation between ever and never smokers. Forty-nine of the associated CpG sites had lower DNA methylation in ever smokers, with a mean difference in methylation of -8.3% (SD: 5.1%, range: -29.5% to -2.2%). The three remaining CpG sites had higher methylation in smokers, with a mean difference of 7.7% (SD: 4.2%, range: 4.7% to 12.6%). **Table 5.2** provides the complete list of all CpGs that were differentially methylated below a multiple testing threshold of $P:2.4x10^{-7}$ (the alpha for the Illumina 450K BeadChip, a predecessor of the EPIC array, common in epidemiological literature, which can assay >450,000 CpG sites compared to >850,000 on the EPIC array). Of the results presented in this table, 37.5% (24/64 CpGs) were CpG sites present on the EPIC array but not its 450K predecessor.

Figure 5.1 - Manhattan plot of EWAS results from a comparison of ever vs. never smoking, showing CpG sites within DMRs in red. Each dot represents a single CpG site, plotting –log10(p) (y-axis) against the genomic position of the CpG site (x-axis). The horizontal red line is at P<5.7x10^{-8} and represents the value below which methylation was deemed to be significantly associated with smoking.

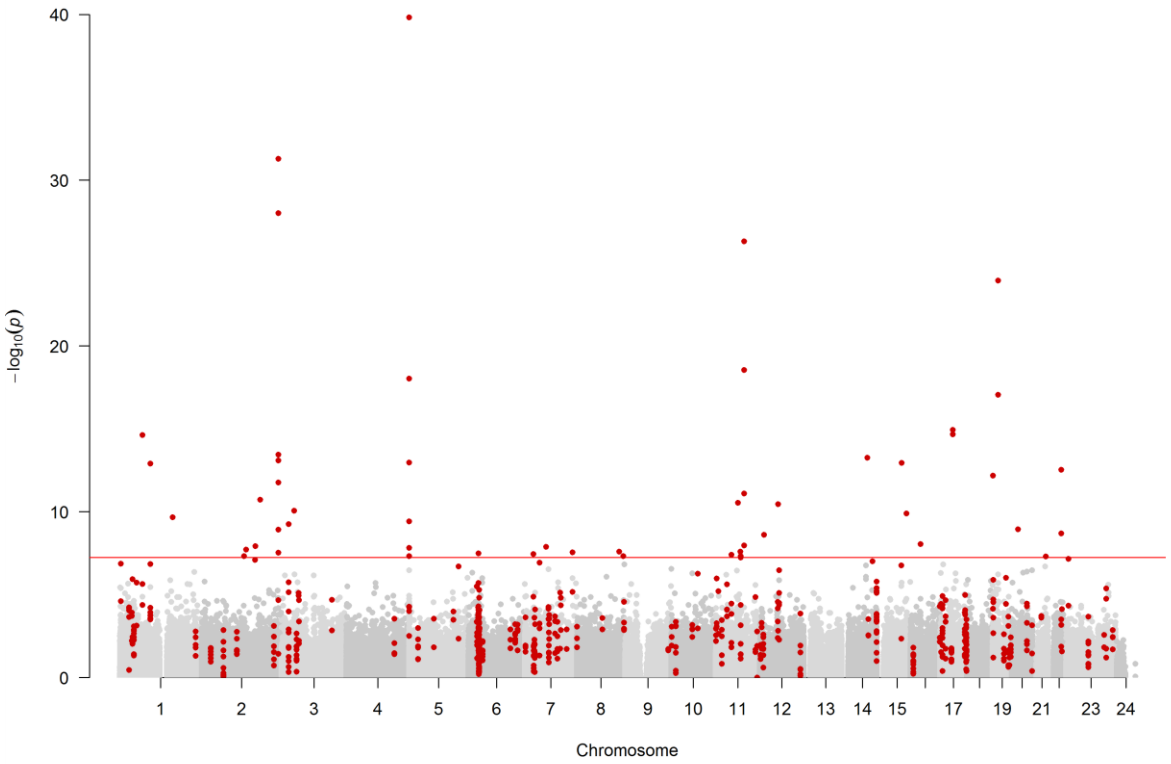


Table 5.2 - Genome-wide differentially-methylated CpG sites associated with smoking status below a multiple testing threshold of P < 5.8e-08. Results are adjusted for age, sex, surrogate variables obtained by SVA, alcohol consumption and HPV16E6 seropositivity

CpG	P-value	Beta	Chromosome	Position	Gene annotation
cg05575921	1.48E-40	-0.295	chr5	373378	*AHRR*
cg21566642	4.94E-32	-0.170	chr2	233284661	-
cg01940273	9.48E-29	-0.123	chr2	233284934	-
cg14391737	4.87E-27	-0.152	chr11	86513429	*PRSS23*
cg03636183	1.09E-24	-0.132	chr19	17000585	*F2RL3*
cg23771366	2.82E-19	-0.071	chr11	86510998	*PRSS23*
cg26703534	8.96E-19	-0.072	chr5	377358	*AHRR*
cg21911711	8.47E-18	-0.101	chr19	16998668	*F2RL3*
cg17739917	1.13E-15	-0.116	chr17	38477572	*RARA*
cg19572487	2.09E-15	-0.085	chr17	38476024	*RARA*
cg25189904	2.30E-15	-0.136	chr1	68299493	*GNG12*

CpG	P-value	Beta	Chromosome	Position	Gene annotation
cg16841366	3.48E-14	-0.104	chr2	233286192	-
cg25001882	5.40E-14	-0.045	chr14	78619077	-
cg03329539	8.04E-14	-0.084	chr2	233283329	-
cg23576855	1.04E-13	-0.213	chr5	373299	*AHRR*
cg18110140	1.09E-13	-0.103	chr15	75350380	-
cg09935388	1.19E-13	-0.185	chr1	92947588	*GFI1*
cg05086879	2.89E-13	-0.089	chr22	39861490	*MGAT3*
cg15187398	6.57E-13	-0.070	chr19	2093896	*MOBKL2A*
cg22812571	1.68E-12	-0.093	chr2	233286229	-
cg11660018	7.90E-12	-0.058	chr11	86510915	*PRSS23*
cg26271591	1.80E-11	-0.070	chr2	178125956	*NFE2L2*
cg21611682	2.80E-11	-0.062	chr11	68138269	*LRP5*
cg02583484	3.45E-11	-0.042	chr12	54677008	*HNRNPA1*
cg09945032	8.34E-11	-0.042	chr3	38871019	-
cg23161492	1.24E-10	-0.058	chr15	90357202	*ANPEP*
cg00045592	2.10E-10	-0.084	chr1	160714299	*SLAMF7*
cg21161138	3.76E-10	-0.098	chr5	399360	*AHRR*
cg04414766	5.53E-10	0.126	chr3	22412963	-
cg06421013	1.13E-09	-0.079	chr20	19194143	*SLC24A3*
cg06644428	1.18E-09	-0.073	chr2	233284112	-
cg09338374	1.96E-09	0.059	chr22	39888390	-
cg07986378	2.44E-09	-0.064	chr12	11898284	*ETV6*
cg07069636	8.63E-09	-0.043	chr16	30671749	-
cg00475490	1.04E-08	-0.030	chr11	86517110	*PRSS23*
cg19965693	1.15E-08	-0.064	chr2	163175743	*IFIH1*
cg10691866	1.31E-08	-0.075	chr7	65817282	*TPST1*
cg17287155	1.51E-08	-0.034	chr5	393347	*AHRR*
cg11866539	1.83E-08	0.047	chr2	135033075	*MGAT5*
cg25305703	2.46E-08	-0.069	chr8	128378218	-
cg01901332	2.47E-08	-0.065	chr11	75031054	*ARRB1*
cg25949550	2.79E-08	-0.022	chr7	145814306	*CNTNAP2*

CpG	P-value	Beta	Chromosome	Position	Gene annotation
cg12956751	2.93E-08	-0.037	chr2	233246922	*ALPP*
cg15342087	3.11E-08	-0.032	chr6	30720209	-
cg07741821	3.46E-08	-0.060	chr7	26577897	*KIAA0087*
cg23337648	3.82E-08	-0.040	chr11	47546192	*CELF1*
cg12075928	4.56E-08	-0.060	chr8	141801307	*PTK2*
cg10012530	4.57E-08	-0.036	chr2	129073706	*HS6ST1*
cg13633560	4.68E-08	-0.060	chr11	76380921	*LRRC32*
cg05934812	4.68E-08	-0.056	chr5	334322	*AHRR*
cg23110422	4.83E-08	-0.044	chr21	40182073	*ETS2*
cg10788371	5.54E-08	-0.046	chr11	76381040	*LRRC32*
cg25558667	6.93E-08	-0.154	chrX	10075110	*WWC3*
cg07995927	7.85E-08	-0.063	chr2	161995135	*TANK*
cg05284742	9.43E-08	-0.045	chr14	93552128	*ITPK1*
cg12803068	1.14E-07	0.124	chr7	45002919	*MYO1G*
cg01431482	1.29E-07	-0.056	chr1	2989085	*PRDM16*
cg12876356	1.41E-07	-0.154	chr1	92946825	*GFI1*
cg05460226	1.44E-07	-0.093	chr17	8804279	*PIK3R5*
cg26361535	1.45E-07	-0.064	chr8	144576604	*ZC3H3*
cg25845814	1.71E-07	-0.040	chr14	74224613	*MIR4505*
cg00310412	1.72E-07	-0.037	chr15	74724918	*SEMA7A*
cg14580211	1.92E-07	-0.073	chr5	150161299	*C5orf62*
cg12919873	1.99E-07	-0.088	chr21	38929815	-

In the differentially methylated region (DMR) analysis of ever vs never smoking, 166 unique DMRs containing 617 measured CpGs and mapping to 156 gene regions were identified (**Figure 5.1**). The DMR with the strongest association contained 3 measured CpGs (cg21566642, cg01072057 and cg13903162) and was located at Chr2:233284661-233285290, an intergenic CpG island on 2q37.1 (P:1.13 x10^{-46}).

Epigenome-wide association study of alcohol

The EWAS of alcohol consumption revealed 3 CpG site associations annotated to 3 unique genes ($P<5.7x10^{-8}$) **(Figure 5.2**). The association with the smallest p-value was cg06690548 ($P:8.3x10^{-16}$), annotating to the SLC7A11 gene region. This CpG site also showed the largest effect size of -0.10% difference in methylation per unit increase in alcohol. All results below the 450K array multiple testing threshold of $2.4x10^{-7}$ are shown in **Table 5.3.** Of the results presented in this table, 40% of the CpGs (2/5 CpGs) were present on the EPIC array but not it's 450K predecessor.

Figure 5.2 - Manhattan plot of EWAS of alcohol consumption, showing CpG sites within DMRs in red. Each dot represents the EWAS result for a single CpG site, plotting –log10(p) (y-axis) against the genomic position of the CpG site (x-axis). The horizontal red line is at $P<5.7x10^{-8}$ and represents the value below which CpG sites were considered to have good evidence of association with alcohol consumption.

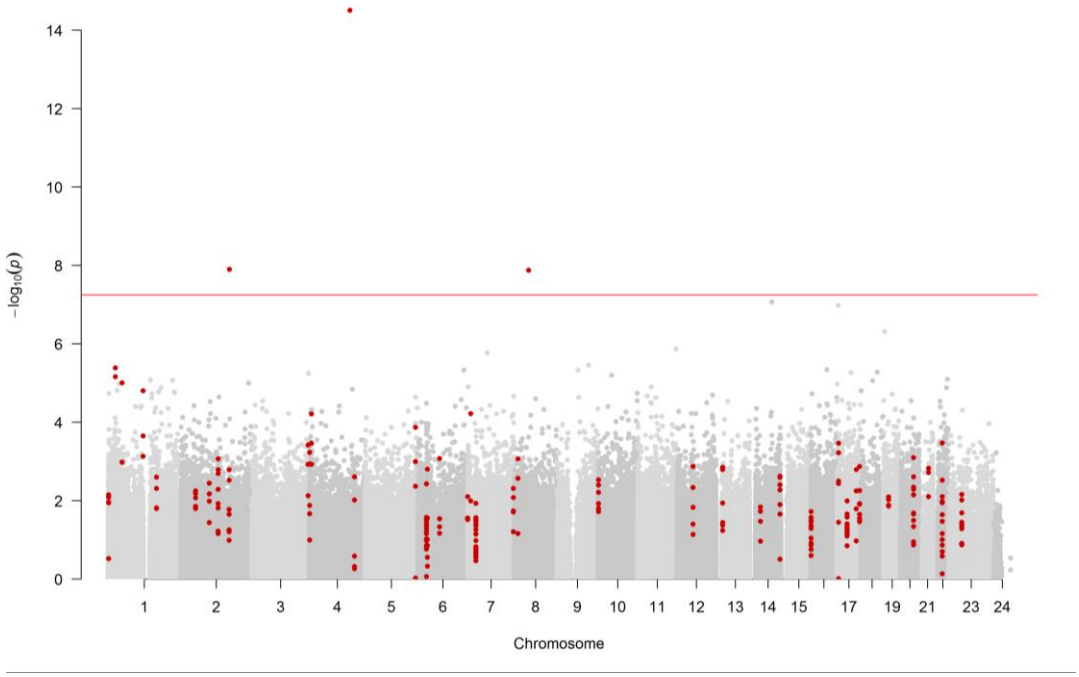


Table 5.3 - Genome-wide differentially-methylated CpG sites associated with alcohol consumption below a multiple testing threshold of P < 5.8e-08. Results are adjusted for age, sex, surrogate variables obtained by SVA, smoking status and HPV16E6 seropositivity

CpG	P-value	Beta	Chromosome	Position	Gene annotation
cg06690548	3.14E-15	-0.0010	chr4	139162808	*SLC7A11*
cg12397071	1.25E-08	0.0004	chr2	166983534	*SCN1A*
cg03137071	1.32E-08	-0.0005	chr8	49496369	-
cg06088069	8.61E-08	-0.0002	chr14	75895604	*JDP2*
cg20871826	1.04E-07	0.0002	chr17	3790402	*CAMKK1*

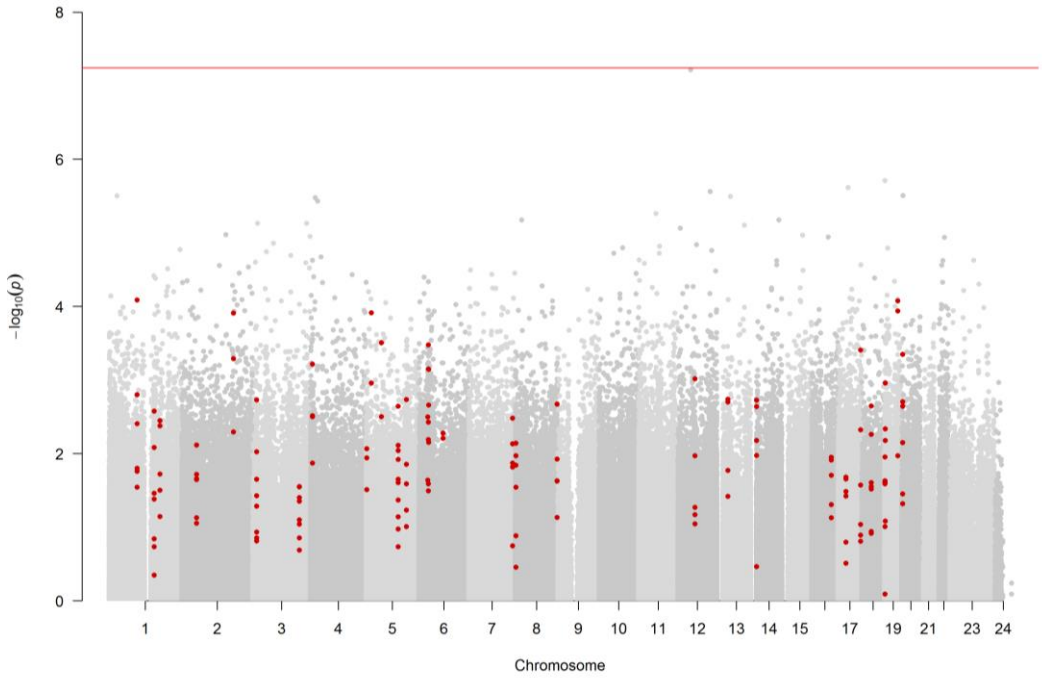
In the DMR analysis of alcohol consumption, 40 unique DMRs containing 238 measured CpGs and mapping to 34 gene regions were identified (**Figure 5.2**). The DMR with the smallest P value was a region containing 2 CpGs (cg06690548 and cg13903162) found at Chr4:139162808-139163020 (P:1.45 x10-^{10}), annotating to the *SLC7A11* gene region.

Epigenome-wide association study of HPV

In the EWAS analysis of HPV16 E6 seropositivity, no CpGs passed the multiple testing p-value threshold (P<5.7x10^{-8}) (**Figure 5.3**). At a suggestive threshold of 2.4x10^{-7}, only 1 CpG site (cg26738437; P:1.3x10^{-7}) was found, annotating to the CCL16 gene. This probe is not found on the 450K array. Methylation at this site was on average 2.3% lower in HPV16 E6 seropositive participants than controls.

In the DMR analysis of HPV16 E6 seropositivity, 31 unique DMRs pertaining to 158 CpGs and annotating to 38 gene regions were identified (**Figure 5.3**). The most associated DMR was a region of 13 CpGs found at Chr5:110062343-110062838 (P:4.10x10-^{6}), annotating to the TMEM232 gene region.

Figure 5.3 - Manhattan plot of EWAS of HPV16E6 seropositivity, showing CpG sites within DMRs in red. Each dot represents the EWAS result for a single CpG site, plotting –log10(p) (y-axis) against the genomic position of the CpG site (x-axis). The horizontal red line is at P<5.7x10^{-8} and represents the value below which CpG sites were considered to have good evidence of association with HPV16 E6 seropositivity.



Epigenome-wide association study of OPC survival

Model 1

In the single-site analysis of survival (adjusting for age, sex and surrogate variables obtained by SVA [312]), three CpGs mapping to three unique loci showed association with survival below a multiple testing p-value threshold ($P<5.7e^{-8}$) (**Figure 5.4**). One CpG site showed lower methylation in those who died vs were alive during follow-up. This site was also the most strongly associated with survival in the survival analysis, annotating to PAQR3 and showed the largest effect size among top hits (cg25864218; β [difference in methylation between those that were dead vs alive before 30th September 2017]: -2.54%; P: $1.04e^{-9}$). Two sites showed higher methylation in those who died vs were alive during follow-up in the analysis, annotating to DNAH11 (cg07377396; β: 0.49%; P: $3.39e^{-8}$) and MYBPC1 (cg12151015; β: 0.11%; P: $7.51e^{-9}$). The mean difference in methylation in these sites was 0.3% (SD: 0.27%, range: 0.11% to 0.49%). All results below a suggestive multiple testing threshold of $2.4e^{-7}$ are shown in **Table 5.4.** Of the results presented in this table, 47% (7/15) were CpG probes specific to the EPIC array.

Figure 5.4 - Manhattan plot of EWAS of survival (model 1 – not adjusted for smoking, alcohol consumption and HPV16E6 seropositivity), showing CpG sites within DMRs in red. Each dot represents the EWAS result for a single CpG site, plotting –log10(p) (y-axis) against the genomic position of the CpG site (x-axis). The horizontal red line is at P<5.7x10-8 and represents the value below which CpG sites were considered to have good evidence of association with survival.

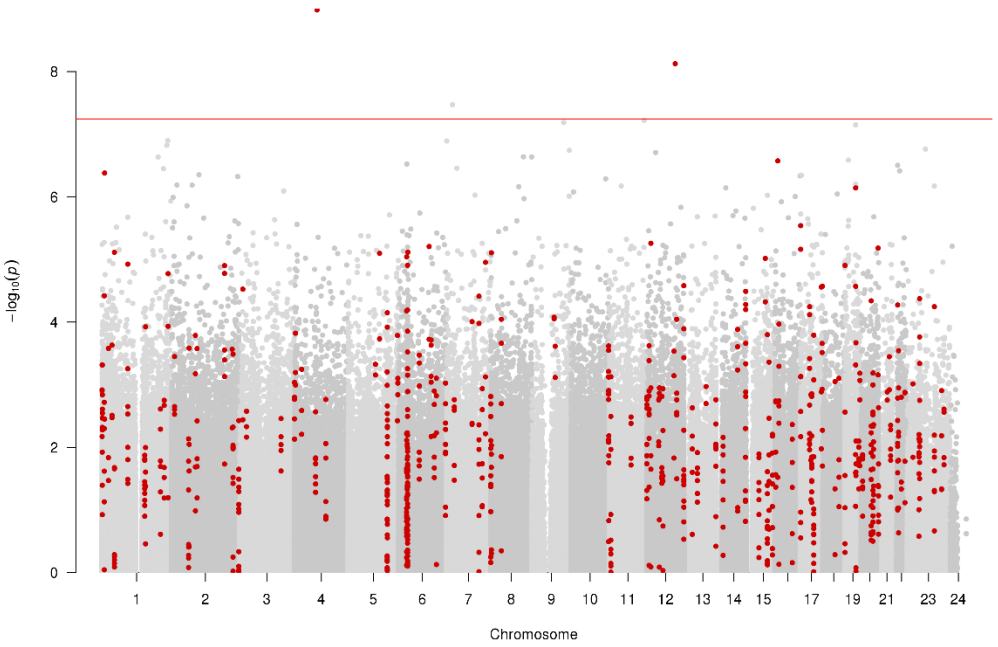


Table 5.4 - Genome-wide differentially-methylated CpG sites associated with ~3-year survival below a multiple testing threshold of P < 5.8e-08. Results are adjusted for age, sex and surrogate variables obtained by SVA

CpG	Beta	P-value	Chromosome	Position	Gene annotation
cg25864218	-2.54	1.04E-09	chr4	79860686	*PAQR3*
cg12151015	0.11	7.51E-09	chr12	102010848	*MYBPC1*
cg07377396	0.49	3.39E-08	chr7	21788428	*DNAH11*
cg15036595	-0.16	6.00E-08	chr11	125818934	-
cg03093995	1.21	6.51E-08	chr9	114393414	*DNAJC25-GNG10*
cg17679548	-0.46	7.11E-08	chr19	39353395	-
cg18236982	-0.15	1.27E-07	chr1	236065291	-
cg00338391	0.76	1.28E-07	chr7	1283981	-
cg11337053	-0.15	1.50E-07	chr1	232651485	*SIPA1L2*
cg09853393	-0.10	1.73E-07	chrX	67995409	-
cg14504586	-0.64	1.81E-07	chr9	134744872	*MED27*
cg19019403	-0.24	1.96E-07	chr12	31899879	-
cg02927174	1.01	2.30E-07	chr8	116681088	*TRPS1*
cg20046119	0.15	2.31E-07	chr8	145981207	*ZNF251*
cg13071729	0.27	2.31E-07	chr1	201617443	*NAV1*

In the DMR analysis of survival, 142 unique DMRs pertaining to 805 CpGs and annotating to 153 gene regions were identified (**Figure 5.4**). The DMR with the lowest p value was a region of 10 CpGs found at Chr17:33814297-33814897 (P:5.26e-^{21}), annotating to the CDK16 gene region.

Model 2

In the single-site analysis of survival using Model 2 (adjusting for age, sex, surrogate variables obtained by SVA, HPV16E6 seropositivity, smoking status and alcohol intake), 6 CpGs annotated to 4 unique loci showed a p-value of association below the multiple testing threshold (P<5.7e^{-8}) (**Figure 5.5**). Three of the 6 CpGs passing multiple testing correction showed lower methylation in those who died vs were alive during follow-up in the analysis, while the other 3 showed higher methylation. Of the 3 sites showing lower methylation, the mean difference in methylation between those that were dead vs alive after ~3-year follow-up was -0.07% (SD: 0.05%, range: -2.54% to -0.16%). For the 3 sites showing higher methylation, the mean difference in methylation was 0.31% (SD: 0.31%, range: 0.11% to 0.67%). The CpG with the smallest P value (cg25864218, P: 1.22x10^{-8}), annotates to the *PAQR3* gene

region. This CpG site also showed the largest effect size of -2.5% difference in methylation between those who are dead vs alive. Other CpGs passing the multiple testing correction which were annotated to genes included *MYBPC1* (cg12151015; β: 0.11%; P: 2.59e-^{8}), *GRIN2A* (cg08204867; β: -0.16%; P: 2.87e-^{8}), and *IL15* (cg26269613; β: 0.67%; P: 5.34e-^{8}). Two CpGs showed an association with survival in both models: cg12151015 (annotating to *MYBPC1*) and cg25864218 (annotating to *PAQR3*). All results below a suggestive multiple testing threshold of 2.4e^{-7} are shown in **Table 5.5.** Interestingly, of the results presented in this table, all 23 associated CpGs were present on the EPIC array but not the 450K predecessor.

Figure 5.5 - Manhattan plot of EWAS of survival (model 2 – adjusted for smoking, alcohol consumption and HPV16E6 seropositivity), showing CpG sites within DMRs in red. Each dot represents the EWAS result for a single CpG site, plotting –log10(p) (y-axis) against the genomic position of the CpG site (x-axis). The horizontal red line is at P<5.7x10-8 and represents the value below which CpG sites were considered to have good evidence of association with survival.

