Social, genetic and behavioural risk factors of Head and Neck Cancers - Elucidating causal pathways

Thekke Purakkal Akhil Soman

Faculty of Dentistry

McGill University

Montreal, Quebec, Canada

May 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

To my parents, Mrs. P. Prabha Devi and Dr. T.P. Somasundaran,

and to my teachers, for leading me from darkness to light.

Table of Contents

List of Tables

List of Figures

List of abbreviations and acronyms

HeNCe Life	The Head and Neck Cancer Life course study
SCCHN	Squamous cell carcinomas or head and neck
SEP	Socioeconomic position
CH	Childhood
EAH	Early adulthood
LAH	Late adulthood
XME	Xenobiotic metabolizing enzyme
CYP	Cytochrome P450
GST	Glutathione -*S* -transferase
ADH	Alcohol dehydrogenase
ALDH	Aldehyde dehydrogenase
SNP	Single nucleotide polymorphism
CNV	Copy number variation
HPV	Human papillomavirus
LD	Linkage disequilibrium
DAG	Directed acyclic graph
OR	Odds ratio
CI	Confidence interval
RR	Risk ratio
RERI	Relative excess risk due to interaction
IPW	Inverse probability weighting
MSM	Marginal structural model
wQIC	Weighted quasi likelihood criterion
2-way decomposition	Standard meditation or mediation
4- way decomposition	Four-way decomposition
CDE	Controlled direct effect
NDE	Natural direct effect
NIE	Natural indirect effect
TE	Total effect
INTref	Reference interaction
INTmed	Mediated interaction
PIE	Pure indirect effect
PM	Proportion mediated
PAI	Proportion attributable to interaction
PE	Proportion eliminated
DNA	Deoxyribonucleic acid
ICD	International Classification of Diseases
ASIR	Age standardised incidence rate
IARC	International Agency for Research on Cancer
PAH	Polycyclic aromatic hydrocarbons
PCR	Polymerase chain reaction
Chr	Chromosome

Acknowledgements

I consider my doctoral training as an exercise in persistence, one that I would not have got through without the help and support of many

First, I would like to thank the **study participants** at both the Indian and Canadian sites of the HeNCe life study. Their cooperation and patience during the long data collection process, laid the foundation for this dissertation.

Words will never be enough to thank my supervisor **Prof. Belinda Nicolau**, without whose sound guidance I would not have got through the 6 years of my doctoral training. I am indebted to her for giving me exposure to the life-course framework, and for the opportunity to work with the rich HeNCe Life case-control data set - two of the three pillars on which my dissertation is built. These, complemented with her unique enthusiasm for multidisciplinary research and collaborations - which she encouraged in her students as well - gave me enough flexibility and freedom to enrich my thesis with elements of complex systems and analytical methodologies. A true leader, her hunger for perfection and excellence, focus on the quality rather than quantity of research, and her student-for-life attitude, motivated the problem solver in me to see every hurdle as an opportunity, and the challenges and days of hard work as paths to future excellence. When going deep into specific challenges obscured the bigger picture of my thesis, she kept me objective and sane, and lent me strong and sometimes the only support to tide over both professional and personal challenges. She corrected my shortcomings at crucial times, always believed in me, allowed me to work independently and gave me immense care and attention through every stage of my doctoral studies, including funding for supporting my student life. My deep appreciation and respect to the genuine, down to earth and fantastic human that I have known her to be throughout the last 10 years of our collaboration.

I extend my sincere gratitude to my co-supervisor **Dr. Nicolas Schlecht** at Roswell Park Cancer Institute, for letting me play with the genetic data of the HeNCe life study, constructive feedback on my work, encouraging me to be specific on exposure definitions and statistical model specifications, and introducing me to the concept of causal mediation, which made me pursue the exciting world of analytical methodologies under the counterfactual causal inference framework - the critical third pillar of my thesis.

I would also like to thank **Prof. Robert Platt** at McGill University, and **Prof. Ashley Issac Naimi** at the University of Pittsburgh, for introducing me to the basics of counterfactual causal inference. Prof. Naimi's classes on inverse probability weighting and marginal structural models revitalized and provided great momentum to my thesis. I also acknowledge **Prof. Tyler J VanderWeele** at Harvard School of Public Health, for his teaching material and workshops, on causal mediation and interaction, as well as personal email communications and conference meetings which proved to be crucial for my work. I express my sincere gratitude to **Prof. Jay Kaufman** at McGill University for his guidance, valuable inputs on my manuscripts, and for encouraging me to pursue and implement the four-way decomposition technique. I thank him for being a major source of enthusiasm and inspiration, and for helping me fill a lot of gaps in my understanding of concepts in epidemiology and biostatistics.

I thank the members of my doctoral thesis committee, **Dr. Marie Claude Rousseau** at the INRS-Institute of Armand-Frappier, **Dr. Paul Allison** at McGill University, and **Dr. Gopalakrishnan Netuveli** at the University of East London, as well as our collaborators **Dr. Amanda Sacker** at the University College London and **Dr. Eduardo Franco** at McGill University for their valuable inputs at various stages of my thesis. Special thanks to Dr. Netuveli for providing greater insight into the concepts of life-course and socioeconomic position, and for helping me better define this exposure.

I am grateful to the research teams at both the Indian and Canadian sites of the HeNCe life study. Specifically, I would like to thank **Dr. Ipe Varghese** and **Dr. Shameena Shiraz** at Government Dental College, Calicut, Kerala, India for making the study at this site feasible, and **Dr. Priya Abraham** at Christian Medical College, Vellore, India for spearheading the HPV analysis at this site. I also thank **Dr. Shahul Hameed K.P** and **Dr. Sreenath Arekunnath Madathil** at McGill University for sharing the responsibilities of conduct and management of various aspects of research in India and data analysis, A debt of gratitude to a great friend and colleague, **Dr. Genevieve Castonguay** for all her advice and support which includes, but is not limited to, conduct of study at both sites, the cleaning, export and management of data, as well as proof reading and providing inputs on my thesis. I would also like to thank **Dr. François Coutlée** at the CHUM, Montreal, and **Dr. Robert Burk** at the Albert Einstein College of Medicine, for spearheading the HPV and genetic analysis respectively at the Canadian site. I extend my thanks to **Mr. Nicolas Drolet** and **Ms. Hilina Hitimana** for their assistance at various stages of the study at the Canadian site.

I am greatly indebted to the Canadian Institute of Health Research (CIHR), Ministère du Développement économique, de l'Innovation et de l'Exportation du Québec (MDEIE), Network for Oral and Bone Health Research (RSBO), Psychosocial Oncology Research Training (PORT), Canadian association of dental Research, and Network for Canadian Oral Health and Research (CADR- NCOHR), INRS-IAF and Faculty of Dentistry, McGill University, for providing me with scholarships, fellowships and awards to support my training.

I also acknowledge the Faculty of Dentistry, McGill University, and the Division of Oral health and Society for providing me with a cordial learning environment amongst a variety of oral health researchers. Special thanks to **Ms. Maria Palumbo**, the student's affairs coordinator at the Faculty - my "Italian mom" - for helping me with the administrative procedures related to my graduate program till date, and also to the library facilities at McGill University. A lot of appreciation to my fellow graduate students, friends and acquaintances who have come into my life, stayed or left during this period, with every interaction providing valuable lessons, and contributing to my professional and personal evolution during the past years.

My pursuit for knowledge and excellence amounts to one objective in life: to sharpen the tools I have procured along the life-course to help and give back to those in need. I consider this core driving force in me as a reflection of the sociopolitical contributions and visions of my **grandfathers** Mr. P. Appukuttan and Mr. T.P. Kuttikrishnan (late), and my **grandmothers** Mrs. P. Bharghavi (late), and Mrs. T.P. Gowri (late) (both of whom I lost during my PhD training period), which brought about revolutionary changes in their societies in Calicut, Kerala, India. I am greatly indebted to them as well as my relatives for instilling these elements of social commitment in me. My heartfelt gratitude to my loving **sister** Ms. Arsha Soman for her friendship, regular infusions of fun and laughter, and for shouldering my part of the duties and responsibilities towards my family in my absence, always assuring me with "big brother, you don't worry, just concentrate on your work there, we will take care of things here". I also thank two fantastic people, my **in-laws** Mrs. Beena Suresh and Mr. Raghavanachari Suresh, my second set of parents, for their enthusiastic and broad outlooks on life, encouragement and for being a major source of positive energy in the past 2 years. I also thank my **brothers-in-law** Mr. Raghunath Suresh and Mr. Arun M.V for their support.

The last 2 years were the toughest, critical and yet the most exciting phase of my PhD. But this period would not have been as smooth as it was without the great companionship and camaraderie of Ms. Gayatri Suresh, **my wife**. I thank her for her easy going and undemanding personality, understanding of the demands on me at this period of PhD, unique sense of humor that would make me break out in laughter even at the most tense situations, teaching of basic genetics and molecular analysis, proof reading certain sections of my dissertation, as well as for keeping our passion for both science and arts alive. I declare her as equal partner in this humble accomplishment.

Each step I have placed in my life towards and throughout my graduate training, would have never been possible without the accumulation of love, support and constant encouragement of **my parents** Mrs. P. Prabha Devi and Dr. T.P. Somasundaran. My mom kept me grounded, and rooted to my core driving force. My dad, has been a constant inspiration for every step towards acquiring knowledge and excellence. He exposed me to the world of research and PhD around 20 years back through the constant visits to his research lab during his doctoral studies at the Indian Institute of Technology, Madras, India. The philosophical as well as objective conversations with him, a professor and graduate student supervisor himself, always threw light on the big picture in life whenever I hit a wall during my PhD training. Words are never enough to thank my parents, and to them, I dedicate this PhD thesis.

Abstract

Introduction: Socioeconomic position (SEP), multiple variants in genes encoding Cytochrome P450 (CYP) and Glutathione S-transferase (GST) enzymes, and smoking and alcohol risk behaviours have been widely investigated in relation to the risk for squamous cell carcinomas of the head and neck (SCCHN). Although an accumulation of disadvantageous SEP over the life-course has been associated with increased risk for oral cancers, this association has not been explored through the lens of multiple life-course models within a single study. The evidence for the effect of multiple genetic variants in CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTP1 and GSTM1 genes on SCCHN is conflicting and their effect, independently or in interaction with smoking, has not been documented among the Canadian Caucasian population. Furthermore, smoking and alcohol use may interact as well as mediate the effect of genetic variants in CYP2A6 and ADH1B respectively on SCCHN. However, their effects, under combined mediation and interaction with the associated risk behaviours, have not been quantified yet. Addressing these gaps in knowledge has the potential to elucidate causal pathways underlying the relationship between these social, genetic and behavioural risk factors, and SCCHN.

Methods: The data for this dissertation was drawn from the Indian and Canadian sites of an international multi-centre hospital-based case-control study: The Head and Neck Cancer (HeNCe) Life study. Information on childhood (0-16 years), early (17-30 years) and late adulthood (above 30 years) SEP, and smoking and alcohol risk behaviours along the life span were collected using interviews and a life-grid technique. The estimation of SEP-oral cancer association used data on 350 incident oral cancer cases and 371 controls frequency-matched by age and sex recruited from two main hospitals in Kozhikode, Kerala, India between 2008 and 2012 SEP and associated confounders were considered as time-varying variables. The estimation of the total effect of genetic variants and their interactive effects with multiple levels of smoking pack-years used data on 389 incident SCCHN cases and 429 age and sex frequency-matched controls recruited from hospitals in Montreal, Canada. Sub-samples of this data on smokers, and alcohol consumers was used to estimate the causal pathways between genetic variants in CYP2A6, ADH1B and SCCHN defined by mediation and interaction with frequencies of smoking and alcohol use. Analytical techniques described under the counterfactual causal framework such as inverse-probability weighted marginal structural models, and mediation and four-way decomposition techniques were used to derive the causal effect estimates.

Results: Childhood and early adulthood SEP (advantageous vs. disadvantageous) were associated with oral cancer risk [(Odds Ratio (OR)=2.76, 95% Confidence interval (CI): 1.99, 3.81) and (OR=1.84, 95% CI: 1.21, 2.79), respectively]. In addition, participants who were in a disadvantageous (vs. advantageous) SEP during the three periods of life had an increased oral cancer risk [OR=4.86, 95% CI: 2.61, 9.06]. The childhood to early adulthood social mobility model and overall life-course trajectories indicated a strong influence of exposure to disadvantageous SEP in childhood on oral cancer risk.

Of all genetic variants analysed in the Canadian sample, carriers of GSTP1 105Val (vs non-carriers) were at 29% (OR=0.71, 95% CI: 0.53, 0.95) decreased risk of SCCHN. Stratum-specific analyses showed that carriers of this variant were at 41% (OR=0.59, 95% CI: 0.36, 0.95) and 51% (OR= 0.49, 95% CI: 0.24, 0.98) decreased risk for SCCHN relative to non-carriers, among the strata of heavy smokers and non-smokers respectively. There was no evidence for statistical interaction on an additive or multiplicative scale for any of the variants analysed. Among smokers, the total effect estimate of the CYP2A6 variant on SCCHN [Relative risk (RR)=1.28, 95% CI: 0.46, 3.59] was composed of a direct effect estimate of 1.22 (95% CI: 0.45, 3.33) and an indirect effect estimate through smoking of 1.05 (95% CI: 0.94, 1.17). Among alcohol users, the total effect estimate of the ADH1B variant on SCCHN (RR= 2.37, 95% CI: 1.12, 4.25) was decomposed into a direct effect estimate of 2.24 (95% CI: 0.88, 5.71) and indirect effect estimate of 1.06 (95% CI: 0.97, 1.16). Approximately 65% and 84% of the excess risk of SCCHN due to CYP2A6 and ADH1B did not involve heavy intensities of smoking and alcohol behaviours respectively.

Conclusion: Our analysis of the Indian data provides empirical evidence that a disadvantageous SEP during childhood is critical for the development of oral cancer later in life. Among the Canadian Caucasian population, GSTP1 105Val decreases the risk for SCCHN independent of smoking, as well as among heavy smokers, and most of the effect of both CYP2A6 and ADH1B genetic variants seems to be though pathways not involving heavy smoking and alcohol risk behaviours respectively, although mediation and interaction by these risk behaviours may play a role in their effects on SCCHN.

Résumé

Introduction : La position socioéconomique (PSE), de multiples variantes dans les gènes encodant les enzymes Cytochrome P450 (CYP) et Glutathione S-transferase (GST), et les comportements à risque liés au tabac et à l'alcool ont été largement étudiés en relation avec le risque de carcinome épidermoïde de la tête et du cou (CETC). Bien qu'une accumulation de PSE désavantageuse au cours de la vie ait été associée à un risque accru de cancer de la bouche, cette association n'a pas été explorée à travers la lentille de multiples modèles du parcours de vie dans une seule étude. Les preuves de l'effet de multiples variantes génétiques dans les gènes CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTP1 et GSTM1 sur le CETC sont conflictuelles est leur effet, indépendamment ou en interaction avec la consommation de tabac, n'a pas été documenté chez la population canadienne caucasienne. En outre, l'utilisation de tabac et d'alcool pourrait interagir ainsi que servir de médiateur de l'effet de variantes génétiques dans CYP2A6 et ADH1B respectivement sur le CETC. Cependant, leurs effets, selon une combinaison de médiation et d'interaction avec les comportements à risque associés, n'ont pas encore été quantifiés. S'attaquer à ces lacunes dans les connaissances a le potentiel d'élucider les mécanismes causaux sous-tendant la relation entre ces facteurs de risque sociaux, génétiques et comportementaux, et le CETC.

Méthodes : Les données de cette thèse ont été tirées des sites indien et canadien d'une étude internationale multicentrique cas-témoin réalisée dans les hôpitaux : l'étude *Head and Neck Cancer (HeNCe) Life*. De l'information sur la PSE dans l'enfance (0-16 ans), la vie de jeune adulte (17-30 ans) et d'adulte (plus de 30 ans), et sur les comportements à risque liés au tabac et à l'alcool tout au long de la vie a été recueillie à l'aide d'entrevues et d'une technique de grille de vie. L'estimation de l'association PSE-cancer de la bouche a utilisé des données sur 350 cas incidents de cancer de la bouche et 371 témoins appariés par fréquence selon l'âge et le sexe recrutés dans deux hôpitaux importants à Kozhikode, Kerala, Inde entre 2008 et 2012. La PSE et les facteurs de confusion ont été considérés comme des variables variant dans le temps. L'estimation de l'effet total des variantes génétiques et de leurs effets d'interaction avec de multiples niveaux de paquets-années (*pack-years*) de tabac a utilisé des données sur 389 cas incidents de CETC et 429 témoins appariés par fréquence selon l'âge et le sexe recrutés dans des hôpitaux de Montréal, Canada. Un sous-échantillon de ces données sur les fumeurs et les consommateurs d'alcool a été utilisé pour estimer les mécanismes causaux entre les variantes génétiques de CYP2A6, ADH1B et le CETC définis par la médiation et l'interaction avec les fréquences d'utilisation de tabac et d'alcool. Des techniques analytiques décrites dans le cadre contrefactuel causal telles que les modèles structurels marginaux pondérés par inverse de probabilité (*inverse-probability weighted marginal structural models*), et la médiation et les techniques de décomposition à quatre niveaux (*mediation and four-way decomposition techniques*) ont été utilisées pour dériver des estimations d'effet causal.

Résultats : La PSE (avantageuse vs désavantageuse) dans l'enfance et dans la vie de jeune adulte étaient associées avec le risque de cancer de la bouche [rapports de cotes (*odds ratios*, RC)=2.76, intervalle de confiance (IC) à 95%: 1.99, 3.81) et (RC=1.84, IC à 95%: 1.21, 2.79), respectivement]. De plus, les participants qui avaient une PSE désavantageuse (vs avantageuse) durant les trois périodes de vie avaient un risque accru de cancer de la bouche [RC=4.86, IC à 95%: 2.61, 9.06]. Le modèle de mobilité sociale de l'enfance à la vie de jeune adulte et les trajectoires du parcours de vie dans l'ensemble indiquaient une forte influence de l'exposition à une PSE désavantageuse dans l'enfance sur le risque de cancer de la bouche.

Parmi toutes les variantes génétiques analysées dans l'échantillon canadien, les porteurs de GSTP1 105Val (vs les non-porteurs) avaient un risque 29% moins élevé (RC=0.71, IC à 95%: 0.53, 0.95) de CETC. Les analyses stratifiées ont montré que les porteurs de cette variante avaient un risque 41% (RC=0.59, IC à 95%: 0.36, 0.95) et 51% (RC= 0.49, IC à 95%: 0.24, 0.98) moins élevé de CETC relativement aux non-porteurs, parmi les strates de gros fumeurs et de non-fumeurs, respectivement. Il n'y avait pas de preuves d'interaction statistique sur une échelle additive ou multiplicative pour aucune des variantes analysées. Parmi les fumeurs, l'estimation de l'effet total de la variante CYP2A6 sur le CETC [risque relatif (RR)=1.28, IC à 95%: 0.46, 3.59] était composée d'une estimation d'effet direct de 1.22 (IC à 95%: 0.45, 3.33) et d'une estimation d'effet indirect via la consommation de tabac de 1.05 (IC à 95%: 0.94, 1.17). Parmi les consommateurs d'alcool, l'estimation de l'effet total de la variante ADH1B sur le CETC (RR= 2.37, IC à 95% : 1.12, 4.25) a été décomposée en une estimation d'effet direct de 2.24 (IC à 95% : 0.88, 5.71) et une estimation d'effet indirect de 1.06 (IC à 95%: 0.97, 1.16). Approximativement 65% et 84% de l'excès de risque de CETC dû à CYP2A6 et ADH1B n'impliquait pas de fortes intensités de comportements liés au tabac et à l'alcool, respectivement.

Conclusion : Notre analyse des données indiennes fournit des preuves empiriques qu'une PSE désavantageuse durant l'enfance est critique pour le développement du cancer de la bouche plus tard au cours de la vie. Chez la population canadienne caucasienne, GSTP1 105Val diminue le risque de CETC indépendamment de la consommation de tabac ainsi que chez les gros fumeurs, et la majorité de l'effet des deux variantes génétiques CYP2A6 and ADH1B semble être via des mécanismes n'impliquant pas les comportements à risque liés au tabac et à l'alcool respectivement, bien que la médiation et l'interaction par ces comportements à risque pourrait jouer un rôle dans leurs effets sur le CETC.

Preface and Contribution of Authors

In this thesis work, I have attempted to address substantive questions related to the pathways underlying the association between specific social, genetic and behavioural risk factors and head and neck cancers. Additionally, I have also demonstrated the application of multiple analytical techniques - originally developed for longitudinal data defined under the counterfactual causal framework - in a case-control study. The dissertation follows a manuscript-based format as outlined by Graduate and Postdoctoral Studies, McGill University. I have organized this dissertation into nine chapters, including introduction, literature review, summary, study objectives, detailed methods, three manuscripts focusing on each objective and corresponding results, followed by an overall discussion, conclusion and appendix files. All chapters of this dissertation are written by me (ThekkePurakkal AS, PhD candidate) under the supervision of Dr. Belinda Nicolau and Dr. Nicolas Schlecht. The following section outlines the contribution of the authors, and elements of originality in each manuscript.

Manuscript I: Socioeconomic position and life-course models

Authors: Akhil Soman ThekkePurakkal, Ashley Isaac Naimi, Sreenath Arekunnath Madathil, Shahul Hameed Kumamangalam Puthiyannal, Gopalakrishnan Netuveli, Amanda Sacker, Nicolas F Schlecht, Belinda Nicolau

Contributions: ThekkePurakkal AS participated in the data collection, designed the objectives and analytical strategy, conducted the data analysis and drafted the manuscript. Naimi AI supervised and participated in the data analysis, interpretation of the results, and preparation of the text. Madathil SA and Shahul HP participated in the data collection and data management. Netuveli G, Sacker A and Nicolas FS participated in the design of the HeNCe Life study and aided in the interpretation of the results. Nicolau B conceived and designed the HeNCe Life study, acquired funding, directed its implementation, quality assurance and control and helped in the interpretation of results as well as the preparation of the draft. All authors critically reviewed and revised the paper, and approved the final manuscript.

Originality: This is the first case-control study investigating the association between socioeconomic position (SEP) over the life-course and oral cancer through the lens of multiple life-course models within a single study, by considering the time-varying nature of SEP and multiple associated confounders over several periods of life. To address analytical challenges, we used a novel approach by combining inverse probability weighting, originally developed to adjust for time-varying variables in longitudinal data, and sampling weights for time-varying exposures in case-control studies, to calculate our associational estimates. We consider this to be a significant contribution to the advancement of analytical methods in life-course research. We also demonstrate the use of a model selection criterion (i.e., weighted quasi-likelihood criterion) for marginal structural models. The results of the study enrich the understanding of pathways underlying the life-course SEP-oral cancer association in the target population. We also contribute annotated software codes that can be used in Stata statistical software to conduct similar analysis in case-control studies, specifically inverse probability weighted marginal structural models.

Manuscript II: Genetic variants in CYP and GST genes, smoking and risk for head and neck cancers: a gene-environment interaction study

Authors: Akhil Soman ThekkePurakkal, Belinda Nicolau, Robert D Burk, Eduardo L Franco, Nicolas F Schlecht

Contributions: ThekkePurakkal AS designed the objectives and analytical strategy, conducted the data analysis and drafted the manuscript. Nicolau B conceived and designed the HeNCe Life study, acquired funding, directed its implementation, quality assurance and control and helped in the interpretation of results as well as the preparation of the manuscript. Burk RD spearheaded the molecular analysis of biological samples. Franco EL contributed to the design of the HeNCe Life study. Nicolas FS participated in the design of the HeNCe life study and directed the genetic component of the study, helped in the interpretation of results and manuscript preparation. All authors critically reviewed and revised the paper.

Originality: This is the first study to investigate the association between multiple genetic variants involved in the metabolism of tobacco carcinogens and head and neck cancers in a Canadian Caucasian population. It is also the first study to consider the CYP2D6*2 single nucleotide polymorphism, and copy number variations of CYP2D6 null in this association. In addition, we also conducted a comprehensive analysis of interaction between these genetic variants and multiple levels of smoking by considering joint, stratum-specific and interaction effects on both additive and multiplicative scales in the target population.

Manuscript II: The effect of interdependencies between CYP2A6 variant and smoking, and ADH1B variant and alcohol, on the risk of head and neck cancers

Authors: Akhil Soman ThekkePurakkal, Belinda Nicolau, Jay S Kaufman, Nicolas F Schlecht

Contributions: ThekkePurakkal AS designed the objectives and analytical strategy, conducted the data analysis and drafted the manuscript. Nicolau B conceived and designed the HeNCe Life study, acquired funding, directed its implementation, quality assurance and control and helped in the interpretation of results as well as preparation of the manuscript. Kaufman JS supervised the data analysis and contributed to the interpretation of the results. Nicolas FS participated in the design of the HeNCe Life study and directed the genetic component of the study, helped in the interpretation of results and manuscript preparation. All authors critically reviewed and revised the manuscript.

Originality: This is the first study to attempt quantification of potential indirect and direct effects of CYP2A6*2 and ADH1B*2 genetic polymorphisms on head and neck cancer risk that may or may not involve an associated risk behaviour (e.g., smoking, alcohol use). This is also the first study to demonstrate the estimation of four potential causal pathways between these genetic exposures and head and neck cancers that is defined by mediation and interaction with associated smoking or alcohol risk behaviours. For this analysis, we used the recently developed four-way decomposition technique under the counterfactual causal inference framework, and this is one of the first case-control studies to demonstrate its application. Due to lack of codes to implement this technique in Stata statistical software, codes were exclusively written for this study using mathematical equations and SAS templates provided by VanderWeele 2014, 2016. The codes are specific to a binary exposure, binary mediator and binary outcome scenario and can be used to calculate various effect estimates and proportions of both mechanistic and policy relevance, and their confidence intervals using a case-control design.

Chapter 1

Introduction

Squamous cell carcinomas of the head and neck (SCCHN) are a heterogeneous group of diseases affecting the oral cavity, pharynx and larynx with wide variation in their incidence and distribution of risk factors across the globe. They are characterised by low survival rates (1) and high morbidity (2) in both developing and developed countries. In addition, these diseases are associated with severe functional, aesthetic and psychosocial consequences, and a considerable economic burden on the patients, families and health care system (3-6).

SCCHN have a multifactorial aetiology that includes social, genetic and behavioural risk factors spread along an individual's life-course. The association between social risk factors such as socioeconomic position (SEP), multiple genetic variants, strong risk behaviours such as tobacco and alcohol consumption, and the risk for SCCHN have been widely investigated. Yet, there is a scarcity of evidence that could provide further insight into "*how*" these risk behaviours may lead to SCCHN later in adult life, and *"who"* among individuals in specific populations with these risk factors are at higher risk for these diseases. Answering these questions has the potential to enrich our knowledge of mechanistic causal pathways between these risk factors and SCCHN, and to inform intervention possibilities and targets. Furthermore, attempting to fill these knowledge gaps requires the quantification of causal pathways by appreciating the basic nature of risk factors along the life-course (e.g., time-varying nature of risk factors such as SEP and behavioural risk factors), and concepts of interaction, mediation and confounding. Utilizing strong theoretical frameworks such as life-course epidemiology that comprehensively articulate these concepts, and complimenting it with analytical tools and techniques developed under the powerful counterfactual causal framework, makes answering these questions feasible. However, in observational studies, these frameworks and techniques rely on a longitudinal design. Hence, there is a pressing need to adapt such existing methods and their application into the context of a case-control study design, commonly used to investigate rare disease outcomes such as SCCHN.

This dissertation attempts to fill some of these knowledge gaps through the amalgamation of life-course and counterfactual causal frameworks, within an international, multicentre, hospital-based case-control study - the Head and Neck Cancer (HeNCe) Life course study - which investigates the aetiology of SCCHN focusing on social, lifestyle, behavioural, biologic and genetic factors. This manuscript-based thesis can be broadly divided into two parts. Manuscript I encompasses the first part and focuses on elucidating causal pathways underlying the relationship between SEP over the life-course and oral cancer risk. Central to this work that explores the SEP-oral cancer association using the accumulation, critical period and social mobility life-course models of risk, is the appreciation of the complex feedback loops between the time-varying SEP exposure and confounders/mediators over multiple periods of life. This work uses data from the Indian site of the parent study, where large socioeconomic disparities and a rapid increase in the incidence of oral cancer in the past few years have been documented.

The second part of the thesis spans manuscripts II and III and focuses on the causal pathways between genetic variants involved in the metabolism of environmental carcinogens, smoking and alcohol use, and SCCHN risk which have not been documented among the Caucasian population in Canada. These manuscripts explore causal interaction and mediation between genetic and associated behavioural risk factors by assuming an accumulation of risks model over the life-course. We also demonstrate the quantification of four potential causal pathways between specific genetic variants and SCCHN under a combined mediation and interaction scenario with the associated risk behaviours. These manuscripts use data from the Canadian site of the parent study where smoking and alcohol are the strongest risk factors for SCCHN.

The objectives of this dissertation are to: (i) estimate the association between life-course SEP measured across three periods of life and oral cancer risk under the critical period, accumulation and social mobility models; (ii) estimate the main effect of genetic variants in CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTM1 and GSTP1 genes involved in the metabolism of environmental carcinogens, as well as their joint, stratum-specific and interactive effects with smoking on the risk of SCCHN; (iii) estimate the extent to which the effect of two specific genetic variants in CYP2A6 and ADH1B genes on SCCHN risk is through smoking and alcohol behaviours respectively, and demonstrate the quantification of proportions of the total excess risk for the outcome due to each genetic variant that is attributable to neither mediation nor interaction, only interaction, only mediation, and both mediation and interaction, with the associated risk behaviour.

Chapter 2

Literature review

This chapter presents current knowledge regarding the epidemiology of squamous cell carcinomas of the head and neck (SCCHN) with special reference to Canada and India, the role of risk factors such as tobacco, alcohol consumption and specific genetic polymorphisms involved in their metabolism and socioeconomic position (SEP), followed by a brief description of life-course epidemiology, case-control study design, counterfactual causal framework, directed acyclic graphs, and an overall summary.

2.1 Squamous cell carcinomas of the head and neck (SCCHN) - Definition

Malignant tumours arising from the squamous cells that line the mucosal surface of the oral cavity, pharynx and larynx [C00-C14, C32 under the International Classification of Diseases (ICD) 10 classification], are commonly referred to as squamous cell carcinomas of the head and neck (7). Histologically, more than 90% of cancers of the oral cavity, pharynx and larynx are of squamous cell origin (8).

2.2 Epidemiology of SCCHN

SCCHN are a heterogeneous group of cancers that differ in distribution, predisposing factors, diagnostic workup and management strategies. According to Globocan 2012 statistics, SCCHN accounted for approximately 599,500 incident cases worldwide, making them the 7th most common cancers in incidence (3.8% of cases) (9). Most of these cancers affect males (70.8%) and are diagnosed above 60 years of age (10). The sub-site with the highest cancer incidence is the oral cavity (300,373), followed by the larynx (156,877) and pharynx (142,387) [Age standardized incidence rates (ASIR) per 100,000 population: oral cavity=4, pharynx=1.9, larynx=2.1]. Globally, these cancers were the 8th most common causes of cancer mortality (3.6% of cases), and were responsible for 300,000 deaths in 2012 (9).

There is wide variation in the geographic distribution of SCCHN incidence across the globe (1, 10). Approximately two-thirds of the burden of incident SCCHN cases is borne by developing countries, with India accounting for 25% of new cases and 35% of deaths occurring worldwide (9). In 2012, approximately 142,000 new SCCHN cases were reported in India, which represents 30% of all incident cancer cases in this country (11). There has been a rapid increase in the incidence of these cancers, specifically oral cancers, in India. A comparison of Globocan 2008 and 2012 reveals that oral cancer surpassed lung cancer in a span of four years to become the 3rd most common cancer in this country after breast and cervical cancers (9, 12).

In developed countries such as Canada, SCCHN accounts for 3% of incident cancer cases (9). An increase in the incidence of SCCHN from 3,000 new cases in 1990 to an estimated 5,650 new cases in 2016 has been reported, leading to 1,650 deaths in this country in 2016 (13). According to Canadian Cancer Statistics 2016, a significant decrease in the incidence rate of oral cavity cancers was noted in males between 1992 and 2003, after which the rates became relatively stable (13). Rates among females did not change significantly between 1992 and 2012. In contrast, the incidence rate of pharyngeal cancers has increased significantly in both males and females since the mid-1990s. In males, the incidence of pharyngeal cancers surpassed that of oral cavity cancers in 2001 while in females, the incidence of oral cavity cancers continues to be higher than that of pharyngeal cancers (13).

A comparison of SCCHN incidence between India and Canada (Table 1) based on Globocan 2012 estimates shows that the age standardised incidence rates (ASIR) for SCCHN overall and nearly all subsites for both males and females are higher in India than in Canada (14).

SCCHN have a significant impact on the quality of life and psychosocial health of the patients and impose a considerable economic burden on their families (5, 15). In the US, patients with SCCHN have more than three times the incidence of suicide compared to the general population (16). Most of these have been reported to occur within the first 5 years of diagnosis and have been attributed to adverse effects on patients' quality of life and resulting psychological distress that may last for decades after successful treatment. The overall 5-year survival rates are low for SCCHN, and vary by cancer sub-site from 35% for oral to 65% for laryngeal cancers (11, 17). Multiple primary tumours developing at the cancer site and a high rate of secondary tumours compared to other

Table 1: Comparison of SCCHN incidence in Canada and India (Globocan 2012)

Type of Cancer	Canada		India	
	Males	Females	Males	Females
SCCHN incidence (total numbers)	3,394	1,347	108,477	32,663
	ASIR per 100,000 population			
SCCHN	11.8	4.2	20.9	6.1
Oral	5.5	2.9	10.1	4.3
Larynx	3	0.6	4.6	0.5
Pharynx	3.2	0.8	6.3	1.3

ASIR- Age standardised incidence rates. Age standardization was performed using the direct method and the World standard population as proposed by Segi (18) and modified by Doll et al (19).

malignancies contribute to this poor prognosis, which has not changed over the past 30 years (4, 20, 21). Although the majority of SCCHN can readily be accessed for visual and tactile examination (e.g., oral cavity cancers), 60% of patients are diagnosed at stages III and IV in North America (22). In India, up to 80% of patients present with advanced disease (11). This situation may be attributed to diagnostic delay (failure in recognizing early signs and symptoms of cancer by patients and/or professionals, delay in accessing professional care) and lack of diagnostic tools with high sensitivity and specificity for the early detection of clinical disease (22, 23). Severe functional and esthetic sequelae, especially for cases diagnosed at late stages, have been reported following treatment for SCCHN. According to a 2007 study, the mean per-patient expense to manage oral cancers in the UK in the first year following diagnosis is 3,500$USD for pre-cancer and 25,000$USD for stage IV cancer patients (6). In North America, SCCHN are responsible for approximately 2.8 billion $USD per year in productivity loss (6). For these reasons, SCCHN have been recognised as a major public health problem in both developed and developing countries.

2.3 Risk factors for SCCHN

SCCHN are complex diseases with multi-factorial aetiology. The discrepancy in the geographic distribution of their incidence has been attributed to variations in the risk factors involved in different locations (1). In developed countries, approximately two-thirds of SCCHN cases are attributed to tobacco smoking and alcohol consumption (10, 24-26) and about 17%-56% of cases may be due to high-risk HPV infection (10, 27-29). In developing countries, such as India and most parts of South Asia, paan chewing is the strongest risk factor (1, 10, 30). Other risk factors include social (e.g., SEP) and psychosocial variables (e.g., acute life events, work stress, depression) (5, 31-35), genetic variants (36-43), diet, sexual behaviour, infection and oral/periodontal health-related factors (1, 10, 44-47). The sections below describe in detail the risk factors for SCCHN; special emphasis is given to tobacco smoking, alcohol consumption, genetic variations (polymorphisms and copy number variations) and SEP because they are central to this dissertation.

2.3.1 Tobacco use and alcohol consumption

2.3.1.1 Tobacco use

Tobacco use is the strongest risk factor for SCCHN. Among the various forms of tobacco consumption (e.g., smoking, chewing and snuffing), smoking (e.g., cigarettes, pipes, cigars, bidi, hookah, chutta, chillam) is the most common (48-50). In its smoked form, tobacco was first used as pipes and cigars, and later as bidis (especially in South Asia), followed by cigarettes in the later half of the nineteenth century (51). Selected characteristics of cigarettes, cigars, pipes and bidis including nicotine content are provided in Table 2 (51-54).

About 50% of men and 9% of women in developing countries, and 35% of men and 22% of women in developed countries smoke tobacco in the form of cigarettes (49). In 2013, the average daily cigarette consumption was 15.2 and 12.5 for male and female smokers respectively in Canada (55). Among the provinces, Quebec reported the highest daily cigarette consumption, at 15.6 overall (males=16.5, females=14.5) (55).

Table 2: Selected physical characteristics of cigarettes, cigars, pipes and bidi

Type	Description	Length (mm)	Diameter (mm)	Average tobacco per stick (g)	Average nicotine per stick (mg)
Cigarette	Roll of tobacco wrapped in paper or non-tobacco material, filter-tipped or untipped[a]	70-120[a]	8[a]	1[b]	1.6 (filter)[b] 2.0 (non-filter)[b] 3.0 (hand-rolled)[b]
Cigar	Roll of tobacco wrapped in tobacco leaf or other substance containing tobacco[a]	110-150[a]	17[a]	5[b]	15[b]
Pipe	Tobacco filled in pipes, lit and puffed[b]	-	-	1.2[b]	5.52[b]
Indian bidi	Roll of raw, dried crushed tobacco flakes (naturally cured), rolled by hand and wrapped in tendu/temburni leaf (Diospyrus mebunoxylon or Diospyrus ebenum). Mostly filterless design[c]	60-100[cd]	< 5[cd]	0.2[c]	2.5-3[d]

[a]IARC, 2004(51); [b]Hoffmann and Hoffmann, 1998(52); [c]Malson, 2001(53); [d]Watson, 2003(54)

In India, approximately 35% of adults use tobacco in some form (49). Paan/betel quid (a combination of tobacco, areca nut and slaked lime wrapped in a betel leaf) chewing is one of the most commonly used forms of tobacco in India, in both males and females (56-59). The prevalence of tobacco smoking is around 14% in India and is much higher in males than females (24% vs 3%) (49). Bidi is the most commonly used smoking product (prevalent in 9% of adults), followed by cigarettes (6%) (49). One bidi produces more nicotine, carbon dioxide, tar, alkaloids and potential carcinogens than a regular cigarette (54, 60-62).

2.3.1.2 Tobacco use and risk for SCCHN

The International Agency for Research on Cancer (IARC) first reported the positive association of tobacco use and alcohol consumption with SCCHN risk in 1985 and 1988, respectively (63, 64). Approximately 69 chemicals identified in tobacco smoke contribute to tumourigenesis, including 10 that are identified as Group 1 human carcinogens by the IARC (65). The most important of these carcinogens, which have also been causally linked to SCCHN, are volatile nitrosamines *[e.g., NDMA (nitrosodimethylamine), NEMA (nitrosoethylamine)]*, nitrosodiethanolamine (NDELA), tobacco-specific nitrosamines (TSNA) *[e.g., 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and N-nitrosonornicotine (NNN)]*, polycyclic aromatic hydrocarbons (PAH) *(e.g., benzo(a)pyrene, benz[a]anthracene),* aromatic amines, benzene and volatile aldehydes *(e.g., acetaldehyde, formaldehyde)* (30, 65, 66).

The oral cavity, pharynx and larynx are directly exposed to tobacco smoke, compared to other sites such as the lungs (51, 60, 67). In the West, approximately 45% of SCCHN cases in men and 75% of cases in women have been attributed to tobacco smoking (26). In a large pooled analysis of case-control studies, Hashibe et al. documented that 30% of SCCHN cases among non-alcohol users was attributed to smoking (16). It is responsible for 60%-90% of deaths from SCCHN in North America (26, 68)(69). The cigarette is the most common form of smoking and thus it is the main route of delivery of tobacco-related carcinogens in most countries. An IARC review documented magnitudes of average relative risk ranging between 4 and 10 for SCCHN risk for ever smokers relative to never smokers (51). However, cigar and pipe smoking may deliver equivalent or higher doses of carcinogens compared to cigarette smoking. Indeed, the largest pooled analysis thus far of 19 studies on SCCHN reported that risk estimates for individuals who smoked only cigarettes, only cigars and only pipes were 3.93, 3.49 and 3.71, respectively, compared to non-smokers (70). Individuals who smoked various combinations of these products were also at approximately 2.5 to 3.5 times the risk for these diseases (70).

The majority of SCCHN cases in India and many Asian countries are attributable to paan/betel quid chewing (10, 58, 59, 71-73). The carcinogenic effect of paan chewing is complex, as it results from an interaction between carcinogens in tobacco, arecoline, the main alkaloid in arecanut, and an increased alkalinity of the oral mucosa due to slaked lime (63, 71, 74-76). In Southern India, approximately 50% of cases in men and 90% in women are attributable to frequent and long-term paan chewing (57). A recent meta-analysis reported a 5-7 times higher risk for oral cancers associated with chewers compared to non-chewers (73) in India. Bidi smoking, reported to deliver approximately 1.5 times the carcinogens of commercial cigarettes, also significantly increases the risk of SCCHN (60). Studies including meta-analytical reviews report a 2-7 times increased risk among bidi smokers compared to non-smokers (58, 59, 77, 78). However, evidence on the association between filtered cigarette smoking and SCCHN in India is mixed. Both case-control and longitudinal studies report little or no association between cigarette smoking and this outcome, mostly due to low exposure (47, 58, 79, 80).

Multiple measures of tobacco use (e.g., frequency, duration, cumulative consumption) have been associated with SCCHN risk, with studies reporting linear or non-linear dose-response relationships with these exposures (26, 58, 59, 72, 78, 81-85). A large pooled analysis in a European male population reported a monotonic increase in risk for SCCHN (from a daily cigarette consumption as low as 2) with increasing frequency of cigarette smoking relative to non-smokers (84). A similar dose-response relationship was demonstrated with a cumulative measure of paan chewing in studies from South India (59) and Taiwan (86). For the association between years since cessation of the habit and SCCHN risk, multiple studies report an inverse relationship (81, 83, 87-90).

2.3.1.3 Alcohol consumption

The World Health Organization (WHO) estimates that there are approximately two billion alcohol consumers worldwide (30). More than half of men (55%) and one third of women (34.4%) consume some form of alcoholic beverage (91) and their drinking patterns vary from occasional to habitual drinking, to alcohol abuse (30). There is wide variation in the type, quality and quantity of alcohol consumed across countries. In Canada, about three-quarters of the population (78%) drink alcohol in the form of wine (10% ethanol), beer (5% ethanol), hard liquor (50% ethanol) and various combinations of spirits (92). In 2011, Quebec reported the highest rate of consumption (82%) in the country (92), and a higher percentage of males consumed alcohol than females (83% vs 74.5%) (92).

Comparatively, the prevalence of alcohol consumption in India is much lower, with only 21% of men and 2% of women who have this habit (93). The state of Kerala located in the Southwest of India reports the highest rates of alcohol consumption in the country (94). In addition to other forms of alcohol, Indians consume high quantities of “toddy”, a beverage produced locally from the fermented and distilled sap of palm and coconut trees (approximately 8-10% ethanol), and a locally brewed liquor known as “arrack”, traditionally produced from fermented palm sap and fruit, grain, or sugarcane (approximately 40-60% ethanol) (72, 75).

2.3.1.4 Alcohol consumption and risk for SCCHN

There is a consensus that alcohol plays the role of a promoter/cocarcinogen in carcinogenesis (38, 64, 95-97). Local exposure to ethanol, the principal type of alcohol found in most alcoholic beverages, is considered to increase the solubility of oral, pharyngeal and laryngeal mucosa, facilitating the penetrance of other carcinogens (26, 96, 98). Heavy drinking induced nutritional deficiencies and a direct toxic effect on the epithelium by alcoholic beverages with high concentrations of ethanol may also contribute to alcohol associated carcinogenesis (97). In addition, certain alcoholic beverages contain low levels of carcinogenic substances (e.g., nitrosamines, urethane, polycyclic hydrocarbons) (64, 89). Furthermore, the primary metabolite of ethanol metabolism in the body, acetaldehyde, is a Group 1 human carcinogen that exerts multiple mutagenic and carcinogenic effects, qualifying alcohol as an initiator of the cancer pathway (26, 99-103). Other potential mechanisms are given in *page no. 29.*

Alcohol consumption accounts for approximately 30% of all SCCHN cases worldwide (104). The greater risk of disease in men is attributed to their higher average alcohol consumption relative to women (101, 104). An increase in the risk of SCCHN with different levels of ethanol consumption, duration, frequency and alcohol types has been documented among never tobacco users (26, 82, 104-106). In a large pooled analysis of case-control studies, Hashibe et al. documented that among never users of tobacco, approximately 7% of SCCHN cases were attributable to alcohol drinking alone (16). A meta-analysis on several cancers reported between 1956 and 2012 documented risk for SCCHN with magnitudes ranging between 1.44 to 1.83, and 2.65 to 5.13 for moderate and heavy drinkers, respectively relative to non-drinkers, among European and North American populations (107). In South India, approximately 26% of the risk for oral cancer is attributable to alcohol consumption, with the risk ranging from 1.2 to 2.8 times higher among moderate to heavy alcohol consumers relative to non-consumers (57, 72, 83).

Similar to tobacco products, a dose-response relationship, either linear or non-linear, has been documented for the association between alcohol consumption and SCCHN (104, 108-111). In a prospective study, Freedman et al. reported an increased risk of SCCHN (1.5 times for males, 2.5 times for females) for 3 drinks per day or more (68). However, a recent meta-analysis reported elevated risks at even lower levels, with risk ratios of 1.29, 3.24, 8.61, 13.2 for 10g (12ml), 50g

(64ml), 100g (127ml), and 125g (160ml) of ethanol per day, respectively (104). Polesel et al. reported a non-linear dose-response relationship in a pooled European study and documented a threshold effect below 50g of ethanol per day for pharyngeal and laryngeal cancers. No such threshold effect was identified for oral cancers and the risk continued to rise with increasing leves of ethanol (110). A recent meta-analysis also reported a non-linear dose-response relationship between frequency of ethanol consumed and SCCHN. However, they did not report a threshold effect for any SCCHN site (112).

2.3.1.5 Combined effect of tobacco and alcohol on risk for SCCHN

The interaction between tobacco and alcohol use in elevating the risk for SCCHN has been well demonstrated. Together they account for approximately 75-80% of SCCHN cases in North America and Europe (24-26) and 50% of oral cancer cases among males in Kerala, India (113). Several studies have considered the nature of the joint effects of smoking and alcohol on SCCHN (82, 104, 107, 108, 111, 114). Positive interactions on both additive and multiplicative scales have been reported between these exposures (101, 104, 108). A non-linear dose-response relationship has been documented for the combined effects of daily alcohol and cigarette consumption (111). For example, a 35-fold increase in risk of SCCHN was observed among those consuming 89g of ethanol and 10 cigarettes daily relative to abstainers of both tobacco and alcohol (111). The risk curve was steeper for increasing daily cigarette consumption among drinkers compared to increasing alcohol consumption among smokers.

To summarise, it has been consistently demonstrated that tobacco use and alcohol consumption in various forms are strong risk factors for SCCHN, and several correlated measures of these exposures (frequency, duration, cumulative measures and time since cessation) are associated with SCCHN risk. The evidence of association also underscores the importance of considering the non-linear functional form of both tobacco and alcoholic beverages in statistical analyses, irrespective of whether they are used as main exposures or confounders. While acetaldehyde is the primary carcinogen derived from alcoholic beverages, carcinogens from tobacco smoke are numerous, including PAH and various nitrosamines.

2.3.2 Genetic polymorphisms and copy number variations

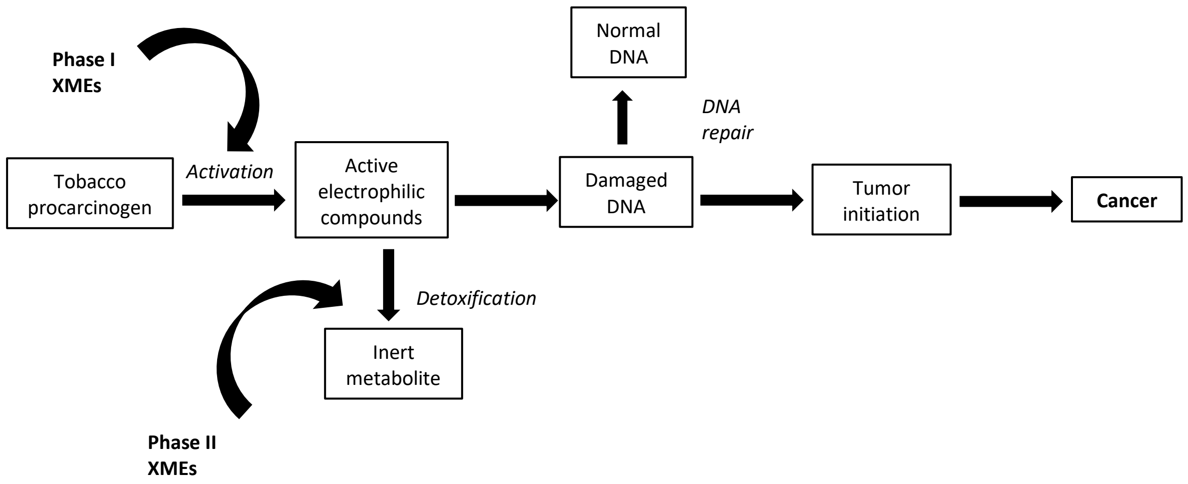
Although tobacco and alcohol are strong risk factors for various cancers (e.g., SCCHN and lung cancer), only a very small proportion of tobacco users and alcohol consumers develop these diseases (35, 115, 116). For example, approximately 10%-15% of smokers develop lung cancers, and even a lesser proportion, SCCHN (115, 116), suggesting inter-individual variation in host susceptibility towards these diseases (35, 117, 118). Investigations of individual genetic makeup have shown that variations in the expression of carcinogen metabolizing enzymes due to variants of genes encoding these enzymes, structural variations in DNA segments, mutagen sensitivity, chromosomal aberrations, DNA repair and apoptosis, contribute alone or in combination to inter-individual variation in susceptibility to cancers including SCCHN (119-123). Single nucleotide polymorphisms (SNPs) are the most common form of variation in the human genome, and SNPs in key genes encoding enzymes involved in the metabolism of specific carcinogens found abundantly in tobacco smoke and alcohol have been the subject of research interest in the past two decades. These SNPs along with risk behaviours are the focus of manuscripts II and III of this thesis. Hence, I describe below the enzymatic pathways underlying the metabolism of tobacco and alcohol carcinogens, specific genes and related SNPs that could alter these pathways and contribute to individual differences in SCCHN susceptibility.

2.3.2.1 Enzymatic pathways in carcinogen metabolism

About 90% of chemical carcinogens from a variety of environmental exposures including tobacco smoke enter the human body as less harmful pro-carcinogens (124). They require bio-activation into reactive molecules for further conjugation, which facilitates their elimination from the body (125, 126). The scenario is similar with constituents of alcoholic beverages. It is hypothesized that part of the susceptibility to tobacco and alcohol related cancers is determined by inter-individual differences in the bio-activation of pro-carcinogens and detoxification of carcinogens derived from these exposures.

The bio-activation and detoxification processes are catalysed by enzymes generally known as Phase I and Phase II xenobiotic metabolizing enzymes (XMEs), respectively (118, 127, 128). Mostly expressed in the liver, these enzymes are also found in the mucosal lining of various organs including the upper aero-digestive tract. The Phase I XMEs that activate pro-carcinogens from environmental sources (including tobacco smoke) into intermediate reactive, electrophilic metabolites belong mainly to the superfamily of cytochrome P450 (CYP) enzymes. Phase I XMEs belonging to the alcohol dehydrogenase family (ADH) oxidise ethanol to acetaldehyde. These reactive moieties (e.g., diol epoxides, arene dioxides, acetaldehyde from ethanol) are genotoxic and can form DNA adducts that may cause mutations in the DNA and result in cell transformation, transcription and translation errors (***Figure 1***) (125). If DNA repair does not occur, these molecular changes persist and mark the earliest events in the pathway leading to cancers. The detoxification and elimination of these reactive moieties are facilitated by Phase II glutathione S-transferase (GST) via conjugation by nucleophilic glutathione, and acetaldehyde dehydrogenase (ALDH) XMEs. The increased water solubility of the substrates from Phase I biotransformation following the action of Phase II enzymes ultimately gets them eliminated through urine and sweat (125).

Figure 1: Pathways involving biotransformation of tobacco carcinogens (adapted from Costa, 2006)(125)



2.3.2.2 Phase I and Phase II enzymes, associated genes and SNPs

With respect to tobacco and alcohol related cancers, several enzymes belonging to the families of CYP, GST, ADH and ALDH enzymes have been studied. Some of the most widely studied are CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTM1, GSTP1, GSTT1, ADH1B, and ALDH2 (129). The specific pro-carcinogenic and carcinogenic substrates of these enzymes are provided in Table 3 (124, 130-132).

The catalytic activity of each of these enzymes is determined by genes (DNA sequence) encoding them. For example, CYP1A1 is encoded by the CYP1A1 gene. Alternative forms of a given gene (or variants of genes) that differ in function, resulting from variations within the nucleotide sequence in DNA at a given gene locus, are termed alleles. DNA sequence variations resulting in alleles that are common in the population (i.e., the least frequent/rare/minor allele occurs in more than 1% of the population due to natural selection of genetic drift) are known as genetic polymorphisms.

Table 3: XMEs and their substrates present in tobacco and alcohol

Enzyme	Substrates
CYP1A1	Polycyclic aromatic hydrocarbons (PAH), heterocyclic aromatic amines (HAA),
CYP2A6	NNK, N-Nitroso-N-Diethylamine (NDEA), nicotine, cotinine, ether
CYP2D6	Amines, nicotine
CYP2E1	Benzene, acrylonitrile, N-Nitroso Diethylamine (NDEA), TSNA (NNK, NNN), ether, ethanol
GSTM1	Arene oxide, diol epoxide
GSTP1	Arene oxide, diol epoxide
ADH1B	Ethanol
ALDH2	Acetaldehyde (from both alcohol and cigarette smoke)

When polymorphic DNA sequences occur due to alterations at a single nucleotide base, they are termed SNPs. Based on the combination of alleles from the maternal and paternal chromosomes, three variable types of individuals can be identified in the population. Genotypes resulting from the presence of the same allele on both chromosomes are referred to as homozygous, whereas those with a wild type allele on one chromosome and a variant allele on the other (maternal or paternal), are termed heterozygous. Homozygous wild type (wild/wild) is usually associated with a functionally normal enzyme, whereas homozygous mutant (variant/variant) or heterozygous (wild/variant) genotypes can result in a functionally different enzyme (e.g., a fast, slow, inactive enzyme).

In summary, SNPs result in different groups of individuals (e.g., carriers: homozygous variant + heterozygous genotypes, non-carriers: homozygous wild type genotypes) with distinct traits (inter-

individual variation) in a given population. These SNPs can lead to functionally different xenobiotic enzymes involved in the biotransformation of tobacco and alcohol pro-carcinogens, which, in turn, underlie the differential SCCHN risks among individuals with different genotypes.

2.3.2.3 Candidate genes and SNPs associated with carcinogen metabolism and risk for SCCHN

The genes encoding Phase I and Phase II XMEs are highly polymorphic and various SNPs are associated with these genes (133). These SNPs can lead to enzyme products with increased, altered, decreased or no activity (132, 134). SNPs enhancing the activity of Phase I CYP enzymes (e.g., CYP1A1*2A, CYP1A1*2C, CYP2E1c2) result in faster conversion of tobacco pro-carcinogens to reactive carcinogenic metabolites (135-138). Similar functional changes to ADH1B enzymes (e.g., ADH1B*2) lead to a higher conversion rate of ethanol to acetaldehyde (38, 139). Certain SNPs related to Phase II XMEs (e.g., GSTP1 105Val) that cause a decreased activity of corresponding enzymes may result in decreased detoxification and excretion of these genotoxic metabolites (140). Overall, these functional changes may result in an overload of reactive carcinogens in the human body, which can lead to an increased risk of SCCHN. Other groups of SNPs that decrease the activity of Phase 1 XMEs (e.g., CYP2D6 null) or Phase II XMEs (e.g., GSTM1null) may result in a decreased production or decreased rate of detoxification of these metabolites respectively, resulting in differential risk for SCCHN (141, 142). Furthermore, because tobacco smoke is one of the richest sources of carcinogenic chemicals that are substrates for these enzymes, the association between these SNPs and SCCHN risk can vary depending on different levels of tobacco smoking. This gene-environment interaction can result in sub-groups with differential risk for SCCHN within a population. The identification of high-risk groups can ultimately aid in targeting prevention activities. Hence, in this work, we focus on the association between several widely-studied SNPs altering the functions of Phase I and Phase II XMEs and SCCHN risk, alone or in interaction with tobacco smoking. We also consider the ADH1B*2 SNP associated with alcohol metabolism. A summary of characteristics of these genetic variants are provided in Table 3 and described in the sub-sections below.

Table 4: Characteristics of SNPs involved in tobacco and alcohol metabolism

Gene	Chr[a]	Designations of the variant allele and rs_number	Nucleotide or amino-acid substitution	Minor allele frequency among Caucasians %	Effect on enzyme activity	Overall evidence on SCCHN risk till date	Evidence on combined effects of variants and smoking/alcohol on SCCHN risk	References
CYP1A1	15	*2A, msp1, m1 rs4646903	3801T>C	5-10	Increased	Increased risk among Asians but not among Caucasians	Overall increased risk among smokers. Inconsistent evidence among Caucasians	(143-146)
CYP1A1	15	*2C, m2, rs1048943	2454 A>G	3-5	Increased	Increased risk among Asians but not among Caucasians	Overall increased risk among smokers. Inconsistent evidence among Caucasians	(143-146)
CYP2E1	10	c2, PstI, *5B rs3813867	1293G>C	< 10	Increased	Increased risk among Asians and mixed population but not among Caucasians	Inconsistent results for stratum specific effects	(138, 147-150)
GSTP1	11	Val, *B, rs1695	105 A>G	10-40	Decreased	No overall association	No conclusive evidence. Increased risk among smokers reported.	(145, 151, 152)
CYP2D6	22	Null (copy number variation identified)	deletion	6-10	No effect	Copy number variations not studied	Copy number variations not studied	(123, 153-156)
GSTM1	1	Null (copy number variation identified)	deletion	40-60 (up to 10% has 2 copies of null allele)	No effect	Increased risk for SCCHN including Caucasians.	Increased risk among smokers	(123, 157, 158).
CYP2A6	19	*2 rs1801272	1799T>A	1-3%	Complete or partial inactivity	Copy number variation not studied with SCCHN	Not studied	(159, 160)
ADH1B	4	*2 Arg48His rs rs1229984	48G>A	1-43%	Decreased or No effect	Decreased risk for SCCHN in all ethnicities	Among drinkers, decreased risk. No effect among, non-drinkers	(38, 161, 162)

[a] Chr: Chromosome; rs_number*- stands for reference SNP cluster ID. It is an accession number that is a stable and unique identifier for SNPs

CYP1A1*2A, CYP1A1*2C and SCCHN risk

The CYP1A1 is a highly active CYP enzyme majorly involved in the activation of pro-carcinogens such as polycyclic aromatic hydrocarbons (e.g., benzo(a)pyrene) and aromatic amines found in tobacco smoke, environmental pollutants and smoked food (136). The enzyme is encoded by the CYP1A1 gene found on chromosome 15. Polycyclic aromatic hydrocarbons induce the expression of this gene (163). SNPs designated as CYP2A1*2A, and CYP1A1*2C, are two of the most widely studied polymorphisms in this gene (143). These SNPs are normally inherited together, resulting in a non-random association (linkage disequilibrium=LD) between them (164). The frequencies of minor alleles (C allele for CYP1A1*2A and G allele for CYP1A1*2C) vary in different ethnicities with 5-10% reported for the C allele and 3-5% for the G allele among Caucasians (146, 148). These SNPs, which occur on the restriction sites that control enzyme activity, result in increased enzyme activity (~ 2-fold) (146, 148). Based on the hypothesis that increased enzyme activity leads to enhanced activation of pro-carcinogens to carcinogens, these SNPs are considered to increase the risk of SCCHN (40, 135, 144, 164-168).

Multiple meta-analytical reviews have aimed to clarify the association between the two CYP1A1 SNPs and SCCHN risk (116, 135, 136, 145, 165, 169-171). An overall increased risk associated with the CYP1A1*2A allele and SCCHN has been reported in all reviews on this variant (116, 136, 165, 169). However, this association is seen only among Asians and not Caucasians. The overall evidence for the association between the CYP1A1*2C allele and SCCHN risk is inconsistent. While none of the reviews conducted to date identified any association among Caucasians, two reviews reported an increased risk among Asians alone.

Meta-analytical reviews have also considered the combined effect of these polymorphisms and smoking for the risk of SCCHN. Valerie and Lema et al. (2008), Liu et al (2013) and He et al (2014) documented an increased risk for SCCHN among the CYP1A1*2A carriers (vs non-carriers) among the smokers alone, but not among the non-smokers. A similar increased association between the CYP1A1*2C allele and SCCHN among the smokers alone was documented by Liu et al. (2013) and Qin et al. (2014). In a joint effect analysis in their review, Liu et al. (2013) reported that, compared to the non-smokers and non-carriers of the CYP1A1*2C allele (AA genotype), carriers (AG/GG genotype) and smokers had the highest risk (approximately 2.4-fold), followed by the non-carriers and smokers (2-fold risk) (165). They also reported similar associations for the joint effects of CYP1A1*2A and smoking. Relative to the non-carriers (TT genotype) and non-smokers, the carriers (TC/CC genotype) and smokers had an approximately 3-fold increase in risk, followed by the non-carriers and smokers (1.78 times the risk). Overall, they reported a positive multiplicative interaction estimate of 1.5 between risky genotype categories of both SNPs and smoking. Similarly, Qin et al. (2014) reported a positive interaction (1.51 times the risk) between carriers of the CYP1A1*2C variant and smoking (135). For CYP1A1*2A, He et al. (2014) reported 2.37 times the risk for SCCHN among carriers who were smokers (136). However, the stratum-specific and joint effect results presented in all these reviews were averaged over multiple ethnicities. Hence, there is a lack of understanding regarding interactive effects of these alleles and smoking on SCCHN risk among specific ethnicities. *Therefore, although extant research suggests that the combined effect of these SNPs and tobacco smoking intensifies the risk for SCCHN, there is a need for studies comprehensively reporting interaction results among Caucasian populations.*

CYP2E1c2 and SCCHN risk

CYP2E1 is involved in the metabolic activation of compounds such as benzene, acrylonitrile, N-dimethyl nitrosamines and ether from tobacco smoke. It is encoded by the CYPE2E1 gene, located on chromosome 10. The gene is inducible by low dose nicotine and ethanol (172). A SNP designated as CYP2E1c2 is widely studied with regards to various tobacco related cancers (137, 147, 149, 150, 173-176); its minor allele (c2 or C) has a frequency of less than 10% among Caucasians (137, 146). This allele is associated with increased enzyme activity [i.e., the c2/c2 genotype (CC genotype) has almost 10 times more carcinogen activating capacity than the c1/c1 genotype (GG genotype)] and hence is hypothesised to increase the risk for SCCHN (137, 138, 150, 166, 177-179). The most recent meta-analysis, conducted on 43 studies, suggested that carriers of the c2 allele are at increased risk for SCCHN among Asians and mixed populations, but not among Caucasians (180). Previous meta-analyses reported similar findings (137, 181). However, individual studies considering the combined effect of CYP2E1c2 and smoking (mainly stratum-specific effects) provide conflicting results (150, 182), *Hence, similar to the CYP1A1 genetic variants, there is a lack of studies comprehensively analysing the possibility of an interaction between CYP2E1c2 and various tobacco smoking levels among Caucasians.*

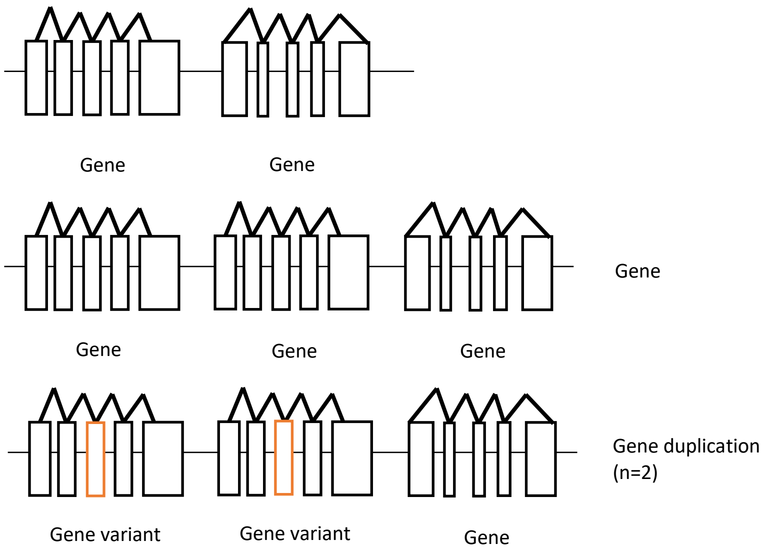
GSTP1 105Val and SCCHN risk

Belonging to a superfamily of multi-functional Phase II XME, the GSTP1 enzyme is the most widely expressed GST enzyme in the head and neck region (183, 184). It detoxifies various electrophilic substrates including active metabolites of carcinogens such as polycyclic aromatic hydrocarbons, monohalomethanes and ethylene oxide, and is encoded by the GSTP1 gene located on chromosome 11 (177). A SNP in this gene designated as GSTP1 105Val has been studied in relation to multiple cancers including SCCHN (147-150, 166). In Caucasians, the frequency of the minor allele (G) is approximately 10-40% (151, 185, 186). Compared to the wild type A allele, the G allele encodes an enzyme that is 2-3 times less stable and hence less efficient in detoxifying its substrates (140, 145, 152). However, three meta-analyses conducted thus far have failed to identify any association between carriers of the G allele and SCCHN risk (140, 145, 187). They also did not identify any conclusive evidence supporting an interaction between ever smoking and the G allele. Nevertheless, increased risk estimates for the joint effect of carrying the G allele and smoking have been reported with increasing levels of daily cigarette consumption and pack-years (36). To complicate matters further, GSTP1 105Val is known to be highly substrate-specific. Although less efficient in detoxifying substrates such as 1-chloro-2,4-dinitrobenzene relative to the stable enzyme, this enzyme is highly efficient in detoxifying carcinogenic epoxides of PAH (e.g., benzo(a)pyrene) (188, 189). Hence, carriers of the 105Val allele have been hypothesized to be less susceptible to PAH-induced DNA damage and carcinogenesis. Indeed, a lower risk for SCCHN and multiple cancers among 105Val allele carriers has been documented (190-195). *In summary, there is an inconsistency in the current evidence on the effect of the GSTP1 105Val allele on SCCHN. Furthermore, more studies are required to comprehensively investigate joint effects and interaction between GSTP1 105Val and tobacco smoking.*

2.3.2.4 Copy number variants

Copy number variants (CNV) or polymorphisms have been defined as DNA segments present in variable copy numbers (repeats) in comparison with a reference genome (196). These segments are 1 kilobase or larger in size (from one kilobase to several mega-bases) and include deletion, duplication, insertion, inversion or complex recombination (***Figure 2***) (123). These structural variants are as important as SNPs in their contribution to genome variation. Genetic variants containing 0-13 gene copies have been reported across human populations (123). CNVs in genes involved in tobacco carcinogen activation and detoxification have been identified and are reported to alter SCCHN susceptibility (123). The identification of CNVs is advantageous in estimating the risk associated with various copy numbers of a variant rather than broad categorizations such as carriers vs non-carriers of the variant. In this work, apart from the SNPs already described, we consider CNVs in two genes, one encoding a Phase I (CYP2D6) and the other a Phase II (GSTM1) enzyme, the null variants of which render these respective enzymes non-functional.

Figure 2: A depiction of copy number variants in the human genome adapted from He et al, 2012 (122)



A gene recombination event between two genes can result in gene duplication or multiplication (n=2, 3,...) or gene deletion (n=0). A duplication of gene could carry mutations from the original copy (red column)

CYP2D6 non-functional (null) CNV and SCCHN risk

CYP2D6 (debrisoquine hydrolase) is the most genetically polymorphic of metabolic enzymes, with approximately 80 variants identified. It is majorly involved in the metabolism of nearly 20-25% clinically used drugs (154, 197) and pro-carcinogens from tobacco (e.g., various amines, nicotine) (198). The XME is encoded by the CYP2D6 gene located on chromosome 22. The variants identified are comprised of SNPs, deletions and insertions, and include normal activity, reduced activity or non-functional alleles (123). There is no detectable activity for this enzyme when encoded by CYP2D6 non-functional alleles (null alleles). Approximately 6-10% of Caucasians harbouring these null alleles are termed poor metabolizers of enzyme substrates (153, 198-200). Due to a lower activation of pro-carcinogens to carcinogens, CYP2D6 null is hypothesised to be associated with a lower risk for tobacco related cancers such as SCCHN compared with highly active functional variants. However, evidence on this association has been inconsistent (141, 155, 156, 201). CNVs exist for CYP2D6 null (123) and individuals with lower numbers of copies of the null variant could have an increased risk for SCCHN compared to those with higher numbers of this variant. *However, this hypothesis has not been explored yet, nor the interaction between CYP2D6 null CNVs and tobacco in the risk for SCCHN.*

GSTM1 CNV and SCCHN risk

Similar to GSTP1, the GSTM1 enzyme is involved in the detoxification of a variety of activated compounds from tobacco smoke with carcinogenic potential. The GSTM1 gene on chromosome 1 encodes the GST-mu enzyme (157). Among the three polymorphisms isolated for this gene, the GSTM1 null gene renders the GST-mu enzyme inactive, and individuals with this allele do not detoxify tobacco related carcinogenic compounds efficiently (202). An accumulation of such compounds that can form DNA adducts could increase the risk for cancers such as SCCHN. The null allele has a frequency of 40-60% among Caucasians (157). Multiple meta-analytical reviews support the hypothesis that GSTM1 null is associated with an increased risk of SCCHN in various ethnicities including Caucasians (40, 116, 142, 145, 203, 204). A higher risk has also been identified among smokers, suggesting an interaction between GSTM1null and tobacco smoking (205, 206). Relative to non-smokers with normally active GSTM1 (non-null), individuals who were smokers and GSTM1 null carriers have up to 5 times greater risk for SCCHN, with the risk increasing with heavier levels of tobacco smoked (6 times for GSTM1 null + more than 20 daily cigarette consumption, 7.4 times for GSTM1 null + more than 40 pack-years of tobacco) (36). CNVs have been identified for GSTM1. Approximately 10% of Caucasians have up to 2 copies of the GSTM1 homozygous deletion (123, 158). Although studies on SCCHN (primary tumours, secondary primary tumours and recurrent tumours), bladder and prostate cancer documented no risk associated with one copy of GSTM1, the presence of at least 2 copies of GSTM1 was associated with a low risk for these outcomes compared to GSTM1 homozygous deletion (207-210). *The interaction between CNVs for GSTM1 and tobacco smoking has yet to be reported comprehensively among Caucasians.*

2.3.2.5 SNPs associated with tobacco and alcohol risk behaviours and risk for SCCHN

The genetic variants discussed so far are hypothesised to be associated with SCCHN risk independently or in interaction with smoking. However, there are SNPs that not only have the potential to interact with risk behaviours, but are documented to affect tobacco and alcohol risk behaviours. CYP2A6*2 and ADH1B*2 are two such variants that influence tobacco and alcohol consumption behaviours respectively. These variants are the focus of manuscript III and are described in the sub-sections below.

CYP2A6*2, intensity of smoking and SCCHN risk

*CYP2A6*2 and nicotine metabolism*

Tobacco smoking is a complex behaviour influenced by social, environmental, psychological and genetic risk factors (211-213). The various phases identified in the continuum of this behaviour include the preparatory stage, initial trying (initiation), repeated irregular/sporadic use (experimentation), regular use, nicotine dependence/ addiction, cessation and relapse (211, 214). Following initiation, this rewarding behaviour is strongly determined by the addictive agent in tobacco called nicotine (52). Within 10-20 seconds of its inhalation, nicotine reaches the brain and starts exerting its psychoactive effects (215). However, nicotine has a short half-life (8 minutes on average) as it is rapidly inactivated and removed from the body, lowering its levels in plasma and tissues (215). Hence, to attain and maintain optimal levels of nicotine in the brain, the individual has to smoke again. *Thus, factors affecting the metabolism of nicotine may influence various phases of smoking behaviour.*

Approximately 70-80% of the nicotine entering the body is metabolized/ inactivated into cotinine through a 2-step process (215, 216): nicotine is first converted to nicotine iminium ion, which is later oxidized into cotinine. The first part of the process is the rate limiting step and is catalyzed by the Phase I CYP2A6 enzyme, mainly in the liver. Overall, 80-90% of the inactivation of nicotine to cotinine is catalyzed by the CYP2A6 enzyme encoded by the CYP2A6 gene on chromosome 19 (216-218). Although several SNPs have been identified in this gene, only a few have been functionally characterised as capable of altering enzyme activity (218-220). Based on their activity, carriers of the functional SNPs have been grouped as slow nicotine metabolizers (individuals hypothesised to smoke less), intermediary metabolizers (individuals hypothesised to be moderate smokers) and normal metabolizers (individuals hypothesised to smoke heavily) (219). Of these genetic variants, the first to be characterised and one of the most widely studied is CYP2A6*2 (217), which is categorized under slow metabolizers. The homozygous variant (AA) and heterozygosity (AT) of this allele result in complete and partial inactivity of the CYP2A6 enzyme, respectively (213, 221). Consequently, relative to homozygous wild type (TT genotype), smokers who are carriers of the variant (AA or AT genotypes) of this allele exhibit higher plasma nicotine levels for a given amount of ingested nicotine (due to a lower conversion rate of nicotine to cotinine). *Based on this mechanism, the CYP2A6*2 allele was hypothesised to have an inverse association with smoking behaviour (e.g., number of cigarettes smoked per day, nicotine dependence).*

*Association of CYP2A6*2 with cigarettes smoked per day*

There is strong evidence for the association between the CYP2A6*2 allele and number of cigarettes smoked per day among Caucasian adult smokers. Inter-ethnic variation has been reported in the frequency distribution of the CYP2A6*2 allele. Although they are rarer (0-0.7%) in the Chinese, Korean and Japanese population, their frequencies range from 1% to 3% among Canadian, American and European Caucasians (217). Several (159, 160, 222-224) but not all (225-227) studies looking into the association between CYP2A6*2 and smoking behaviour among Caucasian adult smokers reported that the CYP2A6*2 allele (AT/AA genotype) protected smokers from becoming nicotine dependent and that carriers smoked fewer cigarettes per day relative to non-carriers (TT genotype). A meta-analysis including observational studies published between 1998 and 2004 documented no overall association between the CYP2A6 gene (multiple variants) and smoking behaviour (213). However, the majority of the studies included in the review used broad definitions of smoking (e.g., ever/never/current/former smoker), which may have led to misclassification of the outcome and obscured significant differences between the groups. Given the existence of a well demonstrated biological mechanism connecting the gene, nicotine metabolism and smoking behaviour, the researchers attributed their results mainly to a lack of methodological rigour in the studies investigated and emphasised the importance of specifically defining the smoking variable (213). Another meta-analysis in the same year provided evidence that smokers who were carriers of at least one CYP2A6*2 allele smoked significantly fewer cigarettes per day, and also had higher chances of quitting smoking (159). The first ever study on CYP2A6 poor metabolizers conducted among Canadian Caucasians in 1998 (222) was reanalysed by Rao et al. (223) using stringent analytical methods. These included: a) a new genotyping method that, unlike the original study, removed the chances of CYP2A6*2 false-positives (228), b) a precise definition of the smoking outcome through multiple indices, and c) proper control for population stratification (confounding by variation in ethnicity) by the restriction of participants to Caucasian smokers who had at least 3 grand-parents of Caucasian ethnicity. This study reported that among smokers, relative to non-carriers of the CYP2A6*2 allele (TT genotype), carriers (AT or AA) smoked fewer cigarettes per day [(13.5 vs 19.5, P<0.03) overall, and at times of heavy smoking (19 vs 29, P<0.001)], had lower breath carbon monoxide levels and lower cotinine levels. A similar study conducted in another North American population reported that among dependent smokers, slow metabolizers (which included carriers of at least one *2 allele) smoked 7 fewer cigarettes per day on average relative to non-carriers (21.3 v 28.3 cigarettes per day) (224). Also, among Caucasians who smoked at least 10 cigarettes per day, those who were slow metabolizers had significantly lower mean and overall puff volume compared to normal or intermediary metabolizers (229).