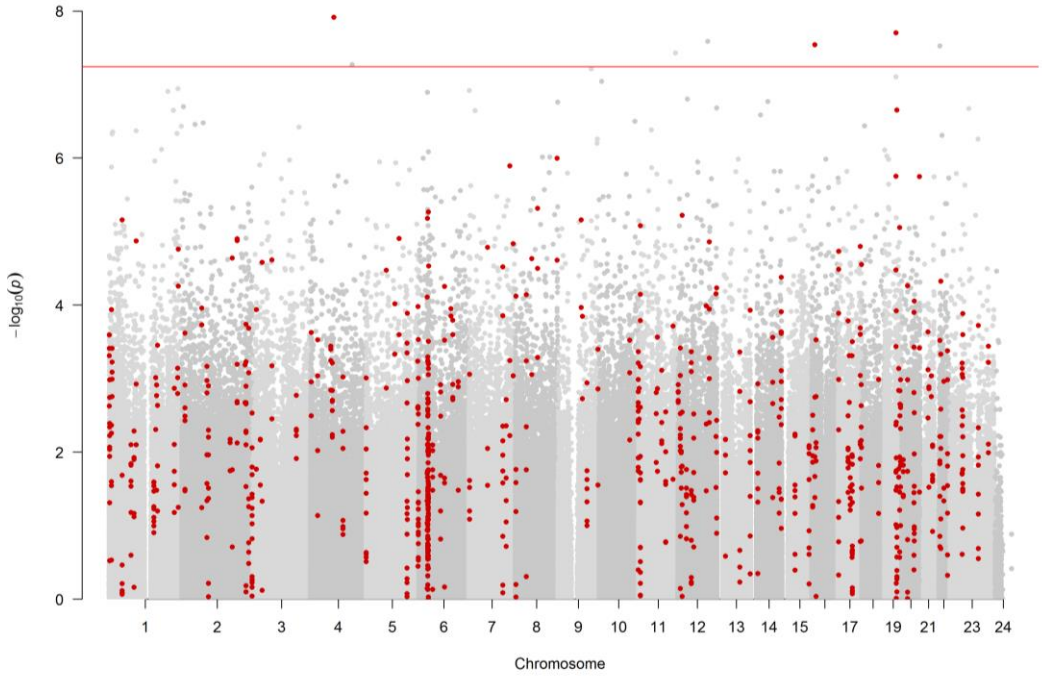


Table 5.5 - Genome-wide differentially-methylated CpG sites associated with ~3-year survival below a multiple testing threshold of P < 5.7e-08. Results are adjusted for age, sex, surrogate variables obtained by SVA, smoking status, alcohol consumption and HPV16E6 seropositivity

CpG	Beta	P-value	Chromosome	Position	Gene annotation
cg25864218	-2.45	1.22E-08	chr4	79860686	*PAQR3*
cg27551718	-0.14	1.99E-08	chr19	39572000	-
cg12151015	0.11	2.59E-08	chr12	102010848	*MYBPC1*
cg08204867	-0.16	2.87E-08	chr16	10208426	*GRIN2A*

CpG	Beta	P-value	Chromosome	Position	Gene annotation
cg18604947	0.16	2.99E-08	chr22	19609793	-
cg15036595	-0.16	3.72E-08	chr11	125818934	-
cg26269613	0.67	5.34E-08	chr4	142558451	*IL15*
cg03093995	1.19	6.13E-08	chr9	114393414	*DNAJC25*
cg17679548	-0.48	7.86E-08	chr19	39353395	-
cg04444399	-0.31	9.02E-08	chr10	8302797	-
cg18236982	-0.16	1.13E-07	chr1	236065291	-
cg00338391	0.75	1.21E-07	chr7	1283981	-
cg13071729	0.29	1.24E-07	chr1	201617443	*NAV1*
cg11588423	-0.20	1.28E-07	chr6	29578191	*GABBR1*
cg19019403	-0.24	1.57E-07	chr12	31899879	-
cg22769651	-0.49	1.70E-07	chr14	59594615	-
cg20046119	0.15	1.73E-07	chr8	145981207	*ZNF251*
cg15334209	-0.25	2.00E-07	chr2	5986281	-
cg13987792	-0.23	2.08E-07	chr12	132431451	-
cg09853393	-0.11	2.12E-07	chrX	67995409	-
cg01555270	-1.24	2.23E-07	chr19	41945599	*ATP5SL*
cg05278651	-0.11	2.24E-07	chr1	220533602	-
cg07377396	0.46	2.26E-07	chr7	21788428	*DNAH11*

In the DMR analysis of survival (model 2), 157 unique DMRs pertaining to 874 CpGs and annotating to 177 gene regions were identified (**Figure 5**). The DMR with the lowest p value was a region of 12 CpGs found at ChrX: 47077168- 47077877 (P:1.08e^{-21}), annotating to the *CDK16* gene region.

5.3.3. DMR overlap between OPC risk factors and survival

Eighteen unique CpGs overlapped between all smoking DMRs and survival DMRs (survival EWAS model 1). These CpGs belonged to 3 unique DMRs (annotated to *GFI1, SPEG* and *PPT2*); five CpGs overlapped between all alcohol DMRs and survival (EWAS Model 1) DMRs, all pertaining to a single DMR (annotated to *KHD3CL*). No CpGs overlapped at the p-value threshold for HPV DMRs and survival (EWAS model 1) DMRs.

Of the 18 CpGs which overlapped between smoking and survival, 15 possessed mQTL proxies in the Generation Scotland summary data with which to conduct MR (see Methods). Of the 5 CpGs which overlapped between alcohol and survival, 3 possessed mQTL proxies in the Generation Scotland summary data (**Table 5.6**).

Table 5.6 - Genetic instrumental variables (IVs) used in Mendelian randomization analyses to assess epigenetic mediation between prognostic factors and ~3-year survival. The final # SNPs denotes genetic IVs which both proxy a CpG and where the same position is available in the genome-wide association study of 3-year mortality

Phenotype	DMR (Gene)	CpG	mQTL	SNP in outcome GWAS?	Final # SNPs
Smoking	Chr1:92946132-92947588 (*GFI1*)	cg18146737	N/A	N/A	8
		cg18316974	rs17131598	Y	
			rs7552212	Y	
		cg09935388	N/A	N/A	
		cg04535902	rs6691038	Y	
			rs116053219	Y	
		cg12876356	rs17518433	Y	
		cg09662411	rs12065617	Y	
			rs2391140	Y	
		cg06338710	rs7549306	N	
			rs17131593	Y	
Smoking	Chr2:220325443-220326041 (*SPEG*)	cg06084174	rs6740770	Y	17
			rs991503	Y	
		cg19179241	rs4674396	Y	
			rs62191875	Y	
			rs3770234	Y	
		cg18963236	rs72965313	Y	
			rs3755059	Y	
		cg18894092	rs13386459	Y	
			rs10932806	N	
		cg14890311	rs366528	Y	
			rs113057402	Y	
			rs6717249	N	

Phenotype	DMR (Gene)	CpG	mQTL	SNP in outcome GWAS?	Final # SNPs
			rs35116888	Y	
			rs73087210	Y	
			rs116399602	Y	
		cg17850359	rs745027	Y	
			rs4674397	Y	
		cg18377670	rs907683	Y	
			rs2046615	Y	
			rs1467116	N	
Smoking	Chr6:32120895-32120907 (*PPT2*)	cg23359665	rs3130283	Y	8
			rs4713534	Y	
		cg02956248	rs6467	N	
			rs9378123	Y	
			rs111911331	Y	
		cg06108383	rs9267551	Y	
			rs7745174	Y	
		cg17113856	rs399950	Y	
			rs9270101	Y	
Alcohol	Chr6:74072255-74072376 (*KHD3CL*)	cg19146112	N/A	N/A	4
		cg27237745	rs488114	Y	
		cg26550861	N/A	N/A	
		cg13001017	rs508770	Y	
			rs538837	Y	
		cg16074228	rs564533	Y	

5.3.4. Mendelian randomization: DNA methylation - OPC survival

Tables 5.7-5.9 and **Figure 5.6** show the results MR analyses for the association of mQTL-proxied DNA methylation, at CpG sites associated with smoking and survival, with 3-year survival in HN5000. In these analyses, there appears to be some evidence for a potential causal effect of decreased DNA methylation on survival at the *SPEG* gene locus (**Table 5.7**; Chr2:22035443-22036041; HR: 1.28; 95% CI: 1.14 to 1.43). Results suggest DNA methylation may mediate part of the association

seen between smoking and decreased survival at this gene region. The *GFI1* (**Table 5.8**) and *PPT2* (**Table 5.9**) gene regions appear to show no consistent evidence of a causal effect of DNA methylation on survival. Multivariable MR Egger analysis using independent SNPs (multivariable MR Egger$_{independent}$: a sensitivity analysis for using multivariable MR Egger with correlated SNPs in the main analysis) could only be conducted at the *SPEG* locus, as other regions did not have sufficient independent SNPs as proxies. Fewer than 3 SNPs greatly reduces the accuracy of MR Egger; therefore, it was only used in analyses with 3 or more SNP proxies. Multivariable MR Egger$_{independent}$ showed a similar effect estimate to normal multivariable MR Egger at this locus, albeit with larger confidence intervals, suggesting the confidence interval for normal multivariable MR Egger is likely to be overly precise in this analysis.

Table 5.7 - Mendelian randomization (MR) analysis results, assessing epigenetic mediation between smoking status and ~3-year survival at the SPEG gene (chromosome 2:220325443-220326041). The number of SNPs per analysis are shown, in addition to the inverse- variance weighted (IVW) and multivariable MR Egger MR results. IVW and MR Egger results are adjusted for genetic correlation between mQTLs are reported as hazard ratios (HR) with 95% confidence intervals (CI). The SPEG locus was the only in our analyses to possess >2 independent SNPs and is therefore the only with multivariable MR Egger analysis conducted on this independent subset in addition to all DMR CpGs.

Region (gene)	MR Method	SNPs	HR	95% CI	P
All DMR CpGs					
Chr2:220325443-220326041 (*SPEG*)	IVW	17	1.28	1.14 to 1.43	$2.12x10^{-05}$
Chr2:220325443-220326041 (*SPEG*)	MR Egger	17	1.28	1.18 to 1.38	$4.04x10^{-10}$
Sentinel CpG only					
cg06084174 (*SPEG*)	IVW	3	1.14	0.90 to 1.45	0.29
CpGs with independent SNPs					
cg19179241 and cg14890311	MR Egger	4	1.27	0.78 to 2.08	0.34

Table 5.8 - Mendelian randomization (MR) analysis results, assessing epigenetic mediation between smoking status and ~3-year survival at the GFI1 gene (chromosome 1:92946132-92947588). The number of SNPs per analysis are shown, in addition to the inverse-variance weighted (IVW) and multivariable MR Egger MR results. IVW and MR Egger results are adjusted for genetic correlation between mQTLs are reported as hazard ratios (HR) with 95% confidence intervals (CI).

Region (gene)	MR Method	SNPs	HR	95% CI	P
All DMR CpGs					
Chr1:92946132-92947588 (*GFI1*)	IVW	8	0.74	0.60 to 0.93	$7.9x10^{-03}$
Chr1:92946132-92947588 (*GFI1*)	MR Egger	8	2.65	0.77 to 9.12	0.12
Sentinel CpG only					
cg06338710 (*GFI1*)	Wald ratio	1	0.93	0.47 to 1.85	0.84

Table 5.9 - Mendelian randomization (MR) analysis results, assessing epigenetic mediation between smoking status and ~3-year survival at the PPT2 gene (chromosome 6:32120895-32120907). The number of SNPs per analysis are shown, in addition to the inverse-variance weighted (IVW) and multivariable MR Egger MR results. IVW and MR Egger results are adjusted for genetic correlation between mQTLs are reported as hazard ratios (HR) with 95% confidence intervals (CI).

Region (gene)	MR Method	SNPs	HR	95% CI	P
All DMR CpGs					
Chr6:32120895-32120907 (*PPT2*)	IVW	8	0.82	0.52 to 1.30	0.40
Chr6:32120895-32120907 (*PPT2*)	MR Egger	8	1.68	0.27 to 10.38	0.58
Sentinel CpG only					
cg17113856 (*PPT2*)	IVW	2	0.67	0.37 to 1.22	0.19

Figure 5.6 - Forest plots showing SNP-specific and overall IV Hazard ratio estimates (95% CI) for Mendelian randomization analyses of smoking-associated methylation at 3 gene loci (GFI1, PPT2, SPEG), against 3-year survival in oropharyngeal cancer.

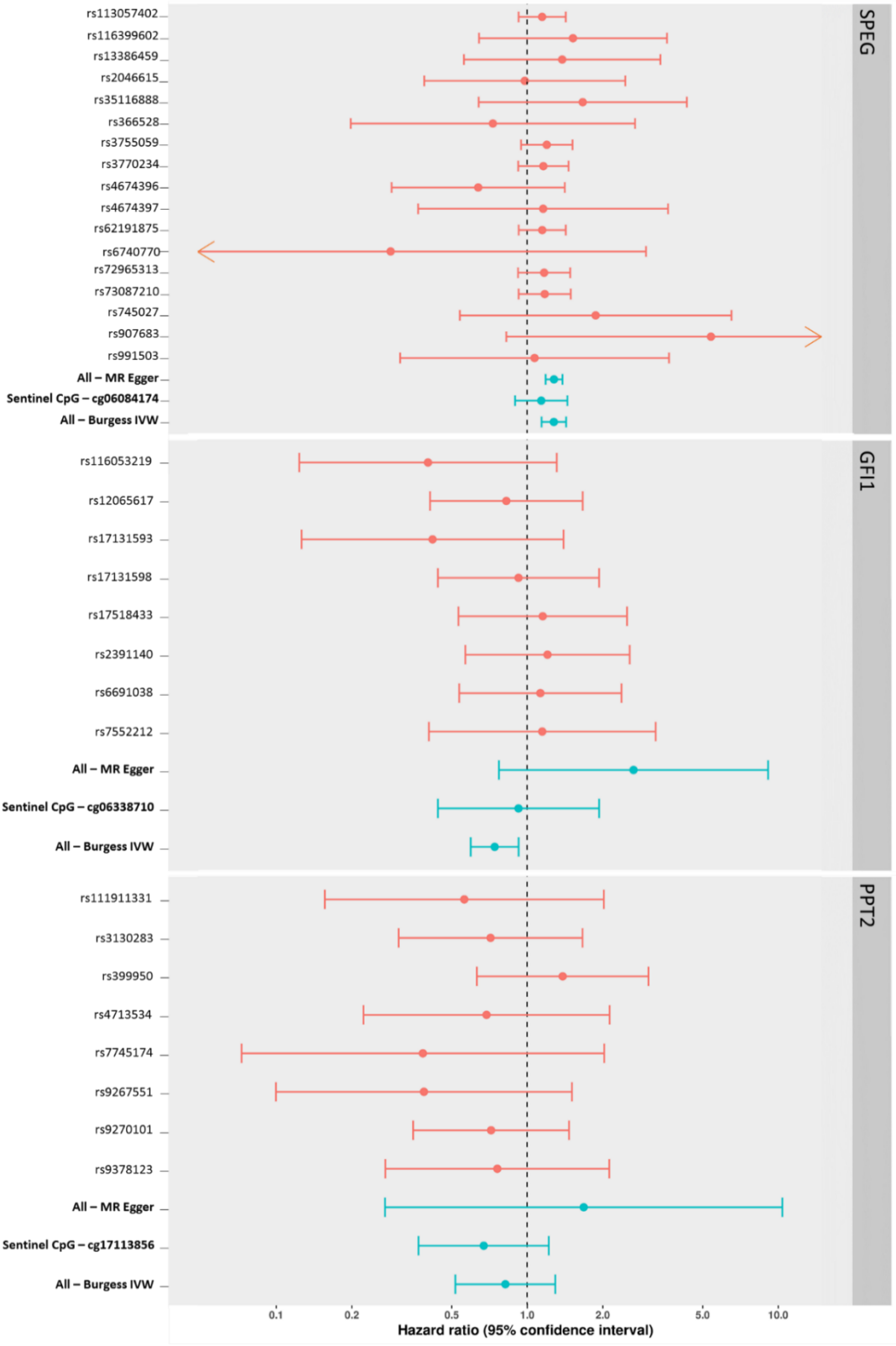
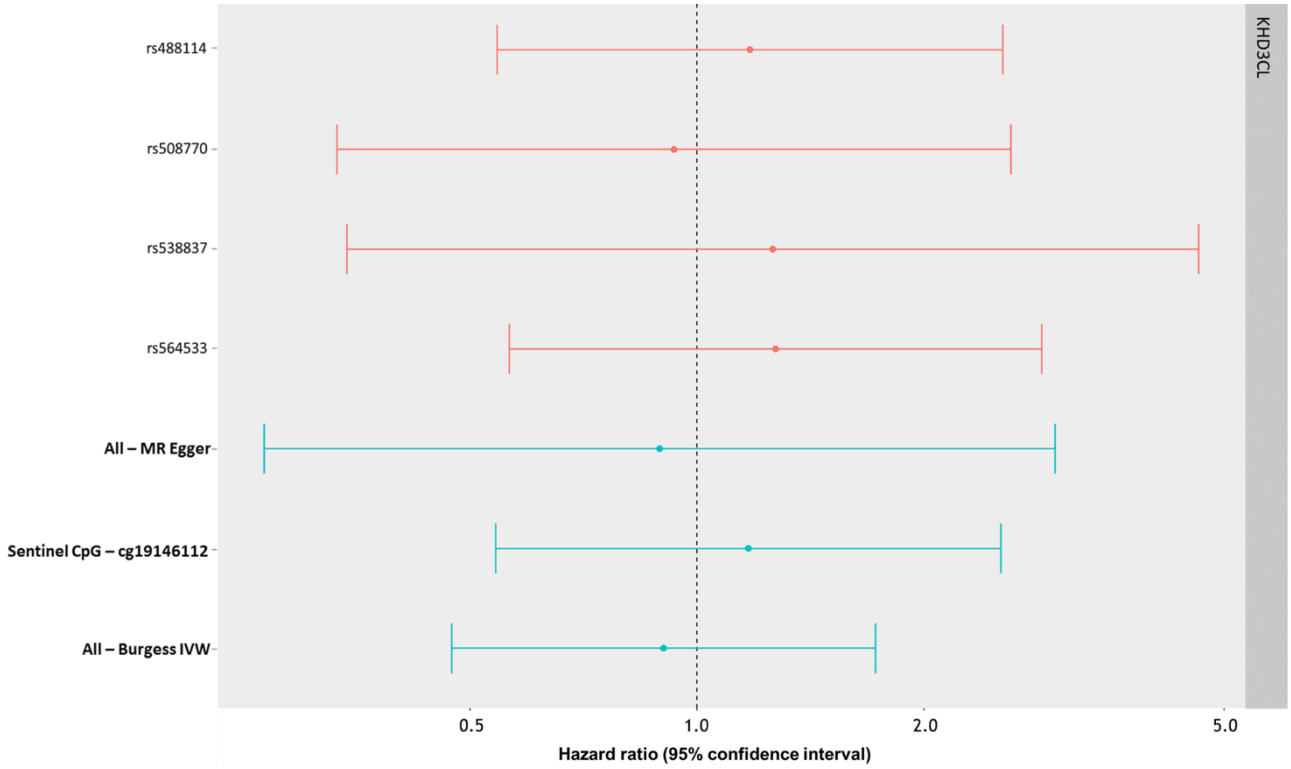


Table 5.10 and **Figure 5.7** show the results of the associations of mQTL-proxied DNA methylation, at CpG sites associated with alcohol and survival with 3-year survival in HN5000. In the analysis, there appears to be no consistent evidence for a causal effect of DNA methylation on survival at the *KHD3CL* gene locus (Chr6:74072255-74072376).

Table 5.10 - Mendelian randomization (MR) analysis results, assessing epigenetic mediation between alcohol consumption and ~3-year survival at the KHD3CL gene (chromosome 6:74072255-74072376). The number of SNPs per analysis are shown, in addition to the inverse-variance weighted (IVW) and multivariable MR Egger MR results. IVW and MR Egger results are adjusted for genetic correlation between mQTLs are reported as hazard ratios (HR) with 95% confidence intervals (CI).

Region (gene)	MR Method	SNPs	HR	95% CI	P
No clumping of final instrument, meta-analysis of mQTLs					
Chr6:74072255-74072376 (*KHD3CL*)	IVW	4	1.17	0.70 to 1.97	0.55
Chr6:74072255-74072376 (*KHD3CL*)	MR Egger	4	0.89	0.27 to 2.98	0.85
Sentinel CpG only					
cg19146112 (*KHD3CL*)	Wald ratio	1	1.17	0.54 to 2.53	0.68

Figure 5.7 - Forest plot showing the SNP-specific and overall IV Hazard ratio estimates (95% CI) for Mendelian randomization analyses of alcohol-associated methylation at the KHDC3L gene locus, against 3-year survival in oropharyngeal cancer.



5.4. Discussion

EWAS analyses were conducted, identifying CpG sites and DMRs associated with smoking and alcohol consumption, but not with HPV infection. Six CpGs were also found to be associated with survival at 3 years post-diagnosis. Twenty-three CpGs at 4 DMRs were identified in both analyses of risk factor and of survival. It is hypothesised that for these CpG sites, DNA methylation could mediate part of the association between risk factor and OPC survival. MR analysis was conducted to test this hypothesis and preliminary evidence was found to support this mediation pathway between smoking and OPC survival at the *SPEG* gene locus.

In relation to smoke exposure, the results include several previously reported loci, notably those mapping to *AHRR* and *PRSS23*. The effect size seen in the EWAS for cg05575921 (29.5%) is markedly stronger than the largest published smoking EWAS analysis; Joehanes et al [182] report 18% lower methylation in current smokers compared to those who have never smoked (P: $4.60e^{-26}$). A potential explanation for this finding could be that the analysis was conducted in a case-only setting where smoking is one of the predominant risk factors for HNC and smoking intensity is likely to be higher in HN5000 smokers compared to non-cancer smoking populations. A lookup of the top smoking CpG sites ($P < 5.7e^{-8}$) was completed using the EWAS Catalog (http://www.ewascatalog.org/) online tool to compare whether the effect sizes were consistently stronger than other published smoking EWAS findings (**Table 5.11**). Of the 52 sites below a multiple-testing correction, 20 had not been previously reported in published EWASs. The other 32 CpG sites which had previously been reported in the literature showed consistently larger effect estimates in response to smoking in the analysis compared to a weighted mean (weighted by sample size) of published EWAS beta values.

Using the same EWAS Catalog resource, an attempt was made to determine whether associations below the multiple testing threshold were stronger than previously published results for all EWAS conducted in this analysis (**Table 5.11**). All 5 associations found in the alcohol consumption analysis had not been previously reported in published EWAS of alcohol consumption; this is probably because they are not measured on the 450K array. *SLC7A11*, the gene annotated to the top CpG site associated with alcohol consumption, is essential for glutathione synthesis, a component of the KEAP1-NRF2-CUL3 axis, and strongly associated with poor prognosis in The Cancer Genome Atlas (TCGA) HNC cohort [331, 332].