A genome-wide meta-analysis conducted in 2010 that analysed 710 SNPs on chromosomes 15, 19, and 8 among adult participants of European ancestry, documented a strong association between the CYP2A6*2 allele and number of cigarettes smoked per day (230). A recent meta-analysis on slow metabolizers of CYP2A6 also reported similar findings (160).

*Overall, findings from observational studies and meta-analyses reported thus far indicate that, due to their involvement in nicotine metabolism, smokers who are homozygous (AA) or heterozygous (AT) for the CYP2A6*2 allele smoke with less intensity (cigarettes per day) relative to homozygous non-carriers (TT).*

*CYP2A6*2 and SCCHN risk*

Based on their involvement in the activation of tobacco pro-carcinogens to carcinogens, CYP2A6 genetic variants have been implicated in the risk for SCCHN. However, the literature on this association is sparse. Three studies have investigated the role of CYP2A6*4 in the risk for tobacco-related cancers (41, 150, 231). This SNP has a lower frequency among Caucasians (0.5-1%)

compared to CYP2A6*2 (217). However, similar to CYP2A6*2, CYP2A6*4 renders the enzyme inactive, resulting in decreased bio-activation of substrates such as nicotinate and NNK, NNN and NDEA pro-carcinogens found in tobacco (217). Carriers of the CYP2A6*4 allele have been associated with a significantly lower risk for tobacco related cancers including those of the upper aerodigestive tract. Furthermore, this variant is suggested to affect cancer risk solely in smokers (41, 232). Based on similarities with CYP2A6*4, it can be hypothesised that smokers who are carriers of the CYP2A6*2 allele (AT or AA) are at a lower risk for SCCHN. *However, no studies have yet investigated the potential role of CYP2A6*2 in SCCHN risk nor its interaction with smoking.*

ADH1B*2, alcohol consumption and SCCHN risk

Alcohol consumption patterns are influenced by social, environmental, psychological and genetic factors with inter- and intra-ethnic variability (233-236). Much of the inter-individual variability in alcohol use is attributable to factors underlying the metabolism of ethanol (162, 235, 236). Ethanol entering the human body is first metabolized into acetaldehyde and later to acetate before being removed from the body. The oxidation of ethanol to acetaldehyde is catalyzed by ADH and its iso-enzymes majorly in the liver. These iso-enzymes are also expressed in the stomach, gut and upper aerodigestive tract in detectable quantities. Similar to nicotine metabolism, the inter-individual variability in alcohol to acetaldehyde metabolism is mostly attributed to the genetic polymorphisms in ADH genes encoding ADH enzymes. Of these, SNPs related to ADH1B and ADH1C iso-enzymes of ADH, namely ADH1B*2 and ADH1C*1, are two of the most functionally polymorphic and well characterised variants in adults. These SNPs are not only associated with alcohol consumption behaviour, but also with altered risk for SCCHN among alcohol consumers in various ethnicities (237). Although they seem to be in LD, studies in both Caucasian and Asian populations suggest that ADH1B*2 has a significant effect on the risk of SCCHN after adjustment for ADH1C*1 (238, 239). Also, among multiple ADH SNPs studied, ADH1B*2 has the strongest association with alcohol consumption behaviour and SCCHN (139, 162, 237). Hence, we will be focusing on the role of ADH1B*2 in relation to both SCCHN risk and alcohol consumption behaviour.

*Association between ADH1B*2 and alcohol metabolism*

The role of ADH1B*2 in alcohol consumption behaviour has been widely investigated (161). The frequency of this allele varies in different ethnicities [Asian: 69% (range: 19%-91%), European: 5.5% (range: 1%-43%), Mexican: 3% (range: 2%-7%)] (161). The homozygous variant (AA genotype) and heterozygosity (AG genotype) of this allele result in an ADH enzyme that rapidly oxidizes ethanol to acetaldehyde (up to 50-100-fold increase in activity) (38, 162, 233, 236). Carriers of this allele (AA or AG genotype) are at decreased risk of alcohol dependence compared to non-carriers (GG genotype). This is hypothesised to be due to the prompt build-up of acetaldehyde (resulting from the rapid oxidation of ethanol), which leads to negative physiological reactions termed alcohol-induced flushing; this condition is characterised by cutaneous flushing, increased skin temperature, decreased blood pressure, tachycardia, dizziness, anxiety, nausea, headache and generalised weakness (240). These aversive reactions lead to decreased alcohol consumption.

*Association of ADH1B*2 with alcohol consumption behaviour*

The association between ADH1B*2 and alcohol consumption behaviour was first investigated in East Asians (241-244), and then later among Europeans and other ethnicities (238, 245, 246). Among East Asians, this allele decreases the risk of alcohol dependence by about 80% relative to non-carriers (243, 245). A study on 4,597 Australian twins (3 studies combined) reported that non-carriers of the ADH1B*2 allele (GG genotype) had fewer negative reactions post alcohol consumption (p=8.2x10-7), consumed a higher number of drinks per day (p=2.7x10-6) and had a greater overall cumulative alcohol consumption (p=8.9x10-8) relative to carriers (162). On average, participants with GG, GA and AA genotypes consumed 5.1, 4.1 and 1.9 drinks per day. A recent meta-analysis (2,298 alcohol-dependent cases and 3,334 non-dependent controls) documented that the ADH1B*2 allele was associated with a significant reduction (by 66%) of alcohol dependence and number of drinks per day among European-Americans. Another meta-analysis on all studies published between 1990 and 2011 reported similar findings with robust associations (161). *Overall, the accumulated evidence is consistent with the hypothesis that an elevation in acetaldehyde leads to an increased sensitivity to alcohol among ADH1B*2 carriers,* *reducing the likelihood for alcohol dependence and number of drinks per day among Caucasian adults.*

*ADH1B*2 and SCCHN risk*

ADH1B*2 has been strongly implicated in the risk for upper aerodigestive tract cancers among various ethnicities. Acetaldehyde, the initial metabolite of ethanol, has been suggested to exert multiple mutagenic and carcinogenic effects, qualifying alcohol as an initiator of the cancer pathway (99-102). Hence, it was hypothesised that fast metabolizers of ethanol (GA or AA genotype) have a higher exposure to acetaldehyde, increasing their risk for SCCHN (38). However, contrary to this hypothesis, the first reported study investigating this association (among Japanese alcoholics) reported an increased risk for SCCHN among the GG genotype relative to the GA or AA genotype (247). Brennan et al. reasoned that this result was due to residual confounding by alcohol consumption (38). However, studies since then have consistently shown a decreased risk (up to a 50% reduction) for SCCHN among carriers of the GA or AA genotype (139, 237, 248). No association was identified among never-drinkers (237, 249) and the protective effect was significant at higher levels of alcohol. These reports hypothesize alternative mechanisms of carcinogenesis *(described in page 29).*

The combined effect of the ADH1B*2 allele and alcohol consumption has also been investigated. A joint effects analysis conducted among the Japanese population reported that when compared to non-drinkers who were AA or GA/AA genotype carriers, GG genotype carriers who were drinkers were at significantly increased risk for the disease. The effect was more pronounced among heavy drinkers (9-26 times higher risk) (239, 250). A Korean study documented a higher risk for the GG genotype compared to the AA genotype within moderate and heavy drinker strata of alcohol consumption (251). Two recent studies among Caucasians did not document any interaction between ADH1B*2 and alcohol consumption levels (248, 252). However, large European studies, which documented significant lower risk among the strata of medium and heavy drinkers among carriers of this allele (GA/AA genotype), do indicate a possibility of negative interaction on an additive or multiplicative scale within this ethnicity (139, 237). Studies among Asian and Caucasian populations have consistently documented no altered risk among never-drinkers who were either carriers or non-carriers of the ADH1B*2 allele. *Overall, studies investigating both* *main effect and stratum-specific effects indicate the possibility of interaction between ADH1B*2 and measures of alcohol consumption.*

*Hypothesis underlying the association between ADH1B*2 and risk for SCCHN*

Multiple potential pathways (not mutually exclusive) underlying the association between ADH1B*2 and SCCHN among alcohol consumers have been proposed. Most of them are based on a direct carcinogenic action of acetaldehyde. Hashibe et al. reasoned that the fast metabolism of ethanol (among GA/AA genotypes) leading to increased acetaldehyde exposure may initiate alternative mechanisms to clear off the peak of acetaldehyde. However, such mechanisms may not be activated among GG genotype carriers who have a moderate initial metabolism, leading to acetaldehyde build up, which in turn increases the risk for cancer (139). In addition, compared to ADH enzymes, the expression of acetaldehyde dehydrogenase (ALDH2) enzymes that majorly degrade acetaldehyde to acetate is extremely weak in the upper aerodigestive tract (253). The resulting inefficient degradation of acetaldehyde may also contribute to additional acetaldehyde exposure among the GG genotype, especially among those consuming moderate to high levels of alcohol (239). Furthermore, apart from ADH enzymes, certain oral microflora can also convert ethanol to acetaldehyde (254-256). Following alcohol consumption, higher levels of acetaldehyde have been found in saliva relative to other parts of the body (especially in individuals with poor oral hygiene) (102, 257, 258). This oral microflora-salivary acetaldehyde pathway can contribute to peak acetaldehyde concentrations among the GG genotype (239, 259). Another hypothesis independent of the acetaldehyde pathway is that the fast metabolism of ethanol may result in lower local exposure (139, 237). Hence, alcohol may not be able to exert its promoter effect (aiding the dissolution of other carcinogens), conferring protection against neoplastic changes in the head and neck region among GA/AA genotypes.

*To summarise, while the evidence for the effect of CYP1A1*2A, CYP1A1*2C, CYP2E1c2, GSTP1 105Val and GSTM1 CNV on SCCHN is inconsistent among Caucasians, the association between CYP2A6*2, CNV in CYP2D6null and SCCHN have not been explored yet. In addition, although these variants are involved in the metabolism of tobacco derived pro-carcinogens and carcinogenic metabolites, a comprehensive characterisation of their interaction (as proposed by recent guidelines) (260) with different levels of smoking incorporating all aspects such as* *interaction on both multiplicative and additive scales, joint effects and stratum-specific risks, has not been reported. Furthermore, because CYP2A6*2 and ADH2B*2 affect tobacco and alcohol consumption behaviours respectively, these behaviours may not only interact but also mediate the causal pathways between these SNPs and SCCHN risk. These pathways have not been quantified yet.*

2.3.3 Human papillomavirus (HPV)

In the past decade, HPV infection has emerged as a strong risk factor for SCCHN. A trend of decreasing incidence of oral cavity cancers (consistent with a decrease in tobacco use), and an increase in the incidence of oropharyngeal cancers (tonsils, base of tongue) have been documented in many developed countries, especially among men (13, 28, 29, 261, 262). The increased incidence of oropharyngeal cancers has been attributed to HPV infection. This infection has been detected in approximately 25% of SCCHN cases worldwide (263), the majority of which are oropharyngeal cancers. This virus is transmitted through skin-to-skin and skin-to-mucosa contact. Hence, unprotected sexual behaviours, notably oral sex, have been identified as routes of HPV transmission with respect to anogenital cancers and SCCHN. More than 100 sub-types of HPV have been identified, among which HPV 16, 18, 31, 33 and 35 have been classified as high-risk sub-types in relation to cancer. More than two-thirds of HPV-positive SCCHN have been attributed to HPV-16 infection. Results from a 2006 meta-analysis show that the association between HPV-16 and SCCHN was strongest for tonsillar (15-fold), followed by oropharyngeal (4-fold), and oral and laryngeal cancers (2-fold) (264). A recent prospective cohort study (2016) conducted in the USA reported an up to 7-fold increase in risk associated with HPV-16 for incident SCCHN cases, with a positive association only for oropharyngeal cancers (265). The researchers also reported that HPV-16 infection preceded SCCHN incidence. HPV-positive SCCHN are clinically distinct from HPV-negative cases and their survival rates are better compared to that of HPV-negative patients (three-year survival of 84% vs. 57%, respectively) (266). Based on recent trends in the incidence of oral cavity and oropharyngeal cancers, the existence of two distinct SCCHN risk groups (tobacco and alcohol related, and HPV related) has been suggested. However, a large study from IARC reported that relative to HPV-negative/non-smokers, HPV-positive/smokers had the greatest risk for both oral cavity and oropharyngeal cancers, greater than

HPV-positive/non-smokers or HPV-negative/smokers (267). Evidence from other studies also indicates interaction between tobacco smoking and HPV status in the risk for SCCHN (268-270).

2.3.4 Socioeconomic position (SEP)

Similar to genetic factors, socioeconomic position (SEP) is a well-documented distal determinant of health outcomes including SCCHN (271-282). In addition, behavioural risk factors such as tobacco use and alcohol consumption are socially patterned (283-289). Hence, whether they are the primary focus or not, it is essential to consider measures of SEP in most epidemiologic studies. In this thesis, different measures of SEP are used, either as the main exposure (manuscript I) or as an important confounder between exposures and the outcome of SCCHN. Therefore, in the following sub-sections, I present an overview of the complex construct of SEP, various methods to measure this exposure and their association with SCCHN risk.

2.3.4.1 Definition and indicators of SEP

Literature identifies social class, status and socio economic position as three entities which are not mutually exclusive (290). Their common feature is the differential access to resources. Social class is related to production and control of production, usually related to occupation. Social status encompasses how others view one's position and includes entitlements and prestige based measures (e.g., education, occupation, and caste[1] in India). SEP refers to the economic and social well-being of a person and is a locus of once resources and endowments. Based on these concepts, correlations with health outcomes, suitability for particular societies and availability of data across the life-course, observational studies use various indicators in an attempt to measure SEP, status or class. This thesis used asset index as a proxy for SEP as it's a measure of one's endowments. An overview of this measure as well as other commonly used measures of SEP are given below.

Asset/wealth index

An asset or wealth index is a measure of the material endowment of an individual or household. It is considered an acceptably reliable proxy for the consumption of goods and services and thus

[1] 1 This includes the castes in the Hindu religion and sections of other religions that have been classified as backward by the state governments of India (here; Kerala) due to discrimination historically faced by them.

SEP, particularly in low to middle income societies (291, 292). The wealth index is calculated using readily-observable household characteristics such as durable assets and household amenities (e.g., owning a car, refrigerator, television, bicycle, livestock, radio, sewing machine), housing characteristics or conditions (household floor, roof, wall material, toilet facilities, water supply), access to services (e.g., electricity supply, drinking water sources), and housing tenure (status of house, land or farm ownership) (292-295). Asset indices were developed based on availability and convenience especially in more agrarian societies and not on a plausible direct causal relationship between wealth or asset possession and health (292). There is also an argument that the index is unlikely to capture the broad concept of SEP (296). However, poor housing is associated with a wide range of health conditions (297). Indicators such as overcrowding in houses have been associated with sanitation and the spread of infections. Moreover, health and mortality are sensitive to fine gradations in neo-material conditions such as access to cars, home ownership, presence of a home garden and healthier food (298, 299). Furthermore, housing tenure, conditions, assets and amenities reflect an individual's educational and occupational status and income (292). The wealth index gained popularity through its use in Demographic and Health Surveys (DHS) data sets to quantify and compare socioeconomic inequalities across approximately 35 countries which mostly included low and middle income countries (291, 300). This measure was utilized because of a lack of reliable data on income and expenditures. Also, household assets are resistant to change in response to short-term economic shocks, which are a feature of low and middle income settings. Based on its slower response to economic shocks, it is also argued that the wealth index captures long term stable aspects of economic status (296, 301). Unlike other indicators such as education and current income, information on components of the wealth index is available across life and hence is an SEP measure available at multiple periods of life.

Education

Education is one of the most widely used individual-level measures of SEP. It marks the transition from childhood to adolescence or early adulthood (302). An individual's educational attainment could determine that individual's health through its influence on decision-making skills, awareness about opportunities, general awareness and interactions with people, access to information and health care, choices of lifestyle behaviours, job and income levels, housing conditions, social status and stress coping mechanisms (33, 303). Relative to other measures of SEP such as income and occupation, education is easier to measure, can be assessed in people who are not in active labour, is equally available to both sexes especially in developed countries, has a high response rate with the exclusion of only a few members of the population and is less subject to negative adult health selection. Together, these attributes make education a useful and important measure of SEP (303-305). However, education is usually acquired early in life and stable after early adulthood, and thus represents SEP only during a short window of the life-course (293, 303). Commonly used markers of education include number of years of formal education and highest level of education attained in life (293, 302). However, the analysis of these markers can be complicated. The number of years of education does not convey any information regarding the quality of the education and its social and economic value. Furthermore, the meaning of a particular level of education and number of years of education are not the same everywhere, and are related to age and birth cohort, social class position, race/ethnicity and cultural norms (290). For example, significant social and educational reforms took place in the state of Kerala in India in the mid-1900s (306). Until that time, a feudalistic system existed for land ownership, wealth, access to education and privileges. Education was considered the privilege of people of the higher caste (hierarchy in the Hindu religion based on occupation) and Syrian Christians, whereas people from the backward caste and most females were denied formal education (306). Completing four years of education was considered a high educational attainment. However, political movements since the Indian independence (1947), especially in the late 1950s, resulted in free and compulsory education until 14 years of age (8 years of education), and education was given a higher importance in the society (307). This educational reform played an important role in lifting people out of poverty by providing the means for upward social mobility. Such features specific to societies and birth cohorts must be considered when using and analysing markers of education as measures of SEP.

Occupation and income

Occupation and income are commonly used measures of SEP. Occupational status is a direct measure of social class in most societies and is the major structural link between education and income (302). Income is a direct indicator of SEP and is the result of an individual’s occupation (308). Occupation plays an important role in positioning an individual within the social structure that directly controls access to resources, interaction with peers, exposure to job related environments and physical exposures, psychological risks and risk behaviours such as tobacco and alcohol consumption (303). Income levels impact health outcomes by influencing the material circumstances of an individual such as quality, type and location of housing, food, clothing, health care, transportation opportunities for cultural, recreational and physical activities, child care and exposure to various toxins (302). Overall, these features make occupation and income suitable indicators of SEP in health research. However, occupation and income can be difficult to measure with precision, especially in low and middle-income societies (276, 291, 292, 301). This can be attributed to issues such as higher non-response rate, missing information on people who are not part of active labour (e.g., home makers) and fluctuations with short term economic shocks (301, 303). Furthermore, most occupational classifications have been developed and validated on working men (303). These factors pose a challenge when using occupation and income as measures of SEP.

2.3.4.2 Association of SEP with risk for SCCHN

As demonstrated with health outcomes such as cardiovascular diseases, mortality, allostatic load, multiple cancers and oral health conditions, cumulative disadvantageous SEP over the life-course has been associated with increased risk for SCCHN, independent of behavioural risk factors (280, 298, 309-313). A large meta-analytical review by Conway et al (2008) on case-control studies that included 24 and 17 studies from high and low income countries, respectively, examined the association between three measures of SEP (income, occupation and education) and oral cancer risk (275). Participants with low educational attainment, low occupational social class and low income had 1.85 (95% CI: 1.60, 2.15), 1.84 (95% CI: 1.47, 2.31) and 2.41 (95% CI: 1.59, 3.65) times the risk, respectively, of developing oral cancer relative to their higher SEP counterparts. In addition, disadvantageous SEP was independently associated with increased oral cancer risk in high and low income countries across the world. Most (277, 314-316) but not all studies (317) conducted subsequently in developed and developing countries have shown that a disadvantageous SEP is independently associated with an increased risk of SCCHN.

2.4 Complex exposures - Need for comprehensive conceptual and analytical framework

Genetic exposures such as SNPs are fixed at birth and are well defined. By contrast, exposures such as behavioural risk factors and SEP have a complex dynamic nature. An individual's SEP may not remain the same from childhood to early to late adulthood stages of their life (284, 293, 318). The situation is similar for behavioural risk factors such as tobacco and alcohol habits, as individuals' behavioural patterns can vary (e.g., frequency, duration, type of tobacco or beverage consumed) over the course of life (319). Thus, these exposures are time-varying. Capturing the dynamic nature of these exposures within an epidemiologic study and addressing it in the analysis is challenging. The challenge is compounded by the bi-directional associations within these exposures at multiple time periods, and between these variables and the health outcome. For example, SEP is considered to affect risk behaviours. However, such behaviours (e.g., alcohol consumption) have also been considered as determinants of socioeconomic consequences, especially in developing societies (320). In addition, these risk behaviours are highly correlated. Hence, SEP in an earlier period of life, for example childhood, may affect risk behaviours in adolescence and early adult life, which can in turn affect social conditions in subsequent late adult life. In short, this time-varying nature produces a complex feedback loop between these variables acting as multiple confounders and mediators in the causal pathways to the health outcome (321). A further concern is the possibility of reverse causality. Based on the social causation perspective, an individual's SEP components can influence their health positively or negatively. For example, following a low educational attainment, one could get a job that exposes them to chemicals and physical hazards including carcinogens, physical and psychological stress, noise, heat, cold, unsafe conditions, and dust, among others. These exposures lead to an increased risk of disease. The same person could also face unemployment, which increases the risk of depression, anxiety and disability, and may lead to unhealthy coping practices (e.g., cigarette smoking and alcohol consumption). In contrast, based on the selection hypothesis, healthy people may obtain and retain their occupational status. These bidirectional associations make collecting repeated data on these exposures at multiple time points and assessing their temporal relationship with the health outcome imperative. Addressing these issues requires a comprehensive theoretical study framework, a study design that is appropriate for the health outcome being investigated, a suitable analytical framework and associated techniques. In this thesis, I used the conceptual framework of life-course epidemiology, a case-control study design that is advantageous to study rare disease outcomes such as SCCHN, a counterfactual causal inference analytical framework to incorporate repeated measures of exposures and causal effects as well as effect decomposition of exposures on the outcome, and causal diagrams. A brief overview of these elements of my thesis are presented in the sub-sections below.

2.5 Life-course epidemiology - Definition and origin

Kuh and Shlomo define life-course epidemiology as “the study of long-term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life” (322). Research in the 1950s by Sir Richard Doll and colleagues suggested that smoking was a strong risk factor for lung cancer (and concomitantly for laryngeal, oesophageal and bladder cancers). This marked a paradigm shift in risk factor research: the focus of chronic disease investigations shifted to an adult lifestyle approach where multiple adult life exposures were implicated in the risk for later life health outcomes (323). However, Forsdahl (1977) documented a strong correlation between infant mortality rates and mortality in middle age for the same generation in specific counties in Norway (324). Similar results linking early life events to adult health outcomes were documented in ecological studies conducted in the USA and Britain, and historical cohort studies (e.g., British birth cohorts) during the following 15 years (325-329). These observations gave rise to the concept of *biological programing* based on the *fetal origins hypothesis.* According to this hypothesis, “environmental exposures such as under-nutrition during critical periods of growth and development in utero may have long term effects on adult chronic disease risk by ‘‘programming’’ the structure or function of organs, tissues, or body systems” (327). In combination, the above observations supported the importance of biological, behavioural, and psychosocial processes that may operate throughout an individual’s life-course, or across generations to influence disease risk, rather than just an adult lifestyle approach to chronic diseases (330). This research became the foundation for the conceptual framework of life-course epidemiology, conceived in the late 1990s, which gives importance to the time (duration) and timing of biological, behavioural and social exposures that may act independently, cumulatively or interactively to influence disease risk (322, 331).

2.5.1 Models under the life-course epidemiology framework

The main aim of the life-course epidemiology framework is to elucidate pathways linking exposures across the life-course to later life health outcomes. To achieve this objective, various theoretical models linking exposures to health outcomes have been proposed. They are described below.

2.5.1.1 Accumulation model

The accumulation model is considered the most fundamental of all life-course models and gives importance to the time (duration) of exposures (332). The model proposes that exposures clustered at different periods of life may accumulate longitudinally over the course of life, leading to differential risk for chronic disease outcomes (331). This concept is in line with the notion of allostatic load, which is the wear and tear on biological systems resulting from chronic over activity or inactivity of normal physiological systems in response to increased exposures (in number and/or duration) from the external environment (331, 333). Indeed, Kuh et al. (1997) describe an individual's biological resources accumulated over the life-course as their 'health capital', which describes and influences current and future health (322). Ben-Schlomo and Khu (2002) propose that risk can accumulate with independent and uncorrelated insults (no interaction between exposures), or with correlated insults (e.g., SEP, smoking, alcohol) that cluster together leading to a health outcome, or similar insults (disadvantageous SEP at different life stages) that form a chain leading to the outcome (331).

2.5.1.2 Critical period model

Stemming directly from the concept of biological programing and fetal origins hypothesis, the critical period model gives importance to the timing of exposures. In its strict sense, the critical period model posits that exposures during specific periods of life can cause irreversible biological damage and have a long-lasting effect on biological systems, irrespective of exposures in prior or later periods of life (331). The sensitive period model is a variation of the critical period model which recognizes that although periods with a higher sensitivity to the effects of an exposure may exist, the effects can be modified or even reversed with prior or later exposure profiles (330).

2.5.1.3 Mobility or pathways model

The mobility or pathways model is considered to be a variation of the accumulation model and is mostly examined in studies with SEP (334). It focuses on the cumulative effect of exposures along life trajectories and implicates differential exposure throughout the life-course in adult disease causation. This model implies the interaction of exposures at multiple periods of life (e.g., SEP in childhood, early and late adulthood). Different hypotheses proposed within the pathways model posit different health effects. For example, under the *natural health selection hypothesis*, less healthy individuals get into a downward mobility (moving from an advantageous to a disadvantageous SEP) and healthier individuals tend to have upward mobility (moving from a disadvantageous to an advantageous SEP) (335, 336). These mobile groups are separated from the individuals who do not show any mobility across life periods as both groups are considered to have distinct traits that make them mobile or non-mobile. In contrast, under a *gradient/health constraint hypothesis,* mobile groups (either upward or downward mobility between different time periods) possess health traits of both the period they leave and the one they join, thus minimizing the health difference between the SEP groups (335-337). The risk associated with mobile groups will be intermediate between the two non-mobile groups (greater than the group with advantageous SEP in all time periods, and lower than the non-mobile groups with disadvantageous SEP at all time points). Interestingly, an elevated risk for a health outcome (e.g., cardiovascular mortality) has been documented among individuals who experience deprivation in early life, followed by later life affluence (324). Forsdahl (1977) hypothesized that this was partly due to risky exposures associated with an affluent lifestyle (e.g., elevation in adult cholesterol levels) (324).

Life-course epidemiology allows considerable overlap between the models specified above. Hence, the models are not mutually exclusive and empirically difficult to disentangle (338). For example, under a social mobility model, a disadvantageous SEP in childhood can interact with an advantageous or disadvantageous SEP in early adulthood to confer a particular risk for SCCHN. However, this is indeed a chain of risk described under the accumulation model. Furthermore, the critical period model with effect modification in prior or later periods (330), or sensitive period model, is reflected in various interactions of the exposure possible under the social mobility concept.

2.5.2 Suitability of the life-course framework to study social, genetic and behavioural risk factors

The life-course framework is particularly well suited for this work exploring genetic, behavioural and social risk factors of SCCHN, as the multiple ways in which exposures can lead to the cancer outcome can be encompassed within this framework. For example, the time-dependent aspect of SEP and associated behavioural risk factors can be effectively captured under this framework and tested under the accumulation, critical and social mobility models. Familial risk factors such as SNPs are already fixed and exert an effect throughout life, which can be visualized under an accumulation model. For example, SNPs such as CYP1A1*2A and CYP2E1c2 increase the risk of SCCHN among Asians independent of smoking. This could be an example of an independent insult causing the health outcome, as explained under the accumulation model. However, the effect of this SNP on the risk of SCCHN among Caucasians might be present only in the presence of heavy smoking (interaction). Yet again, SNPs such as ADH1B*2 and CYP2A6*2 can interact with alcohol and smoking. These also affect alcohol and smoking behaviours respectively. Hence, the effect of these SNPs on the risk of SCCHN can be partly through these risk behaviours, which is referred to as mediation. The concept of interacting and mediating causal pathways leading to a health outcome has been defined under the life-course framework and is reflected in the accumulation model (330). Thus, the possible causal pathways to SCCHN involving potentially confounding, interacting and mediating factors can be tested under the life-course framework. However, this study framework needs to be complemented by a suitable study design incorporating life-course epidemiology to specifically study the relatively rare outcome of SCCHN.

2.6 Study designs for observational epidemiologic studies

Two of the main observational study designs for epidemiologic research are cohort and case-control designs (339). In this thesis, we used a hospital based case-control design and novel approaches to existing analytical techniques originally developed for cohort data. Hence, the principles of these designs are described below with emphasis on case-control studies.

2.6.1 Cohort study design

In a typical cohort study, a group of individuals, sampled based on exposure to certain conditions, are identified and traced over time for the occurrence of health outcomes (340). They can be prospective or retrospective. A commonly used measure of disease frequency is the incidence rate, which is the number of new cases per population at risk in a given time period. The incidence rate can be calculated in both the exposed and unexposed group, from which both absolute and relative measures of association between exposure and outcome can be derived. The difference between the incidence rate in the exposed and that in the unexposed group provides the incidence rate difference (on an absolute scale), whereas the ratio between incidence rates in the exposed to the unexposed group gives the relative risk (RR) (on the relative scale) (341). The calculation of these measures is possible and straightforward in a cohort study, as the probability of the outcome in the non-exposed is known (342). The study design also provides information on multiple exposures and outcomes. Furthermore, prospective cohort studies provide information on variation of these exposures over time, and ascertain temporality (cause precedes effect). However, their time-consuming nature makes them a poor choice to study rare outcomes such as cancers (under a rare disease assumption, health outcomes with a prevalence of less than 10% in the population are considered rare), as following the entire population for long periods of time would be impractical, and the sample would not yield sufficient cases to derive reasonably precise measures of association.

2.6.2 Case-control study design

The case-control design can be advantageous compared to cohort studies, especially when investigating rare disease outcomes, because of its efficient way of sampling individuals from the source population based on the outcome (341, 343). Compared to a cohort study, a case-control design includes a larger fraction of individuals from a source population who develop the outcome (cases) and a lower proportion of those who do not (controls). This design attained significance in the 1920's through studies on rare outcomes such as lip, oral cavity and breast cancers (344). In a case-control study, an adequate number of cases from a source population are first selected and classified as exposed or unexposed. Next, their exposure profile is compared with that of controls, who are sampled from and are representative of (with respect to exposure distribution) the same source population from where the cases were recruited (345, 346). The controls are selected independent of their exposure status. Depending on the method of sampling controls, the two major types of case-control studies are population-based (controls sampled directly from the source population) and hospital-based (controls sampled from the same hospitals as cases) (340)

Because the numbers of cases and controls are fixed by the investigator in a case-control study, the probability of the outcome among the source population remains unknown (342). Hence, relative risk cannot be estimated directly using this study design unless we use techniques to correct for the sampling strategy that gave rise to the data (e.g., correction by the inverse of the probability with which cases and controls are sampled (sampling fraction) into the study from the underlying population) (342, 343). However, since the counts of participants among cases and controls with and without the exposure are available, the measure of association derived from case-control studies is the odds ratio (OR) (342). Basically, the OR is defined as the ratio of odds of the exposure among cases to that among the controls (exposure OR). However, the calculation of the exposure OR and outcome OR are mathematically equivalent, making it a valid measure of association between exposure and outcome (342). For rare outcomes such as certain cancers (incidence of less than 10% in a population), the OR approximates the RR (342, 344).

Although case-control studies are suitable for the investigation of cancer outcomes, the design itself poses challenges with respect to certain research questions. First, unlike cohort studies, data on exposures that change over time (e.g., SEP, smoking) are usually not available from case-control studies. This makes it difficult to assess exposures under a time-varying framework. Second, the estimation of association between exposure and outcome is limited to the health outcome on which the study sampling was based. Hence, a researcher might refrain from exploring research questions that require the use of analytical techniques for which a variable other than the main outcome of interest must be used as a dependent variable (e.g., mediation analysis, multi-step modelling such as inverse probability weighted marginal structural models). However, such scenarios are encountered when research questions aim to elucidate causal pathways and mechanisms underlying exposure-outcome relationships. The case-control design should be explicitly taken into account while answering these questions and appropriate study frameworks such as life-course epidemiology, control sampling techniques and statistical methods are needed to overcome these challenges (343).

2.7 Causal inference and causal effects

"One commonly heard argument is that epidemiologic studies are about association, not causation. According to this proposition, epidemiologists should not worry too much about fishy causal concepts but rather focus their efforts on estimating correct associations. This is certainly a safer strategy but also a dangerous one because it can make much of epidemiology close to irrelevant for both scientists and policy makers". – Hernán (2005) (347)

Information on cause and effect relationships between exposures and health outcomes is the fundamental contribution of epidemiology to the improvement of health (348, 349). Causality and causal inference have been a subject of great interest and contentious debate since the 18th century (350). These concepts further evolved during the 19th century through pioneering works on infectious diseases (Henle-Koch postulates), social causation of disease (Rudolph Virchow), evidence on smoking and various cancers (Richard Doll and colleagues), "Bradford Hill criteria" (1965) and its adaptation into the "Surgeon General's reports" (1964 and 1982) to assess causality, and Rothman's causal pie model (1976) (351-356). Today, causal inference is largely viewed as an exercise in the measurement of the causal effect of an exposure in a target population rather than as a process to be evaluated based on criteria or guidelines (349). This exercise involves: a) defining a clear causal question even if one thinks it is unlikely that estimates will be interpreted as causal, b) choosing causal diagrams, statistical parameters and analytical techniques that help address the causal question, and c) specifying the assumptions under which the statistical parameters we estimate would correspond with the answer to the causal question.

2.7.1 Causal effects under the counterfactual/potential outcomes framework

Apart from substantive knowledge on the outcome and exposures, causal effect estimation requires appropriate causal models/frameworks, causal diagrams depicting assumed relationships between variables, and rigorous analytical techniques based on the study design (357). A statistical association between two random variables X and Y could reflect five possibilities; a) X causes Y, b) Y causes X; c) X and Y have a common cause (confounding), d) random fluctuation, and e) the association was induced by conditioning on a common effect of X and Y (357). Given these possibilities, the statistical association between exposure X and outcome Y can be *defined as causal* if changing the value of X would make a difference in the value of Y, provided nothing else temporally prior to or simultaneous with *X* changed (357). The measurement of a causal effect fundamentally requires contrasting the value of *Y* in the presence of a temporally prior variable *X* (observed) to the potential value of *Y* in the absence (i.e., any other value) of *X* (counter to the fact-unobserved) (358). This understanding, known as the counterfactual concept, originally conceived by Scottish philosopher David Hume in the 18th century, gave rise to the counterfactual/potential outcomes model for causal inference (359). Here, a counterfactual/potential outcome is defined as the outcome *Y* that one would have had, possibility contrary to the fact, under an exposure other than X (359). In an empirical setting, an individual is either exposed or unexposed and one potential outcome is always missing. Hence, although it is not possible to ascertain the causal effect of an exposure on an outcome for an individual, the counterfactual model allows the estimation of the average of individual causal effects in a target population as a parameter in a statistical model using observed data (360). However, this estimation is only possible if three basic identifiability assumptions are met (360): *exchangeability, counterfactual consistency* and *positivity*. Two study groups are *exchangeable* if the probability of the outcome in one group is the same as that of the second group, had the exposures been reversed (i.e., the potential outcome is independent of the exposure). In a well-designed randomized clinical trial (RCT), the exchangeability assumption is met as participants are randomized into groups, which essentially ascertains that the exposure is independent of other covariates and the outcome. *Counterfactual consistency* is the rule that allows the potential outcome to be linked to the observed outcome. It outlines that the potential outcome under the observed exposure is the observed outcome. This assumption is usually considered to be met if the exposure is well-defined and manipulable by intervention (e.g., dose of drugs, dose of specific measure of specific tobacco type rather than dose of tobacco smoking in general) and is violated in the case of under-defined, non-manipulable exposures such as social exposures (e.g., SEP). *Positivity* means that the probability of exposure at every level of all covariates in the model is above 0. Two types of positivity violations are: a) *stochastic or chance positivity violation* in which there is no probability of exposure at a certain level of a covariate due to lower sample size (e.g., genetic polymorphisms with low minor allele frequency), and b) *deterministic positivity violation* in which the individual has no chance of being exposed (e.g., positive exposure to alcohol among non-alcohol consumers).

2.7.2 Causal inference in observational studies

The counterfactual model of causal inference has largely dominated the scientific discourse on the estimation of causal effects in the health sciences since the last century. This model stimulated the development of the randomized trial study design by Ronald A. Fisher, and associated inferential statistics in the 1920's by Fisher, Jerzey Neyman and Egon Pearson (359). Because this study design achieves a valid substitution for counterfactual experience, and the randomization procedure ensures that *exchangeability* and *positivity* assumptions are met, RCTs are the study design of choice to estimate causal effects of well-defined manipulable exposures/interventions on outcomes. However, not all exposures can be manipulated under experimental conditions (e.g., social risk factors) or can be randomized and assigned among humans due to ethical concerns (e.g., smoking). This limitation of RCTs created a need to infer causality utilizing non-experimental observational study designs (e.g., case-control, longitudinal), which have been the mainstay of the majority of epidemiologic studies. However, the greater probability of violating identifiability assumptions in these designs made causal inference from observational studies a challenge. To address this challenge, Rubin (1974) developed the model into a general framework for causal inference that can be applied to non-experimental studies as well, and demonstrated the feasibility of causal inference utilizing these study designs (361).

2.7.3 Causal inference in settings with complex time-dependent feedback loops

As discussed in previous sub-sections, exposures such as SEP and risk behaviours are dynamic and time-varying. These exposures measured at one time point can affect the exposure measured at subsequent time points. Along the way, they can also affect or be affected by other covariates that may bias the causal association between the exposure and the outcome. In other words, time-varying systems are subjected to complex feedback loops that compound the challenge of causal inference. To overcome this problem, James Robins introduced three powerful analytical methods stimulated by the counterfactual framework, under the umbrella term of *G-methods*: the *parametric g-computation formula* (1986), *G-estimation of structural nested models* (1989), and *inverse-probability weighted marginal structural models* (1998) (362-365). These methods made the estimation of causal effects under time-varying feedback conditions achievable with longitudinal data. However, these techniques have not been implemented in a case-control study including a combination of time-varying exposures and confounders or multiple confounders affected by prior exposure. Recent advancements in inferential statistics through the work of Tyler VanderWeele, Stijn Vansteelandt, Miguel Hernan and colleagues have also made the estimation of direct and indirect effects (mediation) as well as the attribution of effects to pathways defined by mediation and interaction (e.g., 4-way decomposition) that may underlie the causal association between exposures and outcome empirically possible with longitudinal data. However, their demonstration within a case-control study is limited and software codes for their easy implementation in commonly used analytical software programs such as Stata are lacking (personal communication- with VanderWeele) (343, 366).

2.7.4 Directed Acyclic Graphs for causal inference

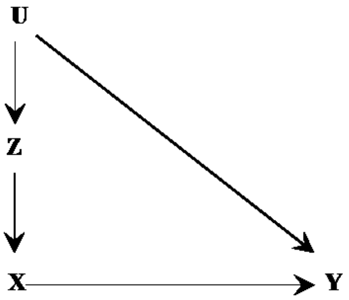
"Epidemiologists are acutely conscious of the danger of over-interpreting associations as causal, and it may be as a consequence of this that they sometimes avoid thinking about the potentially causal nature of associations between exposures of interest and potential confounders. It is all too easy to fall into a purely empirical approach to analysis, where covariates are added to the model one by one and retained if they seem to make a difference. Valid inference would be better served if, perhaps with the aid of causal diagrams, careful consideration were given to whether each factor should be in the model, particularly if the factor may have been caused in part by the exposure under study." Weinberg (1993)(367)

The scientific discourse on causal inference has been supported by a rapid growth in the last two decades in the availability and accessibility of concepts and tools that allow the rigorous and systematic assessment of whether statistical associations are causal. One such important methodological advancement has been the development and increasing adaptation of causal diagrams or directed acyclic graphs (DAGs). The DAG is a graphical tool proposed by Judia Pearl and colleagues, which was introduced into the epidemiology literature in 1995 (368, 369). These graphs are diagrams with formal rules that majorly help in: a) designing epidemiologic studies, b) understanding the causal and non-causal relations among variables related to a specific substantive research question and, c) evaluating structural relationships that may pose a threat to study validity (e.g., confounding, selection/collider bias, information bias) (357). A confounder is most commonly defined as a variable that is 'associated' with both exposure and outcome and is not an intermediate variable between them. Adjusting for traditionally defined confounders when they are in fact non-confounders as revealed through DAGs would induce bias in the estimates (e.g., over adjustment, M-bias) (357). DAGs are extensively used in this thesis to demonstrate underlying causal relations (e.g., time-varying framework, mediation), facilitate various analytical decisions (e.g., identification of confounders for the assessment of total, direct and indirect effects) and explain potential biases (e.g., confounding, selection bias). To facilitate their understanding in the Methods section of this thesis, I describe below the basic terminology, rules and concepts underlying DAGs, structural definitions of confounding and selection bias, steps to follow to estimate the total effect of an exposure on the outcome using DAGs and the special case of time-varying confounding affected by prior exposure.

2.7.4.1 Basic DAG terminology

A DAG consists of a set of random variables (nodes or vertices), both measured (e.g., X, Y, Z in DAG 1) and unmeasured (typically represented by U as in DAG 1), each variable pair connected by a single arrow (directed edges).

Figure 3: Hypothetical DAG 1



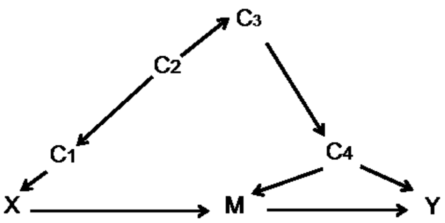
X is the exposure, Y is outcome, Z is a measured variable. U is an unmeasured variable.

The graph is directed as each arrow has only one arrowhead and points from one variable to one other variable. It is also acyclic as no variable can cause itself either directly, or through other variables. An exception to this is a time-varying variable, for which an arrow from the variable measured at one point in time (e.g., SEP in childhood) can point to the same variable measured in a subsequent point in time (SEP in early or late adulthood). Unlike traditional confounder diagrams in which there is uncertainty in the meaning of the arrows used (i.e., whether the arrow represents association, prediction or causation), each arrow in a DAG depicts causation (as per the definition of *cause* provided in *sub-section 2.7.1, page no. 42*). The variable from which an arrow originates (parent) is a direct cause (causative or preventive) of the variable to which the arrow head leads (descendent). In DAG 1, X is a direct cause of Y. Similarly, Z is a cause of X, and U of Z and Y. All common causes of any pair of variables must be included in a causal graph. The arrows do not specify the magnitude or direction of causation.

2.7.4.2 Paths in DAGs

Each path in a DAG goes between the exposure and the outcome without passing through a node more than once. A path can be open or closed. Open paths have an expected causal association flowing along them (e.g., path 1 in DAGs 2, ***Figure 4***).

Figure 4: Hypothetical DAG 2



Paths	Type	Status
1. $X \rightarrow M \rightarrow Y$	Causal	Open
2. $X \leftarrow C_1 \leftarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow Y$	Non-causal	Open
3. $X \leftarrow C_1 \leftarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow M \rightarrow Y$	Non-causal	Open
4. $X \rightarrow M \leftarrow C_4 \rightarrow Y$	Non-causal	Blocked at M

Some paths are open naturally, that is, prior to the intervention of the researcher. Causal paths are naturally open paths in which all arrows point in the same direction from exposure to outcome either directly (e.g., X→Y in DAG 1) or through multiple intermediates (e.g., X→M→Y in DAG 2). All such causal open paths contribute to the total effect of the exposure on the outcome. Such paths can however be closed mistakenly by conditioning (restriction or matching by study design, stratification or covariate adjustment in statistical models during analysis) on the intermediates/mediators. For example, conditioning on M closes the open causal path between X and Y in DAG 2, creating a biased estimate of the total effect of X on Y. On the contrary, certain non-causal paths can be left open naturally (e.g., paths 2 and 3 in DAG 2). Such paths can be used to structurally define confounding paths and naturally include variables that are common causes of the exposure and the outcome (e.g., C2 in path 2 of DAG2). These paths create a bias and the expectation of an association between exposure and outcome that is non-causal. This bias is termed *confounding* and can be removed by conditioning on any variable along non-causal naturally open paths (e.g., conditioning on either C1 or C2 or C3 or C4 can block the non-causal naturally open path 2).

Closed paths are those through which no association flows; they are considered blocked either naturally or by conditioning on variables along them (e.g., conditioning on a confounder makes an open non-causal path closed). For example, path 4 (X → M ← C4 → Y) in DAG 2 is blocked at M which has arrows originating from C4 and X colliding on it. M is termed a collider on path 4. Conditioning on a collider can mistakenly open the blocked non-causal path, creating a biased association to flow between exposure and outcome, and is considered a selection bias. It should be noted that a collider is path-specific. Also, a variable can have different meanings depending on the path. For example, in DAG 2, M is a collider on path 4, but not on paths 1 (X → M→ Y) and 3 (X ← C1 ← C2 → C3 → C4 → M → Y); M is a mediator on path 1 (X → M→ Y), but not on paths 3 and 4. M is a confounder on path 3, but not on paths 1 and 4.

2.7.4.3 Minimally sufficient set of confounders

A set of variables on which conditioning leaves all causal paths open and all non-causal paths blocked is referred to as a sufficient set of confounders. A minimally sufficient set is a sufficient set of which no proper subset is sufficient.

2.7.4.4 Steps to estimate the total effect of an exposure on the outcome using DAGs

To estimate the total causal effect of an exposure on an outcome, 5 steps should be followed: 1) draw a DAG based on the best available data; 2) find all the paths between the exposure and the outcome; 3) separate the causal and non-causal paths; 4) separate open and closed paths; and 5) find the minimally sufficient set(s) of conditioning variables.

In the case of DAG 1, the minimally sufficient set of conditioning variables to estimate the total effect of X on Y will be {Z} as shown in ***Figure 5***. Although U, an measured variable, is also in the confounding path, conditioning on Z turns a confounder U into a non-confounder.

Figure 5: Minimal sufficient set of confounders identified from hypothetical DAGs 1 and 2 to estimate the total causal effect of an exposure on the outcome

Paths	Type	Status
1. $X \rightarrow Y$	Causal	Open
2. $X \leftarrow [Z] \leftarrow U \rightarrow Y$	Non-causal	Blocked

From DAG 1

Paths	Type	Status
1. $X \rightarrow M \rightarrow Y$	Causal	Open
2. $X \leftarrow C_1 \leftarrow C_2 \rightarrow [C_3] \rightarrow C_4 \rightarrow Y$	Non-causal	Blocked at C_3
3. $X \leftarrow C_1 \leftarrow C_2 \rightarrow [C_3] \rightarrow C_4 \rightarrow M \rightarrow Y$	Non-causal	Blocked at C_3
4. $X \rightarrow M \leftarrow C_4 \rightarrow Y$	Non-causal	Blocked at M

From DAG 2

For DAG 2, multiple minimally sufficient sets are possible; e.g., {C1} or {C2} or {C3} or {C4}. ***Figure 5*** depicts conditioning on C3, leaving the only causal path 1 to be open. The selection of any of these 4 variables for conditioning depends on whether the variables have missing data, measurement error or specification error. Although M in DAG 2 is in the confounding path 3, conditioning on it will close the only open causal path between X and Y (M is a mediator in path 1) and will open a blocked non-causal path (path 4).

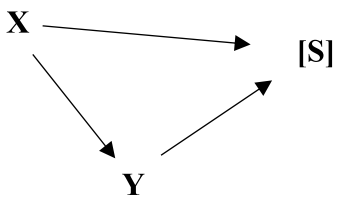
2.7.4.5 DAGs and model selection to estimate the total effect of different exposures

DAGs also indicate if a separate statistical model is required to estimate the total effect of different exposures. For example, because the ADH1B*2 variant affects alcohol use, the latter is a mediator between the genetic variant and SCCHN risk. Let Z, X and Y in DAG 1 represent ADH1B*2, alcohol use and SCCHN respectively. The total effect of alcohol on SCCHN can be estimated from a model where SCCHN is fit on alcohol and ADH1B*2. However, the estimate of ADH1B*2 from this model will not represent the total effect of ADH1B*2 on SCCHN. Estimating the total effect of the variant on SCCHN would require fitting another model that does not contain alcohol use, as the DAG clarifies that including alcohol in the model will block the only causal path between ADH1B*2 and SCCHN.

2.7.4.6 Structural definition of selection bias

DAGs also help us to structurally define selection bias as any bias occurring due to conditioning on the common effect of the exposure and the outcome ***(Figure 6)*** (370).

Figure 6: Structural representation of selection bias

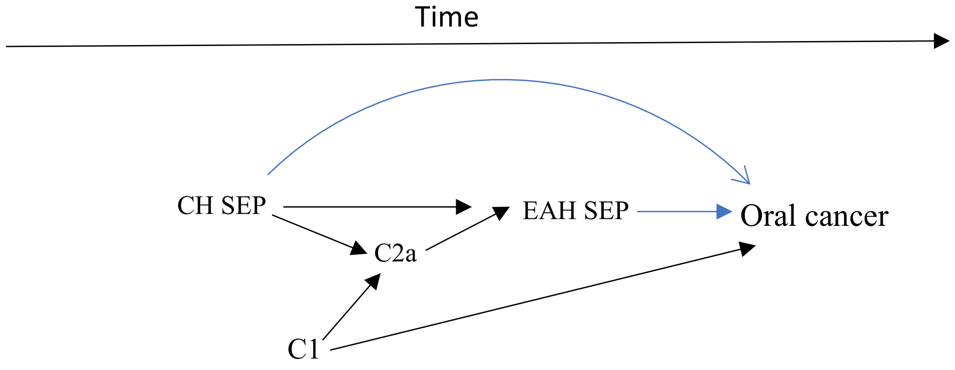


X is the exposure, Y is outcome, S is a measured variable, conditioning on which can lead to selection bias.

2.7.4.7 Time-varying confounding affected by prior exposure

Bias due to confounding and selection bias is compounded when attempting to estimate the total effect of a time-varying exposure in the presence of time-varying confounding affected by prior exposure (i.e., covariates can act as both confounders and mediators) (370). A hypothetical time-varying situation involving SEP in childhood (CH SEP), early adulthood (EAH SEP), confounders measured during childhood (C1) and early adulthood (C2) under a specific temporal relation with respect to the outcome (oral cancer) is depicted in ***Figure** 7*.

Figure 7 : Causal graph representing time-varying confounders affected by prior exposure



Any method that involves conditioning on C2a to estimate the magnitude of the blue lines may induce bias by creating a non-causal association between CH SEP and oral cancer through the path CH SEP → C2a ←C1→ oral cancer (i.e., by opening this naturally blocked non-causal path) (371). However, not adjusting for C2a results in an open non-causal path between EAH SEP ← C2a ← C1→ oral cancer and thus a confounded causal association between EAH SEP and oral cancer. This situation arises because the effect of EAH SEP on oral cancer is confounded by C2a, and C2a is affected by CH SEP (prior exposure); in other words, time-varying confounding affected by prior exposure. Such situations can only be addressed using g-methods described in *sub-section 2.7.3, page no 44.*

2.7.5 Causal interaction

This sub-section gives a brief overview regarding the concept of causal interaction, that is the focus of manuscript II in this thesis. Furthermore, gene-environment interaction is one of the pathways explored in manuscript III. Interaction is defined as the variation in the effect of an exposure on the outcome due to presence or absence of a second exposure or levels of exposure (260). It is an important concept to understand 'whom' among the population with the risk factors are at altered risk for the outcome. Interaction can be measured on both multiplicative and additive scales (343). Multiplicative interaction is the ratio of relative risks (RR) or odds ratios (OR) and can be defined as the ratio of RR or OR of the jointly exposed group to the product of RR or OR of the singly exposed groups. Multiplicative interaction can be supra multiplicative/positive, or sub-multiplicative/ negative. Additive interaction on the other hand measures interaction on the additive scale. It measures how much the joint effect on the difference scale of the two exposures is greater or lesser than the sum of the effects of the two exposures. The additive interaction can be supra/positive, or sub/negative additive interaction. If both exposures have an affect on the outcome, then there will be interaction on either additive or multiplicative scale (340). Additive interaction is expressed through indicators such as relative excess risk due to interaction (RERI), synergy Index (S) and attributable proportion (AP) (343). Of these, RERI is considered to be a stable measure of additive interaction. There can be a positive interaction on the additive scale but a negative interaction on the multiplicative scale or vice versa. However, interaction on the additive scale is the one of policy relevance as, a) it indicates the correct sub-group to treat in both longitudinal or case-control studies, b) multiplicative interaction can indicate the wrong sub-group to treat, c) on a risk difference scale can indicate how many individuals can be treated if intervened (343). However, interaction on multiplicative/relative scale can be important for comparing estimates over time.

It has been documented that there is lack of studies undertaking comprehensive assessment of interaction, with most studies inferring interaction from presentation of stratum-specific estimates with no meaningful comparison across the strata, comparison of p-values between strata, statements on statistical significance without proper statistical tests, assessing overlap of confidence intervals around effect estimates in strata, reporting of interaction estimates without associated confidence intervals, and adjustment for the second exposure in estimating the main effect of the primary exposure (431). Recent guidelines propose that a comprehensive reporting of interaction should consist of reporting joint effects of both exposures with reference to a singe reference group (e.g., doubly unexposed group), effect of the primary exposure in the strata of the secondary exposure (stratum specific effects) and interaction on both additive and multiplicative scales with associated confidence limits (260, 431).

2.8 Summary, rationale and challenges

SEP, genetic variants involved in the metabolism of environmental carcinogens, smoking and alcohol risk factors have been widely studied with respect to risk for SCCHN. However multiple substantive research questions that could provide further insights into the causal pathways underlying the relationship between these exposures and SCCHN remain unanswered. For example, an inverse association between cumulative SEP and oral cancers has been documented in both developed and developing countries. However, SEP varies over the life-course of an individual, a characteristic well documented, but consistently overlooked by SEP-oral cancer studies. Ignoring the basic nature (static vs dynamic) of a risk factor hampers effort to better understand its effects and also results in erroneous inferences on causal mechanisms. Appreciating the time-varying nature of SEP provides us with a unique opportunity to explore deeper into the granularity of its cumulative effect over life on oral cancer. For example, there may exist critical periods in the life of individuals during which exposure to disadvantageous SEP may have a significant impact, or SEP exposure in multiple stages of an individual's life may interact with each other, leading to differential risk for oral cancer incidence in adult life. However, these questions, that can be addressed by testing multiple life-course models within a single epidemiologic study, has not been explored yet. Exploring the SEP-SCCHN association using multiple life-course models may be of special relevance to developing countries such as India were relatively high socio-economic disparity and rapidly increasing oral cancer incidence have been documented.

Genetic variants in candidate genes involved in the bio-activation (e.g., CYP1A1, CYP2E1, CYP2A6, CYP2D6) or detoxification (GSTP1, GSTM1) of environmental carcinogens have been widely investigated as potential SCCHN risk factors. These carcinogens are abundant in tobacco, which is the main risk factor for SCCHN in developed countries such as Canada. While several meta-analytical reviews support a positive association between some SNPs (e.g., CYP1A1*2A, CYP1A1*2C, CYP2E1c2) and SCCHN risk among Asian populations, this association has not been observed among Caucasians. In addition, the evidence for the role of other SNPs (e.g., GSTP1 105Val) in this association is conflicting. The distribution of these genetic variants is ethnicity/population-specific and no studies have yet documented the association between these genetic variants and SCCHN among a Caucasian population in Canada. Furthermore, because these variants alter the kinetics of the metabolism of tobacco derived carcinogens, the estimation of main effects of these variants alone may mask the potential differential risk due to interaction between these genetic variants and various levels of smoking. However, there is a lack of studies undertaking a comprehensive assessment of gene-environment interaction effects, which has the potential to identify high-risk groups in the target population. In addition, these effects, in relation to CYP2A6*2 SNP, CNV in CYP2D6 and SCCHN have not yet been reported in the literature.

The causal effect of genetic variants involved in tobacco and alcohol metabolism on SCCHN risk are not only limited to interaction, but may also involve mediation by associated risk behaviours. For example, CYP2A6*2 and ADH1B*2 may affect smoking and alcohol behaviours respectively, which may in turn lead to altered risk for SCCHN. However, the potential indirect and direct effects possible under this mediation hypothesis has not yet been quantified.

Attempting to address these gaps in the literature primarily requires conceptualizing these causal pathways within a study framework such as life-course epidemiology that integrates the time-varying nature of exposures and confounders, conceptual models such as critical period, accumulation and social mobility, and concepts of interaction and mediation. Primarily, tackling the questions related to SEP and SCCHN requires the exploration of multiple conceptual life-course models under the time-varying framework. Addressing the questions related to genetics and risk behaviours necessitates the quantification of interaction and mediation pathways, assuming an accumulation of risk associated with genetic variants and risk behaviours all described under the life-course framework. Secondly, there is a need for careful consideration of the temporal relationships between multiple exposures, confounders, mediators and SCCHN. This necessitates the use of causal diagrams with formal sets of rules such as DAGs, an increasingly used causal analytical tool for the identification of a minimal sufficient set of confounders and mediators to anchor the analytical strategy, construct analytical models and mitigate potential biases. Thirdly, the quantification of these causal pathways requires that life-course epidemiology be complemented with strong analytical frameworks such as the counterfactual causal framework. Methods described under this framework are superior in handling time-varying exposures, confounders affected by prior exposure, and mediation in the presence of interaction between the exposure and mediator. Utilizing these techniques, one can estimate measures of both mechanistic and policy relevance. For example, the recently developed four-way decomposition technique, enables deciphering the total effect of an exposure on an outcome, in the presence of a third variable that not only mediates but also interacts with the main exposure, into four non-overlapping pathways: a) the pathway that is completely independent of the mediator, b) a pathway that is due to interaction but not mediation, c) a pathway that is due to mediation but not interaction, and d) a pathway that is due to both mediation and interaction. The quantification of these pathways enriches our understating of causal pathways as well as allows us to estimate how much of the risk from the exposure (e.g., ADH1B*2) on the outcome (SCCHN) can be eliminated, if one intervenes on the modifiable mediator variable (alcohol use).

The aforementioned conceptual and analytical techniques mainly suit longitudinal data, whereas rare disease outcomes such as SCCHN are best studied using a case-control design. Hence, integration of life-course epidemiology and the counterfactual causal framework within the case-control study design poses challenges and requires the adaptation of these methods, as well as the availability of software codes for the case-control design.

Chapter 3

Study objectives

The overall objective of the study is to gain further insights into the causal pathways underlying the relationship between SEP, specific genetic polymorphisms, risk behaviours and SCCHN risk. The specific aims of manuscript I, focusing on SEP and oral cancer among an Indian population, and manuscripts II and III, focusing on genetic and behavioural risk factors, are given below.

Manuscript I:

By appreciating the time-varying nature of SEP and associated behavioural confounders,

a) To estimate the extent to which SEP measured over three periods of life is associated with oral cancer risk using data from a case-control study conducted in India.

b) To assess whether the associations conform better to a critical period, accumulation or social mobility model.

Manuscript II:

a) To estimate the total effect of variants in CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTM1 and GSTP1 genes on SCCHN risk in a sample of Caucasians from Montreal, Canada.

c) To estimate the causal interaction between the above-mentioned variants and multiple levels of smoking on SCCHN risk in this target population.

Manuscript III:

a) To estimate the extent to which the effects of two functional SNPs in CYP2A6 and ADH1B on SCCHN risk are mediated by heavy smoking and alcohol consumption respectively, in a case-control sample of Canadian Caucasians.

b) To estimate proportions of excess risk of SCCHN due to each of these genetic variants that are attributable to, (i) both mediation and interaction, (ii) only interaction, (iii) only mediation, and (iv) neither mediation nor interaction, with their associated risk behaviours.

Chapter 4

Methods

This dissertation comprises three manuscripts, each addressing one specific objective of this work utilizing data from an international collaborative study. Although all study sites followed the same study protocol, that is, each site had a similar overall study design and data collection procedures, the distribution and types of risk factors at each study site, variables used in each manuscript and statistical analyses performed to achieve the objectives were different. The overall study design, data and sample collection procedures, as well as specific methodologies for each manuscript are explained in the sections below.

4.1 Overall study design

The Head and Neck Cancer (HeNCe) Life study is an international multi-center hospital based case-control study investigating the aetiology of SCCHN focusing on social, psychosocial, lifestyle, biological and genetic factors, using the life-course framework. This collaborative study was conducted in Canada, India and Brazil. Manuscript I uses data from the Indian site where the incidence of SCCHN, especially oral cancers, is on the rise, and where large social inequalities have been reported (372, 373). Manuscripts II and III rely on data from the Canadian site, where genetic data were available and smoking and alcohol have been the strongest risk factors for SCCHN. Although study sites followed similar protocols, study instruments were culturally adapted through multiple pilot studies.

4.2 Target populations and samples

The target populations for the studies were male and female adult residents of Malabar region of Kerala in India, and Greater Montreal area in Canada. The eligibility criteria of the study were: (i) English, French or Malayalam (Kerala native language) speaking; (ii) to be born in India or Canada; and (iii) to live within a 150 or 50 km radius from the recruiting hospitals in Calicut

(Kerala) and Montreal, respectively. In addition, the participants shouldn't have had any: (iv) previous history of any type of cancer or cancer treatment; (v) mental or cognitive disorders; (vi) communication problems (e.g., inability to speak because of lesions); and (vii) diseases related to immuno-compromise (e.g., HIV/AIDS). Lastly, participants who were too sick or in palliative care were not eligible to participate

In India, cases (N=350) were recruited from the oral pathology clinic at the Government Dental College, and from the cancer outpatient unit of the Government Medical College, Calicut, Kerala, India between 2008 and 2012. Controls (N=371) were recruited from other outpatient clinics in these intuitions during the same study period.

In Canada, cases (N=460) were recruited from Ear, Nose and Throat (ENT) and radio-oncology clinics of four major referral hospitals in Montreal (Jewish General Hospital, Montreal General Hospital, Royal Victoria Hospital, and Notre-Dame Hospital) between 2005 and 2013. Controls (N=458) were recruited from other clinics in the same hospitals.

4.3 Case definition and selection

Incident cases diagnosed with stage I to IV histologically confirmed squamous cell carcinomas of head and neck region, which included cancers of the tongue, gum, floor of the mouth, and other locations in the mouth, oropharynx, hypo-pharynx and larynx (C01-C06, C09, C10, C12- C14, and C32, under the International Statistical Classification of Diseases, 10 Version: 2010), were eligible for this study. Lip (C00), salivary gland (C07-08) and nasopharyngeal (C11) cancers were excluded due to their different aetiologies (374-376). For logistic reasons, only oral cancer cases (C01-C06, and C09 under International Classification of Diseases 10 Version: 2010) were recruited at the Indian site.

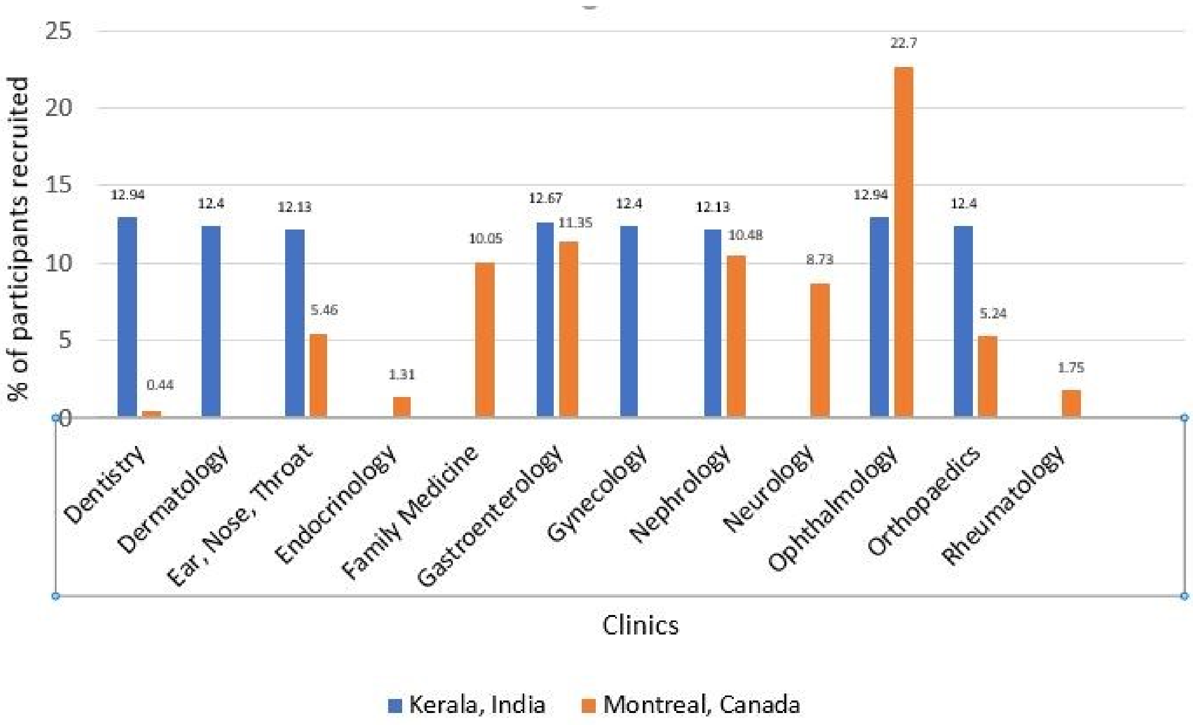
4.4 Control definition and selection

An incident density sampling technique was followed where controls where recruited during the same time period (approximately same risk set) as the cases arose, throughout the course of the study. Non-cancer controls were frequency matched to each identified case by 5-year age group and sex. Controls were sequentially selected from several outpatient clinics that were not typically associated with smoking or alcohol consumption to mitigate Berkson's bias (377). The participation of controls from each clinic was restricted to less than 20% to limit overrepresentation of a single diagnostic/disease group (378). The genetic profile of the participants was not known during recruitment. The list of clinics from which control participants were recruited and the distribution of controls at Indian and Canadian site are given in Figure 8.

4.5 Ethics approval and informed consent

Prior to the commencement of the study, ethical approval was obtained from McGill University's Institutional Review Board (IRB), from Institut National de la Recherche Scientifique (INRS), as well as from all participating hospitals at the Canadian site. At the Indian site, approval was obtained from the IRB and ethics committee of Government Dental and Medical College and Hospitals, Calicut, Kerala. The study procedure was explained to each participant before the start of the interviews and all participants signed informed consent forms before participating in the study (*Appendix 1, page no: 246*). At the Indian site, thumb impressions were obtained from illiterate participants in front of a witness. It was made sure that the research assistant, who explained the study procedures, also signed the consent forms in the presence of the participant and the witness. One copy of the consent form was given to the participant while the other was stored at the study sites. Data were entered into a secure password enabled server, and all participants were given study IDs that were used in all analyses in order to conceal their identities.

Figure 8: Percentage of control participants recruited from participating clinics at Indian and Canadian site



4.6 Participation rate

The participation rate was 54% for controls and 47% for cases in Canada, where as in India the rate was 85.6% and 44.3% among cases and controls, respectively.

4.7 Data collection

The data collection procedures consisted of (i) questionnaire based interviews and (ii) Biological sample collection.

4.7.1 Questionnaire based interviews

One-on-one semi-structured interactive interviews administered by trained interviewers were conducted using a questionnaire and life-grid technique in tandem. The questionnaire (*please refer to Appendix II, page no. 277 for questionnaire used at Indian site)* collected information on several domains of exposures including socioeconomic (e.g., education, occupation, housing conditions), health related behavioural factors (e.g., cigarette, cigar, pipe, bidi smoking, paan chewing, alcohol consumption, diet, sexual behaviour), oral health status, family and environment across multiple stages of an individual's life. For example, data were collected on various housing assets of participant's longest place of residence during 3 periods of their life; childhood (1-16 years), early adulthood (17-30 years) and late adulthood (31 years and above). Information on tobacco and alcohol habits during the entire life of the participant were collected as multiple periods during which where their consumption pattern remained stable. During interviews in Indian site, the help of a proxy respondent was sought for consenting participants who had difficulty speaking due to their disease. The proxies were usually spouses or close relatives who accompanied the participant to the hospital. The questionnaire was developed based on previous studies including British cohort studies - British Civil Servants, Whitehall II (379), British Birth Cohort (BBC) 1946 (326) and BBC 1958 (380) – and studies conducted by the International Association for Cancer Research (IARC). To improve the reliability of the retrospective data collected, a life-grid tool was used interactively with the questionnaire (*please refer to Appendix III, page no. 307*). This instrument helps the participant to recollect past information more precisely by relating them to important events in their life-course (e.g., graduation, marriage, birth of a child, land mark events in the country or region of residence) (381, 382).

In Canada, the research instruments were first developed in English and translated into French and back-translated to confirm the equivalence of the two versions. This was followed by a pilot study conducted among 30 patients at the Montreal Jewish Hospital. In India, the research instruments, which were translated into Malayalam (local language) and back-translated into English, were tested in two pilot studies (between 2006 and 2008) conducted among the target population. Based on the results of the pilot studies, the research instruments were refined before being used in the main study.

Details on the medical history of the cases (e.g., diagnosis, tumour site, TNM staging) and controls (general health information, reason or hospital visit and diagnosis if available) were collected from medical charts.

4.7.2 Biological sample collection

Following the interviews, biological samples were collected from each participant to perform genetic and HPV analyses (383). Although we collected biological samples from both sites, genetic analysis of all samples has been completed only for the Canadian site (no data yet from India). Hence, manuscripts II and III, in which genetic variants associated with bioactivation and detoxification of environmental carcinogens were the main exposures, used the Canadian data. Exposure to HPV was used as a potential confounder in these manuscripts.

Oral epithelial cells, a reliable source of genetic material and HPV DNA, were collected through a validated protocol using mouthwash, and brush biopsies (383-386). The latter was used to collect epithelial cells from the lesion (in cases) as well as normal mucosa in the oral cavity and oro-pharyngeal areas (both cases and controls) (*details of biological specimen collection are available in Appendix IV, page no. 308*) (385). Both mouth wash and brush biopsy methods are simple, non-invasive, inexpensive, and have high acceptance rate among participants. Also, these methods provide great yields of both human DNA and HPV-DNA after purification (383, 387-390). Following collection, the samples were stored at 4ºC as soon as possible and at -20ºC at the sample analysis site. For the Canadian participants, genetic analyses and HPV detection were performed at laboratories at the Albert Einstein College of Medicine in New York, and the CHUM in Montreal, respectively.

4.7.3 Genotyping analysis for DNA polymorphism

To identify SNPs and CNV, genotyping was performed on DNA samples isolated from the mouthwash and brush biopsy samples using real time polymerase chain reaction (PCR) -Taqman gene expression assays. To perform a 10µl Taqman assay, 10 ng of sample DNA was used. Then reactions were set up using 5 µl of the 2X Genotyping Master Mix (Applied Biosystems, Foster City, CA) and assay-specific concentrations of primers and probes. All reactions were set up in a

0.1 ml 96-well plate and sealed with optical cover film (Applied Biosystems). They were spun down at 2000 rpm for 2 min and run in a 7500FAST real-time PCR thermocycler in genotyping mode using default settings. 7500FAST v2.0.6 software was used for allelic discrimination. Polymorphisms identified using validated SNP specific assays (ABI assays) were GSTP1 105Val (rs1695) C_3237198_20, ADH1B*2 (rs1229984) C_2688467_20, CYP2E1c2 (rs3813867) C_2431875_10, and CYP2A6*2 (rs1801272) C_27861808_60. The TaqMan SNP assay for CYP1A1*2C (rs1048943) were specifically designed and manufactured by Applied Biosystems for this study. We ordered all probes from Applied Biosystems, including one that was specifically designed for CYP1A1*2A (rs4646903).

For copy number analysis, the Quantitation-Standard Curve type was used to run duplex reactions in triplicate. Each 10 µl reaction contained FAM-labeled probe for the target gene and VIC-labeled RNase P gene probe to serve as a reference. 5 µl of 2X Universal TaqMan (TaqMan Fast Advanced Master Mix #4444557), no AmpErase UNG master mix (Applied Biosystems) and 0.5 µl of each probe-primers mix assay were added to the total 10 µl reaction with 10 ng of template DNA. Using validated Copy Number assays (ABI assays), copy numbers were identified for GSTM1 and CYP2D6 non-functional (null) variants. TaqMan RNase P assay (Applied Biosystems) was used as a reference for the copy number assays. Data were then imported into CopyCaller v2.0 software (Applied Biosystems) and copy numbers were computed with their probability values.

4.7.4 HPV detection

HPV DNA detection was performed using a standardized PCR protocol (391, 392). The samples were centrifuged (at 1000 x g for 10 minutes), the DNA was extracted from the pellet with a small quantity of supernatant by a modified Gentra Purgene protocol (393). The purified DNA underwent PCR and amplification. To ascertain the integrity of DNA and that there was sufficient sample available for PCR analysis, beta-globin testing was performed. An absence of beta-globin meant that there was insufficient biological material for typing. Samples with positive beta-globin were amplified with primers for HPV and typing (by dot blot assay using radiolabeled probes) was done for HPV -6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 45, 51, 52, 53, 54, 55, 56, 58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 72, 73, 81, 82, 83, and 84 (386, 394)

4.8 Data quality control and management

Interviewers at both sites were trained extensively prior to the commencement of the studies. The training used a manual of procedures and a DVD that explained the data collection steps. Mock interviews were conducted under the supervision of the research coordinator or principal investigator to ensure the procedures were carried out appropriately. Following an interview, the responses were cross-checked with information entered in the life-grid tool. Questionnaires were reviewed by the interviewer and research coordinator in Canada, and by all research assistants at the Indian site to ensure data quality and completeness. To test the reliability of the data collected re-interviews were conducted for 46 randomly selected participants at the Indian site, 6 to 12 weeks after the original interview. At the Canadian site, 12 participants and their 12 siblings matched on age (± 5 years) took part in a validation study evaluating the accuracy of data collected for several variables including housing conditions, education, occupation, and parental education.

Hand-written questionnaire data were entered into a FileMaker (FileMaker Inc.) database using screen layouts that mirrored those of the paper questionnaire to minimise data entry errors. The genetic and HPV data were entered into Microsoft excel files. All data were later exported into the statistical software Stata (Stata, version 13 SE, StataCorp. 2013, College Station, TX: StataCorp LP.) and checked for errors (e.g., missing values, mismatches, inconsistencies, unreasonable values). Errors were fixed by crosschecking with the hand-written questionnaire data before and at various stages of data analysis. Further analysis of the data for each manuscript were performed using Stata, version 13 SE, and SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina).

The above sub-sections presented aspects of the study shared by the Indian and Canadian sites. Below sub-sections describe study measures, analytical frameworks and statistical analyses specific to each manuscript.

4.9 Measures - Manuscript I

In manuscript I, we investigated the association between SEP collected at three periods of the participants' lives and oral cancer risk using the accumulation, critical period and social mobility life-course models. The dependent variable (oral cancer), main exposure (SEP) and potential confounders are described below.

4.9.1 Dependent (outcome) variable – Oral cancer status

Based on the revised ICD classification (ICD 10), oral cancer (C00 – 06) is defined as cancer affecting the lips, tongue, gums, floor of the mouth, palate, cheek mucosa, vestibule of mouth, and retro-molar area. Cancers of the lip were excluded. The diagnosis for squamous cell carcinomas of oral cavity was confirmed using histopathology, which is the gold standard for the diagnosis of malignant lesions (395). The research assistants who collected the data at Indian site (including the 1st author of this manuscript-PhD candidate) were trained dentists who could confirm the status (i.e., presence or absence) and site of the cancer by comparing histopathological information, clinical presentation and site of the lesions. The outcome status, was coded as 1 for cases and 0 for non-cases (controls). This binary variable was used as the dependent variable.

4.9.2 Independent (main exposure) variable: Socioeconomic position (SEP)

As described in sub-section 2.3.4, various measures of SEP have been used in epidemiological research based on theoretical and practical considerations. For this study, the main exposure was an asset/wealth index as, it is a suitable indicator of SEP, especially in developing countries such as India (*please refer to sub-section 2.3.4.2, page no. 31-32, for more details*).

4.9.2.1 Asset/wealth index and principal component analysis (PCA)

The asset/wealth index was created from a list of questions on various assets (housing characteristics, durable assets and access to services) available at the participant's longest place of residence during three time periods: childhood (0-16 years), early adulthood (17-30 years), and late adulthood (above 30 years). I used information on nine assets/items from childhood, eleven from early adulthood and twelve from late adulthood periods (*please refer to manuscript I, supplemental material, eTable 1, page no.121*).

An issue in using housing indicators (which are all correlated) is that each of them could have a different relationship with SEP and may not be sufficient to differentiate household SEP when used individually (291). Hence, different indicators are aggregated to derive a uni-dimensional measure that can be further categorized to reflect different levels of SEP. Summing up the indicators is a common practice (396). However, this assumes an equal weight for each indicator. In this study, we overcame these challenges using principal component analysis (PCA), which is an increasingly employed (e.g., World Bank, Demographic and Health Surveys data sets) data reduction method for creating uni-dimensional SEP measures from data on different assets (291, 292, 301, 397).

Principal component analysis

With PCA, multiple original variables can be summarized with relatively few dimensions that capture the maximum possible information (variation) from the original variables. Mathematically, from an initial set of *n* correlated variables (original), PCA creates uncorrelated components, where each component is a linear weighted combination of the original variables (398). For example, if $X_1, X_2, \ldots, X_n$ are n original indicators, then the first component (PC_1) is given by,

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \ldots. + a_{1n}X_n$$

and m^th component is given by

$$PC_m = a_{m1}X_1 + a_{m2}X_2 + \ldots. + a_{mn}X_n$$

Where a_{mn} is the weight for the m^{th} principal component and the nth variable.

Since PCA aims to maximize the variance, it is sensitive to scale differences in the original variables. For example, in our study, responses to some of the questions on housing were nominal (e.g., type of material for the floor, roof, wall) while others were binary (e.g., presence or absence of radio, clock, TV) or categorical. Hence, the original variables must be standardized and converted to a correlation matrix before performing a PCA (399). The weights for each component are given by eigenvectors of the correlation matrix, and the variance for each component is given by the eigenvalue of corresponding the eigenvector (398). The components are arranged so that the first component explains the largest possible amount of variation in the original data. The second component is uncorrelated with the first and explains a smaller amount additional variance, unexplained by the first component. Subsequent components are uncorrelated with first and second components and explains smaller and smaller additional, unexplained proportion of variation of the original variables (398).

4.9.2.2 Creating the asset index as a measure of SEP using PCA

To standardise the original asset indicators, first, responses to all questions on assets were binary coded into advantageous and disadvantageous SEP based on the type of material used and facilities available, according to the context of Kerala, India. Next, a tetrachoric correlation matrix (399) was created from these binary variables for each life period (*please refer to manuscript I, Supplemental material, eTables 2, 3 and 4, page no 122-123)*. If any variable correlated highly (|0.8|) with other variables, only one variable from the pair of correlated variables was retained for further analysis. In addition, variables were excluded in stepwise manner until a factorable correlation matrix with Kaiser-Meyer-Olkin (KMO) value > 0.7 was attained for each period separately (301). Assets with low test-retest reliability were also removed (*please refer to manuscript I, Supplemental material, etable 5*, *page no 124*). The final variables retained in the matrix for each period were: *Childhood*: crowding, floor, wall, window, piped water, bath, clock, KMO=0.832; *Early adulthood*: crowding, wall, window, piped water, clock, bicycle; KMO=0.771; *Late adulthood*: crowding, wall, window, piped water, clock, radio, television, phone, KMO=0.801. A PCA was conducted on the final correlation matrices to assess the dimensionality of the assets, and the component that explained the maximum variance in each life period (the first component childhood explained 65% of variance, 64% each for early and late adulthood) was extracted (291). Continuous scores were predicted out of these components. for each life period, which were dichotomized using the median of the distribution among controls as cut-off generating respective binary variables representing the SEP exposure (0=advantageous SEP, 1=exposure to disadvantageous SEP) for childhood, early and late adulthood periods of life.

4.9.2.3 SEP exposure measure for critical period models

The binary variables (0-advantageous SEP, 1-disadvantageous SEP) representing SEP in childhood, early, and late adulthood were used as the main exposure in the *critical period model* representing each of these life periods.

4.9.2.4 SEP exposure measure for the accumulation model

A summation of the binary variables representing SEP in each life period generated a variable with four categories with increasing periods of exposure to disadvantageous SEP. This variable represented the *accumulation model*. The variable was coded as: 0=0 period– participants who were in advantageous SEP in all 3 periods of life; 1=1 period-participants who were exposed to disadvantageous SEP in any 1 period and non-exposed in any 2 periods of life; 2=2 periods-participants who were exposed to disadvantageous SEP in any 2 periods and non-exposed in any 1 period of life; and 3=3 periods-participants who were exposed to disadvantageous SEP in all three periods of life.

4.9.2.5 SEP exposure measure for social mobility models

Two models were tested for mobility: childhood to early adulthood mobility, and early to late adulthood mobility.

Childhood to early adulthood mobility - The SEP measure representing this model was a 4-category variable. *Stable advantageous SEP (0, 0)*: Participants who maintained a stable advantageous SEP in both childhood and early adulthood were coded as 0. *Upward mobility (1, 0)*: Participants who were exposed to a disadvantageous SEP in childhood but went on to attain an advantageous SEP in early adulthood were coded as 1. *Downward mobility (0, 1)*: Participants who had an advantageous SEP in childhood but disadvantageous SEP in early adulthood were coded as 2. *Stable disadvantageous SEP (1, 1)*: Participants who maintained a stable disadvantageous SEP in both childhood and early adulthood were coded as 3; all categories were assigned irrespective of their SEP in late adulthood.

Early to late adulthood mobility - A similar strategy was adopted to create the 4-category SEP variable representing social mobility between early and late adulthood by considering participants' SEP in these 2 periods of life.

4.9.3 Covariates used as potential confounders

One of the main challenges addressed in manuscript I is the nature (both static and dynamic) of potential confounders and their temporal ordering with respect to the time-varying exposure of SEP across three time periods and oral cancer. We identified both time-invariant [age, sex, caste i.e., hierarchy in Hindu religion based on occupation, education] and time-varying factors (cigarette smoking, bidi smoking, paan chewing and alcohol consumption) as potential confounders. No HPV was detected in the samples collected from India. Hence this variable was not used in the analysis.

4.9.3.1 Baseline confounders (time- invariant)

Age, sex and caste

Age and sex are strong risk factors for oral cancers. They can also determine an individual's SEP at different periods of life. Hence, to mitigate confounding controls were frequency matched to cases based on 5-year age group and sex. However, there might exist differences within each age group that may result in residual confounding (341). Furthermore, age and sex stand for unknown or unmeasured potential confounders that may determine both the SEP and cancer status of an individual. Hence, these variables were further adjusted in the statistical analysis. Age was used as a continuous variable and sex was binary coded (0= females, 1= males). Caste is a hierarchy in the Hindu religion based on occupation, and may determine an individual's SEP as well as the outcome of cancer. In this study, we collected details on forward caste, backwards caste, other backward caste, scheduled caste scheduled tribe and others as classified by government of Kerala[1]. We adjusted for this variable using a categorical variable (0=higher caste, 1=middle caste comprising of backward caste, 2=other backward[2]/scheduled caste/scheduled tribe/others).

[2] This includes the castes in the Hindu religion and sections of other religions that has been classified as backward by the state governments of India (here; Kerala) due to discrimination faced by them historically

Education (time-invariant)

As discussed previously several indicators are used to measure SEP and they may capture different dimensions of this complex construct. Education may capture a different dimension of SEP than the wealth index. Also, it is an independent risk factor for oral cancers, and the education an individual attains (education is mostly stable after childhood or adolescence) may determine their asset/wealth index in adulthood. Detailed information regarding education was collected from each participant. We used number of years of formal education in the form of a binary variable (0: high education; 1: low education) as an indicator. However, the measure of education is subjected to bias if the differences in birth cohorts of participants from a range of age groups included in a study are unaccounted for (293, 400, 401). With respect to the Kerala study site, considerable educational and sociopolitical reforms took place in the mid1950s, which changed the landscape of education in this state of India (*please refer to page no.33 under sub-section 2.3.4.2 for details*). This information was used to mitigate bias in the categorization of education. The participants were first divided into 2 groups: older: those born before 1950, younger: those born after 1950). For the older cohort, 0-3 years of formal education was considered low level, and 4 years and above was considered as high level of education. For the younger cohort, 8 years of formal education as used as the cut-off for this binary categorization.

4.9.3.2 Time-varying confounders

Tobacco smoking

We used pack-years of tobacco smoked as a measure of tobacco smoking. Extensive information was collected on two commonly used smoked tobacco products in India: cigarette and bidi. This included data on duration (age of cessation - age of initiation) and frequency of consumption (how many cigarettes and /or bidis per day or per week or per month) and for cigarettes, the brand and type used (filtered or non-filtered). This information was collected in the form of multiple stable periods of consumption defined as a specific duration of time (in years) over which a participant followed the same pattern of smoking, either in frequency or brand (for cigarette) or both. Every change in the pattern of smoking marked the start of a different stable period and was recorded with the help of the life-grid tool. From this information, first, the data on frequency of cigarettes smoked per day/week/month was standardised to cigarettes smoked per day. Next, the number of packs of cigarettes smoked per day was calculated by dividing cigarettes per day by 10 (1 commercial cigarette pack in India contains 10 cigarettes) for each stable period of consumption. Next, each year within a stable period was assigned the value of corresponding packs smoked per day. This gave us the flexibility to calculate the cumulative smoking measure of pack-years (product of number of packs smoked per day and duration of smoking) over any specific duration of an individual's life when the participant actually smoked (402, 403). Using this transformed data, we calculated pack-years for five periods of an individual's life (i.e., 0-16 years, 17-23 years, 24-30 years, 31-50 years and above 50 years) which aided in the temporal ordering of variables under study (*please see sub-section 4.9.4 on temporal relationships, page no.72 for further details*).

A similar procedure was performed to calculate pack-years of bidi smoking corresponding to the five periods of life mentioned above. For this calculation, number of bidis per day was divided by 20 to calculate the number of packs of bidi smoked per day (1 pack of bidi contains 20 bidis). In the statistical analyses, the continuous pack-year variables were fitted to restricted cubic spline (*as described in sub-section 4.10.3, page no.76*) to incorporate their non-linear association and adjust for confounding (404).

Paan / betel quid chewing

Similar to tobacco pack-years, a cumulative measure incorporating both duration and frequency of paan chewing was used. Extensive information was collected on paan chewing including duration (age of cessation-age of initiation), frequency (number of quids chewed per day/week/month) and minutes of chewing each time the participant chewed paan, different ingredients used in paan (i.e., combinations of tobacco, slaked lime, betel leaf, areca-nut or any other ingredient) for each stable consumption period. Because paan is used in the form of quids and not packets, we calculated the number of quids chewed per day corresponding to each stable period of consumption. Using a process similar to the one used for pack-year calculation, a cumulative variable known as chew-years was created, and values were calculated for the five periods mentioned under tobacco smoking. Since a non-linear functional form of paan chewing was identified in these data, we used the restricted cubic spline form of this variable in the data analysis to control adequately for confounding.

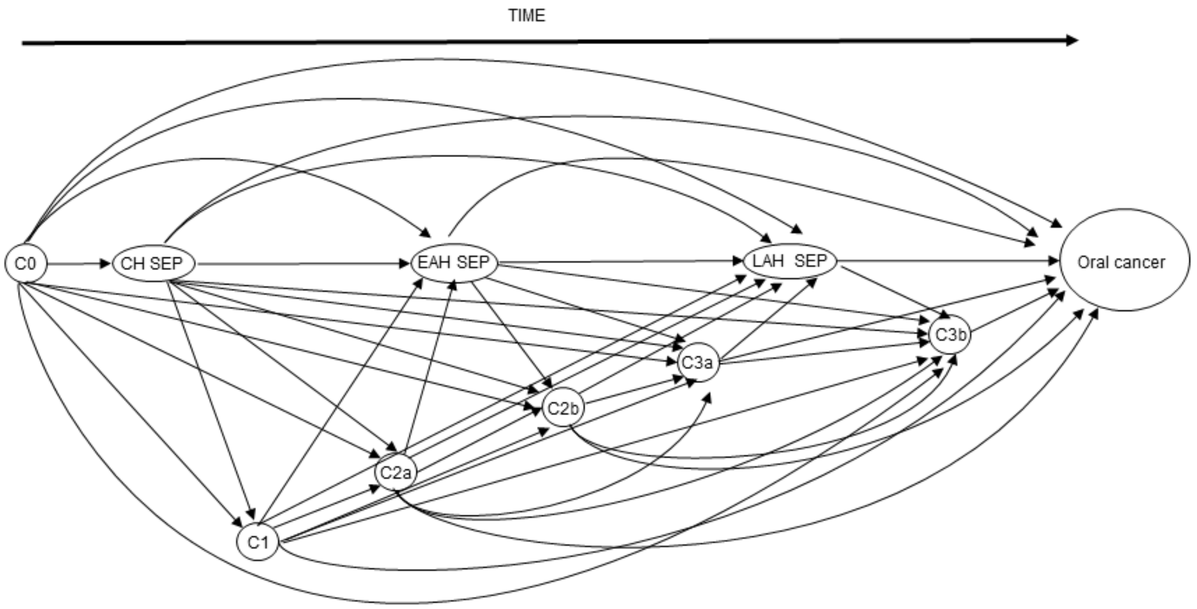
Alcohol consumption

The number of standard drinks of ethanol per week was used as the measure of alcohol consumption. Similar to tobacco smoking and paan chewing, detailed information on type of alcohol beverage [(i) toddy – wine from the coconut tree; (ii) wine; (iii) beer; (iv) hard liquor (arrack, whisky, vodka, brandy, gin, rum, grappa); and (v) any other type of alcohol], age at the start of drinking and age at stop of drinking corresponding to each stable period of consumption, the unit of drinking (small glass (50ml) (1-2oz), medium glass (100ml) (2-3oz), big glass (250ml) (7oz), ½ small bottle (330ml) (1beer), bottle (700-750 ml) (21oz)), as well as consumption frequency (number per day, week, month) were collected. From this information, first, the total amount of alcohol consumption in milliliters corresponding to each period of stable consumption was computed. Using this information, the frequency of alcohol consumption was standardized to milliliters of alcohol consumed per week over the stable period. Next, each type of beverage was standardized based on percentage of ethanol content (5% for beer, 10% for toddy and wine, and 50% for hard liquor) (47, 59) and the number of standard drinks per week for each year of life was calculated. There is no consensus on the definition of a standard drink in India (13 to 28g of pure ethanol) (405). Thus, we divided milliliters of ethanol consumed per week in each stable period by 18 (standard drink=18ml of alcohol containing 14g of pure ethanol) to make it equivalent and comparable to North American standards (47, 59, 406). Using a similar method as with pack-years and chew-years, standard drinks per week corresponding to the five specific periods of life identified above were calculated. This variable was also used in its non-linear functional form for adequate control of confounding.

4.9.4 Temporal relationship of confounders in relation to SEP in three periods of life and oral cancer

The temporal ordering of exposures and covariates with respect to the outcome is imperative when testing life-course models (331). Furthermore, to estimate causal effects (or when applying frameworks for causal inference or associated analytical techniques), the precedence of the causal factor in relation to its effect, is of absolute necessity. Whereas temporal ordering is easier in studies capturing longitudinal data, it is a challenge in case-control studies. But our detailed and comprehensive data collection methods. and techniques to handle the details on confounders (*as described under sub-section 4.9.3.2*) in our life-course based study allowed us to achieve an approximate temporal ordering of variables with respect to SEP in several periods of life and oral cancer diagnosis. As shown in the causal diagram in ***Figure 9***, the vector *C0* represented the time-invariant covariates such as age, sex and caste that temporally precede every other variable under consideration. The vector *C1* represented covariates that were measured for the period between 0-16 years of age. We included education in *C1* because it is usually attained during this period, and could causally affect the subsequent life events of an individual. Other variables represented in *C1* and subsequent vectors *C2a, C2b, C3a* and *C3b* were time-varying risk behaviours (cigarette, bidi, paan and alcohol use). As mentioned previously in the *sub-section 4.9.3.2* of confounders, the cumulative measures of these risk behaviours were calculated for 0-16 years, 17-23 years, 24-30 years, 31 -50 years, and above 50 years. Risk factors collected for the period between 0-16 years might be an effect rather than cause of SEP between 0-16 years of age and were included in *C1*. However, we suspected that the association between *early adulthood SEP* (17-30 years) and habits captured during 17-30 years, was bi-directional, that is, SEP and habits can influence each other causally. Bidirectional arrows cannot occur in causal structures at the same time point (357, 368, 407). To overcome this, we split the habits in this period into vectors C2a (17-23 years) and C2b (24-30 years). This was done assuming that *C2a* would be affected by *C0, C1* and *CH SEP*, but would influence part of SEP in 17-30 years and other subsequent variables. And *C2b* would be affected by *C0, C1, C2a, CH SEP* and *EAH SEP*. The choice of cut-point (i.e., 23 years) was arbitrary. A similar strategy was used with risk behaviours recorded for above 30 years of age. Risk behaviours recorded during the period 31-50 years of age were represented by C3a, and those recorded above for 50 years (the eldest participant was 88 years old) were represented by C3b. This approximate temporal ordering identified complex feed-back loops between the variables under study as any given variable/vector represented in ***Figure 9*** had an arrow pointing from them to any other variable/vector temporally subsequent to it.

Figure 9: Causal graph representing SEP in three periods of life, oral cancer and associated potential confounders



Directed acyclic graph (DAG) representing the relationship between exposure, potential confounders and outcome, in the study from Kerala, India, 2008-2012 (n=684). ***Oral cancer****: Outcome;* ***SEP:*** *Socioeconomic position;* ***CH SEP****: SEP during childhood (0-16 years);* ***EAH SEP****: SEP during early adulthood (17-30 years);* ***LAH SEP****: SEP during late adulthood (above 31 years);* ***C0****: Vector representing baseline covariates, age, sex, caste i.e., hierarchy in Hindu religion (potential time-invariant confounders);* ***C1****: Vector representing education, risk behaviours (time-varying) of cigarette smoking, bidi smoking, paan chewing and alcohol consumption recorded during 0-16 years of age;* ***C2a****: Vector representing risk behaviours recorded during 17-23 years of age;* ***C2b****- Vector representing risk behaviours recorded during 24-30 years age;* ***C3a****- Vector representing risk behaviours recorded during 31-50 years of age;* ***C3b****- Vector representing risk behaviours recorded above 50 years.*

4.10 Measures- Manuscript II

In the manuscript II, we considered the interactive effects of SNPs investigated in this study and smoking on the risk of SCCHN. Hence, the dependent (SCCHN) variable, main exposures (SNPs and smoking) and associated potential confounders are described below.

4.10.1 Dependent (outcome) variable – SCCHN

SCCHN cases were selected as described in *section 4.3, page no.58*. The outcome variable was treated as binary, with the presence of any oral or pharyngeal or laryngeal cancers coded as 1 (cases) and the absence of all coded as 0 (controls).

4.10.2 Independent (primary exposure) variable - Genetic variants

The genetic variants associated with CYP450 genes coding Phase I XMEs are involved in the bio-activation of a variety of tobacco smoke chemicals into electrophilic reactive moieties with carcinogenic potential. The variants associated with GST genes encoding Phase II enzymes are involved in the detoxification of reactive metabolites of Phase I bio transformation. The characteristics of these SNPs and their association with SCCHN have been described in detail under *sub-section 2.3.2, and Tables 3 and 4*. In general, I will consider all genetic exposures as binary variables, with categories coded as 0 considered as reference. The genotypes were collapsed into two categories majorly because the minor allele frequencies of these SNPs in the Caucasian population (except those related to GST enzymes) were low. Specific details on categorization of these genetic measures are given below.

4.10.2.1 Single nucleotide polymorphisms in CYP and GST genes

Dominant models of inheritance were tested for CYP1A1*2A, *2C, CYP2E1c2, CYP2A6*2 and GSTP1 105Val. Dominant model assumes that just the presence of the variant allele, as either homozygous variant or heterozygous variant/wild phenotypes, is enough for the effect of wild allele to be masked. Hence, carriers of these variant alleles, considered as the exposed group (assuming equal risk for homozygous variant and heterozygous wild/variant groups) were compared with non-carriers, assumed unexposed. Thus, CT/CC genotypes for CYP1A1*2A, AG/GG genotypes for CYP1A1*2C, GC/CC genotypes for CYP2E1c2, AT/AA genotypes for CYP2A6*2 and AG/GG for GSTP1 105Val were respectively coded 1 (carriers, exposed), and TT genotypes for CYP1A1*2A, AA genotypes for CYP1A1*2C, GG genotypes for CYP2E1c2, TT genotypes for CYP2A6*2 and AA for GSTP1 105Val were respectively coded 0

4.10.2.2 Copy number variants in CYP2D6 and GSTM1 genes

In this study, we identified 1 to 9 copy numbers of CYP2D6 non-functional null allele among our sample. Individuals with lower number of these null allele are hypothesized to be at relatively higher risk for SCCHN compared to those with higher number of copies of the allele. Based on the distribution of these CNVs in this study, this genetic exposure was binary coded; 1 to 2 copies considered as exposed (coded 1) and 3 to 9 copies as unexposed (coded 0). For GSTM1, we identified 0 to 3 copies. To ascertain sufficient numbers in the categories, the GSTM1CNV classification was limited to null (0 copies, coded 1) and non-null (1-3 copies, coded 0).

4.10.3 Independent (secondary exposure) variable - Pack-years of cigarette smoking

To incorporate the effect of correlated measures such as frequency and duration of smoking and to avoided issues related to collinearity between these measures during statistical analysis, it is recommended to use cumulative measures of smoking in studies investigating the impact of this risk behaviour on cancers (403, 408, 409). Hence, in this study, we used cigarette pack-years to represent tobacco smoking history (402). Pack-years was computed as the product of the average smoking intensity over lifetime, and the total duration of smoking at the time of diagnosis for cases and at the time of interview for controls.

Cigarette pack-years was derived from information on participants' history of cigarette (filtered or unfiltered or hand-rolled), cigar and pipe smoking along the life-course in a similar method as described in *sub-section 4.9.3.2, Tobacco smoking, page no.70*. Hand-rolled cigarettes, cigars and pipes were first converted to their commercial cigarette equivalent (20 commercial cigarettes = 4 hand-rolled cigarette = 4 cigar=5 pipes= 1 pack of commercial cigarettes) (82). This information was used to create total duration of smoking and average packs smoked per day over life time respectively. A product of these two generated a continuous measure of pack-years of cigarettes smoked over life time. Certain participants had a combination of active periods of smoking and periods of abstinence over their life-course. Periods of abstinence were excluded while calculating total duration as we assumed very low probability of misclassification (inclusion vs exclusion of such periods of abstinence gave us similar results [e.g., total duration of smoking including periods of abstinence, (mean=32.25 years ±15.45) and excluding such periods (mean=31.47 years ±15.46)]. Furthermore, from information on time since smoking cessation (age during interview minus age of cessation), we identified that participants who stopped smoking ≤ 2 years prior to recruitment had a higher risk for the outcome than actual current smokers (time since cessation=0) *(Manuscript II, Supplemental material, eTable 1, page no. 150*). Hence, to minimize probability of protopathic /reverse causality bias, we used a cut-off of 2 years' prior interview to define ex-smokers, and excluded details of any exposure (e.g., frequency, duration) during this period for pack-year calculations.

To estimate the effect of various SNPs at different levels of smoking, we categorised the cigarette pack-year variable into 3 categories. The optimal cut-off point for categorization was informed through multiple rigorous modelling approaches.

The first step was to determine the correct functional form of pack-years using dose-response curves. For this, first an outcome model with pack-years entered as linear form was fit following guidelines proposed by Leffondre et al 2002 (403). Subsequently, I fitted multiple logistic regression models, each with pack-years in restricted cubic spline functional form determined by knots at various percentiles of its distribution (5, 50 and 95; 10, 50, 90; 25, 50, 75; 5,25, 75 as well as the modified knot positions recommended by Harrell) (410). Next, among these spline models, the best fit model was chosen by comparing Akaikes information criteria (AIC) values (411). The model with knot positions at 5, 50 and 95 percentiles had the lowest AIC value and was deemed as the best fit. Subsequently, using a likelihood ratio test, fit of this model was compared with that of the linear model under the assumption that the linear model was nested within the model with spline parameters. The spline model had a superior fit. Using this model with spline parameters, the shape of the dose-response curve between pack-years and the SCCHN outcome was constructed, and determined to be non-*linear (Manuscript II, Supplemental material, eFigure 1, page no.151).* The curve indicated that the risk for the outcome increased sharply up to approximately 70 pack-years beyond which the risk plateaued. This informed us that the risk point (optimal cut-off) would lie anywhere between >0 and 70 pack-years.

In the second step, a parametric outcome based approach, developed to identify optimal cut-off for continuous covariates with non-linear functional form as well with respect to a binary outcome, was used to identify the optimal cut-point among smokers (412). This approach, (a) maximized the difference in risk between participants in the two outcome groups, and (b) bonferroni corrected for alpha=5% (to circumvent the possibility of inflation of Type 1 error in the identified cut point, due to multiple comparisons of various cut points possible over the range of >0 and 70 pack years). The optimal cut-off was identified to be at 32 pack-years (defined as smoking 32 packs of commercial cigarette per day for a year, or 16 packs/day for 2 years, or 8 packs/day for 4 years, 4 packs per day for 8 years or 2 packs per day for 16 years or 1 pack/day for 32 years etc.) (412). Using this cut-off, the final smoking variable was categorized as non-smokers (0 pack-years), moderate smokers (>0 to ≤ 32 pack-years) and heavy smokers (> 32 pack-years).

4.10.4 Covariates used as potential confounders

It has been recommended that while assessing interactive effects between two variables, all measured potential confounders for the relation between each exposure variable (i.e., genetic variants, and smoking) and the outcome (SCCHN) must be present in the full confounder set. Variables considered as confounders for estimating the total effect of genetic variants and health outcomes are usually limited to those that address population stratification (biased association between genetic variant and outcome due to heterogeneous ethnicity/ population sub structure), SNPs in linkage disequilibrium, and sex. However, many enzymes coded by SNPs considered in this study are induced by polycyclic aromatic hydrocarbons (CYP1A1), nicotine (CYP2A6, CYP2E1), and ethanol (CYP2E1) found in pollutants, occupational exposures, diet, tobacco smoke, alcohol among others. SNPs under study are noisy proxies for enzymes they code for; thus, exposures that may induce these genes may be confounders for the relation between specific SNPs and SCCHN. Hence, we considered ethnicity, SNPs in LD (e.g., CYP1A1*2A and 2C), age, sex, alcohol (ethanol) and education (SEP proxy) as potential confounders for respective SNPs and SCCHN associations, as depicted in *Appendix V*, DAGs Figures A1 to A7 (*page no. 312*). DAGs were constructed using DAGitty software version 2.3 (413, 414). To mitigate confounding by ethnicity (population stratification), all analyses were restricted to Caucasians.

Based on *a priori* knowledge, measured potential confounders for the relationship between smoking and SCCHN were identified as alcohol consumption, education (as a measure of SEP), HPV risk, CYP2A6*2, age and sex. Using DAGs *(Appendix V, Figure A8, page no. 312)*, we identified the minimal sufficient adjustment set [alcohol consumption, SNPs, education, age and sex] to estimate the total effect of smoking on SCCHN. HPV is a strong risk factor for SCCHN and it has a complex relationship with smoking that may not be limited to interaction. Hence, HPV, which fits the definition of a "potential confounder", was included in the confounder set.

The covariates included in the final set of confounders for any or all gene-environment interaction models in this study are described below.

Alcohol consumption

The frequency of ethanol consumption (average amount of ethanol in ml consumed per day) was used as the measure of alcohol consumption. This measure was derived from detailed information collected in alcoholic beverages as described in sub-section *4.9.3.2, page. no 70.* Each beverage was converted to ethanol equivalents (10% ethanol in wine and aperitif, 5% in beer/cider, and 50% in hard liquor) (81). I calculated total duration and total frequency of ethanol in ml consumed over life time from which I computed the average amount of ethanol consumed per day in ml. Similar to tobacco pack-years, the correct functional form of this ethanol frequency variable was determined (by comparing fit of linear and restricted cubic spline models and fitting dose-response curves) to be non-linear. The two spline parameters in continuous form were used to represent frequency of ethanol consumed per day.

Socioeconomic position - Education

SEP is a determinant of tobacco smoking and a distal risk factor for SCCHN. I used number of years of formal education (415), used as a continuous variable in its linear functional form, as measure of SEP.

HPV status

As described in *sub-section 4.7.4, page. 61,* HPV status was recorded for 36 HPV types. Based on their oncogenic potential, these types were assigned into hierarchical categories as either negative (coded 0), exclusively non-α-9 species types (coded 1), α-9 types other than HPV16, i.e., 31,33,35,52,58 and 67 (coded 2), and HPV16 (coded 3) (386).

Age, sex, SNPs

Age (as a continuous variable) and sex (binary) were adjusted for in all statistical models to mitigate residual confounding. CYP1A1*2A and CYP1A1*2C are known to be in LD. Hence, they were mutually adjusted for in the models in which either of them appeared as the main exposure.

4.11 Measures- Manuscript III

Manuscript III aimed at estimating the effects of CYP2A6*2 and ADH1B*2 on SCCHN through interactive and mediating pathways by smoking and alcohol intensities respectively. Hence, the dependent variable (SCCHN), exposures (CYP2A6*2 and ADH1B*2) and associated potential confounds are described below.

4.11.1 Dependent (outcome) variable – SCCHN

The dependent variable was SCCHN as described in *sub-section 4.3, page no. 58.*

4.11.2 Independent (main exposure) variables – CYP2A6*2 and ADH1B*2

In this study, CYP2A6*2 was genotyped as TT, AA and AT (A = minor allele). Relative to carriers of this allele (AT or AA genotype), non-carriers (TT genotype) are documented to smoke with higher intensities and are hypothesized to be at increased risk for SCCHN among smokers (217, 416). Assuming positive monotonic association between exposure and outcome and exposure and mediator, the TT genotype was considered exposed (coded 1) and compared with AT/ AA genotypes considered unexposed (coded 0). Similarly, relative to carriers, non-carriers of the ADH1B*2 allele are documented to drink with higher intensities and also are at increased risk for

SCCHN among alcohol consumers (161, 162, 237). Hence, GG genotype (coded 1) were compared with AG/AA genotype (coded 0).

4.11.3 Mediators – Intensity of smoking and alcohol consumption

Among the various dimensions of smoking and alcohol consumption behaviour, CY2A6*2 is strongly associated with the intensity of smoking, and ADH1B*2 with intensity of alcohol consumption (248, 416). Hence, we used the intensity measures of these behaviours as mediators.

Details of the smoking data collection are described in *sub-section 4.9.3.2, page no. 70*. All tobacco types were converted to a commercial cigarette equivalent based on their-nicotine content (1/9 cigar = 1/3.5 pipe=1/2 hand rolled cigarettes= 1 commercial cigarette) (52). From the total duration and frequency of a commercial cigarettes used, we calculated the average number of commercial cigarettes smoked per day over the lifetime. Using techniques described in *sub-section 4.10.3, page no.76*, the shape of the dose-response curve for the cigarettes per day-SCCHN relationship was determined to be non-linear *(Manuscript III, Supplemental material eFigure 1, page no. 179)*. Using the parametric outcome based approach described in *sub-section 4.10.3* (412), the optimal cut-off point to categorize the smoking intensity variable among smokers was identified to be 18 cigarettes per day. Then, using this cut-off, a binary variable representing the intensity of smoking was created, with smokers who smoked up to 18 cigarettes per day categorized as low-smokers (coded 0), and those smoking above 18 cigarettes per day as heavy smokers (coded 1).

The data collection on alcohol consumption as well as the creation of an intensity measure of ethanol consumption were described in *sub-sections 4.9.3.2 and 4.10.4.* Using a technique similar to the one employed for the categorization of smoking intensity, the optimal cut-off point to categorize the average amount of ethanol in millilitres consumed per day over the lifetime was identified to be at 25ml of ethanol (*Manuscript III, Supplemental material eFigure 2, page no. 180*). The final intensity measure for alcohol was represented by a binary variable: mild drinkers (coded 0): participants who consumed up to 25ml of ethanol per day and heavy drinkers (coded 1): participants who consumed more than 25ml of ethanol per day considered as

4.11.4 Covariates used as potential confounders

Manuscript III involved analysis related to mediation and interaction based on the counterfactual causal framework. For the estimation and causal interpretation of effects in mediation studies using the counterfactual framework, four no-confounding assumptions are required along with correct model specification (343): there is no unmeasured confounder of the effects of (i) genetic exposure on SCCHN, (ii) genetic exposure on the associated mediating risk behaviour, and (iii) mediating risk behaviour on SCCHN, and (iv) none of the mediating risk behaviour-SCCHN confounders are affected by the associated genetic exposures. We addressed (i) and (ii) by restricting our analysis to Caucasians, thus mitigating confounding due to population stratification (417). For (iii), we adjusted for potential confounders of the relationship between risk behaviours and SCCHN. For the smoking intensity-SCCHN association, we identified duration. and time since cessation of smoking (continuous, mean centred, current smokers recoded to zero), and intensity of alcohol (continuous, adjusted for restricted cubic spline) as confounders. For the alcohol intensity-SCCHN association, time since stoppage of use of alcohol (continuous, mean centred, current users recoded to zero) duration of alcohol, and pack-years of commercial cigarette equivalence were identified. Additionally, we adjusted for age (continuous), sex, number of years of education (continuous) and HPV risk types for both associations. These variables are not known to be affected by the associated genetic exposures that may potentially address the 4th no-confounding assumption (*please refer to sub-section 4.10.4 for details on these confounders*).

4.12 Statistical analysis

This section presents the details of general and specific statistical techniques used to analyse the data for each manuscript.

4.12.1 General considerations

Descriptive statistical analysis was performed to explore the distribution of variables used in the study among cases and controls. T-Tests were used to compare means of continuous variables between the two groups, while chi-square tests based on cross-tabulations were used to describe categorical data (418). For manuscripts II and III which involved genetic variants, deviations from the Hardy-Weinberg equilibrium were assessed among the control population using chi-square tests. Minor allele frequencies were estimated among controls.

The primary dependent/outcome variable investigated in each manuscript was a binary. Furthermore, exposure models used to create inverse probability weights for the marginal structural models in the 1st manuscript, and mediator models fitted in the 3rd manuscript had a binary dependent variable. Hence, all manuscripts depended on a binary logistic regression model to calculate association or effect estimates.

Binary logistic regression

Binary logistic regression is a type of generalized linear model used to estimate the probability of a binary response (dependent) variable as a linear function of any number of independent predictor variables by fitting data to a logistic curve (411). If *P* is the probability of a disease occurring and *1-P* is the probability of the disease not occurring, then *P/1-P* gives the odds of the disease occurring. A log transformation allows the odds of a disease to be expressed as a linear function of the independent variables as:

$$\ln[\frac{Pr}{1-Pr}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots . + \beta_k X_k$$

where β_0 is the y-intercept (log odds of the probability of the outcome when predictor variables have a value of zero), X_1 to X_k represent k independent predictor variables that can be binary, categorical, linear or non-linear functional forms. The corresponding *β's* of any of the *k* predictor variables are coefficients representing regression parameters associated with each variable. The advantage of using a logistic regression over other generalised linear models lies in the predicted probabilities of a binary outcome to fall between 0 and 1, irrespective of the functional form of independent variables used in the model (419). Thus, it is advantageous to use logistic regression in our study, as non-linear functional forms of variables represent risk behaviours. In a case-control study, the measure of association between an independent variable and a dependent variable is usually expressed as odds ratios (OR) (342). In a binary logistic regression model, the OR for the outcome associated with a one unit increase in the predictor variable when all other independent covariates are kept constant is given by $OR = e^{\beta}$. The precision of an OR, usually represented by its 95% confidence limits or interval (95% CI), can be estimated from the value of the β coefficients and associated standard errors. The parameters in a normal logistic regression are estimated using multiple iterations through a maximum likelihood estimation method. However, in the case of weighted regression models (e.g., marginal structural models), the parameters are estimated using a pseudo-loglikelihood method.

The relative risk (RR) of an outcome, which is measurable in cohort studies is related to the OR as

$$RR = \frac{\mathrm{OR}}{[1-Pr_0)+(Pr_0)(OR)}$$

where Pr_0 is the probability of the outcome in the unexposed (341). Hence, if Pr_0 is very small as in a rare disease outcome (usually <10% of the population as in the case of oral cancers or SCCHN), RR approximates OR.

The sub-sections below describe analytical techniques specific to each manuscript.

4.12.2 Analysis specific to Manuscript I

A marginal structural model (MSM) with inverse-probability weighting (IPW) is an appropriate solution to estimate causal effects with complex feedback loops due to time-varying exposures and confounders. This is the third g-method developed by James Robins in 1998 based on the counterfactual/ potential outcomes framework, and is described in a series of papers by Robins and colleagues for application to longitudinal data (363-365, 420). The overall technique involves two steps of regression modeling; first, fitting exposure models based on which IPWs are calculated, and fitting a weighted regression model using the IPWs. Fundamentally, the IPW creates a pseudo-population using original data in which the exposure is independent of the measured confounders, and the coefficients obtained from the MSMs have a causal interpretation provided certain identifiability assumptions are satisfied (*please refer to sub-section 2.7.1, page no.43 for additional details on assumptions*).

With observational studies, the technique is majorly used with cohort data where repeated measurements are readily available, regression models can be fitted on the outcome as well as exposures, and the probability of the outcome among the unexposed is known. These conditions are usually not satisfied with case-control studies (*please refer to page no. 40, sub-section 2.6.2, for additional details*), thus MSM with IPW is rarely used with this study design (316, 421-423). Furthermore, no case-control study has applied this technique to a complex life-course framework such as ours, which involves time-varying exposures as well as confounders at multiple time points. To address these challenges, we applied novel approaches to this existing technique to derive our associational estimates. The steps involving specifying causal contrast for each model, IPW, MSM, and the use of sampling weights to account for the case control design are described below.

4.12.2.1 Average causal contrast for a binary exposure and outcome at one time point

Let A be a dichotomous exposure (e.g., SEP) taking values 0 (unexposed to disadvantageous SEP) and 1 (exposed to disadvantageous SEP), Y be a dichotomous outcome measured at the time of diagnosis / interview (cases=1 and controls=0), and C, a vector of covariates. The random variable A takes two values, a=0 or a=1. Let Y_a denote the counterfactual/potential outcome for a given participant under SEP exposure level a. Then the two potential outcomes under the exposure values are $Y_{a=1}$ and $Y_{a=0}$. Under these notations of exposure contrasts, the average causal effect of SEP at one time point on the outcome of oral cancer can be represented as (359, 424)

$$\text{Average causal effect} = \text{logit}\ [Y|A, C] = E\ [Y_{a=1,\, c}\text{-}Y_{a=0,\, c}]$$

4.12.2.2 Causal contrasts and equations for life-course models

Saturated all trajectories model

In our study, the binary SEP exposure was measured for three time points and we are testing multiple life-course conceptual models, each fit on a specific function/ combination of SEP exposure at these time points (A, the random variable takes values a=000, 111 or any combinations of being exposed/unexposed in childhood (CH), early adulthood (EAH) or late adulthood (EAH) periods. Any binary exposure measured over k time periods can have 2^k potential outcomes. Hence, in our study, there are a maximum of 8 potential outcomes represented by 8 trajectories over the life-course. These 8 potential outcomes are represented by a saturated model (*Manuscript I,* *Supplemental material, eTable 6, page no.125)*. Each category in the saturated model represents a specific life trajectory based on the combination of the specific exposure levels a participant experienced over the three life periods. Corresponding to each life trajectory, these potential outcomes can produce 7 causal contrasts relative to the reference exposure status of being unexposed in all three time periods. Disregarding confounders, the magnitude of these causal contrasts can be obtained from the parameters of the unconditioned logistic regression equation (425):

$$\text{logit}\ (Y|A_{\text{Saturated}}) = \alpha + \beta_1 A_{100} + \beta_2 A_{010} + \beta_3 A_{001} + \beta_4 A_{110} + \beta_5 A_{101} + \beta_6 A_{011} + \beta_7 A_{111}$$

where A_{100} = exposed in childhood (CH) only, A_{010}= exposed in early adulthood (EAH) only, A_{001}= exposed in late adulthood (LAH) only, A_{110}= exposed in both CH and EAH only, A_{101}= exposed in both CH and LAH only, A_{011}= exposed in EAH and LAH only, $A_{111=}$ exposed in CH, EAH and LAH. The intercept represents the logit of the probability of the outcome when the exposure status is A_{000} (i.e., unexposed at all time points).

Similarly, there are 4, 2 and 4 potential outcomes for the accumulation model, each critical period model and each social mobility model respectively. The causal contrasts for each of these models are provided in *Manuscript I, Supplemental material, eTable 6, page no.125*.

4.12.2.3 Inverse probability of exposure weights (IPW)

A simple scenario of time-varying confounding affected by prior exposure has been described in *sub-section 2.7.4.7, page no. 50.* The situation is more complex in our study that involves SEP at three time points and time-varying covariates in several periods with complex associations between them. The problem can be addressed by using IPW.

In IPW, each participant in the study is weighted by their inverse of their probability of being exposed to disadvantageous SEP at specific time periods. For a binary exposure A, a minimally stabilized IP weight SW is given by

$$W = \frac{\Pr\,[\mathrm{A} = 1]}{\Pr[\mathrm{A} = 1/\mathrm{C}]}$$

where Pr [A=1/C] is the probability of being exposed fitted on confounder history C. Pr[A=1] in the numerator stabilizes the IPW which reduces variability in the weights and results in a more efficient estimation of parameters (365, 426, 427).

In our case, for each participant i, SEP=A (A^{CH}, A^{EAH} and A^{LAH}) takes the value of a_i (a_i^{ch}, a_i^{eah}, a_i^{lah}) =1(exposed to disadvantageous SEP), at the three time periods childhood, early and late adulthood, and confounder vectors take the values, C0=$c0_i$, C1=$c1_i$, C2a=$c2a_i$, C2b=$c2b_i$ and C3a=$c3a_i$ across these time periods. Then, the minimally stabilized IPWs, $SWch$, SW_{eah} and SW_{lah} are given by,

$$SW_{ch} = \frac{\Pr\left(A^{\mathrm{CH}} = \mathrm{a}_i{}^{\mathrm{ch}}\right)}{\Pr(A^{\mathrm{CH}} = \mathrm{a}_i{}^{\mathrm{ch}} | \mathrm{C0} = \mathrm{c0}_i)}$$

$$SW_{eah} = \frac{\Pr\left(A^{\mathrm{EAH}} = \mathrm{a}_i{}^{\mathrm{eah}}\right)}{\Pr(A^{\mathrm{EAH}} = \mathrm{a}_i{}^{\mathrm{eah}} | A^{\mathrm{CH}} = \mathrm{a}_i{}^{\mathrm{ch}}, \mathrm{C0} = \mathrm{c0}_i, \mathrm{C1} = \mathrm{c1}_i, \mathrm{C2a} = \ \mathrm{c2a}_i)}$$

$$SW_{lah} = \frac{\Pr\left(A^{\mathrm{LAH}} = \mathrm{a}_i{}^{\mathrm{lah}}\right)}{\Pr(A^{\mathrm{LAH}} = \mathrm{a}_i{}^{\mathrm{lah}} | A^{\mathrm{EAH}} = \mathrm{a}_i{}^{\mathrm{eah}}, A^{\mathrm{CH}} = \mathrm{a}_i{}^{\mathrm{ch}}, \mathrm{C0} = \mathrm{c0}_i, \mathrm{C1} = \mathrm{c1}_i, \mathrm{C2a} = \mathrm{c2a}_i, \mathrm{C2b} = \mathrm{c2b}_i, \mathrm{C3a} = \mathrm{c3a}_i)}$$

An exposure model (logistic regression) for SEP exposure in each period was fitted using participant's covariate history to obtain predicted probabilities of being exposed, from which the denominator for the stabilized IPW was obtained. In these exposure models, we adjusted for categorical and continuous confounders using indicator coding and restricted cubic splines (i.e., adjusted for not only the dose but also their non-linear functional form), respectively. To obtain the predicted probability for the numerator, an intercept only model was fitted (hence leading to the creation of a minimally stabilized IPW). To approximate the distribution of the SEP exposure in our study to that of the underlying population (as in a cohort study) each exposure model was weighted by the inverse of sampling fraction as proposed by Vandeerweele and Vansteelandt (2010) (366). This was achieved by weighting the cases with p/q and controls with (1-p)/(1-q) where p is the prevalence of oral cancer in India during the four-year of study and q the proportion of cases in our study relative to total sample (9, 366). The weights SW_{ch}, SW_{eah} and SW_{lah} accounts for the measured confounding in the relation between SEP measured at childhood, early and late adulthood, respectively, and oral cancer. Estimated weights with a mean far from one or very extreme values are indicative of non-positivity or misspecification of the weight model. This issue can be addressed by truncating the weights (426). Although we did not have extreme values, we did have few higher values (>10) for SW_{eah} and SW_{lah} . Truncation was performed for these weights by attributing the value of the 99th and the 1st percentiles to the 1 % highest and 1 % lowest weights, respectively. The truncation also enabled us to retain the sample size and prevent large bias in estimates due to high weights compared to the small variance induced (bias-variance trade off) (426). The summary statistics for minimally stabilized IPWs are provided in *Manuscript I, Supplemental material, eTable 7, page no. 126.*

In summary, in the pseudo reweighted population created by applying IPW to our original study sample, our SEP exposure at each time point became independent of the measured confounders, whether time-varying or not, establishing exchangeability between the exposed and unexposed groups in this population.

4.12.2.4 Fitting marginal structural models for life-course models

Fitting a logistic regression model in the pseudo-population is equivalent to fitting a weighted model in the study population, the parameters of which are equivalent to that of a MSM (426). The estimates are inferred as marginal because in MSM, the outcome is fitted only on a function of exposure with weights. The marginal estimates (OR) in this study obtained from MSMs correspond to the average causal effect of the exposure in the study sample. Hence, weighted regression models, each representing an MSM, were fitted in way corresponding to the causal contrast equations provided in *Manuscript I, Supplemental material, eTable 6, page no. 125*. In general, the MSMs were defined as:

$$\text{logit}\ \{\Pr[Y_{g(\text{SEP})}=1]\} = \alpha + \beta_1\ g(\text{SEP})$$

Where g(SEP) is a function of SEP exposure specific to the accumulation, critical period and social mobility models as described in *sub-section 4.9.2.3, 4.9.2.4 and 4.9.2.5, pages 65 and 66).* Exponential of the β_1 provides the marginal OR. In addition to these models, we also fit a saturated all trajectories marginal structural model (425, 428). The categories of the SEP exposure variable for this model had all eight possible trajectories formed by binary SEP exposures measured over three periods of life (*please refer to eTable 6, Supplemental material, manuscript I, page no.125*)

An additional challenge related to the case-control study design was addressed while fitting the weighted MSMs. Although we assumed that our data was derived from an underlying cohort, the composition of a case-control study sample may not reflect the composition of the full population risk set at a given point of time. This is due to the over sampling of cases in a case-control design (429). To address this issue, Leffondre et al (2010) (429) developed weighted partial likelihood estimators for a case-control study with time-varying exposures, where the sampling weight (SampW) is given by

$$\text{SampW} = [(1-\prod)/\prod] * \text{ncases/ncontrols},$$

where ∏ is the annual prevalence of oral cancer in India during the four years of the study, and ncases and ncontrols are the numbers of cases and controls in our sample. In our analysis, we weighted MSM for each model using the product of specific combinations of stabilized weights SW_{ch}, SW_{eah} and SWl_{ah} multiplied by the SampW. The weight SW123, used for fitting MSMs for the accumulation, late adulthood critical period, early to late adulthood mobility, and saturated models was obtained using

$$\text{SW123} = SW_{ch} \text{ x } SW_{eah} \text{ x } SW_{lah} \text{ x SampW}$$

The weight SW12, used in MSMs for the early adulthood critical period model, and the childhood to early adulthood mobility model was obtained with

$$\text{SW12} = SW_{ch} \text{ x } SW_{eah} \text{ x SampW}$$

Finally, the weight SW12, used in MSM for the childhood critical period model was obtained using

$$\text{SW1} = SW_{ch} \text{ x SampW}$$

Annotated Stata codes developed to create the exposure weights, a description of SEP exposure weight specification, and outcome marginal structural models in this study are provided in *manuscript I, Supplemental material, etable8, page no.126.*

4.12.2.5 Comparing the fit of life-course MSMs using quasi likelihood criterion

In addition, we compared the fit of the models tested in this study. Because the coefficients of logistic regression models are derived using maximum likelihood estimation, the fit of regression models is usually compared using parameters derived from the estimate of log likelihood (e.g., AIC). However, the coefficients of weighted MSM models are derived using pseudo-maximum likelihood estimates. Hence, model comparison criterion based on likelihood ratio tests and AIC values cannot be used to compare MSM models. To address this issue, Platt et al (2013) developed a weighted quasi-likelihood information criterion (QICw) for marginal structural models (430). We estimated QICw for each MSM in this study to compare their fit to the data. The model with the lowest QICw value was selected as the relatively best fit model.

In addition, the accumulation, critical period and social mobility models are considered to be nested within the saturated all-trajectories model (425). The fit statistic (e.g.,) for this model has been considered as a reference for comparing fit of life-course models (425). Hence, we estimated QICw for this model as well. QICw estimation was conducted in SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina).

Thirty-seven participants (17 controls and 20 cases) had missing values related to the main exposure. Therefore, we present our results on complete case analysis of 684 participants. Analyses were performed using Stata 13SE and SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina). Annotated Stata codes are provided in *Manuscript I, Supplemental material, eTable 8, page. no 124.*

4.12.3 Analysis specific to Manuscript II

The primary aim of manuscript II was to estimate the joint effects, stratum specific effects and measures of interaction between genetic variants involved in the metabolism of tobacco and three levels of smoking for the risk of SCCHN. Before the primary aim was addressed, unconditional logistic regression models were fit to estimate the odds ratios (OR) and 95% confidence intervals (CI) for the total effect of each genetic variants on SCCHN. In the combined effect estimation of genetic variants and smoking, we used unconditional logistic regression to calculate the joint effect estimates, estimates of effect of each genetic variant within the strata of three smoking levels, and the estimate of statistical interaction. Because our outcome was rare, we assessed interaction between the exposures and the outcome on both multiplicative and additive scales using these regression models with a single reference group, and the stratum specific estimates and their respective CIs.

4.12.3.1 Joint effect estimation

Logistic regression models were used to estimate the joint effect ORs and corresponding 95% CIs in the group defined by each level of the variant (G=0,1) and each level of smoking (E=0,1,2), as represented in **Table 6**.

Table 5: Representation of joint effect ORs for SCCHN by strata of smoking and SNP

Genetic variant	Smoking		
	E=0	**E=1**	**E=2**
G =0	OR_{00} =1	OR_{01}	OR_{02}
G =1	OR_{10}	OR_{11}	OR_{12}

Here, G=0 and G=1 represent non-carriers and carriers of a specific genetic variant respectively. E=0, E=1 and E=2 represented non-smokers, moderate smokers and heavy smokers respectively. The ORs for SCCHN corresponding to each group (i.e., E=1, G=0; E=2, G=0; G=1, E=0; G=1, E=1, and G=1, E=2) relative to the reference group (G=0 and E=0) were estimated.

4.12.3.2 Stratum specific effects

Next, the effect of G=1 relative to G=0 on SCCHN, within the strata of E0=0 (calculated as OR_{10}/ OR_{00}), E=1 (calculated as OR_{11}/ OR_{01}) and E2 (calculated as OR_{12}/ OR_{02})OR were calculated (419).

4.12.3.3 Interaction on multiplicative scale

Since our outcome was rare, we assessed interaction on both multiplicative and additive scale using the logistic regression models (419).

Estimates of multiplicative interaction were derived by entering a product term between the two-category genetic variable and three-category smoking variable into a logistic regression model. Measure of each multiplicative interaction were obtained by exponentiating the coefficients corresponding to the 2 levels of the product term. The measure of multiplicative interaction between carrier of a variant, G=1 and moderate smoker E=1 is given by OR_{11}/ OR_{10} $OR_{01.}$ This estimate measures the extent to which, on OR scale, the effect of being a carrier of a genetic variant and moderate smoker together exceeds the product of the effects of being a carrier and being a moderate smoker considered separately (260, 419, 431). The measure of multiplicative interaction between carrier of a variant, G=1 and heavy smoker E=2 is given by OR_{12}/ OR_{10} $OR_{02.}$ This estimate measures the extent to which, on OR scale, the effect of being a carrier of a genetic variant and heavy smoker together exceeds the product of the effects of being a carrier and being a heavy smoker considered separately. Multiplicative interaction estimates less than 1 and greater than 1 would represent negative and positive multiplicative interaction respectively (260, 431).

4.12.3.4 Interaction on additive scale

Relative excess risk due to interaction ($RERI_{OR}$) was calculated as a measure of additive interaction (348). The RERI between moderate smokers and carriers estimate the extent to which, on OR scale, the effect of being a carrier of a genetic variant and moderate smoker together exceeds the sum of the effects of being a carrier and being a moderate smoker considered separately. This is given by

$$RERIe1 = RR_{11}\text{-}RR_{10}\text{-}RR_{01}+1$$

The RERI between heavy smokers and carriers estimate the extent to which, on OR scale, the effect of being a carrier of a genetic variant and heavy smoker together exceeds the sum of the effects of being a carrier and being a heavy smoker considered separately. This is given by

$$\text{RERIe2} = RR_{12}-RR_{10}-RR_{02}+1$$

If e1=0 if E=0, e1=1 if E=1 and G=g, the RERIe1 between carrier G=1 and moderate smoker E=1 was estimated by including a product term between g and e1in the regression model:

$$\text{logit}\ \{P\ (Y=1|G=g,\ E=e_1,\ C=c')\} = b_0 + \beta_1 g + \beta_2 e1 + \beta_3 g*e1 + b_4 c' \qquad \mathbf{Eq1}$$

Similarly, if e2=0 if E=0, e2=1 if E=2 and G=g, the RERIe2 between carrier G=1 and heavy smoker E=2 was estimated by including a product term between g and e2 in the regression model:

$$\text{logit}\ \{P\ (Y=1|G=g,\ E=e_2,\ C=c')\} = b_0 + \beta_1 g + \beta_2 e_2 + \beta_3 g*e_2 + b_4 c' \qquad \mathbf{Eq2}$$

RERIe1 and RERIe2 were calculated by fitting parameters from **Eq 1** and **Eq 2** respectively to the expression (419),

$$RERI_{OR} = e^{(\beta 1+\ \beta 2+\ \beta 3)} - e^{(\beta 1)} - e^{(\beta 2)} + 1$$

$RERI_{OR} > 0$ indicated positive additive interaction and $RERI_{OR} < 0$ indicated negative additive interaction.

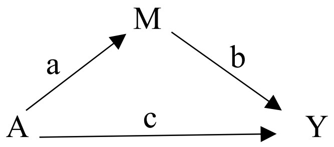
4.12.4 Analysis specific to Manuscript III

Manuscript III was geared towards estimating the extent to which the effects of two functional SNPs in CYP2A6 and ADH1B on SCCHN risk are mediated by heavy smoking and alcohol consumption, respectively. An additional aim was to estimate the proportions of excess risk attributable to the four causal pathways possible when the mediator not only mediates, but also interacts with the exposure. These objectives were achieved through analytical techniques developed for mediation and 4-way decomposition based on the counterfactual framework (432).

4.12.4.1 Causal Mediation

Causal mediation is the process by which an intermediate variable (mediator) explains how or why an exposure variable is related to the outcome variable through a chain of causal relation between the three variables. As represented by the causal diagram in ***Figure 10***, the mediator model hypothesises that an antecedent exposure variable *A (e.g., non-carrier of CYP2A6*2 SNP)* causes the mediator variable *M* (e.g., higher intensity of smoking), which in turn, causes the outcome *Y (e.g., SCCHN) (433-435)*. This chain of relation leads to an *indirect effect* of *A* to *Y* through *M*, and a *direct effect* of *A* on *M* not involving *M*.

Figure 10: Mediation model proposed by *Baron and Kenny* (1986)



Much of the work on mediation in the last 3 decades has been motivated by the seminal paper of Baron and Kenny (1986) (434). According to Baron and Kenny, four criteria are to be satisfied for a variable to function as a mediator: (i) variations in levels of *A* significantly affects the change in presumed mediator *M* (i.e., *Path a*), (ii) variations in the mediator significantly account for variation in the dependent variable *Y* (i.e., *Path b*), and (iii) when path a and b are controlled, a previously significant relation between *A* and *Y* (i.e., *Path c*) is no longer significant, with the strongest demonstration of mediation occurring when Path *c* is zero (complete mediation). They also popularised a parametric regression based approach, generally referred to as the "product method", to quantify mediation using the following regression models:

1) Mediator model: Mediator M regressed on exposure A=a

$$E\ (M|A=a) = \beta_0 + \beta_1 a$$

2) Outcome model: Outcome Y regressed on exposure A=a and mediator M=m

$$E\ (Y|A=a, M=m) = \theta_0 + \theta_{1a} + \theta_{2m}$$

Baron and Kenny proposed that the direct effect is the coefficient of the exposure in the outcome model, i.e., θ_1, and indirect effect is the product of coefficient of the exposure in the mediator model and the coefficient of the mediator in the outcome model, i.e., $\beta_1\theta_2$.

However, the third criteria proposed by Baron and Kenny, which implied that the association between *A* and *Y* need to be statistically significant for *M* to be a mediator and that *Path c* must be zero (complete mediation) when *Paths a* and *c* are controlled for, has been strongly critiqued (435). The assumption is invalid if either direct or indirect effects had opposite signs or if *Path a, or b or c* had different signs and violating a monotonicity assumption[3](343, 436). This is referred to as inconsistent mediation (e.g., one decreases the risk while other increases it). There is consensus that the association between *A* and *Y* needs not be statistically significant for *M* to be a mediator, and that mediation can be complete or partial (343, 435, 437). In addition, although Baron and Kenny allude to possibilities of both interaction and mediation occurring simultaneously, their product method is limited in not being applicable in the presence of exposure-mediator interaction or nonlinearities where the direct and indirect effect cannot be separated using the product method (437-439)}. These limitations have been addressed by recent advances in causal mediation and interaction analysis based on counterfactual causal framework provided certain no-confounding assumptions are met (*please refer to sub-section 4.11.4, page. no 80*) (343, 366, 438-440).

4.12.4.2 Direct and indirect effects in the presence of interaction under counterfactual framework

In this study, we defined the components of mediation using counterfactual framework. As per this framework, total effect of an exposure on the outcome can be decomposed into two non-overlapping components; the direct and indirect effect. These definitions were originally proposed on a risk difference scale and later defined on the odds ratio or relative risk scale (366, 437).

Total effect (TE)

Let *A* be an exposure (e.g., carrier or non-carrier of ADH1B*2 SNP) M a mediator (e.g., low vs high intensity of alcohol consumption), *Y* the outcome (SCCHN) and *C* a set of confounders. Let *A, M* and *Y* be binary variables. Within the counterfactual framework, for each individual in a study population, we may define Y_a as the potential outcome *Y* we would have observed if *A* had been set, possibly contrary to the fact, to *a* (*a* can have values 0 and 1). Under this scenario, on the

[3] Monotonicity assumption: The effect of exposure on the mediator or outcome, and the effect of the mediator on the outcome all have the same sign, i.e., either all not preventive (positive monotonicity assumption) or all not preventive (negative monotonicity assumption).

odds ratio scale, the total effect (*TE*) of *A* on *Y* conditional on *C=c*, comparing exposures *a1* and *a0* can be defined as:

$$OR^{TE} = \frac{\Pr(Y_{a1} = 1|c)/\{1-\Pr(Y_{a1} = 1|c)\}}{\Pr(Y_{a0} = 1|c)/\{1-\Pr(Y_{a0} = 1|c)\}}$$

In our context of ADH1B*2 and SCCHN, if we let a1 denote GG genotype and a0 denote AG/AA genotype, then the TE would be the OR for SCCHN comparing GG genotype with AG/AA genotype for individuals with covariate value *c*. For CYP2A6*2 and SCCHN, TE would be the effect of TT genotype on SCCHN if all individuals in the sample in fact had TT genotype with covariates *c* instead of AT/AA genotype.

Controlled direct effect (CDE)

Now, let Y_{am} be the potential outcome *Y* if, possibly contrary to the fact, *A* were set to *a* and *M* were set to *m* (*m* can have values 0 and 1). Similarly, we define *M*a as the potential mediator *M* if, possibly contrary to the fact, *A* were set to *a*.

The controlled direct effect (CDE) expresses how much the outcome would change on average if *M* was intervened to be fixed at level *m* (either 0 or 1) uniformly in the study sample, but the exposure was changed from level a0 to a1in the population. On the odds ratio scale, we can define the CDE of *A* on *Y* conditional on *C=c*, comparing exposures *a1* and *a0*, fixing *M* on *m* as:

$$OR^{CDE} = \frac{\Pr(Y_{a1m} = 1|c)\ /\{1 - \Pr(Y_{a1m} = 1|c)\}}{\Pr(Y_{a0m} = 1|c)\ /\{1 - \Pr(Y_{a0m} = 1|c)\}}$$

In our scenario of ADH1B, alcohol intensity and SCCHN, for example, the CDE at low intensity of alcohol consumption would be the OR for SCCHN comparing GG genotype with AG/AA genotype under a condition were, through an intervention, we managed to set the level of alcohol consumption at either high or low intensity in the whole sample. For CYP2A6*2, smoking intensity and SCCHN, CDE would be the effect of TT genotype on SCCHN if level of smoking intensity in the sample was intervened to be set up to either high or moderate.

Natural direct effect (NDE)

The natural direct effect (NDE) captures what effect of the exposure on the outcome would remain if we were to disable the pathway from the exposure to the mediator. Counterfactually, (NDE) expresses how much the outcome would change if the exposure were set at a level *A=a1* vs *A=a0* but for each individual, the mediator *M* was kept at the level it would have taken for that individual, in the absence of the exposure (*A=a0*). On the odds ratio scale, we can define the NDE of *A* on *Y* conditional on *C=c*, comparing exposures *a1* and *a0* and *M* takes the value *Ma0* (i.e., value of *M* if *A=a0*)

$$OR^{NDE} = \frac{\Pr(\mathrm{Y}_{a1Ma0} = 1|\mathrm{c})\,/\{1 - \Pr(\mathrm{Y}_{a1Ma0} = 1|\mathrm{c})\}}{\Pr(Y_{a0Ma0} = 1|\mathrm{c})\,/\{1 - \Pr(\mathrm{Y}_{a0Ma0} = 1|\mathrm{c})\}}$$

In our scenario of ADH1B, alcohol intensity and SCCHN, the NDE is the estimated effect of GG genotypes on SCCHN risk operating through pathways other than heavy alcohol intensities. For CYP2A6*2 and SCCHN, the NDE would be the effect of TT genotype on SCCHN risk through pathways other than heavy smoking intensity.

Natural indirect effect (NIE)

The natural indirect effect (NIE) captures the effect of the exposure on the outcome that operates through the mediator. Counterfactually, NIE is the extent to which the outcome would change on average if the exposure were fixed at level *A=a1* but the mediator *M* was changed from the level it would take if *A=a0* (i.e., *Ma0*) to the level it would take if A=a1 (i.e., *Ma1*). The conditional NIE of *A* on *Y* on the odds ratio scale, comparing the effect of mediator at levels Ma1 and Ma0 can be expressed as:

$$OR^{NIE} = \frac{\Pr(\mathrm{Y}_{a1Ma1} = 1|\mathrm{c})\,/\{1 - \Pr(\mathrm{Y}_{a1Ma1} = 1|\mathrm{c})\}}{\Pr(Y_{a1Ma0} = 1|\mathrm{c})\,/\{1 - \Pr(\mathrm{Y}_{a1Ma0} = 1|\mathrm{c})\}}$$

With respect to ADH1B, alcohol intensity and SCCHN, NIE estimates the effect of GG genotypes on SCCHN risk through heavy drinking intensities. For CYP2A6*2, smoking and SCCHN, NIE would be the effect of TT genotype on SCCHN risk mediated through heavy smoking intensity.

In the counterfactual scenario, on the ratio scale, the TE decomposes into a NDE and NIE as TE= NDE*NIE (366).

4.12.4.3 Regression models for direct and indirect effects

In a case-control study with binary outcome, mediator and exposure, TE, NDE, NIE and CDE can be estimated from the coefficients of the mediator and outcome parametric logistic regression models. The outcome model differs from the one specified under the product method in that this model, under the counterfactual setting can incorporate an exposure-mediator interaction term as given below (343, 437).

1) Mediator model: Mediator *M* regressed on exposure *A*=a

$$\text{logit}\ [\Pr\ (M=1|A=a, C=c)] = \beta_0 + \beta_1 a + \beta'_2 c \qquad \textbf{Eq3}$$

2) Outcome model: Outcome Y regressed on exposure A=a and mediator M=m

$$\text{logit}\ [\Pr\ (Y=1|A=a, M=m, C=c)] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c \qquad \textbf{Eq4}$$

where *am* is the product of the exposure and mediator representing interaction between them, and *C* is the set of confounders with value *c*. The estimates for TE, NDE, NIE and CDE can be derived by combining the coefficients of these equations as illustrated in the mathematical equations by VanderWeele and Vanstelandt 2010 (366, 441) .

4.12.4.4 Four-way decomposition

In our study, we hypothesized that smoking and alcohol intensities not only mediated the effect of CYP2A6*2 and ADH1B*2 on SCCHN respectively, but they also interacted with these respective exposures. Under this combined mediation and interaction scenario, the total effect for the outcome among those exposed can be decomposed into four non-overlapping pathways. The quantification of these causal pathways has been made possible by a recently developed four-way decomposition method by VanderWeele (2014, 2015, 2016) (432, 441, 442). In our case with a case-control study design, binary rare outcome, binary mediator and binary exposures, the effect measures are calculated on the risk ratio/odds ratio scale. In such a scenario, the decomposition involves decomposing the total excess risk (TE-1) for SCCHN among those carrying the specific genotype (i.e., TT for CYP2A6*2 and GG for ADH1B*2) into four components on the excess relative risk scale that represent: (i) *the controlled direct effect (CDE)* among exposed when the associated risk behaviour intensity is set to low/mild levels (i.e., component of excess risk attributed to neither mediation nor interaction with heavy intensity of the risk behaviour); (ii) *the reference interaction (INTref),* that represents the portion of the effect of the genetic exposure on SCCHN that requires the joint presence of high intensity of the associated risk behaviour (interaction alone), with the high intensity behaviour arising independently of the associated genetic exposure; (iii) *the mediated interaction (INTmed), that* represents the portion of the effect of genetic exposure that requires the joint presence of associated heavy intensity behaviour, with the heavy intensity behaviour arising as a consequence of the associated genetic exposure (both interaction and mediation), and (iv) *the pure indirect effect (PIE),* which represents the portion of the effect of genetic exposure on SCCHN that is due to genetic exposure -induced high intensity behaviour (mediation alone). The components representing CDE, INTref, INTmed and PIE are represented as cde_comp, intref_comp, intmed_comp and pie_comp, respectively. As described by VanderWeele (2015,2016) (343, 442) using mathematical equations, these four components add up to the total excess risk of the outcome among the exposed as:

$$TE_{a1} - 1 = cde\_comp + intref\_comp + intmed\_comp + pie\_comp$$

Alternatively, the proportions attributed to the CDE, INTref, INTmed and PIE can also be calculated under the ratio scale by dividing each of the components by total excess risk.

Proportions attributed to mediation and interaction and proportion eliminated

A summation of proportions related to PIE and INTmed components will give the *overall proportion* of the effect of genetic variant on SCCHN *mediated (PM)* by the heavy intensity risk behaviour (441). *The proportion of the effect attributable to interaction (PAI)* between the genetic variant and heavy intensity of risk behaviour is given by summation of proportions corresponding to INTref and INTmed components (441). *Proportion eliminated (PE)* is the proportion of effect of the genetic variant on SCCHN risk that can be eliminated in the population if the level of the risk behaviour was decreased to that of low/mild intensity in the population. This is given by summing up the proportions attributable to INTref, INTmed and PIE components (441).

The components of the four-way decomposition can be derived by specific combinations of coefficients from the mediator **(Eq3)** and outcome **(Eq4)** logistic regression models as illustrated through the mathematical equations by VanderWeele 2014 and 2016 (personal communication, June 2016) (432) (442).

Overall analytic techniques used in Manuscript III

The CYP2A6*2–smoking-SCCHN and ADH1B*2-alcohol-SCCHN analyses were performed only among smokers and alcohol consumers, respectively. The exposure, mediator and outcome variables were coded to maintain positive monotonicity and calculate excess risks. In our study, participants were sampled based on SCCHN outcome and not based on smoking or alcohol intensities. Hence, although outcome regression parameters of Y that are required for calculation of CDE, NDE, NIE and the four-way components could be consistently estimated through regression model **Eq4**, the mediator regression (**Eq3**) cannot be fit on the full sample. Since SCCHN is a rare outcome, to make the distribution of mediator M among controls approximate the distribution in the population, we fit the mediator models among controls only (343). For CYP2A6*2-smoking-SCCHN, SCCHN was regressed on the CYP2A6*2, cigarettes per day, their product term (denoting interaction) and associated potential confounders (outcome model). Next, cigarettes per day was regressed on CYP2A6*2 and potential confounders among controls. For ADH1B*2-alcohol-SCCHN, the outcome model was fit on ADH1B*2, ethanol per day, their product term and associated potential confounders. For the mediator model, ethanol per day was fit on ADH1B*2 and potential confounders among controls. Alternatively, a waiting approach was used to address the case-control study design where we fit the mediator regression model in the full sample of cases and controls weighted on inverse of sampling fraction. This was achieved by weighting the cases with p/q and controls with (1-p)/(1-q) where p is the prevalence of SCCHN in Canada during the study period, and q the proportion of cases in the study (9, 366).

An indicator variable for ex-smokers and ex-alcohol consumers was added in the smoking and alcohol related models, respectively, to account for time since cessation/stoppage of use of the respective products (403). Effect estimates and associated proportions were obtained by combining parameters from these two models according to their corresponding analytical equations (343, 441). For a rare disease outcome, ORs approximate RR and hence estimates for TE, direct and indirect effects were interpreted as RRs.

Bootstrapping for calculating 95% CI

No automated statistical codes were available to retrieve the standard errors and confidence limits of the estimated parameters in this analysis. The 95% CI for the parameters derived using the regression models were obtained using bootstrapping (443,444). In this procedure, 2000 bootstrapping replications with replacement were taken from the original sample (each replication the size of the original sample). Each parameter was estimated in each of these replications and the 95% CI for CDE, NDE, NIE, TE and the components of the four-way decomposition were estimated as the 2.5th and 97.5th percentile of the resulting distribution.

4.12.4.5 Software codes written for mediation and 4-way decomposition for Stata statistical package

Statistical analyses are carried out by encoding theory associated mathematical formulas into software codes specific to various statistical software programs (e.g., Stata, SAS, PASW). The codes for standard mediation (2-way) and 4-way decomposition were written for SAS software by Valarie, VanderWeele (432, 437, 441) .Although Stata macros were available for the 2-way mediation analysis (e.g., PARAMED), there was lack of codes of any form for carrying out the 4-way decomposition in Stata. Since our analysis mostly depended on Stata, codes for 4-way decomposition where exclusively written for carrying out this analysis in this thesis work using the mathematical equations provided by VanderWeele (2014, 2016) for a binary outcome, binary mediator and binary exposure scenario (ratio scale) (432, 442). In this code, we also included an alternative method to conduct the 2-way mediation analysis using mathematical equations provided by VanderWeele and Vansteelandt (direct and indirect effects) (343, 366) For retrieving the 95% CI's for all estimates we used the codes for bootstrapping in Stata (443). Stata codes are provided in *manuscript III, Supplemental material, eAppendix, pages 182-186.*

Chapter 5

Manuscript 1

Original Research Report

Socioeconomic position and oral cancer: life-course models

Akhil Soman ThekkePurakkal[1], Ashley Isaac Naimi[2], Sreenath Arekunnath Madathil[1,3], Shahul Hameed Kumamangalam Puthiyannal[1], Gopalakrishnan Netuveli[4], Amanda Sacker[5], Nicolas F Schlecht[6], Belinda Nicolau[1]*

[1]Division of Oral Health and Society, Faculty of Dentistry, McGill University, Montreal, Canada; [2]Department of Epidemiology, University of Pittsburgh, Pennsylvania, United States of America. [3]Epidemiology and Biostatistics Unit, INRS-Institut Armand-Frappier, Laval, Canada, [4] Institute of Health and Human Development, University of East London, London, United Kingdom, [5]UCL Research Department of Epidemiology and Public Health, University College London, United Kingdom, [6]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, United States of America.

***Corresponding author**

Division of Oral Health and Society - Faculty of Dentistry - McGill University
2001 McGill College, suite 527, Montréal, QC, Canada H3A 1G1
Tel: 514-3987203 ext. 094655 Fax: 514-3987220
Email: belinda.nicolau@mcgill.ca

Keywords: case-control study; childhood; inverse probability weighting; marginal structural models; quasi-likelihood criterion; socioeconomic position; time-varying.

Abstract

Disadvantageous life-course socioeconomic position (SEP) is associated with the risk of several chronic oral diseases including oral cancer. However, most studies investigating these association do not take into consideration the time-varying nature of SEP and associated behavioural confounders, which limits our understanding of these links. Using analytical strategies developed to address these limitations, we estimated the association between life-course SEP and oral cancer risk under three life-course models: critical period, accumulation and social mobility.