Table 5.11 - Lookup of CpG sites in the MRCIEU EWAS Catalog across all EWAS analyses below a Bonferroni p-value threshold of 5.7e-08. Betas for all studies reporting beta values are calculated as a weighted mean, weighted by sample size

Phenotype	CpG name	Gene annotation	Beta	P	Novel finding?	Mean beta across published literature	# PMIDs with finding
Alcohol	cg06690548	*SLC7A11*	-0.0010	3.14E-15	Y	NA	NA
Alcohol	cg12397071	*SCN1A*	0.0004	1.25E-08	Y	NA	NA
Alcohol	cg03137071	-	-0.0005	1.32E-08	Y	NA	NA
Alcohol	cg06088069	*JDP2*	-0.0002	8.61E-08	Y	NA	NA
Alcohol	cg20871826	*CAMKK1*	0.0002	1.04E-07	Y	NA	NA
Smoking	cg05575921	*AHRR*	-0.295	1.48E-40	N	-0.191	30
Smoking	cg21566642	-	-0.170	4.94E-32	N	-0.191	19
Smoking	cg01940273	-	-0.123	9.48E-29	N	-0.107	22
Smoking	cg14391737	*PRSS23*	-0.152	4.87E-27	Y	NA	NA
Smoking	cg03636183	*F2RL3*	-0.132	1.09E-24	N	-0.132	19
Smoking	cg23771366	*PRSS23*	-0.071	2.82E-19	N	-0.055	14
Smoking	cg26703534	*AHRR*	-0.072	8.96E-19	N	-0.076	15
Smoking	cg21911711	*F2RL3*	-0.101	8.47E-18	Y	NA	NA
Smoking	cg17739917	*RARA*	-0.116	1.13E-15	Y	NA	NA
Smoking	cg19572487	*RARA*	-0.085	2.09E-15	N	-0.039	17
Smoking	cg25189904	*GNG12*	-0.136	2.30E-15	N	-0.050	21
Smoking	cg16841366	-	-0.104	3.48E-14	Y	NA	NA
Smoking	cg25001882	-	-0.045	5.40E-14	Y	NA	NA
Smoking	cg03329539	-	-0.084	8.04E-14	N	-0.044	15
Smoking	cg23576855	*AHRR*	-0.213	1.04E-13	N	-0.146	12
Smoking	cg18110140	-	-0.103	1.09E-13	Y	NA	NA
Smoking	cg09935388	*GFI1*	-0.185	1.19E-13	N	-0.076	19
Smoking	cg05086879	*MGAT3*	-0.089	2.89E-13	Y	NA	NA
Smoking	cg15187398	*MOBKL2A*	-0.070	6.57E-13	N	-0.029	11
Smoking	cg22812571	-	-0.093	1.68E-12	Y	NA	NA
Smoking	cg11660018	*PRSS23*	-0.058	7.90E-12	N	-0.038	18
Smoking	cg26271591	*NFE2L2*	-0.070	1.80E-11	N	-0.032	10
Smoking	cg21611682	*LRP5*	-0.062	2.80E-11	N	-0.026	15

Phenotype	CpG name	Gene annotation	Beta	P	Novel finding?	Mean beta across published literature	# PMIDs with finding
Smoking	cg02583484	*HNRNPA1*	-0.042	3.45E-11	N	-0.026	15
Smoking	cg09945032	-	-0.042	8.34E-11	Y	NA	NA
Smoking	cg23161492	*ANPEP*	-0.058	1.24E-10	N	-0.040	12
Smoking	cg00045592	*SLAMF7*	-0.084	2.10E-10	Y	NA	NA
Smoking	cg21161138	*AHRR*	-0.098	3.76E-10	N	-0.053	18
Smoking	cg04414766	-	0.126	5.53E-10	Y	NA	NA
Smoking	cg06421013	*SLC24A3*	-0.079	1.13E-09	Y	NA	NA
Smoking	cg06644428	-	-0.073	1.18E-09	N	-0.04	16
Smoking	cg09338374	-	0.059	1.96E-09	Y	NA	NA
Smoking	cg07986378	*ETV6*	-0.064	2.44E-09	N	-0.029	9
Smoking	cg07069636	-	-0.043	8.63E-09	N	-0.017	9
Smoking	cg00475490	*PRSS23*	-0.030	1.04E-08	Y	NA	NA
Smoking	cg19965693	*IFIH1*	-0.064	1.15E-08	Y	NA	NA
Smoking	cg10691866	*TPST1*	-0.075	1.31E-08	N	-0.031	10
Smoking	cg17287155	*AHRR*	-0.034	1.51E-08	N	-0.027	10
Smoking	cg11866539	*MGAT5*	0.047	1.83E-08	Y	NA	NA
Smoking	cg25305703	-	-0.069	2.46E-08	N	-0.035	9
Smoking	cg01901332	*ARRB1*	-0.065	2.47E-08	N	-0.030	11
Smoking	cg25949550	*CNTNAP2*	-0.022	2.79E-08	N	-0.020	17
Smoking	cg12956751	*ALPP*	-0.037	2.93E-08	Y	NA	NA
Smoking	cg15342087	-	-0.032	3.11E-08	N	-0.032	18
Smoking	cg07741821	*KIAA0087*	-0.060	3.46E-08	Y	NA	NA
Smoking	cg23337648	*CELF1*	-0.040	3.82E-08	Y	NA	NA
Smoking	cg12075928	*PTK2*	-0.060	4.56E-08	N	-0.033	13
Smoking	cg10012530	*HS6ST1*	-0.036	4.57E-08	N	-0.011	4
Smoking	cg13633560	*LRRC32*	-0.060	4.68E-08	N	-0.015	5
Smoking	cg05934812	*AHRR*	-0.056	4.68E-08	Y	NA	NA
Smoking	cg23110422	*ETS2*	-0.044	4.83E-08	N	-0.021	7
Smoking	cg10788371	*LRRC32*	-0.046	5.54E-08	N	-0.018	9

Phenotype	CpG name	Gene annotation	Beta	P	Novel finding?	Mean beta across published literature	# PMIDs with finding
Survival Model 1	cg25864218	*PAQR3*	-2.54	1.04E-09	Y	NA	NA
Survival Model 1	cg12151015	*MYBPC1*	0.11	7.51E-09	Y	NA	NA
Survival Model 1	cg07377396	*DNAH11*	0.49	3.39E-08	Y	NA	NA
Survival Model 1	cg15036595	-	-0.16	6.00E-08	Y	NA	NA
Survival Model 1	cg03093995	*DNAJC25-GNG10*	1.21	6.51E-08	Y	NA	NA
Survival Model 1	cg17679548	-	-0.46	7.11E-08	Y	NA	NA
Survival Model 1	cg18236982	-	-0.15	1.27E-07	Y	NA	NA
Survival Model 1	cg00338391	-	0.76	1.28E-07	Y	NA	NA
Survival Model 1	cg11337053	*SIPA1L2*	-0.15	1.50E-07	Y	NA	NA
Survival Model 1	cg09853393	-	-0.10	1.73E-07	Y	NA	NA
Survival Model 1	cg14504586	*MED27*	-0.64	1.81E-07	Y	NA	NA
Survival Model 1	cg19019403	-	-0.24	1.96E-07	Y	NA	NA
Survival Model 1	cg02927174	*TRPS1*	1.01	2.30E-07	Y	NA	NA
Survival Model 1	cg20046119	*ZNF251*	0.15	2.31E-07	Y	NA	NA
Survival Model 1	cg13071729	*NAV1*	0.27	2.31E-07	Y	NA	NA
Survival Model 2	cg13071729	*NAV1*	0.29	1.24E-07	Y	NA	NA
Survival Model 2	cg05278651	-	-0.11	2.24E-07	Y	NA	NA
Survival Model 2	cg18236982	-	-0.16	1.13E-07	Y	NA	NA
Survival Model 2	cg04444399	-	-0.31	9.02E-08	Y	NA	NA
Survival Model 2	cg15036595	-	-0.16	3.72E-08	Y	NA	NA
Survival Model 2	cg19019403	-	-0.24	1.57E-07	Y	NA	NA
Survival Model 2	cg12151015	*MYBPC1*	0.11	2.59E-08	Y	NA	NA
Survival Model 2	cg13987792	-	-0.23	2.08E-07	Y	NA	NA
Survival Model 2	cg22769651	-	-0.49	1.70E-07	Y	NA	NA
Survival Model 2	cg08204867	*GRIN2A*	-0.16	2.87E-08	Y	NA	NA
Survival Model 2	cg17679548	-	-0.48	7.86E-08	Y	NA	NA
Survival Model 2	cg27551718	-	-0.14	1.99E-08	Y	NA	NA
Survival Model 2	cg01555270	*ATP5SL*	-1.24	2.23E-07	Y	NA	NA
Survival Model 2	cg15334209	-	-0.25	2.00E-07	Y	NA	NA

Phenotype	CpG name	Gene annotation	Beta	P	Novel finding?	Mean beta across published literature	# PMIDs with finding
Survival Model 2	cg18604947	-	0.16	2.99E-08	Y	NA	NA
Survival Model 2	cg25864218	*PAQR3*	-2.45	1.22E-08	Y	NA	NA
Survival Model 2	cg26269613	*IL15*	0.67	5.34E-08	Y	NA	NA
Survival Model 2	cg11588423	*GABBR1*	-0.20	1.28E-07	Y	NA	NA
Survival Model 2	cg00338391	-	0.75	1.21E-07	Y	NA	NA
Survival Model 2	cg07377396	*DNAH11*	0.46	2.26E-07	Y	NA	NA
Survival Model 2	cg20046119	*ZNF251*	0.15	1.73E-07	Y	NA	NA
Survival Model 2	cg03093995	*DNAJC25*	1.19	6.13E-08	Y	NA	NA
Survival Model 2	cg09853393	-	-0.11	2.12E-07	Y	NA	NA

In the EWAS of 3-year survival none of the 15 (model 1) or 23 (model 2) reported associations have previously been reported in published studies of OPC survival. Both survival EWAS models gave a top hit annotating to the *PAQR3* gene. Aberrant promotor methylation at this gene has been shown to be associated with prostate cancer [333], with the gene itself being an established tumour suppressor [334]. Within the context of HNC, *PAQR3* has been associated with tumorigenesis in oesophageal cancer [335, 336], although currently no literature has examined whether this gene affects oropharyngeal cancer specifically.

A limitation of the ~3-year survival EWAS is that it was conducted using whole-blood-based DNA methylation rather than tumour tissue-based methylation. Whole-blood methylation acts as a robust "biosocial archive" of exposure, but lacks the specificity to detect local, tumour-based epigenetic change. Accordingly, there may be valuable CpG sites undiscovered in this analysis which better characterise OPC survival and predict prognosis in tumour tissue. Furthermore, many whole-blood-based methylation changes in response to cancer survival are likely associated with immune response or inflammation, which are almost certainly confounded by cell composition effects rather than a direct epigenetic effect. However, whole-blood-based EWAS may also report associations which are not explained by cell composition effects, making it still a valuable tissue to assay; one that is also comparatively inexpensive and non-invasive to obtain and process versus tumour tissue. Moreover, despite the increased likelihood of confounding using blood-based methylation assays, the

cell composition effects themselves and their potential interactions with other processes may prove a particularly lucrative avenue to explore.

The consistent direction of effect between MR Egger, MR $Egger_{independent}$ and Burgess IVW estimates for the *SPEG* locus provide us with greater confidence that the IV is reliable and that there is sufficient statistical power to demonstrate preliminary evidence for a causal association with decreased survival. Expression of the *SPEG* gene shows specificity to vascular smooth muscle cells – the major cell type in blood vessel walls, in which smoking has been shown to produce abnormal function throughout the human body [337]. Functional annotations show the *SPEG* gene to be essential for cardiac function in particular, with deficiency of this gene reported to result in heart failure [338]. A lookup in the BIOS QTL Browser (https://genenetwork.nl/biosqtlbrowser/) confirms 20 cis-expression quantitative trait methylations (eQTMs) showing evidence of correlation between gene expression and methylation at this locus in whole blood, though further work evaluating tissue-specific expression is required. People with head and neck squamous cell carcinoma (HNSCC) have an elevated risk of non-HNSCC mortality that persists over their lifetime. Among people with HNSCC, the 5-year incidence of non-cancer mortality is 13% [339], with a high baseline risk of cardiovascular disease compared to matched controls [340, 341].

This is currently the first EWAS study investigating oropharyngeal cancer survival using a cox proportional-hazards model to investigate DNA methylation in relation to survival at ~3 years. A key strength of the study relates to the use of the EPIC array which profiles methylation at approximately twice as many CpG sites as its 450k predecessor. Across the EWASs of smoking, alcohol, HPV and both survival models, 39.4% of the CpG sites showing association at $P<2.4e^{-7}$ were specific to the EPIC array (43/109). However, proportionally, the results suggest that associations are not enriched with the inclusion of novel enhancer region CpGs from the EPIC array. A one-sided Fisher's exact test for enrichment of EPIC probes vs 450K probes in CpG sites below P: $2.4e^{-7}$ confirms this; $P > 0.99$, suggesting no evidence of enrichment.

It should be noted that, despite being an established biomarker with high sensitivity (>93%) and specificity (>94%) for oral HPV16 infection [21], HPV16 E6 seropositivity may underestimate the number of individuals in the data with a current HPV16 infection. It has been reported that HPV can colocalise to biofilm (a community of immotile bacteria encased in a self-produced glycocalyx matrix) in tonsillar crypts, representing a reservoir of latent oncovirus undetected by the immune system [342]. Therefore, it is possible that individuals in the data have a historically HPV16-driven OPC without

evidence of infection at time of assessment. As such, the EWAS results for HPV16 infection may be biased toward the null.

Collider bias may influence associations between the prognostic factors and progression in a case-only setting [343]. HPV, smoking and alcohol are all associated with OPC incidence; by only examining cases, incidence is conditioned on, potentially inducing an association between HPV, smoking, alcohol and any unmeasured confounding. Unmeasured, unknown, confounding cannot be adjusted for here, so if any unmeasured confounding is associated with survival, it may be that an association between a prognostic factor and survival is simply a result of the induced association of the prognostic factor and unmeasured confounding.

Some of the MR analyses highlight potential violations of its methodological assumptions. Primarily, those analyses where the MR Egger estimate shows an opposite direction of effect to the IVW estimate (*GFI1, PPT2, KHDC3L*) could indicate an IV where one or more of the genetic variants proxying methylation is disproportionately skewing the effect in a certain direction (horizontal pleiotropy). However, for each of these analyses, the MR Egger intercept test of heterogeneity (explained elsewhere [329, 330]) spans 0 (*GFI1* intercept: -0.25, 95% CI: -0.54 to 0.05, p-value: 0.10*; PPT2* intercept: -0.18, 95% CI: -0.58 to 0.23*,* p-value: 0.40; *KHDC3L* intercept: 0.07, 95% CI: -0.09 to 0.23, p-value: 0.37), indicating that directional pleiotropy is not causing the difference between the MR Egger and IVW estimates. A possible explanation of this finding, and one that cannot be ruled out, is that these analyses suffer from a type of bias known as weak instrument bias; a bias where the chance difference in confounders may explain more of the difference in phenotype between genotype subgroups than the instrument, thereby confounding the true causal estimate. Finally, in these three analyses, the true direction of effect cannot be determined with confidence, given that the confidence intervals span a null line of Y = 1; this is likely an artefact of low statistical power.

One notable limitation of the MR analysis is that it is likely particularly conservative as overlap was assessed between prognostic factor DMRs and survival DMRs only if they surpassed the multiple correction threshold in both analyses. This approach was chosen (rather than to test corrected prognostic factor DMRs for association with all survival DMRs, only correcting for a number of tests equal to the number of prognostic factor DMRs) to improve confidence that regional methylation was associated with *both* a prognostic factor and survival. In order to reduce the possibility that regional methylation was only associated with a prognostic factor (and only spuriously associated with survival), genuine causal mediation may have been missed at less-stringent p-value thresholds.

5.5. Conclusion

Within the context of OPC, novel epigenetic biomarkers from whole blood measured by the EPIC array were found to be associated with the prognostic factors of smoking and alcohol and with survival. Of these biomarkers, overlapping signals between prognostic factor and survival analyses were used in an MR framework to appraise the causal role of DNA methylation. Using a novel IVW approach, derived to investigate the causal effect of DNA methylation at DMRs against mortality for this project, a stretch of CpGs located within a DMR, found to be associated with smoking (located at Chr2:220325443-220326041; annotating to the *SPEG* gene), shows preliminary evidence of a causal effect on mortality (HR: 1.28, 95% CI: 1.14 to 1.43, P: 2.12x10-05). DNA methylation at this locus could potentially mediate some of the association between smoking and OPC survival. To strengthen the validity of these findings, replication analyses and a longer follow-up period in Head and Neck 5000 are necessary. Additionally, conducting an EWAS of OPC-specific mortality (when possible) and applying may provide more sights with greater clinical relevance.

Finally, the discovery of novel CpG sites associated jointly with prognostic factors (smoking and alcohol) and mortality, whilst not causal intermediates, highlight the potential of DNA methylation to be used as an objective *predictor* of prognostic-factor-specific mortality in those with OPC. By applying robust epigenetic signatures from EWAS in a weighted “methylation score” approach to predict phenotypes from OPC epigenetic data, DNA methylation may possess utility as an exposure indicator, notably where phenotype information is missing, or a significant misreporting bias is suspected.

CHAPTER 6. EPIGENETIC PREDICTION OF COMPLEX TRAITS IN INDIVIDUALS WITH OROPHARYNGEAL CANCER

6.1. Introduction

This chapter describes DNA methylation as a predictor of complex health and lifestyle factors in OPC and assesses its validity as a tool for phenotype prediction and risk stratification. Firstly, the chapter introduces four phenotypes (with robust EWAS data) reported to be associated with OPC prognosis in observational literature; two explaining a high fraction of attributable risk for OPC incidence (smoking and alcohol consumption) and two commonly associated with adverse outcomes (BMI and educational attainment). Secondly, epigenetic scores for the aforementioned phenotypes, from previously published EWAS, are applied to epigenetic data in HN5000. The predictive accuracy of these scores are assessed via area under a ROC curve. Finally, the prognostic value of epigenetically-predicted phenotypes are determined by assessing their effect on 5-year survival using Cox proportional hazards regression models in HN5000. This chapter seeks to examine the clinical utility of epigenetic prediction in the case of OPC (above and beyond self-reported phenotypes), and to further investigate potential pathways associated with established phenotype-mortality relationships.

Multiple examples of the utility of DNA methylation in trait prediction exist in the epidemiological literature. DNAm has been shown to serve as both a sensitive and specific biomarker of tobacco smoke exposure, with methylation status at one cytosine-phosphate-guanine (CpG) site in the aryl hydrocarbon repressor (*AHRR*) gene (cg05575921) having a predictive area under the receiver operating characteristic curve (AUC) for smoking status of 0.99 for current vs never smokers [344]. Moreover, previous studies have found that DNAm markers at smoking-related CpG sites, both individually and combined in "epigenetic scores" (methylation values derived from a weighted average of multiple CpG sites associated with a trait of interest), may have potential for improving lung cancer risk and mortality prediction over and above self-reported smoking information [345-347]. Epigenetic scores of other lifestyle characteristics, including alcohol consumption, body mass index (BMI) and educational attainment, have recently been developed in large training datasets [348] and have been shown to independently explain relatively large degrees of phenotypic variance (up to 60%). These too have been shown to serve as predictors of disease outcomes, as well as all-cause mortality in general population-based cohort studies [346, 348-350].

The utility of epigenetic predictors in estimating mortality risk in clinical cohorts of individuals diagnosed with disease has not been thoroughly investigated and epigenetic predictors may be able to reliably supplement prognostication. In the setting of a large prospective head and neck cancer cohort (the Head and Neck 5000 Study [209]), epigenetic and self-report data associated with four complex traits, namely alcohol consumption, smoking, body mass index (BMI), and educational

attainment, were used to assess whether externally-derived DNAm predictors could provide an accurate prediction of phenotype in a subset of participants with oropharyngeal tumours. Furthermore, the validity of these predictors was assessed against mortality, given that all of these factors have been shown to be related to HNC mortality in previous studies [351-357]. The methylation predictors were then compared with the self-reported measures for the four phenotypes in terms of their predictive ability. Though smoking is the only phenotype predicted here which shows evidence of possessing a causal epigenetic intermediate between phenotype and mortality (see Chapter 5), the phenotypes proxied by epigenetic scores are important factors in OPC prognosis and may be able to augment the resolution of existing phenotypic information to improve risk stratification. Moreover, in the absence of directly measured phenotypic information, DNA methylation predictors may provide reliable estimates of unmeasured phenotype.

6.2. Methods

Study population

The study population for this analysis were individuals enrolled in the Head and Neck 5000 clinical cohort study (HN5000). Full details of enrolment criteria, phenotype measurement and data availability for this study can be found in the Methods section of this thesis.

Epigenetic risk score generation

DNA methylation scores for alcohol consumption, smoking, BMI and educational attainment were generated based on independently identified CpG sites from the largest or most recent EWAS in epidemiological literature. Details of regression model, sample size, year of publication and number of CpGs for each EWAS used to derive epigenetic risk scores are shown in **Table 6.1**.

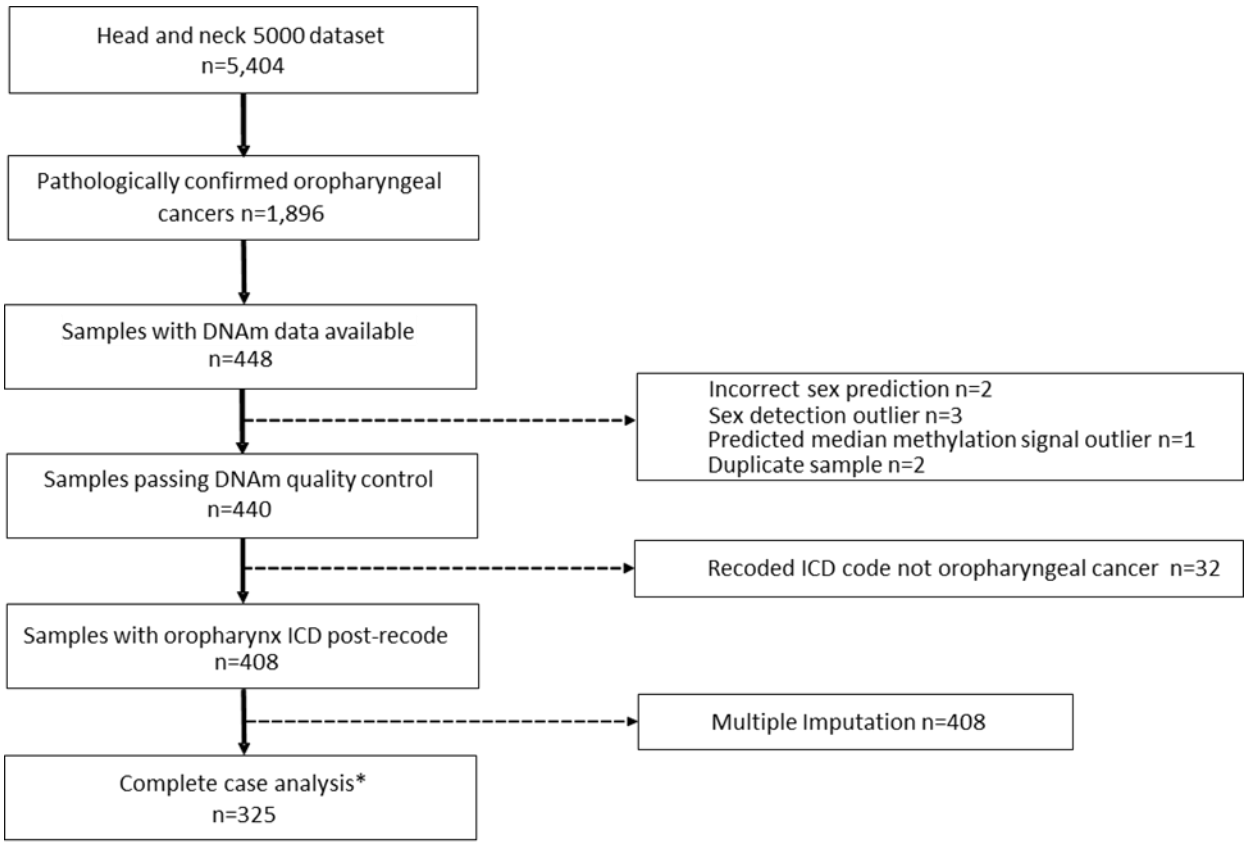
DNA extracted from blood samples taken from 448 individuals from HN5000 were run on the Illumina Infinium EPIC array. Of these, 440 passed quality control (2 samples with incorrect sex prediction, 3 samples with sex detection outliers, 1 sample with an outlier in predicted median methylated vs unmethylated signal, 2 duplicate samples). An additional 32 individuals were subsequently removed from the analysis owing to pathological re-classification, leaving 408 participants with DNAm data available (**Figure 6.1**). The primary analysis included 364 individuals with epigenetic data and covariate data (excluding BMI measures; see 6.3 – Results – Sensitivity Analyses) available.

Table 6.1 - Details of regression model, sample size, year of publication and number of CpGs for each EWAS used to derive epigenetic risk scores

Phenotype	Origin publication	EWAS model	# CpG sites
Alcohol consumption	"A DNA methylation biomarker of alcohol consumption" Liu et al. 2018 [307]	EWAS were conducted initially using linear models per cohort. Next, an inverse variance-weighted random-effects model was used to meta-analyse 8 European-ancestry cohorts. CpGs from the meta-analysis were taken forward and included in a least absolute shrinkage and selection operator (LASSO) regression in an independent cohort, with four selection criteria used to select CpGs with predictive value of alcohol consumption	Model 1: 5 Model 2: 23 Model 3: 78 Model 4: 144
	"Epigenetic prediction of complex traits and death" McCartney et al. 2018 [348]	EWAS were conducted using a LASSO regression model with k-fold (k=10) cross-validation.	450
BMI	"Epigenetic prediction of complex traits and death" McCartney et al. 2018 [348]	EWAS were conducted using a LASSO regression model with k-fold (k=10) cross-validation.	1109
	"Bayesian reassessment of the epigenetic architecture of complex traits" Trejo Banos et al. 2018 [358]	EWAS were conducted using a Baysian framework with BayesRR	144
Educational attainment	"Epigenetic prediction of complex traits and death" McCartney et al. 2018 [348]	EWAS were conducted using a LASSO regression model with k-fold (k=10) cross-validation.	373
Smoking	"Epigenetic Signatures of Cigarette Smoking" Joehanes et al. 2016 [182]	Linear mixed models were conducted, then combined in a random-effects model meta-analysis. After meta-analysis, one set of CpGs was selected based on a Bonferroni p-value of $P<1x10-7$ (485,381 tests) and another was selected based on a genome-wide false discovery rate P-value<0.05	Bonferroni model: 2623 FDR model: 18760

	"Self-reported smoking, serum cotinine, and blood DNA methylation" Zhang et al. 2016 [359]	An EWAS of cotinine concentration was conducted using median quantile regression, then CpG sites were individually validated against estimated average cigarettes per day using restricted cubic spline regression. Results were filtered by optimising AUCs derived from logistic regression for smoking status (current vs never; former vs never).	4
	"Bayesian reassessment of the epigenetic architecture of complex traits" Trejo Banos et al. 2018 [358]	EWAS were conducted using a Baysian framework with BayesRR	59

*Figure 6.1 - Flow diagram of HN5000 participants included in the analysis *Data available for age, gender, TNM stage, HPV status, comorbidity, education, self-reported smoking status and alcohol consumption.*



For each individual, methylation scores were calculated per individual in HN5000 as the product-sum of each CpG beta value from the respective EWAS in Box 1, multiplied by the normalised methylation value of the same CpG site in the HN500 MethylationEPIC data.