We recruited incident oral cancer cases (N=350) and controls (N=371) frequency-matched by age and sex from two main referral hospitals in Kozhikode, Kerala, India between 2008 and 2012. We collected information on childhood (0-16 years), early adulthood (17-30 years) and late adulthood (above 30 years) SEP and behavioural factors along the life span using interviews and a life-grid technique. Odds ratios (OR) and 95% confidence intervals (CI) were estimated for the association between life-course SEP and oral cancer risk using inverse probability weighted marginal structural models to adjust for confounding. Fit of the models were compared using a quasi-likelihood criterion.

Childhood and early adulthood SEP (advantageous vs. disadvantageous) were associated with oral cancer risk [(OR=2.76, 95% CI: 1.99, 3.81) and (OR=1.84, 95% CI: 1.21, 2.79), respectively]. In addition, participants who were in a disadvantageous (vs. advantageous) SEP during the three periods of life, had an increased oral cancer risk [OR=4.86, 95% CI: 2.61, 9.06]. Childhood to early adulthood social mobility model and over-all life-course trajectories indicated strong influence of exposure to disadvantageous SEP in childhood on the risk for oral cancer. The childhood critical period model fit our data best relative to other models.

Using several rigorous modelling approaches that can be applied to other oral health problems, our study provides empirical evidence that disadvantageous childhood SEP is critical for oral cancer risk.

INTRODUCTION

Oral cancer, a disease with low survival rates, and high morbidity, affects roughly 300,000 people each year, leading to approximately 145,000 deaths worldwide (Ferlay et al.; Warnakulasuriya 2009). Developing countries bear two-thirds of the global burden, with India accounting for 25% of new cases and 35% of deaths and where incidence rates have increased considerably in the last decade (Ferlay et al.). Risk behaviours such as paan chewing, cigarette and bidi smoking, and alcohol consumption, which are more common among those in a disadvantageous socioeconomic position (SEP), are the strongest risk factors for oral cancer (Petti 2009). Yet exposure to cumulative disadvantageous SEP has been independently associated with increased risk for the disease (Conway et al. 2008). The reason for the independent effect of SEP might be that most studies looking into this association have overlooked the dynamic nature of SEP and behavioural risk factors over the individual's life-course (Stringhini et al. 2010). The life-course framework takes into account the effects of multiple risk factors spread across multiple points in life (Ben-Shlomo and Kuh 2002). Although imperative, the relation between SEP and oral cancer has not yet been explored through the lens of multiple life-course models within a single study.

The life-course framework is not only beneficial to estimate the cumulative effect of SEP on the outcome (accumulation model), but also the timing of exposure to disadvantageous SEP at key periods of life (critical period model) that could cause initiation of oral cancer. In addition, the framework allows the investigation of life trajectories caused by interaction of SEP exposures between multiple periods of life (social mobility model) that can alter an individual's risk of cancer (Ben-Shlomo and Kuh 2002).

Furthermore, the life-course framework implies that the relation between SEP at different time points and behavioural risk factors are likely subject to complex time-varying feedback loops (VanderWeele et al. 2016). Yet, investigators often fail to account for these relations between SEP over the life-course and other time-varying covariates (Conway et al. 2008; Hallqvist et al. 2004; Mishra et al. 2013). Therefore, by appreciating the time-varying nature of SEP and these variables, we estimated the association between SEP measured over three periods of life and oral cancer risk using a case-control study from India. We further assessed whether the associations conformed better to a critical period, accumulation or social mobility model.

METHODS

Data for this analysis were drawn from the Head and Neck Cancer (HeNCe) Life course study, an IRB approved multicentre hospital-based case-control study investigating the aetiology of head and neck cancers. Adult participants (N=721) were recruited from the outpatient clinics at two major teaching hospitals, the Government Dental and Medical College and Hospitals, Kozhikode, Kerala, South India between 2008 and 2012. The study design and sample have been described in detail elsewhere (Laprise et al. 2016). Briefly, cases (N=350) included incident, histologically confirmed stage I to IV squamous cell carcinomas (C01-C06, and C09 under International Classification of Diseases 10 Version:2010) of oral cavity diagnosed during the study period. Non-cancer controls (N=371), frequency matched to each identified case by 5-year age group and sex, were randomly selected from 8 outpatient clinics in the same hospitals from a list of non-chronic diseases (distribution reported elsewhere) (Madathil et al. 2016), not strongly associated with tobacco and alcohol consumption (with no single diagnostic group contributing to more than 20% of the total). Recruitment of controls followed an incident density sampling technique.

Data were collected through one-on-one semi-structured interviews using a questionnaire with life-grid technique. Help of a proxy respondent was sought for consenting participants who had difficulty speaking due to disease status. Re-interviews were conducted for 46 randomly selected participants, 6 to 12 weeks after the original interview to test the reliability of the data collected. Informed consent was obtained from all participants prior to inclusion in this study.

Life-course socioeconomic position

Information on housing conditions was used to derive the SEP exposures. We created an asset index to represent SEP using responses to questions about various assets (housing characteristics, durable assets and access to services) (Gwatkin et al. 2000), available at the participant’s longest place of residence during three time periods: childhood (0-16 years), early adulthood (17-30 years), and late adulthood (above 30 years). Responses to each question were binary coded (Supplemental Appendix file, eTable 1) and a tetrachoric correlation matrix was created for each period (Supplemental Appendix file eTable 2-4). Principal component analysis was conducted on the correlation matrices and the first component that explained maximum variance (approximately 65%) was extracted (Filmer and Pritchett 2001). Continuous scores were predicted from these components. The scores for each period were then dichotomized (cut-off at 50th percentile among controls), generating a binary SEP variable (0= advantageous SEP, 1= disadvantageous SEP) for childhood, early adulthood and late adulthood periods each. This variable represented the SEP exposure for each of the three respective critical period models. A four-category variable representing the accumulation model was created by summing the number of periods of disadvantageous SEP (0, 1, 2 and 3). Finally, to test the social mobility models (childhood to early adulthood, and early to late adulthood) we combined the binary SEP variables in respective periods into two variables with four categories representing stable advantageous SEP, upward mobility, downward mobility, and stable disadvantageous SEP. Additional details are provided in Supplemental appendix file eAppendix and eTables 1-5.

Potential confounders

Information on potential confounders and mediators was collected from a set of time-invariant and time-varying factors. The factors included baseline exposures [age (continuous), sex (binary), caste i.e., hierarchy in Hindu religion based on occupation, (higher, middle, low)], education, and time-varying exposures (cigarette smoking, bidi smoking, paan chewing and alcohol consumption). Education was measured by the number of years of schooling and dichotomized based on the participants' birth cohort (participant's year of birth in our study ranged from 1921 to 1979) to account for the major social and educational reforms in Kerala in the 1950's (Kerala Education bill. 2009). We collected detailed lifetime information on risk behaviours (e.g., duration, quantity, and type of cigarette and bidi smoking, paan chewing, and alcohol consumption) as described elsewhere (Madathil et al. 2016). This information was used to compute continuous measures of pack-years of cigarette and bidi smoked, chew-years of paan, and number of standard drinks of alcohol per week corresponding to multiple life periods (Madathil et al. 2016).

The directed acyclic graph in Figure 1 represents the assumed temporal relations between these variables. Although this is a case-control study, our unique data collection procedure allowed us flexibility to appreciate the temporal relation between vectors representing potential confounders (C0: baseline covariates, C1: 0-16 years, C2a: 17-23 years, C2b: 24-30 years, C3a: 31-50 years, C3b: above 50 years), SEP exposure in the three periods of life and oral cancer. We adjusted for categorical and continuous confounders using indicator coding and restricted cubic splines respectively. Although biological samples were collected and analysed for HPV, no HPV was detected in any of the samples. Hence this variable was not included in the analysis (Laprise et al, 2016).

Statistical methods

Our primary aim was to assess the relation between life-course SEP and oral cancer under the three conceptual life-course models. Due to their time-varying nature, SEP and related confounders may also act as mediators. Consequently, standard regression methods may produce biased estimates of exposure-outcome association, regardless of the method used to adjust for confounders. We therefore used inverse probability weighted marginal structural models to account for such confounding and derive our estimates (Robins et al. 2000). The inverse probability weighting creates a pseudo reweighted sample where the exposure is independent of the measured potential confounders. We assumed that our case-control data arose from an underlying cohort representing the population of interest (Langholz 2007).

Weights were derived by fitting a separate exposure model for each period of life and were computed as the inverse of the conditional probability of falling in the disadvantageous SEP category at each time period. To account for the case-control design, each exposure model was weighted by sampling fraction. The weights were stabilized by the marginal probability of falling in the disadvantageous SEP category at each time period. Once the stabilized inverse probability weights were computed, they were further combined with time dependent sampling weights to account for the case-control design (Leffondre et al. 2010). Sampling weights were defined as:

$$\text{Sampling weight} = [(1-\prod)/\prod] * \text{ncases}/\text{ncontrols},$$

where $\prod$ is the annual prevalence of oral cancer in India during the four-years of study, and ncases and ncontrols are the number of cases and controls in our sample. Finally, unadjusted logistic regression marginal structural outcome models were fit for each life-course model. In general, the outcome model took the form:

$$\text{Logit } \{\Pr[Yg(A)=1]\} = \propto + \beta 1\ g(A)$$

where g(A) is a function of exposure, SEP, specific to each model. Additional technical details including those on exposure models and characteristics of stabilized weights are provided in Supplemental Appendix file eTables 6 - 8.

We also fit a saturated all-trajectories model in which the other three models are nested (Mishra et al. 2009). This model contained eight possible trajectories formed by binary SEP exposures measured over three periods of life. The fit statistic for this model has been considered as a reference for comparing other life-course models (Mishra et al. 2009). To assess how each model fit our data, we used a weighted quasi-likelihood information criterion (QICw) proposed for marginal structural models (Platt et al. 2013). The model with the lowest QICw value was selected as the relatively best fit model.

Thirty-seven participants (17 controls and 20 cases) had missing values related to the main exposure. Therefore, we present our results on complete case analysis of 684 participants. Analyses were performed using Stata, version 13 SE (StataCorp. 2013, College Station, TX: StataCorp LP.) and SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina). Annotated Stata codes are provided in Supplemental Appendix file eTable 8.

RESULTS

Table 1 shows socio-demographic characteristics and measured potential confounders among cases and controls. The participants' age ranged from 32 to 88 years (mean=61 years) and the majority of the cases had a low level of education (78% of cases vs 50% of controls). The help of a proxy respondent was sought more rarely for controls (3%) than cases (14%). The majority of the participants belonged to the middle caste (81% of controls, 70% of cases). On an average, cases had a higher propensity for practicing all habits in each life period except for cigarette smoking. A higher proportion of cases than controls were exposed to disadvantageous SEP (60% vs 36% in childhood, 63% vs 35% in early adulthood, and 62% vs 33% in late adulthood).

Table 1 Descriptive characteristics of oral cancer cases and controls from Kerala, India, 2008-2012, (n=684)

	Controls (n=354)		**Cases (n=330)**	
	N (%)	**mean (SD)**	**N (%)**	**mean (SD)**
Age in years		61 (11)		61 (11)
Sex				
Female	163 (46)		149 (45)	
Male	191 (54)		181 (55)	
Education				
High	178 (50)		74 (22)	
Low	176 (50)		256 (78)	
Respondent type				
Use of proxy	11 (3)		46 (14)	
No use of proxy	343 (97)		284 (86)	
Caste				
Higher	51 (14)		26 (8)	
Middle	285 (81)		231 (70)	
Lower	18 (5)		73 (22)	
During childhood (0-16 years)				
Cigarette smoking (pack-years)		0.08 (0.59)		0.05 (0.36)
Bidi smoking (pack-years)		0.14 (0.59)		0.23 (0.78)
Paan chewing (chew-years)		0.41 (3.50)		4.09 (9.29)
Alcohol consumption (drinks per week)		0.17 (2.29)		0.38 (2.70)
During early adulthood (17-23 years)				
Cigarette smoking (pack-years)		0.57 (1.73)		0.36 (1.87)
Bidi smoking (pack-years)		0.59 (1.94)		0.88 (2.10)
Paan chewing (chew-years)		1.72 (6.80)		12.44 (18.39)
Alcohol consumption (drinks per week)		2.54 (12.52)		4.30 (15.15)
During early adulthood (24-30 years)				
Cigarette smoking (pack-years)		1.26 (3.46)		0.88 (3.26)
Bidi smoking (pack-years)		0.91 (2.73)		1.60 (3.44)
Paan chewing (chew-years)		3.51 (11.67)		22.27 (28.24)
Alcohol consumption (drinks per week)		5.52 (33.18)		11.17 (43.33)

Table 1 continued…

	Controls (n=354)		**Cases (n=330)**	
	N (%)	**mean (SD)**	**N (%)**	**mean (SD)**
During late adulthood (31-50 years)				
Cigarette smoking (pack-years)		5.56 (14.33)		3.32 (9.58)
Bidi smoking (pack-years)		2.20 (7.59)		4.45 (9.19)
Paan chewing (chew-years)		15.76 (49.06)		94.95 (94.06)
Alcohol consumption (drinks per week)		6.65 (40.27)		15.37 (48.38)
Alcohol consumption (drinks per week)		2.47 (11.97)		11.71 (44.82)
SEP over the life-course				
Childhood SEP (0-16 years)				
Advantageous SEP	227 (64)		131 (40)	
Disadvantageous SEP	127 (36)		199 (60)	
Early adulthood SEP (17-30 years)				
Advantageous SEP	230 (65)		121 (37)	
Disadvantageous SEP	124 (35)		209 (63)	
Late adulthood SEP (above 30 years)				
Advantageous SEP	237 (67)		125 (38)	
Disadvantageous SEP	117 (33)		205 (62)	

SD: Standard deviation; SEP: Socioeconomic position.

Table 2 presents the association between life-course SEP and oral cancer under each conceptual model. Among the critical period models, being exposed to disadvantageous (vs. advantageous) SEP in childhood and early adulthood was associated with an increased risk for oral cancer (childhood: OR = 2.76, 95% CI: 1.99, 3.81; early adulthood: OR=1.84, 95% CI: 1.21, 2.79). In contrast, relative to an advantageous SEP, exposure to disadvantageous SEP in late adulthood was not associated with the disease (OR=0.92, 95% CI: 0.55, 1.54).

For the accumulation model, the risk of oral cancer increased with additional periods of socioeconomic disadvantage. Relative to never experiencing a period of disadvantageous SEP, experiencing one, two, and three periods of disadvantageous SEP yielded ORs of 2.56 (95% CI: 1.34, 4.87), 2.71 (95% CI: 1.44, 5.09), and 4.86 (95% CI: 2.61, 9.06), respectively.

Table 2 Odds ratios and 95% confidence intervals for risk of oral cancer under different life-course socioeconomic models in the study sample from Kerala, India, 2008-2012 (n=684)

Life-course SEP models	**Levels of SEP**	**Controls**	**OR (95% CI)**	**QICw**
	(0 = Advantageous,	**/Cases**		
	1= Disadvantageous)	**N**		
Critical period models				
Childhood SEP	0[a]	227/131	Ref	
	1	127/199	2.76 (1.99, 3.81)	6597.1
Early adulthood SEP	0[a]	230/121	Ref	
	1	124/209	1.84 (1.21, 2.79)	7180.5
Late adulthood SEP	0[a]	237/125	Ref	
	1	117/205	0.92 (0.55, 1.54)	8659.4
Accumulation model				
Number of periods spent	0 periods[a]	162/53	Ref	
in disadvantageous SEP	1 period	71/63	2.56 (1.34, 4.87)	
over the life-course	2 periods	66/92	2.71 (1.44, 5.09)	8629.5
	3 periods	55/122	4.86 (2.61, 9.06)	
Social mobility models				
Childhood to early adulthood SEP				
Stable advantageous	0,0[a]	190/79	Ref	
Upward mobility	1, 0	40/42	3.19 (1.83, 5.55)	
Downward mobility	0, 1	37/52	2.75 (1.57, 4.83)	7120.5
Stable disadvantageous	1,1	87/157	4.06 (2.62, 6.28)	
Early to late adulthood SEP				
Stable advantageous	0,0[a]	183/71	Ref	
Upward mobility	1, 0	54/54	1.52 (0.80,2.87)	
Downward mobility	0, 1	47/50	0.81 (0.40, 1.62)	8632.5
Stable disadvantageous	1,1	70/155	1.53 (0.68, 3.41)	
Saturated all-trajectories	0, 0, 0[a]	162/53	Ref	
model[b]	1, 0, 0	21/18	4.37 (1.83,10.85)	
(All SEP trajectories	0, 1, 0	22/19	2.58 (1.15,5.80)	
across 3 life periods)	0, 0, 1	28/26	1.00 (0.40,2.53)	
	1, 1, 0	32/35	3.36 (1.61,6.99)	8514.2
	1, 0, 1	19/24	2.61 (1.16,5.89)	
	0, 1, 1	15/33	2.25 (0.82,6.21)	
	1, 1, 1	55/122	4.86 (2.61, 9.06)	

SEP: socioeconomic position; QICw: weighted quasi likelihood criterion.

Note: ORs presented are adjusted for time-invariant and time-varying covariates using IPW

[a] Reference category/ level within each SEP variable representing the specific life-course model.

[b] Categories/levels in the saturated all-trajectories model variable represents all possible 8 trajectories created from each binary SEP measure representing the three time periods.

Under the social mobility models, for childhood to early adulthood mobility, compared to stable advantageous SEP group, downward mobile (OR=2.75, 95% CI: 1.57, 4.83), upward mobile (OR=3.19, 95% CI: 1.83, 5.55) and stable disadvantageous (OR=4.06, 95% CI: 2.62, 6.28) trajectories were associated with increased risk for oral cancer.

The all-trajectories model (Table 2) showed that compared to non-exposure to disadvantageous SEP in all periods (0, 0, 0), the magnitude of ORs associated with trajectories in which individuals were exposed to disadvantageous SEP in childhood (1, 0, 0: OR= 4.37, 95% CI:1.83, 10.85 ; 1, 1, 0: OR=3.36, 95% CI 1.61, 6.99; 1, 0, 1: OR= 2.61, 95% CI: 1.16, 5.89; 1, 1, 1: OR=4.86, 95% CI: 2.61, 9.06) were larger than those of trajectories where participants were never exposed in childhood (0, 1, 0: OR=2.58, 95% CI: 1.15,5.80; 0, 0, 1: OR =1.00, 95% CI: 0.40, 2.53; 0, 1, 1: OR=2.25, 95% CI: 0.82, 6.21).

Comparing the QICw values for all statistical models tested, the childhood critical period model had the lowest QICw value.

DISCUSSION

In this study, we compared results from multiple life-course models to explore the pathways underlying the association between SEP across life-course and oral cancer risk. Our findings indicate that an exposure to disadvantageous SEP in childhood may play a critical role in the development of oral cancer later in life.

Considered as the most fundamental of all life-course models (Blane et al. 2007), the accumulation model implies cross-sectional clustering of (dis)advantages driven by social structure that accumulate longitudinally (Blane 1995). In our study, we found that the risk for oral cancer increased with the accumulation of disadvantageous SEP periods over the course of life. This finding is similar to the monotonically increasing risk pattern identified in life-course studies investigating other health outcomes (Bernabe et al. 2011; Peres et al. 2011; Pollitt et al. 2005). However, exploring other life-course models within this study provided further insight into this overall exposure-outcome relationship.

In line with studies investigating other chronic diseases including cancers (Krishna Rao et al. 2015; Nicolau et al. 2007; Pollitt et al. 2005; Vohra et al. 2016), our findings indicated that disadvantageous SEP during childhood and early adulthood increased the risk of oral cancer. The magnitude of association was higher for childhood. Interestingly, our findings from other models tested as well converged to indicate that childhood is a critical period for the risk of oral cancer. In our social mobility analyses, the magnitude of the OR associated with upward mobility from childhood to early adulthood was higher than that of downward mobility. This reflects the higher impact of disadvantageous SEP in childhood compared to the same exposure in early adulthood as observed from the critical period models. Also, the estimates from the all-trajectories model provided further evidence for the critical role an exposure to disadvantageous SEP in childhood may play in the increased risk for oral cancer later in life. The QICw values indicated that relative to other models, the childhood critical period model, fit our data best.

The risk behaviours considered as time-varying variables in our study are usually considered to be affected by SEP and hence as mediators of the relationship between SEP and adult health outcomes. However, such behaviours (e.g., alcohol consumption) have been considered as determinants of socioeconomic consequences, especially in developing societies (WHO 2014). Although the state of Kerala ranks high in social development relative to other states, the state has one of the highest alcohol consumption levels in India (Kerala's human development report, 2005). Furthermore, alcohol consumption is highly correlated with tobacco habits. This strengthens our methodology considering the dual nature of these exposures as potential confounders and mediators.

The statistical evidence presented here does have biological plausibility. The adverse effects of an accumulation of socioeconomic disadvantage over an individual’s life span can manifest biologically through increased allostatic load, impaired immune response, and specific genetic or epigenetic changes resulting in oral cancer (Ben-Shlomo and Kuh 2002; McEwen 1998; Stringhini et al. 2015). Of particular relevance to the critical period model, childhood represents a specific time of rapid development and vulnerability when exposures produce irreversible biological damage (Barker 1990). Childhood SEP captures different dimensions of adversity (e.g., poor nutrition) that may initiate the above carcinogenic processes in the oral cavity (Borghol et al. 2012; Fagundes and Way 2014).

There are several challenges in interpreting the results of our study. For example, although the results from the mobility models were in line with the gradient constraint hypothesis of social mobility (Blane et al. 1999), the empirical difficulty in defining social mobility and associated life-course trajectories from limited time periods has been discussed in the literature (Pollitt et al. 2005). In addition, although our sample size did exceed that of the majority of case-control studies exploring the SEP-oral cancer association (Conway et al. 2008; Krishna Rao et al. 2015), the results from social mobility and all-trajectories models tested were limited by the low numbers in some of the trajectories. There is also the potential for measurement error affecting our results. Our measure of SEP, an asset index (Gwatkin et al. 2000), may not have captured all aspects of SEP. However, asset indices serve as indicators of wealth and are particularly relevant to less industrialized societies (Howe et al. 2009). Developing countries like India are more prone to high rates of short-term economic shock, and lack concrete socioeconomic classification systems such as those used in developed countries (Galobardes et al. 2006; Gwatkin et al. 2000; Howe et al. 2009). In addition, the cut-off points chosen to divide confounder vectors C2 into C2a and C2b (23 years) and C3 into C3a and C3b (50 years) were not based on statistical modelling, which might be a source of potential misclassification. However, we expect this to be negligible as our cut-off selection was based on the assumption that disadvantageous SEP during earlier stages of life (e.g., 17-23 years for early adulthood and 31-50 for late adulthood) is less likely to drive risk behaviours during these early stages (Figure 1). Moreover, behavioural factors in these earlier stages have higher probability to causally effect SEP later in life. Finally, although recall bias is a well-recognized problem in case-control studies, we attempted to mitigate this by using a life-grid tool that has been shown to improve recall (Berney et al. 1997). Relative measures of test-retest reliability for housing assets from this study are presented in Supplemental Appendix file, eTables5.

Several methodological strengths of our study also merit consideration. Cohort studies are not always feasible to investigate rare health outcomes. Our rigorous data collection procedures allowed us to analyse the temporal associations between exposures and potential confounders, as well as their time-varying nature. Leffondre et al (Leffondre et al. 2010), developed weighted partial likelihood estimators for time-dependent exposures in a case-control setting. However, these estimators have not been extended to time-varying confounders. In this study, we employed a novel approach by combining these estimators with inverse probability weighting to account for both time-varying exposures and confounders. An additional complication in social epidemiology stems from the non-manipulable nature of social exposures such as SEP. We believe that, as there are numerous ways in which an individual may be "assigned" to a given level of SEP, each of which may have different impacts on the risk of oral cancer, interpreting associational estimates as causal effects is not possible (Naimi and Kaufman 2015). However, our results do provide valid estimates of the socioeconomic distribution of oral cancer risk in Kerala, India.

CONCLUSION

Using several rigorous modelling approaches under the time-varying framework, we investigated the association between SEP and oral cancer risk using different conceptual life-course models. Multiple life-course models provided empirical evidence for the independent association of SEP during childhood on oral cancer risk. Addressing issues related to unfavourable social circumstances early on in life may be beneficial in reducing the long-term burden of oral cancer in high-risk regions like India.

Acknowledgements

We would like to thank our collaborators at the Indian site Dr I V Varghese, Dr Shameena Shiraz, and the HeNCe Life Study team, especially Dr Geneviève Castonguay for her valuable assistance in coordinating and monitoring the study. We would also like to thank Dr. Jay Kaufman, Professor, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada, whose guidance helped in improving the manuscript.

Funding

This work was supported by the Canadian Institutes of Health Research [grant number MOP 81172, MOP 111207]; Ministère du Développement économique, de l'Innovation et de l'Exportation du Québec: Programme de soutien à la recherche (PSR), volet: Soutien à des initiatives internationales de recherche et d'innovation (SIIRI). Belinda Nicolau holds a Canada Research Chair in Life Course Oral Epidemiology. Nicolas Schlecht is a Miriam Mandel Faculty Scholar in Cancer Research.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Ethical approval: The study was approved by IRB and ethics committees of Government Dental and Medical colleges, Calicut, Kerala, India. All procedures performed in this study which involved human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

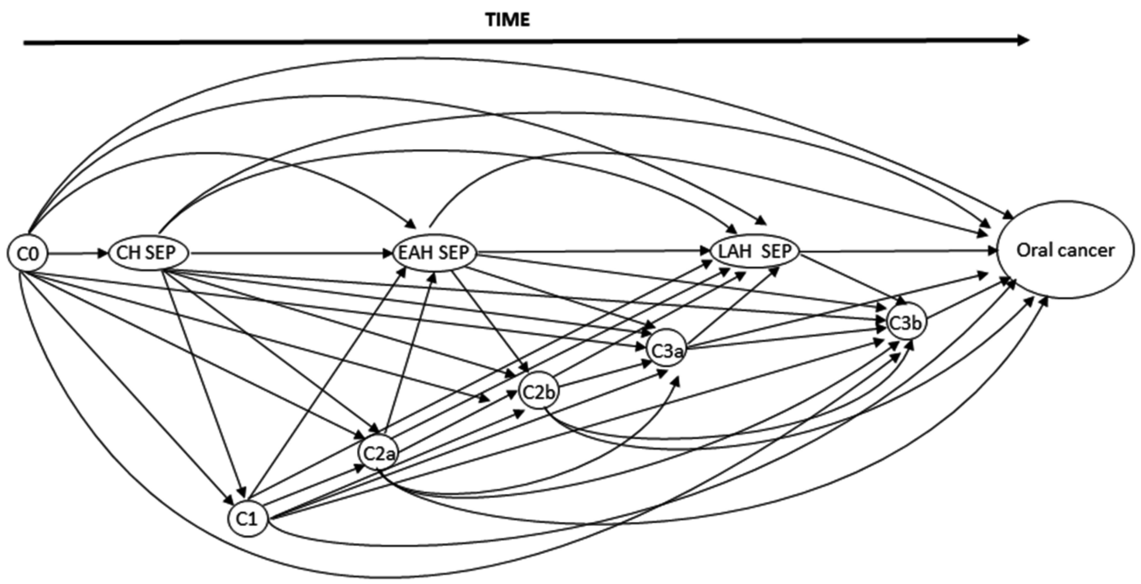


Figure caption

Figure 1. Directed acyclic graph (DAG) representing the relationship between exposure, covariates and outcome, in the study from Kerala, India, 2008-2012 (n=684). **Oral cancer**: Outcome; **SEP:** Socioeconomic position, main exposure; **CH SEP**: SEP during childhood; **EAH SEP**: SEP during early adulthood; **LAH SEP**: SEP during late adulthood; **C0**: Vector representing baseline covariates, age, sex, caste i.e., hierarchy in Hindu religion (potential time-invariant confounders); **C1**: Vector representing education, health related behaviours (time-varying) of cigarette smoking, bidi smoking, paan chewing and alcohol consumption recorded during 0-16 years of age; **C2a**: Vector representing health related behaviours recorded during 17-23 years of age; **C2b**- Vector representing health related behaviours recorded during 24-30 years age; **C3a**- Vector representing health related behaviours recorded during 31-50 years of age; **C3b**- Vector representing health related behaviours recorded above 50 years

5.1 Supplemental material - Manuscript I

Socioeconomic position and oral cancer: life-course models

Supplemental appendix file

eAppendix

Measurement of socioeconomic position (SEP)

Asset/wealth index was created from a list of questions on various assets (housing characteristics, durable assets and access to services) available at the participant's longest place of residence during three time periods: childhood (0-16 years), early adulthood (17-30 years), and late adulthood (above 30 years). As given in eTable 1, information on nine assets/items from childhood, eleven from early adulthood and twelve from late adulthood were used. The nominal responses to each of these questions were binary coded based on type of material used and facilities available, contextual to Kerala, India. A tetrachoric correlation matrix (Debelak and Tran 2013)[4] was created from these binary variables for each life period (eTables 2,3,4). If any variable correlated highly (|0.8|) with other variables, only one variable from the group of correlated variables were retained for further analysis. In addition, variables were excluded in stepwise manner until a factorable correlation matrix with Kaiser-Meyer-Olkin (KMO) value > 0.7 was attained for each period separately (Balen et al. 2010)[5]. Assets with low test-retest reliability (inter class correlation) were also removed (eTable 5). Final variables retained in the matrix for each period were; Childhood: crowding, floor, wall, window, water, bath, clock, KMO=0.832; Early adulthood: crowding, wall, window, water, clock, bicycle; KMO=0.771; Late adulthood: Crowding, wall, window, water, clock, radio, television, phone, KMO=0.801. A principal component analysis was conducted without rotation on the final correlation matrices to assess dimensionality of the assets, and the first component that explained maximum variance in each life period (childhood 1st component explained 65% of variance, 64% each for early and late

[4] Debelak R, Tran US. 2013. Principal component analysis of smoothed tetrachoric correlation matrices as a measure of dimensionality. Educational and Psychological Measurement. 73(1):63-77.

[5] Balen J, McManus DP, Li Y-S, Zhao Z-Y, Yuan L-P, Utzinger J, Williams GM, Li Y, Ren M-Y, Liu Z-C et al. 2010. Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of hunan province, china. Emerging Themes in Epidemiology. 7:7-7.

adulthood) was extracted (Filmer and Pritchett 2001)[6]. Scores were predicted out of these components. Each of the continues score for each life period was then dichotomized using the median of the distribution as cut-off generating respective binary variable representing SEP (0= advantageous SEP, 1= disadvantageous SEP) for childhood, early and late adulthood**.**

SEP exposure measure for critical period models

The binary variable (0-advantageous SEP, 1-disadvantageous SEP) representing SEP in childhood, early, and late adulthood were used as the main exposure in the critical period model representing each of these life periods.

SEP exposure measure for accumulation model

A summation of the binary variables representing SEP in each life period generated a variable with four categories with increasing periods of exposure to disadvantageous SEP. This variable represented the accumulation model. The variable was coded as: 0=0 period– participants who were in advantageous SEP in all 3 periods of life; 1=1 period-participants who were exposed to disadvantageous SEP in any 1 period and non-exposed in any 2 periods of life; 2=2 periods - participants who were exposed to disadvantageous SEP in any 2 periods and non-exposed in any 1 period of life; and 3= 3 periods-participants who were exposed to disadvantageous SEP in all three periods of life.

SEP exposure measure for social mobility models

Two models were tested for mobility; childhood to early adulthood mobility, and early to late adulthood mobility.

Childhood to early adulthood mobility - The SEP measure representing this model was a 4-category variable. *Stable advantageous SEP (0, 0)*: Participants who maintained a stable advantageous SEP in both childhood and early adulthood irrespective of their SEP in late adulthood, were coded as 0. *Upward mobility (1, 0)*: Participants who were exposed to a

[6] Filmer D, Pritchett LH. 2001. Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of india. Demography. 38(1):115-132**.**

disadvantageous SEP in childhood but went on to attain an advantageous SEP in early adulthood irrespective of their SEP in late adulthood were coded as 1. *Downward mobility (0, 1)*: Participants who had an advantageous SEP in childhood but disadvantageous SEP in early adulthood irrespective of their SEP in late adulthood were coded as 2. *Stable disadvantageous SEP (1, 1)*: Participants who maintained a stable disadvantageous SEP in both childhood and early adulthood irrespective of their SEP in late adulthood, were coded as 3;

Early to late adulthood mobility - A similar strategy was adopted to create the 4 category SEP variable representing social mobility between early and late adulthood by considering participants' SEP in these 2 periods of life.

eTable 1. List of housing assets/items, their categories, corresponding binary Stata codes and their proportion, if selected for creating the SEP measure for childhood, early or late adulthood.

Assets/items	**Categories (Stata code)**	**Proportion, if used in childhood (%)**	**Proportion, if used in early adulthood**	**Proportion, if used in late adulthood**
Crowding	Absent (0)	45.61	41.81	50.58
	Present (1)	54.39	58.19	49.42
Material of floor used	High cost (0)	85.23	65.50	17.69
	Low cost (1)	14.77	34.50	82.31
Material of roof used	High cost 0)	68.57	50.44	12.72
	Low cost (1)	31.43	49.56	87.28
Material of wall used	High cost (0)	77.92	65.06	23.54
	Low cost (1)	22.08	34.94	76.46
Windows	High cost (0)	34.06	19.15	06.14
	Low cost (1)	65.94	80.85	93.86
Water source	Protected (0)	46.35	34.50	09.06
	Unprotected (1)	53.65	65.50	90.94
Bathroom	Present (0)	79.39	51.46	06.29
	Absent (1)	20.61	48.54	93.71
Clock	Present (0)	84.06	58.19	10.67
	Absent (1)	15.94	41.81	89.33
Radio	Present (0)	91.08	73.39	31.58
	Absent (1)	08.92	26.61	68.42
Bicycle	Present (0)		90.64	
	Absent (1)		09.36	
Electricity	Present (0)		75.15	19.30
	Absent (1)		24.85	80.70
Television	Present (0)			42.11
	Absent (1)			57.89
Phone	Present (0)			32.31
	Absent (1)			67.69

0 represents advantageous SEP, and 1 represents disadvantageous SEP

eTable 2. Tetrachoric correlation matrix for items recorded in childhood

```
             | CH_crowd CH_floor  CH_roof  CH_wall  CH_wind CH_water  CH_bath CH_clock CH_radio
-------------+----------------------------------------------------------------------------------
    CH_crowd |   1.0000
    CH_floor |   0.4674   1.0000
     CH_roof |   0.5912   0.8038   1.0000
     CH_wall |   0.5362   0.7876   0.8618   1.0000
     CH_wind |   0.4474   0.6791   0.7352   0.6613   1.0000
    CH_water |   0.4203   0.5282   0.5891   0.5981   0.5493   1.0000
     CH_bath |   0.4827   0.7544   0.7556   0.6522   0.4896   0.5396   1.0000
    CH_clock |   0.5790   0.7576   0.7432   0.7788   0.4623   0.4433   0.7568   1.0000
    CH_radio |   0.5581   0.7296   0.7272   0.7147   0.5562   0.5295   0.7673   0.9068   1.0000
```

eTable 3. Tetrachoric correlation matrix for items recorded in early adulthood

```
             | EAH_crowd  EAH_floor EAH_roof  EAH_wall  EAH_wind  EAH_water  EAH_bath  EAH_elect EAH_clock EAH_radio EAH_cycle
-------------+----------------------------------------------------------------------------------------------------------------
   EAH_crowd |   1.0000
   EAH_floor |   0.5091   1.0000
    EAH_roof |   0.5045   0.8311   1.0000
    EAH_wall |   0.4791   0.8464   0.8302   1.0000
    EAH_wind |   0.3798   0.6649   0.6962   0.8196   1.0000
   EAH_water |   0.3618   0.6284   0.6184   0.6894   0.5987   1.0000
    EAH_bath |   0.4455   0.7554   0.7512   0.6827   0.5469   0.6093   1.0000
   EAH_elect |   0.4269   0.7462   0.7441   0.7474   0.5224   0.5759   0.8046   1.0000
   EAH_clock |   0.3843   0.6544   0.6289   0.6650   0.3359   0.4505   0.6566   0.8448   1.0000
   EAH_radio |   0.4532   0.7263   0.6860   0.6790   0.5206   0.5628   0.8108   0.8114   0.8704   1.0000
   EAH_cycle |   0.3946   0.5108   0.5326   0.5448   0.3879   0.6253   0.5962   0.5736   0.7142   0.8102   1.0000
```

eTable 4. Tetrachoric correlation matrix for items recorded in late adulthood

```
             | LAH_crowd LAH_floor  LAH_roof  LAH_wall  LAH_wind LAH_water  LAH_bath  LAH_clock LAH_radio  LAH_elect LAH_tv LAH_phone
-------------+-------------------------------------------------------------------------------------------------------------------------
   LAH_crowd |   1.0000
   LAH_floor |   0.3263   1.0000
    LAH_roof |   0.3178   0.8622   1.0000
    LAH_wall |   0.3237   0.8811   0.8672   1.0000
    LAH_wind |   0.2523   0.5743   0.6523   0.5789   1.0000
   LAH_water |   0.2568   0.4108   0.4918   0.4424   0.4123   1.0000
    LAH_bath |   0.2943   0.7639   0.7493   0.7337   0.5949   0.5375   1.0000
   LAH_clock |   0.1781   0.6312   0.5693   0.6312   0.3599   0.3192   0.7153   1.0000
   LAH_radio |   0.3373   0.4644   0.5405   0.4725   0.3582   0.3729   0.5895   0.7428   1.0000
   LAH_elect |   0.3161   0.7030   0.7312   0.6030   0.5455   0.4411   0.8708   0.6296   0.5432   1.0000
      LAH_tv |   0.3371   0.7417   0.6730   0.6848   0.5621   0.4188   1.0000   0.7670   0.5826   0.8759   1.0000
   LAH_phone |   0.2706   0.5839   0.6039   0.5992   0.5165   0.4240   1.0000   0.7521   0.5584   0.7088   0.8120   1.0000
```

eTable 5. Relative measures of test-retest reliability for housing based assets used to create SEP measures for childhood, early and late adulthood periods.

		Childhood			**Early adulthood**			**Late adulthood**		
Assets/items	**N**	**Pearson correlation**	**Intra class correlation**	**95% confidence interval**	**Pearson correlation**	**Intra class correlation**	**95% confidence interval**	**Pearson correlation**	**Intra class correlation**	**95% confidence interval**
Crowding	46[a]	0.91	0.95	0.92, 0.98	0.86	0.93	0.87, 0.96	0.74	0.85	0.73, 0.92
Material of floor	46 [a]	0.91	0.95	0.92, 0.98	0.99	0.99	0.99, 0.99	0.63	0.79	0.58, 0.87
Material of roof	46 [a]	0.99	0.99	0.99, 0.99	0.99	0.99	0.99, 0.99	0.99	0.99	0.99, 0.99
Material of wall	46 [a]	0.99	0.99	0.99, 0.99	0.95	0.98	0.96, 0.99	0.92	0.96	0.93, 0.98
Windows	46 [a]	0.86	0.93	0.86, 0.56	0.99	0.99	0.99, 0.99	0.99	0.99	0.99, 0.99
Water source	46 [a]	0.92	0.96	0.92, 0.98	0.90	0.95	0.90, 0.97	0.85	0.98	0.94, 0.99
Bathroom	46 [a]	0.99	0.99	0.98, 0.99	0.87	0.93	0.87, 0.96	0.70	0.80	0.63, 0.89
Clock	46 [a]	0.83	0.76	0.64, 0.95	0.82	0.79	0.55, 0.94	0.84	0.77	0.59, 0.96
Radio	46 [a]	0.75	0.85	0.72, 0.91	0.79	0.87	0.76, 0.93	0.69	0.79	0.62, 0.88
Bicycle	46 [a]				0.73	0.82	0.67, 0.90	0.64	0.80	0.62, 0.88
Electricity	46 [a]				0.92	0.97	0.95, 0.98	0.91	0.95	0.92, 0.96
Television	46 [a]							0.85	0.92	0.85, 0.54
Phone	46 [a]							0.87	0.93	0.87, 0.96

[a] Among the sample of 721 participants recruited in total at the Indian site, re-interviews were conducted for 46 randomly selected participants, 6 to 12 weeks after the original interview. The above measures were estimated among these participants.

eTable 6. Conceptual life-course models, corresponding trajectories, causal contrasts and regression models

Conceptual Model	Levels of exposure (0=No, 1=Yes)	Contrast for each trajectory	Marginal structural regression models
All-trajectories saturated model			
Never exposed	0, 0, 0		
Exposed in CH (A_{100}) vs never exposed	1, 0, 0	$E[Y_{100}- Y_{000}]$	$\text{logit}\ \{\Pr[Y_g(SEP)]\} = \alpha + \beta_1\, g(SEP)$ g(SEP)=function of 8 category variable involving all 8 life-course trajectories. Betas correspond to estimates for each contrast in column 3.
Exposed in EAH (A_{010}) vs never exposed	0, 1, 0	$E[Y_{010}- Y_{000}]$	
Exposed in LAH (A_{001}) vs never exposed	0, 0, 1	$E[Y_{001}- Y_{000}]$	
Exposed in CH & EAH (A_{110}) vs never	1, 1, 0	$E[Y_{110}- Y_{000}]$	
Exposed in CH & LAH (A_{101}) vs never	1, 0, 1	$E[Y_{101}- Y_{000}]$	
Exposed in EAH & LAH (A_{011}) vs never	0, 1, 1	$E[Y_{011}- Y_{000}]$	
Exposed in CH, EAH&LAH (A_{111}) vs never	1, 1, 1	$E[Y_{111}- Y_{000}]$	
Accumulation model			
Never exposed	0 periods		
Exposed at 1 time point vs never exposed	1 period	$E(Y_{100,010,001}-Y_{000})$	$\text{logit}\ \{\Pr[Y_g(SEP)]\} = \alpha + \beta_1\, g(SEP)$ g(SEP)= function of 4 category variable involving specific combination of
Exposed at 2 time points vs never	2 periods	$E(Y_{110,101,011}-Y_{000})$	
Exposed at 3 time points vs never	3 periods	$E(Y_{111}-Y_{000})$	
Critical period			
Exposed in CH vs unexposed in CH	1 vs 0	$E(Y_{1**}-Y_{0**})^{a}$	$\text{logit}\ \{\Pr[Y(csep)=1]\} = \alpha + \beta_1 * g(csep)$
Exposed in EAH vs unexposed in EAH	1 vs 0	$E(Y_{*1*}-Y_{*0*})^{a}$	$\text{logit}\ \{\Pr[Y(esep)=1]\} = \alpha + \beta_1 * g(esep)$
Exposed in LAH vs unexposed in LAH	1 vs 0	$E(Y_{**1}-Y_{**0})^{a}$	$\text{logit}\ \{\Pr[Y(lsep)=1]\} = \alpha + \beta_1 * g(lsep)$
Social mobility model			
Childhood to early adulthood			$\text{logit}\ \{\Pr[Y_g(SEP)]\} =\alpha +\beta 1\ g(SEP)$ g(SEP)= function of 4 category variable involving specific combination of trajectories over CH and EAH SEP
Stable advantageous	0,0		
Upward mobility	1, 0	$E[Y_{10*}-Y_{00*}]^{a}$	
Downward mobility	0, 1	$E[Y_{01*}-Y_{00*}]^{a}$	
Stable disadvantageous	1, 1	$E[Y_{11*}-Y_{00*}]^{a}$	
Early to Late adulthood			
Stable advantageous	0,0		$\text{logit}\ \{\Pr[Y_g(SEP)]\} =\alpha +\beta 1\ g(SEP)$ g(SEP)= function of 4 category variable involving specific combination of trajectories over EAH and LAH SEP
Upward mobility	1, 0	$E[Y_{*10}-Y_{*00}]^{a}$	
Downward mobility	0, 1	$E[Y_{*01}-Y_{*00}]^{a}$	
Stable disadvantageous	1,1	$E[Y_{*11}-Y_{*00}]^{a}$	

Abbreviations: CH=childhood; EAH= early adulthood; LAH= late adulthood; SEP= Socioeconomic position; csep= childhood SEP; esep= early adulthood SEP; lsep= late adulthood SEP
If A is the exposure level, A=1 would represent exposed to disadvantageous SEP, and A=0 would be non-exposure.

[a] *can take any value between 0 and 1

eTable 7. Summary statistics for minimally stabilized inverse probability weights based on which the final weights for the outcome marginal structural models were created.

Stabilized inverse probability weights	N	Mean	SD	Min	Max
W1	684	1.02	0.41	0.42	4.10
W2	684	1.01	0.84	0.26	5.66
W3	684	1.13	1.29	0.33	7.19
W12	684	1.03	0.90	0.16	5.97
W123	684	1.16	1.64	0.06	13.47

W1 – Stabilized inverse probability weight for childhood; W2 – Stabilized inverse probability weight for early adulthood; W3 –Stabilized inverse probability weight for late adulthood; W12- Product of W1 and W2; W123- Product of W1, W2 and W3. Please see Table 8 for details.

eTable 8. Annotated Stata code for exposure weights, description of SEP exposure weight specification, and outcome marginal structural models

```
// For creating inverse probability weight for childhood (W1)
logit csep [pw=Sampfrac] // [for PP of numerator]
logit csep age* sex caste [pw=Sampfrac] // (for PP of denominator)

// For creating inverse probability weight for early adulthood (W2)
logit esep [pw=Sampfrac] // (for numerator PP)
logit esep csep age* sex caste edu C1Cig* C1Bidi* C1Chew* C1Drink* C2aCig* C2aBidi* C2aChew* ///
C2aDrink* [pw=Sampfrac] // (for PP denominator)

// For creating inverse probability weight for late adulthood (W3)
logit lsep [pw=Sampfrac] // (for numerator PP)
logit lsep esep csep age* sex caste edu C1Cig* C1Bidi* C1Chew* C1Drink* C2aCig* C2aBidi* ///
C2aChew* C2aDrink* C2bCig* C2bBidi* C2bChew* C2bDrink* C3aCig* C3aBidi* C3aChew* C3aDrink* [pw=Sampfrac]
// (for PP denominator)

/*csep=childhood SEP, esep=early adulthood SEP, lsep= late adulthood SEP, edu = education, Cig=cigarette
smoking, Bidi=Bidi smoking, Chew=paan chewing, Drink= alcohol consumption, *= entered as spline variable,
PP=predicted probability; Sampfrac =sampling fraction, (VanderWeele & Vansteelandt 2010). Please refer to
Figure 1 to understand the confounder variables from their respective prefix (e.g., C1, C2a, C2b, C3a) */

// Weight multiplication [SampW=time dependent sampling weight (Leffondre et al, 2010]
sW1=W1*SampW
sW12= W12*SampW // where W12=W1*W2
sW123= W123*SampW // where W123=W1*W2*W3

// Outcome marginal structural models (unadjusted logistic regression models on the pseudo-population)
logistic Status i.Traj_SEP [pw=sW123 ] //  (saturated all-trajectories model)
logistic Status i.AccSEP [pw=sW123] //(accumulation model)
logistic Status csep [pw=sW1] // (childhood critical period)
logistic Status esep [pw=sW12] // (early adulthood critical period)
logistic Status lsep [pw=sW123] // (late adulthood critical period)
logistic Status i.ce_mob [pw=sW12] // ce_mob= childhood to early adulthood mobility SEP variable
logistic Status i.ea_mob [pw=sW123] // ea_mob= early to late adulthood mobility SEP variable
```

References

Akg center for research and studies, communist party of india (marxist), state committee, kerala. Education bill. 2009. Kerala,India: AKG Center for Research and Studies; [accessed 2015 17 Nov]. http://www.cpimkerala.org/eng/education-23.php?n=1.

Barker DJ. 1990. The fetal and infant origins of adult disease. BMJ : British Medical Journal. 301(6761):1111-1111.

Ben-Shlomo Y, Kuh D. 2002. A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. International Journal of Epidemiology. 31(2):285-293.

Bernabe E, Suominen AL, Nordblad A, Vehkalahti MM, Hausen H, Knuuttila M, Kivimaki M, Watt RG, Sheiham A, Tsakos G. 2011. Education level and oral health in finnish adults: Evidence from different lifecourse models. J Clin Periodontol. 38(1):25-32.

Berney L, Blane DB, Soc Sci M, Blane B. 1997. Collecting retrospective data: Accuracy of recall after 50 years judged against historical records. Social Science & Medicine. 45(10):1519-1525.

Blane D. 1995. Social determinants of health--socioeconomic status, social class, and ethnicity. American Journal of Public Health. 85(7):903-905.

Blane D, Harding S, Rosato M. 1999. Does social mobility affect the size of the socioeconomic mortality differential?: Evidence from the office for national statistics longitudinal study. Journal of the Royal Statistical Society Series A, (Statistics in Society). 162(Pt. 1):59-70.

Blane D, Netuveli G, Stone J. 2007. The development of life course epidemiology. Revue d'epidemiologie et de sante publique. 55(1):31-38.

Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, Hertzman C, Power C, Szyf M. 2012. Associations with early-life socio-economic position in adult DNA methylation. Int J Epidemiol. 41(1):62-74.

Conway DI, Petticrew M, Marlborough H, Berthiller J, Hashibe M, Macpherson LM. 2008. Socioeconomic inequalities and oral cancer risk: A systematic review and meta-analysis of case-control studies. International journal of cancer. 122(12):2811-2819.

Fagundes CP, Way B. 2014. Early-life stress and adult inflammation. Current Directions in Psychological Science. 23(4):277-283.

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. 2013. Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. Lyon, france: International agency for research on cancer; 2013. Available from: Http://globocan.Iarc.Fr, accessed on january 05, 2015.

Filmer D, Pritchett LH. 2001. Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of india. Demography. 38(1):115-132.

Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. 2006. Indicators of socioeconomic position (part 1). Journal of Epidemiology and Community Health. 60(1):7-12.

Global status report on alcohol and health, world health organisation. Who:Management of substance abuse, 2014. Available form: Http://www.Who.Int/substance_abuse/publications/global_alcohol_report/en/. 2014.

Gwatkin D, Rutstein S, Johnson K, Pande R, Wagstaff A. 2000. Socioeconomic differences in health, nutrition and population. Health, nutrition and population discussion paper, washington (d.C.): The world bank.

Hallqvist J, Lynch J, Bartley M, Lang T, Blane D. 2004. Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the stockholm heart epidemiology program. Soc Sci Med. 58(8):1555-1562.

Howe LD, Hargreaves JR, Gabrysch S, Huttly SRA. 2009. Is the wealth index a proxy for consumption expenditure? A systematic review. Journal of Epidemiology and Community Health. 63(11):871-877.

Kerala's 2005 human development report, united nations development program (undp). Available from: Http://www.In.Undp.Org/content/india/en/home/library/hdr/human-development-reports/state_human_development_reports/kerala.Html, accessed on june 2nd, 2016. 2005.

Krishna Rao S, Mejia GC, Roberts-Thomson K, Logan RM, Kamath V, Kulkarni M, Mittinty MN. 2015. Estimating the effect of childhood socioeconomic disadvantage on oral cancer in india using marginal structural models. Epidemiology (Cambridge, Mass). 26(4):509-517.

Langholz B. 2007. Use of cohort information in the design and analysis of case-control studies. Scandinavian Journal of Statistics. 34(1):120-136.

Laprise C, Shahul HP, Madathil SA, Thekkepurakkal AS, Castonguay G, Varghese I, Shiraz S, Allison P, Schlecht NF, Rousseau MC et al. 2016. Periodontal diseases and risk of oral cancer in southern india: Results from the hence life study. International journal of cancer. 139(7):1512-1519.

Laprise C, Madathil SA, Allison P, Abraham P, Raghavendran A, Shahul HP, et al. No role for human papillomavirus infection in oral cancers in a region in southern India. International journal of cancer. 2016;138(4):912-7.

Leffondre K, Wynant W, Cao Z, Abrahamowicz M, Heinze G, Siemiatycki J. 2010. A weighted cox model for modelling time-dependent exposures in the analysis of case-control studies. Stat Med. 29(7-8):839-850.

Madathil SA, Rousseau MC, Wynant W, Schlecht NF, Netuveli G, Franco EL, Nicolau B. 2016. Nonlinear association between betel quid chewing and oral cancer: Implications for prevention. Oral oncology. 60:25-31.

McEwen BS. 1998. Stress, adaptation, and disease. Allostasis and allostatic load. Annals of the New York Academy of Sciences. 840:33-44.

Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R. 2009. A structured approach to modelling the effects of binary exposure variables over the life course. International Journal of Epidemiology. 38(2):528-537.

Mishra GD, Chiesa F, Goodman A, De Stavola B, Koupil I. 2013. Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50-87 years: Results from a swedish birth cohort. European journal of epidemiology. 28(2):139-147.

Naimi A, Kaufman J. 2015. Counterfactual theory in social epidemiology: Reconciling analysis and action for the social determinants of health. Current Epidemiology Reports. 2(1):52-60.

National sample survey office: Household consumption of various good and services in india 2011-2012, nss 68th round, ministry of statistics and programme implementation, government of india, june 2014. Available from: Http://mospi.Nic.In/mospi_new/upload/report_no558_rou68_30june14.Pdf. 2014. In: Implementation MoSaP, editor. India: Government of India.

Nicolau B, Netuveli G, Kim JW, Sheiham A, Marcenes W. 2007. A life-course approach to assess psychosocial factors and periodontal disease. J Clin Periodontol. 34(10):844-850.

Peres MA, Peres KG, Thomson WM, Broadbent JM, Gigante DP, Horta BL. 2011. The influence of family income trajectories from birth to adulthood on adult oral health: Findings from the 1982 pelotas birth cohort. Am J Public Health. 101(4):730-736.

Petti S. 2009. Lifestyle risk factors for oral cancer. Oral oncology. 45(4-5):340-350.

Platt RW, Brookhart MA, Cole SR, Westreich D, Schisterman EF. 2013. An information criterion for marginal structural models. Statistics in Medicine. 32(8):1383-1393.

Pollitt R, Rose K, Kaufman J. 2005. Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: A systematic review. BMC Public Health. 5(1):7.

Robins JM, Hernán Ma, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass). 11(5):550-560.

Stringhini S, Polidoro S, Sacerdote C, Kelly RS, van Veldhoven K, Agnoli C, Grioni S, Tumino R, Giurdanella MC, Panico S et al. 2015. Life-course socioeconomic status and DNA methylation of genes regulating inflammation. International Journal of Epidemiology. 44(4):1320-1330.

Stringhini S, Sabia S, Shipley M, et al. 2010. Association of socioeconomic position with health behaviors and mortality. Jama. 303(12):1159-1166.

VanderWeele TJ, Jackson JW, Li S. 2016. Causal inference and longitudinal data: A case study of religion and mental health. Social Psychiatry and Psychiatric Epidemiology.1-10.

Vohra J, Marmot MG, Bauld L, Hiatt RA. 2016. Socioeconomic position in childhood and cancer in adulthood: A rapid-review. J Epidemiol Community Health. 70(6):629-634.

Warnakulasuriya S. 2009. Global epidemiology of oral and oropharyngeal cancer. Oral oncology. 45(4-5):309-316.

Chapter 6

Manuscript II

Genetic variants in CYP and GST genes, smoking and risk for head and neck cancers: a gene-environment interaction study

Akhil Soman ThekkePurakkal[1], Belinda Nicolau[1], Robert D Burk[2], Eduardo L Franco[3], Nicolas F Schlecht[4]

[1]Division of Oral Health and Society, Faculty of Dentistry, McGill University, Montreal, Canada; [2]Departments of Pediatrics (Genetics), Microbiology & Immunology, Obstetrics Gynecology & Women's Health, and Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA; [3]Departments of Oncology and Epidemiology & Biostatistics, McGill University, Montreal, Quebec, Canada; [4]Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA

Keywords: GSTP1 105Val, Canadian Caucasian, copy number variation, joint effect' stratum specific effect, additive interaction, multiplicative interaction, case-control

List of abbreviations and acronyms: SCCHN: Squamous cell carcinomas of the head and neck, PAH: Polycyclic aromatic hydrocarbons, CYP: Cytochrome P450 enzyme, GST: Glutathione S-transferase, SNP: Single nucleotide polymorphism; CNV: copy number variation; HPV: Human papillomavirus, DNA: Deoxyribonuclic acid; PCR: Polymerase chain reaction; OR: Odds ratio; CI: Confidence Interval; RERI: Relative excess risk due to interaction.

Abstract

Background: Genetic variants in Cytochrome P450 (CYP) and Glutathione S-transferase (GST) genes involved in the metabolism of environmental carcinogens have been widely studied as potential risk factors for squamous cell carcinomas of the head and neck (SCCHN). However, evidence for the effect of many of these variants is conflicting. Furthermore, their effects on SCCHN in interaction with multiple levels of smoking have not been documented among Canadian Caucasian population.

Objective: We aimed to estimate the total effect of CYP1A1*2A, CYP1A1*2C, CYP2A6*2, CYP2E1c2, GSTP1 105Val, copy number variations in CYP2D6 null and GSTM1 on SCCHN risk in a sample of Caucasians from Montreal, Canada. In addition, we conducted an analysis of causal interaction between these variants and multiple levels of smoking on SCCHN risk.

Methods: Analysis was conducted on 389 incident SCCHN cases and 429 controls, frequency-matched by age and sex, recruited from four main hospitals in Montreal, Canada between 2005 and 2013. Life-course based interviews collected information on several domains of exposures. DNA was isolated from oral exfoliated cells and genotyped for multiple genetic variants. A dominant model of inheritance for CYP1A1*2A, CYP1A1*2C, CYP2A6*2, CYP2E1c2 and GSTP1 105Val, and binary categories of copy number variants of CYP2D6 null (1 to 2 vs 3 to 9 copies) as well as GSTM1 (0 vs 1 to 3 copies) were analysed. Cigarette equivalence of tobacco smoking was categorized into no, moderate (>0 to 32 pack-years) and heavy (> 32 pack-years) smokers. Unconditional logistic regression models estimated odds ratios (OR) and 95% confidence intervals (CI) for main, joint effect, stratum specific and interaction estimates.

Results: Of all variants analysed, carriers of GSTP1 105Val (vs non-carriers) were at 29% (OR=0.71, 95% CI: 0.53, 0.95) decreased risk of SCCHN. Stratum specific analyses showed that carriers of this variant were at 41% (OR=0.59, 95% CI: 0.36, 0.95) and 51% (OR= 0.49, 95% CI: 0.24, 0.98) decreased risk for SCCHN relative to non-carriers, among the strata of heavy smokers and non-smokers respectively. There was no evidence for statistical interaction on additive or multiplicative scale for any of the variants analysed.

Conclusion: GSTP1 105Val decreased the risk for SCCHN independent of smoking, as well as among heavy smokers in this Caucasian population.

Introduction

Squamous cell carcinomas of head and neck (SCCHN) are chronic multi-factorial diseases with both environmental and genetic risk factors. Single nucleotide polymorphisms (SNPs) in key genes encoding enzymes involved in either the bio-activation (e.g., Cytochrome or CYP450 enzymes) or detoxification (e.g., Glutathione S-transferase or GST enzymes) of various environmental carcinogens such as polycyclic aromatic hydrocarbons (PAH), volatile nitrosamines, aromatic amines and tobacco specific nitrosamines have been widely investigated as potential SCCHN risk factors (1-6). The frequencies of these variants and consequently their effects vary across ethnicities. Several meta-analytical reviews support a positive association between some SNPs (e.g., CYP1A1*2A, CYP1A1*2C, CYP2E1c2) and SCCHN risk among Asian populations. However, this association is not observed among Caucasians (7-10). While the evidence for the role of other SNPs (e.g., GSTP1 105Val) in this association is conflicting (6, 11-13), the effect of certain variants in CYP2A6 and CYP2D6 (e.g., CYP2A6*2 and copy number variation in CYP2D6) on SCCHN risk has not been documented. No studies have yet looked into the association between these CYP and GST genetic variants and SCCHN risk among Caucasian population in Canada.

The carcinogenic substrates of CYP and GST genes are abundant in tobacco smoke, the major risk factor for SCCHN in the West (14-16). Given the strong biological plausibility of involvement of genetic variants in these genes in altering metabolism of tobacco derived carcinogens, estimation of main effects of these variants alone may mask the potential differential risk due to their interaction with various levels of smoking. Hence, the objective of this study was two-fold. First, we aimed to estimate the total effect of variants in CYP1A1, CYP2E1, CYP2A6, CYP2D6, GSTM1 and GSTP1 genes on SCCHN risk in a sample of Caucasians from Montreal, Canada. Second, we conducted a comprehensive analysis of causal interaction between these variants and multiple levels of smoking on SCCHN risk.

Methods

Population, Sample and data collection

Data for this study were drawn from the Canadian site of an IRB approved, international multi-center hospital-based case-control study: The Head and Neck Cancer (HeNCe) Life study. A total of 918 participants were recruited between 2005 and 2013 from four major referral hospitals in Montreal. The eligibility criteria were: (i) born in Canada; (ii) aged ≥ 18 years; (iii) English or French speaking; and (iv) living within a 50km radius from the recruiting hospitals. Cases (N=460) included consecutive, incident, histologically confirmed, stage I to IV squamous cell carcinomas of the mouth, oropharynx and larynx (C01-C06, C09, C10, C12- C14, and C32, under the International Statistical Classification of Diseases, 10th Revision). Cancer-free controls (N=458), frequency-matched by 5-year age group and sex to cases, were randomly selected from several outpatient clinics (that are not typically associated with smoking and alcohol) at the same hospitals as the cases. No single diagnostic group contributed to more than 20% of the total. The recruitment of controls followed an incident density sampling technique.

Trained interviewers conducted one-on-one semi-structured interviews using a questionnaire and life-grid technique to improve recall. Data collected included information on several domains of exposures such as sociodemographic factors, life time tobacco smoking and alcohol consumption.

Sample collection and analysis

To identify genetic polymorphisms, genotyping was performed on DNA from oral epithelial samples collected from normal buccal mucosa by brush following standardized protocols (17, 18). Briefly, reactions were set up using 5 µl of 2X Genotyping Master Mix (Applied Biosystems, Foster City, CA) combined with assay-specific concentrations of primers and probes, and 10 ng of sample DNA. The reactions were then spun down at 2000 rpm for 2 min, and run in the 7500FAST real-time PCR thermocycler in Genotyping mode under default settings. 7500FAST v2.0.1 (updated to v2.0.6) software was used for allelic discrimination. Five SNPs (CYP1A1*2A, CYP1A1*2C, CYP2A6*2, CYP2E1c2, GSTP1 105Val) and two structural variations in genes known as copy number variations (CNV) related to two SNPs (CYP2D6null and GSTM1) were selected for this analysis (19-21).

Testing for presence of human papillomavirus (HPV) in the oral cavity as well as typing for HPV genotypes were done as described elsewhere (22).

Data Analysis

Genetic exposures

All variants were dichotomized for analysis to ensure adequate numbers in each category. CYP1A1*2A (variant allele=C), CYP1A1*2C (variant allele=G) and CYP2E1c2 (variant allele=C) SNPs encode respective enzymes with faster activity (8-10). A dominant model of inheritance for these SNPs was tested. Hence, the non-carrier and carrier genotype groups were TT and CT/CC for CYP1A1*2A, AA and AG/GG for CYP1A1*2C, and GG and GC/CC for CYP2E1c2, respectively. While the CYP2A6*2 (variant allele=A) results in inactivity of the encoding enzyme, the GSTP1 105Val (minor allele=G) is associated with slower enzyme activity (12). The genetic status for these SNPs was for CYP2A6*2: TT and AT/AA, and for GSTP1 105Val: AA and AG/GG for non-carriers and carriers, respectively. Lower copy numbers of CYP2D6null are associated with a faster enzyme relative to higher copies (20). One to nine copies were identified in our sample. So, participants with higher CNV (3 to 9 copies) were grouped together and compared to those with 1 to 2 copies. A deletion of GSTM1 gene resulting in null genotype renders the encoded enzyme inactive (21). We identified 0 to 3 copies of GSTM1. Participants with 0 copies were coded null, where as those with 1-3 copies were grouped as non-null.

Tobacco smoking exposure

Smoking history was based on lifetime exposure. For each period with consistent smoking habits, the following information was collected for cigarette, cigar and pipe smoking: (a) age at start; (b) age at end; (c) type (filtered, unfiltered, hand-rolled); (d) brand, grams or pipes, respectively; (d) number of units per day. The duration of each smoking period was calculated as the difference between age at beginning and end of the period in years. A continuous variable representing lifetime exposure to smoking was created using the pack-year unit, one pack-year being equivalent to smoking one pack of cigarettes daily during one year. The following equivalence was assumed for the computation of pack-year smoking exposure: 1 manufactured cigarette pack = 20 manufactured cigarettes = 4 hand-rolled cigarettes = 4 cigars = 4 pipes (23). We also used information on time since smoking cessation (age at recruitment minus age at cessation) to identify participants who stopped smoking ≤ 2 years prior to recruitment. This strategy was used because these participants had a higher risk for the outcome than current smokers (time since cessation=0) (Supplemental material, eTable1). Hence, to mitigate reverse causality bias, we used a cut-off of 2 years prior to interview to define ex-smokers, and excluded details of any exposure during this period for the above calculations (24). Based on the pack-years variable, we created a three-category tobacco smoking variable: non-, moderate and heavy smoker. To categorize the pack-year variable, we first determined the shape of the dose-response curve between pack-years and SCCHN risk. A non-linear association with the risk increasing up to 70 pack-years beyond which the curve plateaued (Supplemental material, eFigure 1) indicated that the optimal cut-off would be between 0 and 70 pack-years. Since smoking was the secondary exposure, next, we used a parametric outcome based approach (25) to identify an optimal cut-off point of 32 pack-years among smokers. The final smoking variable was categorized as non- (0 pack-years), moderate (>0 to ≤ 32 pack-years) and heavy (> 32 pack-years) smokers.

Potential confounders

Causal graphs where used to identify the minimal set of potential confounders for estimating the total effect of each genetic variant on SCCHN[7] (26). All variants were adjusted for age, sex and education (measure of socioeconomic position). Additionally, models for GSTP1105Val, CYP2D6null and GSTM1 were adjusted for smoking, and those for CYP2E1c2, CYP1A1*2A, CYP1A1*2C for both smoking and ethanol. CYP1A1*2A and *2C were mutually adjusted for each other in their respective models. Ethanol frequency variable was calculated from comprehensive information on multiple alcohol beverages as described in the supplemental material, eAppendix. A continuous frequency variable representing average amount of ethanol consumed per day over life time was created and was adjusted using restricted cubic splines. For assessing joint effects, stratum specific effects and causal interactions, all potential confounders for the relation between specific variants and SCCHN, and smoking and SCCHN were considered

[7] *Please refer to Appendix V, page 312* for the causal graphs, and for the minimal set of potential confounders identified through these graphs, specific to each genetic variant a nd SCCHN association.

in the models. This included age, sex, alcohol, education and HPV. HPV risk types were categorised as described elsewhere (22). All analysis was limited to Caucasians to mitigate bias through population stratification.

Statistical methods

Deviations from the Hardy-Weinberg equilibrium were assessed among the controls using Chi-square tests. Allele frequencies were estimated among controls. Unconditional logistic regression models were fit to estimate the odds ratios (OR) and 95% confidence intervals (CI) for the association between SNPs and SCCHN. For interaction analysis, we used unconditional logistic regression to calculate the joint effect estimates using a single reference group, and effect of genotypes within the strata of smoking levels, interaction estimates on both multiplicative and additive scales and their respective CIs. The estimates of multiplicative interaction (ratio of odds ratios) were derived by including a product term for the two exposures in the models. The estimate of multiplicative interaction >1 was considered positive and <1 as negative (27). Relative excess risk due to interaction ($RERI_{OR}$) was calculated as a measure of additive interaction (28). $RERI_{OR} > 0$ indicated positive additive interaction and $RERI_{OR} < 0$ indicated negative additive interaction. Because we included ex-smokers, time since cessation in years (mean centered, current and non-smokers recoded to zero) was added to the list of confounders (24), and an indicator for ex-smokers was also included in the models. All analyses were carried out using Stata, version 13SE (StataCorp. 2013, College Station, TX).

Results

Of the 918 participants enrolled, 818 (429 controls, 389 cases) reported to be Caucasians who were genotyped as well as had complete data on smoking history. Table 1 provides the sample characteristics and total effect of non-genetic variables on SCCHN risk. The mean age of the participants was 61 years for both cases and controls, and almost three quarters were male. On average, cases had fewer years of formal education compared to controls (12 years vs 14 years), and greater number of years of education was protective for SCCHN. Approximately 86% of controls were HPV negative compared to 58% of cases. In addition, a larger proportion of cases compared to controls (27% vs 2 %) were HPV16 positive. Individuals who were HPV16 positive were at considerably higher risk for SCCHN (OR=20.26, 95% CI: 10.11, 40.59) relative to those who were HPV negative. On average, cases smoked more compared to controls (mean pack-years = 41 vs 25). Based on our categorization of smoking, 27% and 43% of controls were non-smokers and moderate smokers respectively, whereas 54% of cases were heavy smokers. Relative to non-smokers, heavy smokers (OR=2.76, 95% CI: 1.36, 5.58), but not moderate smokers were at increased risk for SCCHN.

Table 1: Sample characteristics of SCCHN cases and controls and multi-variate associations with head and neck cancer risk in a sample from Montreal, Canada, HeNCe Life study 2005-13, (n=818)

Non-genetic variables	**Controls (n=429)**		**Cases (n=389)**		
	N (%)	**mean (SD)**	**N (%)**	**mean (SD)**	**Adjusted OR (95% CI)**
Age, years		61 (11)		61 (10)	
Sex					
Female	131 (31)		98 (25)		
Male	298 (69)		291 (75)		
Education, years		14 (4)		12 (4)	0.93 (0.90, 0.97)
Tobacco smoking					
Pack-years		25 (39)		41 (46)	1.03 (1.01, 1.05)
Non-smoker	117(27)		68(17)		1
>0-32 pack years	183 (43)		112 (29)		1.37 (0.71, 2.63)
>32 pack-years	129 (30)		209 (54)		2.76 (1.36, 5.58)
Alcohol consumption					
Frequency (ml per day)	36 (103)		49 (98)		1.01 (1.00, 1.02)
HPV status					
HPV negative	369 (86)		226 (58)		1
HPV other	35 (8)		33 (9)		1.19 (0.70, 2.05)
HPV alpha-9 other than HPV16	15 (4)		25 (6)		2.16 (1.08, 4.33)
HPV-16	10 (2)		104 (27)		20.26 (10.11, 40.59)

OR: Odds ratio; CI: Confidence interval; SD: Standard deviation; HPV: Human papillomavirus
OR for education conditioned on age, sex, RC spline of tobacco pack-years and ethanol frequency, and HPV risk
OR for tobacco conditioned on time since cessation of smoking, indicator for ex-smoker, age, sex, education, RC spline of ethanol frequency, CYP2A6*2 and HPV risk
OR for ethanol conditioned on time since stoppage of use, indicator for ex-drinker, age, sex, education, RC spline of pack-years, ADH1B*2 and HPV risk
OR for HPV risk conditioned on age, sex, RC spline of tobacco pack-years and ethanol frequency, and education
HPV other: 6, 11, 18, 26, 34, 39, 40, 42, 44, 45, 51, 53, 54, 56, 59, 61, 62, 66, 68, 69, 70, 71, 72, 73, 81, 82, 83, 84, and 89
HPV alpha-9 other than HPV 16: 31,33,35,52,58 and 67

Of the 818 participants, genotypes could not be determined on 3, 10, 3, 47, 3, 2 and 3 participants for CYP1A1*2A, CYP1A1*2C, CYP2E1c2, CYP2A6*2, GSTP1 105Val, CYP2D6null and GSTM1null variants, respectively. Hence, we present genetic main effect and interaction results based on complete case analysis for each of these variants. Among controls, no deviation from Hardy-Weinberg equilibrium was observed for any of the SNPs tested. The total effect estimates for each variant on SCCHN risk are presented in Table 2. The minor allele frequencies of all SNPs among controls were similar to those reported in the literature among Caucasians. Of all the variants tested, carriers of GSTP1 105 Val were at ~29% decreased risk for SCCHN relative to non-carriers (OR= 0.71 95% CI: 0.53, 0.95).

Table 3 shows the joint effect, stratum specific and interaction estimates and associated confidence intervals. The joint effect estimates showed that relative to non-smokers who were non-carriers of CYP1A1*2A allele, risk for SCCHN was higher for both carriers and non-carriers who smoked heavily (carrier+ heavy smoker: OR=3.09, 95% CI: 1.27, 7.51; non-carrier + heavy smoker: OR=2.22, 95% CI: 1.05, 4.70). A similar pattern was identified for GSTM1 CNV. Relative to non-smokers who carried 1-3 copies (non-null) of GSTM1, the joint effect of heavy smoking and both null and non-null carriers (heavy smoking + null: OR=2.88, 95% CI= 1.23, 6.69), and heavy smoking + non-null (OR=2.72, 95% CI= 1.15, 6.42) conferred increased risk for SCCHN. For CYP2A6*2, relative to non-carriers and non-smokers, non-carriers who smoked heavily showed approximately 2.5-fold increased risk for SCCHN (OR=2.56, 95% CI: 1.21, 5.44). A similar pattern of increased risk was seen for the joint effect of heavy smoking and non-carriers of CYP1A1*2C (vs non-smoker + non-carrier of CYP1A1*2C group), heavy smoking and non-carriers of CYP2E1c2 (vs non-smoker + non-carrier of CYP2E1c2 group), and heavy smoking and non-carriers of GSTP1 105Val (vs non-smoker + non-carrier of GSTP1 105Val group).

Table 2: Genetic markers, allele frequencies, and their association with head and neck cancer risk, Montreal, Canada, 2005-13, (n=818)

Genetic variant	MAF in controls (%)	Genotype comparisons	Controls * N (%)	Cases * N (%)	Adjusted OR (95%CI)
CYP1A1*2A	C=11.9	TT	339 (79)	312 (81)	
		CT/CC	89 (21)	75 (19)	0.93 (0.61, 1.41)[a]
CYP1A1*2C	G=3.8	AA	393(92)	356 (93)	
		AG/GG	33(8)	26 (7)	0.90 (0.47, 1.70)[b]
CYP2E1(c2)	C=3.7	GG	390 (91)	366 (94)	
		GC/CC	37 (9)	22 (6)	0.67 (0.37, 1.19)[c]
CYP2A6*2	A=3.5	TT	385 (94)	341 (94)	
		AT/AA	24 (6)	21 (6)	0.94 (0.50, 1.77)[d]
GSTP1 (105 Val)	G=31.2	AA	194 (45)	203 (52)	
		AG/GG	234 (55)	184 (47)	**0.71 (0.53, 0.95)**[e]
CYP2D6null CNV		3 to 9	41(10)	33 (9)	
		1 to 2	387 (90)	355 (91)	1.21 (0.74, 2.00)[e]
GSTM1 CNV		1-3 (Non-Null)	193 (45)	178 (46)	
		0 (Null)	234 (55)	210 (54)	0.94 (0.70, 1.26)[e]

SNP: Single nucleotide polymorphism; MAF: Minor allele frequency; CNV: Copy number variation, OR: Odds ratio; CI: Confidence interval

*Numbers may not add up to 818 due to missing genotypes among variants studied

[a] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education, CYP1A1*2C

[b] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education, CYP1A1*2A

[c] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education

[d] OR conditioned on age, sex, education

[e] OR conditioned on age, sex, RC spline of smoking, education

Table 3: Joint effects for genetic variants and smoking on the risk of SCCHN, stratum specific effects and measures of interaction on multiplicative and additive scale in a sample from Montreal, Canada, 2005-13, (n=818)

Genetic variant	**N**	**Non-smoker**	**N**	**Moderate smoker**	**N**	**Heavy smoker**
	Co/Ca	**0 pack-years**	**Co/Ca**	**>0 to 32 pack-years**	**Co/Ca**	**> 32 pack-years**
		OR (95% CI)		**OR (95% CI)**		**OR (95% CI)**
CYP1A1*2A						
TT	92/54	1	141/93	1.31 (0.66, 2.61)	106/165	**2.22 (1.05, 4.70)**
CT/CC	25/13	0.55 (0.21, 1.43)	41/19	0.94 (0.37, 2.37)	23/43	**3.09 (1.27, 7.51)**
ORs (95% CI) for CT/CC within strata of smoking level		0.55 (0.21, 1.43)		0.72 (0.34, 1.49)		1.39 (0.72, 2.68)
Measure of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.30 (0.41, 4.09)		2.50 (0.82, 7.60)
Measure of interaction on additive scale: RERI (95% CI)				0.24 (-0.47, 0.95)		1.59 (-1.63, 4.82)
CYP1A1*2C						
AA	110/62	1	163/102	1.39 (0.71, 2.72)	120/192	**2.66 (1.29, 5.48)**
AG/GG	7/4	0.72 (0.14, 3.62)	17/8	1.32 (0.41, 4.18)	9/14	2.81 (0.83, 9.49)
ORs (95% CI) for AG/GG within strata of smoking levels		0.72 (0.14, 3.62)		0.95 (0.35, 2.57)		1.06 (0.38, 2.91)
Measure of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.31 (0.21, 8.20)		1.47 (0.23, 9.17)
Measure of interaction on additive scale: RERI (95% CI)				0.36 (-1.66, 2.35)		0.06 (-4.12, 4.23)
CYP2E1c2						
GG	105/64	1	164/107	1.40 (0.72, 2.71)	121/195	**2.52 (1.23, 5.16)**
GC/CC	12/3	0.40 (0.09, 1.74)	17/5	0.59 (0.17, 2.13)	8/14	2.80 (0.88, 8.84)
ORs (95% CI) for GC/CC within strata of smoking levels		0.40 (0.09, 1.74)		0.42 (0.13, 1.36)		1.11 (0.42, 2.92)
Measure of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.06 (0.16, 6.86)		2.76 (0.47, 16.06)
Measure of interaction on additive scale: RERI (95% CI)				0.04 (-0.97, 1.05)		1.03 (-3.19,5.25)

Table 3 continued …

Genetic variant	N Co/Ca	Non-smoker 0 pack-years OR (95% CI)	N Co/Ca	Moderate smoker >0 to 32 pack-years OR (95% CI)	N Co/Ca	Heavy smoker > 32 pack-years OR (95% CI)
CYP2A6*2						
TT	105/55	1	164/98	1.39 (0.69, 2.79)	116/188	**2.56 (1.21, 5.44)**
AT/AA	5/5	1.96 (0.49, 7.84)	11/7	0.93 (0.25, 3.50)	8/9	1.96 (0.57, 6.68)
ORs (95% CI) for AT/AA within strata of smoking levels		1.96 (0.49, 7.84)		0.67 (0.21, 2.18)		0.77 (0.27, 2.12)
Measure of interaction on multiplicative scale: Ratio of ORs (95% CI)				0.34 (0.05, 2.12)		0.39 (0.07, 2.18)
Measure of interaction on additive scale: RERI (95% CI)				-1.49 (-4.68, 1.69)		-1.84 (-6.22,2.53)
nCYP2D6 CNV						
3 to 9	11/9	1	13/6	0.96 (0.21, 4.47)	17/18	1.38 (0.34, 5.49)
1 to 2	106/58	0.86 (0.29, 2.53)	169/106	1.23 (0.38, 3.95)	112/191	2.56 (0.78, 8.36)
ORs (95% CI) for 1 to 2 CNV within strata of smoking levels		0.86 (0.29, 2.53)		1.28 (0.43, 3.75)		1.86 (0.82, 4.17)
Measure of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.49 (0.32, 6.80)		2.15 (0.56, 8.29)
Measure of interaction on additive scale: RERI (95% CI)				0.20 (-1.04, 1.45)		2.08 (-0.88, 5.05)
GSTP1 105 Val						
AA	51/39	1	95/58	0.98 (0.45, 2.12)	48/106	**2.75 (1.18, 6.44)**
AG/GG	66/28	**0.49 (0.24, 0.98)**	87/53	1.02 (0.47, 2.24)	81/103	1.62 (0.73, 3.62)
ORs (95% CI) for AG/GG within strata of smoking levels		**0.49 (0.24, 0.98)**		1.04 (0.61, 1.77)		**0.59 (0.36, 0.95)**
Measures of interaction on multiplicative scale: Ratio of ORs (95% CI)				2.14 (0.89, 5.16)		1.20 (0.51, 2.83)
Measures of interaction on additive scale: RERI (95% CI)				**0.51 (0.01, 1.01)**		-1.59 (-4.89,1.70)

Table 3 continued …

Genetic variant	**N Co/Ca**	**Non-smoker 0 pack-years OR (95% CI)**	**N Co/Ca**	**Moderate smoker >0 to 32 pack-years OR (95% CI)**	**N Co/Ca**	**Heavy smoker > 32 pack-years OR (95% CI)**
GSTM1 CNV						
1- 3 (Non-Null)	51/27	1	87/58	1.47 (0.66, 3.29)	55/93	**2.72 (1.15, 6.42)**
0 (Null)	65/40	1.09 (0.53, 2.21)	95/54	1.37 (0.60, 3.11)	74/116	**2.88 (1.23, 6.69)**
ORs (95% CI) for 0 (Null) CNV within strata of smoking levels		1.09 (0.53, 2.21)		0.93 (0.55, 1.57)		1.06 (0.65, 1.71)
Measures of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.17 (0.48, 2.81)		1.10 (0.47, 2.57)
Measures of interaction on additive scale: RERI (95% CI)				-0.13 (-1.08,0.82)		0.22 (-1.96,2.41)

Co: controls; Ca: cases; OR: Odds ratio; CI: Confidence interval; RERI: Relative excess risk due to interaction

All models conditioned on age, sex, education, time since cessation of smoking (mean centered, non-smokers coded 0), indicator for ex-smoker, RC spline of ethanol frequency, and HPV

Stratum specific analyses showed that carriers of GSTP1 105Val were at 41% (OR=0.59, 95% CI: 0.36, 0.95) and 51% (OR= 0.49, 95% CI: 0.24, 0.98) decreased risk for SCCHN relative to non-carriers, among the strata of heavy smokers and non-smokers respectively. In contrast, a positive interaction on the additive scale was seen for carriers of 105Val who smoked moderately (RERI=0.51, 95% CI: 0.01, 1.01).

Discussion

In this study, we aimed to estimate the total effect of genetic variants (SNPs and CNVs) in six genes involved in the metabolism of tobacco related carcinogens whose evidence for association with SCCHN is conflicting, or, has not been documented among Caucasian population. In addition, an analysis of gene-smoking interaction was also conducted. Our analysis showed a lower risk for SCCHN among carriers of GSTP1 105Val allele, overall and among heavy smokers. Overall, there was no evidence of statistical interaction on either multiplicative or additive scales between any of the variants tested at any level of smoking.

The GST enzymes are Phase II detoxifying enzymes involved in the detoxification of various electrophilic substrates including active metabolites of carcinogens such as PAH, monohalomethanes and ethylene oxide. GSTP1 enzyme and corresponding genes have been considered as important biomarkers for differential susceptibility to SCCHN as it is the most widely expressed GST enzyme in the head and neck region (29, 30). The GSTP1 105Val allele encodes an enzyme that is 2-3 times less stable than normally active GSTP1, and hence is considered less efficient in detoxifying its substrates (12, 31). However, the evidence for the association between the 105Val allele and SCCHN risk has been inconsistent. The three meta-analytical reviews conducted till date failed to document an association between carriers of 105Val allele and SCCHN risk (6, 11, 12`). In our study, the carriers of the 105Val allele showed a lower risk for SCCHN relative to non-carriers in the main effect analysis adjusted for smoking. This association persisted in our analysis stratified by HPV risk-types (28% decreased risk in the sample without HPV-16, 32% decreased risk in the sample without high risk HPV including HPV-16, and 35% decreased risk in the sample without HPV risk types except HPV-16) (Supplemental material, eTable 2). Indeed, a lower risk for multiple cancers including SCCHN among 105Val allele carriers relative to non-carriers, independent of tobacco smoking, has been documented (13, 32-36). The enzyme encoded by 105Val allele is highly substrate specific. The unstable enzyme, although less efficient in detoxifying substrates such as 1-chloro-2,4-dinitrobenzene relative to the stable enzyme, is highly efficient in detoxifying carcinogenic epoxides of PAH (e.g., benzo(a)pyrene) (37, 38). Hence, carriers of 105Val allele have been hypothesized to be less susceptible to PAH induced DNA damage and carcinogenesis. Apart from tobacco smoke, the upper aero digestive tract can be exposed to PAH from other sources such as diet, vehicle exhaust and wood combustion. Our results also showed a decreased risk for SCCHN among 105Val carriers (vs non-carriers) in the strata of non-smokers in the full sample and in the sample excluding all HPV risk types except HPV-16 (Supplemental material, eTable 3). However, a decreased risk for carriers of GSTP1 105Val among heavy smokers was consistently identified in the overall sample, as well as in all samples stratified by HPV risk types indicating that this finding is independent of HPV risk status. Indeed, the GSTP1 enzyme among 105Val carriers has been documented to show up to 3-fold increase in detoxification activity relative to non-carriers, in the presence of bulky diol epoxides of benzo(a)pyrene or structurally related PAH, which are readily available during heavy smoking (39). Analysis stratified by HPV also indicated that the models used in this study adjusted for HPV is a valid statistical model for estimating the genetic effects and gene-environment interaction effects.

Multiple meta-analytical reviews have failed to document strong associations between genetic variants such as CYP1A1*2A, *2C and CYP2E1c2 among Caucasians (5, 7, 9-11, 40-44). However, an increased for SCCHN among carriers of GSTM1 null relative to non-null carriers, in various ethnicities including Caucasians is well documented in the literature (4, 11, 45-48). Our study did not identify any association between these CYP and GST genetic variants and SCCHN risk in the main effect as well as stratum specific analyses.

To our knowledge, this is the first study to investigate the association between CNVs of CYP2D6 null, CYP2A6*2 SNP, and SCCHN risk (19, 20). Null variants of CYP2D6 have been associated with decreased risk for SCCHN (20). Hence, individuals with lower copy number of the null variant may be at increased risk for SCCHN relative to those who carry higher copy numbers of the variant. However, we did not find any evidence to support this hypothesis in our study. Previous studies have investigated the association between CYP2A6 variants with similar functional consequences as CYP2A6*2 (e.g. CYP2A6*4 SNP) and tobacco related cancers. These studies document a reduced risk for these cancers among Caucasians who are carriers of these variants relative to non-carriers (49, 50). This has been hypothesised to be due to slower procarcinogen to carcinogen conversion activity of CYP2A6*4 compared to the wild-type allele (51). In this study, we did not identify any association between this variant and SCCHN risk in both the main effect and stratum specific analysis. However, relative to non-smokers and non-carriers of the CYP2A6*2 variant, the joint effect of non-carriers and heavy smokers conferred a 2.5-fold increased risk for SCCHN.

The joint effect analysis in our study suggested the presence of a differential effect of various genotypes with different levels of smoking. Specifically, the joint effects between heavy smoking and multiple genetic variants (e.g., both carriers and non-carriers of CYP1A1*2A, GSTM1null, non-carriers of CYP1A1*2C, CYP2A6*2, CYP2E1c2, GSTP1 105Val) relative to the single reference group, indicated increased risk for SCCHN. However, these differential effects did not translate into conclusive evidence for interaction on additive or multiplicative scales. Nevertheless, we noted a positive interaction between GSTP1 105Val and moderate smoking on the additive scale. It has been demonstrated that there will be interaction on either multiplicative or additive scale if both exposures have an effect on the outcome (52). Hence, in the absence of an association between moderate smoking, and SCCHN, and a negative risk association between 105Val and the outcome in our study, our result indicating excess absolute risk among 105Val carriers who smoked moderately must be interpreted with caution.

The aim of most interaction studies is to identify high-risk sub-groups for targeted public health interventions, especially when resources are limited and cannot target the entire population. The measure of interaction on the additive scale (e.g., RERI), which informs us about absolute risk, is more relevant to identify which group to intervene on, than the multiplicative scale (28). For example, in this study, there was consistency in the direction of interaction on both multiplicative and additive scales for carriers of all variants and both levels of smoking except for carriers of GSTP1 105Val and heavy smoking and carriers of GSTM1null and moderate smoking. Although the imprecision associated with these interaction estimates limits their meaningful interpretation, for public health intervention decision making purposes, one must choose estimates on additive scale, as the multiplicative scale may indicate the wrong subgroup to intervene (28, 53). It is to be cautioned that in a case-control study, it will be erroneous to make inferences about the relative magnitudes of the additive interaction for risks as the probability of risk among the unexposed is not known in this study design (54, 55). Nevertheless, the direction of RERI, which can be consistently estimated from a case-control study, is sufficient to draw conclusions about the public health relevance of interaction (53). For example, if the positive RERI estimate between carriers of GSTP1 variant and moderate smoking, and negative RERI between GSTP1 and heavy smoking were indeed valid, this would indicate that the public health consequence of an intervention on moderate smokers to reduce the risk of SCCHN in this Caucasian population would be larger among carriers of GSTP1 105Val allele carriers while that on heavy smokers would be larger among non-carriers of 105Val.

Certain limitations of this study need to be outlined. Firstly, although our sample size was large relative to many previous studies, our study was underpowered in detecting effects of multiple variants (e.g., CYP1A1*2A, *2C, CYP2E1c2, CYP2A6*2). However, our association results for SNPs in CYP1A1 and CYP2E1c2 were consistent with results from meta-analytical reviews. Furthermore, most estimates for joint effect and interaction analysis were in the expected direction based on the biological mechanisms of these SNPs. The sample size limitation also prevented us from testing co-dominant and recessive models of inheritance for SNP's tested, as well as performing the analyses stratified by oral, pharyngeal and laryngeal cancers. Secondly, smoking, the secondary exposure, was categorized using an outcome based approach which indeed has the potential to induce differential misclassification bias. However, for gene-environment interaction studies, it has been demonstrated that differential misclassification of exposure need not produce a bias under two conditions: 1) there is no association between the genotype and the environmental exposure among controls, 2) there is no association between the genotype and the exposure among cases (56). The bias will be non-differential and towards the null if condition 1 is satisfied but condition 2 is not. A bias analysis conducted in this data showed no evidence of association between the genetic variants and smoking among controls or cases (Supplemental material, eTable 4).

Few strengths of the study are also to be mentioned. First, our outcome based methods for categorising smoking has been documented to best separate the exposure with respect to the outcome by additionally applying Bonferroni correction to account for multiple comparisons of various cut points possible over the range of values identified using the dose response curves. Secondly, a comprehensive list of potential confounders identified using causal graphs, were included in the regression models to mitigate confounding. Attempt was also made to adjust for the appropriate functional forms (e.g., non-linear form) of continuous confounders such as alcohol and time since cessation of smoking. In addition, careful consideration was also given to specify the regression models. Thirdly, although recall error is inherent in case-control studies, we used a life-grid tool which is documented to improve an individual's recall (57).

In conclusion, our study suggests that GSTP1 105Val SNP alters susceptibility to SCCHN among non-smokers, as well as heavy smokers in this Caucasian population. For variants such as GSTP1 105Val whose frequency is high among Caucasians, even moderate effects on cancer risk may be of significant population impact. The lower power to detect main effects and gene-environment interactions for most variants, along with the rigorous approaches used to mitigate bias due to confounding may explain why we could not document any statistical evidence for interaction on additive or multiplicative scale in this study. Larger studies utilizing similar methodology are required for a more definitive investigation of causal gene-environment interactions for high-risk group identification and facilitating targeted smoking interventions for reducing overall risk of SCCHN in this population.

6.1 Supplemental material: Manuscript II

Genetic variants in CYP and GST genes, smoking and risk for head and neck cancers: a gene-environment interaction study

Supplemental material

eAppendix

Calculation of frequency variable for ethanol

Ethanol frequency variable was calculated from information collected on: (a) type of beverage (wine/cider, beer, hard liquor, aperitif or any other), (b) duration (from age to age), (c) quantity (small glass-50ml, medium glass-100ml, big glass-250 ml, half bottle-330 ml, bottle-700-750ml), and (d) frequency of consumption (how many per day, week or month) corresponding to multiple stable consumption periods across life. Each beverage type was converted to ethanol assuming 10% ethanol content in wine and aperitif, 5% for beer/cider and 50% for hard liquor (1).

Reference

1. Schlecht NF, Franco EL, Pintos J, Negassa A, Kowalski LP, Oliveira BV, et al. Interaction between Tobacco and Alcohol Consumption and the Risk of Cancers of the Upper Aero-Digestive Tract in Brazil. American Journal of Epidemiology. 1999;150(11):1129-37.

eTable 1: Impact of definition of an “ex-smoker” on estimated odds ratios for head and neck cancers, Montreal, Quebec, n=818

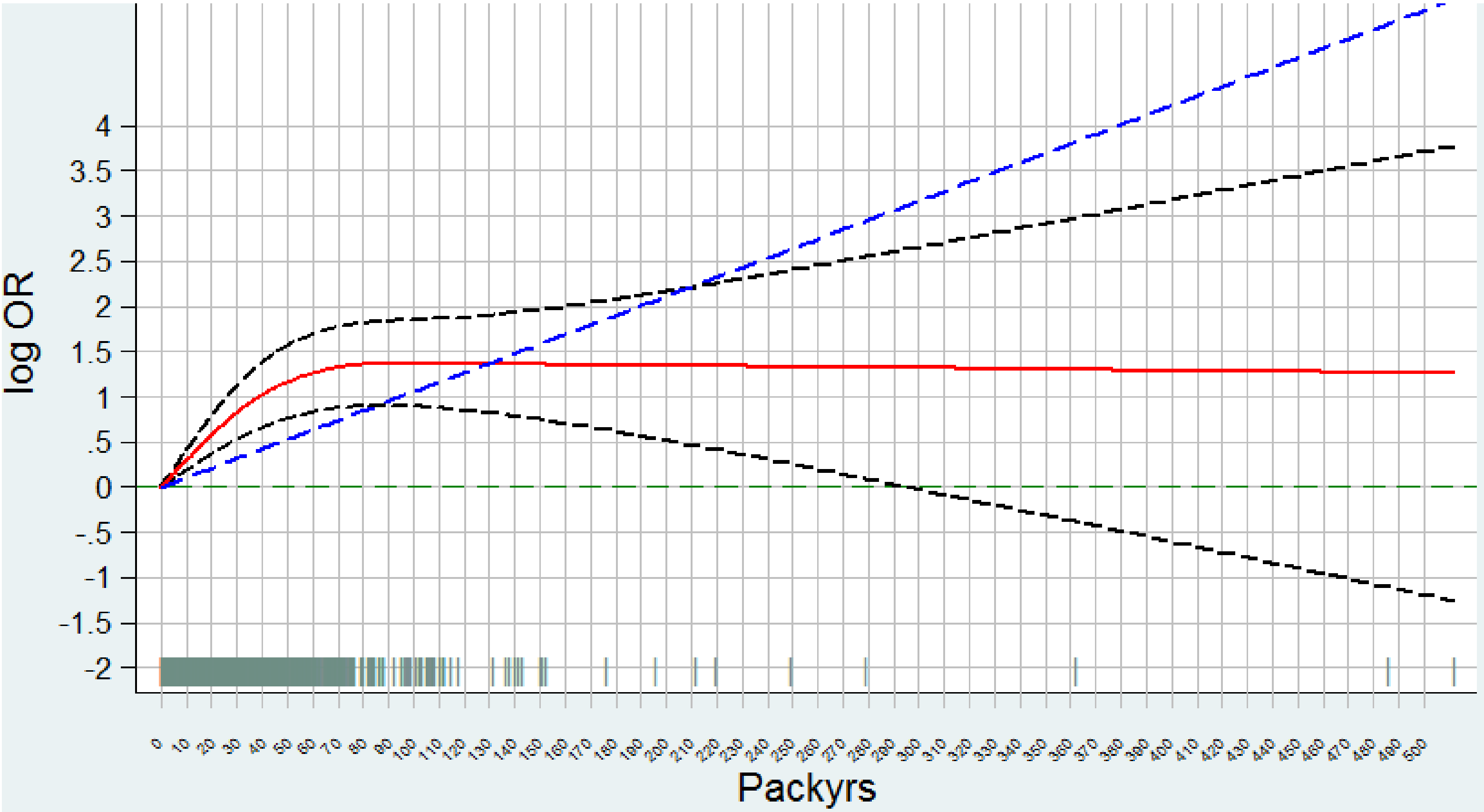
Model	Cutoff[a]	Smoking status	Controls (n =429)	Cases (n = 389)	OR (95% CI)[b]	AIC[c]
			N (%)	N (%)		
1	NA	Ever smoker	312(72)	321(83)	1.73(1.18,2.56)	982
2	One day	Current	94(22)	83(21)	1.35 (0.84,2.20)	991
		Ex smoker	218(51)	238(61)	1.89 (1.27,2.83)	
3	Within 1 year	Current	96(22)	132(34)	2.30 (1.45,3.63)	989
		Ex smoker	216(50)	189(49)	1.50 (1.00,2.27)	
4	1 year	Current	100(22)	154(21)	2.61 (1.66,4.10)	981
		Ex smoker	213(51)	167(61)	1.32 (0.87,1.99)	
5	**2 year**	**Current**	**103(24)**	**158(41)**	**2.68 (1.71-4.19)**	978
		Ex smoker	**209(49)**	**163(42)**	**1.32 (0.87,2.00)**	
6	3 year	Current	108(25)	164(42)	2.55 (1.64,3.99)	980
		Ex smoker	205(48)	157(40)	1.29 (0.85,1.97)	
7	5 year	Current	120(28)	173(44)	2.41 (1,55,3.75)	983
		Ex smoker	193(45)	157(38)	1.31 (0.86,1.99)	
8	10 year	Current	146(34)	203(52)	2.35 (1.52,3.60)	982
		Ex smoker	167(39)	118(30)	1.22 (0.79,1.88)	

[a] Cutoff corresponding to the minimum time interval for which the participants were required to have stopped smoking to be considered ex-smoker

[b] All estimates relative to non-smokers

[c] Akaike’s Information Criteria. A lower AIC indicates best fit to data

efigure1: Restricted cubic spline graph for the association between pack-years of smoking and head and neck cancers, HeNCe



The solid red line represents the estimates from restricted cubic spline and black dashed lines represent associated 95% confidence intervals. Blue dash line represents the association between pack-years and SCCHN when assuming linear functional form of smoking. The rug plot over x-axis represents the distribution of pack-years among participants. Model was conditioned on time since cessation of smoking, indicator for ex-smoker, age (continuous), sex, RC spline of ethanol frequency, and HPV risk

eTable 2: Total effect of genetic markers on head and neck cancer risk, stratified by HPV risk types, Montreal, Canada, 2005-13

Genetic variant	Genotype	Full sample	Sample A*	Sample B*	Sample C*
		OR (95%CI)	OR (95%CI)	OR (95%CI)	OR (95%CI)
CYP1A1*2A	TT				
	CT/CC	0.93 (0.61, 1.41)[a]	0.94 (0.61, 1.45)[a]	0.82 (0.51, 1.31)[a]	0.83 (0.51, 1.34)[a]
CYP1A1*2C	AA				
	AG/GG	0.90 (0.47, 1.70)[b]	0.87 (0.44, 1.70)[b]	1.05 (0.52, 2.14)[b]	1.02 (0.48, 2.14)[b]
CYP2E1(c2)	GG				
	GC/CC	0.67 (0.37, 1.19)[c]	0.56 (0.33, 1.37)[c]	0.64 (0.33, 1.21)[c]	0.56 (0.28, 1.12)[c]
CYP2A6*2	TT				
	AT/AA	0.94 (0.50, 1.77)[d]	0.77 (0.38, 1.54)[d]	1.02 (0.51, 2.05)[d]	0.99 (0.48, 2.02)[d]
GSTP1 Val	AA				
	AG/GG	**0.71 (0.53, 0.95)**[e]	**0.65 (0.48, 0.89)**[e]	**0.72 (0.52, 0.98)**[e]	**0.68 (0.49, 0.95)**[e]
CYP2D6null	3 to 9				
	1 to 2	1.21 (0.74, 2.00)[e]	1.24 (0.73, 2.12)[e]	1.33 (0.75, 2.37)[e]	1.30 (0.72, 2.35)[e]
GSTM1 cnv	1-3				
	0 (Null)	0.94 (0.70, 1.26)e	0.97 (0.71, 1.33)[e]	1.01 (0.73, 1.40)[e]	1.12 (0.80, 1.57)[e]

*****HPV variable was a 4-category variable with HPV-ve, HPV other, HPV alpha-9 other than HPV 16 and HPV 16 as categories. Form these, 3 samples were created for the analysis stratified by HPV. Sample A= Sample without HPV risk types except HPV-16; Sample B= Sample without HPV-16; Sample C= Sample without HPV alpha-9 and HPV-16

CNV: Copy number variation, OR: Odds ratio; CI: Confidence interval

*Numbers may not add up to 818 due to missing genotypes among variants studied

[a] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education, CYP1A1*2C

[b] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education, CYP1A1*2A

[c] OR conditioned on age, sex, RC spline of smoking, RC spline of ethanol, education

[d] OR conditioned on age, sex, education

[e] OR conditioned on age, sex, RC spline of smoking, education

eTable 3: Joint, stratum specific and interaction effects of GSTP1105 Val and smoking levels, on the risk of head and neck cancers, stratified by HPV risk types, Montreal, Canada, 2005-13

Genetic variant	**N**	**Non-smoker**	**N**	**Moderate smoker**	**N**	**Heavy smoker**
	Co/Ca	**OR (95% CI)**	**Co/Ca**	**OR (95% CI)**	**Co/Ca**	**OR (95% CI)**
Sample A = Sample without HPV risk types except HPV-16						
AA	46/37	1	68/36	0.70 (0.32, 1.51)	52/102	1.79 (0.78, 4.12)
AG/GG	63/26	**0.49 (0.26, 0.92)**	64/36	0.79 (0.36, 1.74)	85/93	0.98 (0.45, 2.15)
ORs (95% CI) for AG/GG within strata of smoking levels		**0.49 (0.26, 0.92)**		1.13 (0.63, 2.04)		**0.55 (0.35, 0.87)**
Measures of interaction on multiplicative scale: Ratio of ORs (95% CI)				2.32 (0.97, 5.54)		1.13 (0.51, 2.49)
Measures of interaction on additive scale: RERI (95% CI)				**0.55 (0.17, 0.93)**		-0.54 (-2.08,0.98)
Sample B= Sample without HPV-16						
AA	50/24	1	95/40	0.86 (0.39, 1.92)	46/83	**2.70 (1.13, 6.45)**
AG/GG	63/17	0.52 (0.25, 1.09)	86/37	0.91 (0.41, 2.05)	78/83	1.63 (0.71, 3.70)
ORs (95% CI) for AG/GG within strata of smoking levels		0.52 (0.25, 1.09)		1.06 (0.61, 1.82)		**0.60 (0.37, 0.98)**
Measures of interaction on multiplicative scale: Ratio of ORs (95% CI)				2.01 (0.80, 5.03)		1.14 (0.47, 2.77)
Measures of interaction on additive scale: RERI (95% CI)				0.48 (-0.04, 1.00)		-1.12 (-3.77,1.52)
Sample C= Sample without HPV alpha-9 and HPV-16						
AA	49/22	1	92/37	0.92 (0.40, 2.09)	44/79	**2.82 (1.14, 6.96)**
AG/GG	62/15	0.50 (0.23, 1.08)	84/32	0.91 (0.39, 2.11)	72/75	1.68 (0.72, 3.96)
ORs (95% CI) for AG/GG within strata of smoking levels		0.50 (0.23, 1.08)		0.99 (0.56, 1.75)		**0.60 (0.36, 0.98)**
Measures of interaction on multiplicative scale: Ratio of ORs (95% CI)				1.97 (0.76, 5.12)		1.19 (0.48, 2.97)
Measures of interaction on additive scale: RERI (95% CI)				0.44 (-0.09, 0.98)		-1.52 (-4.89,1.85)

Co: controls; Ca: cases; OR: Odds ratio; CI: Confidence interval; RERI: Relative excess risk due to interaction
All models conditioned on age, sex, education, time since cessation of smoking (mean centered, non-smokers coded 0), indicator for ex-smoker, RC spline of ethanol frequency, and HPV

Bias analysis to identify bias due to potential differential misclassification of smoking and alcohol exposures in gene-environment interaction analysis

eTable 4: Association between genetic variants and smoking levels among controls and cases using multinomial logistic regression model

Genetic variant	Smoking pack-years (0=never, 1=moderate, 2=heavy) OR (95% CI)	
	Among controls	**Among Cases**
CYP1A1*2A	(1) 0.85 (0.43, 1.69)	(1) 0.86 (0.35, 2.09)
	(2) 0.62 (0.27, 1.40)	(2) 1.17 (0.51, 2.70)
CYP1A1*2C	(1) 1.81 (0.60, 5.42)	(1) 1.12 (0.28, 4.48)
	(2) 1.75 (0.48, 6.38)	(2) 0.82 (0.21, 3.15)
CYP2E1(c2)	(1) 0.95 (.42, 2.13)	(1) 0.98 (0.22, 4.38)
	(2) 0.63 (0.24, 1.69)	(2) 1.30 (0.31, 5.32)
CYP2A6*2	(1) 1.49 (0.49, 4.44)	(1) 0.83 (0.24, 2.85)
	(2) 1.60 (0.49, 5.20)	(2) 0.49 (0.14, 1.61)
CYP2D6null CNV	(1) 1.39 (0.59, 3.24)	(1) 2.61 (0.86, 7.92)
	(2) 0.77 (0.33, 1.76)	(2) 1.87 (0.76, 4.60)
GSTP1 105Val	(1) 0.73 (0.45, 1.17)	(1) 1.29 (0.69, 2.41)
	(2) 1.38 (0.81, 2.34)	(2) 1.47 (0.82, 2.63)
GSTM1null CNV	(1) 0.84 (0.52, 1.36)	(1) 0.73 (0.38, 1.37)
	(2) 1.05 (0.62, 1.36)	(2) 1.10 (0.61, 2.20)

All models were adjusted for age, sex, education, alcohol use

No evidence for association between genetic variants and smoking levels among controls.

No evidence for association identified between genetic variants and smoking levels among cases.

Inference= no evidence for bias due to potential differential misclassification of smoking exposure

References

1. Hoffmann D, Hoffmann I. Chemistry and toxicology.In: US Department of Health and Human Services. Cigars: health effects and trends (Smoking and Tobacco Control Monograph 9). DHHS (Publ No. NIH 98-4302), 1998:55–104.
2. Bartsch H, Nair U, Risch A. Genetic Polymorphism of CYP Genes , Alone or in Combination , as a Risk Modifier of Tobacco-related Cancers 2000:3-28.
3. Geisler SA, Olshan AF. GSTM1, GSTT1, and the Risk of Squamous Cell Carcinoma of the Head and Neck: A Mini-HuGE Review. American Journal of Epidemiology. 2001;154(2):95-105.
4. Ho T, Wei Q, Sturgis EM. Epidemiology of carcinogen metabolism genes and risk of squamous cell carcinoma of the head and neck. Head Neck. 2007;29(7):682-99.
5. Varela-Lema L, Taioli E, Ruano-Ravina A, Barros-Dios JM, Benhamou S, Bhisey RA, et al. Meta- and pooled analysis of GSTM1 and CYP1A1 polymorphisms and oropharyngeal cancer: a HuGE-GSEC review. Genetics in medicine : official journal of the American College of Medical Genetics. 2008;10(6):369-84.
6. Zhang Z-j, Hao K, Shi R, Zhao G, Jiang G-x, Song Y, et al. Glutathione S-transferase M1 (GSTM1) and glutathione S-transferase T1 (GSTT1) null polymorphisms, smoking, and their interaction in oral cancer: a HuGE review and meta-analysis. American journal of epidemiology. 2011;173(8):847-57.
7. He X-F, Wei W, Liu Z-Z, Shen X-L, Yang X-B, Wang S-L, et al. Association between the CYP1A1 T3801C polymorphism and risk of cancer: Evidence from 268 case–control studies. Gene. 2014;534(2):324-44.
8. Liu L, Wu G, Xue F, Li Y, Shi J, Han J, et al. Functional CYP1A1 genetic variants, alone and in combination with smoking, contribute to development of head and neck cancers. European Journal of Cancer. 2013;49(9):2143-51.
9. Wang Y, Yang H, Duan G, Wang H. The association of the CYP1A1 Ile462Val polymorphism with head and neck cancer risk: evidence based on a cumulative meta-analysis. OncoTargets and therapy. 2016;9:2927-34.
10. Zhuo X, Song J, Liao J, Zhou W, Ye H, Li Q, et al. Does CYP2E1 RsaI/PstI polymorphism confer head and neck carcinoma susceptibility?: A meta-analysis based on 43 studies. Medicine (Baltimore). 2016;95(43):e5156.
11. Hashibe M, Brennan P, Strange RC, Bhisey R, Cascorbi I, Lazarus P, et al. Meta- and pooled analyses of GSTM1, GSTT1, GSTP1, and CYP1A1 genotypes and risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev. 2003;12(12):1509-17.
12. Lang J, Song X, Cheng J, Zhao S, Fan J. Association of GSTP1 Ile105Val Polymorphism and Risk of Head and Neck Cancers : A Meta-Analysis of 28 Case- Control Studies. 2012;7(11).
13. Singh M, Shah PP, Singh AP, Ruwali M, Mathur N, Pant MC, et al. Association of genetic polymorphisms in glutathione S-transferases and susceptibility to head and neck cancer. 2008;638:184-94.
14. Iarc LF. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans1985.
15. Freedman ND, Abnet CC, Leitzmann MF, Hollenbeck AR, Schatzkin A. Prospective investigation of the cigarette smoking-head and neck cancer association by sex. Cancer. 2007;110(7):1593-601.

16. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chen C, Curado MP, 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):777-89.

17. Egan KM, Abruzzo J, Cytobrush B, Newcomb PA, Titus-ernstoff L, Franklin T, et al. Collection of Genomic DNA from Adults in Epidemiological Studies by Buccal Cytobrush and Mouthwash. 2001:687-96.

18. D'Souza G, Sugar E, Ruby W, Gravitt P, Gillison M. Analysis of the effect of DNA purification on detection of human papillomavirus in oral rinse samples by PCR. J Clin Microbiol. 2005;43(11):5526-35.

19. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature reviews Genetics. 2006;7(2):85-97.

20. He Y, Hoskins JM, McLeod HL. Copy Number Variants in pharmacogenetic genes. Trends in molecular medicine. 2011;17(5):244-51.

21. Huang RS, Chen P, Wisel S, Duan S, Zhang W, Cook EH, et al. Population-specific GSTM1 copy number variation. Human molecular genetics. 2009;18(2):366-72.

22. Laprise C, Madathil SA, Schlecht NF, Castonguay G, Soulières D, Nguyen-Tan PF, et al. Human papillomavirus genotypes and risk of head and neck cancers: Results from the HeNCe Life case-control study. Oral oncology. 2017;69:56-61.

23. Schlecht NF, Franco EL, Pintos J, Negassa A, Kowalski LP, Oliveira BV, et al. Interaction between Tobacco and Alcohol Consumption and the Risk of Cancers of the Upper Aero-Digestive Tract in Brazil. American Journal of Epidemiology. 1999;150(11):1129-37.

24. Leffondre K. Modeling Smoking History: A Comparison of Different Approaches. American Journal of Epidemiology. 2002;156(9):813-23.

25. Williams BA, Madrekar JN, Madrekar SJ, Cha SS, Furth AF. Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes Mayo Clinic, Rochester, Minnesota Division of Biostatistics DoHSR; June 2006. Contract No.: 10027230.

26. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology (Cambridge, Mass). 1999;10(1):37-48.

27. VanderWeele TJ. Explanation in Causal inference: Methods for mediation and Interaction. Press OU, editor. USA2015. 706 p.

28. Rothman KJ. Modern epidemiology. Boston: Little, Brown; 1986.

29. Mulder TPJ, Manni JJ, Roelofs HMJ, Peters WHM, Wiersma A. Glutathione S-transferases and glutathione in human head and neck cancer. Carcinogenesis. 1995;16(3):619-24.

30. Jourenkova-Mironova N, Voho A, Bouchardy C, Wikman H, Dayer P, Benhamouand S, et al. Glutathione S-transferase GSTM1, GSTM3, GSTP1 and GSTT1 genotypes and the risk of smoking-related oral and pharyngeal cancers. International journal of cancer. 1999;81(1):44-8.

31. Ryberg D, Skaug V, Hewer A, Phillips DH, Harries LW, Wolf CR, et al. Genotypes of glutathione transferase M1 and P1 and their significance for lung DNA adduct levels and cancer risk. Carcinogenesis. 1997;18(7):1285-9.

32. Ruwali M, Pant MC, Shah PP, Mishra BN, Parmar D. Polymorphism in cytochrome P450 2A6 and glutathione S-transferase P1 modifies head and neck cancer risk and treatment outcome. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2009;669(1–2):36-41.

33. Ruzzo A, Canestrari E, Maltese P, Pizzagalli F, Graziano F, Santini D, et al. Polymorphisms in genes involved in DNA repair and metabolism of xenobiotics in individual susceptibility to sporadic diffuse gastric cancer. Clin Chem Lab Med. 2007;45(7):822-8.

34. Vlaykova T, Miteva L, Gulubova M, Stanilova S. Ile105Val GSTP1 polymorphism and susceptibility to colorectal carcinoma in Bulgarian population. Int J Colorectal Dis. 2007;22(10):1209-15.

35. Jiao L, Bondy ML, Hassan MM, Chang DZ, Abbruzzese JL, Evans DB, et al. Glutathione S-transferase Gene Polymorphisms and Risk and Survival of Pancreatic Cancer. Cancer. 2007;109(5):840-8.

36. Chen JB, Wang F, Wu JJ, Cai M. Glutathione S-transferase pi polymorphism contributes to the treatment outcomes of advanced non-small cell lung cancer patients in a Chinese population. Genet Mol Res. 2016;15(3).

37. Hu X, Xia H, Srivastava SK, Herzog C, Awasthi YC, Ji X, et al. Activity of four allelic forms of glutathione S-transferase hGSTP1-1 for diol epoxides of polycyclic aromatic hydrocarbons. Biochem Biophys Res Commun. 1997;238(2):397-402.

38. Saarikoski ST, Voho A, Reinikainen M, Anttila S, Karjalainen A, Malaveille C, et al. Combined effect of polymorphic GST genes on individual susceptibility to lung cancer. International journal of cancer. 1998;77(4):516-21.

39. Sundberg K, Johansson AS, Stenberg G, Widersten M, Seidel A, Mannervik B, et al. Differences in the catalytic efficiencies of allelic variants of glutathione transferase P1-1 towards carcinogenic diol epoxides of polycyclic aromatic hydrocarbons. Carcinogenesis. 1998;19(3):433-6.

40. Liu L, Wu G, Xue F, Li Y, Shi J, Han J. Functional CYP1A1 genetic variants , alone and in combination with smoking , contribute to development of head and neck cancers. European Journal of Cancer. 2013;49(9):2143-51.

41. Qin J, Zhang J-X, Li X-P, Wu B-Q, Chen G-B, He X-F. Association between the CYP1A1 A2455G polymorphism and risk of cancer: evidence from 272 case–control studies. Tumor Biology. 2014;35(4):3363-76.

42. Xie S, Luo C, Shan X, Zhao S, He J, Cai Z. CYP1A1 MspI polymorphism and the risk of oral squamous cell carcinoma: Evidence from a meta-analysis. Mol Clin Oncol. 2016;4(4):660-6.

43. Lu D, Yu X, Du Y. Meta-analyses of the effect of cytochrome P450 2E1 gene polymorphism on the risk of head and neck cancer. Mol Biol Rep. 2011;38(4):2409-16.

44. Tang K, Li Y, Zhang Z, Gu Y, Xiong Y, Feng G, et al. The PstI/RsaI and DraI polymorphisms of CYP2E1and head and neck cancer risk: a meta-analysis based on 21 case-control studies. BMC Cancer. 2010;10(1):575.

45. Tripathy CB, Roy N. Meta-analysis of glutathione S-transferase M1 genotype and risk toward head and neck cancer. Head & neck. 2006;28(3):217-24.

46. Hiyama T, Yoshihara M, Tanaka S, Chayama K. Genetic polymorphisms and head and neck cancer risk (Review). 2008:945-73.

47. Zhuo W, Wang Y, Zhuo X, Zhu Y, Wang W, Zhu B, et al. CYP1A1 and GSTM1 polymorphisms and oral cancer risk: association studies via evidence-based meta-analyses. Cancer Invest. 2009;27(1):86-95.

48. Brunotto M, Zarate AM, Bono A, Barra JL, Berra S. Risk genes in head and neck cancer: a systematic review and meta-analysis of last 5 years. Oral oncology. 2014;50(3):178-88.

49. Kamataki T, Fujieda M, Kiyotani K, Iwano S, Kunitoh H. Genetic polymorphism of CYP2A6 as one of the potential determinants of tobacco-related cancer risk. Biochemical and Biophysical Research Communications. 2005;338(1):306-10.

50. Canova C, Richiardi L, Merletti F, Pentenero M, Gervasio C, Tanturri G, et al. Alcohol, tobacco and genetic susceptibility in relation to cancers of the upper aerodigestive tract in northern Italy. Tumori. 2010;96(1):1-10.

51. Raunio H, Rautio A, Gullstén H, Pelkonen O. Polymorphisms of CYP2A6 and its practical consequences. British Journal of Clinical Pharmacology. 2001;52(4):357-63.

52. Greenland S, Lash TL, Rothman KJ. Concepts of Interaction. In: Modern Epidemiology. 3rd ed. Rothman KJ, Greenland S, Lash TL, editors. Philadelphia: Lippincott Williams and Wilkins; 2008.

53. VanderWeele TJ. An introduction to interaction analysis. Explanation in causal inference: methods of mediation and interaction. United States of America: Oxford University press; 2015. p. 250-73.

54. Greenland S. Additive risk versus additive relative risk models. Epidemiology (Cambridge, Mass). 1993;4(1):32-6.

55. Richardson DB, Kaufman JS. Estimation of the Relative Excess Risk Due to Interaction and Associated Confidence Bounds. American Journal of Epidemiology. 2009;169(6):756-60.

56. Garcia-Closas M, Rothman N, Lubin J. Misclassification in Case-Control Studies of Gene-Environment Interactions: Assessment of Bias and Sample Size. Cancer Epidemiology Biomarkers &amp;amp; Prevention. 1999;8(12):1043.

57. Berney L, Blane DB, Soc Sci M, Blane B. Collecting retrospective data: accuracy of recall after 50 years judged against historical records. Social Science & Medicine. 1997

Chapter 7

Manuscript III

The effect of interdependencies between CYP2A6 variant and smoking, and ADH1B variant and alcohol, on the risk of head and neck cancers

Akhil Soman ThekkePurakkal[1], Belinda Nicolau[1], Robert D Burk[2], Jay S Kaufman[3], Nicolas F Schlecht[4]

[1]Division of Oral Health and Society, Faculty of Dentistry, McGill University, Montreal, Canada; [2]Departments of Pediatrics (Genetics), Microbiology & Immunology, Obstetrics Gynecology & Women's Health, and Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA Departments of Oncology and Epidemiology & Biostatistics, McGill University, Montreal, Quebec, Canada; [3]Department of Epidemiology, Biostatistics, & Occupational Health, McGill University, Montreal, Canada;[4]Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA

Keywords: Mediation, 4-way decomposition, counterfactual causal framework, Canadian Caucasian, direct and indirect effects; case-control

List of abbreviations and acronyms: SCCHN: Squamous cell carcinomas of the head and neck, CYP2A6: Cytochrome P450 2A6 enzyme, ADH1B: Alcohol dehyderogenase 1B enzyme, SNP: Single nucleotide polymorphism; HPV: Human papillomavirus, DNA: Deoxyribonuclic acid; PCR: Polymerase chain reaction; RR: Relative risk; CI: Confidence Interval; CDE: Controlled direct effect, TE: Total effect, NDE: Natural direct effect, NIE: Natural indirect effect, PIE: Pure indirect effect, INTref: Reference interaction, INTmed: Mediated interaction, PM: Proportion mediated, PAI: Proportion attributable to interaction, PE: Proportion eliminated

Abstract

Background: Smoking and alcohol risk behaviours may interact as well as mediate the effect of genetic variants in CYP2A6 and ADH1B respectively on SCCHN risk. However, their mediated effects on SCCHN, as well as effects under combined mediation and interaction with the associated risk behaviours, have not been quantified yet. In this study, we aimed to estimate the extent to which the effect of CYP2A6*2 and ADH12B*2 on SCCHN is mediated by smoking and alcohol risk behaviours respectively. In addition, we use this data to demonstrate how much of the effect of these genetic exposures on SCCHN is through four potential causal pathways which may or may not involve the associated risk behaviour under a combined mediation and interaction scenario.

Methods: A subsample of Caucasian smokers (controls=312, cases=321), and alcohol consumers (controls=353, cases=325) with genetic data, obtained from a hospital based case-control study were analysed. Incident SCCHN cases, frequency matched by age and sex, were recruited from four main hospitals in Montreal. Interviews collected information on several domains of exposures. DNA was isolated from oral exfoliated cells. CYP2A6*2 was genotyped as AT/AA (CYP2A6*2 carriers) and TT (non-carriers), and ADH1B*2 as GA/AA (carriers) and GG (non-carriers). Smoking and drinking intensities were measured by cigarettes smoked and amount (ml) of ethanol consumed per day, respectively. Mediation and 4-way decomposition analysis based on counterfactual causal framework was used to derive risk estimates and proportions.

Results: Among smokers, the total effect estimate of CYP2A6 variant on SCCHN [Relative risk (RR) =1.28, 95% CI: 0.46, 3.59] was composed of a direct effect estimate of 1.22 (95% CI: 0.45, 3.33) and an indirect effect estimate through smoking of 1.05 (95% CI: 0.94, 1.17). Among alcohol, the total effect estimate of the ADH1B variant on SCCHN [RR= 2.37, 95% CI: 1.12, 4.25) was decomposed into a direct effect estimate of 2.24 (95% CI: 0.88, 5.71), and indirect effect estimate of 1.06 (95% CI: 0.97, 1.16). Approximately 65% and 84% of excess risk of SCCHN due to CYP2A6 and ADH1B did not involve heavy intensities of smoking and alcohol behaviours respectively.

Conclusion: The majority of the effect of each genetic variant on SCCHN risk seemed to operate through pathways other than changing the associated risk behaviour. However, mediation and interaction by the risk behaviours may play a role in their effects on SCCHN.

Introduction

Smoking and alcohol consumption are well-established risk factors for squamous cell carcinomas of the head and neck (SCCHN) (1). Single nucleotide polymorphisms (SNP) in genes encoding tobacco and alcohol metabolising enzymes [Cytochrome P450 2A6 (CYP2A6) and Alcohol dehydrogenase 1B (ADH1B), respectively] have been implicated in the risk for SCCHN among smokers and alcohol consumers, respectively (2-4). However, the underlying potential casual pathways that may or may not involve these risk behaviours have not yet been quantified.

CYP2A6 is the primary enzyme responsible for the oxidation of more than 80% of nicotine entering the body. Carriers of the slow metabolizing CYP2A6*2 variant (A allele) metabolize nicotine at slower rates, display decreased nicotine clearance, higher plasma nicotine levels and consequently smoke at lower intensities (to maintain optimal nicotine levels) relative to wildtype (T allele) carriers, lowering the risk for tobacco related cancers (5-7). Studies on CYP2A6 SNPs with similar functional consequences as CYP2A6*2 support this hypothesis for SCCHN risk (2). Similarly, the ADH1B*2 polymorphism (A allele) encodes a version of the ADH1B enzyme that converts ethanol to acetaldehyde 50-100 times faster compared to wildtype (G allele) carriers (8). Individuals who lack this SNP do not exhibit aversive physiological reactions (alcohol induced flushing) associated with prompt build-up of acetaldehyde, documented among carriers, and consequently are associated with increased likelihood of heavy alcohol consumption (9-11). Consequently, non-carriers of these variants may have a higher risk for SCCHN relative to carriers (4, 12). This raises the possibility of an indirect causal pathway, whereby absence of CYP2A6*2 and ADH1B*2 variants lead to higher intensities of smoking and alcohol consumption, which may mediate the risk for SCCHN. However, the potential indirect and direct causal pathways from each of these genetic exposures to the risk for SCCHN have not been quantified yet.

Alternatively, there is also evidence that relative to carriers, individuals who lack the CYP2A6 and ADH1B SNPs have increased susceptibility to the carcinogenic effects of tobacco smoking and alcohol consumption, respectively suggesting interaction (2, 4, 6, 13). Under this scenario where a single exposure may interact with a potential single mediator, the total effect of the exposure on the outcome can be deciphered into four non-overlapping causal pathways: one that does not involve mediation or interaction, one that involves only mediation but not interaction, one that involves only interaction but not mediation, and one that involves both mediation and interaction

(14). In a case-control study, it is possible to consistently estimate the proportions of the excess risk for the outcome attributable to the four potential causal pathways (14). Quantifying these pathways in our scenario can provide greater insights into the direct and indirect effects of tobacco/alcohol metabolizing genes on risk of SCCHN which may or may not involve the associated smoking/alcohol consumption behaviours.

The primary objective of this study was to estimate the extent to which the total effects of two functional SNPs in CYP2A6 and ADH1B on SCCHN risk are mediated by heavy smoking and alcohol consumption, respectively, in a case-control sample of Canadian Caucasians. In addition, we used this data to demonstrate the estimation of proportions of excess risk attributable to four underlying pathways possible involving each genetic exposure, associated risk behaviour and SCCHN.

Methods

Population, study design and data collection

The data were drawn from the Canadian site of an international hospital-based case-control study, Head and Neck Cancer (HeNCe) Life, investigating the aetiology of head and neck cancers in relation to social, behavioural, lifestyle, biologic and genetic risk factors. Adult participants (N= 918) were recruited from the outpatient clinics of four major referral hospitals in Montreal between 2005 and 2013. Participant eligibility criteria for the study were: (I) born in Canada, (II) aged ≥18 years, (III) English or French speaking; (IV) living within 50 Km from the hospitals, (V) without history of cancer, and (VI) without mental or immune suppression disorders. Cases (N=460) included consecutive, incident, histologically confirmed, stage I to IV squamous cell carcinomas of mouth, oropharynx and larynx (C01-C06, C09, C10, C12- C14, and C32, under the International Statistical Classification of Diseases, 10th Revision). Cancer-free controls (N=458), frequency matched to each identified case by 5-year age group and sex, were randomly selected from 10 outpatient clinics in the same hospitals from a list of non-chronic diseases not strongly associated with tobacco and alcohol consumption (with no single diagnostic group contributing to more than 20% of the total). Ethics approval was obtained from review boards of McGill University, Institut National de la Recherche Scientifique, and all participating hospitals. All participants signed an informed consent form prior to enrolment in the study.

Face-to face interviews using a questionnaire with a life-grid technique (15) were used to collect data on several domains of exposure including sociodemographic factors, lifetime history of tobacco smoking and alcohol consumption.

Sample collection and analysis: Genetic polymorphisms and Human papillomavirus

Oral epithelial samples for genetic and human papillomavirus (HPV) analysis were collected using brush biopsy and oral rinse following standardized protocols (16, 17). To identify genetic polymorphisms, genotyping was performed on DNA isolated from the samples using real-time Taqman PCR expression assays. Briefly, reactions were set up using 5 µl of 2X Genotyping Master Mix (Applied Biosystems, Foster City, CA) combined with assay-specific concentrations of primers and probes, and 10 ng of sample DNA. The reactions were then spun down at 2000 rpm for 2 min, and run in the 7500FAST real-time PCR thermocycler in Genotyping mode under default settings. 7500FAST v2.0.1 (updated to v2.0.6) software was used for allelic discrimination. The 2 SNPs selected for this analysis were CYP2A6*2 (rs1801272) and ADH1B*2 (rs1229984). HPV DNA detection and typing was performed as described elsewhere (18).

Data Analysis

*Genetic exposure definition: CYP2A6*2 and ADH1B*2*

CYP2A6*2 was genotyped as TT, AA and AT (A = minor allele), and ADH1B*2 was genotyped as GG, AA and AG (A = minor allele). We used both the genetic exposures as binary based on, a) the underlying biological mechanism of action of the enzymes they encode (4, 6, 9, 10, 13), and b) proportion of participants homozygous for the minor allele of both variants, which were low. Carriers of CYP2A6*2 (AT/AA) and non-carriers (TT) were coded as 0 and 1 respectively. For ADH1B*2, carriers (AG/AA) and non-carriers (GG) were coded as 0 and 1 respectively.

Risk behaviour- mediator- definition: Tobacco smoking, alcohol consumption

CY2A6*2 and ADH1B*2 SNPs are associated with intensity of smoking and alcohol consumption respectively (6, 19). Therefore, we used frequency of smoking and alcohol use as measures for these risk behaviours. For tobacco smoking, we collected detailed information on commercial cigarettes, hand rolled cigarettes, cigars and pipes [duration (age of cessation minus age of initiation), frequency (how many per day)] during multiple stable smoking periods over an individual's life. All tobacco types were then converted to the commercial cigarette equivalent based on nicotine content (1/9 cigar = 1/3.5 pipe=1/2 hand rolled cigarettes= 1 commercial cigarette) (20). From the total duration and frequency of commercial cigarettes used, we calculated average number of commercial cigarettes smoked per day over the lifetime. A non-linear dose-response relationship was identified between cigarettes smoked per day and SCCHN risk with risk increasing till 35 cigarettes per day and plateauing thereafter (Supplemental material, eFigure 1). Using a parametric outcome-based approach (21), we identified 18 cigarettes per day as the optimal cut point and used this threshold to dichotomize participants into moderate smokers (>0 to 18 cigarettes per day, coded 0) and heavy smokers (>18 cigarettes per day, coded 1).

We collected similar information for alcohol consumption: multiple beverages [type (wine/cider, beer, hard liquor, aperitif), duration (age of cessation minus age of initiation), quantity (small glass-50ml, medium glass-100ml, big glass-250 ml, half bottle-330 ml, bottle-700-750ml), and frequency of consumption (how many per day or per week or per month)] for multiple time periods of stable consumption across life. Each beverage was converted to ethanol equivalents (10% ethanol in wine and aperitif, 5% in beer/cider, and 50% in hard liquor) (22). Similar to tobacco intensity, this information was used to calculate the average amount of ethanol (in millilitres) consumed per day over the lifetime. The risk for SCCHN increased till approximately 100ml of ethanol per day and then plateaued (Supplemental material, eFigure 2). An amount of 25ml of ethanol per day was identified as the optimal cut point using the parametric outcome based approach (21), and participants were then grouped into moderate drinkers (>0 to 25ml per day, coded 0) and heavy drinkers (>25ml per day, coded 1).

Mediation and 4-way-decomposition

The mediation and 4-way decomposition models used in this analysis were based on the counterfactual framework for causal inference (23). Under this framework, the *average total effect* (TE) of the genetic exposure on SCCHN in the population can be decomposed into the product of overall *direct (NDE)* and *indirect effects (NIE) on the relative risk scale (24, 25)*. In our scenarios, TE reflects the change in risk of SCCHN for an overall change in the exposure in the population from AT/AA to TT genotype for CYP2A6*2, and AG/AA to GG genotype for ADH1B*2 SNPs.

The NDE is the estimated effect of TT and GG genotypes on SCCHN risk operating through pathways other than heavy smoking and heavy alcohol intensities, respectively. By contrast, the NIE estimated the effects of TT and GG genotypes through heavy smoking and alcohol intensities, respectively.

Alternatively, the four-way decomposition involves segregating the total excess relative risk (i.e., TE-1) for SCCHN among those exposed into four non-overlapping components on the excess relative risk scale (26). These included: (i) *the controlled direct effect (CDE)*; the portion of the effect of genetic exposure on SCCHN risk when the associated risk behaviour intensity is set to moderate/mild levels (i.e., component of excess risk attributed to neither mediation nor interaction with heavy intensity of risk behaviour); (ii) *the reference interaction (INTref)*; the portion of the effect of the genetic exposure that requires the joint presence of high intensity of the associated risk behaviour (interaction alone), with the high intensity behaviour arising independently of the associated genetic exposure; (iii) *the mediated interaction (INTmed);* the portion of the effect of genetic exposure that requires the joint presence of associated heavy intensity behaviour, with the heavy intensity behaviour arising as a consequence of the associated genetic exposure (both interaction and mediation), and (iv) *the pure indirect effect (PIE);* the portion of the effect of genetic exposure that is due to genetic exposure-induced high intensity behaviour (mediation alone). The *overall proportion* of the effect of genetic exposure on SCCHN risk *mediated (PM)* by associated heavy intensity risk behaviour can be calculated as the sum of PIE and INTmed components, divided by the excess relative risk. The *proportion of the effect attributable to interaction (PAI)* between the genetic variant and associated heavy intensity of risk behaviour is given by sum of INTref and INTmed components, divided by the excess relative risk. *Proportion eliminated (PE)* is the proportion of effect of the genetic variant on SCCHN risk that can be eliminated in the population if the level of the associated risk behaviour was decreased to that of moderate/mild intensity in the population. This is given by the sum of INTref, INTmed and PIE, divided by the excess relative risk.

Assumptions for causal interpretation and potential confounders

Causal interpretation of our results through the counterfactual framework rely on four no-confounding assumptions as well as correct model specifications (26): no unmeasured confounding of the effects of (i) genetic exposure on SCCHN risk, (ii) genetic exposure on associated risk behaviour, and (iii) risk behaviour on SCCHN risk, and (iv) none of the risk behaviour-SCCHN confounders are affected by the associated genetic exposures. We addressed assumptions (i) and (ii) by adjusting for age, sex, and education, as well as restricting our analysis to Caucasians, thus mitigating confounding due to population stratification (27). Regarding assumption (iii), we adjusted for potential measured confounders of the relationship between each risk behaviour and SCCHN risk. For the smoking intensity-SCCHN association, we identified duration and time since cessation (continuous, mean centred, current smokers recoded to zero) of smoking, and intensity of alcohol (continuous, adjusted for restricted cubic spline) as confounders. For the alcohol intensity-SCCHN association, time since cessation of use (continuous, mean centred, current users recoded to zero) of alcohol, and pack-years of commercial cigarette equivalence (continuous, adjusted for restricted cubic spline, 20 commercial cigarettes = 4 hand-rolled cigarettes = 4 cigars = 5 pipes = 1 pack of commercial cigarettes) were identified (28). Additionally, we adjusted for age (continuous), sex, number of years of education (continuous) and HPV risk types[8] (as described elsewhere) (18) for both associations. These variables are not known to be affected by the associated genetic exposures which may potentially address assumption (iv).

Statistical analysis

The CYP2A6*2-smoking-SCCHN and ADH1B*2-alcohol-SCCHN analyses were performed only

[8] Our overall results were estimated through models adjusted for HPV risk types where as HPV 16 associated SCCHN cancers have been documented as clinically distinct entities. However, as previously described (*please refer to discussion section of manuscript II, page 145*), statistical model adjusted for HPV is a valid model in estimating the effects of the genetic variants on SCCHN risk.

among smokers and alcohol consumers, respectively as no association has been documented between these genetic variants and SCCHN among non-consumers. The direct and indirect effects and decomposition estimates were obtained by fitting logistic regression models on the binary outcome and mediator (26). For CYP2A6*2-smoking-SCCHN, SCCHN was regressed on the CYP2A6*2, cigarettes per day, their product term (denoting interaction) and associated potential confounders (outcome model). Next, cigarettes per day was regressed on CYP2A6*2 and potential confounders (mediator model, fit only among controls). For ADH1B*2-alcohol-SCCHN, the outcome model was fit on ADH1B*2, ethanol per day, their product term and associated potential confounders. For the mediator model, ethanol per day was fit on ADH1B*2 and potential confounders among controls. An indicator variable for ex-smokers and ex-alcohol consumers was added in the smoking and alcohol related models, respectively, to account for time since cessation of use of the respective products (29). For both scenarios, the mediator model in the full sample (i.e., in both cases and controls) weighted on the sampling fraction (30) gave quantitatively similar estimates as the model fit among controls (26). Effect estimates and associated proportions were obtained by combining parameters from these two models according to their corresponding analytical equations (26, 31). The 95% Confidence Intervals (CI) for the estimates were obtained using bootstrapping (2000 replications). All analyses were carried out using Stata, version 13SE (StataCorp. 2013, College Station, TX). Due to unavailability of Stata codes for carrying out the 4-way decomposition analysis, codes were exclusively written for this work using mathematical equations provided by VanderWeele 2015, 2016 (and personal communication) for the binary outcome, binary mediator and binary exposure scenario (14, 32). Stata codes are provided in *Supplemental material, eAppendix*.

Results

Of the total 918 participants, 818 were genotyped on CYP2A6*2 and ADH1B*2. Of these, 633 and 678 were Caucasian smokers and alcohol consumers among whom 32 (13 controls and 19 cases) and 3 (2 controls and 1 case) had data missing on CYP2A6*2 and ADH1B*2, respectively. Therefore, we present the CYP2A6*2-smoking-SCCHN analysis on a sub-sample of 601 participants (only smokers), and ADH1B*2-alcohol-SCCHN on 675 participants (only alcohol consumers).

Descriptive characteristics of smoker and alcohol consumer sub-samples are given in Tables 1. Briefly, both samples had similar sociodemographic characteristics. Among the smoker sub-sample, cases had higher proportions of TT genotype (CYP2A6*2 non-carriers) and of heavy smokers compared to controls. Similarly, in the alcohol user sub-sample, cases had a higher proportion of GG genotype and of heavy drinkers compared to controls. The estimates (risk ratio, RR) for genetic exposure-SCCHN, mediator-SCCHN, and genetic exposure-mediator (among controls) associations in both samples were all above 1 (Tables 1 and 2).

The results of the standard mediation analysis (2-way decomposition) are given in Table 3. The average TE estimate for SCCHN for a change from AT/AA to TT genotype in the sample of smokers was RR=1.28 (95% CI: 0.46, 3.59) which was composed of a direct effect estimate (NDE) of 1.22 (95% CI: 0.45, 3.33) and an indirect effect estimate (NIE) through smoking of 1.05 (95% CI: 0.94, 1.17). The TE estimate amounted to an excess RR estimate of 0.28 (95% CI: -1.8, 2.37). Among the sample of alcohol consumers, the average TE estimate for a change from AG/AA to GG genotype was RR= 2.37 (95% CI: 1.12, 4.25) with an excess RR estimate of 1.37 (95% CI: 1.62, 4.38). The TE estimate decomposed into a NDE estimate of 2.24 (95% CI: 0.88, 5.71), and NIE estimate of 1.06 (95% CI: 0.97, 1.16).

Table 3 displays the results of the 4-way decomposition demonstration. Among smokers, 65% of the excess RR estimate for SCCHN due to TT genotype (CYP2A6*2 non-carriers) was attributable to the CDE component, 14% to INTref, 11% to PIE, and 10% to INTmed. The overall proportion of risk due to TT genotype mediated by heavy smoking was 21% and the proportion attributable to interaction with heavy smoking was 24%. The overall proportion eliminated was estimated at 35%. Among alcohol consumers, approximately 84% of the excess RR for SCCHN due to GG genotype (ADH1B*2 non-carriers) was attributable to the CDE component. Proportions attributable to the other 3 components were about 5% each. The proportion of risk due to GG genotype mediated by heavy alcohol use was 10%, that attributable to interaction was 11% and the proportion eliminated was approximately 16%.

Discussion

In this study, we aimed to quantify the causal pathways from two functional SNPs in CYP2A6 and ADH1B genes, leading to SCCHN risk which may or may not be mediated by heavy smoking and alcohol consumption behaviours respectively. We further demonstrated the potential for existence of 4 causal pathways between these genetic exposures and SCCHN risk combining mediation and interaction hypotheses. Albeit imprecise, the point estimates seem to indicate that effects of TT genotype (CYP2A6) and GG genotype (ADH1B) on SCCHN risk were mainly through pathways not mediated by heavy smoking or alcohol intensities, respectively.

Before interpreting the results, it is important to consider the limitations of this study. Firstly, our analysis was limited by sample size and confidence intervals of most estimates were wide. This limits our capability to assert that inference based on these estimates are true of the population parameters. Nevertheless, in this work, we intended to demonstrate the technique of decomposition analysis with respect to these genetic variants, respective risk behaviours and SCCHN, that has not been explored in the oral health literature. These analyses were performed based on the positive and monotonic point estimates for the exposure-outcome, mediator-outcome and exposure-mediator associations, whose directions were as documented in the literature based on underlying biological mechanisms. In addition, our point estimates for joint effects and interaction between the genotypes and heavy intensities of respective risk behaviours (Supplementary material, etable 1), were also in the expected direction. Secondly, it is possible that the effect of ADH1B*2 on SCCHN documented in this study is a reflection of it being in linkage disequilibrium with the variants in ADH1C gene (with similar functional consequences). We did not have information on the ADH1C variant to adjust for in the models. However, studies in both Caucasian and Asian populations suggest that the associations between ADHIB*2, intensity of alcohol consumption and SCCHN risk are independent of variants in the ADH1C gene, and were strongest among all alcohol dehydrogenase (ADH) related genes studied (3, 4, 33). Also, the direction of estimates for the association between ADH1B*2, alcohol consumption and SCCHN risk were similar to what has been documented before (3, 19, 34). Thirdly, since the CYP2A6*2-smoking-SCCHN analysis, and ADH1B*2-alcohol-SCCHN was restricted to smokers and alcohol consumers respectively, there is possibility of a collider stratification bias, as selection into the study is affected by both exposure and outcome. This may have led to an underestimation of true causal effects (35). However, not restricting may lead to a higher variance in estimates because a large proportion of the controls have no direct exposure effect (the non-smokers and non-alcohol consumers). Hence, restriction was performed assuming a small bias vs large variance in estimates. Larger studies or data simulations are required to quantify this bias-variance trade-off.

*CYP2A6*2, smoking and SCCHN risk*

Genetic studies have hypothesised that the effect of variants in CYP2A6 on the risk of tobacco related cancers (specifically of squamous cell origin) may be due to interaction, mediation, both interaction and mediation, or independent of smoking (5, 7, 13, 36, 37). However, these potential pathways have not been quantified yet. In our study, among smokers, the total effect point estimate and positive excess risk suggest that TT genotype could confer a higher risk for SCCHN risk. Our results also indicate that this total effect could be composed of a large direct effect and a small indirect effect.

Analysis among controls indicated that the TT genotype had a positive, albeit imprecise, excess risk for being heavy smokers and smoked 6 cigarettes more per day on average relative to AT/AA genotype (17±9 cigarettes per day vs 23±15 cigarettes per day) (Supplementary material, etable 2).This is consistent with other studies conducted among North American Caucasian smokers including Canadians (7, 38) and is based on the mechanism that, relative to AT/AA genotype, the TT genotype metabolizes nicotine faster, increasing the need for smoking more cigarettes to maintain optimal nicotine levels in plasma. Based on this mechanism, a small proportion of the risk for SCCHN due to TT genotype being mediated by heavy smoking (PM = 21%) is a possibility. As explained in other clinical and biological settings (39), this overall proportion mediated could comprise of two distinct components; a) the increase in risk for SCCHN due to the total pool of carcinogens supplied by the excess number of cigarettes smoked per day as a consequence of CYP2A6 enzyme activity among TT genotype i.e., pure mediation (PIE), and b) the increase in risk for SCCHN due higher levels of carcinogenic products resulting from metabolism of pro-carcinogenic substrates specific to CYP2A6 enzyme, among the total pool of procarcinogens/carcinogens supplied by the excess number of cigarettes smoked per day due to TT genotype, i.e., both interaction and mediation (INTmed). The overall PM in our study was half attributable to PIE and half attributable to INTmed.

The majority of the effect of TT genotype seemed to be through a pure direct effect (NDE) which is a combination of controlled direct effect (CDE) and reference interaction (INTref) components. The CYP2A6 enzyme, expressed in both hepatic and extra hepatic tissues (e.g., upper aerodigestive tract), is mainly associated with the activation of tobacco specific nitrosamines and other nitrosamine procarcinogens to carcinogens (37). This gives rise to the possibility of interaction between CYP2A6 variants such as CYP2A6*2 and smoking which is a major source of these pro-carcinogenic substrates. Relative to AT/AA genotype, the rate of conversion of these pro-carcinogens to carcinogens is faster for individuals with TT genotype. The strength of association between CYP2A6 variants and cancers is stronger among heavy smokers, and a lack of association among non-smokers has been reported (36, 40). These suggest potential interaction between non-carriers of the variant and heavy smoking (that need not be induced by the CYP2A6 variant itself, i.e., INTref) Out of about 79% of excess risk due to NDE, 14% was attributed to INTref; the effect of TT genotype operating only in the presence of heavy smoking (not induced by the variant). Interaction results of our study may lend support to this finding (Supplementary material, etable 1); we estimated the overall proportion attributable to interaction (INTref + INTmed) at 24%.

Research on biologic mechanisms majorly support the view that the effect of TT genotype on SCCHN may involve mediation or interaction pathways with smoking intensity. However, approximately 65% of the excess risk for SCCHN due to TT genotype seemed to be attributed to CDE, i.e., TT has an effect on SCCHN risk even without the presence of heavy smoking, and without changing smoking intensity. However, given the wider confidence intervals and lack of studies exploring other mechanistic pathways through which CYP2A6 variants could lead to SCCHN risk (e.g., interaction/mediation with other sources of nitrosamines such as diet, environmental pollutants, gene-gene interactions), it may be speculative to interpret these results further.

ADH1B, alcohol and SCCHN

There is strong evidence for the impact of ADH1B*2 on intensity of alcohol consumed, as well as SCCHN risk among alcohol consumers. This is based on its involvement in the metabolism of ethanol to acetaldehyde (8). Alcohol consumers of Caucasian descent with GG genotype have been associated with 1.5 to 2 times the risk of being heavy alcohol consumers (including increased frequency) relative to AG/AA genotype (9-11). Among controls in our study, GG and AG/AA genotypes consumed 51 and 40 ml ethanol per day on average respectively (Supplementary material, etable 2). Although small, we documented an indirect effect between ADH1B*2 and

SCCHN risk (constituting about 10% of the total excess risk) supporting the mediation hypothesis. Half of this indirect excess risk was attributable to excess amount of ethanol consumed due to GG genotype (PIE), and half due to interaction between GG genotype and excess amount of ethanol consumed due to GG genotype (INTmed=5%). In our study, and as documented by a pooled study among Australian twins (9), the average difference in frequency of ethanol consumed between GG genotype and AG/AA genotype is approximately 10 ml per day. Ethanol at this small dose may be of limited biologic relevance for SCCHN and could probably explain the relatively small magnitude of the indirect effect.

The majority of the total effect of GG genotype on SCCHN risk seems to be direct, as supported by both INTref and CDE proportions. Acetaldehyde, the primary metabolite of ethanol, is a human carcinogen (41-43). Contrary to what is expected among alcohol consumers with AG/AA genotype in whom acetaldehyde builds up promptly, studies in various ethnicities show higher risk for SCCHN among GG genotype carriers. It is hypothesised that AG/AA genotype exhibit alternative mechanisms to clear off acetaldehyde peak formed following alcohol ingestion (3). On the contrary, lower acetaldehyde peaks in GG genotype results in excess alcohol consumption. This may result in higher local exposure to ethanol in the head and neck region which is acted up on my ADH1B enzyme present in this region as well as oral microflora, leading to slower but increased build-up of acetaldehyde (4, 33, 44, 45). These findings suggest possibility for strong interaction between GG genotype and heavy alcohol consumption for SCCHN risk. In our results, only 11% of the total effect was attributed to interaction of which only about half was due to interaction with higher alcohol intensity not due to GG genotype (INTref=6%). Although, secondary analysis was suggestive of positive additive interaction, confidence intervals were wide (Supplementary material, etable 2).

Out of 90% of excess risk on SCCHN due to GG genotype through direct effect, more than 80% seem to be independent of heavy alcohol consumption, i.e., controlled direct effect. Ethanol being the only documented substrate for ADH1B enzyme, and sample size limitation, prevents interpretation of this result. It is to be noted, however, that we used ethanol per day as the measure of mediator, and it is possible that other aspects of alcohol consumption behaviour such as overall cumulative alcohol consumption and alcohol dependence (with which ADH1B*2 is also strongly associated) could potentially be important mediators of GG genotype on SCCHN risk.

A few strengths of this study are worth mentioning. Our work is one of the first to apply the 4-way decomposition causal analytical technique in a case-control setting, which provides the maximum insight into the interrelationship between a single exposure and mediator, and their effect on an outcome (26). This analytical strategy is based on the counterfactual framework, which allows for mediation and 4-way decomposition analysis in the presence of exposure-mediator interaction, and within a case-control design. This is an advantage over other methods proposed in the literature (46, 47). Although the possibility of indirect effect of variants in CYP2A6 gene on SCCHN risk through smoking behaviour has been hypothesised, they have not been previously explored. Also, it has been proposed that any effect of the ADH1B*2 variant on SCCHN risk is likely due to interaction alone. Our study encompasses these possibilities in the relationship between these genetic variants and SCCHN risk. Furthermore, based on our point estimates, the potential for the existence of four pathways which may or may not include interaction, mediation or both with associated heavy intensity risk behaviours may not be ruled out.

Conclusion

Most of the effect of both the genetic variants investigated here on SCCHN risk seems to be through pathways that does not involve their associated risk behaviours. The 4-way decomposition approach not only has the potential for deciphering mechanistically relevant pathways for rare disease outcomes, but also quantify measures of policy relevance (14). For example, the majority of the Caucasian population carries the TT (CYP2A6*2) and GG (ADH1B*2) genotypes. Direct modification of these variants to reduce their effect may not be possible nor economical. However, their effect on SCCHN can be modified by changing the level of modifiable risk behaviours such as smoking and alcohol consumption. For example, if the effect estimates in this study were indeed valid, approximately 37% of the effect of TT genotype on SCCHN risk could be eliminated, if the level of smoking was brought down to that of moderate smokers (i.e., under a packet of cigarettes per day). And among GG carriers, about 16% risk could be eliminated if the alcohol consumption level was brought down to that of mild drinkers (<25ml ethanol per day) (48). The overall proportion eliminated is higher than proportion mediated because of the additional risk for SCCHN due to interaction between the exposure and the mediator (31). Future studies with large sample size should explore these findings further.