6.2.1. Statistical analysis

Associations of epigenetic scores with self-reported phenotypes

Linear regression analyses of epigenetic risk scores were performed against directly measured phenotype data (score ~ phenotype) to determine the variance of each phenotype explained by the methylation score attempting to predict that phenotype. The R-squared statistic generated by the "lm" function of the core Stats package in R (v3.4.1) was used as a measure of variance explained per model. R-squared statistics were derived as follows:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The r-squared value is the quotient of the variances of the fitted values and observed values of the dependent variable. For the equation above, let y_i be the observed values of the dependent variable, $\bar{y}$ be the mean, and $\hat{y}_i$ be the fitted regression value

The epigenetic prediction scores (derived using EWAS results from **Table 6.1**, in HN5000 epigenetic data) explaining the greatest proportion of variance for each phenotype were taken forward as exposure variables in survival analyses. Furthermore, scores showing the greatest variance explained were considered the most predictive epigenetic risk scores of alcohol consumption, BMI, educational attainment and smoking, respectively. In these scores, the degree of variance explained was investigated further, to determine if it changed notably (that is, greater than additively) with the inclusion of a corresponding polygenic risk score as a variable in each epigenetic risk score model.

Survival analyses

All-cause mortality, defined as the time in days from study enrolment to date of death from any cause or the date of censorship (i.e., the last date of follow-up), was the failure variable in the survival analyses. Primary analyses included cases with complete data only (i.e. participants with complete data for all the covariates used in the adjusted models and epigenetic data available; N=364)). Kaplan-Meier curves and the log-rank test were first used to investigate the univariate impact

of covariates on mortality. A test for nonproportional hazards using the Schoenfeld residuals was performed. Mortality risk was then assessed in relation to each of the directly measured phenotypes (i.e. self-reported smoking, alcohol drinking, BMI and education level), DNAm scores and polygenic scores using Cox proportional-hazards models. DNAm scores and polygenic scores were standardised (z-scored) to allow direct comparison. Hazard ratios (HRs) and 95% confidence intervals (CIs) for mortality were calculated for each standard deviation (SD) increase in score. For phenotypic measures, HRs represent the increase in mortality risk for current smokers versus never-smokers, hazardous to harmful drinkers versus non-drinkers or degree educated versus school educated individuals; BMI was treated as a continuous variable and therefore the HR represents the difference in mortality risk per unit increase in BMI.

To assess potential associations of directly measured phenotypes with mortality, three regression models were fitted: (1) a minimally adjusted model that controlled for age and sex; (2) a model that additionally adjusted for clinical factors (TNM stage, HPV status and comorbidity); and (3) a fully adjusted model that adjusted for the other directly-measured phenotypes of interest. Owing to issues of missing data, models examining the associations of self-reported smoking, alcohol drinking and education with mortality were not adjusted for directly measured BMI (model 3) because this would have reduced the sample size, and therefore the statistical power. The aforementioned clinical factors were selected on the basis of the strength of prior evidence linking them with HNC survival. Higher TNM stage is consistently associated with poorer survival [360]. HPV positivity, despite being a risk factor for OPC (that is, tumors driven by HPV infection, in particular HPV16) confers a marked survival advantage to those with OPC without HPV-driven tumors [76]. Comorbidity greatly affects all-cause mortality in both general populations and cancer populations [361, 362]. Finally, ethnicity was not included as a potential covariate in this study because the cohort is almost exclusively White British (97.1%).

Four separate models were fit to examine the relationship between DNAm scores with mortality: (1) A minimally adjusted model that adjusted for age, sex, cell counts and batch effects (epigenetic scores); (2) a 'clinical model', as above; (3) a model that additionally adjusted for the corresponding directly-measured phenotype (e.g. models that examined the association of smoking-related DNAm scores with mortality adjusted for self-reported smoking status) and 4) a model that additionally adjusted for the other directly measured phenotypes (excluding BMI to preserve sample numbers-see above).

As described above, it was decided *a priori* not to restrict the complete case analysis to participants with directly measured BMI data available due to the amount of missing data, as this would decrease the statistical power to detect an effect of our exposures on mortality. Therefore, as a sensitivity analysis, the same dataset was analysed as above, but restricted to complete data for BMI (measured at baseline).

Predictive accuracy of epigenetic risk scores against mortality

To assess the efficacy of epigenetic risk scores in prediction of mortality (that is, the accuracy with which the epigenetic risk scores for phenotypes could independently predict mortality, rather than directly affect it), ROC curves of epigenetic risk scores against our all-cause mortality variable were derived using the pROC R package [363]. Area under the ROC curves was computed using the trapezoidal rule. With all-cause mortality as the response variable, 3 ROC curves were generated per phenotype (alcohol consumption, BMI, educational attainment, smoking):

1. Epigenetic risk score of the phenotype as the predictor
2. Directly-measured phenotype as the predictor
3. A generalized linear model combining both epigenetic risk score of the phenotype and the directly measured phenotype

6.3. Results

The baseline descriptive statistics of included participants are presented in **Table 6.2**. Descriptive statistics stratified by human papilloma virus (HPV) status in are shown in **Table 6.3**. Seventy-eight of the 364 individuals died during a median follow-up period of 5.3 years (IQR: 4.9-5.9). OPC associated with HPV has been established in literature as a different pathological subtype with improved prognosis, compared to the more “traditional” OPC associated with alcohol consumption and smoking [364]. The Kaplan-Meier survival curves for mortality based on the covariates of interest are shown in **Figure 2a** and **Figure 2b**.

Table 6.2 - Baseline descriptive statistics of included participants (n=364), by gender, age at enrolment, TNM stage, HPV status, BMI, education, smoking and alcohol intake. Created by Rhona Beynon

	Alive (n=273)		Dead (n=91)		
Characteristic	N	Frequency	N	Frequency	*P*-value
Gender					
Male	209	76.6%	75	82.4%	
Female	64	23.4%	16	17.6%	0.242
Age at enrolment					
< 44	20	7.3%	3	3.3%	
45 to 54	83	30.4%	22	24.2%	
55 to 64	113	41.4%	34	37.4%	
65 to 74	48	17.6%	22	24.2%	
75 +	9	3.3%	10	11.0%	0.016
TNM stage					
Low (1-2)	39	14.3%	8	8.8%	
High (3-4)	234	85.7%	83	91.2%	0.176
HPV status					
Negative	61	22.3%	48	52.7%	
Positive	212	77.7%	43	47.3%	<0.001
BMI group					
not overweight	73	38.0%	31	55.4%	
overweight or obese	119	62.0%	25	44.6%	0.021
Comorbidity					
None	164	60.1%	34	37.4%	
Mild	73	26.7%	29	31.9%	
Moderate/Severe	36	13.2%	28	30.8%	<0.001
Education level					
School education	116	42.5%	45	49.5%	
College	111	40.7%	34	37.4%	
Degree	46	16.8%	12	13.2%	0.470
Self-reported smoking status					
Never	96	35.2%	11	12.1%	
Former	140	51.3%	49	53.8%	
Current	37	13.6%	31	34.1%	<0.001
Self-reported alcohol intake					
Non-drinker	75	27.5%	22	24.2%	
Moderate	68	24.9%	15	16.5%	
Hazardous-harmful	130	47.6%	54	59.3%	0.119

Abbreviations: BMI, body mass index; HPV, human papillomavirus; N, number.* Comorbidity was defined using the Adult Comorbidity Evaluation-27 (ACE-27) index [365]. For the purposes of analysis, moderate and severe comorbidity groups were combined.

Table 6.3 - Baseline descriptives of included participants as in table 6.2, stratified by HPV status. Created by Rhona Beynon

	HPV-negative n=109		HPV-positive n=255			
Variable	**N**	**Frequency**	**N**	**Frequency**	**Total**	***p*-value***
Gender						
Male	87	79.8%	197	77.3%	284	
Female	22	20.2%	58	22.7%	80	0.589
Age at enrolment						
< 44	4	3.7%	19	7.5%	23	
45 to 54	27	24.8%	78	30.6%	105	
55 to 64	46	42.2%	101	39.6%	147	
65 to 74	23	21.1%	47	18.4%	70	
75 +	9	8.3%	10	3.9%	19	0.216
TNM stage						
I	9	8.3%	7	2.7%	16	
II	17	15.6%	14	5.5%	31	
III	21	19.3%	31	12.2%	52	
IV	62	56.9%	203	79.6%	265	<0.001
BMI group						
Not overweight (BMI ≤25)	43	39.4%	61	23.9%	104	
Overweight or obese (BMI >25)	66	60.6%	194	76.1%	260	0.003
Comorbidity						
None	44	40.4%	154	60.4%	198	
Mild	35	32.1%	67	26.3%	102	
Moderate/Severe	30	27.5%	34	13.3%	64	<0.001
Education level						
School education	50	45.9%	111	43.5%	161	
College	42	38.5%	103	40.4%	145	
Degree	17	15.6%	41	16.1%	58	0.918
Self-reported smoking status						
Never	17	15.6%	90	35.3%	107	
Former	46	42.2%	143	56.1%	189	
Current	46	42.2%	22	8.6%	68	<0.001
Self-reported alcohol intake						
Non-drinker	29	26.6%	68	26.7%	97	
Moderate	16	14.7%	67	26.3%	83	
Hazardous-harmful	64	58.7%	120	47.1%	184	0.038

Figure 6.2a - Kaplan-Meier survival curves based on demographic and clinical covariates. Created by Rhona Beynon

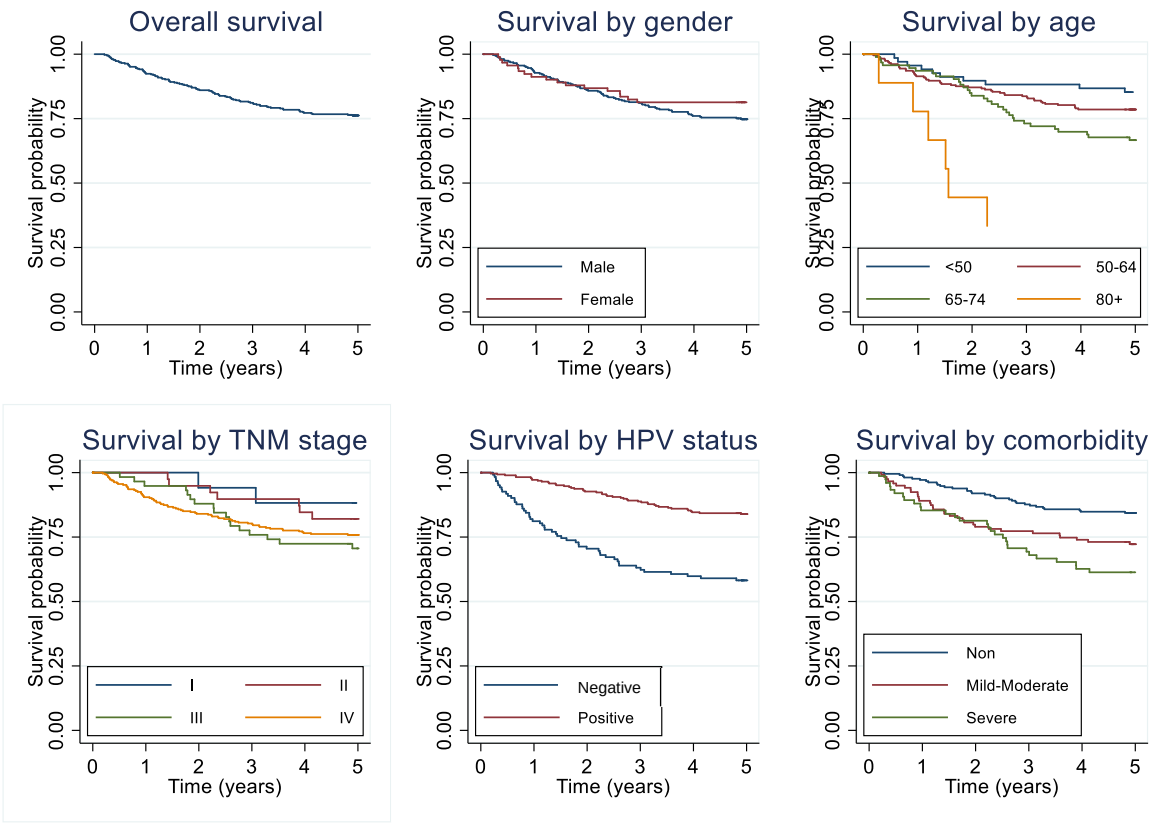
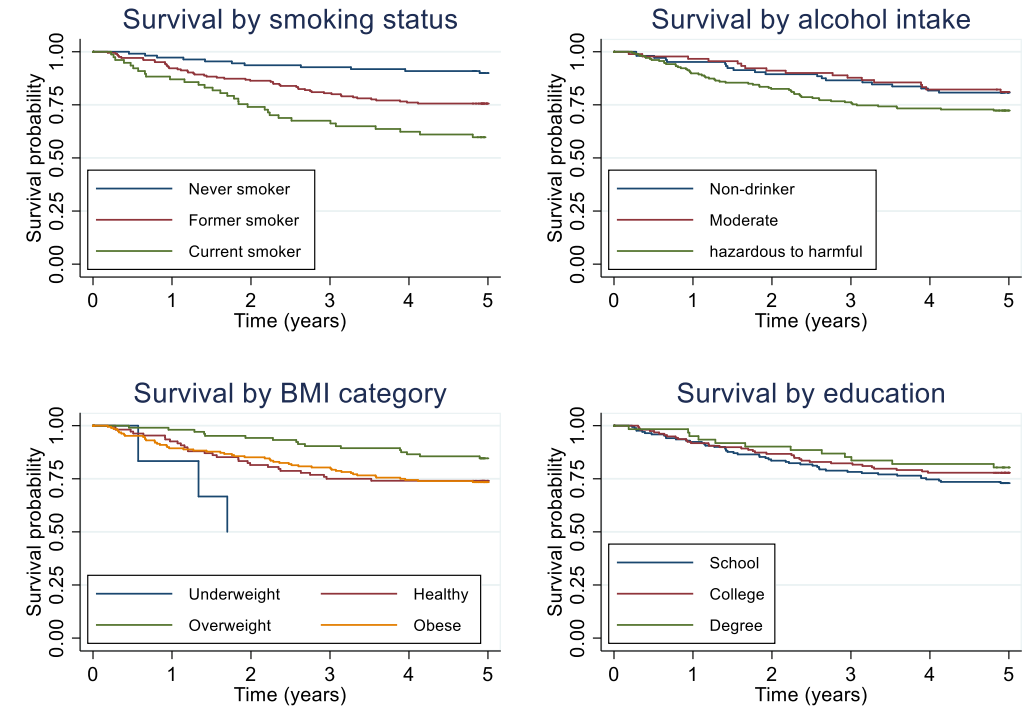


Figure 6.2b: Kaplan-Meier survival curves based on phenotypes of interest. Created by Rhona Beynon



Proportion of phenotypic variance explained for DNAm-based and genetic predictors.

A comparison of phenotypic variance explained by DNAm predictors can be found in **Table 6.4**. Where available, the Bayesian-derived epigenetic risk scores [358] seemed to explain a higher proportion of variance than their LASSO and glm-derived counterparts. The Bayesian-derived epigenetic score explained the greatest proportion of variance for smoking, explaining 51.06% of phenotypic variance. The epigenetic risk score of educational attainment explained the least variance explained of our phenotypes, of 0.40%.

Table 6.4 - Proportions of phenotypic variance explained by the epigenetic risk scores employed in this analysis

Methylation score	Variance explained in phenotype
Smoking	
Trejo Bayesian (59 CpG sites)	51.06%
AHRR (cg05575921)	48.96%
McCartney LASSO (233 CpG sites)	44.33%
Joehanes (Bonferroni) (2623 CpG sites)	39.08%
Joehanes (FDR) (18,670 CpG sites)	22.71%
Zhang (4 CpG sites)	5.15%
Alcohol	
Liu Model 4 (144 CpG sites)	16.48%
Liu Model 3 (78 CpG sites)	16.18%
Liu Model 1 (5 CpG sites)	14.96%
Liu Model 2 (23 CpG sites)	11.56%
McCartney LASSO (450 CpG sites)	7.90%
BMI	
Trejo Bayesian (144 CpG sites)	22.70%
McCartney LASSO (1109 CpG sites)	21.05%
Educational attainment	
McCartney LASSO (373 CpG sites)	0.40%

Relationship between directly-measured phenotype and mortality

A comparison of multivariable Cox proportional hazards outputs for minimally adjusted and fully adjusted models is presented in **Table 6.5**. In minimally adjusted models, smoking was positively associated with mortality (ever vs never smoking HR: 2.34, 95% CI=1.71, 3.21; p=1.21x10^{-07}) while BMI appeared to be protective (HR per kg/m^2: 0.93, 95% CI= 0.87, 0.99; p=0.028). There was weak evidence to suggest that alcohol consumption was related to mortality risk (hazardous-to-harmful drinking vs moderate-to-non-drinking HR:1.26, 95% CI=0.97, 1.64; p=0.089). Educational attainment was not associated with mortality. The association of self-reported smoking status with all-cause mortality remained on full adjustment (HR: 1.72, 95% CI= 1.21, 2.45; p=0.003).

Relationship between DNAm scores and mortality

All the epigenetic predictors were related to mortality in the minimally adjusted models (**Table 6.5**), except for the BMI predictor derived by McCartney et al [348]. After adjusting for clinical factors and directly measured phenotypes (**Table 6.6**), only the smoking-derived DNAm scores developed by Joehanes *et al* (Bonferroni) and Zhang *et al* were associated with mortality risk (HRs: 1.36, 95% CI: 1.02 to 1.82; *p*: 0.034 and 1.28, 95% CI: 1.02 to 1.61; p: 0.036, respectively). There was additional evidence of a relationship between *AHRR* methylation status and mortality in the imputed analysis, whereby a SD unit decrease in cg05575921 methylation (smoking is associated with hypomethylation at this loci) was associated with a 25% decrease in risk of death (HR: 0.75, 95% CI: 0.56 to 0.98; *p*: 0.044) in the fully adjusted model.

Table 6.5 - Association of phenotypic and DNAm-based predictors of smoking, alcohol drinking, BMI and education with mortality. Created by Rhona Beynon

	Minimally adjusted*					Fully adjusted**				
Exposure	**N**	**HR**	**ll**	**ul**	**p-value**	**N‡**	**HR**	**ll**	**ul**	**p-value**
Directly measured phenotype										
Ever vs. never smoker	364	3.29	1.75	6.18	2.22E-04	364	2.21	1.14	4.30	0.019
Hazardous to harmful drinker vs. not	364	1.62	1.06	2.49	0.027	364	1.34	0.86	2.09	0.202
Higher education vs school education	364	0.81	0.54	1.22	0.320	364	0.87	0.57	1.31	0.503
BMI	248	0.93	0.87	0.99	0.028	248	0.98	0.92	1.06	0.664
DNAm score										
Smoking										
McCartney LASSO (233 CpG sites)	364	1.53	1.24	1.88	7.89E-05	364	1.20	0.94	1.52	0.144
Trejo Bayesian (59 CpG sites)	364	1.70	1.37	2.11	1.49E-06	364	1.26	0.93	1.72	0.140
AHRR (cg05575921)	364	0.59	0.48	0.74	1.72E-06	364	0.79	0.58	1.07	0.125
Joehanes (FDR) (18760 CpG Sites)	364	1.70	1.34	2.15	1.27E-05	364	1.35	0.99	1.84	0.056
Joehanes (Bonferroni) (2623 CpG Sites)	364	1.67	1.36	2.05	7.57E-07	364	1.38	1.04	1.83	0.025
Zhang (4 CpG Sites)	364	1.48	1.16	1.88	1.48E-03	364	1.28	1.02	1.60	0.036
Alcohol										
Liu (5 CpG sites)	364	1.32	1.10	1.57	2.50E-03	364	1.19	0.97	1.47	0.094
Liu (23 CpG sites)	364	1.26	1.04	1.52	0.019	364	1.10	0.89	1.36	0.357
Liu (78 CpG sites)	364	1.25	1.07	1.45	5.02E-03	364	1.20	0.99	1.45	0.067
Liu (144 CpG sites)	364	1.24	1.07	1.44	5.31E-03	364	1.21	1.00	1.46	0.052
McCartney LASSO (450 CpG Sites)	364	1.28	1.03	1.60	0.024	364	1.05	0.79	1.41	0.723
BMI										
Trejo Bayesian (144 CpG Sites)	364	0.78	0.63	0.97	0.024	248	0.77	0.56	1.08	0.132
McCartney LASSO (1109 CpG Sites)	364	0.85	0.68	1.06	0.146	248	0.77	0.57	1.04	0.093
Education										
McCartney LASSO (373 CpG Sites)	364	0.76	0.61	0.96	0.021	364	0.87	0.68	1.12	0.270

Abbreviations: *N, number; HR, hazard ratio; ll, lower confidence interval; ul, upper confidence interval. *Directly measured phenotypes adjusted for age and gender; epigenetic scores adjusted for age, gender, cell counts and batch effects. **phenotypes additionally adjusted for clinical variables (TNM stage, HPV status and co-morbidity), and a combination of smoking, alcohol intake, education and BMI, as appropriate to the model; risk scores additionally adjusted for clinical variables, the corresponding phenotype predicted by the score of interest and the remaining directly measured phenotypes (excluding BMI). ‡ sample numbers vary due to missing BMI data.*

Table 6.6 - Multivariable Cox proportional hazards results for model 2 (clinical) and model 3 (respective phenotype). Created by Rhona Beynon

	Model 2*					Model 3**			
Exposure	**N**	**HR**	**ll**	**ul**	**p-value**				
				Directly measured phenotype					
Ever- vs. never-smoker	364	2.47	1.29	4.72	0.006				
Hazardous to harmful drinker vs. not	364	1.47	0.95	2.27	0.084				
Higher education vs. school education	364	0.81	0.53	1.22	0.306				
BMI	248	0.96	0.89	1.02	0.169				
				DNAm score					
Smoking									
McCartney smoking (233 CpG sites)	364	1.34	1.07	1.67	0.011	1.05	0.79	1.41	0.726
AHRR (cg05575921)	364	0.66	0.52	0.83	4.11E-04	0.79	0.58	1.07	0.125
Joehanes (FDR) (18670 CpG sites)	364	1.59	1.22	2.08	6.11E-04	1.35	0.99	1.84	0.056
Joehanes (Bonferroni) (2623 CpG sites)	364	1.59	1.26	2.00	9.82E-05	1.38	1.04	1.83	0.025
Trejo Bayesian smoking (59 CpG sites)	364	1.51	1.2	1.91	4.34E-04	1.26	0.93	1.72	0.14
Zhang (4 CpG sites)	364	1.38	1.1	1.74	6.26E-03	1.28	1.02	1.6	0.036
Alcohol consumption									
Liu Model 1 (5 CpG sites)	364	1.25	1.03	1.51	0.023	1.19	0.97	1.47	0.094
Liu Model 2 (23 CpG sites)	364	1.17	0.96	1.42	0.13	1.1	0.89	1.36	0.357
Liu Model 3 (78 CpG sites)	364	1.25	1.05	1.49	0.014	1.2	0.99	1.45	0.067
Liu Model 4 (144 CpG sites)	364	1.26	1.06	1.49	9.99E-03	1.21	1.00	1.46	0.052
McCartney alcohol (450 CpG sites)	364	1.26	1.00	1.57	0.046	1.2	0.94	1.52	0.144
BMI									
McCartney BMI (1109 CpG sites)	364	0.82	0.66	1.02	0.075	0.77	0.57	1.04	0.093
Trejo Bayesian BMI (144 CpG sites)	364	0.77	0.61	0.97	0.025	0.78	0.56	1.08	0.132
Educational attainment									
McCartney education (373 CpG sites)	364	0.87	0.67	1.11	0.26	0.87	0.68	1.12	0.27

** Adjusted for age, gender, TNM stage, HPV status and comorbidity (plus cell count and batch effects for epigenetic (DNAm models)). ** additionally, adjusted for the respective directly-measured phenotype Cox proportional hazards results for model 2 (clinical) and model 3 (respective phenotype)*

Predictive accuracy of epigenetic risk scores against mortality

Across all four phenotypes assessed, the AUC when DNAm risk scores (**see Table 6.1**) were used to predict mortality was greater than when directly-measured phenotype attempted to do the same (**Figure 6.3**). However, none of these findings were found to improve upon the predictive value of directly-measured phenotype to a level of statistical significance (Z-test p-value for comparison of epigenetic AUC and directly-measured AUC for: smoking = 0.19, alcohol = 0.41, BMI = 0.62, educational attainment = 0.49). When a generalized linear model of epigenetic risk score and corresponding directly-measured phenotype was used to predict mortality, AUC improved over directly-measured phenotype alone, but also below a level of statistical significance (Z-test p-value for combined epigenetic risk score and directly-measured phenotype AUC vs directly-measured phenotype AUC for: smoking = 0.30, alcohol = 0.38, BMI = 0.71, educational attainment = 0.26). The most predictive epigenetic risk score for mortality was that of smoking with an area under the curve (AUC) of 0.70. The weakest epigenetic risk score predictor of mortality was the educational attainment score, with an AUC of 0.57.