Table 1: Sample characteristics among cases and controls, HeNCe Life study, Montreal, Canada, 2005-2013

	Smokers, n=601				**Alcohol consumers, n=675**				
	Controls n=299		**Cases n=302**		**Controls n=351**		**Cases n=324**		**RR (95% CI)**
	n (%)	**Mean (+- SD)**	**n (%)**	**Mean (+- SD)**	**n (%)**	**Mean (+- SD)**	**n (%)**	**Mean (+- SD)**	
Age, years		61 (10)	61 (9)			61 (10)	61 (10)		
Sex									
Female	81 (27)		61 (20)		95 (27)		66 (20)		
Male	218 (73)		241 (80)		256 (73)		258 (80)		
Education, years		14 (4)		12 (4)		14 (4)			
CYP2A6 (minor allele frequency, A allele)	3								
AT/AA	19 (6)		16 (5)						1
TT	280 (94)		286 (95)						1.28 (0.46, 3.59)[a]
ADH1B (minor allele frequency, A allele)					5				
AG/AA					30 (9)		14 (4)		1
GG					321 (91)		310 (96)		2.37 (1.12, 4.25)[b]
Tobacco smoking (cigarette equivalence)									
Frequency (average cig/day)									
Moderate smoker (> 0 to 18 cig/day)	116 (39)		66 (22)						1
Heavy smoker (> 18 cig/day)	183 (61)		236 (78)						1.84 (1.21, 2.79)[c]
Duration in years		28 (15)		35 (14)					
Years since cessation		16 (15)		10 (13)					
Pack-years						27 (41)		41 (41)	
Alcohol use (ethanol) (average ml/day)		45 (120)		57 (108)					
Mild drinkers (> 0 to 25 ml per day)					230 (66)		146 (45)		1
Heavy drinkers (> 25 ml per day)					121 (34)		178 (55)		1.68 (1.13, 2.48)[d]
Duration in years						35 (14)		35 (13)	
Years since stopping use						3 (8)		3 (8)	

Table 1: Continued

HPV risk				
HPV negative	254 (85)	176 (58)	398 (85)	183 (57)
HPV other	27 (9)	31 (10)	31 (9)	31 (9)
HPV alpha-9 other than HPV16	12 (4)	20 (7)	12 (3)	22 (7)
HPV-16	6 (2)	74 (25)	10 (3)	87 (27)

RR: Risk ratio; CI: Confidence interval; SD: Standard deviation; HPV: Human papillomavirus
a Adjusted for age, sex, years of education (continuous)
b Adjusted for age, sex, years of education (continuous)
c Adjusted for age, sex, years of education (continuous), HPV risk (categorical), duration of smoking (continuous), years since cessation of smoking (mean centred, current users recoded as 0), indicator for ex-smoker, RC spline of ethanol frequency, CYP2A6*2
d Adjusted for age, sex, years of education (continuous), HPV risk (categorical), duration of drinking (continuous), years since stopping use of alcohol (mean centred, current users recoded as 0), indicator for ex-drinker, RC spline of pack-years of tobacco, ADH1B*2
HPV other: 6, 11, 18, 26, 34, 39, 40, 42, 44, 45, 51, 53, 54, 56, 59, 61, 62, 66, 68, 69, 70, 71, 72, 73, 81, 82, 83, 84, and 89
HPV alpha-9 other than HPV 16: 31,33,35,52,58 and 67

Table 2: Effect of SNPs on risk behaviours (exposure-mediator association models fitted among controls)

	Effect of CYP2A6*2 on heavy smoking among Caucasian smokers control participants, n=299		
SNP	**Moderate smokers (> 0 to 18 cig/day) n=116**	**Heavy smoker (> 18 cig/day) n=183**	**RR[a] (95% CI)**
CYP2A6			
AT/AA	10 (9)	9 (5)	1
TT	106 (91)	174 (95)	1.67 (0.63, 4.35)[a]

Effect of ADH1B*2 on heavy drinking among Caucasian alcohol consumers control participants, n=351			
	Mild drinkers (> 0 to 25 ml per day) n=230	**Heavy drinkers (> 25 ml per day) n=121**	**RR[a] (95% CI)**
ADH1B			
AG/AA	23 (10)	7 (6)	1
GG	207 (90)	114 (94)	2.01 (0.84, 5.15)[b]

RR: Relative risk; CI: Confidence interval; SD: Standard deviation;
a Adjusted for age, sex, years of education (continuous)
b Adjusted for age, sex, years of education (continuous)

Table 3: Total, direct and indirect effects, as well as 4-way decomposition of total effect between SNPs, risk behaviours and head and neck cancers, HeNCe Life study, Montreal, Canada, 2005-2013

	Smokers, n=601		**Alcohol consumers, n=675**	
	CYP2A6*2 (TT vs AT/AA) and heavy smoking		**ADH1B*2 (GG vs AG/AA) and heavy alcohol consumption**	
	RR (95% CI)		RR (95% CI)	
Total effect (TE)	1.28 (0.46, 3.59)		2.37 (1.12, 4.25)	
Excess relative risk (coeff)	0.28 (-1.8, 2.37)		1.37 (1.62, 4.38)	
	2 - Way decomposition of total effect			
Direct effect of variant (NDE)[a]	1.22 (0.45, 3.33)		2.24 (0.88, 5.71)	
Indirect effect through risk behaviour (NIE)[b]	1.05 (0.94, 1.17)		1.06 (0.97, 1.16)	
	4- Way decomposition of excess relative risk			
Component	**Coeff. (95% CI)**	**Proportions**	**Coeff. (95% CI)**	**Proportions**
CDE	0.18 (-1.79, 2.16)	65 %	1.15 (-1.57, 3.87)	84%
INTref	0.04 (-0.82, 0.90)	14 %	0.09 (-0.95, 1.13)	6%
INTmed	0.03 (-0.33, 0.38)	10 %	0.07 (-0.59, 0.73)	5%
PIE	0.03 (-0.43, 0.49)	11 %	0.06 (-0.63,0.77)	5%
Total excess risk	0.28 (-1.80,2.37)	100%	1.37 (-1.62, 4.38)	100%
	Overall proportion attributable to interaction = 24%		Overall proportion attributable to interaction = 11%	
	Overall proportion attributable to meditation = 21%		Overall proportion attributable to meditation = 10%	
	Overall proportion eliminated = 35%		Overall proportion eliminated = 16%	

RR: Odds ratio; CI: Confidence interval; Coeff: Regression coefficient, not exponential.

Outcome and mediator models for CYP2A6-smoking-cancer analysis among smokers adjusted for age, sex, years of education (continuous), HPV risk (categorical), duration of smoking (continuous), years since cessation of smoking (mean centred, current users recoded as 0), indicator for ex-smoker, RC spline of ethanol frequency

Outcome and mediator models for ADH1B-alcohol-cancer analysis among alcohol users adjusted for age, sex, years of education (continuous), HPV risk (categorical), duration of drinking (continuous), years since stopping use of alcohol (mean centred, current users recoded as 0), indicator for ex-drinker, RC spline of pack-years of tobacco

a NDE= Natural Direct Effect, is also referred as pure direct effect, of just direct effect in the literature; b NIE= Natural indirect effect, is also referred as total indirect effect or just indirect effect in the literature. On the ratio scale, the product of NDE and NIE = TE

Note: Please refer to the methods section of the manuscript for definitions of CDE, INTref, INTmed and PIE.

7.1 Supplemental material: Manuscript III

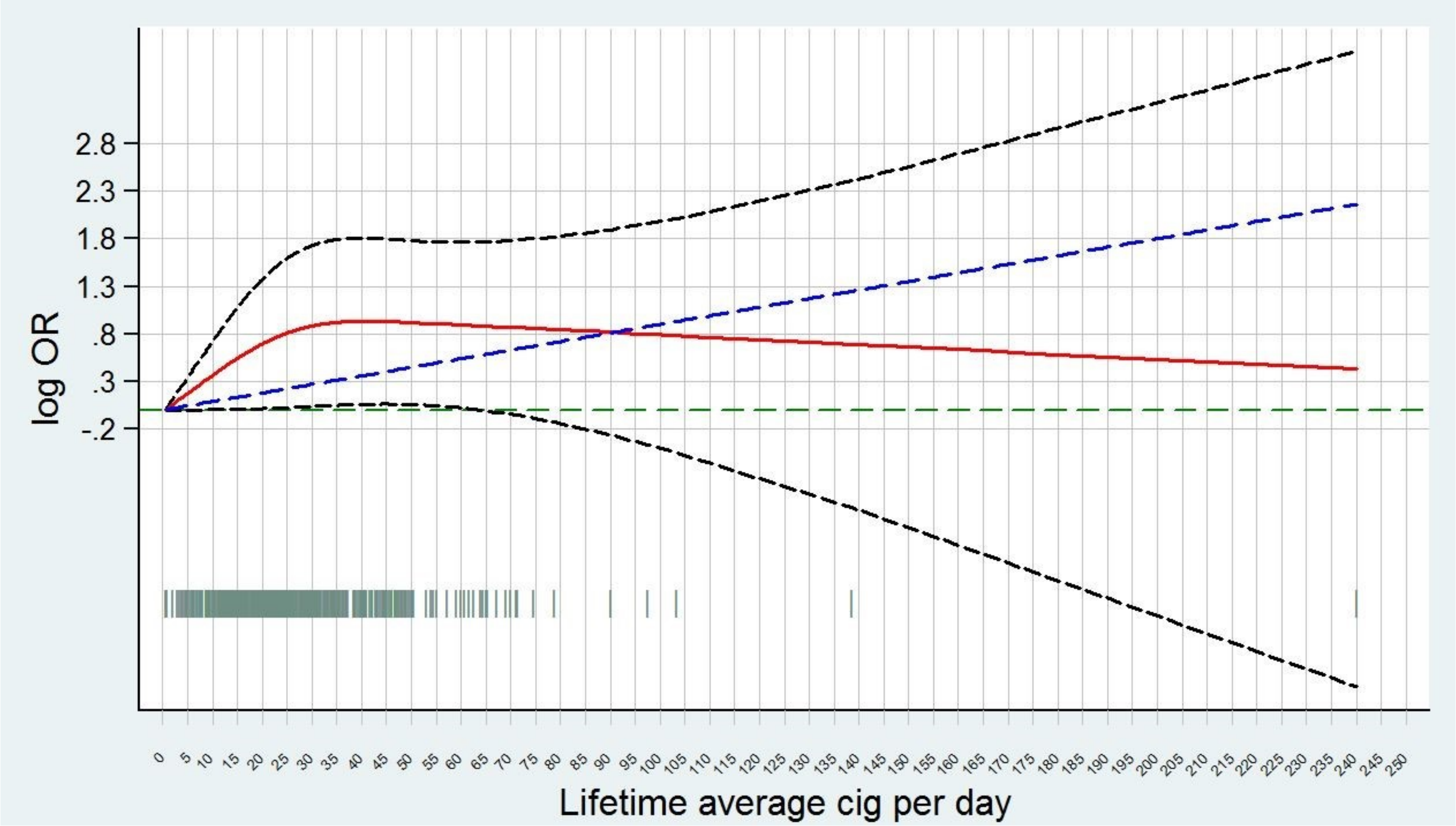
The effect of Interdependencies between CYP2A6 variant and smoking, and ADH1B variant and alcohol, on the risk of head and neck cancers

Supplemental material

Akhil Soman ThekkePurakkal[1], Belinda Nicolau[1], Robert D Burk[2], Jay S Kaufman[3], Nicolas F Schlecht[4]

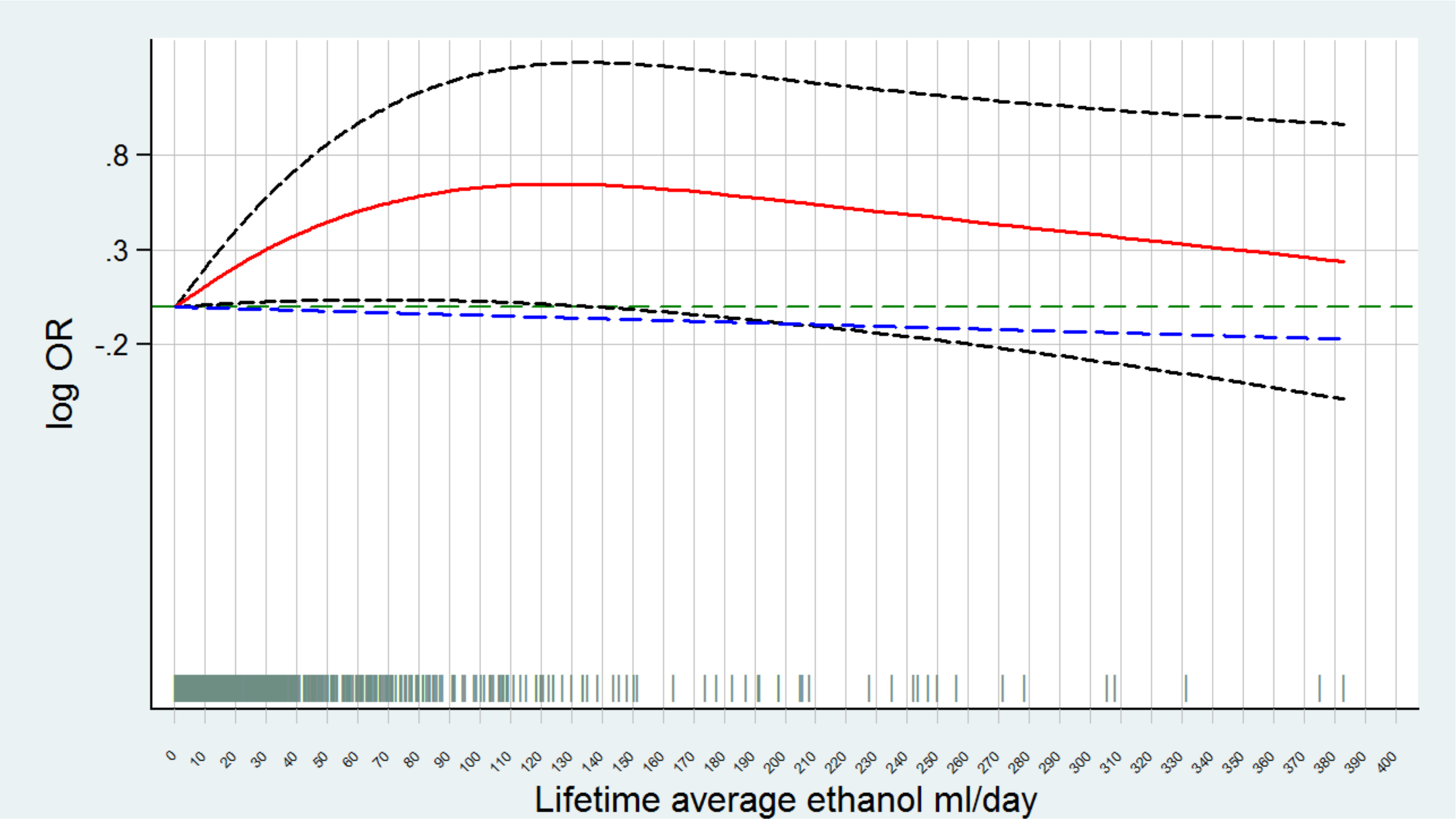
[1]Division of Oral Health and Society, Faculty of Dentistry, McGill University, Montreal, Canada; [2]Departments of Pediatrics (Genetics), Microbiology & Immunology, Obstetrics Gynecology & Women's Health, and Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA Departments of Oncology and Epidemiology & Biostatistics, McGill University, Montreal, Quebec, Canada; [3]Department of Epidemiology, Biostatistics, & Occupational Health, McGill University, Montreal, Canada;[4]Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA

efigure1: Restricted cubic spline graph for the association between frequency of smoking and head and neck cancers, HeNCe Canada, 2005-2013, n=601



The solid red line represents the estimates from restricted cubic spline and black dashed lines represent associated 95% confidence intervals. Blue dash line represents the association between cig/day and SCCHN when assuming linear functional form of smoking. The rug plot over x-axis represents the distribution of cig/day among participants (smokers). Model was conditioned on age, sex, years of education (continuous), HPV risk (categorical), duration of smoking (continuous), years since cessation of smoking (mean centered, current users recoded as 0), indicator for ex-smoker, RC spline of ethanol frequency, CYP2A6*2.

efigure2: Restricted cubic spline graph for the association between frequency of alcohol use and head and neck cancers, HeNCe Canada, 2005-2013, n=675



The solid red line represents the estimates from restricted cubic spline and black dashed lines represent associated 95% confidence intervals. Blue dash line represents the association between ethanol frequency and SCCHN when assuming linear functional form of smoking. The rug plot over x-axis represents the distribution of ethanol ml/day among participants (alcohol users). Model was conditioned on age, sex, years of education (continuous), HPV risk (categorical), duration of drinking (continuous), years since stopping use of alcohol (mean centered, current users recoded as 0), indicator for ex-drinker, RC spline of pack-years of tobacco, ADH1B*2.

eTable 1: Joint effects of CYP2A6*2 and smoking intensities among smokers, ADH1B*2 and alcohol consumption intensities among alcohol users on head and neck cancer risk, stratum specific effects, and interaction on multiplicative and additive scales

SNP	N Co/Ca	Low smoker <=18 cig/day	N Co/Ca	Heavy smoker >18 cig/day
CYP2A6*2				
AT/AA	10/5	1	9/11	1.55 (0.33, 7.25)[a]
TT	106/61	1.19 (0.34, 4.17)[a]	174/225	2.23 (0.66, 7.57)[a]
Interaction	Multiplicative scale			1.21 (0.24, 5.98)[a]
	Additive scale (RERI)			0.49 (-1.27,2.25)[a]
	N Co/Ca	**Mild drinker <=25 ml/day**	**N Co/Ca**	**Heavy drinker >25ml/day**
ADH1B*2				
AG/AA	23/9	1	7/5	1.68 (0.32, 8.82)[b]
GG	207/137	2.25 (0.87, 5.83) [b]	114/173	3.65 (1.41, 9.47)[b]
Interaction	Multiplicative scale			0.96 (0.17, 5.17)[b]
	Additive scale (RERI)			0.72 (-1.90, 3.34)[b]

[a] Adjusted age, sex, years of education (continuous), HPV risk (categorical), duration of smoking (continuous), years since cessation of smoking (mean centered, current users recoded as 0), indicator for ex-smoker, RC spline of ethanol frequency

[b] Adjusted for age, sex, years of education (continuous), HPV risk (categorical), duration of drinking (continuous), years since stopping use of alcohol (mean centered, current users recoded as 0), indicator for ex-drinker, RC spline of pack-years of tobacco

eTable 2: Distribution of risk behaviours among SNPs

	CYP2A6*2 among smokers, controls, n=299	
	AT/AA	**TT**
	Mean (+- SD)	Mean (+- SD)
Average cigarettes smoked per day	17 (9)	23 (15)
	ADH1B*2 among alcohol consumers, controls, n=351	
	AG/AA	**GG**
Alcohol use (ethanol) (average ml/day)	40 (154)	51 (105)

eAppendix

Stata codes for mediation and 4-way decomposition analysis under counterfactual causal framework for case-control study design

Thekke Purakkal, Akhil Soman[1], Kaufman, Jay S[2]

[1]Division of Oral Health and Society, Faculty of Dentistry, McGill University, Montreal Quebec, [2]Department of Epidemiology, Biostatistics and Occupation Health, McGill University, Montreal, Quebec

Date: August 1, 2016; Version 1

Stata macros (e.g., PARAMED) for estimating mediation effects in the presence of exposure mediator interaction under counterfactual causal framework already exist (1). For conducting 4-way decomposition analysis, although the mathematical equations and SAS codes have been provided by VanderWeele 2014 (2), Stata codes have not been written. The below Stata codes were exclusively written for conducting mediation analysis under exposure-mediator interaction (alternative method using mathematical equations), as well as 4-way decomposition analysis for this thesis work. Although the codes given here is specific for binary outcome, mediator and exposure variables, the code can be easily extended to the case where the exposure, mediator and outcome are continuous, categorical, binary or their combinations. The codes have been written using the mathematical formulas for total effect, mediation effects, and 4-way decomposition effects, as well as for calculating various proportions provided by VanderWeele 2015 and 2016 (3, 4). The user is encouraged to cross check these codes with the formulas in these references. The codes for bootstrapping procedure given at the end of the codes can be used to derive the confidence limits for these estimates.

Let Y be a binary outcome. In a case-control study, Y=1 may represent cases and Y=0 may represent controls or non-cases. Let A be a binary exposure, and M a binary mediator. Let C1, C2, be continues covariates, and C3, C4 be binary or categorical. If there are more or fewer covariates, one can add or remove scalars under "//Covariates", "//Assigning levels of covariates" and "//calculating bcc" in the below given code. For a case-control study with rare disease outcome, the line of code for mediator model may be fit only among controls. Alternatively, or, if the outcome is not rare, one can weight the mediator model using sampling weights as suggested by VanderWeele and Vansteelandt 2010 (5).

References

1. VanderWeele TJ. Mediation: Introduction and regresion -based approaches. Explanation in Causal inference: Methods for mediation and Interaction. USA: Oxford University Press; 2015. p. 40-1.
2. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. Epidemiology (Cambridge, Mass). 2014;25(5):749-61.
3. VanderWeele TJ. A unification of mediation and interaction. Explanation in Causal inference: Methods for mediation and interaction. USA: Oxford University Press; 2015. p. 371-96.
4. Erratum: A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology (Cambridge, Mass). 2016;27(5):e36.
5. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American journal of epidemiology. 2010;172(12):1339-48.

```
****** Run code from the line below till the end at a single stretch *******


**Start of code **

set varabbrev off, perm
cap prog drop calc2
prog calc2, rclass

logit Y A##M C1 C2 C3 C4 // Outcome model

scalar t1=_b[1.A]
scalar t2=_b[1.M]
scalar t3=_b[1.A#1.M]

logit M A C1 C2 C3 C4 if Y==0 // Mediator model fit among among controls

scalar b0=_b[_cons]
scalar b1=_b[A]

//Covariates

scalar bc1 = _b[C1]
scalar bc2 = _b[C2]
scalar bc3 = _b[1.C3]
scalar bc4 = _b[1.C4]

// Assigning level of covariates // for continuous covariates, just take the mean of
// their distribution in full sample

sum C1
scalar cc1= r(mean)
sum C2
scalar cc2= r(mean)
scalar cc3=1 // at level 1 of binary covariate C3. Any level can be assigned based on
// requirement
scalar cc4=3 // at level 3 of 3 category covariate C4



// Calculating bcc - Calculating sum of products of coefficients of covariates from
// mediator model and level of covariate

scalar bcc = bc1*cc1 + bc2*cc2 + bc3*cc3+ bc4*cc4

// Additional values assigned

scalar a1=1 // level 1 of exposure A
scalar a0=0 // level 0 of exposure A
scalar m0=0 // level 0 of binary mediator
scalar mstar=0 // level of mediator at which CDE is calculated


// 2-way decomposition or mediation – Calculating coefficients for natural direct
effect (NDE), natural indirect effect (NIE) and total effect (total)

scalar lnde    = ln((exp(t1*a1)*(1+exp(t2+t3*a1+b0+b1*a0+bcc))) ///
                  /(exp(t1*a0)*(1+exp(t2+t3*a0+b0+b1*a0+bcc))))

scalar lnie    = ln(((1+exp(b0+b1*a0+bcc))*(1+exp(t2+t3*a1+b0+b1*a1+bcc))) ///
                  /((1+exp(b0+b1*a1+bcc))* (1+exp(t2+t3*a1+b0+b1*a0+bcc))))

scalar ltotal  = ln((exp(t1*a1)*(1+exp(b0+b1*a0+bcc))* ///
                 (1+exp(b0+b1*a1+bcc+t2+t3*a1)))/(exp(t1*a0)* ///
                 (1+exp(b0+b1*a1+bcc))* (1+exp(b0+b1*a0+bcc+t2+t3*a0))))

// 4- way decomposition - Calculating coefficients for controlled direct effect (CDE),
// reference interaction(INTref), mediated interaction(INTmed), pure indirect effect
// (PIE)

scalar lcde    = ln(exp(t1 + t3*mstar)*(a1-a0))

scalar lIntref = ln((exp(t1*(a1-a0)-t2*mstar-t3*a0*mstar)* ///
                 (1+exp(b0+b1*a0+bcc+t2+t3*a1))) /(1+exp(b0+b1*a0+bcc)) ///
                 -(exp(-t2*mstar-t3*a0*mstar)*(1+exp(b0+b1*a0+bcc+t2+t3*a0))) ///
                 /(1+exp(b0+b1*a0+bcc)) - exp((t1+t3*mstar)*(a1-a0)) + 1)

scalar lIntmed = ln((exp(t1*(a1-a0)-t2*mstar-t3*a0*mstar)* ///
                 (1+exp(b0+b1*a1+bcc+t2+t3*a1)) /(1+exp(b0+b1*a1+bcc))) ///
                 - (exp(-t2*mstar-t3*a0*mstar)*(1+exp(b0+b1*a1+bcc+t2+t3*a0)) ///
                 /(1+exp(b0+b1*a1+bcc)))- exp(t1*(a1-a0)-t2*mstar-t3*a0*mstar) ///
                 *(1+exp(b0+b1*a0+bcc+t2+t3*a1))/(1+exp(b0+b1*a0+bcc)) ///
                 + exp(-t2*mstar-t3*a0*m0)*(1+exp(b0+b1*a0+bcc+t2+t3*a0)) ///
                  /(1+exp(b0+b1*a0+bcc)))

scalar lpie    = ln((1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a0)) ///
                 /((1 + exp(b0+b1*a1+bcc))*(1+exp(b0+b1*a0+bcc+t2+t3*a0))))

// Calculating coefficients for each 4-way component and total effect

scalar cde_comp    = (exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc))/ ///
                     (1+exp(b0+b1*a0+bcc+t2+t3*a0)))-(exp(t2*mstar+t3*a0*mstar)* ///
                     (1+exp(b0+b1*a0+bcc))/(1+exp(b0+b1*a0+bcc+t2+t3*a0)))
scalar INTref_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a0+bcc+t2+t3*a1)) ///
                     /(1+exp(b0+b1*a0+bcc+t2+t3*a0)) - (1) ///
                     -exp(t1*(a1-a0)+t2*mstar+t3*a1*mstar)*(1+exp(b0+b1*a0+bcc)) ///
                     /(1+exp(b0+b1*a0+bcc+t2+t3*a0))+ exp(t2*mstar+t3*a0*mstar)* ///
                     (1+exp(b0+b1*a0+bcc))/(1+exp(b0+b1*a0+bcc+t2+t3*a0))
scalar INTmed_comp = exp(t1*(a1-a0))*(1+exp(b0+b1*a1+bcc+t2+t3*a1))* ///
                     (1+exp(b0+b1*a0+bcc))/((1+exp(b0+b1*a0+bcc+t2+t3*a0)) ///
                     *(1+exp(b0+b1*a1+bcc)))-(1+exp(b0+b1*a1+bcc+t2+t3*a0))* ///
                     (1+exp(b0+b1*a0+bcc)) /((1+exp(b0+b1*a0+bcc+t2+t3*a0))* ///
                     (1+exp(b0+b1*a1+bcc))) - exp(t1*(a1-a0))* ///
                     (1+exp(b0+b1*a0+bcc+t2+t3*a1)) ///
                      /(1+exp(b0+b1*a0+bcc+t2+t3*a0)) + (1)
scalar pie_comp    = (1+exp(b0+b1*a0+bcc))*(1+exp(b0+b1*a1+bcc+t2+t3*a0)) ///
                      /((1 + exp(b0+b1*a1+bcc))*(1+exp(b0+b1*a0+bcc+t2+t3*a0))) -(1)
scalar total       = (exp(t1*a1)*(1+exp(b0+b1*a0+bcc))* ///
                     (1+exp(b0+b1*a1+bcc+t2+t3*a1))) /(exp(t1*a0)* ///
                     (1+exp(b0+b1*a1+bcc))* (1+exp(b0+b1*a0+bcc+t2+t3*a0)))


// Retrieving the values of each coefficient calculated above

return scalar lnie=lnie
return scalar lnde=lnde
return scalar ltotal=ltotal

return scalar lcde=lcde
return scalar lIntref=lIntref
return scalar lIntmed=lIntmed
return scalar lpie=lpie

return scalar cde_comp = cde_comp
return scalar INTref_comp = INTref_comp
return scalar INTmed_comp = INTmed_comp
return scalar pie_comp  = pie_comp
return scalar total=total
// Calculating value for total excess relative risk (terr)
scalar terr   = cde_comp +INTref_comp + INTmed_comp + pie_comp

// Calculating the values for each of the 4 components of the total excess risk
scalar errCDE                   = cde_comp*(total-1)/terr
scalar errINTref                = INTref_comp*(total-1)/terr
scalar errINTmed                = INTmed_comp*(total-1)/terr
scalar errPIE                   = pie_comp*(total-1)/terr
// Assigning the values for proportions of total excess risk that is due to each of
// the component

scalar PropCDE                  = cde_comp/terr
scalar PropINTref               = INTref_comp/terr
scalar PropINTmed               = INTmed_comp/terr
scalar PropPIE                  = pie_comp/terr

// Assigning the values of overall proportions of risk attributable to mediation,
// interaction, and proportion eliminated
scalar PropMediated             = (pie_comp+INTmed_comp)/terr
scalar PropAttribInteraction    = (INTref_comp+INTmed_comp)/terr
scalar PropEliminated           = (INTref_comp+INTmed_comp+pie_comp)/terr

// Retrieving the values for each of the above

return scalar terr              = terr
return scalar errCDE            = errCDE
return scalar errINTref         = errINTref
return scalar errINTmed         = errINTmed
return scalar errPIE            = pie_comp
return scalar PropCDE           = PropCDE
return scalar PropINTref        = PropINTref
return scalar PropINTmed        = PropINTmed
return scalar PropPIE           = PropPIE
return scalar PropMediated                    = PropMediated
return scalar PropAttribInteraction = PropAttribInteraction
return scalar PropEliminated        = PropEliminated
end
calc2
return list

****End of code ***

***Run the code from start till the line above****

// Running the above code will give an output of estimates (coefficients) only, of all
// parameters
```

Code for calculating confidence intervals and exponentiated risk estimates

Stata codes for - bootstrap - procedure to calculate the 95% CIs for estimate of each parameter. Any number of repetition (reps) can be assigned. Random seed number (seed) should be assigned.

```
*Code:

bootstrap lcde=r(lcde) lIntref=r(lIntref) lIntmed = r(lIntmed)  lpie=r(lpie) ///
    ltotal=r(ltotal)  lnde=r(lnde) lnie=r(lnie), reps(2000) seed(438766) nodrop: calc2

// Running the above command will create an output of results with estimates of
// observed coefficients, Bootstrap standard error, z, P>|z| and 95% CI in a table of
// rows and 6 columns

// Steps to exponentiate the values of each observed coefficient and corresponding CIs
// in the results table

matrix T= r(table) // captures the real matrix returned by -bootstrap-
matrix list T

// Calculating the exponentiated results

display "CDE=" exp(T[1,1]), "LB=" exp(T[5,1]), "UB=" exp(T[6,1])
display "INTref=" exp(T[1,2]), "LB=" exp(T[5,2]), "UB=" exp(T[6,2])
display "INTmed=" exp(T[1,3]), "LB=" exp(T[5,3]), "UB=" exp(T[6,3])
display "PIE="exp(T[1,4]), "LB=" exp(T[5,4]), "UB=" exp(T[6,4])
display "TE=" exp(T[1,5]), "LB=" exp(T[5,5]), "UB=" exp(T[6,5])
display "NDE=" exp(T[1,6]), "LB=" exp(T[5,6]), "UB=" exp(T[6,6])
display "NIE=" exp(T[1,7]), "LB=" exp(T[5,7]), "UB=" exp(T[6,7])


// Calculating the CIs using bootstrap for 4- way components, 4 components of excess

// relative risks, and proportions

bootstrap cde_comp = r(cde_comp) INTref_comp = r(INTref_comp) ///
          INTmed_comp = r(INTmed_comp) pie_comp = r(pie_comp) ///
          terr = r(terr) errCDE = r(errCDE) errINTref = r(errINTref) ///
          errINTmed = r(errINTmed) errPIE = r(errPIE) PropCDE = r(PropCDE) ///
          PropINTref = r(PropINTref) PropINTmed = r(PropINTmed) ///
          PropPIE = r(PropPIE) PropMediated = r(PropMediated) ///
          PropAttribInteraction = r(PropAttribInteraction) ///
          PropEliminated =r(PropEliminated), reps(2000) seed(438766) ///
          saving(`boot_results') nodrop: calc2
```

eTable 3: Association between genetic variants with smoking and alcohol outcomes among controls and cases (Manuscript III) using logistic regression model

Genetic variant	**Smoking frequency (0= moderate, 1=heavy) OR (95% CI)**		
	Among controls	**Among Cases**	
CYP2A6*2 (TT genotype)	1.64 (0.62, 4.34)	1.57 (0.48, 5.14)	**Inference**= no evidence for bias due to potential differential misclassification of exposures
	Alcohol frequency (0= moderate, 1=heavy) OR (95% CI)		
ADH1B*2 (GG genotype)	2.05 (0.81, 5.15)	2.97 (0.86, 10.29)	

References

1. Hashibe M, Brennan P, Chuang SC, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiol Biomarkers Prev. 2009;18(2):541-50.
2. Ruwali M, Pant MC, Shah PP, Mishra BN, Parmar D. Polymorphism in cytochrome P450 2A6 and glutathione S-transferase P1 modifies head and neck cancer risk and treatment outcome. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2009;669(1–2):36-41.
3. Hashibe M, Boffetta P, Zaridze D, Shangina O, Szeszenia-Dabrowska N, Mates D, et al. Evidence for an important role of alcohol- and aldehyde-metabolizing genes in cancers of the upper aerodigestive tract. Cancer Epidemiol Biomarkers Prev. 2006;15(4):696-703.
4. Hashibe M, McKay JD, Curado MP, Oliveira JC, Koifman S, Koifman R, et al. Multiple ADH genes are associated with upper aerodigestive cancers. Nat Genet. 2008;40(6):707-9.
5. Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking. Nature. 1998;393(6687):750.
6. Tyndale RF, Sellers EM. Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. Drug metabolism and disposition: the biological fate of chemicals. 2001;29(4 Pt 2):548-52.
7. Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, Sellers EM, et al. Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol Pharmacol. 2000;58(4):747-55.
8. Brennan P, Lewis S, Hashibe M, Bell DA, Boffetta P, Bouchardy C, et al. Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a HuGE review. Am J Epidemiol. 2004;159(1):1-16.
9. Macgregor S, Lind PA, Bucholz KK, Hansell NK, Madden PA, Richter MM, et al. Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. Human molecular genetics. 2009;18(3):580-93.
10. Li D, Zhao H, Gelernter J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. Biol Psychiatry. 2011;70(6):504-12.
11. Bierut LJ, Goate AM, Breslau N, Johnson EO, Bertelsen S, Fox L, et al. ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. Molecular psychiatry. 2012;17(4):445-50.
12. Guo H, Zhang G, Mai R. Alcohol Dehydrogenase-1B Arg47His Polymorphism and Upper Aerodigestive Tract Cancer Risk: A Meta-Analysis Including 24,252 Subjects. Alcoholism: Clinical and Experimental Research. 2012;36(2):272-8.
13. Raunio H, Rautio A, Gullstén H, Pelkonen O. Polymorphisms of CYP2A6 and its practical consequences. British Journal of Clinical Pharmacology. 2001;52(4):357-63.
14. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. Epidemiology (Cambridge, Mass). 2014;25(5):749-61.
15. Berney L, Blane DB, Soc Sci M, Blane B. Collecting retrospective data: accuracy of recall after 50 years judged against historical records. Social Science & Medicine. 1997;45(10):1519-25.

16. Egan KM, Abruzzo J, Cytobrush B, Newcomb PA, Titus-ernstoff L, Franklin T, et al. Collection of Genomic DNA from Adults in Epidemiological Studies by Buccal Cytobrush and Mouthwash. 2001:687-96.

17. D'Souza G, Sugar E, Ruby W, Gravitt P, Gillison M. Analysis of the effect of DNA purification on detection of human papillomavirus in oral rinse samples by PCR. J Clin Microbiol. 2005;43(11):5526-35.

18. Laprise C, Madathil SA, Schlecht NF, Castonguay G, Soulières D, Nguyen-Tan PF, et al. Human papillomavirus genotypes and risk of head and neck cancers: Results from the HeNCe Life case-control study. Oral oncology. 2017;69:56-61.

19. Hakenewerth AM, Millikan RC, Rusyn I, Herring AH, North KE, Barnholtz-Sloan JS, et al. Joint effects of alcohol consumption and polymorphisms in alcohol and oxidative stress metabolism genes on risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev. 2011;20(11):2438-49.

20. Hoffmann D, Hoffmann I. Chemistry and toxicology.In: US Department of Health and Human Services. Cigars: health effects and trends (Smoking and Tobacco Control Monograph 9). DHHS (Publ No. NIH 98-4302), 1998:55–104.

21. Williams BA, Madrekar JN, Madrekar SJ, Cha SS, Furth AF. Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes Mayo Clinic, Rochester, Minnesota Division of Biostatistics DoHSR; June 2006. Contract No.: 10027230.

22. Schlecht NF, Franco EL, Pintos J, Kowalski LP. Effect of smoking cessation and tobacco type on the risk of cancers of the upper aero-digestive tract in Brazil. Epidemiology (Cambridge, Mass). 1999;10(4):412-8.

23. Brady, Henry. “Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory.” In Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, editors, The Oxford Handbook of Political Methodology New York: Oxford University Press, 2008.

24. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology (Cambridge, Mass). 1992;3(2):143-55.

25. Pearl J. Direct and indirect effects. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence; Seattle, Washington. 2074073: Morgan Kaufmann Publishers Inc.; 2001. p. 411-20.

26. VanderWeele TJ. A unification of mediation and interaction. Explanation in Causal inference: Methods for mediation and interaction. USA: Oxford University Press; 2015. p. 371-96.

27. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. Journal of the National Cancer Institute. 2000;92(14):1151-8.

28. Schlecht NF, Franco EL, Pintos J, Negassa A, Kowalski LP, Oliveira BV, et al. Interaction between Tobacco and Alcohol Consumption and the Risk of Cancers of the Upper Aero-Digestive Tract in Brazil. American Journal of Epidemiology. 1999;150(11):1129-37.

29. Leffondre K. Modeling Smoking History: A Comparison of Different Approaches. American Journal of Epidemiology. 2002;156(9):813-23.

30. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American journal of epidemiology. 2010;172(12):1339-48.

31. VanderWeele TJ. Explanation in Causal inference: Methods for mediation and Interaction. Press OU, editor. USA2015. 706 p.

32. Erratum: A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology (Cambridge, Mass). 2016;27(5):e36.

33. Asakage T, Yokoyama A, Haneda T, Yamazaki M, Muto M, Yokoyama T, et al. Genetic polymorphisms of alcohol and aldehyde dehydrogenases, and drinking, smoking and diet in Japanese men with oral and pharyngeal squamous cell carcinoma. Carcinogenesis. 2006;28(4):865-74.

34. Hashibe M, McKay JD, Curado MP, Oliveira JC, Koifman S, Koifman R, et al. Multiple ADH genes are associated with upper aerodigestive cancers.

35. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology (Cambridge, Mass). 2004;15(5):615-25.

36. Kamataki T, Fujieda M, Kiyotani K, Iwano S, Kunitoh H. Genetic polymorphism of CYP2A6 as one of the potential determinants of tobacco-related cancer risk. Biochemical and Biophysical Research Communications. 2005;338(1):306-10.

37. Rossini A, de Almeida Simao T, Albano RM, Pinto LF. CYP2A6 polymorphisms and risk for tobacco-related cancers. Pharmacogenomics. 2008;9(11):1737-52.

38. Schoedel Ka, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF. Ethnic variation in CYP2A6 and association of genetically slow nicotine metabolism and smoking in adult Caucasians. Pharmacogenetics. 2004;14(9):615-26.

39. Ikram MA, VanderWeele TJ. A proposed clinical and biological interpretation of mediated interaction. European Journal of Epidemiology. 2015;30(10):1115-8.

40. Canova C, Richiardi L, Merletti F, Pentenero M, Gervasio C, Tanturri G, et al. Alcohol, tobacco and genetic susceptibility in relation to cancers of the upper aerodigestive tract in northern Italy. Tumori. 2010;96(1):1-10.

41. Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, Bouvard V, et al. Carcinogenicity of alcoholic beverages. The Lancet Oncology. 2007;8(4):292-3.

42. Secretan B, Straif K, Baan R, Grosse Y, El Ghissassi F, Bouvard V, et al. A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. The Lancet Oncology. 2009;10(11):1033-4.

43. Salaspuro M. Acetaldehyde as a common denominator and cumulative carcinogen in digestive tract cancers. Scand J Gastroenterol. 2009;44(8):912-25.

44. Chang JS, Straif K, Guha N. The role of alcohol dehydrogenase genes in head and neck cancers: a systematic review and meta-analysis of ADH1B and ADH1C. Mutagenesis. 2012;27(3):275-86.

45. Homann N, Jousimies-Somer H, Jokelainen K, Heine R, Salaspuro M. High acetaldehyde levels in saliva after ethanol consumption: methodological aspects and pathogenetic implications. Carcinogenesis. 1997;18(9):1739-43.

46. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173-82.

47. Hayes AF. Beyond Baron and Kenny : Statistical Mediation Analysis in the New Millennium Beyond Baron and Kenny : Statistical Mediation Analysis in the New Millennium. 2009(July 2014):37-41.

48. VanderWeele TJ. Policy-relevant proportions for direct effects. Epidemiology (Cambridge, Mass). 2013;24(1):175-

Chapter 8
Discussion

The specific results related to each objective of this thesis have been discussed in individual manuscripts. This chapter will provide a brief overview of the rationale, results, and plausible explanations of the findings in each manuscript. Potential non-causal explanations of findings and steps taken to mitigate bias will be discussed in the limitations section. The overall strengths and contributions of the project, public health implications and future directions are also discussed.

8.1 Manuscript 1

The first manuscript in this dissertation explored the association between SEP over the life-course and oral cancer through the lens of multiple life-course models using data from a single case-control study among a sample of participants from Kerala, India.

Extant research shows that a cumulative disadvantageous SEP over the life-course is associated with increased risk for multiple chronic disease outcomes including oral cancers (33). This inverse association between cumulative SEP and oral cancers has been documented in both developed and developing countries. Thus, disadvantageous SEP can be considered as a component cause of oral cancer. However, SEP is a complex construct that varies over the life-course of individuals. This phenomenon, although well recognised, has been consistently overlooked by SEP-oral cancer studies. Ignoring the basic nature (static vs dynamic) of an exposure results in erroneous associational or causal estimates, leading to biased epidemiologic evidence and a flawed understanding of the development of health outcomes. Furthermore, critical questions that remain unanswered with respect to SEP exposure and oral cancer risk can be addressed only by appreciating the time-varying nature of the exposure. For example, ***"how"*** does a disadvantageous SEP associate with oral cancer risk, i.e., within the cumulative effect of the SEP exposure, are there specific time periods in an individual's life during which experiencing disadvantageous SEP may be etiologically relevant compared to other time periods? ***"Who"*** are the high-risk groups:

i.e., can oral cancer risk profile within a given population be characterised based on differential SEP exposure experienced by individuals (disadvantageous or advantageous) at multiple periods of their lives? A quantitative assessment of these questions has the potential to enrich the understanding of causal mechanisms underlying the association between SEP and oral cancer risk.

Manuscript I aimed to address these mechanistically relevant questions. Using data from the Indian site of the HeNCe Life case-control study, we estimated the association between SEP and oral cancer using accumulation, critical period and social mobility models specified under the life-course framework. We used asset index/wealth index, a measure of material endowment of the individual or household (similar to income, which is a direct indicator of social class) as the measure of SEP. The objectives of manuscript I were achieved by utilizing causal analytical techniques based on the counterfactual framework, to account for the time-varying nature of both SEP over three periods of life, and associated confounders over multiple periods with respect to oral cancer in adult life. These techniques involved adapting inverse probability weighted marginal structural models, originally developed for longitudinal data, to the case-control study design through novel approaches.

Our study confirms the association between exposure to cumulative disadvantageous SEP over the life-course and oral cancer incidence, with relative risks and gradients going in the well-known direction. We observed an inverse graded association between the accumulation of SEP across childhood, early and late adulthood periods and marginal odds for oral cancer risk (after conditioning on both dose and functional form of associated confounders- categorical, linear, non-linear). Within this cumulative risk association, exposure to disadvantageous SEP during early life periods [childhood (0-16 years) and early adulthood (17-30 years)] increased the risk for the disease. However, the magnitude of association was higher for the childhood critical period model than the early adulthood model (marginal OR of 2.75 vs 1.8). The critical nature of exposure during the childhood period was reflected in the social mobility (childhood to early adulthood) and the saturated all-trajectories model. Both models indicated that, compared to non-exposure to disadvantageous SEP in all periods considered in each model, the ORs associated with trajectories in which individuals were exposed to disadvantageous SEP in childhood were larger than those of trajectories where participants were never exposed in childhood. Furthermore, using QICw values, we concluded that the childhood critical period model fit our data best relative to other models. A recent smaller study of 180 oral cancer cases and 272 controls from another region of South India assessing mediation, reported that a disadvantageous socioeconomic condition (measured using occupation of head of the household) in childhood had a significant controlled direct effect on oral cancer risk that was not mediated through smoking, alcohol or chewing habits (316). Future studies should document how much of the effect of exposure to disadvantageous SEP in childhood on oral cancer would remain following mediation by this exposure in early and late adulthood periods and other potential causal mediators such as oral health indicators.

8.1.1 Biologic plausibility of our findings

The biologic plausibility of associations between socioeconomic factors and cancers may be embedded within the growing body of research related to psychosocial, neuro-immunological, genetic and epigenetic pathways (35). Experimental and observational studies in these fields identify *stress* as the central concept in understanding the direct biological embodiment of socioeconomic disadvantages leading to adverse health outcomes (445-448).

An asset or wealth index, used as a proxy measure of SEP in this work, is a stockpile of financial resources; a lack of wealth translates into the absence of a social safety net which, over time, generates chronic stress (273, 449). In general, cumulative disadvantageous SEP translates into different stressors leading to general anxiety in response to day-to-day events and challenges, low social capital and community cohesion, lack of social support, stressful work environments, adverse social and living conditions, job insecurities, unemployment, fear of crime, etc. (271, 450-452). The accumulation of stressors affects the hippocampal region of the brain, dysregulates the hypothalamic-pituitary-adrenal (HPA) and sympathetic-adrenal-medullary (SAM) axes, produces stress hormones such as glucocorticoids, epinephrine and norepinephrine, subsequently leading to the disruption of neuroendocrine, immune, cardiovascular and metabolic systems (333, 447, 453, 454). Chronic over-activity, failure to shut down or inadequate response of these regulatory physiological systems leads to an increase in ‘allostatic load’ (333, 454). Although complex, non-linear, dynamic and interactive, the pathways involving stress and allostatic regulatory systems resulting in increased allostatic load have been cited as one of the most compelling explanations (mediator) of how cumulative social adversity results in chronic diseases (35, 281, 331, 454). Indeed, measures of allostatic load (e.g., levels of cortisol, glucose, blood pressure throughout the day or in response to a challenge) are socially patterned; an accumulation of socioeconomic disadvantages is associated with a higher allostatic load (310, 333). But how could this psychosocial stress pathway mediate the carcinogenic process?

At the cellular level, genetic events such as the shortening of telomeres (regions of repetitive DNA fragments at the end of chromosomes that protect them from replication failure) and increased telomerase activity in various cells (e.g., leucocytes) in response to increased stress hormone secretion have been reported (35, 455, 456). These genetic changes have been associated with an increased risk of oral cancers as well (456-459).

Yet another hypothesis to explain how oral cancer development may be due to cumulative adverse social exposures and stress is through epigenetic modifications specific to the head and neck region (460). Epigenetics refers to the study of heritable changes in gene expression that occur without a change in DNA sequence. Epigenetic changes are specific to anatomic sites and are potentially reversible. A clear correlation between epigenetic-driven dysregulation of gene expression and SCCHN (e.g., oral cancers) progression is not fully demonstrated at present. However, epigenetic modifications (e.g., hypermethylation of gene promoters and consequent silencing of several tumour suppressor genes, hypomethylation resulting in the activation of oncogenes) leading to chromosomal instability, increased proliferation and growth advantage have been identified in oral cancers (461-463). Several studies have implicated SEP in the development of epigenetic patterns that may contribute to cancer. McGuinnes et al (2012) showed that individuals living in disadvantageous compared to advantageous social conditions had 17% lower DNA methylation levels, which was in turn associated with higher interleukin-6 and fibrinogen, both implicated in risk of cancers (464). Subramanyam et al (2013) documented that lower levels of wealth were associated with lower methylation levels in leucocyte DNA, although no association was identified with income or education (465). Genes involved in the inflammatory response are less methylated in individuals accumulating socioeconomic adversities over life in a dose-dependent fashion (466). These findings implicate epigenetic changes as a potential mediator of the association between cumulative socioeconomic disadvantage and risk of oral cancer. The differential effect of SEP on oral cancer demonstrated through the results of our social mobility models may be a reflection of the reversible nature of epigenetic modifications.

With respect to SEP in childhood, a disadvantageous childhood SEP has been associated with higher psychological stress, and altered lymphocyte activity through heightened cytokinin production (467). This is significant, as tumour infiltrating lymphocytes represent an immune response against tumour antigens and the distribution of lymphocytes in lymph nodes serves as a significant prognostic biomarker for SCCHN (468). Furthermore, Tang et al (2013) documented that children living in adverse social conditions had significantly higher methylation levels in genes associated with multiple cancers. Borghol et al (2011) identified hypermethylation clusters in specific parts of DNA in adults who experienced a disadvantageous SEP in childhood (469). Associations between early life stress and epigenetic modifications through stress response pathways have also been reported (35, 448, 470).

In summary, the biological pathways that could connect socioeconomic disadvantage to oral cancer is a field of active research, and they have not been conclusively demonstrated for oral cancers. Nevertheless, psychosocial stress and the consequent physiological response cascade resulting in various neuro-immuno-endocrine, genetic and epigenetic changes is a plausible biological explanation for our findings and the long-lasting effect of adverse childhood SEP on oral cancer development in adult life.

8.2 Manuscript II

The second manuscript investigated the association between several genetic variants involved in the metabolism of tobacco derived carcinogens and SCCHN risk among a sample of the Montreal Caucasian population recruited at the HeNCe Canada site. Furthermore, a comprehensive analysis of joint effects, stratum-specific effects and interaction between these genetic variants and three levels of smoking was also carried out. We identified that, relative to non-carriers, carriers of GSTP1 105Val were associated with 29% decreased risk for SCCHN, independent of tobacco smoking. Although results from three meta-analytical reviews did not show an association between carriers of the Val allele and SCCHN risk (140, 145, 187), a lower risk for SCCHN and other cancers among Val allele carriers has been documented (190-195). The lower risk for SCCHN among carriers of GSTP1 105Val has been attributed to the increased efficiency of the enzyme coded by this variant in detoxifying the carcinogenic epoxide of PAH (e.g., benzo(a)pyrene). Although tobacco smoke was adjusted for in the model, we did not have measures in our data to adjust for other sources of PAH such as diet, vehicle exhaust and wood combustion. Indeed, the protective effect of the Val allele (51% decreased risk) was also seen among the strata of non-smokers as well relative to non-carriers. In addition, we identified a 41% decreased risk among heavy smokers who were Val allele carriers relative to non-carriers. The documentation that the GSTP1 enzyme among Val allele carriers shows increased detoxification activity in the presence of PAH metabolites may provide one explanation for this finding (202).

None of the other variants tested were associated with SCCHN risk. Multiple meta-analytical reviews have failed to document strong associations between genetic variants such as CYP1A1*2A, *2C and CYP2E1c2 and SCCHN risk among Caucasians (135-137, 145, 165, 169-171, 180, 181). Furthermore, although the joint effect analysis indicated differential risk for various genotypes tested (e.g., increased risk for joint effect of heavy smoking and both carriers and non-carriers of CYP1A1*2A, GSTM1null, non-carriers of CYP1A1*2C, CYP2A6*2, CYP2E1c2, GSTP1 105Val, relative to non-smokers and non-carriers in each group), we did not identify any evidence of statistical interaction on either a multiplicative or additive scale for any of the variants tested.

Interestingly, there was an indication of excess absolute risk among carriers of GSTP1 105Val who smoked moderately. However, given the lower risk of SCCHN for GSTP1 105Val carriers and the lack of association between moderate smoking and SCCHN risk, this result of positive additive interaction should be interpreted with caution. Larger studies in this target population employing similar methods as ours to control of confounding, and model specification, are required to validate this finding.

8.3 Manuscript III

Manuscript III explored the scenario in which the causal effect of two genetic variants involved in tobacco and alcohol metabolism on SCCHN risk occurs through pathways involving not only interaction, but also mediation with heavy tobacco smoking and alcohol consumption.

It is well-documented that non-carriers of ADH1B*2 SNP (GG genotype) are at an increased risk for SCCHN among alcohol consumers in various ethnicities including Caucasians. Similarly, non-carriers of SNPs that result in slower activity of the CYP2A6 enzyme (e.g., CYP2A6*4, CYP2A6*2) are hypothesised to be at higher risk for tobacco-related cancers such as SCCHN. These genetic associations with SCCHN may be brought about either directly or through the interdependencies between CYP2A6*2 and smoking, and ADH1B*2 and alcohol consumption behaviours.

Nicotine is one of the major determinants of smoking frequency and the CYP2A6 enzyme is involved in the metabolism of 80-90% of nicotine entering the body. Non-carriers of the CYP2A6*2 variant (TT genotype) of the CYP2A6 gene coding this enzyme smoke with higher intensities (160, 223, 224, 229, 230). Similarly, through a mechanism that involves the conversion of ethanol to acetaldehyde, non-carriers of the *2 variant of the ADH1B gene have a relatively increased risk for alcohol consumption. The overall evidence suggests an indirect effect, in which the effect of being a non-carrier of CYP2A6*2 or ADH1B*2 on SCCHN is mediated by heavy smoking and alcohol consumption behaviours, respectively.

However, there is also evidence that the effect of these SNPs on SCCHN is brought about through interaction with smoking and alcohol risk behaviours. Apart from nicotine, CYP2A6 also metabolises pro-carcinogenic nitrosamines, found abundantly in tobacco smoke, to carcinogenic compounds. Hence, non-carriers of slow-metabolizing genotypes display higher concentrations of carcinogenic compounds, increasing the susceptibility of these individuals to tobacco-related cancers. This gene-smoking interaction pathway is supported by the higher risk for tobacco-related cancers documented among non-carriers of the CYP2A6*4 variant (or a combination of such slow-metabolizing variants of the CYP2A6 enzyme) among smokers but not among non-smokers (41, 190, 231). The association between CYP2A6*2 and SCCHN risk has not yet been documented. However, the joint effect analysis from manuscript II of this dissertation indicated that, relative to non-carriers of CYP2A6*2 (TT genotype) who never smoked, non-carriers who smoked heavily were associated with approximately 2.5-fold risk (OR= 2.56, 95% CI=1.21, 5.44) for the disease, whereas carriers who smoked heavily did not have a higher risk for SCCHN. Similarly, through a mechanism that involves acetaldehyde, non-carriers of ADH1B*2 are associated with an increased risk for SCCHN among alcohol consumers. This gives rise to the possibility of an interaction pathway in which being non-carriers of CYP2A6*2 and ADH1B*2 alters an individual's susceptibility to the carcinogenic effects of tobacco and alcohol, respectively, increasing the risk for SCCHN.

This combined mediation and interaction scenario presents a total of four non-overlapping pathways between the variants and SCCHN. Hence, apart from segregating the direct and indirect effects (mediated by heavy intensities of the respective risk behaviours) of TT (CYP2A6*2 non-carriers) and GG (ADH1B*2 non-carriers) on SCCHN, manuscript III also aimed to quantify the proportion of the excess risk for the outcome due to these genetic exposures that is completely independent of the mediator, is due to interaction but not mediation, is due to mediation but not interaction, and is due to both mediation and interaction, in the case-control sample from the HeNCe Canada site. This was achieved using mediation and four-way decomposition techniques introduced in the recent literature employing the counterfactual causal framework.

Our results indicated that both the TT genotype (CYP2A6*2 non-carriers) among smokers, and the GG genotype (ADH1B*2 non-carriers) among alcohol consumers conferred a positive excess risk for SCCHN. Overall, our estimates, albeit imprecise, indicated that most of the effect of TT and GG genotypes on SCCHN risk was through pathways not mediated by heavy smoking and alcohol intensity, respectively. However, a measurable portion of interaction (24%) and mediation (21%) by heavy smoking did play a role in the positive excess risk of CYP2A6*2 on SCCHN. For the GG genotype as well, smaller but measurable proportions of interaction (11%) and mediation (10%) with heavy drinking also contributed to SCCHN risk. Although our analysis was limited by sample size, our estimates indicated that the direction of associations (exposure-outcome, mediator-outcome and exposure-mediator) were as documented in the literature. Moreover, within the context of genetic variants, associated risk behaviours and an oral health outcome, this manuscript demonstrates a powerful causal analytical technique that can be used to decipher causal pathways defined by both meditation and interaction between the exposure and the mediator.

8.4 Project limitations - Validity of the findings

This section examines potential non-causal explanations of our results. I discuss three main structural threats to the validity of results in this work commonly classified as systematic errors: selection bias, information bias and confounding.

8.4.1 Selection bias

The causal diagrams framework identifies a common structure underlying all scenarios referred to as selection bias: conditioning on a "common effect" of the exposure and the outcome, or their causes (370). In a case-control study, the OR estimate is conditional on having been selected into the study. Selection bias in case-control studies, majorly dependent on properties of control participants selected, can arise due to the lack of a proper definition of the study base, dependence between controls and exposure, differential participation rates and degree to which exposure distribution among controls represents that of the underlying source population. Multiple strategies adopted to mitigate selection bias in our hospital-based case-control study are discussed below.

8.4.1.1 Steps taken to define the secondary study base

For case-control studies, selection bias can ensue if controls do not arise from the same precisely defined and identified population i.e., the source population, from which cases are identified. In a hospital-based case-control study, the source population is termed the secondary study base (340). Ensuring that controls are indeed members of the secondary study base and represent the experience (e.g., exposure distribution) of this study base is a challenge. To address this challenge in the HeNCe Life study, controls were recruited from several outpatient clinics in the same hospitals as the cases. The study base was refined using a set of strict exclusion criteria that applied equally to the selection of cases and controls. These criteria included restricting participants to the specific geographic location served by the participating hospitals (150 and 50km radius around participating hospitals in Kerala and Montreal, respectively), which reduced the likelihood of recruiting participants from another study base.

8.4.1.2 Steps taken to ensure independency between control selection and exposure status

The central tenet in case-control studies to avoid selection bias is to sample controls independent of the exposure under study (471). To ensure this, participants were sampled irrespective of their SEP, genetic, smoking or alcohol behaviour. The recruitment of controls from clinics treating conditions related to tobacco or alcohol consumption can lead to an overestimation of true causal effects in tobacco and alcohol related studies on head and neck cancers (378). Hence, in our study controls were recruited from several clinics that are not typically associated with smoking and alcohol. In addition, recruiting from several clinics dilutes the biasing effect of overrepresentation of any particular disease group (340). To rigorously limit the overrepresentation of a single diagnostic/disease group and consequently any particular risk factor, the participation of controls from each clinic was restricted to approximately 20%. The participating hospitals at each site were tertiary medical facilities and catered to individuals irrespective of their SEP (main exposure in manuscript I), genetic, smoking and alcohol consumption profiles (main exposures in manuscripts II and III), mitigating selection bias due to differential referral systems (471).

8.4.1.3 Bias associated with participation rate

A significant decline in participation rates for epidemiological studies has been documented over the past 30-40 years (472, 473). In addition, a lower participation rate among controls relative to cases has been consistently recorded in case-control studies. However, a low participation rate or differential rate between cases and controls carries the potential for selection bias only if the reason for refusal is a common effect of both the exposure and outcome status.

At the Indian study site, the participation rate was 85.6% and 44.3% for cases and controls, respectively. Considerable efforts were made to ensure maximum participation at both sites. However, at the Indian site, oral cancer cases showed willingness to participate immediately after diagnosis and were interviewed at the clinics by dentists (including the PhD candidate). This ensured that no cases were missed during recruitment and that cases did not undergo any cancer related treatment that may impede biological sample collection. From a cultural point of view, patients found it necessary to talk to dentists (interviewers) about their disease status, which may also have contributed to a higher response rate among controls. In contrast, although controls were also selected from outpatient clinics in the same hospitals, most were interviewed in their homes after making an appointment by telephone. Although the method of screening eligible participants on the phone increased the feasibility of the study, it may have increased the chance of refusal (473, 474). The higher consent rate among cases may also be attributed to the salience, importance and dread individuals associate with “cancers” (473) Controls do not exhibit this feature and therefore may not consider their participation worthwhile (473). These reasons show how disease status could affect participation rate at this study site. There is evidence that socioeconomic factors may affect participation in case-control studies (474). However, except for age and sex we did not have access to any data on non-participants to assess the magnitude and direction of potential bias in our hospital-based study.

At the Canadian site, the participation rate was 54% for cases and 47% for controls. Eligible participants were selected from outpatient clinics and both cases and controls were mostly interviewed in their homes following appointments over the telephone. This may be a reason as to why the participation rate was low in Canada. Here, although case-control status can affect the decision to participate, we believe that it is improbable that genotypes of an individual (unknown to the individual or interviewer) could have affected their participation (248).

8.4.1.4 Comparability of the distribution of exposure between controls and the source population

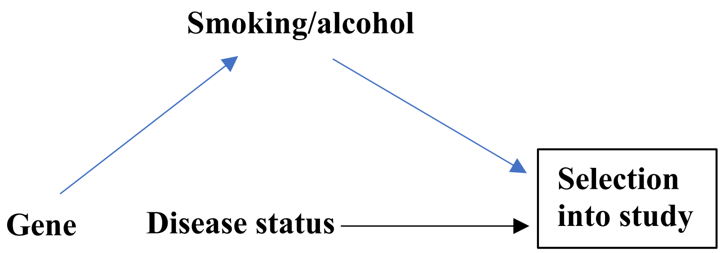
As mentioned earlier, selection bias can ensue if the exposure distribution of controls in the study differs from that of the source population. Participants at the Indian site were recruited from tertiary hospitals in the Calicut district of Kerala, India. A comparison of the housing assets of controls in our study (longest residence in late adulthood) and data from the Census of India 2011, Calicut district, Kerala, showed that the distributions were similar (475) (*please refer to Appendix VI, page no.317*). Unlike other indicators of SEP, housing assets are not prone to change due to short-term economic shocks, which increases the comparability of both sets of data. In addition, the distribution of religions among the controls in our study (Hindus=61%, Muslims=35%, Christians=3%) was comparable to that in Calicut in 2011 (Hindus=57%, Muslims=39%, Christians=4%). The majority of the Hindus were from the middle caste (other backward caste) and belonged to the Thiyya/Ezhava sub-caste, as in the source population. These factors are strongly correlated with SEP in India (476). Altogether, the results of these comparisons indicate that the controls in our sample were a good representation of the source population with respect to the main exposure.

At the Canadian site, data analyses for both manuscripts II and III were restricted to Caucasians. The frequencies of minor alleles of SNPs among controls were similar to those reported by other studies among the general Caucasian population. The proportion of current cigarette smokers among controls in our data was 21%, which is comparable to the range of 22.2% to 17.8% reported in Quebec during the study period (between 2005 and 2013). However, the proportions of current users (68.3%) and ever-users (82.3%) of alcohol among controls in our sample were lower than those reported in Quebec (current users=82%, ever-users=92.8%) and Canada (current users=77.2%, ever-users=89.35%) between 2010 and 2012. However, unlike data collection methods for smoking, those used for alcohol in these surveys were not well defined and detailed, reducing comparability with our data. Overall, the controls in the Canadian sample analysed were a fairly good representation of the Quebec population with respect to smoking and alcohol. The overall representativeness of study samples at both the study sites to the underlying population increases the confidence in our estimates, against any probable bias due to differences in response rates.

8.4.1.5 Selection bias-variance trade-off due to the analytical strategy in manuscript III

In manuscript III, the CYP2A6*2-smoking-SCCHN analysis was restricted to ever smokers, and the ADH1B*2-alcohol-SCCHN analysis was restricted to alcohol consumers. Here, the genotypes affect the risk behaviour and hence, as depicted in ***Figure 11***, will lead to a biased association between each genotype and SCCHN through the risk behaviour and the selection node.

Figure 11: Illustration of selection bias in manuscript III due to restriction



Restricting these analyses to smokers only or alcohol consumers only leads to an underestimation of the true causal effect as, a) the control series will have a higher level of the risk behaviours relative to the source population with cases (370), b) cases have a higher likelihood of smoking, and controls will be more similar to cases. However, the restriction was carried out as there is no effect of CYP2A6*2 on SCCHN among non-smokers and no effect of ADH1B*2 on SCCHN among non-alcohol consumers. In other words, there is no direct effect of genetic variants on SCCHN unless the risk behaviours are involved. Therefore, restricting the analysis to those who had the risk behaviours may be the preferred approach. Restriction may induce some collider stratification bias for the association between the variant and SCCHN, which is likely to be minimal as the genotypes have a weak effect on the respective risk behaviours. However, not restricting may lead to a higher variance because a large proportion of the controls have no direct exposure effect (the never smokers and never alcohol consumers). Hence, the approach used was adopted assuming a small bias vs large variance trade-off. Due to low power, most of our estimates were imprecise. Hence, performing a sensitivity analysis with our data would not have added to the understating of this trade-off. Simulation studies or larger studies in future should consider quantification of this trade-off by inclusion and exclusion of never smokers and never alcohol consumers.

8.4.2 Information bias

Information bias occurs when the variables (e.g., outcome, exposure, confounders) used for analysis in a study do not represent the true values of these variables (measurement error) or the variables are wrongly categorized (misclassification) (477, 478).

Measures adopted in this study to mitigate these biases are discussed below.

8.4.2.1 Misclassification of the outcome

The outcome in our study was SCCHN. Participants were classified as cases based on histological confirmation, which is the gold standard for the diagnosis of malignant lesions. Hence, the potential for outcome misclassification in this work is highly improbable.

8.4.2.2 Measurement error in exposures and confounders during data collection and management

Risk of biased information on exposures due to participants' inaccurate recall of experiences from their past is inherent in case-control studies (340). With respect to measurement error in tobacco and alcohol related variables, recall of past smoking status after 20 years has been documented to be valid (kappa=0.80), while the amount smoked (kappa=0.63) was not recalled as well as smoking status (479). Alcohol status and consumption were reported with an accuracy similar to that for smoking (479). However, recall bias on exposures dependent on outcome status has been cited as a major concern in this study design. Among cases, the diagnosis may affect the reporting of these variables either by increased sensitivity (due to improved memory), decreased sensitivity (clouded memory) or reduced specificity (false memory) towards recall of exposure status. The direction of this differential misclassification cannot be predicted (340).

However, at both study sites, multiple measures were adopted to reduce measurement error in exposure and confounder variables *(please refer to sub-section 4.8, page no. 64)*. Interviewers were blinded to the study hypothesis during interview sessions. The extensive interviews averaged 1.5 to 2 hours in length at both study sites and were conducted in private settings. This, along with interviewers who were well versed in the local language and culture at each study site, contributed to build a good rapport with the participants. In addition, interviewers at both sites were comprehensively trained in the data collection procedures and were advised on strict adherence to the study protocol. Multiple questions concerning a single exposure were used in different parts of the questionnaire, which helped to ensure consistency in the reported information. Moreover, a life grid tool was used in tandem with the questionnaire interview at multiple steps during the interview process for both cases and controls. This technique substantially improves recall of residential, behavioural and occupational information (381, 382). In addition, information collected from hospital-based controls are expected to have less error due to recall compared to population-based controls, since controls and cases are both diseased (340). The strategies described above may decrease the possibility of differential misclassification in hospital case-control studies and increase the expectation of non-differential misclassification of exposures, resulting in an overall bias towards the null (340, 478).

Among the total sample of 721 participants recruited at the Indian site, re-interviews were conducted with 46 randomly selected participants, 6 to 12 weeks after the original interview. Relative measures of test-retest reliability for housing-based assets used to create SEP measures for childhood, early and late adulthood periods showed very good test-retest reliability (inter class correlation >0.85) (please refer to *manuscript I, Supplemental material, eTable5, page no. 124*). A validation study conducted at the Canadian site among a random sample of cases and their matched siblings also showed very good agreement between the two groups validating our measures *(please refer to Appendix VII, page no.318)*. Finally, for variables such as age and sex, there is less expectation of measurement error.

8.4.2.3 Misclassification of exposures due to the categorization of continuous variables

In this study, data on SEP, education, tobacco and alcohol behaviours were initially derived as continuous variables. Analyses that were conducted by conditioning on these variables used them in their continuous form. However, they were summarised as categorical variables when used as the main exposure in each manuscript. The categorization of continuous variables has been criticized in the literature. However, we used categorical exposures in this study for multiple reasons. First, analyses for manuscripts I and III were based on the counterfactual framework and the use of categorical variables makes it easier to demonstrate the causal contrasts being compared for the exposures. Second, a non-linear dose-response relationship was identified for smoking and alcohol variables. However, it was difficult to infer any difference in the shape of the dose-response curve when stratifying by genotype. Genetic variants with low minor allele frequencies may exert their effect on SCCHN only at elevated levels of exposure to smoking or alcohol. Such differential effects may be difficult to identify and interpret using dose-response curves. One of the objectives of manuscript II was the comprehensive characterisation of gene-environment interactions in which genetic variants were the primary exposure and smoking the secondary exposure. Using categorical data makes it easier to interpret joint, stratum-specific and interaction effects. Third, this work introduces multiple complex analytical techniques such as IP weighted marginal structural modeling for time-varying exposure and several confounders over multiple time periods, mediation analysis and decomposition of effects into the oral health field. The use of categorical exposures is a starting point where by these techniques can be demonstrated and results better interpreted, before moving into more complex continuous data (480).

There is an absence of *a priori* knowledge regarding the prognostic value of different cut-off levels for the categorization of SEP, tobacco and alcohol variables with relevance to the SCCHN outcome. Hence, although criticized in the literature (340, 481, 482), we had to adopt a data-dependent method to ascertain the cut-off levels to categorize our exposures. The qualitative meaning of SEP related variables such as education may be subjected to a cohort effect, leading to misclassification during their categorization. Although we attempted to deal with the possibility of a cohort effect on education at the Indian site (*please refer to page. no 68, sub-section 4.9.3.1*), it is a relatively crude attempt to adjust for the significant secular, economic and political changes that occurred in the last century in Kerala, India. To create the asset index, only housing indicators with very high test-retest reliability were used for PCA analysis to create the continuous SEP scores for each period of life (*please refer to manuscript I, Supplementary material, eAppendix, page no. 124*). The SEP scores were dichotomized using the median of the distribution of these scores among controls (representative of the underlying population) as the cut-off. We expect any misclassification in the SEP variable to be non-differential, which will, on average, bias the estimates towards the null.

In manuscripts II and III, I identified optimal cut-off points for secondary exposures and mediators respectively using a combination of splines that inform about the potential range of exposure in which the cut-point might lie, and a parametric outcome-based approach that corrects for multiple testing. This approach is advantageous over other methods documented in the gene-environment interaction literature (e.g., cut-offs based on percentile distribution of these exposures among controls without consideration of a non-linear dose-response relationship, maximising sensitivity and specificity between the groups with respect to the outcome and minimizing p-value, that can induce error due to multiple testing over several cut-points possible, visual methods using splines alone which may induce subjectivity (36, 248, 252, 483). Most variants investigated in this study do not have an effect on SCCHN in the absence of smoking/alcohol exposures. This suggests that the cut-off points are at a higher level than in the normal population. Hence, the outcome-based approach was conducted on samples from which non-users of corresponding risk behaviours were excluded. Any outcome-based exposure classification approach indeed has the potential to induce differential misclassification bias. However, for gene-environment interaction studies, it has been demonstrated that differential misclassification of exposure need not produce a bias under two conditions: 1) there is no association between the genotype and the environmental exposure among controls, 2) there is no association between the genotype and the exposure among cases (484). A bias analysis conducted in the data used for manuscripts II and III showed no evidence of association between the genetic variants and smoking/alcohol exposures among controls or cases (*Manuscript II, Supplementary material, eTable2, page no.154, and Manuscript III, Supplementary material, eTable 3, page no. 187*). Thus, for both these manuscripts, random error appears to be the major threat to validity, rather than any structural bias.

8.4.2.4 Misclassification in HPV and genetic data

Although validated and reliable methods were used to collect, isolate and genotype the DNA samples, misclassification in HPV and genetic variant genotyping could potentially have occurred due to variations in the sensitivity and specificity of the genotyping assays used (485). However, the laboratory personnel were unaware of the case control status of the samples and case and control samples were processed together. Moreover, molecular analytical procedure was used for all samples. Hence, we expect any misclassification to be non-differential, biasing estimates towards the null.

8.4.3 Confounding bias

The structural pattern of confounding bias when estimating the total effect of an exposure on an outcome has been explained under *section 2.7.4 DAGs, page no.45*. Minimal sufficient sets of confounders required to estimate the total effect of each exposure on the outcome in each manuscript were identified using DAGs. Continuous variables, especially tobacco and alcohol consumption, were adjusted for both their dose and non-linear functional forms in all analyses. The use of IPW in manuscript I ensured that the SEP exposure was independent of all measured covariates and exchangeability was ensured within the measured potential confounders. There is potential for confounding due to time-varying occupation or income-based SEP measures, which were not used in this study. However, these variables would exert their effect only during early or late adulthood periods of life. Hence, our overall finding that disadvantageous SEP in childhood is critical for oral cancer risk in adult life is robust to effects of confounding due to these SEP measures.

Because manuscript II aimed to quantify causal interaction rather than effect modification or effect heterogeneity, all measured confounders of genetic variants-SCCHN, and smoking-SCCHN were included in the models. The analyses of manuscript III relied on stronger assumptions related to confounding as the causal interpretation of results from mediation and decomposition analysis; using the counterfactual framework for these analyses required the identification of all confounders between the exposure-outcome, exposure-mediator, mediator-outcome and any mediator-outcome confounder affected by the exposure. However, the possibility of bias in the estimates due to unmeasured confounding in both manuscripts II (between genetic variant and

SCCHN, or smoking and SCCHN) and III (mediator and outcome) cannot be ruled out. For example, the proxy for SEP used in both these manuscripts was the number of years of education, which cannot account for all facets of SEP (e.g., occupation, housing conditions). However, SEP has an inverse association with both risk behaviour and SCCHN risk, i.e., the direction of association is the same. In manuscript II, we explored interactions which is a form of direct effect, and the results in manuscript II indicated that most of the effect of the variants on SCCHN was direct. It has been demonstrated using sensitivity analysis although not accounting for measures of SEP is likely to result in an overestimation of indirect effects, the direct effect will be under estimated, biasing estimates towards null (486).

8.4.4 Integrity of inverse probability weights and marginal structural models

Under the assumptions of correct specification of the model used to estimate weights, exchangeability, positivity and consistency, IP weighted MSMs can produce unbiased causal estimates. It should be acknowledged that the effect estimates are sensitive to misspecification of weight models and, in observational studies, it is not possible to accurately ensure the other three assumptions. In our study, first, all possible measured time-varying confounders were identified using DAGs and used in the weight models. The vector C3b (*please refer to **Figure 9**, page no:74*), representing risk behaviours identified above 50 years of age, was a complete mediator and hence not used in the models. Second, because flexible modeling of time-varying confounders mitigates residual confounding and reduces the potential bias due to model misspecification from strong linearity assumptions, three-knot restricted cubic splines were added to all linear terms (e.g., paan chewing, smoking, alcohol) in the weight models. Third, the weights that were finally used had a mean close to 1 and a small range compared to weights from other models (e.g., models in which confounders were included as linear terms or categorical). Weights were truncated at the first and 99th percentile to ensure proper weight behaviour. Reporting estimates using weights constructed by these techniques has been demonstrated to be reasonably robust to alternative weight model specifications, as well as to mitigate violation of exchangeability and positivity (426). Hence, we did not explore this further using specific sensitivity analysis. Whether the assumption of consistency holds with social exposures is a controversial topic in the current literature (284, 487, 488). In this study, we laid out causal contrasts for our SEP exposure in all the life-course models tested (under the simple binary categorization of the childhood, early and late adulthood SEP exposures). Under our definition of causal effects, for an association to be causal, an intervention on the exposure should produce a change in the outcome. However, measures of SEP (e.g., education, income, wealth) are composed of multiple components (e.g., multiple assets, sources of income, education indicators) and the effect on SCCHN may vary based on the component intervened up on. This makes most social exposures, which are “compound treatments”, not-well defined (relative to specific doses of medications, carcinogens, viral load), and non-manipulable (487, 489), consequently violating the consistency assumption. Hence, we report our estimates in manuscript I without overstating their causal nature.

8.5 Strengths and contributions of the study

The unique feature of this work is the amalgamation of two powerful theoretical and analytical frameworks, life-course epidemiology and the counterfactual causal framework, within a case-control design to elucidate the causal pathways underlying the relationship between social, genetic and behavioural exposures and SCCHN risk. The use of data from the HeNCe Life study ensured that our strong analytical techniques were mounted on rigorous retrospective data collection, management and quality control procedures at both study sites. Throughout this work, I have used a structural approach (causal diagrams) to illustrate causal associations between variables, and addressed potential confounding, selection and information bias. All exposures and confounders used in this thesis were measured in a rigorous and comprehensive fashion through two-hour face-to-face private interviews. The life-grid technique was used to improve the quality and reliability of retrospective data collected. Despite the case-control design, the life-course conceptual framework underlying study procedures allowed us to approximate the temporal relations between variables. This feature was thoroughly exploited in manuscript I, which aimed to identify which life-course hypothesis pertaining to SEP over the life-course explained the most variance in the oral cancer outcome in the target population in Kerala, India. To our knowledge, this is the first case-control study to investigate an exposure-outcome association by appreciating the complex system of time-dependent feedback loops between variables, including not only the exposures but also potential confounders measured over several periods of life. To address this issue, we employed novel approaches to existing methods in the causal inference literature, incorporating IP weighting to control for time-varying confounding (originally developed for longitudinal data), and weighted partial likelihood estimators for time-dependent exposures developed for case-control data (429, 490). We consider this to be a significant contribution to the advancement of analytical methods in life-course research. After satisfying several identifiability assumptions (*please refer to page no. 43, sub-section 2.7.1*), the estimates of MSM can be given a causal interpretation even in the presence of time-varying confounding. Without overstating the causal nature of the association between SEP and oral cancer, we believe this work produced improved estimates relative to past studies that used single-step regression techniques.

The genetic component of the HeNCe Life study was utilized in manuscripts II and III, which explored causal pathways from genetic risk factors to SCCHN defined by interaction and mediation with associated risk behaviours. In manuscript II, we adopted an elaborate strategy to provide sufficient information to interpret gene-environment interaction results on both multiplicative and additive scales.

The work undertaken in manuscript III is one of the first demonstrations of the four-way decomposition analytical technique within a case-control study on an oral health outcome. This method provides maximum insight into the interrelationship between two specific variables and their effect on an outcome based on interaction and mediation (441). In our work, we used this technique in the context of a binary genetic exposure and single binary mediating risk behaviour for SCCHN. This analytical strategy is based on the counterfactual framework, which allows for mediation and four-way decomposition analysis in the presence of exposure-mediator interaction. This is an advantage of the counterfactual-based technique over other methods proposed in the literature (434, 440). It should be noted that, in a case-control study, the total effect of the genetic variants on SCCHN risk cannot be decomposed into four non-overlapping components (defined by mediation and interaction), unlike cohort data that provides risk estimates on an absolute scale. However, the four-way decomposition technique does provide valid estimates of proportions of total excess risk attributable to the four causal pathways in a case-control study (441). Using this technique, we also demonstrated how much of the risk for SCCHN from the direct non-modifiable genetic risk factors can be eliminated from the target population, if the intermediate risk behaviour, which is modifiable, is intervened upon.

A characteristic of case-control studies often cited as a limitation of this design is that they provide information only about the one outcome the study is sampled on, and therefore analytical models cannot be fit on any other exposure as a dependent variable (340). However, the possibility of carrying out such analyses has been argued in the literature. Analytical techniques such as IP weighting consist of a two-step process, the first of which is fitting a regression model with the exposure of interest as the dependent variable, i.e., the exposure model. Furthermore, mediation analysis requires parameters from two models, one of which is a mediation regression model in which the mediator variable is the dependent variable. This is one of the reasons why researchers do not undertake these analytical techniques in case-control designs. We addressed these challenges by fitting these models (i.e., the exposure model and the mediator model described above) among controls only, under the rare disease assumption (343). The robustness of this technique was ensured by alternatively weighting these models with the inverse of sampling fractions (491, 492). This technique corrects for the biased sampling in case-control studies and approximates the distribution of exposures in our study to that of the underlying population (as in a cohort study), even when the rare disease assumption does not hold (366). Furthermore, our controls were non-cases at the time corresponding cases occurred. This incident density sampling pattern (340) ensured that our odds ratio was equivalent to relative risks that would have been obtained from a cohort study.

The lack of statistical codes for commonly used statistical packages is one of the challenges faced by researchers while applying complex statistical techniques. In this study, we used Stata statistical software, for which codes had not been adapted to execute the IP weighted MSM technique (manuscript I) in case-control studies. Moreover, there were no automated Stata codes available to execute the four-way decomposition analytical technique in Stata, irrespective of study design. Thus, the codes for manuscript I had to be written by combining exposure and outcome logistic regression models and sampling weights for time-varying exposures in case-control studies. For manuscript III, parameters from the mediator and outcome models had to be combined and inputted in mathematical formulas for the four-way decomposition technique. The point estimates for risk ratios were calculated from the coefficients retrieved using these formulas. Confidence intervals were estimated using the bootstrapping procedure (443, 444). Codes for Stata incorporating these steps were exclusively written for this analysis (*provided in manuscript I, Supplemental material, eTable8, page 126, and manuscript III, Supplemental material, eAppendix, pages 182-186*). The Stata codes written for the four-way decomposition are specific to a binary outcome, binary mediator and binary exposure scenario (ratio scale), and can be extended to scenarios with other functional forms of these variables. Overall, this thesis contributes new insights into the mechanisms involving social, genetic and behavioural risk factors that lead to the outcome of SCCHN.

8.6 Public health implications

It is arguably dangerous to put forth direct public health implications of results from a single epidemiologic study. Nevertheless, the validation and replication of our findings by future studies using similar methodology and analytical techniques can contribute to opportunistic screening for SCCHN and the identification of high-risk groups. For example, the prevention of oral cancer is largely dependent on screening and the early detection of lesions. Dentists, nurses and hygienists can support the reduction of oral cancer risk factors through clinics, as well as through regular dental camps/outreach programs targeting populations with disadvantageous SEP. In its current form (e.g., use of toluidine blue dye, fluorescent imaging, brush biopsy), population-based screening for oral cancer is not cost-effective (493). Furthermore, evidence on the effectiveness of the visual screening method is insufficient. However, the systematic examination of the oral cavity by dentists and physicians with particular attention to high-risk sub-groups (e.g., those exposed to a disadvantageous childhood SEP, and tobacco or alcohol risk behaviours) is largely recommended. The opportunistic screening of high-risk group individuals and their referral to secondary prevention programs (e.g., alcohol and tobacco cessation) by these medical service providers play a central role in a multi-disciplinary approach to the prevention of oral cancer (493). Also, valid indicators of childhood SEP may be incorporated into oral cancer risk calculations and screening tools. Factoring in the negative effect of low SEP on oral cancer, specifically childhood SEP, can increase the precision of risk calculations and enhance the effectiveness of opportunistic screening.

The elucidation of carcinogenesis pathways to SCCHN is not immediately useful for public health purposes, but may eventually lead to high-risk group identification for targeted recommendations and interventions. The analyses of manuscript II were undertaken to identify the sub-group of individuals in whom genetic variants may exert their effect based on differential smoking patterns. High-risk group identification can be useful to target smoking related public health interventions to achieve a greater reduction in SCCHN risk. Because smoking at any level increases the risk of

SCCHN, ideally any tobacco-related intervention such as cessation programs should target the whole population and not only individuals with high-risk genes. Channelling such interventions to a fraction of the population may raise ethical questions. However, there are practical scenarios (e.g., lack of funding, political priorities) in which it may not be possible to intervene on the entire population and resources may be available to target only a fraction of the population. High-risk group identification is of greater help in such resource-limited situations. But it should be acknowledged that, despite its promising initiation, rapid development and growth in the past two decades, the field of gene-environment interaction studies has provided little in the way of practical benefits thus far. This is mainly due to inconsistent results, which may be related to inadequate study power, poor control selection, failure to obtain accurate environmental exposure data, failure to take these exposures into account in the analyses, or excessive false positive reports. Nevertheless, gene-environment interaction studies with comprehensive result reporting are fundamental to identify complex pathways to cancer as well as high-risk groups.

Finally, although the results of manuscript III were imprecise due to sampling variability, the methodology used is beneficial to derive estimates of policy relevance (494). For example, the majority of the Caucasian population carries the TT (CYP2A6*2) and GG (ADH1B*2) genotypes. The direct modification of these variants to reduce their effect may not be possible nor economical. However, their effect on SCCHN risk can be modified by changing the level of modifiable risk behaviours such as smoking and alcohol consumption. If the estimates in this study were indeed valid, then our results suggest that approximately 37% of the effect of the TT genotype on SCCHN risk could be eliminated if the level of smoking was brought down to that of moderate smokers (i.e., under a pack of cigarettes per day). And among GG carriers, about 16% of SCCHN risk could be eliminated if the alcohol consumption level was brought down to that of mild drinkers (< 25ml of ethanol per day, equivalent to < 250ml wine or aperitif, < 500ml beer or cider, or < 50ml hard liquor, per day). The examination of a larger study population is needed to clarify these relationships.

8.7 Future directions

The scientific method is essentially reductionist, as complex causal systems are broken down into single exposures and investigated to quantify their effect on outcomes. Complex disorders such as

SCCHN result from interdependencies between social, behavioural/environmental, genetic and biological risk factors. In this work, we attempted to move one step up the ladder of complexity by conceptualizing complex relations within a single exposure (e.g., SEP), or two exposures within the life-course framework. In manuscript I, we considered the complex time-varying nature of the SEP exposure and associated confounders. The results of this study provide a platform for future research questions; for example, how much of the effect of disadvantageous SEP in childhood is mediated by adulthood SEP as well as adult risk behaviours? Recently, an integration of structured life-course approach, and mediation and four-way decomposition analysis under the counter-factual framework has been proposed in which each life-course hypothesis is viewed under the lens of a specific set of mediation and/or interaction terms (495). It will be interesting to investigate the association of SEP with oral cancer using this novel approach.

Manuscripts II and III explored genetic associations as well as the interrelation between a single genetic exposure and risk behaviour in the risk for SCCHN through interaction, mediation and a combination of these causal mechanisms. However, these may involve multiple gene-gene and gene-environment interactions. For example, the two Phase I genetic variants CYP1A1*2A and *2C, which are in linkage disequilibrium in Caucasians, can be considered as haplotypes. A thorough investigation of their effects on SCCHN would require a haplotype analysis (i.e., to assess their combined effect). However, this was not undertaken in our study due to small numbers across genotype groups. Furthermore, the combination of Phase I (e.g., CYP1A1) and Phase II (e.g., GSTM1) genetic variants has been associated with altered SCCHN risk (496). The next steps using studies with larger samples would aim to address the effect of these gene-gene interactions on SCCHN. Heterogeneity has been identified for the effect of variants studied in this work (e.g., CYP1A1*2C) with respect to subsites of the head and neck region (497). Furthermore, in this work, due to the low minor allele frequencies of many genetic variants, the genetic exposure variable was collapsed into two categories assuming a dominant model of inheritance. However, studies have shown differential risk associated with co-dominant and recessive models of inheritance of multiple variants (e.g., CYP1A1, GSTP1) and SCCHN risk (135, 136, 498, 499). Future studies with larger sample sizes conducted in the target population should attempt to explore these gene-gene interactions, haplotype analysis, heterogeneities with respect to anatomic sub-sites and other models of genetic inheritance.

Manuscript III investigated mediation of the effect of cumulative genetic exposure on SCCHN by accumulation of associated risk behaviour. However, both smoking and alcohol are time-varying exposures. The mediating effects of these variables may be different at different periods of life. Exploring the mediation question under a time-varying framework will aid not only to identify how much of the genetic risk can be eliminated by intervening on the mediator, but also the time period that must be chosen to maximize the efficiency of such interventions (500). In addition, the quantification of proportions of the genetic effect mediated by multiple mediators (e.g., smoking, alcohol, diet, occupational exposures) should also be explored in future studies.

Lastly, throughout this work we used tools and analytical techniques described under the causal inference framework. The essence of epidemiology lies in finding ways to improve population health and causal inference, although not the only step, is essential to that path. All causal inference rests on unverifiable assumptions and hence the estimates from an analysis may not be interpretable as causal. However, it should be noted that not all violations of assumptions are the same. Some are modest and their impact may not be strong enough to qualitatively change the research findings. Sadly, evidence and action are not tightly tied together in many domains of public health. Most often, it is the direction of the effect that matters rather than extremely precise effect estimates. Furthermore, the mere fact that we cannot verify that the assumptions are true is not a justification for not adopting causal analytical tools that provide rigour in the methods adopted to answer important research questions.

Chapter 9

Conclusions and contributions

- Disadvantageous SEP in both childhood and early adulthood were associated with increased risk for oral cancer in the Indian sample.

- The increased risk association identified between an accumulation of disadvantageous SEP over the life-course and oral cancer was largely explained by a disadvantageous SEP in childhood.

- The GSTM1 105Val polymorphism decreased the risk for SCCHN among the sample of Canadian Caucasians independent of smoking, as well as among heavy smokers.

- There was no evidence for statistical interaction between any of the genetic variants tested and any levels of smoking on both multiplicative or additive scales.

- Although our estimates were imprecise, the results indicated that most of the effect of CYP2A6*2 and ADH1B*2 on SCCHN in the target population seemed to be through pathways not involving heavy intensities of smoking or alcohol risk behaviours, respectively.

- We demonstrated the application of inverse probability weighted marginal structural models for both time-varying exposure and confounders across multiple periods of life, and the four-way decomposition technique, in a case-control study utilizing life-course data.

- Stata software codes adapted to a case-control design, to implement the inverse probability weighted marginal structural models (for a binary exposures and binary outcome scenario), mediation and four-way decomposition (binary exposure, mediator and outcome scenario) techniques explained under the counterfactual causal framework are provided

References

1. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral oncology. 2009;45(4-5):309-16.
2. Lippman SM, Hong WK. Molecular markers of the risk of oral cancer. New England Journal of Medicine. 2001;344(17):1323-6.
3. List MA, Bilir SP. Functional outcomes in head and neck cancer. Seminars in Radiation Oncology. 2004;14(2):178-89.
4. SEER Cancer Statistics Review; 1973–1998. https://seer.cancer.gov/archive/csr/1973_1998/oralcav.pdf.
5. Buchmann L, Conlee J, Hunt J, Agarwal J, White S. Psychosocial distress is prevalent in head and neck cancer patients. The Laryngoscope. 2013;123(6):1424-9.
6. Menzin J, Lines LM, Manning LN. The economics of squamous cell carcinoma of the head and neck. Current opinion in otolaryngology & head and neck surgery. 2007;15(2):68-73.
7. Who. International Statistical Classification of Diseases and related Health Problems 10th Revision - ICD-10 Version:2010. 2010.
8. Sanderson RJ, Ironside JAD. Squamous cell carcinomas of the head and neck. BMJ : British Medical Journal. 2002;325(7368):822-7.
9. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, Accessed on January 05, 2015.2013.
10. Shield KD, Ferlay J, Jemal A, Sankaranarayanan R, Chaturvedi AK, Bray F, et al. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA Cancer J Clin. 2017;67(1):51-64.
11. Kulkarni MR. Head and Neck Cancer Burden in India. International Journal of Head and Neck Surgery. 2013;4(1):29-35.
12. Ferlay J, Shin H-R, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer. 2010;127(12):2893-917.
13. Canadian Cancer Statistics 2016. Toronto, ON: Canadian Cancer Society; 2016.; 2016.
14. http://globocan.iarc.fr/ 2012. 2012.
15. de Oliveira C, Bremner KE, Pataky R, Gunraj N, Chan K, Peacock S, et al. Understanding the costs of cancer care before and after diagnosis for the 21 most common cancers in Ontario: a population-based descriptive study. CMAJ Open. 2013;1(1):E1-8.
16. Kam D, Salib A, Gorgy G, et al. Incidence of suicide in patients with head and neck cancer. JAMA Otolaryngology–Head & Neck Surgery. 2015;141(12):1075-81.
17. Canadian Cancer Society, 2011.
18. Segi M, Tōhoku D, Nippon Taigan K. Cancer mortality for selected sites in 24 countries. Cancer mortality for selected sites in twenty four countries. 1950:6 v.
19. Doll R, Payne P. Cancer Incidence in Five Continents, Vol. I Union Internationale Contre le Cancer, Geneva. 1966.

20. G.L. Day WJB. Second primary tumors in patients with oral cancer. Cancer. 1992;70(1):14-9.
21. Lippman Sm HWK. Second malignant tumors in head and neck squamous cell carcinoma: the overshadowing threat for patients with early-stage disease. Int J Radiat Oncol Biol Phys. 1989;Sep;17((3)):691-4.
22. Lingen MW, Kalmar JR, Karrison T, Speight PM. Critical evaluation of diagnostic aids for the detection of oral cancer. Oral oncology. 2008;44(1):10-22.
23. Carvalho AL, Pintos J, Schlecht NF, et al. PRedictive factors for diagnosis of advanced-stage squamous cell carcinoma of the head and neck. Archives of Otolaryngology–Head & Neck Surgery. 2002;128(3):313-8.
24. Blot WJ. Oral and Pharyngeal cancers. Cancer Surv. 1994;19-20:23-42.
25. Marron M, Boffetta P, Zhang Z-F, Zaridze D, Wunsch-Filho V, Winn DM, et al. Cessation of alcohol drinking, tobacco smoking and the reversal of head and neck cancer risk. 1 ed. England: International Agency for Research on Cancer, Lyon, France.; 2010. p. 182-96.
26. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chen C, Curado MP, 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):777-89.
27. International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Volume 100B. Biological Agents. Lyon, France: International Agency for Research on Cancer; 2009.
28. Forte T, Niu J, Lockwood Ga, Bryant HE. Incidence trends in head and neck cancers and human papillomavirus (HPV)-associated oropharyngeal cancer in Canada, 1992-2009. Cancer causes & control : CCC. 2012;23(8):1343-8.
29. Chaturvedi AK, Anderson WF, Lortet-Tieulent J, Curado MP, Ferlay J, Franceschi S, et al. Worldwide Trends in Incidence Rates for Oral Cavity and Oropharyngeal Cancers. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013;31(36).
30. Petti S. Lifestyle risk factors for oral cancer. Oral oncology. 2009;45(4-5):340-50.
31. The relation of socioeconomic status to oral and pharyngeal cancer. Epidemiology. 1991;2(3):194-200.
32. Conway DI, McKinney PA, McMahon AD, Ahrens W, Schmeisser N, Benhamou S, et al. Socioeconomic factors associated with risk of upper aerodigestive tract cancer in Europe. Eur J Cancer. 2010;46(3):588-98.
33. Conway DI, Petticrew M, Marlborough H, Berthiller J, Hashibe M, Macpherson LM. Socioeconomic inequalities and oral cancer risk: a systematic review and meta-analysis of case-control studies. International journal of cancer. 2008;122(12):2811-9.
34. Pruyn JF, de Jong PC, Bosman LJ, van Poppel JW, van Den Borne HW, Ryckman RM, et al. Psychosocial aspects of head and neck cancer--a review of the literature. Clinical otolaryngology and allied sciences. 1986;11(6):469-74.
35. Knox SS, Cancer Cell I. From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. Cancer Cell Int. 2010(1475-2867).
36. Olshan AF, Weissler MC, Watson MA, Bell DA. GSTM1, GSTT1, GSTP1, CYP1A1, and NAT1 Polymorphisms, Tobacco Use, and the Risk of Head and Neck Cancer. 2000:185-91.

37. Hashibe M, Brennan P, Strange RC, Bhisey R, Cascorbi I, Lazarus P, et al. CYP1A1 Genotypes and Risk of Head and Neck Cancer Genotypes and Risk of Head and Neck Cancer. 2003:1509-17.
38. Brennan P, Lewis S, Hashibe M, Bell DA, Boffetta P, Bouchardy C, et al. Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a HuGE review. Am J Epidemiol. 2004;159(1):1-16.
39. Hung RJ, van der Hel O, Tavtigian SV, Brennan P, Boffetta P, Hashibe M. Perspectives on the molecular epidemiology of aerodigestive tract cancers. Mutation research. 2005;592(1-2):102-18.
40. Brunotto M, Zarate AM, Bono A, Barra JL, Berra S. Risk genes in head and neck cancer: a systematic review and meta-analysis of last 5 years. Oral oncology. 2014;50(3):178-88.
41. Canova C, Richiardi L, Merletti F, Pentenero M, Gervasio C, Tanturri G, et al. Alcohol, tobacco and genetic susceptibility in relation to cancers of the upper aerodigestive tract in northern Italy. Tumori. 2010;96(1):1-10.
42. Sreelekha TT, Ramadas K, Pandey M, Thomas G, Nalinakumari KR, Pillai MR. Genetic polymorphism of CYP1A1, GSTM1 and GSTT1 genes in Indian oral cancer. Oral oncology. 2001;37(7):593-8.
43. Vincent-Chong VK, Ismail SM, Rahman Zaa, Sharifah Na, Anwar a, Pradeep PJ, et al. Genome-wide analysis of oral squamous cell carcinomas revealed over expression of ISG15, Nestin and WNT11. Oral diseases. 2012;18(5):469-76.
44. International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Volume 100E. Personal Habits and Indoor Combustions. Lyon, France: International Agency for Research on Cancer. 2009 http://monographs.iarc.fr/ENG/Monographs/vol100E/mono100E.pdf.
45. Heck JE, Berthiller J, Vaccarella S, Winn DM, Smith EM, Shan'gina O, 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. 1 ed. England: Lifestyle, Environment, and Cancer Group, International Agency for Research on Cancer, Lyon, France.; 2010. p. 166-81.
46. Tezal M, Sullivan Ma, Reid ME, Marshall JR, Hyland A, Loree T, et al. Chronic periodontitis and the risk of tongue cancer. Archives of otolaryngology--head & neck surgery. 2007;133(5):450-4.
47. Laprise C, Shahul HP, Madathil SA, Thekkepurakkal AS, Castonguay G, Varghese I, et al. Periodontal diseases and risk of oral cancer in Southern India: Results from the HeNCe Life study. International journal of cancer. 2016;139(7):1512-9.
48. Giovino GA, Henningfield JE, Tomar SL, Escobedo LG, Slade J. Epidemiology of Tobacco Use and Dependence. Epidemiologic Reviews. 1995;17(1):48-65.
49. Report GI. Global Adult Tobacco Survey; India 2009-2010, Ministry of Health & Family Welfare, Government of India, New Delhi. India: International Institute for Population Sciences, Deonar, Mumbai 2010 http://www.searo.who.int/tobacco/documents/2010-pub2.pdf?ua=1.
50. Tobacco use in Canada: Patterns and trends, 2013.
51. International Agency for Research on Cancer.IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Volume 83: Tobacco Smoke and Involuntary Smoking. Lyon, France: International Agency for Research on Cancer. 2004.

52. Hoffmann D, Hoffmann I. Chemistry and toxicology.In: US Department of Health and Human Services. Cigars: health effects and trends (Smoking and Tobacco Control Monograph 9). DHHS (Publ No. NIH 98-4302), 1998:55–104.
53. Malson JL, Sims K, Murty R, Pickworth WB. Comparison of the nicotine content of tobacco used in bidis and conventional cigarettes. Tobacco control. 2001;10(2):181-3.
54. Watson CH, Polzin GM, Calafat AM, Ashley DL. Determination of tar, nicotine, and carbon monoxide yields in the smoke of bidi cigarettes. Nicotine Tob Res. 2003;5(5):747-53.
55. Reid J, Hammond D, Rynard V, Burkhalter R. Tobacco Use in Canada: Patterns and Trends, 2015 Edition. Waterloo, ON: Propel Centre for Population Health Impact, University of Waterloo.; 2015.
56. Pindborg Jj KJGPC, Chawla TN. Studies in oral leukoplakias. Prevalence of leukoplakia among 10 000 persons in Lucknow, India, with special reference to use of tobacco and betel nut. Bull World Health Organ. 1967;37(1):109-16.
57. Balaram P, Sridhar H, Rajkumar T, Vaccarella S, Herrero R, Nandakumar A, et al. Oral cancer in southern India: the influence of smoking, drinking, paan-chewing and oral hygiene. International journal of cancer. 2002;98(3):440-5.
58. Jayalekshmi PA, Gangadharan P, Akiba S, Nair RRK, Tsuji M, Rajan B. Tobacco chewing and female oral cavity cancer risk in Karunagappally cohort, India. British Journal of Cancer. 2009;100(5):848-52.
59. Madathil SA, Rousseau MC, Wynant W, Schlecht NF, Netuveli G, Franco EL, et al. Nonlinear association between betel quid chewing and oral cancer: Implications for prevention. Oral oncology. 2016;60:25-31.
60. D. Hoffmann LDS, Wynder EL. Comparative chemical analysis of indian bidi and American cigarette smoke. International journal of cancer. 1974;14:49-55.
61. Pakhale ea. Chemical analysis of smoke of Indian cigarettes, bidis and other indigenous forms of smoking-levels of steam-volatile phenol, hydrogen cyanide and benzo(a)pyrene. The Indian journal of chest diseases and allied sciences. 1990;32(2).
62. Pakhale SS, Sarkar S, Jayant K, Bhide SV. Carcinogenicity of Indian bidi and cigarette smoke condensate in Swiss albino mice. J Cancer Res Clin Oncol. 1988;114(6):647-9.
63. Iarc LF. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans1985 http://monographs.iarc.fr/ENG/Monographs/vol1-42/.
64. WHO. IARC Monographs on the evaluation of carcinogenic risks to humans- Alcohol drinking IARC, Lyon, France; 1988.
65. Iarc. IARC monographs on evaluation of human carcinogens: Agents classified by IARC Monographs, Volumes1-109.
66. De Stavola BL, Daniel RM. Marginal structural models: the way forward for life-course epidemiology? Epidemiology (Cambridge, Mass). 2012;23(2):233-7.
67. Kumar R, Prakash S, Kushwah AS, Vijayan VK. Breath carbon monoxide concentration in cigarette and bidi smokers in India. Indian J Chest Dis Allied Sci. 2010;52(1):19-24.
68. Freedman ND, Abnet CC, Leitzmann MF, Hollenbeck AR, Schatzkin A. Prospective investigation of the cigarette smoking-head and neck cancer association by sex. Cancer. 2007;110(7):1593-601.
69. CDC. Smoking and tobacco use- The health consequences of Smoking. Center for Disease, Control Prevention; 2004.

70. Wyss A, Hashibe M, Chuang SC, Lee YC, Zhang ZF, Yu GP, 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):679-90.
71. Shiu M, Chen TH, Eur JCP. Impact of betel quid, tobacco and alcohol on three-stage disease natural history of oral leukoplakia and cancer: implication for prevention of oral cancer. Eur J Cancer Prev. 2004;13(1).
72. Muwonge R, Ramadas K, Sankila R, Thara S, Thomas G, Vinoda J, et al. Role of tobacco smoking, chewing and alcohol drinking in the risk of oral cancer in Trivandrum, India: A nested case-control design using incident cancer cases. Oral oncology. 2008;44(5):446-54.
73. Gupta B, Johnson NW. Systematic Review and Meta-Analysis of Association of Smokeless Tobacco and of Betel Quid without Tobacco with Incidence of Oral Cancer in South Asia and the Pacific. PLoS ONE. 2014;9(11):e113385.
74. Awang MN, Singapore Med J. Betel quid and oral carcinogenesis. Singapore Med J. 1988;29:589-93.
75. Balaram P, Sridhar H, Rajkumar T, Vaccarella S, Herrero R, Nandakumar A, et al. Oral cancer in southern India: The influence of smoking, drinking, paan-chewing and oral hygiene. International journal of cancer. 2002;98(3):440-5.
76. Chandra et.al P, Drug Alcohol D. Prevalence and correlates of areca nut use among psychiatric patients in India. Drug and Alcohol Dependence. 2003;69(3):311-6.
77. Sankaranarayanan R, Duffy SW, Nair MK, Padmakumary G, Day NE. Tobacco and alcohol as risk factors in cancer of the larynx in Kerala, India. International journal of cancer. 1990;45(5):879-82.
78. Rahman M, Sakamoto J, Fukui T. Bidi smoking and oral cancer: a meta-analysis. International journal of cancer Journal international du cancer. 2003;106(4):600-4.
79. Abdoul Hossain MMDDB. Risk for oral cancer associated to smoking, smokeless and oral dip products. Indian J Public Health. 2012;Jan-Mar;56(1):57-60.
80. Tobacco chewing and female oral cavity cancer risk in Karunagappally cohort, India. Br J Cancer. 2009;100(5):848-52.
81. Schlecht NF, Franco EL, Pintos J, Kowalski LP. Effect of smoking cessation and tobacco type on the risk of cancers of the upper aero-digestive tract in Brazil. Epidemiology (Cambridge, Mass). 1999;10(4):412-8.
82. Schlecht NF, Franco EL, Pintos J, Negassa A, Kowalski LP, Oliveira BV, et al. Interaction between Tobacco and Alcohol Consumption and the Risk of Cancers of the Upper Aero-Digestive Tract in Brazil. American Journal of Epidemiology. 1999;150(11):1129-37.
83. Znaor A, Brennan P, Gajalakshmi V, Mathew A, Shanta V, Varghese C, et al. Independent and combined effects of tobacco smoking, chewing and alcohol drinking on the risk of oral, pharyngeal and esophageal cancers in Indian men. International journal of cancer Journal international du cancer. 2003;105(5):681-6.
84. Polesel J, Talamini R, La Vecchia C, Levi F, Barzan L, Serraino D, et al. Tobacco smoking and the risk of upper aero-digestive tract cancers: A reanalysis of case-control studies using spline models. International journal of cancer Journal international du cancer. 2008;122(10):2398-402.
85. Guha N, Warnakulasuriya S, Vlaanderen J, Straif K. Betel quid chewing and the risk of oral and oropharyngeal cancers: a meta-analysis with implications for cancer control. International journal of cancer. 2014;135(6):1433-43.

86. Lee CH, Lee KW, Fang FM, Wu DC, Tsai SM, Chen PH, et al. The neoplastic impact of tobacco-free betel-quid on the histological type and the anatomical site of aerodigestive tract cancers. International journal of cancer. 2012;131(5):E733-43.

87. Blot WJ, McLaughlin JK, Winn DM, Austin DF, Greenberg RS, Preston-Martin S, et al. Smoking and drinking in relation to oral and pharyngeal cancer. Cancer Res. 1988;48(11):3282-7.

88. Lewin F, Norell SE, Johansson H, Gustavsson P, Wennerberg J, Biorklund A, et al. Smoking tobacco, oral snuff, and alcohol in the etiology of squamous cell carcinoma of the head and neck: a population-based case-referent study in Sweden. Cancer. 1998;82(7):1367-75.

89. Castellsague et.al X, Int JC. The role of type of tobacco and type of alcoholic beverage in oral carcinogenesis. International journal of cancer. 2004;108(714-749).

90. Bosetti C, Gallus S, Peto R, Negri E, Talamini R, Tavani A, et al. Tobacco smoking, smoking cessation, and cumulative risk of upper aerodigestive tract cancers. Am J Epidemiol. 2008;167(4):468-73.

91. WHO. Global status report on alcohol and health. Switzerland; 2011.

92. Canada H. Canadian Alcohol and Drug Use Monitoring Survey Canada2011 [updated 2012-08-02. Available from: http://www.hc-sc.gc.ca/hc-ps/drugs-drogues/stat/_2011/tables-tableaux-eng.php#t10.

93. Prasad R. Alcohol use on the rise in India. The Lancet.373(9657):17-8.

94. Economic review 2012- State planning board, Trivandrum, India, http://spb.kerala.gov.in/~spbuser/images/pdf/er12/Chapter4/chapter04.html. 2012.

95. Weisburger JH, Wynder EL. The role of genotoxic carcinogens and of promoters in carcinogenesis and in human cancer causation. Acta Pharmacol Toxicol (Copenh). 1984;55 Suppl 2:53-68.

96. Poschl G, Pöschl G, Seitz HK. Alcohol and Cancer. Alcohol and Alcoholism. 2004;39(3):155-65.

97. Seitz HK, Stickel F, Homann N. Pathogenetic mechanisms of upper aerodigestive tract cancer in alcoholics. International journal of cancer. 2004;108(4):483-7.

98. Doll R, Forman D, Vecchia Cl, Woutersen R. Alcoholic beverages and cancers of the digestive tract and larynx. Health issues related to alcohol consumption. 2nd ed. Brussel: Oxford : Blacwell Science; 1999. p. 351-93.

99. Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, Bouvard V, et al. Carcinogenicity of alcoholic beverages. The Lancet Oncology. 2007;8(4):292-3.

100. Secretan B, Straif K, Baan R, Grosse Y, El Ghissassi F, Bouvard V, et al. A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. The Lancet Oncology. 2009;10(11):1033-4.

101. Boffetta P, Hashibe M. Alcohol and cancer. The Lancet Oncology. 2006;7(2):149-56.

102. Chang JS, Straif K, Guha N. The role of alcohol dehydrogenase genes in head and neck cancers: a systematic review and meta-analysis of ADH1B and ADH1C. Mutagenesis. 2012;27(3):275-86.

103. Connor J. Alcohol consumption as a cause of cancer. Addiction. 2017;112(2):222-8.

104. Turati F, Garavello W, Tramacere I, Pelucchi C, Galeone C, Bagnardi V, et al. A meta-analysis of alcohol drinking and oral and pharyngeal cancers: results from subgroup analyses. Alcohol and alcoholism (Oxford, Oxfordshire). 2012;48(1):107-18.

105. Talamini R, La Vecchia C, Levi F, Conti E, Favero A, Franceschi S. Cancer of the oral cavity and pharynx in nonsmokers who drink alcohol and in nondrinkers who smoke tobacco. J Natl Cancer Inst. 1998;90(24):1901-3.
106. Schlecht NF, Pintos J, Kowalski LP, Franco EL. Effect of type of alcoholic beverage on the risks of upper aerodigestive tract cancers in Brazil. Cancer Causes Control. 2001;12(7):579-87.
107. Bagnardi V, Blangiardo M, Vecchia CL, Corrao G. A meta-analysis of alcohol drinking and cancer risk. Br J Cancer. 2001;85(11):1700-5.
108. Hashibe M, Brennan P, Chuang SC, Boccia S, Castellsague X, Chen C, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiol Biomarkers Prev. 2009;18(2):541-50.
109. Maasland DH, van den Brandt PA, Kremer B, Goldbohm RA, Schouten LJ. Alcohol consumption, cigarette smoking and the risk of subtypes of head-neck cancer: results from the Netherlands Cohort Study. BMC Cancer. 2014;14(1471-2407 (Electronic)):187.
110. Polesel J, Dal Maso L, Bagnardi V, Zucchetto A, Zambon A, Levi F, et al. Estimating dose-response relationship between ethanol and risk of cancer using regression spline models. International journal of cancer. 2005;114(5):836-41.
111. Dal Maso L, Torelli N, Biancotto E, Di Maso M, Gini A, Franchin G, et al. Combined effect of tobacco smoking and alcohol drinking in the risk of head and neck cancers: a re-analysis of case-control studies using bi-dimensional spline models. Eur J Epidemiol. 2016;31(4):385-93.
112. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. Br J Cancer. 2015;112(3):580-93.
113. Health sector: Chapter 4: Government of Kerala Vision 2030. RCC, TVM; 2011 http://www.indiaenvironmentportal.org.in/files/file/Kerala%20Perspective%20Plan%202030.pdf 2011.
114. Lubin JH, Purdue M, Kelsey K, Zhang ZF, Winn D, Wei Q, et al. Total exposure and exposure rate effects for alcohol and smoking and risk of head and neck cancer: a pooled analysis of case-control studies. Am J Epidemiol. 2009;170(8):937-47.
115. Hsu TC, Spitz MR, Schantz SP. Mutagen sensitivity: a biological marker of cancer susceptibility. Cancer Epidemiology Biomarkers &amp;amp; Prevention. 1991;1(1):83.
116. Ho T, Wei Q, Sturgis EM. Epidemiology of carcinogen metabolism genes and risk of squamous cell carcinoma of the head and neck. Head Neck. 2007;29(7):682-99.
117. Harris CC, Mulvihill JJ, Thorgeirsson SS, Minna JD. Individual differences in cancer susceptibility. Ann Intern Med. 1980;92(6):809-25.
118. Luch A. Nature and nurture - lessons from chemical carcinogenesis. Nat Rev Cancer. 2005;5(2):113-25.
119. Perera FP. Molecular Epidemiology: Insights Into Cancer Susceptibility, Risk Assessment, and Prevention. Journal of the National Cancer Institute. 1996;88(8):496-509.
120. Friedlander PL. Genomic instability in head and neck cancer patients. Head Neck. 2001;23(8):683-91.
121. Sturgis EM, Wei Q. Genetic susceptibility--molecular epidemiology of head and neck cancer. Curr Opin Oncol. 2002;14(3):310-7.

122. Chen YC, Hunter DJ. Molecular epidemiology of cancer. CA Cancer J Clin. 2005;55(1):45-54.

123. He Y, Hoskins JM, McLeod HL. Copy Number Variants in pharmacogenetic genes. Trends in molecular medicine. 2011;17(5):244-51.

124. Hecht SS, Hoffmann D. Tobacco-specific nitrosamines, an important group of carcinogens in tobacco and tobacco smoke. Carcinogenesis. 1988;9(6):875-84.

125. Costa LED. Gene–Environment Interactions: Fundamentals of Ecogenetics. National Institute of Environmental Health Science; 2006. p. A382-A.

126. Hunter DJ. Gene-environment interactions in human diseases. Nature reviews Genetics. 2005;6(4):287-98.

127. Singh MS, Michael M. Role of xenobiotic metabolic enzymes in cancer epidemiology. Methods Mol Biol. 2009;472:243-64.

128. Ingelman-Sundberg M, Oscarson M, McLellan RA. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends in Pharmacological Sciences. 1999;20(8):342-9.

129. Dong LM, Potter JD, White E, Ulrich CM, Cardon LR, Peters U. CLINICIAN ' S CORNER The Role of Polymorphisms in Candidate Genes. 2013;299(20):2423-36.

130. Ma Q, Lu AYH. CYP1A Induction and Human Risk Assessment: An Evolving Tale of in Vitro and in Vivo Studies. Drug Metabolism and Disposition. 2007;35(7):1009-16.

131. Le Gal A, Dreano Y, Lucas D, Berthou F. Diversity of selective environmental substrates for human cytochrome P450 2A6: alkoxyethers, nicotine, coumarin, N-nitrosodiethylamine, and N-nitrosobenzylmethylamine. Toxicol Lett. 2003;144(1):77-91.

132. Bozina N, Bradamante V, Lovric M. Genetic polymorphism of metabolic enzymes P450 (CYP) as a susceptibility factor for drug response, toxicity, and cancer risk. Arh Hig Rada Toksikol. 2009;60(2):217-42.

133. Sim SC, Ingelman-Sundberg M. The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Human genomics. 2010;4(4):278-81.

134. Costa CG, Eaton DL. Chapter 1. Introduction; Gene-Environment Interactions: Fundamentals of Ecogenetics2006. 2-6 p.

135. Qin J, Zhang JX, Li XP, Wu BQ, Chen GB, He XF. Association between the CYP1A1 A2455G polymorphism and risk of cancer: evidence from 272 case-control studies. Tumour Biol. 2014;35(4):3363-76.

136. He X-F, Wei W, Liu Z-Z, Shen X-L, Yang X-B, Wang S-L, et al. Association between the CYP1A1 T3801C polymorphism and risk of cancer: Evidence from 268 case–control studies. Gene. 2014;534(2):324-44.

137. Lu D, Yu X, Du Y. Meta-analyses of the effect of cytochrome P450 2E1 gene polymorphism on the risk of head and neck cancer. Mol Biol Rep. 2011;38(4):2409-16.

138. Tang K, Li Y, Zhang Z, Gu Y, Xiong Y, Feng G, et al. The PstI/RsaI and DraI polymorphisms of CYP2E1 and head and neck cancer risk: a meta-analysis based on 21 case-control studies. BMC Cancer. 2010;10:575.

139. Hashibe M, Boffetta P, Zaridze D, Shangina O, Szeszenia-Dabrowska N, Mates D, et al. Evidence for an important role of alcohol- and aldehyde-metabolizing genes in cancers of the upper aerodigestive tract. Cancer Epidemiol Biomarkers Prev. 2006;15(4):696-703.

140. Lang J, Song X, Cheng J, Zhao S, Fan J. Association of GSTP1 Ile105Val polymorphism and risk of head and neck cancers: a meta-analysis of 28 case-control studies. PLoS One. 2012;7(11):e48132.
141. Shukla P, Gupta D, Pant MC, Parmar D. CYP 2D6 polymorphism: a predictor of susceptibility and response to chemoradiotherapy in head and neck cancer. J Cancer Res Ther. 2012;8(1):40-5.
142. Tripathy CB, Roy N. Meta-analysis of glutathione S-transferase M1 genotype and risk toward head and neck cancer. Head & neck. 2006;28(3):217-24.
143. San Jose C, Cabanillas A, Benitez J, Carrillo JA, Jimenez M, Gervasini G. CYP1A1 gene polymorphisms increase lung cancer risk in a high-incidence region of Spain: a case control study. BMC Cancer. 2010;10:463.
144. Rojas M, Cascorbi I, Alexandrov K, Kriek E, Auburtin G, Mayer L, et al. Modulation of benzo[a]pyrene diolepoxide-DNA adduct levels in human white blood cells by CYP1A1, GSTM1 and GSTT1 polymorphism. Carcinogenesis. 2000;21(1):35-41.
145. Hashibe M, Brennan P, Strange RC, Bhisey R, Cascorbi I, Lazarus P, et al. Meta- and pooled analyses of GSTM1, GSTT1, GSTP1, and CYP1A1 genotypes and risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev. 2003;12(12):1509-17.
146. Garte S, Gaspari L, Alexandrie AK, Ambrosone C, Autrup H, Autrup JL, et al. Metabolic gene polymorphism frequencies in control populations. Cancer Epidemiol Biomarkers Prev. 2001;10(12):1239-48.
147. Zhang MX, Liu K, Wang FG, Wen XW, Song XL. Association between CYP2E1 polymorphisms and risk of gastric cancer: An updated meta-analysis of 32 case-control studies. Mol Clin Oncol. 2016;4(6):1031-8.
148. Boccia S, Cadoni G, Sayed-Tabatabaei FA, Volante M, Arzani D, De Lauretis A, et al. CYP1A1, CYP2E1, GSTM1, GSTT1, EPHX1 exons 3 and 4, and NAT2 polymorphisms, smoking, consumption of alcohol and fruit and vegetables and risk of head and neck cancer. J Cancer Res Clin Oncol. 2008;134(1):93-100.
149. Yao K, Qin H, Gong L, Zhang R, Li L. CYP2E1 polymorphisms and nasopharyngeal carcinoma risk: a meta-analysis. Eur Arch Otorhinolaryngol. 2016;274(1):253-9.
150. Ruwali M, Khan AJ, Shah PP, Singh AP, Pant MC, Parmar D. Cytochrome P450 2E1 and head and neck cancer: interaction with genetic and environmental risk factors. Environ Mol Mutagen. 2009;50(6):473-82.
151. Buchard A, Sanchez JJ, Dalhoff K, Morling N. Multiplex PCR detection of GSTM1, GSTT1, and GSTP1 gene variants: simultaneously detecting GSTM1 and GSTT1 gene copy number and the allelic status of the GSTP1 Ile105Val genetic variant. The Journal of molecular diagnostics : JMD. 2007;9(5):612-7.
152. Ryberg D, Skaug V, Hewer A, Phillips DH, Harries LW, Wolf CR, et al. Genotypes of glutathione transferase M1 and P1 and their significance for lung DNA adduct levels and cancer risk. Carcinogenesis. 1997;18(7):1285-9.
153. Hoskins JM, Carey LA, McLeod HL. CYP2D6 and tamoxifen: DNA matters in breast cancer. Nat Rev Cancer. 2009;9(8):576-86.
154. Zhuo W, Wang Y, Zhuo X, Zhu Y, Wang W, Zhu B, et al. CYP1A1 and GSTM1 polymorphisms and oral cancer risk: association studies via evidence-based meta-analyses. Cancer investigation. 2009;27(1):86-95.

155. Agundez JA, Gallardo L, Ledesma MC, Lozano L, Rodriguez-Lescure A, Pontes JC, et al. Functionally active duplications of the CYP2D6 gene are more prevalent among larynx and lung cancer patients. Oncology. 2001;61(1):59-63.

156. Yadav SS, Ruwali M, Pant MC, Shukla P, Singh RL, Parmar D. Interaction of drug metabolizing cytochrome P450 2D6 poor metabolizers with cytochrome P450 2C9 and 2C19 genotypes modify the susceptibility to head and neck cancer and treatment response. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2010;684(1–2):49-55.

157. Tripathy CB, Roy N. Meta-analysis of glutathione S-transferase M1 genotype and risk toward head and neck cancer. Head Neck. 2006;28(3):217-24.

158. Huang RS, Chen P, Wisel S, Duan S, Zhang W, Cook EH, et al. Population-specific GSTM1 copy number variation. Human molecular genetics. 2009;18(2):366-72.

159. Munafò MR, Clark TG, Johnstone EC, Murphy MFG, Walton RT. The genetic basis for smoking behavior: A systematic review and meta-analysis. Nicotine & Tobacco Research. 2004;6(4):583-97.

160. Pan L, Yang X, Li S, Jia C. Association of CYP2A6 gene polymorphisms with cigarette consumption: a meta-analysis. Drug Alcohol Depend. 2015;149:268-71.

161. Li D, Zhao H, Gelernter J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. Biol Psychiatry. 2011;70(6):504-12.

162. Macgregor S, Lind PA, Bucholz KK, Hansell NK, Madden PA, Richter MM, et al. Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. Human molecular genetics. 2009;18(3):580-93.

163. Khlifi R, Messaoud O, Rebai A, Hamza-Chaffai A. Polymorphisms in the Human Cytochrome P450 and Arylamine N-Acetyltransferase: Susceptibility to Head and Neck Cancers. BioMed Research International. 2013;2013:20.

164. Hashibe M, Brennan P, Strange RC, Bhisey R, Cascorbi I, Lazarus P, et al. Meta- and Pooled Analysis of GSTM1, GSTT1, GSTP1 and CYP1A1 Genotypes and Risk of Head and Neck Cancer. 2003:1509-17.

165. Liu L, Wu G, Xue F, Li Y, Shi J, Han J. Functional CYP1A1 genetic variants , alone and in combination with smoking , contribute to development of head and neck cancers. European Journal of Cancer. 2013;49(9):2143-51.

166. Olivieri EHR, da Silva SD, Mendonça FF, Urata YN, Vidal DO, Faria MDAM, et al. CYP1A2*1C, CYP2E1*5B, and GSTM1 polymorphisms are predictors of risk and poor outcome in head and neck squamous cell carcinoma patients. Oral oncology. 2009;45(9):e73-9.

167. Anantharaman D, Chaubal PM, Kannan S, Bhisey Ra, Mahimkar MB. Susceptibility to oral cancer by genetic polymorphisms at CYP1A1, GSTM1 and GSTT1 loci among Indians: tobacco exposure as a risk modulator. Carcinogenesis. 2007;28(7):1455-62.

168. Chatterjee S, Dhar S, Sengupta B, Ghosh A, De M, Roy S, et al. Polymorphisms of CYP1A1, GSTM1 and GSTT1 Loci as the Genetic Predispositions of Oral Cancers and Other Oral Pathologies: Tobacco and Alcohol as Risk Modifiers. Indian journal of clinical biochemistry : IJCB. 2010;25(3):260-72.

169. Varela-Lema L, Taioli E, Ruano-Ravina A, Barros-Dios JM, Benhamou S, Bhisey RA, et al. Meta- and pooled analysis of GSTM1 and CYP1A1 polymorphisms and oropharyngeal

cancer: a HuGE-GSEC review. Genetics in medicine : official journal of the American College of Medical Genetics. 2008;10(6):369-84.

170. Wang Y, Yang H, Duan G, Wang H. The association of the CYP1A1 Ile462Val polymorphism with head and neck cancer risk: evidence based on a cumulative meta-analysis. OncoTargets and therapy. 2016;9:2927-34.

171. Xie S, Luo C, Shan X, Zhao S, He J, Cai Z. CYP1A1 MspI polymorphism and the risk of oral squamous cell carcinoma: Evidence from a meta-analysis. Mol Clin Oncol. 2016;4(4):660-6.

172. Howard LA, Micu AL, Sellers EM, Tyndale RF. Low doses of nicotine and ethanol induce CYP2E1 and chlorzoxazone metabolism in rat liver. J Pharmacol Exp Ther. 2001;299(2):542-50.

173. Fu P, Yang F, Li B, Zhang B, Guan L, Sheng J, et al. Meta-analysis of CYP2E1 polymorphisms in liver carcinogenesis. Dig Liver Dis. 2017;49(1):77-83.

174. Boccia S, Cadoni G, Sayed-Tabatabaei FA, Volante M, Arzani D, De Lauretis A, et al. CYP1A1, CYP2E1, GSTM1, GSTT1, EPHX1 exons 3 and 4, and NAT2 polymorphisms, smoking, consumption of alcohol and fruit and vegetables and risk of head and neck cancer. J Cancer Res Clin Oncol. 2008;134.

175. Soya SS, Vinod T, Reddy KS, Gopalakrishnan S, Adithan C. CYP2E1 polymorphisms and gene-environment interactions in the risk of upper aerodigestive tract cancers among Indians. Pharmacogenomics. 2008;9(5):551-60.

176. Gattá GJF, de Carvalho MB, Siraque MS. Genetic polymorphisms of CYP1A1, CYP2E1, GSTM1, and GSTT1 associated with head and neck cancer. Head & Neck. 2006;28.

177. Yamazaki H, Inui Y, Yun CH, Guengerich FP, Shimada T. Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes. Carcinogenesis. 1992;13(10):1789-94.

178. Hung HC, Chuang J, Chien YC, Hildesheim A. Genetic polymorphisms of CYP2E1, GSTM1, and GSTT1;environmental factors and risk of oral cancer. 1997:901-5.

179. Maurya SS, Anand G, Dhawan A, Khan AJ, Jain SK, Pant MC, et al. Polymorphisms in drug-metabolizing enzymes and risk to head and neck cancer: evidence for gene-gene and gene-environment interaction. Environ Mol Mutagen. 2014;55(2):134-44.

180. Zhuo X, Song J, Liao J, Zhou W, Ye H, Li Q, et al. Does CYP2E1 RsaI/PstI polymorphism confer head and neck carcinoma susceptibility?: A meta-analysis based on 43 studies. Medicine (Baltimore). 2016;95(43):e5156.

181. Tang K, Li Y, Zhang Z, Gu Y, Xiong Y, Feng G, et al. The PstI/RsaI and DraI polymorphisms of CYP2E1and head and neck cancer risk: a meta-analysis based on 21 case-control studies. BMC Cancer. 2010;10(1):575.

182. Cury NM, Russo A, Galbiatti AL, Ruiz MT, Raposo LS, Maniglia JV, et al. Polymorphisms of the CYP1A1 and CYP2E1 genes in head and neck squamous cell carcinoma risk. Mol Biol Rep. 2012;39(2):1055-63.

183. Mulder TPJ, Manni JJ, Roelofs HMJ, Peters WHM, Wiersma A. Glutathione S-transferases and glutathione in human head and neck cancer. Carcinogenesis. 1995;16(3):619-24.

184. Jourenkova-Mironova N, Voho A, Bouchardy C, Wikman H, Dayer P, Benhamouand S, et al. Glutathione S-transferase GSTM1, GSTM3, GSTP1 and GSTT1 genotypes and the risk of smoking-related oral and pharyngeal cancers. International journal of cancer. 1999;81(1):44-8.

185. Hezova R, Bienertova-Vasku J, Sachlova M, Brezkova V, Vasku A, Svoboda M, et al. Common polymorphisms in GSTM1, GSTT1, GSTP1, GSTA1 and susceptibility to colorectal cancer in the Central European population. European Journal of Medical Research. 2012;17(1):17.

186. Agúndez JAG, García-Martín E, Martínez C, Benito-León J, Millán-Pascual J, Díaz-Sánchez M, et al. The GSTP1 gene variant rs1695 is not associated with an increased risk of multiple sclerosis. Cellular and Molecular Immunology. 2015;12(6):777-9.

187. Zhang Z-j, Hao K, Shi R, Zhao G, Jiang G-x, Song Y, et al. Glutathione S-transferase M1 (GSTM1) and glutathione S-transferase T1 (GSTT1) null polymorphisms, smoking, and their interaction in oral cancer: a HuGE review and meta-analysis. American journal of epidemiology. 2011;173(8):847-57.

188. Hu X, Xia H, Srivastava SK, Herzog C, Awasthi YC, Ji X, et al. Activity of four allelic forms of glutathione S-transferase hGSTP1-1 for diol epoxides of polycyclic aromatic hydrocarbons. Biochem Biophys Res Commun. 1997;238(2):397-402.

189. Saarikoski ST, Voho A, Reinikainen M, Anttila S, Karjalainen A, Malaveille C, et al. Combined effect of polymorphic GST genes on individual susceptibility to lung cancer. International journal of cancer. 1998;77(4):516-21.

190. Ruwali M, Pant MC, Shah PP, Mishra BN, Parmar D. Polymorphism in cytochrome P450 2A6 and glutathione S-transferase P1 modifies head and neck cancer risk and treatment outcome. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2009;669(1–2):36-41.

191. Singh M, Shah PP, Singh AP, Ruwali M, Mathur N, Pant MC, et al. Association of genetic polymorphisms in glutathione S-transferases and susceptibility to head and neck cancer. 2008;638:184-94.

192. Ruzzo A, Canestrari E, Maltese P, Pizzagalli F, Graziano F, Santini D, et al. Polymorphisms in genes involved in DNA repair and metabolism of xenobiotics in individual susceptibility to sporadic diffuse gastric cancer. Clinical Chemical Laboratory Medicine2007. p. 822.

193. Vlaykova T, Miteva L, Gulubova M, Stanilova S. Ile105Val GSTP1 polymorphism and susceptibility to colorectal carcinoma in Bulgarian population. Int J Colorectal Dis. 2007;22(10):1209-15.

194. Jiao L, Bondy ML, Hassan MM, Chang DZ, Abbruzzese JL, Evans DB, et al. Glutathione S-transferase Gene Polymorphisms and Risk and Survival of Pancreatic Cancer. Cancer. 2007;109(5):840-8.

195. Chen JB, Wang F, Wu JJ, Cai M. Glutathione S-transferase pi polymorphism contributes to the treatment outcomes of advanced non-small cell lung cancer patients in a Chinese population. Genet Mol Res. 2016;15(3).

196. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature reviews Genetics. 2006;7(2):85-97.

197. CYP2D6 allele nomenclature [Internet]. Available from: http://www.cypalleles.ki.se/cyp2d6.htm.

198. Neafsey P, Ginsberg G, Hattis D, Sonawane B. Genetic Polymorphism in Cytochrome P450 2D6 (CYP2D6): Population Distribution of CYP2D6 Activity. Journal of Toxicology and Environmental Health, Part B. 2009;12(5-6):334-61.

199. Solus JF, Arietta BJ, Harris JR, Sexton DP, Steward JQ, McMunn C, et al. Genetic variation in eleven phase I drug metabolism genes in an ethnically diverse population. Pharmacogenomics. 2004;5(7):895-931.

200. Carol B, Suzanne C. CYP2D6 and smoking behaviour 1997.pdf. Pharmacogenetics (1997). 1997;7:411-4.

201. Gajecka M, Rydzanicz M, Jaskula-Sztul R, Kujawski M, Szyfter W, Szyfter K. CYP1A1, CYP2D6, CYP2E1, NAT2, GSTM1 and GSTT1 polymorphisms or their combinations are associated with the increased risk of the laryngeal squamous cell carcinoma. Mutat Res. 2005;574.

202. Sundberg K, Dreij K, Seidel A, Jernström B. Glutathione Conjugation and DNA Adduct Formation of Dibenzo[a,l]pyrene and Benzo[a]pyrene Diol Epoxides in V79 Cells Stably Expressing Different Human Glutathione Transferases. Chemical Research in Toxicology. 2002;15(2):170-9.

203. Hiyama T, Yoshihara M, Tanaka S, Chayama K. Genetic polymorphisms and head and neck cancer risk (Review). 2008:945-73.

204. Zhuo W, Wang Y, Zhuo X, Zhu Y, Wang W, Zhu B, et al. CYP1A1 and GSTM1 polymorphisms and oral cancer risk: association studies via evidence-based meta-analyses. Cancer Invest. 2009;27(1):86-95.

205. Suzen HS, Guvenc G, Turanli M, Comert E, Duydu Y, Elhan A. The role of GSTM1 and GSTT1 polymorphisms in head and neck cancer risk. Oncol Res. 2007;16(9):423-9.

206. Choudhury JH, Singh SA, Kundu S, Choudhury B, Talukdar FR, Srivasta S, et al. Tobacco carcinogen-metabolizing genes CYP1A1, GSTM1, and GSTT1 polymorphisms and their interaction with tobacco exposure influence the risk of head and neck cancer in Northeast Indian population. Tumour Biol. 2015;36(8):5773-83.

207. Zhang X, Huang M, Wu X, Kadlubar S, Lin J, Yu X, et al. GSTM1 copy number and promoter haplotype as predictors for risk of recurrence and/or second primary tumor in patients with head and neck cancer. Pharmacogenomics and Personalized Medicine. 2013;6:9-17.

208. Sharma R, Ahuja M, Panda NK, Khullar M. Interactions among genetic variants in tobacco metabolizing genes and smoking are associated with head and neck cancer susceptibility in North Indians. DNA Cell Biol. 2011;30(8):611-6.

209. Zhang X, Lin J, Wu X, Lin Z, Ning B, Kadlubar S, et al. Association between GSTM1 copy number, promoter variants and susceptibility to urinary bladder cancer. International Journal of Molecular Epidemiology and Genetics. 2012;3(3):228-36.

210. Emeville E, Broquere C, Brureau L, Ferdinand S, Blanchet P, Multigner L, et al. Copy number variation of GSTT1 and GSTM1 and the risk of prostate cancer in a Caribbean population of African descent. PLoS One. 2014;9(9):e107275.

211. Flay BR. Youth tobacco use: risk patterns, and control J. Slade, C.T. Orleans (Eds.), Nicotine Addiction: Principles and Management, Oxford University Press, New York 1993:653–61.

212. Tyas SL, Pederson LL. Psychosocial factors related to adolescent smoking: a critical review of the literature. Tobacco Control. 1998;7(4):409-20.

213. Carter B, Long T, Cinciripini P. A meta-analytic review of the CYP2A6 genotype and smoking behavior. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2004;6(2):221-7.

214. Mayhew KP, Flay BR, Mott JA. Stages in the development of adolescent smoking. Drug and Alcohol Dependence. 2000;59, Supplement 1:61-81.

215. Benowitz NL, Hukkanen J, Jacob P. Nicotine Chemistry, Metabolism, Kinetics and Biomarkers. Handbook of experimental pharmacology. 2009(192):29-60.

216. Messina ES, Tyndale RF, Sellers EM. A Major Role for CYP2A6 in Nicotine C-Oxidation by Human Liver Microsomes. Journal of Pharmacology and Experimental Therapeutics. 1997;282(3):1608.

217. Raunio H, Rautio A, Gullstén H, Pelkonen O. Polymorphisms of CYP2A6 and its practical consequences. British Journal of Clinical Pharmacology. 2001;52(4):357-63.

218. Malaiyandi V, Sellers EM, Tyndale RF. Implications of CYP2A6 Genetic Variation for Smoking Behaviors and Nicotine Dependence. Clinical Pharmacology & Therapeutics. 2005;77(3):145-58.

219. Benowitz NL, Swan GE, Jacob P, Lessov-Schlaggar CN, Tyndale RF. CYP2A6 genotype and the metabolism and disposition kinetics of nicotine. Clinical Pharmacology & Therapeutics. 2006;80(5):457-67.

220. Nakajima M, Fukami T, Yamanaka H, Higashi E, Sakai H, Yoshida R, et al. Comprehensive evaluation of variability in nicotine metabolism and CYP2A6 polymorphic alleles in four ethnic populations. Clinical Pharmacology & Therapeutics. 2006;80(3):282-97.

221. Ho MK. Impact of CYP2A6 genetic variation on nicotine metabolism and smoking behaviours in light smoking populations of Black-African descent. Toronto: University of Toronto; 2011.

222. Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking. Nature. 1998;393(6687):750.

223. Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, Sellers EM, et al. Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol Pharmacol. 2000;58(4):747-55.

224. Schoedel Ka, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF. Ethnic variation in CYP2A6 and association of genetically slow nicotine metabolism and smoking in adult Caucasians. Pharmacogenetics. 2004;14(9):615-26.

225. Sabol SZ, Hamer DH. An Improved Assay Shows No Association Between the CYP2A6 Gene and Cigarette Smoking Behavior. Behavior Genetics. 1999;29(4):257-61.

226. Tiihonen J, Pesonen U, Kauhanen J, Koulu M, Hallikainen T, Leskinen L, et al. CYP2A6 genotype and smoking. Molecular psychiatry. 2000;5(4):347-8.

227. Loriot MA, Rebuissou S, Oscarson M, Cenee S, Miyamoto M, Ariyoshi N, et al. Genetic polymorphisms of cytochrome P450 2A6 in a case-control study on lung cancer in a French population. Pharmacogenetics. 2001;11(1):39-44.

228. Oscarson M, McLellan RA, Gullsten H, Yue QY, Lang MA, Bernal ML, et al. Characterisation and PCR-based detection of a CYP2A6 gene deletion found at a high frequency in a Chinese population. FEBS Lett. 1999;448(1):105-10.

229. Strasser AA, Malaiyandi V, Hoffmann E, Tyndale RF, Lerman C. An association of CYP2A6 genotype and smoking topography. Nicotine Tob Res. 2007;9(4):511-8.

230. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010;42(5):448-53.

231. Canova C, Hashibe M, Simonato L, Nelis M, Metspalu A, Lagiou P, et al. Genetic Associations of 115 Polymorphisms with Cancers of the Upper Aerodigestive Tract across 10 European Countries: The ARCAGE Project. Cancer Research. 2009;69(7):2956.

232. Kamataki T, Fujieda M, Kiyotani K, Iwano S, Kunitoh H. Genetic polymorphism of CYP2A6 as one of the potential determinants of tobacco-related cancer risk. Biochemical and Biophysical Research Communications. 2005;338(1):306-10.

233. Bosron WF, Li T-K. Genetic polymorphism of human liver alcohol and aldehyde dehydrogenases, and their relationship to alcohol metabolism and alcoholism. Hepatology. 1986;6(3):502-10.
234. Meyers JL, Dick DM. Genetic and Environmental Risk Factors for Adolescent-Onset Substance Use Disorders. Child and adolescent psychiatric clinics of North America. 2010;19(3):465-77.
235. Bierut Laura J. Genetic Vulnerability and Susceptibility to Substance Dependence. Neuron. 2011;69(4):618-27.
236. Bierut LJ, Goate AM, Breslau N, Johnson EO, Bertelsen S, Fox L, et al. ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. Molecular psychiatry. 2012;17(4):445-50.
237. Hashibe M, McKay JD, Curado MP, Oliveira JC, Koifman S, Koifman R, et al. Multiple ADH genes are associated with upper aerodigestive cancers. Nat Genet. 2008;40(6):707-9.
238. Borras E, Coutelle C, Rosell A, Fernandez-Muixi F, Broch M, Crosas B, et al. Genetic polymorphism of alcohol dehydrogenase in europeans: the ADH2*2 allele decreases the risk for alcoholism and is associated with ADH3*1. Hepatology. 2000;31(4):984-9.
239. Asakage T, Yokoyama A, Haneda T, Yamazaki M, Muto M, Yokoyama T, et al. Genetic polymorphisms of alcohol and aldehyde dehydrogenases, and drinking, smoking and diet in Japanese men with oral and pharyngeal squamous cell carcinoma. Carcinogenesis. 2006;28(4):865-74.
240. Wall TL. Genetic associations of alcohol and aldehyde dehydrogenase with alcohol dependence and their mechanisms of action. Ther Drug Monit. 2005;27(6):700-3.
241. Chen WJ, Loh EW, Hsu YP, Chen CC, Yu JM, Cheng AT. Alcohol-metabolising genes and alcoholism among Taiwanese Han men: independent effect of ADH2, ADH3 and ALDH2. Br J Psychiatry. 1996;168(6):762-7.
242. Osier M, Pakstis AJ, Kidd JR, Lee JF, Yin SJ, Ko HC, et al. Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am J Hum Genet. 1999;64(4):1147-57.
243. Chen CC, Lu RB, Chen YC, Wang MF, Chang YC, Li TK, et al. Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. American Journal of Human Genetics. 1999;65(3):795-807.
244. Thomasson HR, Edenberg HJ, Crabb DW, Mai XL, Jerome RE, Li TK, et al. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. American Journal of Human Genetics. 1991;48(4):677-81.
245. Whitfield JB. Alcohol Dehydrogenase and Alcohol Dependence: Variation in Genotype-Associated Risk between Populations. American Journal of Human Genetics. 2002;71(5):1247-50.
246. Edenberg HJ, Xuei X, Chen HJ, Tian H, Wetherill LF, Dick DM, et al. Association of alcohol dehydrogenase genes with alcohol dependence: a comprehensive analysis. Human molecular genetics. 2006;15(9):1539-49.
247. Yokoyama A, Muramatsu T, Omori T, Yokoyama T, Matsushita S, Higuchi S, et al. Alcohol and aldehyde dehydrogenase gene polymorphisms and oropharyngolaryngeal, esophageal and stomach cancers in Japanese alcoholics. Carcinogenesis. 2001;22(3):433-9.
248. Hakenewerth AM, Millikan RC, Rusyn I, Herring AH, North KE, Barnholtz-Sloan JS, et al. Joint effects of alcohol consumption and polymorphisms in alcohol and oxidative stress metabolism genes on risk of head and neck cancer. Cancer Epidemiol Biomarkers Prev. 2011;20(11):2438-49.

249. Guo H, Zhang G, Mai R. Alcohol Dehydrogenase-1B Arg47His Polymorphism and Upper Aerodigestive Tract Cancer Risk: A Meta-Analysis Including 24,252 Subjects. Alcoholism: Clinical and Experimental Research. 2012;36(2):272-8.

250. Hiraki A, Matsuo K, Wakai K, Suzuki T, Hasegawa Y, Tajima K. Gene–gene and gene–environment interactions between alcohol drinking habit and polymorphisms in alcohol-metabolizing enzyme genes and the risk of head and neck cancer in Japan. Cancer Science. 2007;98(7):1087-91.

251. Ji YB, Lee SH, Kim KR, Park CW, Song CM, Park BL, et al. Association between ADH1B and ADH1C polymorphisms and the risk of head and neck squamous cell carcinoma. Tumor Biology. 2015;36(6):4387-96.

252. Garcia SM, Curioni OA, de Carvalho MB, Gattas GJ. Polymorphisms in alcohol metabolizing genes and the risk of head and neck cancer in a Brazilian population. Alcohol Alcohol. 2010;45(1):6-12.

253. Dong Y-J, Peng T-K, Yin S-J. Expression and activities of class IV alcohol dehydrogenase and class III aldehyde dehydrogenase in human mouth. Alcohol. 1996;13(3):257-62.

254. Muto M, Hitomi Y, Ohtsu A, Shimada H, Kashiwase Y, Sasaki H, et al. Acetaldehyde production by non-pathogenic Neisseria in human oral microflora: implications for carcinogenesis in upper aerodigestive tract. International journal of cancer. 2000;88(3):342-50.

255. Salaspuro M. Interrelationship between alcohol, smoking, acetaldehyde and cancer. Novartis Found Symp. 2007;285:80-9; discussion 9-96, 198-9.

256. Tillonen J, Homann N, Rautio M, Jousimies-Somer H, Salaspuro M. Role of yeasts in the salivary acetaldehyde production from ethanol among risk groups for ethanol-associated oral cavity cancer. Alcohol Clin Exp Res. 1999;23(8):1409-15.

257. Homann N, Jousimies-Somer H, Jokelainen K, Heine R, Salaspuro M. High acetaldehyde levels in saliva after ethanol consumption: methodological aspects and pathogenetic implications. Carcinogenesis. 1997;18(9):1739-43.

258. Homann N, Tillonen J, Meurman JH, Rintamaki H, Lindqvist C, Rautio M, et al. Increased salivary acetaldehyde levels in heavy drinkers and smokers: a microbiological approach to oral cavity cancer. Carcinogenesis. 2000;21(4):663-8.

259. Tsai ST, Wong TY, Ou CY, Fang SY, Chen KC, Hsiao JR, et al. The interplay between alcohol consumption, oral hygiene, ALDH2 and ADH1B in the risk of head and neck cancer. International journal of cancer. 2014;135(10):2424-36.

260. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. International journal of epidemiology. 2012;41(2):514-20.

261. Nasman A, Attner P, Hammarstedt L, Du J, Eriksson M, Giraud G, et al. Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: an epidemic of viral-induced carcinoma? International journal of cancer. 2009;125(2):362-6.

262. Hwang T-Z, Hsiao J-R, Tsai C-R, Chang JS. Incidence trends of human papillomavirus-related head and neck cancer in Taiwan, 1995–2009. International journal of cancer. 2015;137(2):395-408.

263. Kreimer AR, Clifford GM, Boyle P, Franceschi S. Human Papillomavirus Types in Head and Neck Squamous Cell Carcinomas Worldwide: A Systematic Review. Cancer Epidemiology Biomarkers & Prevention. 2005;14(2):467-75.

264. Hobbs CG, Sterne JA, Bailey M, Heyderman RS, Birchall MA, Thomas SJ. Human papillomavirus and head and neck cancer: a systematic review and meta-analysis. Clin Otolaryngol. 2006;31(4):259-66.
265. Agalliu I, Gapstur S, Chen Z, et al. Associations of oral α-, β-, and γ-human papillomavirus types with risk of incident head and neck cancer. JAMA Oncology. 2016;2(5):599-606.
266. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. New England Journal of Medicine. 2010;363(1):24-35.
267. Herrero R, Castellsagué X, Pawlita M, Lissowska J, Kee F, Balaram P, et al. Human Papillomavirus and Oral Cancer: The International Agency for Research on Cancer Multicenter Study. JNCI: Journal of the National Cancer Institute. 2003;95(23):1772-83.
268. Smith EM, Ritchie JM, Summersgill KF, Hoffman HT, Wang DH, Haugen TH, et al. Human Papillomavirus in Oral Exfoliated Cells and Risk of Head and Neck Cancer. JNCI Journal of the National Cancer Institute. 2004;96(6):449-55.
269. Smith EM, Rubenstein LM, Haugen TH, Pawlita M, Turek LP. Complex Etiology Underlies Risk and Survival in Head and Neck Cancer Human Papillomavirus, Tobacco, and Alcohol: A Case for Multifactor Disease. Journal of Oncology. 2012;2012:9.
270. Sinha P, Logan HL, Mendenhall WM. Human papillomavirus, smoking, and head and neck cancer. American journal of otolaryngology. 2012;33(1):130-6.
271. Marmot M. Social determinants of health inequalities. The Lancet. 2005;365(9464):1099-104.
272. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't. Annals of the New York Academy of Sciences. 1999;896:3-15.
273. Bartley M, Blane D, Montgomery S. Health and the life course: why safety nets matter. BMJ (Clinical research ed). 1997;314(7088):1194-6.
274. Blane D. Social determinants of health--socioeconomic status, social class, and ethnicity. American Journal of Public Health. 1995;85(7):903-5.
275. Conway DI, Petticrew M, Marlborough H, Berthiller J, Hashibe M, Macpherson LMD. Socioeconomic inequalities and oral cancer risk: A systematic review and meta-analysis of case-control studies. International journal of cancer. 2008;122(12):2811-9.
276. Hajat A, Kaufman JS, Rose KM, Siddiqi A, Thomas JC. Do the wealthy have a health advantage? Cardiovascular disease risk factors and wealth. Soc Sci Med. 2010;71(11):1935-42.
277. Hwang E, Johnson-Obaseki S, McDonald JT, Connell C, Corsten M. Incidence of head and neck cancer and socioeconomic status in Canada from 1992 to 2007. Oral oncology. 2013;49(11):1072-6.
278. Link BG, Phelan J. Social conditions as fundamental causes of disease. Journal of Health and Social Behavior. 1995:80-94.
279. Nicolau B, Netuveli G, Kim JW, Sheiham A, Marcenes W. A life-course approach to assess psychosocial factors and periodontal disease. J Clin Periodontol. 2007;34(10):844-50.
280. Pollitt R, Rose K, Kaufman J. Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. BMC Public Health. 2005;5(1):7.
281. Szanton SL, Candidate CD. Allostatic Load : A Mechanism of Socioeconomic Health. 2010;7(1):7-15.

282. Warnakulasuriya S. Significant oral cancer risk associated with low socioeconomic status. Evid Based Dent. 2009;10(1):4-5.
283. Bhan N, Srivastava S, Agrawal S, Subramanyam M, Millett C, Selvaraj S, et al. Are socioeconomic disparities in tobacco consumption increasing in India? A repeated cross-sectional multilevel analysis. BMJ Open. 2012;2(5).
284. Kaufman JS, Cooper RS. Seeking Causal Explanations in Social Epidemiology. American Journal of Epidemiology. 1999;150(2):113-20.
285. Corsi DJ, Boyle MH, Lear Sa, Chow CK, Teo KK, Subramanian SV. Trends in smoking in Canada from 1950 to 2011: progression of the tobacco epidemic according to socioeconomic status and geography. Cancer causes & control : CCC. 2013.
286. Droomers M, Schrijvers CTM, Stronks K, van de Mheen D, Mackenbach JP. Educational Differences in Excessive Alcohol Consumption: The Role of Psychosocial and Material Stressors. Preventive Medicine. 1999;29(1):1-10.
287. Fone DL, Farewell DM, White J, Lyons RA, Dunstan FD. Socioeconomic patterning of excess alcohol consumption and binge drinking: a cross-sectional study of multilevel associations with neighbourhood deprivation. BMJ Open. 2013;3(4).
288. Hiscock R, Bauld L, Amos A, Fidler JA, Munafò M. Socioeconomic status and smoking: a review. Annals of the New York Academy of Sciences. 2012;1248(1):107-23.
289. Thankappan KR, Thresia CU. Tobacco use & social status in Kerala. Indian J Med Res. 2007;126(4):300-8.
290. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annual review of public health. 1997;18(16):341-78.
291. Filmer D, Pritchett LH. Estimating Wealth Effects without Expenditure Data-or Tears: An Application to Educational Enrollments in States of India. Demography. 2001;38(1):115-32.
292. Gwatkin DR, Rutstein S, Johnson K, Suliman E, Wagstaff A, Amouzou A. Socio-economic differences in health, nutrition, and population within developing countries: an overview. Niger J Clin Pract. 2007;10(4):272-82.
293. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 1). Journal of Epidemiology and Community Health. 2006;60(1):7-12.
294. Shaw M, Annu Rev Public H. Housing and public health. Annu Rev Public Health. 2004;25:397-418.
295. McKenzie DJ. Measuring inequality with asset indicators. Journal of Population Economics. 2005;18(2):229-60.
296. Howe LD, Hargreaves JR, Gabrysch S, Huttly SRA. Is the wealth index a proxy for consumption expenditure? A systematic review. Journal of Epidemiology and Community Health. 2009;63(11):871-7.
297. Krieger J, Higgins DL. Housing and Health: Time Again for Public Health Action. American Journal of Public Health. 2002;92(5):758-68.
298. Smith GD, Hart C, Blane D, Gillis C, Hawthorne V. Lifetime socioeconomic position and mortality: prospective observational study. BMJ (Clinical research ed). 1997;314(7080):547-52.
299. Berkman LF, Macintyre S. The measurement of social class in health studies: old measures and new formulations. IARC Sci Publ. 1997(138):51-64.