Sensitivity analysis

A summary of the baseline descriptive characteristics of participants included in the sensitivity analysis is provided in **Table 6.7**. When the analysis was restricted to participants with data available for BMI (**Table 6.8**), the results of models examining the association of self-reported phenotypes with mortality were broadly comparable; only self-reported smoking was associated on full adjustment. When the relationships of DNAm scores with mortality were examined, there was evidence that the alcohol-related DNAm scores developed by Liu et al. were associated with mortality, in addition to the Joehanes and Zhang scores identified in the primary analysis.

Figure 6.3 - ROC curves detailing the predictive accuracy of epigenetic risk scores, directly-measure phenotype and a combination of the two, against 5-year mortality in HN5000. ROC curves are provided for smoking, alcohol consumption, BMI and educational attainment

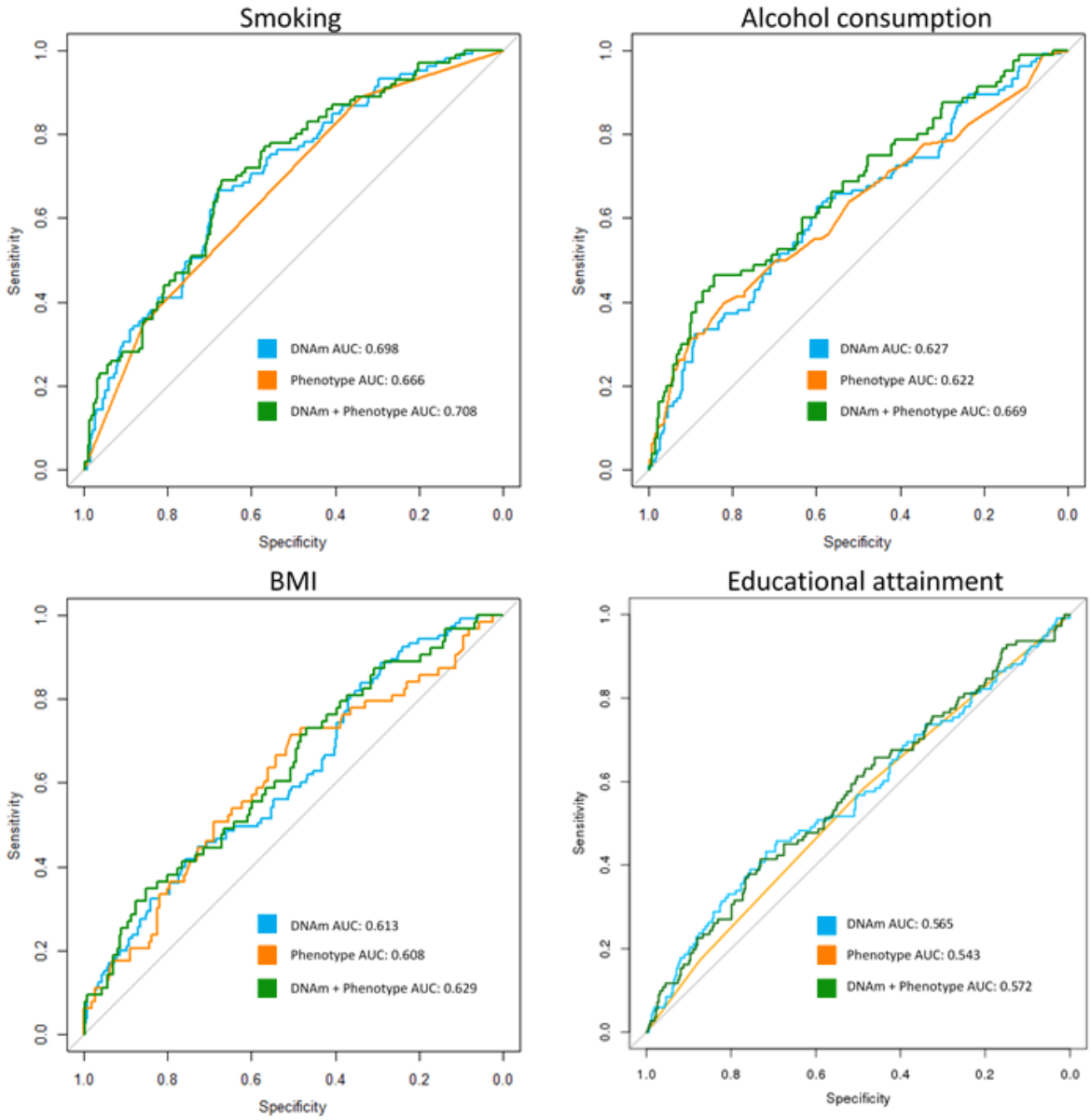


Table 6.7 - Baseline descriptives of participants included in the sensitivity analysis (n=248). Created by Rhona Beynon

	Alive (n=192)		Dead (n=56)		
Characteristic	N	Frequency	N	Frequency	p-value
Gender					
Male	147	76.6%	45	80.4%	
Female	45	23.4%	11	19.6%	0.550
Age at enrolment					
< 44	16	8.3%	1	1.8%	
45 to 54	61	31.8%	16	28.6%	
55 to 64	76	39.6%	17	30.4%	
65 to 74	32	16.7%	14	25.0%	
75 +	7	3.6%	8	14.3%	0.009
TNM stage					
Low	29	15.1%	7	12.5%	
High	163	84.9%	49	87.5%	0.626
HPV status					
Negative	43	22.4%	31	55.4%	
Positive	149	77.6%	25	44.6%	<0.001
BMI group					
not overweight	73	38.0%	31	55.4%	
overweight or obese	119	62.0%	25	44.6%	0.021
Comorbidity					
None	119	62.0%	25	44.6%	
Mild	51	26.6%	18	32.1%	
Moderate/Severe	22	11.5%	13	23.2%	0.031
Education level					
School education	84	43.8%	30	53.6%	
College	80	41.7%	20	35.7%	
Degree	28	14.6%	6	10.7%	0.414
Self-reported smoking status					
Never	72	37.5%	7	12.5%	
Former	98	51.0%	29	51.8%	
Current	22	11.5%	20	35.7%	<0.001
Self-reported alcohol intake					
Non-drinker	53	27.6%	15	26.8%	
Moderate	47	24.5%	9	16.1%	
Hazardous-harmful	92	47.9%	32	57.1%	0.349

Table 6.8 - Results of the sensitivity analysis, restricted to those with complete data (including BMI). Created by Rhona Beynon

	Minimally adjusted					Fully adjusted				
Exposure	**N**	**HR**	**ll**	**ul**	**p-value**	**N**	**HR**	**ll**	**ul**	**p-value**
Directly measured phenotype										
Ever vs. never smoker	248	3.64	1.65	8.07	**1.42E-03**	248	2.47	1.07	5.73	**0.035**
Hazardous to harmful drinker vs. not	248	1.48	0.86	2.56	0.16	248	1.24	0.70	2.21	0.46
Higher education vs school education	248	0.74	0.44	1.25	0.26	248	0.94	0.54	1.64	0.84
BMI	248	0.93	0.87	0.99	**0.028**	248	0.98	0.92	1.06	0.66
DNAm score										
McCartney smoking	248	1.49	1.13	1.97	**4.31E-03**	248	1.32	0.97	1.81	0.076
McCartney alcohol	248	1.31	0.98	1.76	0.067	248	1.04	0.71	1.51	0.85
McCartney BMI	248	0.76	0.57	1.01	0.059	248	0.77	0.57	1.04	0.093
McCartney Education	248	0.86	0.65	1.14	0.29	248	0.94	0.69	1.28	0.69
Liu 5 CpG alcohol	248	1.36	1.08	1.73	**9.39E-03**	248	1.43	1.07	1.92	**0.017**
Liu 23 CpG alcohol	248	1.33	1.03	1.72	**0.029**	248	1.33	0.98	1.80	0.068
Liu 78 CpG alcohol	248	1.23	1.02	1.49	**0.028**	248	1.32	1.03	1.69	**0.027**
Liu 144 CpG alcohol	248	1.22	1.01	1.46	**0.037**	248	1.29	1.02	1.63	**0.036**
AHRR (cg05575921)	248	0.63	0.47	0.83	**1.28E-03**	248	0.89	0.60	1.32	0.55
Joehanes	248	1.84	1.36	2.49	**7.43E-05**	248	1.59	1.09	2.32	**0.012**
Joehanes strict	248	1.72	1.32	2.24	**5.24E-05**	248	1.50	1.06	2.12	**0.022**
Zhang	248	1.41	1.04	1.91	**0.029**	248	1.33	1.00	1.77	**0.047**
Bayes smoking	248	1.61	1.21	2.14	**1.17E-03**	248	1.12	0.75	1.67	0.58
Bayes alcohol	248	0.76	0.59	0.99	**0.045**	248	0.77	0.56	1.08	0.13

Abbreviations: *N, number; HR, hazard ratio; ll, lower confidence interval; ul, upper confidence interval. *Directly measured phenotypes adjusted for age and gender; risk scores adjusted for age, gender, cell counts and batch effects. **phenotypes additionally adjusted for clinical variables (TNM stage, HPV status, co-morbidity and BMI) and a combination of smoking, alcohol intake, education and BMI, as appropriate to the model; risk scores additionally adjusted for clinical variables and corresponding phenotype.*

6.4. Discussion

Analyses were undertaken to attempt to define the predictive accuracy of epigenetic risk scores for smoking, alcohol drinking, BMI and educational attainment, in comparison with directly measured or self-reported phenotypes. These epigenetic scores were used to assess ~3-year mortality risk in a clinical cohort of individuals with oropharyngeal cancer. In all models, the epigenetic risk scores explaining the largest amount of phenotypic variance yielded similar mortality estimates to directly measured or self-reported phenotypes and may therefore provide a useful measure of these exposures in future epidemiological studies, particularly if directly measured phenotypic data is not available.

Results from the fully-adjusted model show that phenotypically, smoking is the only trait strongly associated with mortality risk after adjustment for age, sex, TNM stage, HPV status, comorbidity, alcohol consumption and educational attainment (HR: 2.21, 95% CI: 1.14 to 4.30; *P:* 0.019 for ever versus never-smokers). Similarly, when investigating DNAm risk scores, only smoking was associated with mortality after adjustment for the covariates above, cell counts and BeadChip ID for batch (Joehanes Bonferroni HR: 1.38, 95% CI: 1.04 to 1.83; *P:* 0.025, Zhang HR: 1.28, 95% CI: 1.02 to 1.60; *P:* 0.036). Finally, it should be noted that alcohol consumption showed weak evidence of association with mortality (with similar magnitude to smoking analyses) both phenotypically (HR: 1.34; 95% CI: 0.86 to 2.09; *P*: 0.202) and as an epigenetic risk score (HR: 1.21; 95% CI: 1.00 to 1.46; *P:* 0.052) when adjusting for the same covariates in our full-adjusted model.

For the prediction of mortality using epigenetic scores, the two predictors that were derived using results from a Bayesian framework explained the most phenotypic variance, thus were employed over other epigenetic scores derived using a LASSO/linear mixed-effects regression. Interestingly, despite explaining the largest amount of phenotypic variance, neither Bayesian predictor was associated with mortality as strongly as the respective phenotypic measures. For smoking, the Joehanes Bonferroni epigenetic score was most strongly associated with mortality; for BMI, the McCartney LASSO epigenetic score was most strongly associated with mortality. One potential explanation for this finding is that the Bayesian scores are proxying characteristics of smoking and BMI which might not necessarily directly impact survival. However, whilst not affecting survival, the aspects of smoking that the Bayesian scores proxy may better predict the phenotype. An example of such a characteristic may be a susceptibility to social peer pressure [366, 367], though this is currently only speculative.

In the minimally adjusted model, phenotypic BMI, epigenetic risk scores for alcohol and an epigenetic risk score for education all show an association with mortality. However, when adjusted for clinical covariates and co-adjusted for other phenotypes in the fully adjusted models, the associations attenuate. This observation may be explained by inadequately measurement of each of the above phenotypes. As such, in the minimally adjusted model, the above phenotypes may be closely associated with one or more of the covariates adjusted for in our fully adjusted model, and the observed association with mortality actually driven by the association of a covariate with mortality. Alternatively, the attenuation seen in the fully adjusted model results could reflect a true association masked by an inflated standard error due to addition of multiple covariates. A notable limitation of this analysis is the small sample size; in the survival analysis, it cannot be said with certainty that the observed change in effect size between models is a true attenuation. A larger sample size or independent replication is necessary to resolve this issue.

This analysis has several strengths including the availability of MethylationEPIC epigenetic data and comprehensive mortality follow-up data in the same cohort, as well as the ability to adjust for multiple biological, clinical and lifestyle covariates, including HPV, presents a major strength; it enables investigation of the association of methylation scores with mortality within a cancer cohort, a novel application of epigenetic prediction of phenotypes which may have clinical utility in the future.

A notable limitation of the analysis is that all-cause mortality was used as the mortality phenotype. This was because cause-of death data was not available for all participants in the current HN5000 data release. Moreover, previous work has shown that the cause of death information on a death certificate is often inaccurate [368]. Whilst all-cause mortality will be impacted by cancer status, it will not show specificity to OPC. Deaths could arise from competing causes such as cardiovascular disease, secondary cancer or age, preventing us from estimating phenotype risk on OPC-specific death. Interestingly, however, hazard ratio estimates are larger in this analysis compared to another study examining the association of epigenetic scores against mortality in a healthy population. McCartney *et al*. [348] report a HR of 1.29, 95% CI of 1.05 to 1.57 and *P* of 0.013 for their smoking epigenetic risk score (vs our smoking epigenetic score HR: 1.72; 95% CI: 1.21 to 2.45; $P = 2.50 \times 10^{-03}$). All-cause mortality estimates in those with OPC likely reflects the effect of sustained heavy tobacco and alcohol use (a hallmark demographic of HNC populations), in addition to presence of cancer on mortality. The marked HR differences seen between those with and without OPC illustrate a need to separately risk-stratify those with the disease from those without.

Another limitation is that sample sizes differ in models examining the effect of BMI on mortality risk, owing to missing data. As a result, these models are not directly comparable to those estimating the mortality risk associated with smoking, drinking and education because the individuals included differ. However, the baseline descriptive statistics of participants included in the models did not appear to be different, presumably because BMI data was missing at random. If BMI had been included as a covariate in the fully adjusted models, this would have reduced the statistical power further, as shown by the loss of precision in sensitivity analyses (N = 248), which adjusted for directly-measured BMI in all instances.

6.5. Conclusion

In summary, in the context of OPC, DNAm predictors are able to predict complex traits with a relatively high degree of variance explained for smoking, alcohol consumption and BMI; however, the educational attainment DNAm predictor did not display the same high degree of variance explained. Comparing the effect on mortality of both DNAm predictors and directly measured phenotype yielded similar results between the two, with methylation displaying similar effect sizes and variance explained of ~3-year mortality, across all traits assessed. Including genetic predictors as a covariate in our Cox regression analyses did not significantly affect our results. Findings suggest DNAm predictors can be used to supplement phenotypic prediction of mortality, potentially even providing reliable insight into smoking, alcohol consumption, BMI and educational attainment in situations where directly-measured phenotype information is not available.

CHAPTER 7. THESIS DISCUSSION AND CONCLUSION

7.1.1. Introduction

This body of work aimed to augment understanding of OPC by investigating aetiological factors and biological pathways with progressively greater granularity than had been available in the published literature. First, existing risk factors from observational epidemiological literature were collated. This was undertaken to form an evidence base from which to prioritise hypotheses and highlight potential gaps in current OPC research. The use of MELODI as a literature mining tool allowed for high-throughput retrieval and quantification of the proportion of all PubMed literature which mentioned an enriched, directional association from any pre-defined risk factor to OPC.

Next, using a hypothesis-free approach, the causality of a vast number of phenotypes was appraised against risk of OPC in an MR-PheWAS framework. The purpose of this analysis was firstly to quantify the effect of established risk factors on OPC-specific risk (rather than OPC in combination with other HNC sub-types, as is currently prevalent in observational literature) using a robust causal inference method which can circumvent the innate biases of observational studies (namely confounding and reverse causation). Secondly, the MR-PheWAS was conducted to ascertain whether any novel, genetically-proxied phenotypes were associated with OPC risk outside of those already established. By utilising the vast scope and scale of published summary GWAS data in an MR approach, a comprehensive "scan" of the causal landscape of OPC could be achieved, investigating many previously untested associations between phenotype and OPC risk.

Finally, DNA methylation has been established as an exposure indicator and causal molecular intermediate in epidemiological literature [369] EWAS, DMR and a novel 2-step MR approach were employed to assess whether risk factors for OPC affected disease aetiology via mediation by DNA methylation. Assessing mediation could provide novel therapeutic targets and biological pathways to prioritise, that is the methylation variable locus could be targeted as opposed to the risk factor itself to modulate OPC risk. Furthermore, ascertainment of DNA methylation patterns associated with OPC mortality or prognostic factors could shed insight into OPC progression pathways and allow for greater accuracy when stratifying OPC subgroups (such as HPV-IMU, HPV-KRT and HPV-negative OPC). This is of particular interest in light of unexpected de-escalation clinical trial findings for HPV-positive OPC [370]. Patient segmentation for entry into trials might be a fruitful strategy in future to increase the likelihood of success of new treatments and that DNA methylation biomarkers may be helpful to group patients in this way. Phenotypic prediction using DNA methylation also has clinical translation to circumvent bias in the self-report of relevant health behaviours or in the prediction of a missing phenotype in the absence of direct measurement.

Below, each chapter is discussed with respect to its contribution to wider epidemiological literature. Overall strengths and limitations of this thesis are then discussed, before exploration of future research directions informed by the work completed.

7.1.2. Discussion

Novel contributions to epidemiological literature and OPC aetiology

Systematic retrieval of OPC risk factors enriched in epidemiological literature

The use of MELODI to retrieve risk factors for disease is a novel application of this platform. The use of SemMedDB "triples" to retrieve OPC risk factors (i.e. risk factor → PREDICATE → OPC) allowed for ascertainment of the degree of evidence, semantically, to which a risk factor was related to OPC risk ("associated with", "predisposes" or "causes"). Furthermore, this method provided an evidence base of risk factors which reflected existing literature, albeit with "noise" filtered out through use of an enrichment step. Finally, MELODI provided large numbers of potential intermediates between smoking and OPC, alcohol and OPC, HPV and OPC, and oral sex and OPC.

MR-PheWAS

MR is recognised as a robust causal inference method in epidemiological literature, to the degree that MR study findings have contributed evidence to IARC Monographs of cancer risk factors. Furthermore, MR-PheWAS have been able to establish novel risk factors for other cancers in peer-reviewed epidemiological literature. Here, findings from 17,4449 phenotypes were largely in concordance with existing literature, finding a concentration of alcohol, sexual activity and smoking phenotypes to be most associated with risk of OPC. However, using a two-sample MR approach to interrogate OPC-specific risk is a key strength and novel contribution to epidemiological literature. Furthermore, MR is known to circumvent the issues of confounding and reverse causation typically seen in observational literature, thus provides new, robust evidence to augment the established effects of smoking, weight and sexual activity on OPC risk. The finding of weight as inversely associated with OPC risk provides greater evidence that the direction of effect for this observation may indeed be from low weight to OPC (ambiguous in observational literature due to study design), rather than an artefact of reverse causation (OPC causing a decrease in weight). Because genotype is inherited at conception, a genetically-proxied phenotype will always precede OPC in this framework. Finally, a key finding of this analysis which should not be overlooked is the number of findings which were not associated with risk of OPC. By virtue of the sheer number of phenotypes investigated, many of them

will not have previously been appraised against OPC risk. Such phenotypes will now have evidence for de-prioritisation (and a comparative effect estimate) in future risk factor studies.

DNA methylation as a mediator of OPC prognostic factors and mortality

Due to its proximity to the genome, DNA methylation patterns can reveal important gene expression pathways for OPC aetiology. Epigenetic signatures have been established in epidemiological literature as exposure indicators and causal mediators for risk of multiple diseases, including OPC. However, DNA methylation patterns for OPC prognostic factors have not been investigated in a case-only context before and Illumina MethylationEPIC-derived signatures have not been investigated in relation to genome-wide OPC mortality. In 2017, Degli Esposti et al. investigated Illumina 450K for association with HPV status in HNSCCs, but this study combined oral, oropharyngeal, hypopharyngeal and laryngeal cancers. Given the uniquely-large effect of HPV on OPC, combining these cancer sites has potentially biased methylation findings unique to the OPC anatomical site towards the null. The specificity of HPV to OPC can be seen within this study when comparing the multidimensional scaling (MDS) plot generated for HNC anatomical site and MDS plot for HPV status (**Figure 7.1**); points describing samples with positive HPV status correlate almost perfectly with points describing the OPC subtype and not with oral cavity, hypopharyngeal and laryngeal sites.

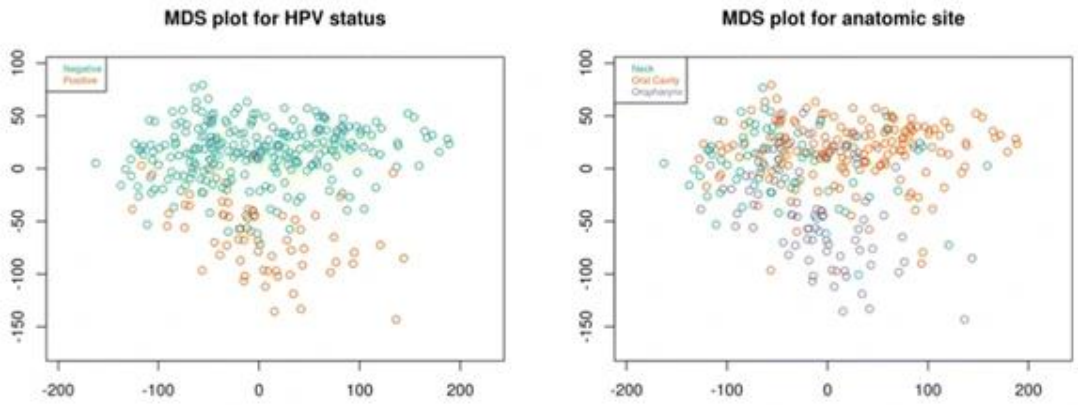


Figure 7.1 – Adapted from Degli Eposti et al.: MDS plots for HPV status (left) and HNC anatomical site (right) show significant correlation between OPC and HPV status

Chapter 5 of this thesis employs one of the largest OPC epigenetic data resources in epidemiological literature (N = 448), investigating DNA methylation data from the current BeadChip platform with the greatest coverage (Illumina MethylationEPIC). Novel methylation signatures relating to smoking, alcohol consumption and ~4-year OPC mortality were discovered in single-site EWAS. Moreover, DMRs of novel gene regions were found to be associated with smoking, alcohol

consumption, HPV and OPC mortality in a separate analysis. A novel application of MR, investigating DMR methylation against OPC mortality, identified that shared DNA methylation between smoking and OPC mortality at the *SPEG* gene locus and indicated a causal role for this gene locus in OPC mortality. This finding may have clinical relevance for OPC and prognostic studies more broadly if methylation and/or expression of *SPEG* is confirmed to be causally related to with survival. For example, there may be scope to target DNA methylation at this gene region therapeutically if a proportion of the effect of smoking on mortality is mediated through this pathway. However, appropriate validation and replication studies need to be conducted to establish the true effect of smoking-related DNAm at the *SPEG* gene region on mortality. Furthermore, quantification of the proportion of smoking-related mortality risk at this gene region will be crucial in determining whether targeting it is a cost-effective therapeutic target.

Epigenetic prediction of complex traits and mortality

The utility of DNA methylation as an exposure indicator is an area of increasing research interest, exemplified in the burgeoning literature around the prediction of age through the use of epigenetic clocks [371]. As DNA methylation can reflect both a variety of risk factors and simultaneously capture information on the early stages of disease, it has high potential for impact in clinical settings. Prediction of phenotypes, improvement of risk estimation/prognostication and insight into biological processes are all potential applications of DNA methylation [372, 373]. McCartney et al. found that DNA predictors of alcohol consumption, BMI, education and smoking correlate with lifestyle factors associated with health outcomes and mortality [348].

In OPC, comparing the effect on mortality of both peripheral blood DNAm predictors and self-reported phenotypes yielded similar results between the two. DNAm indices of established risk factors (alcohol, smoking, BMI and educational attainment) predicted 3-year mortality with the same accuracy as self-reported data, displaying similar effect sizes and variance explained of mortality across all traits assessed. Findings from this analysis suggest peripheral blood DNAm predictors can be used to supplement a prediction model of mortality in those with oropharyngeal cancer, potentially providing reliable insight into smoking, alcohol consumption and BMI measures in situations, particularly where self-reported phenotype information is not available for these individuals. This avenue of enquiry could be explored further with the advent of more extensive measurement of DNA methylation across the genome. The current analysis was restricted to probes present on the Illumina Methylation EPIC BeadChip, whereas a technology such as whole genome bisulphite sequencing may provide more nuanced exposure or OPC indices for application in prediction or prognosis.

Strengths of work presented

Perhaps one of the greatest strengths of this thesis is that the methodologies employed throughout use data which is specific to OPC, as opposed to a heterogenous mix of multiple head and neck cancers. There is a paucity of epidemiological literature which investigates the effect of risk factors in an OPC-specific context. As shown in the introduction, most observational epidemiological findings present the effects of risk factors with respect to "pharyngeal" cancers (nasopharyngeal, hypopharyngeal and oropharyngeal cancers [or a combination thereof]) or "oral cancer" (oral cavity and oropharyngeal combined). Whilst combining HNC anatomical sites boosts statistical power, translation or clinical utility of these findings presents an issue; the quantitative effect of a given risk factor on an OPC patient is ambiguous when effect estimates are derived for a combination of cancers. Despite their biological proximity, these cancers are not the same and therefore should not be treated so. A key example of this disparity is the comparison of hypopharyngeal cancer and OPC. HPV, in an observational context, confers a huge increased risk to OPC (OR: 147.3; 95% CI: 83.07 to 361.24) in large-scale multi-centre studies, whereas any link between HPV and hypopharyngeal risk is only currently proposed in a few small-scale studies, with few, if none, presenting ORs for hypopharyngeal cancer incidence [374]. Given the predominance of HPV as a risk factor for OPC and the lack of robust association with hypopharyngeal cancer, it would appear the causal landscape of these cancers differ. Pathologically-confirmed OPC cases in the HN5000 genetic and epigenetic data allow for OPC-specific risk estimates and appraisal of biological pathways with minimisation of bias from other primary sites.

Secondly, the use of Mendelian randomization in particular is a key strength throughout this thesis. Observational epidemiological literature is prone to innate biases that prevent true causality from being established, the most notable being unmeasured confounding. Mendelian randomization circumvents the issue of unmeasured confounding by relying on genetic variants to proxy a phenotype of interest. Genetic variants which are uniquely associated with an exposure of interest, independent of OPC, cannot be confounded by unmeasured confounding (with the exception of genetic confounding by population stratification, which in a well-designed study in greatly minimised). In a MR analysis, genetic variants that influence levels of a trait of interest are randomly allocated at conception, as per the Laws of Segregation and Independent Assortment. Groups defined by a certain genotype (proxying a risk factor) should then be wholly comparable in the distribution of both genetic and environmental confounding factors, except for their exposure to a trait of interest. Any observed differences in OPC between groups defined by genotype can then be solely attributed to differences in lifelong exposure to the proxied risk factor. However, this type of analysis is not without its

limitations, which have been reviewed at length elsewhere [254] but are briefly summarised below where they are pertinent to the findings in this thesis.

In addition to instrumental variables having to fulfil a strict set of assumptions, MR is susceptible to a number of practical and theoretical limitations. Firstly, not every trait has a genetic polymorphism associated with it. Despite the 17,449 phenotypes examined in Chapter 4, some phenotypes considered important to OPC aetiology, such as HPV16 infection, could not be appraised in an MR framework. However, despite this limitation, GWAS have been (and continue to be) published at an exponential rate since their conception, with more traits per GWAS and a greater frequency of GWAS published year-on-year. Future MR-PheWAS can therefore be expected to investigate many more traits than those seen in Chapter 4 and provide increasingly comprehensive information regarding the causal landscape of a disease. A key development in GWAS design, and by extension the scope of MR methodology, is the appraisal of molecular intermediates. It is important to know which risk factors causally affect a disease; once this has been determined, it is then important to know *how* a risk factor affects a disease. By conducting MR from OPC intermediates to OPC mortality, biological pathways of disease aetiology can begin to be investigated. Moreover, if a molecular intermediate (i.e. methylation) can be established as a causal mediator of the two, these findings may be of clinical use as either risk stratification biomarkers or therapeutic targets. This thesis has shed light on one such pathway/target in methylation at the *SPEG* gene (Chapter 5), potentially mediating the causal association seen between smoking and OPC mortality. Whilst more analysis is needed to confirm a true causal effect of smoking → methylation at the *SPEG* locus, methylation in response to smoking appears to show a causal effect on OPC mortality.

Additional developments are currently underway in the application of Mendelian randomization to understand the causal factors in disease progression including in the context of cancer [256]. This presents some additional methodological challenges and is limited at the current time by the paucity of GWAS that have looked specifically at disease survival or outcomes post diagnosis. However, this is an interesting and active area of research and may be a fruitful avenue to explore going forward.

Thirdly, this thesis was able to address a key gap in observational literature by employing causal inference methods (namely two-sample and two-step MR) associated with genetic and epigenetic data. As mentioned previously, a pitfall of observational epidemiology is a lack of ability to control for unmeasured confounding. In Chapters 1 and 3, it was uncertain whether effect estimates

for smoking, alcohol and HPV infection/oral sex were mediating a "risk-taking" phenotype. Co-adjusting for smoking, alcohol and HPV infection/oral sex in observational analyses would likely be insufficient if this was the case, as the breadth of impact "risk-taking" would have on an individual's life beyond these 3 phenotypes cannot be completely accounted for. Therefore, the observational effect of smoking, alcohol and HPV infection, respectively, could be biased by any of the other two risk factors, or by any other commonality between "risk-takers" (e.g. low SEP, high BMI, poor dental hygiene). Using genetic data to appraise smoking and alcohol consumption allowed for an independent effect of these respective factors, free of traditional confounding.

Limitations of work presented

The use of an approach based on text mining of published literature to derive novel risk factors for OPC only clarified the existence of risk factors which are already established for this disease in epidemiological literature, therefore this approach inevitably suffers from publication bias. It appears that a concentration of articles investigating alcohol, smoking and HPV against OPC risk, and a paucity of literature investigating OPC as a specific HNC sub-type (rather than grouping it with other sub-types) severely limited MELODI's ability to infer novel links between publications. This is evident in the apparent mediation of smoking, alcohol and HPV infection by each other, with no other suggestions of modifiable risk factors for OPC. To this end, MELODI worked in its application in this thesis much more as a "review" of current literature than a "hypothesis generator", as intended. Nevertheless, using MELODI to retrieve potential intermediates retrieved hundreds of concepts. A "concept" in the context of MELODI refers to any of: an amino acid, peptide, protein, bacteria, biologically active substance, carbohydrate, fungus, gene, hormone, lipid, neuroreactive substance or biogenic amine, pharmacologic substance or virus, as per the criteria outlined in Table 3.1 of Chapter 3.

A key limitation of this thesis relates to epidemiological study design in general; behavioural phenotypes (e.g. smoking and alcohol consumption) are typically derived via self-report questionnaire. Accordingly, interpretation of observational results and MR results necessitates caution. Various factors affect how accurately a study can determine behavioural phenotypes from self-report questionnaires, including intentional patient misreporting (e.g. a patient defining themselves as an ex-smoker when they are actually a current smoker), recall bias (e.g. an OPC patient over-reporting historic units of alcohol consumed per week due to a belief that this behaviour caused their cancer) and a general inaccuracy in questionnaire design for the measurement of behaviours (e.g. using "current" smokers as a category doesn't account for people in this category smoking vastly

different numbers of cigarettes per day, and reporting alcohol consumed in units per week doesn't necessarily account for binge-drinking).

These limitations can also extend to some extent to the application of MR because these biased measures are the basis of GWAS conducted to generate SNP-trait associations. Poor phenotype measurements can potentially compromise the quality of GWAS outputs. Furthermore, when proxying a phenotype for MR using GWAS results, knowledge of both how the phenotype was measured and the biological plausibility SNPs used is essential. For example, when selecting SNPs for a smoking IV (e.g. cigarettes per day), some or all of these may reside within the *CHRNA3* gene region. However, SNPs in this region are reported to proxy for smoking heaviness *amongst smokers* rather than being representative of cigarettes per day in a general population [375, 376]. As such, the outcome OPC GWAS data would have to be restricted to current smokers to produce a meaningful effect-estimate, which is not possible using summary (rather than individual-level) genetic data. Some of these limitations are overcome by utilising well-established and validated genetic instruments for application in MR i.e. those that have been identified in a reliably phenotyped, representative population that is independent of the OPC case series in which the MR analysis is being applied.

A limitation of the epigenetic results chapters presented in this thesis is the lack of availability of tumour tissue-based methylation data, thus analyses being conducted in whole blood. Whole blood methylation has been established as a "biosocial archive", changing robustly in response to long-term behaviours and exposures. However, use of blood rather than tumour tissue prevents identification of aberrant expression at the cancer site itself and is prone to confounding by cell composition. It has been used here to more precisely define risk or prognostic *factors* in prediction models, including their effect on mortality, rather than diagnosis or prognosis of OPC specifically. Substantial changes in methylation can be seen across a variety of tumour tissues, including those of OPC. Having access to tumour tissue methylation and the specificity it affords may prove particularly lucrative in discovering predictive diagnostic or prognostic methylation patterns to better improve clinical outcomes. However, in the absence of the availability of these data, robust relationships of blood-based methylation patterns with prognostic factors and OPC survival have been established in Chapters 5 and 6, creating novel avenues and gene regions to be investigated and appraised with respect to the aetiology of this cancer.

7.1.3. Future directions

'Omic' MR analyses

One of the key strengths of this thesis was the ability to interrogate the causality of a vast number of phenotypes, many of which have not previously been appraised or hypothesised to be associated with OPC. The advancement of GWAS studies to discover SNPs associated with gene expression levels, methylation levels, metabolite levels and protein levels, in addition to novel phenotypes, allows for an exponentially greater number of hypotheses to be tested and for vastly improved insight into disease aetiology. It also allows for many of the intermediates identified by MELODI in Chapter 3 to be investigated. Despite molecular intermediates not typically possessing a large number of SNPs from GWAS to proxy them, precise measurement of these phenotypes (e.g. measurement of metabolites using nuclear magnetic resonance and liquid chromatography-mass spectronomy) result in relatively large GWAS per-allele effect sizes and phenotypic variance explained, both of which improve power to detect a true causal association. These "omic" associations may prove invaluable in the progression of OPC research, aiding the discovery and appraisal of important biological pathways with the ability to translate to biomarkers and therapeutic targets with clinical importance.

MR of HPV status on OPC risk

At the time the analyses were conducted no data were available that allowed MR to be applied to explore the causal role between HPV status and OPC. GWAS of HPV status do exist [377] but do not provide beta values that are suitable for use in MR analysis. The identification of genetic instruments to proxy HPV status will be useful to probe this question but these are likely to be challenging to identify and be free from pleiotropic action, given what is known about the biological complexity of infection and immunity.

Epigenetic signatures in saliva

In Chapter 5, blood-based epigenetic data from individuals from HN5000 was used to determine differentially methylated sites and regions associated with different risk factors. Chapter 6 investigated the utility of methylation as a predictor of OPC mortality in the context of commonly reported risk factors. Methylation appeared to be associated with both alcohol and smoking status, with methylation at some novel gene regions measured by the MethylationEPIC BeadChip highlighted for these risk factors. Methylation also appeared to predict risk factor-specific OPC mortality with the

same efficacy as directly-measured phenotypes, showing promise as an objective archive of health behaviours; methylation is able to predict the exposure history of an individual with OPC independent of direct phenotype measurement. Although the use of minimally invasive samples as a source of DNA are commonly used (e.g. blood or saliva), there is logic in aiming to combine this with sources of DNA from the affected tissue. Indeed, many investigators have sought to identify DNA methylation signatures in circulating tumour DNA [373]. This is a promising avenue of exploration given the very profound changes to the methylome seen in tumour tissue (even if only tiny trace amounts can be isolated) when compared to the more subtle shifts in methylation patterns seen in somatic tissues.

Saliva is present as a resource in HN5000, which may plausibly contain small amounts of shed tumour DNA; a potentially lucrative resource for early detection of cancer. DNA extraction from saliva samples (usually a mix of buccal epithelial cells and lymphocytes) is relatively routine. It would be of interest to compare DNA methylation data sources (blood vs saliva) to evaluate their respective performance in OPC prognosis. Saliva-based methylation translates to a less-invasive clinical test if notable methylation pattern differences in this tissue type occur between people with OPC, exposed vs unexposed to a risk factor.

MR of risk factors for OPC progression

As mentioned above, methodological developments in the application of Mendelian randomization will in the future allow the better discrimination of causal factors associated with disease progression as opposed to those that cause disease to occur in the first instance. This has important implications for both treatment (causal factors affecting prognosis should be the targets of any therapy) and for secondary and tertiary prevention. For example, understanding whether smoking cessation impacts OPC prognosis or survival could be an important issue for those diagnosed with this cancer; if smoking does not causally impact survival after OPC has been diagnosed then it would not be a priority to stop smoking. It is plausible that a very different suite of factors influence disease (OPC) progression compared to those that influence disease onset so it would be advantageous to recapitulate the MR-PheWAS approach that I adopted in this thesis to investigate causal factors for OPC progression or survival once the requisite data are available.

Case-control EWAS of OPC

The EWAS analysis undertaken as part of this thesis focused on an analysis of the relationship between established risk factors and DNA methylation variable loci. An EWAS of OPC was not

undertaken as this would require a case-control study design and the H&N5000 study is a case only cohort. DNA methylation data generation is highly prone to batch effects, it would therefore not be appropriate (as has been the case for GWAS) to use a separately sourced control population in an EWAS study design as any comparison would likely result in differentially methylated loci being confounded by batch. Future work is warranted to undertake a well-powered EWAS of OPC and this would require the recruitment (or identification) of an appropriately matched control group and contiguous data generation with both case and control samples randomised across the arrays analysed.

Epigenetic prediction

Possibly one or the most exciting areas of epigenetic epidemiology currently is the use of DNA methylation in prediction and prognosis. Although the use of DNA methylation as a tumour biomarker and its detection in circulating tumour DNA has been explored in various cancers [373], the potential of DNA methylation to harness information on lifelong exposure to a wide range of risk factors has not really been exploited. Work in this field has begun to apply machine learning approaches to maximise the informative component of multi-dimensional DNA methylation data [378] but there is far more that could be done to explore this further. This will be aided by the generation of more granular DNA methylation data on well phenotyped samples from patients with a detailed case history.

7.1.4. Conclusion

This body of work highlights the potential of genetic and epigenetic data to augment understanding of OPC aetiology. It has shown the potential of two-sample MR analyses using genetic data to establish causal associations between risk factors and OPC. Smoking, alcohol, educational attainment, sexual behaviour and weight were all seen to be causally associated with OPC, thereby corroborating the observational literature findings for these risk factors. Unfortunately, HPV could not be appraised in an MR framework, and so no causal association could be established. Epigenetic data highlighted methylation pattern differences at novel gene regions for smoking, alcohol and HPV status. Methylation associated with OPC mortality was also appraised, allowing deeper insight into biological pathways associated with OPC. Using 2-step MR, causality was established between smoking-associated methylation at the *SPEG* gene region and OPC mortality. No evidence was found for methylation between HPV or alcohol and OPC mortality, respectively. Finally, epigenetic data shows promise of clinical or translational utility as a predictor of phenotype. Overall, the workflow

and methodologies employed in this thesis have served to improve knowledge of the causal and biological landscape of OPC beyond that of observational literature. Advanced statistical and bioinformatic analyses have allowed for further evaluation with genetic and DNA methylation data to understand causality, investigate molecular pathways, and evaluate prediction.

References

1. Lowe, J.S. and P.G. Anderson, *Chapter 11 - Alimentary Tract*, in *Stevens & Lowe's Human Histology (Fourth Edition) (Fourth Edition)*, J.S. Lowe and P.G. Anderson, Editors. 2015, Mosby: Philadelphia. p. 186-224.
2. Fossum, C.C., et al., *Characterization of the oropharynx: anatomy, histology, immunology, squamous cell carcinoma and surgical resection.* Histopathology, 2017. **70**(7): p. 1021-1029.
3. Matsuo, K. and J.B. Palmer, *Anatomy and physiology of feeding and swallowing: normal and abnormal.* Physical medicine and rehabilitation clinics of North America, 2008. **19**(4): p. 691-vii.
4. Campagnolo, A.M., et al., *Laryngopharyngeal reflux: diagnosis, treatment, and latest research.* International archives of otorhinolaryngology, 2014. **18**(2): p. 184-191.
5. Wolford, R.W. and T.J. Schaefer, *Pharyngitis*, in *StatPearls*. 2019, StatPearls Publishing LLC.: Treasure Island FL.
6. Morrison, W.H., A.S. Garden, and K.K. Ang, *Chapter 11 - The Oropharynx*, in *Radiation Oncology (Ninth Edition)*, J.D. Cox and K.K. Ang, Editors. 2010, Content Repository Only!: Philadelphia. p. 224-249.
7. Goel, A.N., et al., *Minor Salivary Gland Carcinoma of the Oropharynx: A Population-Based Analysis of 1426 Patients.* Otolaryngol Head Neck Surg, 2018. **158**(2): p. 287-294.
8. Carpenter, D.H., S.K. El-Mofty, and J.S. Lewis, Jr., *Undifferentiated carcinoma of the oropharynx: a human papillomavirus-associated tumor with a favorable prognosis.* Mod Pathol, 2011. **24**(10): p. 1306-12.
9. Li, Y.X., et al., *Variable clinical presentations of nasal and Waldeyer ring natural killer/T-cell lymphoma.* Clin Cancer Res, 2009. **15**(8): p. 2905-12.
10. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.
11. Chaturvedi, A.K., et al., *Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers.* J Clin Oncol, 2013. **31**(36): p. 4550-9.
12. Schache, A.G., et al., *HPV-Related Oropharynx Cancer in the United Kingdom: An Evolution in the Understanding of Disease Etiology.* Cancer Res, 2016. **76**(22): p. 6598-6606.
13. Network, N.C.I., *Profile of Head and Neck Cancers in England: Incidence, Mortality and Survival. Oxford Cancer Intelligence Unit.* 2010.
14. Shaw, R. and N. Beasley, *Aetiology and risk factors for head and neck cancer: United Kingdom National Multidisciplinary Guidelines.* J Laryngol Otol, 2016. **130**(S2): p. S9-S12.
15. Lee, Y.C., et al., *Active and involuntary tobacco smoking and upper aerodigestive tract cancer risks in a multicenter case-control study.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(12): p. 3353-61.
16. Sadri, G. and H. Mahjub, *Tobacco smoking and oral cancer: a meta-analysis.* J Res Health Sci, 2007. **7**(1): p. 18-23.
17. Bagnardi, V., et al., *Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis.* Br J Cancer, 2015. **112**(3): p. 580-93.
18. Turati, F., et al., *A meta-analysis of alcohol drinking and oral and pharyngeal cancers. Part 2: results by subsites.* Oral Oncol, 2010. **46**(10): p. 720-6.
19. Hashibe, M., et al., *Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: Pooled analysis in the international head and neck cancer epidemiology consortium.* Journal of the National Cancer Institute, 2007. **99**(10): p. 777-789.

20. 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.* J Natl Cancer Inst, 2008. **100**(6): p. 407-20.
21. Lang Kuhs, K.A., et al., *Human papillomavirus 16 E6 antibodies are sensitive for human papillomavirus-driven oropharyngeal cancer and are associated with recurrence.* Cancer, 2017. **123**(22): p. 4382-4390.
22. Rampias, T., et al., *E6 and e7 gene silencing and transformed phenotype of human papillomavirus 16-positive oropharyngeal cancer cells.* J Natl Cancer Inst, 2009. **101**(6): p. 412-23.
23. D'Souza, G., et al., *Case-control study of human papillomavirus and oropharyngeal cancer.* N Engl J Med, 2007. **356**(19): p. 1944-56.
24. You, E.L., M. Henry, and A.G. Zeitouni, *Human papillomavirus-associated oropharyngeal cancer: review of current evidence and management.* Curr Oncol, 2019. **26**(2): p. 119-123.
25. 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, 2009. **18**(2): p. 541-50.
26. Anantharaman, D., et al., *Combined effects of smoking and HPV16 in oropharyngeal cancer.* Int J Epidemiol, 2016. **45**(3): p. 752-61.
27. Kumar, R., et al., *Alcohol and Tobacco Increases Risk of High Risk HPV Infection in Head and Neck Cancer Patients: Study from North-East Region of India.* PLoS One, 2015. **10**(10): p. e0140700.
28. Oh, H.Y., et al., *Synergistic effect of viral load and alcohol consumption on the risk of persistent high-risk human papillomavirus infection.* PLoS One, 2014. **9**(8): p. e104374.
29. Brown, K.F., et al., *The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015.* Br J Cancer, 2018. **118**(8): p. 1130-1141.
30. *World Development Indicators*, T.W. Bank, Editor. 2012, The World Bank: Washington, D.C.
31. Goldstein, B.Y., et al., *Alcohol consumption and cancers of the oral cavity and pharynx from 1988 to 2009: an update.* European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP), 2010. **19**(6): p. 431-465.
32. Boeing, H., *Alcohol and risk of cancer of the upper gastrointestinal tract: first analysis of the EPIC data.* Nutrition and lifestyle: Opportunities for cancer prevention, 2002: p. 151-154.
33. Boffetta, P. and L. Garfinkel, *Alcohol drinking and mortality among men enrolled in an American Cancer Society prospective study.* Epidemiology, 1990: p. 342-348.
34. Kjærheim, K., M. Gaard, and A. Andersen, *The role of alcohol, tobacco, and dietary factors in upper aerogastric tract cancers: a prospective study of 10,900 Norwegian men.* Cancer Causes & Control, 1998. **9**(1): p. 99-108.
35. Llewellyn, C.D., N.W. Johnson, and K.A. Warnakulasuriya, *Risk factors for oral cancer in newly diagnosed patients aged 45 years and younger: a case-control study in Southern England.* J Oral Pathol Med, 2004. **33**(9): p. 525-32.
36. Holmes, A.J. and K. Anderson, *Convergence in National Alcohol Consumption Patterns: New Global Indicators.* Journal of Wine Economics, 2017. **12**(2): p. 117-148.
37. Bower, J., *The Evolution of the UK Wine Market: From Niche to Mass-Market Appeal.* Beverages, 2018. **4**: p. 87.
38. 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, 2009. **169**(2): p. 132-42.
39. Johansen, D., et al., *Food buying habits of people who buy wine or beer: cross sectional study.* BMJ, 2006. **332**(7540): p. 519-22.
40. Tjonneland, A., et al., *Wine intake and diet in a random sample of 48763 Danish men and women.* Am J Clin Nutr, 1999. **69**(1): p. 49-54.

41. Klatsky, A.L., M.A. Armstrong, and H. Kipp, *Correlates of alcoholic beverage preference: traits of persons who choose wine, liquor or beer.* Br J Addict, 1990. **85**(10): p. 1279-89.
42. Turati, F., et al., *A meta-analysis of alcohol drinking and oral and pharyngeal cancers: results from subgroup analyses.* Alcohol Alcohol, 2013. **48**(1): p. 107-18.
43. Marron, M., et al., *Cessation of alcohol drinking, tobacco smoking and the reversal of head and neck cancer risk.* Int J Epidemiol, 2010. **39**(1): p. 182-96.
44. De Leon, J., et al., *Association between smoking and alcohol use in the general population: stable and unstable odds ratios across two years in two different countries.* Alcohol Alcohol, 2007. **42**(3): p. 252-7.
45. Joossens, L. and M. Raw, *The Tobacco Control Scale: a new scale to measure country activity.* Tob Control, 2006. **15**(3): p. 247-53.
46. Willemsen, M.C. and G.E. Nagelhout, *Country Differences and Changes in Focus of Scientific Tobacco Control Publications between 2000 and 2012 in Europe.* Eur Addict Res, 2016. **22**(1): p. 52-8.
47. IARC, *IARC Monographs on the evaluation of carcinogenic risk to humans.* 2004. **83**(Tobacco smoke and involuntary smoking).
48. Anantharaman, D., et al., *Population attributable risk of tobacco and alcohol for upper aerodigestive tract cancer.* Oral Oncol, 2011. **47**(8): p. 725-31.
49. 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, 2013. **178**(5): p. 679-90.
50. Wyss, A.B., et al., *Smokeless Tobacco Use and the Risk of Head and Neck Cancer: Pooled Analysis of US Studies in the INHANCE Consortium.* American journal of epidemiology, 2016. **184**(10): p. 703-716.
51. Organization, W.H., *WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship*. 2013: World Health Organization.
52. Agaku, I.T., et al., *Poly-tobacco use among adults in 44 countries during 2008–2012: evidence for an integrative and comprehensive approach in tobacco control.* Drug and alcohol dependence, 2014. **139**: p. 60-70.
53. Boffetta, P., et al., *Smokeless tobacco and cancer.* Lancet Oncol, 2008. **9**(7): p. 667-75.
54. Siddiqi, K., et al., *Global burden of disease due to smokeless tobacco consumption in adults: analysis of data from 113 countries.* BMC medicine, 2015. **13**: p. 194-194.
55. Martin-Hernan, F., et al., *Oral cancer, HPV infection and evidence of sexual transmission.* Med Oral Patol Oral Cir Bucal, 2013. **18**(3): p. e439-44.
56. Akram, N., et al., *Oncogenic Role of Tumor Viruses in Humans.* Viral Immunol, 2017. **30**(1): p. 20-27.
57. Wood, N.H., et al., *The pathobiology and mechanisms of infection of HPV.* SADJ, 2010. **65**(3): p. 124-6.
58. Ndiaye, C., et al., *HPV DNA, E6/E7 mRNA, and p16INK4a detection in head and neck cancers: a systematic review and meta-analysis.* Lancet Oncol, 2014. **15**(12): p. 1319-31.
59. Sturgis, E.M. and K.K. Ang, *The epidemic of HPV-associated oropharyngeal cancer is here: is it time to change our treatment paradigms?* J Natl Compr Canc Netw, 2011. **9**(6): p. 665-73.
60. de Martel, C., et al., *Worldwide burden of cancer attributable to HPV by site, country and HPV type.* Int J Cancer, 2017. **141**(4): p. 664-670.
61. Oh, J.E., et al., *Molecular genetic characterization of p53 mutated oropharyngeal squamous cell carcinoma cells transformed with human papillomavirus E6 and E7 oncogenes.* Int J Oncol, 2013. **43**(2): p. 383-93.
62. Herrero, R., et al., *Human papillomavirus and oral cancer: the International Agency for Research on Cancer multicenter study.* J Natl Cancer Inst, 2003. **95**(23): p. 1772-83.
63. Carter, J.J., et al., *The natural history of human papillomavirus type 16 capsid antibodies among a cohort of university women.* J Infect Dis, 1996. **174**(5): p. 927-36.

64. Carter, J.J., et al., *Comparison of human papillomavirus types 16, 18, and 6 capsid antibody responses following incident infection.* J Infect Dis, 2000. **181**(6): p. 1911-9.
65. Forman, D., et al., *Global burden of human papillomavirus and related diseases.* Vaccine, 2012. **30 Suppl 5**: p. F12-23.
66. Plummer, M., et al., *Global burden of cancers attributable to infections in 2012: a synthetic analysis.* Lancet Glob Health, 2016. **4**(9): p. e609-16.
67. Jit, M., Y.H. Choi, and W.J. Edmunds, *Economic evaluation of human papillomavirus vaccination in the United Kingdom.* BMJ, 2008. **337**: p. a769.
68. Papillomaviruses, H., *IARC monographs on the evaluation of carcinogenic risks to humans.*
69. van Monsjou, H.S., et al., *Oropharyngeal squamous cell carcinoma: a unique disease on the rise?* Oral Oncol, 2010. **46**(11): p. 780-5.
70. Hosmer, D.W. and S. Lemeshow, *Confidence interval estimation of interaction.* Epidemiology, 1992. **3**(5): p. 452-6.
71. Zhang, L., et al., *Variants of human papillomavirus type 16 predispose toward persistent infection.* Int J Clin Exp Pathol, 2015. **8**(7): p. 8453-9.
72. Clifford, G.M., et al., *Worldwide distribution of human papillomavirus types in cytologically normal women in the International Agency for Research on Cancer HPV prevalence surveys: a pooled analysis.* Lancet, 2005. **366**(9490): p. 991-8.
73. Dillner, J., *The serological response to papillomaviruses.* Semin Cancer Biol, 1999. **9**(6): p. 423-30.
74. Buck, C.B., P.M. Day, and B.L. Trus, *The papillomavirus major capsid protein L1.* Virology, 2013. **445**(1-2): p. 169-74.
75. Elrefaey, S., et al., *HPV in oropharyngeal cancer: the basics to know in clinical practice.* Acta Otorhinolaryngol Ital, 2014. **34**(5): p. 299-309.
76. Ang, K.K., et al., *Human papillomavirus and survival of patients with oropharyngeal cancer.* N Engl J Med, 2010. **363**(1): p. 24-35.
77. Warnakulasuriya, S., *Significant oral cancer risk associated with low socioeconomic status.* Evid Based Dent, 2009. **10**(1): p. 4-5.
78. Conway, D.I., et al., *Socioeconomic factors associated with risk of upper aerodigestive tract cancer in Europe.* Eur J Cancer, 2010. **46**(3): p. 588-98.
79. Sharpe, K.H., et al., *Association between socioeconomic factors and cancer risk: a population cohort study in Scotland (1991-2006).* PLoS One, 2014. **9**(2): p. e89513.
80. Nieto, A., et al., *Lifetime body mass index and risk of oral cavity and oropharyngeal cancer by smoking and drinking habits.* Br J Cancer, 2003. **89**(9): p. 1667-71.
81. Maasland, D.H., et al., *Body mass index and risk of subtypes of head-neck cancer: the Netherlands Cohort Study.* Sci Rep, 2015. **5**: p. 17744.
82. Gaudet, M.M., et al., *Body mass index and risk of head and neck cancer in a pooled analysis of case-control studies in the International Head and Neck Cancer Epidemiology (INHANCE) Consortium.* Int J Epidemiol, 2010. **39**(4): p. 1091-102.
83. Mehanna, H., et al., *Oropharyngeal cancer: United Kingdom National Multidisciplinary Guidelines.* J Laryngol Otol, 2016. **130**(S2): p. S90-S96.
84. Huang, S.H. and B. O'Sullivan, *Overview of the 8th Edition TNM Classification for Head and Neck Cancer.* Curr Treat Options Oncol, 2017. **18**(7): p. 40.
85. Vainshtein, J.M., et al., *Long-term quality of life after swallowing and salivary-sparing chemo-intensity modulated radiation therapy in survivors of human papillomavirus-related oropharyngeal cancer.* Int J Radiat Oncol Biol Phys, 2015. **91**(5): p. 925-33.
86. Oates, J., et al., *The effect of cancer stage and treatment modality on quality of life in oropharyngeal cancer.* Laryngoscope, 2014. **124**(1): p. 151-8.
87. Al-Mamgani, A., et al., *A prospective evaluation of patient-reported quality-of-life after (chemo)radiation for oropharyngeal cancer: which patients are at risk of significant quality-of-life deterioration?* Radiother Oncol, 2013. **106**(3): p. 359-63.

88. Ojo, B., et al., *A systematic review of head and neck cancer quality of life assessment instruments.* Oral Oncol, 2012. **48**(10): p. 923-937.
89. Amini, A. and S.D. Karam, *Concurrent chemotherapy in oropharyngeal cancer: Cisplatin wins.* Oncotarget, 2019. **10**(6): p. 624-625.
90. Liu, Y., et al., *Altered fractionation radiotherapy with or without chemotherapy in the treatment of head and neck cancer: a network meta-analysis.* Onco Targets Ther, 2018. **11**: p. 5465-5483.
91. Bird, T., et al., *Outcomes of intensity-modulated radiotherapy as primary treatment for oropharyngeal squamous cell carcinoma - a European singleinstitution analysis.* Clin Otolaryngol, 2017. **42**(1): p. 115-122.
92. Teguh, D.N., et al., *Quality of life of oropharyngeal cancer patients treated with brachytherapy.* Curr Oncol Rep, 2009. **11**(2): p. 143-50.
93. Roets, E., et al., *Quality of life in oropharyngeal cancer: a structured review of the literature.* Support Care Cancer, 2018. **26**(8): p. 2511-2518.
94. Gershenwald, J., et al., *AJCC cancer staging manual.* Switzerland: Springer, 2017: p. 563-89.
95. Albers, A.E., et al., *Meta analysis: HPV and p16 pattern determines survival in patients with HNSCC and identifies potential new biologic subtype.* Sci Rep, 2017. **7**(1): p. 16715.
96. Leemans, C.R., P.J.F. Snijders, and R.H. Brakenhoff, *The molecular landscape of head and neck cancer.* Nat Rev Cancer, 2018. **18**(5): p. 269-282.
97. Mehanna, H., et al., *Radiotherapy plus cisplatin or cetuximab in low-risk human papillomavirus-positive oropharyngeal cancer (De-ESCALaTE HPV): an open-label randomised controlled phase 3 trial.* Lancet, 2019. **393**(10166): p. 51-60.
98. Gillison, M.L., et al., *Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): a randomised, multicentre, non-inferiority trial.* Lancet, 2019. **393**(10166): p. 40-50.
99. Keck, M.K., et al., *Integrative analysis of head and neck cancer identifies two biologically distinct HPV and three non-HPV subtypes.* Clin Cancer Res, 2015. **21**(4): p. 870-81.
100. Zhang, Y., et al., *Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures.* Clin Cancer Res, 2016. **22**(18): p. 4735-45.
101. Williams, V.M., et al., *Human papillomavirus type 16 E6* induces oxidative stress and DNA damage.* J Virol, 2014. **88**(12): p. 6751-61.
102. Lefevre, M., et al., *Epithelial to mesenchymal transition and HPV infection in squamous cell oropharyngeal carcinomas: the papillophar study.* Br J Cancer, 2017. **116**(3): p. 362-369.
103. Klussmann, J.P., et al., *Genetic signatures of HPV-related and unrelated oropharyngeal carcinoma and their prognostic implications.* Clin Cancer Res, 2009. **15**(5): p. 1779-86.
104. King, E.V., C.H. Ottensmeier, and G.J. Thomas, *The immune response in HPV(+) oropharyngeal cancer.* Oncoimmunology, 2014. **3**(1): p. e27254.
105. Ward, M.J., et al., *Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer.* Br J Cancer, 2014. **110**(2): p. 489-500.
106. Stephen, J.K., et al., *Significance of p16 in Site-specific HPV Positive and HPV Negative Head and Neck Squamous Cell Carcinoma.* Cancer Clin Oncol, 2013. **2**(1): p. 51-61.
107. Kumar, B., et al., *EGFR, p16, HPV Titer, Bcl-xL and p53, sex, and smoking as indicators of response to therapy and survival in oropharyngeal cancer.* J Clin Oncol, 2008. **26**(19): p. 3128-37.
108. Adams, J.M. and S. Cory, *The Bcl-2 protein family: arbiters of cell survival.* Science, 1998. **281**(5381): p. 1322-6.
109. Bauer, J.A., et al., *Reversal of cisplatin resistance with a BH3 mimetic, (-)-gossypol, in head and neck cancer cells: role of wild-type p53 and Bcl-xL.* Mol Cancer Ther, 2005. **4**(7): p. 1096-104.
110. Pikor, L., et al., *The detection and implication of genome instability in cancer.* Cancer Metastasis Rev, 2013. **32**(3-4): p. 341-52.

111. Tanaka, K. and T. Hirota, *Chromosomal instability: A common feature and a therapeutic target of cancer.* Biochim Biophys Acta, 2016. **1866**(1): p. 64-75.
112. Vargas-Rondon, N., V.E. Villegas, and M. Rondon-Lagos, *The Role of Chromosomal Instability in Cancer and Therapeutic Responses.* Cancers (Basel), 2017. **10**(1).
113. Sato, H., et al., *Prognostic utility of chromosomal instability detected by fluorescence in situ hybridization in fine-needle aspirates from oral squamous cell carcinomas.* BMC Cancer, 2010. **10**: p. 182.
114. Kawamura, K., et al., *Centrosome hyperamplification and chromosomal instability in bladder cancer.* Eur Urol, 2003. **43**(5): p. 505-15.
115. Villepelet, A., et al., *Effects of tobacco abuse on major chromosomal instability in human papilloma virus 16-positive oropharyngeal squamous cell carcinoma.* Int J Oncol, 2019. **55**(2): p. 527-535.
116. Zagradisnik, B., et al., *Identification of genomic copy number variations associated with specific clinical features of head and neck cancer.* Mol Cytogenet, 2018. **11**: p. 5.
117. Leemans, C.R., et al., *Recurrence at the primary site in head and neck cancer and the significance of neck lymph node metastases as a prognostic factor.* Cancer, 1994. **73**(1): p. 187-90.
118. Suda, T., et al., *Copy number amplification of the PIK3CA gene is associated with poor prognosis in non-lymph node metastatic head and neck squamous cell carcinoma.* BMC cancer, 2012. **12**: p. 416-416.
119. Resteghini, C., et al., *Prognostic role of PIK3CA and TP53 in human papillomavirus-negative oropharyngeal cancers.* Tumori, 2018. **104**(3): p. 213-220.
120. Sharma, S., T.K. Kelly, and P.A. Jones, *Epigenetics in cancer.* Carcinogenesis, 2010. **31**(1): p. 27-36.
121. Idelevich, A., et al., *Neuronal hypothalamic regulation of body metabolism and bone density is galanin dependent.* The Journal of clinical investigation, 2018. **128**(6): p. 2626-2641.
122. Misawa, K., et al., *Galanin Has Tumor Suppressor Activity and Is Frequently Inactivated by Aberrant Promoter Methylation in Head and Neck Cancer.* Translational Oncology, 2013. **6**(3): p. 338-346.
123. Misawa, K., et al., *Site-specific methylation patterns of the GAL and GALR1/2 genes in head and neck cancer: Potential utility as biomarkers for prognosis.* Mol Carcinog, 2017. **56**(3): p. 1107-1116.
124. Millard, L.A., et al., *MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization.* Sci Rep, 2015. **5**: p. 16645.
125. Burgess, S., et al., *Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors.* Eur J Epidemiol, 2015. **30**(7): p. 543-52.
126. Relton, C.L. and G. Davey Smith, *Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease.* Int J Epidemiol, 2012. **41**(1): p. 161-76.
127. Lau, F.Y.Y. and C. Kuziemsky, *Handbook of EHealth Evaluation: An Evidence-Based Approach*. 2016: University of Victoria.
128. Paré, G., et al., *Synthesizing information systems knowledge: A typology of literature reviews.* Information & Management, 2015. **52**(2): p. 183-199.
129. Skogen, J.C. and S. Overland, *The fetal origins of adult disease: a narrative review of the epidemiological literature.* JRSM Short Rep, 2012. **3**(8): p. 59.
130. Harborne, L., et al., *Descriptive review of the evidence for the use of metformin in polycystic ovary syndrome.* Lancet, 2003. **361**(9372): p. 1894-901.
131. Glover Takahashi, S., M. Nayer, and L.M.M. St Amant, *Epidemiology of competence: a scoping review to understand the risks and supports to competence of four health professions.* BMJ Open, 2017. **7**(9): p. e014823.

132. Zhou, Y., et al., *Epidemiology of spider mite sensitivity: a meta-analysis and systematic review.* Clin Transl Allergy, 2018. **8**: p. 21.
133. Beasant, L., et al., *Treatment preference and recruitment to pediatric RCTs: A systematic review.* Contemp Clin Trials Commun, 2019. **14**: p. 100335.
134. Veronese, N., et al., *Is chocolate consumption associated with health outcomes? An umbrella review of systematic reviews and meta-analyses.* Clin Nutr, 2019. **38**(3): p. 1101-1108.
135. Grover, S., F. Del Greco M, and I.R. König, *Evaluating the current state of Mendelian randomization studies: a protocol for a systematic review on methodological and clinical aspects using neurodegenerative disorders as outcome.* Systematic Reviews, 2018. **7**(1): p. 145.
136. Mogre, V., et al., *A realist review of educational interventions to improve the delivery of nutrition care by doctors and future doctors.* Systematic reviews, 2014. **3**: p. 148-148.
137. Chang, E.T., et al., *A critical review of the epidemiology of Agent Orange or 2,3,7,8-tetrachlorodibenzo-p-dioxin and lymphoid malignancies.* Annals of Epidemiology, 2015. **25**(4): p. 275-292.e30.
138. Elsworth, B., et al., *MELODI: Mining Enriched Literature Objects to Derive Intermediates.* Int J Epidemiol, 2018.
139. Cairelli, M.J., et al., *Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox.* AMIA Annu Symp Proc, 2013. **2013**: p. 164-73.
140. Kilicoglu, H., et al., *SemMedDB: a PubMed-scale repository of biomedical semantic predications.* Bioinformatics, 2012. **28**(23): p. 3158-60.
141. Lindberg, D.A., B.L. Humphreys, and A.T. McCray, *The Unified Medical Language System.* Methods Inf Med, 1993. **32**(4): p. 281-91.
142. Humphreys, B.L. and D.A. Lindberg, *The UMLS project: making the conceptual connection between users and the information they need.* Bull Med Libr Assoc, 1993. **81**(2): p. 170-7.
143. Fisher, R.A., *Statistical methods for research workers*. 2006: Genesis Publishing Pvt Ltd.
144. Halushka, M.K., et al., *Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.* Nat Genet, 1999. **22**(3): p. 239-47.
145. Hallfors, J., et al., *Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway.* Addict Biol, 2019. **24**(3): p. 549-561.
146. Strawbridge, R.J., et al., *Genome-wide analysis of self-reported risk-taking behaviour and cross-disorder genetic correlations in the UK Biobank cohort.* Transl Psychiatry, 2018. **8**(1): p. 39.
147. Jonnalagadda, M., et al., *A Genome-Wide Association Study of Skin and Iris Pigmentation among Individuals of South Asian Ancestry.* Genome Biol Evol, 2019. **11**(4): p. 1066-1076.
148. Tian, C., et al., *Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections.* Nat Commun, 2017. **8**(1): p. 599.
149. Lesseur, C., et al., *Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer.* Nat Genet, 2016. **48**(12): p. 1544-1550.
150. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-D1012.
151. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
152. Staley, J.R., et al., *A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design.* European Journal of Human Genetics, 2017. **25**(7): p. 854-862.
153. Weisstein, E.W., *Bernoulli distribution.* sigma, 2002. **19**: p. 20.

154. MacKenzie, D.I., et al., *Chapter 3 - Fundamental Principals of Statistical Inference*, in *Occupancy Estimation and Modeling (Second Edition)*, D.I. MacKenzie, et al., Editors. 2018, Academic Press: Boston. p. 71-111.
155. Sul, J.H., L.S. Martin, and E. Eskin, *Population structure in genetic studies: Confounding factors and mixed models.* PLoS Genet, 2018. **14**(12): p. e1007309.
156. Lander, E.S. and N.J. Schork, *Genetic dissection of complex traits.* Science, 1994. **265**(5181): p. 2037-48.
157. Vilhjalmsson, B.J. and M. Nordborg, *The nature of confounding in genome-wide association studies.* Nat Rev Genet, 2013. **14**(1): p. 1-2.
158. Haworth, S., et al., *Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis.* Nat Commun, 2019. **10**(1): p. 333.
159. Yu, J., et al., *A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.* Nat Genet, 2006. **38**(2): p. 203-8.
160. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
161. Price, A.L., et al., *New approaches to population stratification in genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 459-63.
162. Lever, J., M. Krzywinski, and N. Altman, *Principal component analysis.* Nature Methods, 2017. **14**(7): p. 641-642.
163. Vansteelandt, S., et al., *On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects.* Genet Epidemiol, 2009. **33**(5): p. 394-405.
164. Holmes, M.V. and G. Davey Smith, *Problems in interpreting and using GWAS of conditional phenotypes illustrated by 'alcohol GWAS'.* Molecular psychiatry, 2019. **24**(2): p. 167-168.
165. Aschard, H., et al., *Adjusting for heritable covariates can bias effect estimates in genome-wide association studies.* American journal of human genetics, 2015. **96**(2): p. 329-339.
166. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genet Epidemiol, 2008. **32**(3): p. 227-34.
167. Marees, A.T., et al., *A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.* Int J Methods Psychiatr Res, 2018. **27**(2): p. e1608.
168. Bland, J.M. and D.G. Altman, *Multiple significance tests: the Bonferroni method.* BMJ, 1995. **310**(6973): p. 170.
169. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
170. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
171. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
172. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.
173. Scheet, P. and M. Stephens, *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.* Am J Hum Genet, 2006. **78**(4): p. 629-44.
174. Servin, B. and M. Stephens, *Imputation-based analysis of association studies: candidate regions and quantitative traits.* PLoS Genet, 2007. **3**(7): p. e114.
175. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
176. Nicolae, D.L., *Testing untyped alleles (TUNA)-applications to genome-wide association studies.* Genet Epidemiol, 2006. **30**(8): p. 718-27.
177. Zaitlen, N., et al., *Leveraging the HapMap correlation structure in association studies.* Am J Hum Genet, 2007. **80**(4): p. 683-91.

178. Browning, S.R., *Multilocus association mapping using variable-length Markov chains.* Am J Hum Genet, 2006. **78**(6): p. 903-13.
179. Li, Y., et al., *Genotype imputation.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 387-406.
180. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.
181. Imboden, M., et al., *Epigenome-wide association study of lung function level and its change.* Eur Respir J, 2019. **54**(1).
182. Joehanes, R., et al., *Epigenetic Signatures of Cigarette Smoking.* Circ Cardiovasc Genet, 2016. **9**(5): p. 436-447.
183. Eckstein, M., M. Rea, and Y.N. Fondufe-Mittendorf, *Transient and permanent changes in DNA methylation patterns in inorganic arsenic-mediated epithelial-to-mesenchymal transition.* Toxicol Appl Pharmacol, 2017. **331**: p. 6-17.
184. Relton, C.L., F.P. Hartwig, and G. Davey Smith, *From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive.* Int J Epidemiol, 2015. **44**(4): p. 1083-93.
185. Hotta, K., et al., *Identification of differentially methylated region (DMR) networks associated with progression of nonalcoholic fatty liver disease.* Sci Rep, 2018. **8**(1): p. 13567.
186. Suderman, M., et al., *dmrff: identifying differentially methylated regions efficiently with power and control.* 2018: p. 508556.
187. Iwasiow, R., A. Desbois, and H. Birnboim, *Long-term stability of DNA from saliva samples stored in the Oragene self-collection kit.* DNA Genotek, 2011.
188. Bulla, A., et al., *Blood DNA Yield but Not Integrity or Methylation Is Impacted After Long-Term Storage.* Biopreserv Biobank, 2016. **14**(1): p. 29-38.
189. Kokkat, T.J., et al., *Archived formalin-fixed paraffin-embedded (FFPE) blocks: A valuable underexploited resource for extraction of DNA, RNA, and protein.* Biopreserv Biobank, 2013. **11**(2): p. 101-6.
190. Abraham, J.E., et al., *Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping.* BMC medical genomics, 2012. **5**: p. 19-19.
191. Sarnecka, A.K., et al., *DNA extraction from FFPE tissue samples - a comparison of three procedures.* Contemp Oncol (Pozn), 2019. **23**(1): p. 52-58.
192. Langie, S., et al., *Environmental programming of respiratory allergy in childhood: the applicability of saliva.* Allergy, 2015. **70**: p. 236.
193. Houseman, E.A., et al., *DNA Methylation in Whole Blood: Uses and Challenges.* Curr Environ Health Rep, 2015. **2**(2): p. 145-54.
194. Moran, S., et al., *Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the Infinium HumanMethylation450 Microarray.* Epigenetics, 2014. **9**(6): p. 829-33.
195. Langie, S.A.S., et al., *Salivary DNA Methylation Profiling: Aspects to Consider for Biomarker Identification.* Basic Clin Pharmacol Toxicol, 2017. **121 Suppl 3**: p. 93-101.
196. Huang, L.H., et al., *The effects of storage temperature and duration of blood samples on DNA and RNA qualities.* PLoS One, 2017. **12**(9): p. e0184692.
197. Patel, P.G., et al., *Reliability and performance of commercial RNA and DNA extraction kits for FFPE tissue cores.* PLoS One, 2017. **12**(6): p. e0179732.
198. Gaunt, T.R., et al., *Systematic identification of genetic influences on methylation across the human life course.* Genome Biol, 2016. **17**(1): p. 61.
199. Volkov, P., et al., *A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits.* PLoS One, 2016. **11**(6): p. e0157776.
200. Veyrieras, J.B., B. Goffinet, and A. Charcosset, *MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments.* BMC Bioinformatics, 2007. **8**: p. 49.

201. Richmond, R.C., et al., *DNA methylation as a marker for prenatal smoke exposure in adults.* Int J Epidemiol, 2018. **47**(4): p. 1120-1130.
202. Yousefi, P.D., et al., *Validation and characterization of a DNA methylation alcohol biomarker across the life course.* bioRxiv, 2019: p. 591404.
203. Davey Smith G, E.S., *"Mendelian randomisation": can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiology, 2003. **32**: p. 1-22.
204. Yarmolinsky, J., et al., *Circulating Selenium and Prostate Cancer Risk: A Mendelian Randomization Analysis.* J Natl Cancer Inst, 2018.
205. Davey Smith G, E.S., *Mendelian randomization: prospects, potentials, and limitations.* Int J Epi, 2004. **33**: p. 30-42.
206. Zheng, J., et al., *Recent Developments in Mendelian Randomization Studies.* Curr Epidemiol Rep, 2017. **4**(4): p. 330-345.
207. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
208. Langdon, R.J., et al., *A phenome-wide Mendelian randomization study of pancreatic cancer using summary genetic data.* Cancer Epidemiology Biomarkers & Prevention, 2019: p. cebp.0036.2019.
209. 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, 2014. **14**.
210. Min, J., et al., *Meffil: efficient normalisation and analysis of very large DNA methylation samples.* bioRxiv, 2017.
211. Ambatipudi, S., et al., *DNA methylation derived systemic inflammation indices are associated with head and neck cancer development and survival.* Oral Oncol, 2018. **85**: p. 87-94.
212. Triche, T.J., Jr., et al., *Low-level processing of Illumina Infinium DNA Methylation BeadArrays.* Nucleic acids research, 2013. **41**(7): p. e90-e90.
213. Fortin, J.P., T.J. Triche, Jr., and K.D. Hansen, *Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi.* Bioinformatics, 2017. **33**(4): p. 558-560.
214. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC bioinformatics, 2012. **13**: p. 86-86.
215. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PLoS One, 2012. **7**(7): p. e41361.
216. Tozzi, V., et al., *Global, pathway and gene coverage of three Illumina arrays with respect to inflammatory and immune-related pathways.* Eur J Hum Genet, 2019. **27**(11): p. 1716-1723.
217. 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, 2018. **143**(5): p. 1114-1127.
218. Buckley, L., et al., *HPV-related Oropharyngeal Carcinoma: A Review of Clinical and Pathologic Features With Emphasis on Updates in Clinical and Pathologic Staging.* Adv Anat Pathol, 2018. **25**(3): p. 180-188.
219. Lubin, J.H., et al., *An examination of male and female odds ratios by BMI, cigarette smoking, and alcohol consumption for cancers of the oral cavity, pharynx, and larynx in pooled data from 15 case-control studies.* Cancer Causes Control, 2011. **22**(9): p. 1217-31.
220. Petersen, A.M., D. Rotolo, and L. Leydesdorff, *A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of Medical Subject Headings.* Research Policy, 2016. **45**(3): p. 666-681.
221. Humphreys, B.L., et al., *The Unified Medical Language System: an informatics research collaboration.* J Am Med Inform Assoc, 1998. **5**(1): p. 1-11.
222. Rindflesch, T.C. and M. Fiszman, *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.* J Biomed Inform, 2003. **36**(6): p. 462-77.

223. Simou, E., J. Britton, and J. Leonardi-Bee, *Alcohol and the risk of pneumonia: a systematic review and meta-analysis.* BMJ Open, 2018. **8**(8): p. e022344.
224. Aune, D., et al., *Tobacco smoking and the risk of sudden cardiac death: a systematic review and meta-analysis of prospective studies.* Eur J Epidemiol, 2018. **33**(6): p. 509-521.
225. Tam, S., et al., *The epidemiology of oral human papillomavirus infection in healthy populations: A systematic review and meta-analysis.* Oral Oncol, 2018. **82**: p. 91-99.
226. Marston, C. and E. King, *Factors that shape young people's sexual behaviour: a systematic review.* Lancet, 2006. **368**(9547): p. 1581-6.
227. Kreimer, A.R., et al., *Evaluation of human papillomavirus antibodies and risk of subsequent head and neck cancer.* J Clin Oncol, 2013. **31**(21): p. 2708-15.
228. Zhang, Y., et al., *Different levels in alcohol and tobacco consumption in head and neck cancer patients from 1957 to 2013.* PLoS One, 2015. **10**(4): p. e0124045.
229. Gandini, S., et al., *Tobacco smoking and cancer: a meta-analysis.* Int J Cancer, 2008. **122**(1): p. 155-64.
230. Li, S., et al., *Oral sex and risk of oral cancer: a meta-analysis of observational studies.* J Evid Based Med, 2015. **8**(3): p. 126-33.
231. Chancellor, J.A., S.J. Ioannides, and J.M. Elwood, *Oral and oropharyngeal cancer and the role of sexual behaviour: a systematic review.* Community Dent Oral Epidemiol, 2017. **45**(1): p. 20-34.
232. D'Souza, G., T.S. McNeel, and C. Fakhry, *Understanding personal risk of oropharyngeal cancer: risk-groups for oncogenic oral HPV infection and oropharyngeal cancer.* Ann Oncol, 2017. **28**(12): p. 3065-3069.
233. Majchrzak, E., et al., *Oral cavity and oropharyngeal squamous cell carcinoma in young adults: a review of the literature.* Radiol Oncol, 2014. **48**(1): p. 1-10.
234. Hearnden, V., et al., *Oral human papillomavirus infection in England and associated risk factors: a case-control study.* BMJ Open, 2018. **8**(8): p. e022497.
235. Clark, L., et al., *Striatal dopamine D(2)/D(3) receptor binding in pathological gambling is correlated with mood-related impulsivity.* Neuroimage, 2012. **63**(1): p. 40-6.
236. Weintraub, D., et al., *Association of dopamine agonist use with impulse control disorders in Parkinson disease.* Archives of neurology, 2006. **63**(7): p. 969-973.
237. Wing, V.C., et al., *Measuring cigarette smoking-induced cortical dopamine release: A [(1)(1)C]FLB-457 PET study.* Neuropsychopharmacology, 2015. **40**(6): p. 1417-27.
238. Ma, H. and G. Zhu, *The dopamine system and alcohol dependence.* Shanghai archives of psychiatry, 2014. **26**(2): p. 61-68.
239. MacKillop, J., et al., *Behavioral economic decision making and alcohol-related sexual risk behavior.* AIDS Behav, 2015. **19**(3): p. 450-8.
240. Hu, Z., et al., *Human papillomavirus 16 oncoprotein regulates the translocation of beta-catenin via the activation of epidermal growth factor receptor.* Cancer, 2015. **121**(2): p. 214-25.
241. Garcia-Pedrero, J.M., et al., *Prognostic significance of E-cadherin and beta-catenin expression in HPV-negative oropharyngeal squamous cell carcinomas.* Head Neck, 2017. **39**(11): p. 2293-2300.
242. Mercer, K.E., L. Hennings, and M.J. Ronis, *Alcohol consumption, Wnt/beta-catenin signaling, and hepatocarcinogenesis.* Adv Exp Med Biol, 2015. **815**: p. 185-95.
243. Santoro, A., et al., *Beta-catenin and epithelial tumors: a study based on 374 oropharyngeal cancers.* Biomed Res Int, 2014. **2014**: p. 948264.
244. Tandon, S., et al., *A systematic review of p53 as a prognostic factor of survival in squamous cell carcinoma of the four main anatomical subsites of the head and neck.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(2): p. 574-87.
245. DasGupta, T., et al., *Human papillomavirus oncogenic E6 protein regulates human beta-defensin 3 (hBD3) expression via the tumor suppressor protein p53.* Oncotarget, 2016. **7**(19): p. 27430-44.

246. Sass, V., et al., *Mode of action of human beta-defensin 3 against Staphylococcus aureus and transcriptional analysis of responses to defensin challenge.* Int J Med Microbiol, 2008. **298**(7-8): p. 619-33.
247. Spence, T., et al., *HPV Associated Head and Neck Cancer.* Cancers (Basel), 2016. **8**(8).
248. Elsevier, *Scopus*. 2019.
249. Franco, A., N. Malhotra, and G. Simonovits, *Social science. Publication bias in the social sciences: unlocking the file drawer.* Science, 2014. **345**(6203): p. 1502-5.
250. Davey Smith, G. and S. Ebrahim, *"Mendelian randomisation": can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiology, 2003. **32**: p. 1-22.
251. Davey Smith, G. and S. Ebrahim, *Mendelian randomization: prospects, potentials, and limitations.* Int J Epi, 2004. **33**: p. 30-42.
252. Lewis, S.J. and G. Davey Smith, *Alcohol, ALDH2, and Esophageal Cancer: A Meta-analysis Which Illustrates the Potentials and Limitations of a Mendelian Randomization Approach.* Cancer Epidemiology Biomarkers & Prevention, 2005. **14**(8): p. 1967-1971.
253. Lawlor, D., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**: p. 1133-1163.
254. Davey Smith, G. and G. Hemani, *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Human molecular genetics, 2014. **23**(R1): p. R89-98.
255. Pierce, B.L., P. Kraft, and C. Zhang, *Mendelian Randomization Studies of Cancer Risk: a Literature Review.* Current Epidemiology Reports, 2018: p. pp 1-13.
256. Yarmolinsky, J., et al., *Causal inference in cancer epidemiology: what is the role of Mendelian randomization?* bioRxiv, 2017.
257. Timpson, N.J., et al., *Does Greater Adiposity Increase Blood Pressure and Hypertension Risk?: Mendelian Randomization Using the FTO/MC4R Genotype.* Hypertension, 2009. **54**(1): p. 84-90.
258. Benn, M., et al., *Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study.* J Natl Cancer Inst, 2011. **103**(6): p. 508-19.
259. Theodoratou, E., et al., *Instrumental Variable Estimation of the Causal Effect of Plasma 25-Hydroxy-Vitamin D on Colorectal Cancer Risk: A Mendelian Randomization Analysis.* PLoS ONE, 2012. **7**(6): p. e37662.
260. Hägg, S., et al., *Adiposity as a cause of cardiovascular disease: a Mendelian randomization study.* International Journal of Epidemiology, 2015. **44**(2): p. 578-586.
261. Pei, Y., Y. Xu, and W. Niu, *Causal relevance of circulating adiponectin with cancer: a meta-analysis implementing Mendelian randomization.* Tumor Biology, 2015. **36**(2): p. 585-594.
262. Telomeres Mendelian Randomization, C., et al., *Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study.* JAMA Oncol, 2017. **3**(5): p. 636-651.
263. Millard, L.A.C., et al., *Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank.* Int J Epidemiol, 2017.
264. Das, S., et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-7.
265. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2011. **9**(2): p. 179-81.
266. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.
267. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
268. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.

269. Hemani, G., et al., *The MR-Base platform supports systematic causal inference across the human phenome.* Elife, 2018. **7**.
270. Shin, S.Y., et al., *An atlas of genetic influences on human blood metabolites.* Nat Genet, 2014. **46**(6): p. 543-50.
271. Kettunen, J., et al., *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet, 2012. **44**(3): p. 269-76.
272. Lemaitre, R.N., et al., *Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium.* PLoS Genet, 2011. **7**(7): p. e1002193.
273. Guan, W., et al., *Genome-wide association study of plasma N6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium.* Circ Cardiovasc Genet, 2014. **7**(3): p. 321-331.
274. Wu, J.H., et al., *Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.* Circ Cardiovasc Genet, 2013. **6**(2): p. 171-83.
275. Mozaffarian, D., et al., *Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium.* Am J Clin Nutr, 2015. **101**(2): p. 398-406.
276. Dastani, Z., et al., *Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals.* PLoS Genet, 2012. **8**(3): p. e1002607.
277. Paterson, A.D., et al., *A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose.* Diabetes, 2010. **59**(2): p. 539-49.
278. Kilpelainen, T.O., et al., *Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels.* Nat Commun, 2016. **7**: p. 10494.
279. Roederer, M., et al., *The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis.* Cell, 2015. **161**(2): p. 387-403.
280. Burgess, S. and J. Bowden, *Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods.* arXiv, 2015. **arXiv:1512.04486v1**.
281. International Consortium for Blood Pressure Genome-Wide Association, S., et al., *Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk.* Nature, 2011. **478**(7367): p. 103-9.
282. Wald, A., *Tests of statistical hypotheses concerning several parameters when the number of observations is large.* Trans. Am. Math. Soc., 1943. **54**: p. 426-482.
283. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
284. Kreimer, A.R., et al., *Diet and body mass, and oral and oropharyngeal squamous cell carcinomas: analysis from the IARC multinational case-control study.* Int J Cancer, 2006. **118**(9): p. 2293-7.
285. D'Avanzo, B., et al., *Anthropometric measures and risk of cancers of the upper digestive and respiratory tract.* Nutr Cancer, 1996. **26**(2): p. 219-27.
286. Jager-Wittenaar, H., et al., *Critical weight loss in head and neck cancer--prevalence and risk factors at diagnosis: an explorative study.* Support Care Cancer, 2007. **15**(9): p. 1045-50.
287. Lees, J., *Incidence of weight loss in head and neck cancer patients on commencing radiotherapy treatment at a regional oncology centre.* Eur J Cancer Care (Engl), 1999. **8**(3): p. 133-6.

288. van Bokhorst-de van der Schueren, M.A., et al., *Assessment of malnutrition parameters in head and neck cancer and their relation to postoperative complications.* Head Neck, 1997. **19**(5): p. 419-25.
289. Matthews, T.W., H.B. Lampe, and K. Dragosz, *Nutritional status in head and neck cancer patients.* J Otolaryngol, 1995. **24**(2): p. 87-91.
290. Plata-Salaman, C.R., *Central nervous system mechanisms contributing to the cachexia-anorexia syndrome.* Nutrition, 2000. **16**(10): p. 1009-12.
291. Riedel, F., et al., *Serum levels of interleukin-6 in patients with primary head and neck squamous cell carcinoma.* Anticancer Res, 2005. **25**(4): p. 2761-5.
292. Nakano, Y., et al., *Expression of tumor necrosis factor-alpha and interleukin-6 in oral squamous cell carcinoma.* Jpn J Cancer Res, 1999. **90**(8): p. 858-66.
293. Bierut, L.J., et al., *ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry.* Mol Psychiatry, 2012. **17**(4): p. 445-50.
294. Bonevski, B., et al., *Associations between alcohol, smoking, socioeconomic status and comorbidities: evidence from the 45 and Up Study.* Drug Alcohol Rev, 2014. **33**(2): p. 169-76.
295. Schwartz, S.M., et al., *Oral cancer risk in relation to sexual history and evidence of human papillomavirus infection.* J Natl Cancer Inst, 1998. **90**(21): p. 1626-36.
296. Dahlstrom, K.R., et al., *Differences in history of sexual behavior between patients with oropharyngeal squamous cell carcinoma and patients with squamous cell carcinoma at other head and neck sites.* Head & neck, 2011. **33**(6): p. 847-855.
297. Heck, J.E., et al., *Sexual behaviours and the risk of head and neck cancers: a pooled analysis in the International Head and Neck Cancer Epidemiology (INHANCE) consortium.* Int J Epidemiol, 2010. **39**(1): p. 166-81.
298. Cannon, T.L., et al., *Squamous cell carcinoma of the oral cavity in nonsmoking women: a new and unusual complication of chemotherapy for recurrent ovarian cancer?* Oncologist, 2012. **17**(12): p. 1541-6.
299. Pastorino, R., et al., *Genetic Contributions to The Association Between Adult Height and Head and Neck Cancer: A Mendelian Randomization Analysis.* Sci Rep, 2018. **8**(1): p. 4534.
300. Kachuri, L., et al., *Mendelian Randomization and mediation analysis of leukocyte telomere length and risk of lung and head and neck cancers.* Int J Epidemiol, 2018.
301. Toporcov, T.N., et al., *Risk factors for head and neck cancer in young adults: a pooled analysis in the INHANCE consortium.* Int J Epidemiol, 2015. **44**(1): p. 169-85.
302. Worsham, M.J., *Identifying the risk factors for late-stage head and neck cancer.* Expert Rev Anticancer Ther, 2011. **11**(9): p. 1321-5.
303. Reyes-Gibby, C.C., et al., *Survival patterns in squamous cell carcinoma of the head and neck: pain as an independent prognostic factor for survival.* J Pain, 2014. **15**(10): p. 1015-22.
304. Ragin, C.C. and E. Taioli, *Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: review and meta-analysis.* Int J Cancer, 2007. **121**(8): p. 1813-20.
305. Fakhry, C., et al., *Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial.* J Natl Cancer Inst, 2008. **100**(4): p. 261-9.
306. Liang, C., et al., *Biomarkers of HPV in head and neck squamous cell carcinoma.* Cancer Res, 2012. **72**(19): p. 5004-13.
307. Liu, C., et al., *A DNA methylation biomarker of alcohol consumption.* Mol Psychiatry, 2018. **23**(2): p. 422-433.
308. Prawdzic Senkowska, A., et al., *Impact of HPV infection on gene expression and methylation in oral cancer patients.* J Med Microbiol, 2019. **68**(3): p. 440-445.
309. 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, 2016. **41**(6): p. 804-809.

310. Waterboer, T., et al., *Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins.* Clin Chem, 2005. **51**(10): p. 1845-53.
311. Lang Kuhs, K.A., et al., *Human Papillomavirus 16 E6 Antibodies in Individuals without Diagnosed Cancer: A Pooled Analysis.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(4): p. 683-9.
312. Leek, J.T. and J.D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.* PLoS Genet, 2007. **3**(9): p. 1724-35.
313. Smith, B.H., et al., *Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness.* Int J Epidemiol, 2013. **42**(3): p. 689-700.
314. McCartney, D.L., et al., *Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease.* Alzheimers Dement (Amst), 2018. **10**: p. 429-437.
315. Syed, H., A.L. Jorgensen, and A.P. Morris, *SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes.* BMC Bioinformatics, 2017. **18**(1): p. 265.
316. Saffari, A., et al., *Estimation of a significance threshold for epigenome-wide association studies.* Genet Epidemiol, 2018. **42**(1): p. 20-33.
317. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.
318. Li, Y., et al., *The DNA methylome of human peripheral blood mononuclear cells.* PLoS Biol, 2010. **8**(11): p. e1000533.
319. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.* BMC Bioinformatics, 2010. **11**: p. 587.
320. Lehne, B., et al., *A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies.* Genome Biol, 2015. **16**: p. 37.
321. Robins, C., et al., *Testing Two Evolutionary Theories of Human Aging with DNA Methylation Data.* Genetics, 2017. **207**(4): p. 1547-1560.
322. Xia, C., et al., *Correction: Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation.* PLoS Genet, 2017. **13**(2): p. e1006608.
323. Zeng, Y., et al., *Shared Genetics and Couple-Associated Environment Are Major Contributors to the Risk of Both Clinical and Self-Declared Depression.* EBioMedicine, 2016. **14**: p. 161-167.
324. Xia, C., et al., *Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation.* PLoS Genet, 2016. **12**(2): p. e1005804.
325. Zaitlen, N., et al., *Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits.* PLoS Genet, 2013. **9**(5): p. e1003520.
326. Viechtbauer, W., *Conducting Meta-Analyses in R with the metafor Package.* 2010, 2010. **36**(3): p. 48 %J Journal of Statistical Software.
327. Machiela, M.J. and S.J. Chanock, *LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants.* Bioinformatics, 2015. **31**(21): p. 3555-7.
328. Burgess, S., F. Dudbridge, and S.G. Thompson, *Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods.* Stat Med, 2016. **35**(11): p. 1880-906.
329. Burgess, S. and S.G. Thompson, *Interpreting findings from Mendelian randomization using the MR-Egger method.* Eur J Epidemiol, 2017. **32**(5): p. 377-389.

330. Bowden, J., G.D. Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* International Journal of Epidemiology, 2015. **44**(2): p. 512-525.
331. Namani, A., et al., *Gene-expression signature regulated by the KEAP1-NRF2-CUL3 axis is associated with a poor prognosis in head and neck squamous cell cancer.* BMC Cancer, 2018. **18**(1): p. 46.
332. Ma, Z., et al., *SLC7A11, a component of cysteine/glutamate transporter, is a novel biomarker for the diagnosis and prognosis in laryngeal squamous cell carcinoma.* Oncol Rep, 2017. **38**(5): p. 3019-3029.
333. Lounglaithong, K., A. Bychkov, and P. Sampatanukul, *Aberrant promoter methylation of the PAQR3 gene is associated with prostate cancer.* Pathology - Research and Practice, 2018. **214**(1): p. 126-129.
334. Yu, X., et al., *PAQR3: a novel tumor suppressor gene.* Am J Cancer Res, 2015. **5**(9): p. 2562-8.
335. Bai, G., et al., *PAQR3 overexpression suppresses the aggressive phenotype of esophageal squamous cell carcinoma cells via inhibition of ERK signaling.* Biomed Pharmacother, 2017. **94**: p. 813-819.
336. Zhou, F., S. Wang, and J. Wang, *PAQR3 Inhibits the Proliferation and Tumorigenesis in Esophageal Cancer Cells.* Oncol Res, 2017. **25**(5): p. 663-671.
337. Starke, R.M., et al., *Cigarette smoke modulates vascular smooth muscle phenotype: implications for carotid and cerebrovascular disease.* PLoS One, 2013. **8**(8): p. e71954.
338. Quick, A.P., et al., *SPEG (Striated Muscle Preferentially Expressed Protein Kinase) Is Essential for Cardiac Function by Regulating Junctional Membrane Complex Activity.* Circ Res, 2017. **120**(1): p. 110-119.
339. Rose, B.S., et al., *Population-based study of competing mortality in head and neck cancer.* J Clin Oncol, 2011. **29**(26): p. 3503-9.
340. Okoye, C.C., et al., *Cardiovascular risk and prevention in patients with head and neck cancer treated with radiotherapy.* Head Neck, 2017. **39**(3): p. 527-532.
341. Wei, M., et al., *Cardiovascular disease risks among head and neck cancer survivors in a large, population-based cohort study.* 2018. **36**(15_suppl): p. 6051-6051.
342. Rieth, K.K.S., et al., *Prevalence of High-Risk Human Papillomavirus in Tonsil Tissue in Healthy Adults and Colocalization in Biofilm of Tonsillar Crypts.* JAMA Otolaryngol Head Neck Surg, 2018. **144**(3): p. 231-237.
343. Paternoster, L., K. Tilling, and G. Davey Smith, *Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges.* PLoS Genet, 2017. **13**(10): p. e1006944.
344. Philibert, R., et al., *A quantitative epigenetic approach for the assessment of cigarette consumption.* Front Psychol, 2015. **6**: p. 656.
345. Zhang, Y., et al., *Smoking-associated DNA methylation markers predict lung cancer incidence.* Clin Epigenetics, 2016. **8**: p. 127.
346. Bojesen, S.E., et al., *AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality.* Thorax, 2017. **72**(7): p. 646-653.
347. Guida, F., et al., *Lung cancer risk prediction using DNA methylation markers* Cancer Research, 2018.
348. McCartney, D.L., et al., *Epigenetic prediction of complex traits and death.* Genome Biol, 2018. **19**(1): p. 136.
349. Zhang, Y., et al., *Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality.* International Journal of Cancer, 2016. **139**(11): p. 2482-2492.
350. Zhang, Y., et al., *DNA methylation signatures in peripheral blood strongly predict all-cause mortality.* Nature Communications, 2017. **8**.

351. Sharp, L., et al., *Smoking at diagnosis is an independent prognostic factor for cancer-specific survival in head and neck cancer: findings from a large, population-based study.* Cancer Epidemiol Biomarkers Prev, 2014. **23**(11): p. 2579-90.
352. Duffy, S.A., et al., *Pretreatment health behaviors predict survival among patients with head and neck squamous cell carcinoma.* J Clin Oncol, 2009. **27**(12): p. 1969-75.
353. Hilgert, E., et al., *Tobacco abuse relates to significantly reduced survival of patients with oropharyngeal carcinomas.* Eur J Cancer Prev, 2009. **18**(2): p. 120-6.
354. Mayne, S.T., et al., *Alcohol and tobacco use prediagnosis and postdiagnosis, and survival in a cohort of patients with early stage cancers of the oral cavity, pharynx, and larynx.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(12): p. 3368-74.
355. Gama, R.R., et al., *Body mass index and prognosis in patients with head and neck cancer.* Head Neck, 2017.
356. Hollander, D., E. Kampman, and C.M. van Herpen, *Pretreatment body mass index and head and neck cancer outcome: A review of the literature.* Crit Rev Oncol Hematol, 2015. **96**(2): p. 328-38.
357. Choi, S.H., et al., *Socioeconomic and Other Demographic Disparities Predicting Survival among Head and Neck Cancer Patients.* PLoS One, 2016. **11**(3): p. e0149886.
358. Banos, D.T., et al., *Bayesian reassessment of the epigenetic architecture of complex traits.* 2018: p. 450288.
359. Zhang, Y., et al., *Self-reported smoking, serum cotinine, and blood DNA methylation.* Environ Res, 2016. **146**: p. 395-403.
360. Brierley JD, G.M., Wittekind C, *TNM Classification of Malignant Tumours* 2017, Union for International Cancer Control (UICC): Oxford, UK.
361. Schimansky, S., et al., *Association between comorbidity and survival in head and neck cancer: Results from Head and Neck 5000.* Head Neck, 2019. **41**(4): p. 1053-1062.
362. Sogaard, M., et al., *The impact of comorbidity on cancer survival: a review.* Clin Epidemiol, 2013. **5**(Suppl 1): p. 3-29.
363. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.
364. O'Sullivan, B., et al., *Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study.* Lancet Oncol, 2016. **17**(4): p. 440-451.
365. *Adult Comorbidity Evaluation-27*. 2018; Available from: https://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/adult_comorbidity_evaluation_-_27_de.asp?shownav=1.
366. Steinmetz-Wood, M., et al., *Do social characteristics influence smoking uptake and cessation during young adulthood?* Int J Public Health, 2018. **63**(1): p. 115-123.
367. Blok, D.J., et al., *The role of smoking in social networks on smoking cessation and relapse among adults: A longitudinal study.* Prev Med, 2017. **99**: p. 105-110.
368. Smith Sehdev, A.E. and G.M. Hutchins, *Problems with proper completion and accuracy of the cause-of-death statement.* Arch Intern Med, 2001. **161**(2): p. 277-84.
369. Richmond, R.C., et al., *Challenges and novel approaches for investigating molecular mediation.* Hum Mol Genet, 2016. **25**(R2): p. R149-R156.
370. Ventz, S., L. Trippa, and J.D. Schoenfeld, *Lessons Learned from De-escalation trials in favorable risk HPV-associated Squamous Cell Head and Neck Cancer - A Perspective on future trial designs.* Clinical Cancer Research, 2019: p. clincanres.0945.2019.
371. Horvath, S. and K. Raj, *DNA methylation-based biomarkers and the epigenetic clock theory of ageing.* Nat Rev Genet, 2018. **19**(6): p. 371-384.
372. Rosen, A.D., et al., *DNA methylation age is accelerated in alcohol dependence.* Transl Psychiatry, 2018. **8**(1): p. 182.

373. Berdasco, M. and M. Esteller, *Clinical epigenetics: seizing opportunities for translation.* Nat Rev Genet, 2019. **20**(2): p. 109-127.
374. Dahm, V., et al., *Cancer stage and pack-years, but not p16 or HPV, are relevant for survival in hypopharyngeal and laryngeal squamous cell carcinomas.* Eur Arch Otorhinolaryngol, 2018. **275**(7): p. 1837-1843.
375. Lassi, G., et al., *The CHRNA5-A3-B4 Gene Cluster and Smoking: From Discovery to Therapeutics.* Trends Neurosci, 2016. **39**(12): p. 851-861.
376. Ware, J.J., M.B. van den Bree, and M.R. Munafo, *Association of the CHRNA5-A3-B4 gene cluster with heaviness of smoking: a meta-analysis.* Nicotine Tob Res, 2011. **13**(12): p. 1167-75.
377. Chen, D., et al., *Genome-wide association study of HPV seropositivity.* Hum Mol Genet, 2011. **20**(23): p. 4714-23.
378. Bollepalli, S., et al., *EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data.* Epigenomics, 2019. **11**(13): p. 1469-1486.