300. International Institute for Population Sciences (IIPS), 1995. National Family Health Survey (MCH and Family Planning), India http://dhsprogram.com/pubs/pdf/FRIND1/FRIND1.pdf [Internet]. IIPS. 1992-93. Available from: http://dhsprogram.com/Publications/Publication-Search.cfm?ctry_id=57&c=India&Country=India&cn=India.
301. Balen J, McManus DP, Li Y-S, Zhao Z-Y, Yuan L-P, Utzinger J, et al. Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of Hunan province, China. Emerging Themes in Epidemiology. 2010;7:7-.
302. Lynch J, Kaplan G. Socioeconomic position: Oxford University Press; 2000.
303. Shavers VL. Measurement of socioeconomic status in health disparities research. J Natl Med Assoc. 2007;99(9):1013-23.
304. Liberatos P, Link BG, Kelsey JL. The measurement of social class in epidemiology. Epidemiol Rev. 1988;10:87-121.
305. Hauser RM. Measuring Socioeconomic status in childhood development: Blackwell Publishing; 1994. 1541-5 p.
306. Nair PRG. Education and Socio-Economic Change in Kerala, 1793-1947. Social Scientist. 1976;4(8):28-43.
307. AKG center for research and studies, Communist party of India (Marxist), State committee, Kerala. Education Bill. Kerala,India: AKG Center for Research and Studies; 2009 [updated 2012. Available from: http://www.cpimkerala.org/eng/education-23.php?n=1.
308. Galobardes BF, Lynch J, Smith GD, Br Med B. Measuring socioeconomic position in health research. British medical bulletin. 2007;81-82(1):21-37.
309. Lynge E. Unemployment and cancer: a literature review. IARC Sci Publ. 1997(138):343-51.
310. Robertson T, Popham F, Benzeval M. Socioeconomic position across the lifecourse & allostatic load: data from the West of Scotland Twenty-07 cohort study. BMC Public Health. 2014;14(1):184.
311. Galobardes B, Lynch JW, Davey Smith G. Childhood Socioeconomic Circumstances and Cause-specific Mortality in Adulthood: Systematic Review and Interpretation. Epidemiologic Reviews. 2004;26(1):7-21.
312. Bernabe E, Suominen AL, Nordblad A, Vehkalahti MM, Hausen H, Knuuttila M, et al. Education level and oral health in Finnish adults: evidence from different lifecourse models. J Clin Periodontol. 2011;38(1):25-32.
313. Brennan DS, Spencer AJ. Income-based life-course models of caries in 30-year-old Australian adults. Community Dent Oral Epidemiol. 2015;43(3):262-71.
314. Johnson S, McDonald JT, Corsten M, Rourke R. Socio-economic status and head and neck cancer incidence in Canada: A case-control study. Oral oncology. 2010;46(3):200-3.
315. Madani AH, Dikshit M, Bhaduri D, Jahromi AS. Relationship between Selected Socio-Demographic Factors and Cancer of Oral Cavity - A Case Control Study. Cancer Inform. 2010;9:163-8.
316. Krishna Rao S, Mejia GC, Roberts-Thomson K, Logan RM, Kamath V, Kulkarni M, et al. Estimating the effect of childhood socioeconomic disadvantage on oral cancer in India using marginal structural models. Epidemiology (Cambridge, Mass). 2015;26(4):509-17.
317. Conway et.al DI, Conway DI, McMahon AD, Smith K, Black R, Robertson G, et al. Components of socioeconomic risk associated with head and neck cancer: a population-based case-control study in Scotland. Br J Oral Maxillofacial Surg. 2010;48(1):11-7.
318. Kaufman J. Progress and pitfalls in the social epidemiology of cancer. Cancer Causes & Control. 1999;10(6):489-94.

319. Stringhini S, Sabia S, Shipley M, et al. ASsociation of socioeconomic position with health behaviors and mortality. Jama. 2010;303(12):1159-66.
320. Global status report on alcohol and health, World Health Organisation. WHO:Management of substance abuse, 2014. Available form: http://www.who.int/substance_abuse/publications/global_alcohol_report/en/. 2014.
321. VanderWeele TJ, Jackson JW, Li S. Causal inference and longitudinal data: a case study of religion and mental health. Social Psychiatry and Psychiatric Epidemiology. 2016:1-10.
322. Kuh D, Ben-Shlomo Y. A life course approach to chronic disease epidemiology. Oxford; New York: Oxford University Press; 1997.
323. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. Journal of Epidemiology and Community Health. 2003;57(10):778.
324. Forsdahl A. Are poor living conditions in childhood and adolescence an important risk factor for arteriosclerotic heart disease? British Journal of Preventive &amp;amp; Social Medicine. 1977;31(2):91.
325. Wadsworth ME, Cripps HA, Midwinter RE, Colley JR. Blood pressure in a national birth cohort at the age of 36 related to social and familial factors, smoking, and body mass. Br Med J (Clin Res Ed). 1985;291(6508):1534-8.
326. M.Wadsworth. The imprint of time: childhood, history and adult life: Oxford: Clarendon Press.; 1991.
327. Barker DJ. The fetal and infant origins of adult disease. BMJ : British Medical Journal. 1990;301(6761):1111-.
328. Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. Lancet. 1986;1(8489):1077-81.
329. Osmond C, Barker DJ, Winter PD, Fall CH, Simmonds SJ. Early growth and death from cardiovascular disease in women. BMJ : British Medical Journal. 1993;307(6918):1519-24.
330. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. Journal of Epidemiology and Community Health. 2003;57(10):778-83.
331. Ben-Shlomo Y, Kuh D. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. International Journal of Epidemiology. 2002;31(2):285-93.
332. Blane D, Netuveli G, Stone J. The development of life course epidemiology. Revue d'epidemiologie et de sante publique. 2007;55(1):31-8.
333. McEwen BS. Stress, adaptation, and disease. Allostasis and allostatic load. Annals of the New York Academy of Sciences. 1998;840:33-44.
334. Power C, Hertzman C. Social and biological pathways linking early life and adult disease. British medical bulletin. 1997;53(1):210-21.
335. Hart CL, Davey Smith G, Blane D. Social mobility and 21 year mortality in a cohort of Scottish men. Social Science & Medicine. 1998;47(8):1121-30.
336. Blane D, Harding S, Rosato M. Does social mobility affect the size of the socioeconomic mortality differential?: evidence from the Office for National Statistics Longitudinal Study. Journal of the Royal Statistical Society Series A, (Statistics in Society). 1999;162(Pt. 1):59-70.
337. Bartley M, Plewis I. Increasing social mobility: an effective policy to reduce health inequalities. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2007;170(2):469-81.

338. Hallqvist J, Lynch J, Bartley M, Lang T, Blane D. Can we disentangle life course processes of accumulation, critical period and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. Soc Sci Med. 2004;58(8):1555-62.
339. Mayo NE, Goldberg MS. When is a case-control study not a case-control study? J Rehabil Med. 2009;41(4):209-16.
340. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
341. Szklo M, Nieto FJ. Epidemiology : beyond the basics. Burlington, Mass.: Jones & Bartlett Learning; 2014.
342. Sistrom CL, Garvan CW. Proportions, odds, and risk. Radiology. 2004;230(1):12-9.
343. VanderWeele TJ. Mediation: Introduction and Regression-Based Approaches. Explanation in Causal inference: Methods for mediation and Interaction. New York: NY: Oxford University Press; 2015. p. 20-65.
344. Breslow NE. Statistics in epidemiology: the case-control study. J Am Stat Assoc. 1996;91(433):14-28.
345. Miettinen OS. The "case-control" study: valid selection of subjects. J Chronic Dis. 1985;38(7):543-48.
346. Schlesselman JJ, editor. Case-control studies: design, conduct, analysis. New York: Oxford University Press; 1982.
347. Hernán MA. Invited Commentary: Hypothetical Interventions to Define Causal Effects—Afterthought or Prerequisite? American Journal of Epidemiology. 2005;162(7):618-20.
348. Rothman KJ. Modern epidemiology. Boston: Little, Brown; 1986.
349. Savitz DA, Wellenius GA. Interpreting epidemiologic evidence. 2nd ed. New York: Oxford University Press; 2016. 226 p.
350. Bunge M. Causality in Modern Science. New York: Dover. 434 pp. 3rd.1979.
351. Kaufman JS, Poole C. Looking back on "causal thinking in the health sciences". Annu Rev Public Health. 2000;21:101-19.
352. Doll R, Hill AB. The Mortality of Doctors in Relation to Their Smoking Habits. British Medical Journal. 1954;1(4877):1451-5.
353. Doll R, Hill AB. Lung Cancer and Other Causes of Death in Relation to Smoking. British Medical Journal. 1956;2(5001):1071-81.
354. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58.
355. Parascandola M, Weed DL, Dasgupta A. Two Surgeon General's reports on smoking and cancer: a historical investigation of the practice of causal inference. Emerging Themes in Epidemiology. 2006;3(1):1.
356. Rothman KJ. CAUSES. American Journal of Epidemiology. 1976;104(6):587-92.
357. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes MJ, Kaufman JS, editors. Methods in social epidemiology: Jossey-Bass; 2006. p. 387-418.
358. Rothman KJ. Epidemiology : an introduction. Oxford; New York: Oxford University Press; 2002.
359. Hofler M. Causal inference based on counterfactuals. BMC Med Res Methodol. 2005;5:28.
360. Hernan M, Robins JM. Causal inference 2016 [Available from: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

361. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701.
362. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986;7(9):1393-512.
363. Robins J. Proceedings of the Section on Bayesian Statistical Science. Alexandria, VA, American Statistical Association; 1998. Marginal structural models. 1997:1-10.
364. Robins JM. Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In: Halloran ME, Berry D, editors. Statistical Models in Epidemiology, the Environment, and Clinical Trials. New York, NY: Springer New York; 2000. p. 95-133.
365. Robins JM, Hernán Ma, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass). 2000;11(5):550-60.
366. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American journal of epidemiology. 2010;172(12):1339-48.
367. Weinberg CR. Toward a clearer definition of confounding. Am J Epidemiol. 1993;137(1):1-8.
368. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-88.
369. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology (Cambridge, Mass). 1999;10(1):37-48.
370. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology (Cambridge, Mass). 2004;15(5):615-25.
371. Naimi AI. Causal inference in occupational epidemiology: Asbestos, lung cancer mortality, and the healthy worker survivor effect: The University of North Carolina , Chapel Hill; 2012.
372. Gupta B, Ariyawardana A, Johnson NW. Oral cancer in India continues in epidemic proportions: evidence base and policy initiatives. International dental journal. 2013;63(1):12-25.
373. Bhan N, Rao KD, Kachwaha S. Health inequalities research in India: a review of trends and themes in the literature since the 1990s. International Journal for Equity in Health. 2016;15:166.
374. Moore N, Pierce A, Wilson DS, Johnson. The epidemiology of lip cancer: a review of global incidence and aetiology. Oral Dis. 1999;5(3):185-95.
375. Zarbo RJ. Salivary gland neoplasia: a review for the practicing pathologist. Mod Pathol. 2002;15(3):298-323.
376. Chang ET, Adami HO. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 2006;15(10):1765-77.
377. Westreich D. Berkson’s bias, selection bias, and missing data. Epidemiology (Cambridge, Mass). 2012;23(1):159-64.
378. Nishimoto, Pintos J, Schlecht NF, Torloni H, Carvalho AL, Kowalski LP, et al. Assessment of control selection bias in a hospital-based case-control study of upper aero-digestive tract cancers. J Cancer Epidemiol Prev. 2002;7(3):131-41.
379. Marmot MG, Stansfeld S, Patel C, North F, Head J, White I, et al. Health inequalities among British civil servants: the Whitehall II study. The Lancet. 1991;337(8754):1387-93.
380. Power C Fox J MO. Health and Class: The Early Years:London: Chapman Hall; 1991.
381. Berney L, Blane DB, Soc Sci M, Blane B. Collecting retrospective data: accuracy of recall after 50 years judged against historical records. Social Science & Medicine. 1997;45(10):1519-25.

382. Bell AJ. "Oh yes, I remember it well!" Reflections on using the the life-grid in in qualitative interviews with couples. Qualitative Sociology Review. 2005;1(1):67-.
383. Heath EM, Morken NW, Campbell Ka, Tkach D, Boyd Ea, Strom Da. Use of buccal cells collected in mouthwash as a source of DNA for clinical testing. Archives of pathology & laboratory medicine. 2001;125(1):127-33.
384. D'Souza G, Sugar E, Ruby W, Gravitt P, Gillison M. Analysis of the effect of DNA purification on detection of human papillomavirus in oral rinse samples by PCR. J Clin Microbiol. 2005;43(11):5526-35.
385. Sciubba JJ. Improving detection of precancerous and cancerous oral lesions. Computer-assisted analysis of the oral brush biopsy. U.S. Collaborative OralCDx Study Group. J Am Dent Assoc. 1999;130(10):1445-57.
386. Laprise C, Madathil SA, Schlecht NF, Castonguay G, Soulières D, Nguyen-Tan PF, et al. Human papillomavirus genotypes and risk of head and neck cancers: Results from the HeNCe Life case-control study. Oral oncology. 2017;69:56-61.
387. Egan KM, Abruzzo J, Cytobrush B, Newcomb PA, Titus-ernstoff L, Franklin T, et al. Collection of Genomic DNA from Adults in Epidemiological Studies by Buccal Cytobrush and Mouthwash. 2001:687-96.
388. Lawton G, Thomas, Schonrock, Monsour, Frazer. Human papillomaviruses in normal oral mucosa: a comparison of methods for sample collection. J Oral Pathol Med. 1992;Jul;(21(6)):265-9.
389. Walling DM, Flaitz CM, Adler-Storthz K, Nichols CM. A non-invasive technique for studying oral epithelial Epstein-Barr virus infection and disease. Oral Oncol. 2003;39(5):436-44.
390. Muñoz N, Bosch FX. Biomarkers for biological agents. IARC Sci Publ. 1997;142:127-42.
391. Coutlée F, Mayrand MH, Provencher D, Franco E. The future of HPV testing in clinical laboratories and applied virology research. Clinical and diagnostic virology. 1997;8(2):123-41.
392. Kornegay JR, Roger M, Davies PO, Shepard AP, Guerrero NA, Lloveras B, et al. International Proficiency Study of a Consensus L1 PCR Assay for the Detection and Typing of Human Papillomavirus DNA: Evaluation of Accuracy and Intralaboratory and Interlaboratory Agreement. Journal of Clinical Microbiology. 2003;41(3):1080-6.
393. London SJ, Xia J, Lehman TA, Yang J-h, Granada E, Chunhong L. Collection of Buccal Cell DNA in Seventh-Grade Children Using Water and a Toothbrush. 2001:1227-30.
394. Laprise C, Madathil SA, Allison P, Abraham P, Raghavendran A, Shahul HP, et al. No role for human papillomavirus infection in oral cancers in a region in southern India. International journal of cancer. 2016;138(4):912-7.
395. Mehrotra ea, Mol C. Application of cytology and molecular biology in diagnosing premalignant or malignant oral lesions. Molecular Cancer 2006. 2006;5:1-9.
396. Montgomery MR, Gragnolati M, Burke KA, Paredes E. Measuring living standards with proxy variables. Demography. 2000;37(2):155-74.
397. DHS. Demographic and Health Survey http://dhsprogram.com/topics/wealth-index/Wealth-Index-Construction.cfm [
398. Jolliffe IT. Principal component analysis. New York: Springer; 2002.
399. Debelak R, Tran US. Principal Component Analysis of Smoothed Tetrachoric Correlation Matrices as a Measure of Dimensionality. Educational and Psychological Measurement. 2013;73(1):63-77.

400. Beebe-Dimmer J, Lynch JW, Turrell G, Lustgarten S, Raghunathan T, Kaplan GA. Childhood and Adult Socioeconomic Conditions and 31-Year Mortality Risk in Women. American Journal of Epidemiology. 2004;159(5):481-90.
401. Hadden WC. Annotation: the use of educational attainment as an indicator of socioeconomic position. Am J Public Health. 1996;86(11):1525-6.
402. National Cancer I. http://www.cancer.gov/dictionary?cdrid=306510.
403. Leffondre K. Modeling Smoking History: A Comparison of Different Approaches. American Journal of Epidemiology. 2002;156(9):813-23.
404. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM. Adjustment for continuous confounders: an example of how to prevent residual confounding. CMAJ : Canadian Medical Association Journal. 2013;185(5):401-6.
405. Nayak MB, Kerr W, Greenfield TK, Pillai A. Not All Drinks Are Created Equal: Implications for Alcohol Assessment in India. Alcohol and Alcoholism. 2008;43(6):713-8.
406. Madathil SA. Paan chewing : intergenerational habit transmission and lifetime dose-response relationship with oral cancer among a subset of South Indian population [Manuscript-Based]. Montreal: McGill University; 2013.
407. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology (Cambridge, Mass). 1999;10(1):37-48.
408. Dietrich T, Hoffmann K. A comprehensive index for the modeling of smoking history in periodontal research. J Dent Res. 2004;83(11):859-63.
409. Leffondr K, Abrahamowicz M, Xiao Y. Modelling smoking history using a comprehensive smoking index : Application to lung cancer. 2006(September):4132-46.
410. Harrell FE. Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis. 2001.
411. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 1989.
412. Williams BA, Madrekar JN, Madrekar SJ, Cha SS, Furth AF. Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes Mayo Clinic, Rochester, Minnesota Division of Biostatistics DoHSR; June 2006. Contract No.: 10027230.
413. Textor J. Drawing and Analyzing Causal DAGs with DAG itty User Manual for Version 2 . 3. 2015. p. 1-15.
414. Textor J, Hardt J, Knuppel S. DAGitty: A graphical tool for analysisng causal diagrams. Epidemiology (Cambridge, Mass). 2011;5(22):745-.
415. Farsi N. Epidemiology of human papilloma virus related head and neck cancers [Manuscript-based]. Montreal, Canada: McGill University; 2014.
416. Tyndale RF, Sellers EM. Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. Drug metabolism and disposition: the biological fate of chemicals. 2001;29(4 Pt 2):548-52.
417. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. Journal of the National Cancer Institute. 2000;92(14):1151-8.
418. Rosner B. Fundamentals of Biostatistics. Seventh ed ed2010. 2-888 p.
419. VanderWeele TJ. An Introduction to Interaction Analysis. Explanation in Causal Inference: Methods of Mediation and Interaction. New York: NY:Oxford University press; 2015. p. 249-84.

420. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology (Cambridge, Mass). 2000;11(5):561-70.
421. Hogue CJR, Parker CB, Willinger M, Temple JR, Bann CM, Silver RM, et al. The Association of Stillbirth with Depressive Symptoms 6–36 Months Post-Delivery. Paediatric and Perinatal Epidemiology. 2015;29(2):131-43.
422. Menvielle G, Franck J-e, Radoï L, Sanchez M, Févotte J, Guizard A-V, et al. Quantifying the mediating effects of smoking and occupational exposures in the relation between education and lung cancer: the ICARE study. European Journal of Epidemiology. 2016;31(12):1213-21.
423. Xu X, Ritz B, Cockburn M, Lombardi C, Heck JE. Maternal Preeclampsia and Odds of Childhood Cancers in Offspring - A California Statewide Case-Control Study. Paediatr Perinat Epidemiol. 2017.
424. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health. 2004;58(4):265-71.
425. Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R. A structured approach to modelling the effects of binary exposure variables over the life course. International Journal of Epidemiology. 2009;38(2):528-37.
426. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology. 2008;168(6):656-64.
427. Nandi A, Glymour MM, Kawachi I, VanderWeele TJ. Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes, and stroke. Epidemiology (Cambridge, Mass). 2012;23(2):223-32.
428. Murray ET, Mishra GD, Kuh D, Guralnik J, Black S, Hardy R. Life course models of socioeconomic position and cardiovascular risk factors: 1946 birth cohort. Ann Epidemiol. 2011;21(8):589-97.
429. Leffondre K, Wynant W, Cao Z, Abrahamowicz M, Heinze G, Siemiatycki J. A weighted Cox model for modelling time-dependent exposures in the analysis of case-control studies. Stat Med. 2010;29(7-8):839-50.
430. Platt RW, Brookhart MA, Cole SR, Westreich D, Schisterman EF. An information criterion for marginal structural models. Statistics in Medicine. 2013;32(8):1383-93.
431. Knol MJ, Egger M, Scott P, Geerlings MI, Vandenbroucke JP. When one depends on the other: reporting of interaction in case-control and cohort studies. Epidemiology (Cambridge, Mass). 2009;20(2):161-6.
432. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. Epidemiology (Cambridge, Mass). 2014;25(5):749-61.
433. Judd CM, Kenny DA. Process Analysis. Evaluation Review. 1981;5(5):602-19.
434. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173-82.
435. MacKinnon DP, Fairchild AJ. Current Directions in Mediation Analysis. Current directions in psychological science. 2009;18(1):16-20.
436. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the Mediation, Confounding and Suppression Effect. Prevention Science. 2000;1(4):173-81.

437. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological methods. 2013;18(2):137-50.
438. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology (Cambridge, Mass). 1992;3(2):143-55.
439. Pearl J. Direct and indirect effects. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence; Seattle, Washington. 2074073: Morgan Kaufmann Publishers Inc.; 2001. p. 411-20.
440. Hayes AF. Beyond Baron and Kenny : Statistical Mediation Analysis in the New Millennium Beyond Baron and Kenny : Statistical Mediation Analysis in the New Millennium. 2009(July 2014):37-41.
441. VanderWeele TJ. A Unification of Mediation and Interaction. Explanation in Causal inference: Methods for Mediation and Interaction. New York: NY: Oxford University Press; 2015. p. 371-96.
442. Erratum: A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology (Cambridge, Mass). 2016;27(5):e36.
443. http://www.stata.com/manuals13/rbootstrap.pdf [Internet].
444. DiCiccio TJ, Efron B. Bootstrap confidence intervals. 1996:189-228.
445. Carlson LE, Speca M, Patel KD, Goodey E. Mindfulness-Based Stress Reduction in Relation to Quality of Life, Mood, Symptoms of Stress, and Immune Parameters in Breast and Prostate Cancer Outpatients. Psychosomatic Medicine. 2003;65(4):571-81.
446. Sapolsky RM. The Influence of Social Hierarchy on Primate Health. Science. 2005;308(5722):648-52.
447. Adler NE, Stewart J. Preface to The Biology of Disadvantage: Socioeconomic Status and Health. Annals of the New York Academy of Sciences. 2010;1186(1):1-4.
448. Kelly-Irving M, Mabile L, Grosclaude P, Lang T, Delpierre C. The embodiment of adverse childhood experiences and cancer development: potential biological mechanisms and pathways across the life course. Int J Public Health. 2013;58(1):3-11.
449. Buckley R, Cartwright K, Struyk R, Szymanoski E. Integrating housing wealth into the social safety net for the Moscow elderly: an empirical essay. Journal of Housing Economics. 2003;12(3):202-23.
450. Kawachi I, Kennedy BP. Health and social cohesion: why care about income inequality? BMJ (Clinical research ed). 1997;314(7086):1037-40.
451. Berton HK, Cassel JC, Gore S. Social Support and Health. Medical Care. 1977;15(5):47-58.
452. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Social Science & Medicine. 2000;51(6):843-57.
453. McEwen BS. Allostasis and Allostatic Load: Implications for Neuropsychopharmacology. Neuropsychopharmacology. 2000;22(2):108-24.
454. McEwen BS. Interacting Mediators of Allostasis and Allostatic Load: Towards an Understanding of Resilience in Aging. 2003;52(10):10-6.
455. Epel ES, Blackburn EH, Lin J, Dhabhar FS, Adler NE, Morrow JD, et al. Accelerated telomere shortening in response to life stress. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(49):17312-5.
456. Willeit P, Willeit J, Mayr A, Weger S, Oberhollenzer F, Brandstätter A, et al. Telomere Length and Risk of Incident Cancer and Cancer Mortality. JAMA: The Journal of the American Medical Association. 2010;304(1):69-75.

457. Patel et.al MM, Patel MM, Parekh LJ, Jha FP, Sainger RN, Patel JB, et al. Clinical usefulness of telomerase activation and telomere length in head and neck cancer. Head & Neck. 2002;24(12):1060-7.
458. Sainger. Clinical significance of telomere length and associated proteins in oral cancer. Biomark In. 2007;14(2):9-19.
459. Sebastian S, Grammatica L, Paradiso A. Telomeres, telomerase and oral cancer (Review). Int J Oncol. 2005;27(6):1583-15896.
460. Bakhtiar SM, Ali A, Barh D. Epigenetics in Head and Neck Cancer. In: Verma M, editor. Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis. New York, NY: Springer New York; 2015. p. 751-69.
461. Jithesh PV, Risk JM, Schache AG, Dhanda J, Lane B, Liloglou T, et al. The epigenetic landscape of oral squamous cell carcinoma. British Journal of Cancer. 2013;108(2):370-9.
462. Paper IR, Oncology N, Shaw R. The epigenetics of oral cancer. International journal of oral and maxillofacial surgery. 2006;35(2):101-8.
463. Shaw R. The epigenetics of oral cancer. International Journal of Oral and Maxillofacial Surgery. 2006;35(2):101-8.
464. McGuinness D, McGlynn LM, Johnson PCD, MacIntyre A, Batty GD, Burns H, et al. Socio-economic status is associated with epigenetic differences in the pSoBid cohort. International journal of epidemiology. 2012;41(1):151-60.
465. Subramanyam MA, Diez-Roux AV, Pilsner JR, Villamor E, Donohue KM, Liu Y, et al. Social Factors and Leukocyte DNA Methylation of Repetitive Sequences: The Multi-Ethnic Study of Atherosclerosis. PLOS ONE. 2013;8(1):e54018.
466. Stringhini S, Polidoro S, Sacerdote C, Kelly RS, van Veldhoven K, Agnoli C, et al. Life-course socioeconomic status and DNA methylation of genes regulating inflammation. International Journal of Epidemiology. 2015;44(4):1320-30.
467. Chen E, Hanson MD, Paterson LQ, Griffin MJ, Walker HA, Miller GE. Socioeconomic status and inflammatory processes in childhood asthma: The role of psychological stress. Journal of Allergy and Clinical Immunology. 2006;117(5):1014-20.
468. Pretscher D, Distel LV, Grabenbauer GG, Wittlinger M, Buettner M, Niedobitek G. Distribution of immune cells in head and neck cancer: CD8+ T-cells and CD20+B-cells in metastatic lymph nodes are associated with favourable outcome in patients with oro- and hypopharyngeal carcinoma. BMC Cancer. 2009;9(1):292.
469. Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, et al. Associations with early-life socio-economic position in adult DNA methylation. Int J Epidemiol. 2012;41(1):62-74.
470. Fagundes CP, Way B. Early-Life Stress and Adult Inflammation. Current Directions in Psychological Science. 2014;23(4):277-83.
471. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol. 1992;135(9):1019-28.
472. Olson SH. Reported Participation in Case-Control Studies: Changes over Time. American Journal of Epidemiology. 2001;154(6):574-81.
473. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. Annals of Epidemiology. 2007;17(9):643-53.
474. Conway DI. Socioeconomic factors influence selection and participation in a population-based case-control study of head and neck cancer in Scotland. Journal of Clinical Epidemiology. 2008;61(11):1187-93.

475. Inda, 2011 Census Data [Internet]. Office of the Registrar General & Census Commissioner, India. 2011 http://www.censusindia.gov.in/2011census/dchb/KerlaA.html. Available from: http://www.censusindia.gov.in/.
476. Hardgrave RL. Caste in Kerala: A preface to the Elections. Kerala; November 21,1964.
477. Hernán MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. American Journal of Epidemiology. 2009;170(8):959-62.
478. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. 2009.
479. Krall EA, Valadian I, Dwyer JT, Gardner J. Accuracy of recalled smoking data. Am J Public Health. 1989;79(2):200-2.
480. VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction. New York: NY: Oxford University Press; 2015. 706 p.
481. Prince Nelson SL, Viswanathan R, Paul J N, Diane L K, Paula S R, Bethany J W. An evaluation of common methods for dichotomization of continuous variables to discriminate disease status. Communications in Statistics - Theory and Methods. 2016.
482. Altman DG. Categorising continuous variables. British Journal of Cancer. 1991;64(5):975-.
483. Greenland S. Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. Epidemiology (Cambridge, Mass). 1995;6(4):356-65.
484. Garcia-Closas M, Rothman N, Lubin J. Misclassification in Case-Control Studies of Gene-Environment Interactions: Assessment of Bias and Sample Size. Cancer Epidemiology Biomarkers &amp;amp; Prevention. 1999;8(12):1043.
485. Brennan P. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis. 2002;23(3):381-7.
486. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology (Cambridge, Mass). 2010;21(4):540-51.
487. Naimi A, Kaufman J. Counterfactual Theory in Social Epidemiology: Reconciling Analysis and Action for the Social Determinants of Health. Current Epidemiology Reports. 2015;2(1):52-60.
488. Rehkopf DH, Glymour MM, Osypuk TL. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose Is Not a Rose. Current Epidemiology Reports. 2016;3(1):63-71.
489. Hernan MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology (Cambridge, Mass). 2011;22(3):368-77.
490. Robins JM. Association, Causation, And Marginal Structural Models. Synthese. 1999;121(1):151-79.
491. VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, Han Y, Spitz MR, Shete S, et al. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. American journal of epidemiology. 2012;175(10):1013-20.
492. Richardson DB, Rzehak P, Klenk J, Weiland SK. Analyses of case-control data for additional outcomes. Epidemiology (Cambridge, Mass). 2007;18(4):441-5.
493. Kujan O, Glenny AM, Duxbury J, Thakker N, Sloan P. Evaluation of screening strategies for improving oral cancer mortality: a Cochrane systematic review. J Dent Educ. 2005;69(2):255-65.
494. VanderWeele TJ. Policy-relevant proportions for direct effects. Epidemiology (Cambridge, Mass). 2013;24(1):175-6.

495. Howe LD, Smith AD, Macdonald-Wallis C, Anderson EL, Galobardes B, Lawlor DA, et al. Relationship between mediation analysis and the structured life course approach. International Journal of Epidemiology. 2016;45(4):1280-94.

496. Oude Ophuis MB, van Lieshout EM, Roelofs HM, Peters WH, Manni JJ. Glutathione S-transferase M1 and T1 and cytochrome P4501A1 polymorphisms in relation to the risk for benign and malignant head and neck lesions. Cancer. 1998;82(5):936-43.

497. Liu L, Wu G, Xue F, Li Y, Shi J, Han J, et al. Functional CYP1A1 genetic variants, alone and in combination with smoking, contribute to development of head and neck cancers. European Journal of Cancer. 2013;49(9):2143-51.

498. Matthias C, Bockmühl U, Jahnke V, Harries LW, Wolf CR, Jones PW, et al. The glutathione S-transferase GSTP1 polymorphism: effects on susceptibility to oral/pharyngeal and laryngeal carcinomas. Pharmacogenetics. 1998;8(1):1-6.

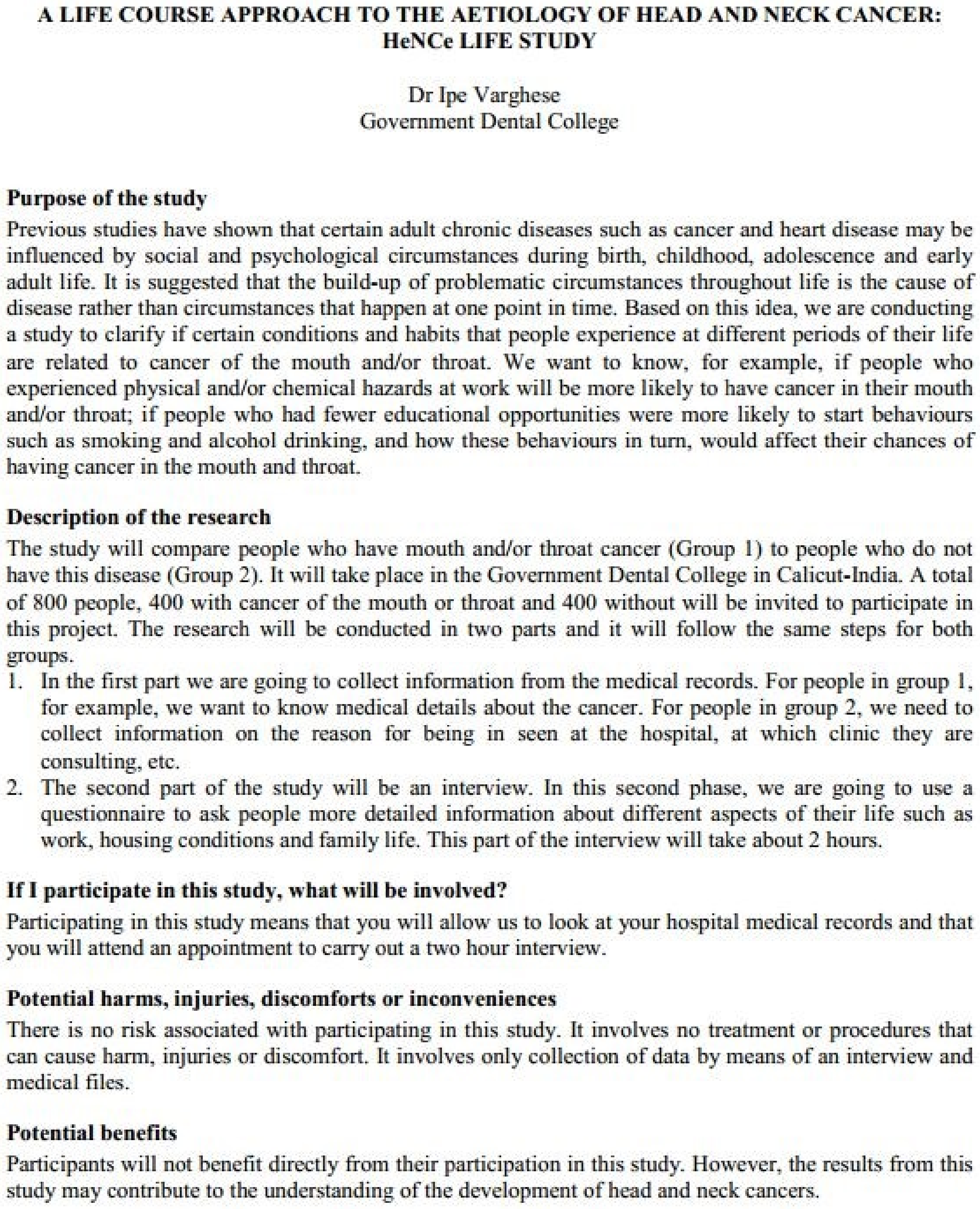
499. Russo A, Francelin PR, Galbiatti AL, Raposo LS, Maniglia JV, Pavarino EC, et al. Association between GSTP1, GSTM1 and GSTT1 polymorphisms involved in xenobiotic metabolism and head and neck cancer development. Mol Biol Rep. 2013;40(7):4181-8.

500. O'Loughlin J, Paradis G, Kim W, DiFranza J, Meshefedjian G, McMillan-Davey E, et al. Genetically decreased CYP2A6 and the risk of tobacco dependence: a prospective study of novice smokers. Tobacco control. 2004;13(4):422-8.

Appendix I

Participant consent forms for interview and biological sample collection

India site	**:**	**English and Malayalam (local language of study site) versions**
Canada site	**:**	**English and French versions**



A LIFE COURSE APPROACH TO THE AETIOLOGY OF HEAD AND NECK CANCER: HeNCe LIFE STUDY

Dr Ipe Varghese
Government Dental College

Purpose of the study

Previous studies have shown that certain adult chronic diseases such as cancer and heart disease may be influenced by social and psychological circumstances during birth, childhood, adolescence and early adult life. It is suggested that the build-up of problematic circumstances throughout life is the cause of disease rather than circumstances that happen at one point in time. Based on this idea, we are conducting a study to clarify if certain conditions and habits that people experience at different periods of their life are related to cancer of the mouth and/or throat. We want to know, for example, if people who experienced physical and/or chemical hazards at work will be more likely to have cancer in their mouth and/or throat; if people who had fewer educational opportunities were more likely to start behaviours such as smoking and alcohol drinking, and how these behaviours in turn, would affect their chances of having cancer in the mouth and throat.

Description of the research

The study will compare people who have mouth and/or throat cancer (Group 1) to people who do not have this disease (Group 2). It will take place in the Government Dental College in Calicut-India. A total of 800 people, 400 with cancer of the mouth or throat and 400 without will be invited to participate in this project. The research will be conducted in two parts and it will follow the same steps for both groups.

1. In the first part we are going to collect information from the medical records. For people in group 1, for example, we want to know medical details about the cancer. For people in group 2, we need to collect information on the reason for being in seen at the hospital, at which clinic they are consulting, etc.
2. The second part of the study will be an interview. In this second phase, we are going to use a questionnaire to ask people more detailed information about different aspects of their life such as work, housing conditions and family life. This part of the interview will take about 2 hours.

If I participate in this study, what will be involved?

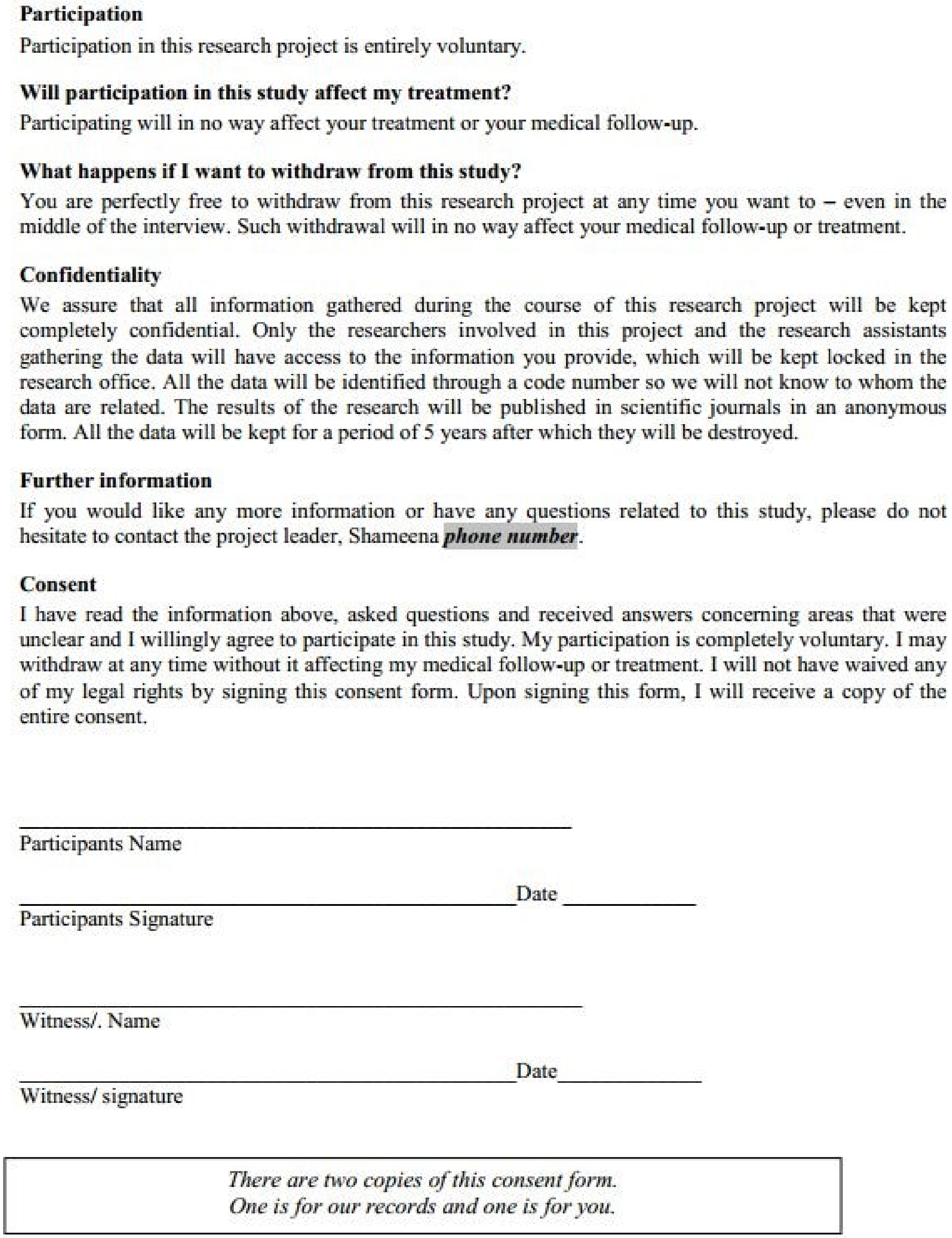
Participating in this study means that you will allow us to look at your hospital medical records and that you will attend an appointment to carry out a two hour interview.

Potential harms, injuries, discomforts or inconveniences

There is no risk associated with participating in this study. It involves no treatment or procedures that can cause harm, injuries or discomfort. It involves only collection of data by means of an interview and medical files.

Potential benefits

Participants will not benefit directly from their participation in this study. However, the results from this study may contribute to the understanding of the development of head and neck cancers.



Participation

Participation in this research project is entirely voluntary.

Will participation in this study affect my treatment?

Participating will in no way affect your treatment or your medical follow-up.

What happens if I want to withdraw from this study?

You are perfectly free to withdraw from this research project at any time you want to – even in the middle of the interview. Such withdrawal will in no way affect your medical follow-up or treatment.

Confidentiality

We assure that all information gathered during the course of this research project will be kept completely confidential. Only the researchers involved in this project and the research assistants gathering the data will have access to the information you provide, which will be kept locked in the research office. All the data will be identified through a code number so we will not know to whom the data are related. The results of the research will be published in scientific journals in an anonymous form. All the data will be kept for a period of 5 years after which they will be destroyed.

Further information

If you would like any more information or have any questions related to this study, please do not hesitate to contact the project leader, Shameena ***phone number***.

Consent

I have read the information above, asked questions and received answers concerning areas that were unclear and I willingly agree to participate in this study. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment. I will not have waived any of my legal rights by signing this consent form. Upon signing this form, I will receive a copy of the entire consent.

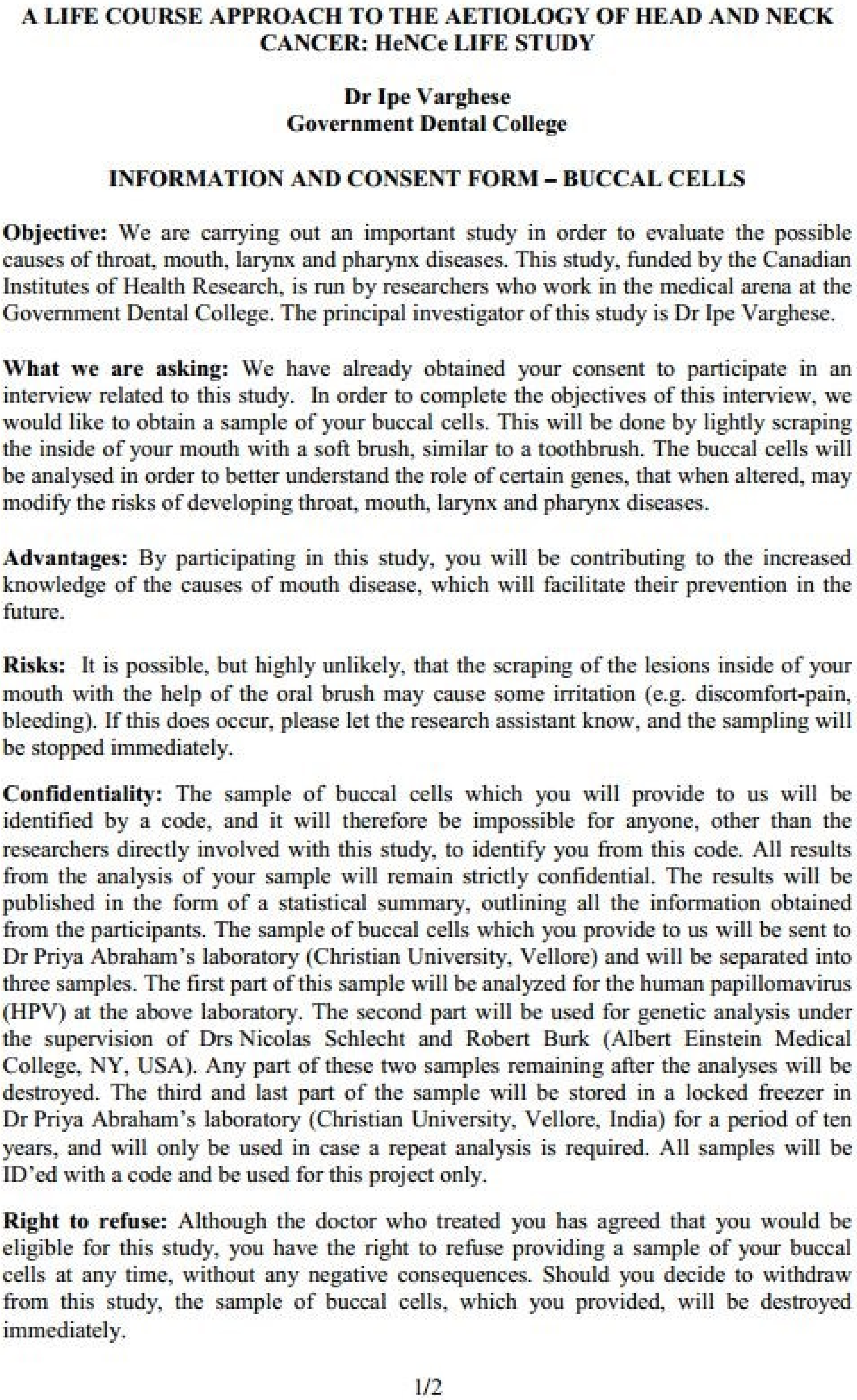
____
Participants Name

____ Date ___________
Participants Signature

____
Witness/. Name

____ Date ___________
Witness/ signature

There are two copies of this consent form.
One is for our records and one is for you.



A LIFE COURSE APPROACH TO THE AETIOLOGY OF HEAD AND NECK CANCER: HeNCe LIFE STUDY

Dr Ipe Varghese
Government Dental College

INFORMATION AND CONSENT FORM – BUCCAL CELLS

Objective: We are carrying out an important study in order to evaluate the possible causes of throat, mouth, larynx and pharynx diseases. This study, funded by the Canadian Institutes of Health Research, is run by researchers who work in the medical arena at the Government Dental College. The principal investigator of this study is Dr Ipe Varghese.

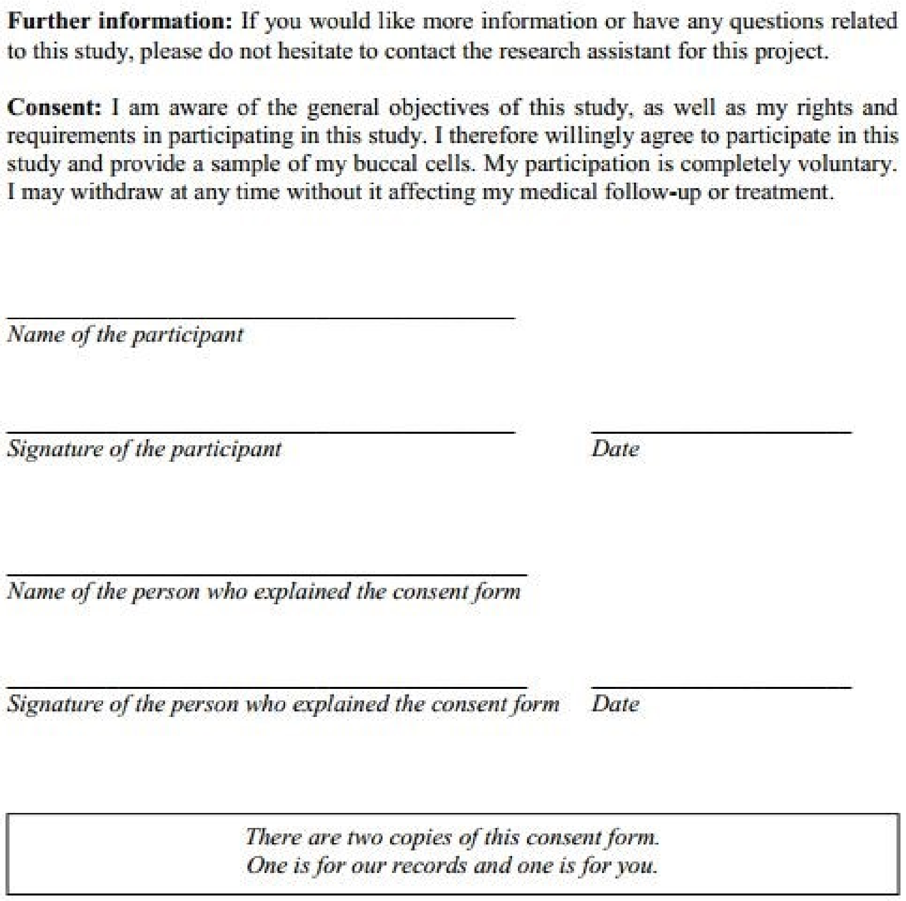
What we are asking: We have already obtained your consent to participate in an interview related to this study. In order to complete the objectives of this interview, we would like to obtain a sample of your buccal cells. This will be done by lightly scraping the inside of your mouth with a soft brush, similar to a toothbrush. The buccal cells will be analysed in order to better understand the role of certain genes, that when altered, may modify the risks of developing throat, mouth, larynx and pharynx diseases.

Advantages: By participating in this study, you will be contributing to the increased knowledge of the causes of mouth disease, which will facilitate their prevention in the future.

Risks: It is possible, but highly unlikely, that the scraping of the lesions inside of your mouth with the help of the oral brush may cause some irritation (e.g. discomfort-pain, bleeding). If this does occur, please let the research assistant know, and the sampling will be stopped immediately.

Confidentiality: The sample of buccal cells which you will provide to us will be identified by a code, and it will therefore be impossible for anyone, other than the researchers directly involved with this study, to identify you from this code. All results from the analysis of your sample will remain strictly confidential. The results will be published in the form of a statistical summary, outlining all the information obtained from the participants. The sample of buccal cells which you provide to us will be sent to Dr Priya Abraham's laboratory (Christian University, Vellore) and will be separated into three samples. The first part of this sample will be analyzed for the human papillomavirus (HPV) at the above laboratory. The second part will be used for genetic analysis under the supervision of Drs Nicolas Schlecht and Robert Burk (Albert Einstein Medical College, NY, USA). Any part of these two samples remaining after the analyses will be destroyed. The third and last part of the sample will be stored in a locked freezer in Dr Priya Abraham's laboratory (Christian University, Vellore, India) for a period of ten years, and will only be used in case a repeat analysis is required. All samples will be ID'ed with a code and be used for this project only.

Right to refuse: Although the doctor who treated you has agreed that you would be eligible for this study, you have the right to refuse providing a sample of your buccal cells at any time, without any negative consequences. Should you decide to withdraw from this study, the sample of buccal cells, which you provided, will be destroyed immediately.



Further information: If you would like more information or have any questions related to this study, please do not hesitate to contact the research assistant for this project.

Consent: I am aware of the general objectives of this study, as well as my rights and requirements in participating in this study. I therefore willingly agree to participate in this study and provide a sample of my buccal cells. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment.

____
Name of the participant

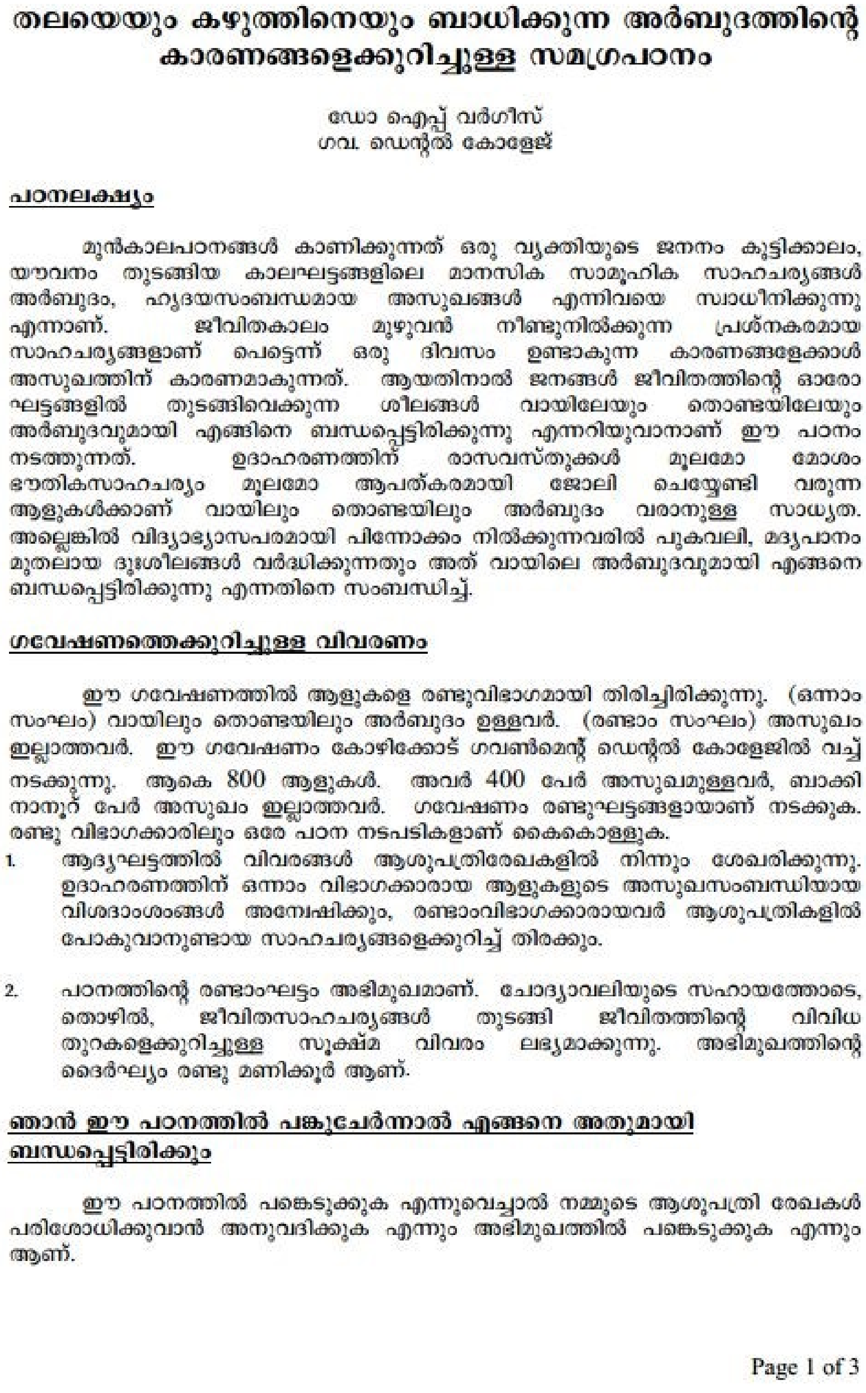
________________________________ ________________
Signature of the participant *Date*

____
Name of the person who explained the consent form

________________________________ ________________
Signature of the person who explained the consent form *Date*

There are two copies of this consent form.
One is for our records and one is for you.

2/2



തലയെയും കഴുത്തിനെയും ബാധിക്കുന്ന അർബുദത്തിന്റെ കാരണങ്ങളെക്കുറിച്ചുള്ള സമഗ്രപഠനം

ഡോ ഐപ്പ് വർഗീസ്
ഗവ. ഡെന്റൽ കോളേജ്

പഠനലക്ഷ്യം

മുൻകാലപഠനങ്ങൾ കാണിക്കുന്നത് ഒരു വ്യക്തിയുടെ ജനനം കുട്ടിക്കാലം, യൗവനം തുടങ്ങിയ കാലഘട്ടങ്ങളിലെ മാനസിക സാമൂഹിക സാഹചര്യങ്ങൾ അർബുദം, ഹൃദയസംബന്ധമായ അസുഖങ്ങൾ എന്നിവയെ സ്വാധീനിക്കുന്നു എന്നാണ്. ജീവിതകാലം മുഴുവൻ നീണ്ടുനിൽക്കുന്ന പ്രശ്നകരമായ സാഹചര്യങ്ങളാണ് പെട്ടെന്ന് ഒരു ദിവസം ഉണ്ടാകുന്ന കാരണങ്ങളേക്കാൾ അസുഖത്തിന് കാരണമാകുന്നത്. ആയതിനാൽ ജനങ്ങൾ ജീവിതത്തിന്റെ ഓരോ ഘട്ടങ്ങളിൽ തുടങ്ങിവെക്കുന്ന ശീലങ്ങൾ വായിലേയും തൊണ്ടയിലേയും അർബുദവുമായി എങ്ങിനെ ബന്ധപ്പെട്ടിരിക്കുന്നു എന്നറിയുവാനാണ് ഈ പഠനം നടത്തുന്നത്. ഉദാഹരണത്തിന് രാസവസ്തുക്കൾ മൂലമോ മോശം ഭൗതികസാഹചര്യം മൂലമോ ആപത്കരമായി ജോലി ചെയ്യേണ്ടി വരുന്ന ആളുകൾക്കാണ് വായിലും തൊണ്ടയിലും അർബുദം വരാനുള്ള സാധ്യത. അല്ലെങ്കിൽ വിദ്യാഭ്യാസപരമായി പിന്നോക്കം നിൽക്കുന്നവരിൽ പുകവലി, മദ്യപാനം മുതലായ ദുഃശീലങ്ങൾ വർദ്ധിക്കുന്നതും അത് വായിലെ അർബുദവുമായി എങ്ങനെ ബന്ധപ്പെട്ടിരിക്കുന്നു എന്നതിനെ സംബന്ധിച്ച്.

ഗവേഷണത്തെക്കുറിച്ചുള്ള വിവരണം

ഈ ഗവേഷണത്തിൽ ആളുകളെ രണ്ടുവിഭാഗമായി തിരിച്ചിരിക്കുന്നു. (ഒന്നാം സംഘം) വായിലും തൊണ്ടയിലും അർബുദം ഉള്ളവർ. (രണ്ടാം സംഘം) അസുഖം ഇല്ലാത്തവർ. ഈ ഗവേഷണം കോഴിക്കോട് ഗവൺമെന്റ് ഡെന്റൽ കോളേജിൽ വച്ച് നടക്കുന്നു. ആകെ 800 ആളുകൾ. അവർ 400 പേർ അസുഖമുള്ളവർ, ബാക്കി നാനൂറ് പേർ അസുഖം ഇല്ലാത്തവർ. ഗവേഷണം രണ്ടുഘട്ടങ്ങളായാണ് നടക്കുക. രണ്ടു വിഭാഗക്കാരിലും ഒരേ പഠന നടപടികളാണ് കൈകൊള്ളുക.

1. ആദ്യഘട്ടത്തിൽ വിവരങ്ങൾ ആശുപത്രിരേഖകളിൽ നിന്നും ശേഖരിക്കുന്നു. ഉദാഹരണത്തിന് ഒന്നാം വിഭാഗക്കാരായ ആളുകളുടെ അസുഖസംബന്ധിയായ വിശദാംശങ്ങൾ അന്വേഷിക്കും, രണ്ടാംവിഭാഗക്കാരായവർ ആശുപത്രികളിൽ പോകുവാനുണ്ടായ സാഹചര്യങ്ങളെക്കുറിച്ച് തിരക്കും.

2. പഠനത്തിന്റെ രണ്ടാംഘട്ടം അഭിമുഖമാണ്. ചോദ്യാവലിയുടെ സഹായത്തോടെ, തൊഴിൽ, ജീവിതസാഹചര്യങ്ങൾ തുടങ്ങി ജീവിതത്തിന്റെ വിവിധ തുറകളെക്കുറിച്ചുള്ള സൂക്ഷ്മ വിവരം ലഭ്യമാക്കുന്നു. അഭിമുഖത്തിന്റെ ദൈർഘ്യം രണ്ടു മണിക്കൂർ ആണ്.

ഞാൻ ഈ പഠനത്തിൽ പങ്കുചേർന്നാൽ എങ്ങനെ അതുമായി ബന്ധപ്പെട്ടിരിക്കും

ഈ പഠനത്തിൽ പങ്കെടുക്കുക എന്നുവെച്ചാൽ നമ്മുടെ ആശുപത്രി രേഖകൾ പരിശോധിക്കുവാൻ അനുവദിക്കുക എന്നും അഭിമുഖത്തിൽ പങ്കെടുക്കുക എന്നും ആണ്.

പഠനവുമായി ബന്ധപ്പെട്ട് എന്തെങ്കിലും അപകടകരമായ സാഹചര്യങ്ങളോ, അസൗകര്യങ്ങളോ നിലനിൽക്കുന്നുണ്ടോ ?

ഈ പഠനവുമായി ബന്ധപ്പെട്ട് യാതൊരു അപകടവും നിലനിൽക്കുന്നില്ല. അപകടമോ, അസ്വസ്ഥതയോ ഉളവാക്കുന്ന ഒരു ചികിത്സാരീതിയും ഇതിൽ ഇല്ല. ആശുപത്രിരേഖകളും അഭിമുഖവും വഴി വിവരങ്ങൾ ശേഖരിക്കുക മാത്രമെ ചെയ്യുന്നുള്ളു.

പഠനവുമായി ബന്ധപ്പെട്ട് എനിക്ക് എന്തെങ്കിലും മെച്ചം ലഭിക്കുമോ ?

പങ്കെടുക്കുന്നവർക്ക് നേരിട്ട് യാതൊരു മെച്ചവും ലഭിക്കുന്നതല്ല. എങ്കിലും ഈ പഠനത്തിന്റെ ഫലം വായിലേയും, തൊണ്ടയിലേയും അർബുദസംബന്ധമായി കൂടുതൽ വിവരങ്ങൾ നമുക്ക് പ്രദാനം ചെയ്യുമെന്ന് പ്രത്യാശിക്കാം.

പഠനത്തിൽ പങ്കെടുക്കുന്നത് സംബന്ധിച്ച്

ഈ പഠനപദ്ധതിയിൽ പങ്കെടുക്കേണ്ടത് സ്വമേധയാ ആണ്

ഈ ഗവേഷണത്തിൽ പങ്കെടുക്കുന്നത് എന്റെ ചികിത്സയെ ബാധിക്കുമോ ?

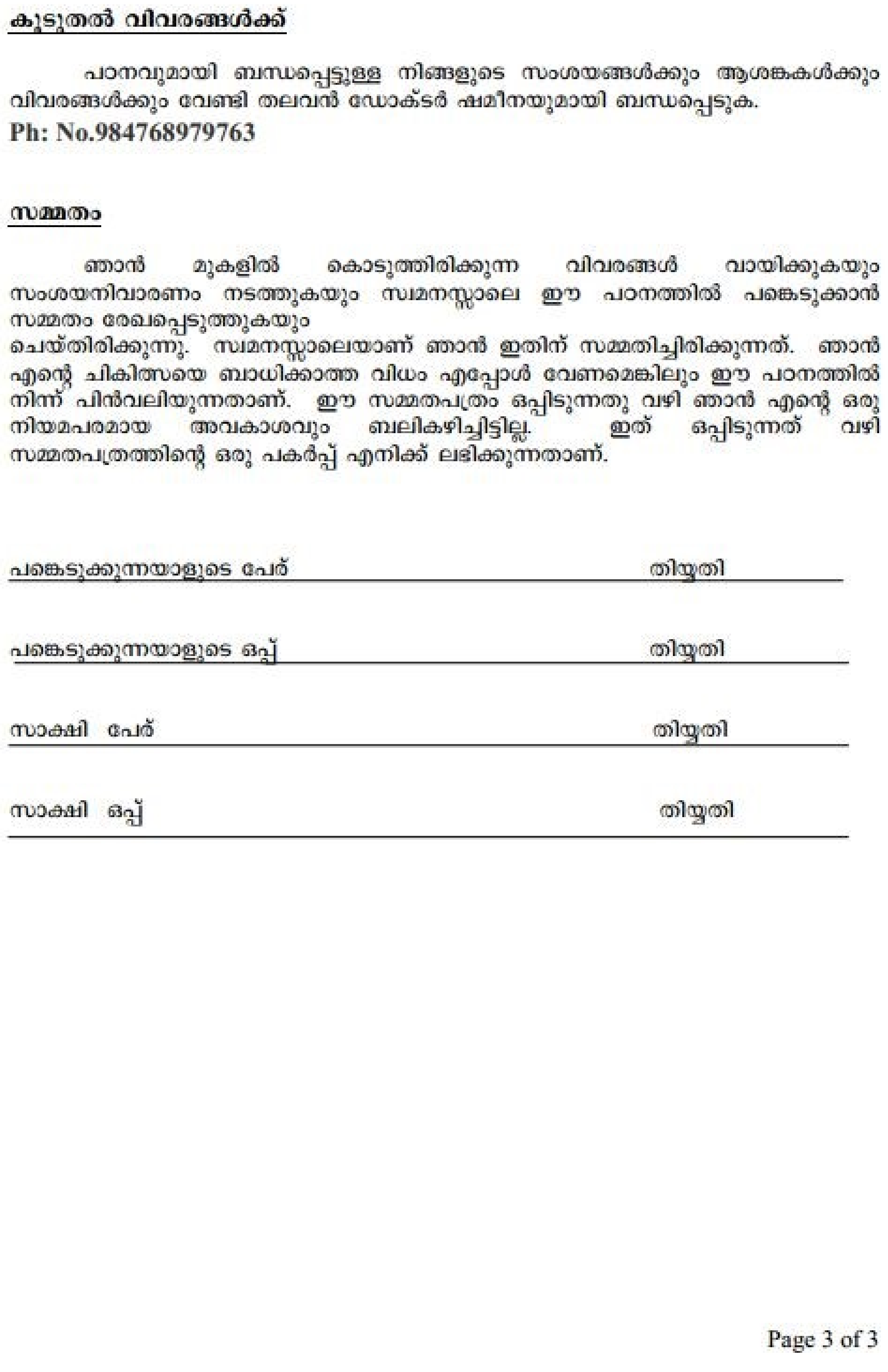
ചികിത്സയേയോ ചികിത്സാനന്തരനടപടികളെയും പഠനം യാതൊരു കാരണവശാലും ബാധിക്കുന്നതല്ല.

ഈ പഠനത്തിൽ നിന്ന് പിൻവാങ്ങണമെന്ന് കരുതിയാൽ അതിന്റെ അനന്തരഫലങ്ങൾ എന്തായിരിക്കും ?

ഈ ഗവേഷണപദ്ധതിയിൽ നിന്ന് ഏതു സമയത്തും അതായത് അഭിമുഖത്തിന്റെ പകുതിയിൽ വെച്ച് പോലും പിൻമാറാനുള്ള പൂർണ്ണ അവകാശം നിങ്ങൾക്കുണ്ട്. അത് നിങ്ങളുടെ ചികിത്സയെ ഒരു കാരണവശാലും ബാധിക്കുന്നതല്ല.

പഠനത്തിന്റെ വിശ്വസ്തത

ഗവേഷണവേളയിൽ നിങ്ങൾ നൽകുന്ന വിവരങ്ങൾ പൂർണ്ണരഹസ്യസ്വഭാവത്തോടെ സൂക്ഷിക്കുന്നതായിരിക്കും എന്ന് ഉറപ്പ് തരുന്നു. നിങ്ങൾ നൽകിയ വിവരങ്ങളുമ ായി ഗവേഷകർക്കും, വിവരം ശേഖരിക്കുന്ന ഗവേഷകസഹായികൾക്കുമല്ലാതെ മറ്റാർക്കും പ്രാപ്യത ഉണ്ടായിരിക്കുന്നതല്ല. പ്രസ്തുത വിവരങ്ങൾ ഗവേഷണകാര്യാലയത്തിൽ ഭദ്രമായി പൂട്ടി സൂക്ഷിക്കുന്നതാണ്. കൂടാതെ നിങ്ങൾ നൽകുന്ന വിവരങ്ങൾ ഒരു രഹസ്യ അക്കം ഉപയോഗിച്ച് ഏകോപിപ്പിക്കുന്നതിനാൽ അവയെ വ്യക്തിപരമായി ആരുടേതെന്ന് തിരിച്ചറിയാൻ സാധ്യമല്ല. ഗവേഷണഫലം ശാസ്ത്രമാസികകളിൽ പ്രസിദ്ധീകരണത്തിന് നൽകുമ്പോൾ വ്യക്തിപരമായി തിരിച്ചറിയാത്ത രീതിയിലാണ് നൽകുക. ഗവേഷണസംബന്ധമായ എല്ലാ വിവരങ്ങളും 5 വർഷത്തേക്ക് സൂക്ഷിച്ച് വെക്കുകയും അതിന് ശേഷം നശിപ്പിച്ചു കളയുകയും ചെയ്യുന്നതാണ്.



കൂടുതൽ വിവരങ്ങൾക്ക്

പഠനവുമായി ബന്ധപ്പെട്ടുള്ള നിങ്ങളുടെ സംശയങ്ങൾക്കും ആശങ്കകൾക്കും വിവരങ്ങൾക്കും വേണ്ടി തലവൻ ഡോക്ടർ ഷമീനയുമായി ബന്ധപ്പെടുക.

Ph: No.984768979763

സമ്മതം

ഞാൻ മുകളിൽ കൊടുത്തിരിക്കുന്ന വിവരങ്ങൾ വായിക്കുകയും സംശയനിവാരണം നടത്തുകയും സ്വമനസ്സാലെ ഈ പഠനത്തിൽ പങ്കെടുക്കാൻ സമ്മതം രേഖപ്പെടുത്തുകയും
ചെയ്തിരിക്കുന്നു. സ്വമനസ്സാലെയാണ് ഞാൻ ഇതിന് സമ്മതിച്ചിരിക്കുന്നത്. ഞാൻ എന്റെ ചികിത്സയെ ബാധിക്കാത്ത വിധം എപ്പോൾ വേണമെങ്കിലും ഈ പഠനത്തിൽ നിന്ന് പിൻവലിയുന്നതാണ്. ഈ സമ്മതപത്രം ഒപ്പിടുന്നതു വഴി ഞാൻ എന്റെ ഒരു നിയമപരമായ അവകാശവും ബലികഴിച്ചിട്ടില്ല. ഇത് ഒപ്പിടുന്നത് വഴി സമ്മതപത്രത്തിന്റെ ഒരു പകർപ്പ് എനിക്ക് ലഭിക്കുന്നതാണ്.

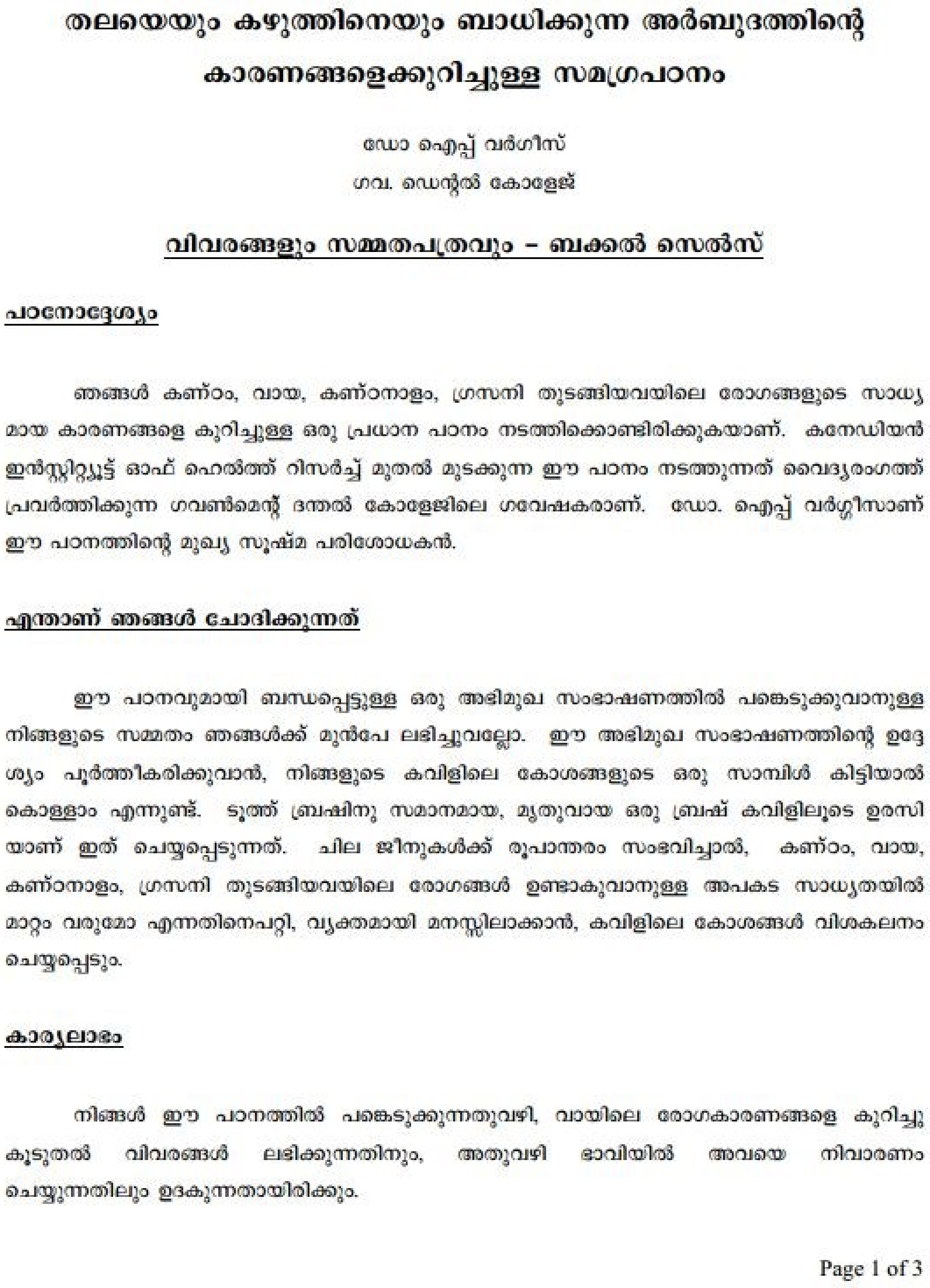
പങ്കെടുക്കുന്നയാളുടെ പേര് തിയ്യതി

പങ്കെടുക്കുന്നയാളുടെ ഒപ്പ് തിയ്യതി

സാക്ഷി പേര് തിയ്യതി

സാക്ഷി ഒപ്പ് തിയ്യതി

Page 3 of 3



തലയെയും കഴുത്തിനെയും ബാധിക്കുന്ന അർബുദത്തിന്റെ കാരണങ്ങളെക്കുറിച്ചുള്ള സമഗ്രപഠനം

ഡോ ഐപ്പ് വർഗീസ്
ഗവ. ഡെന്റൽ കോളേജ്

വിവരങ്ങളും സമ്മതപത്രവും – ബക്കൽ സെൽസ്

പഠനോദ്ദേശ്യം

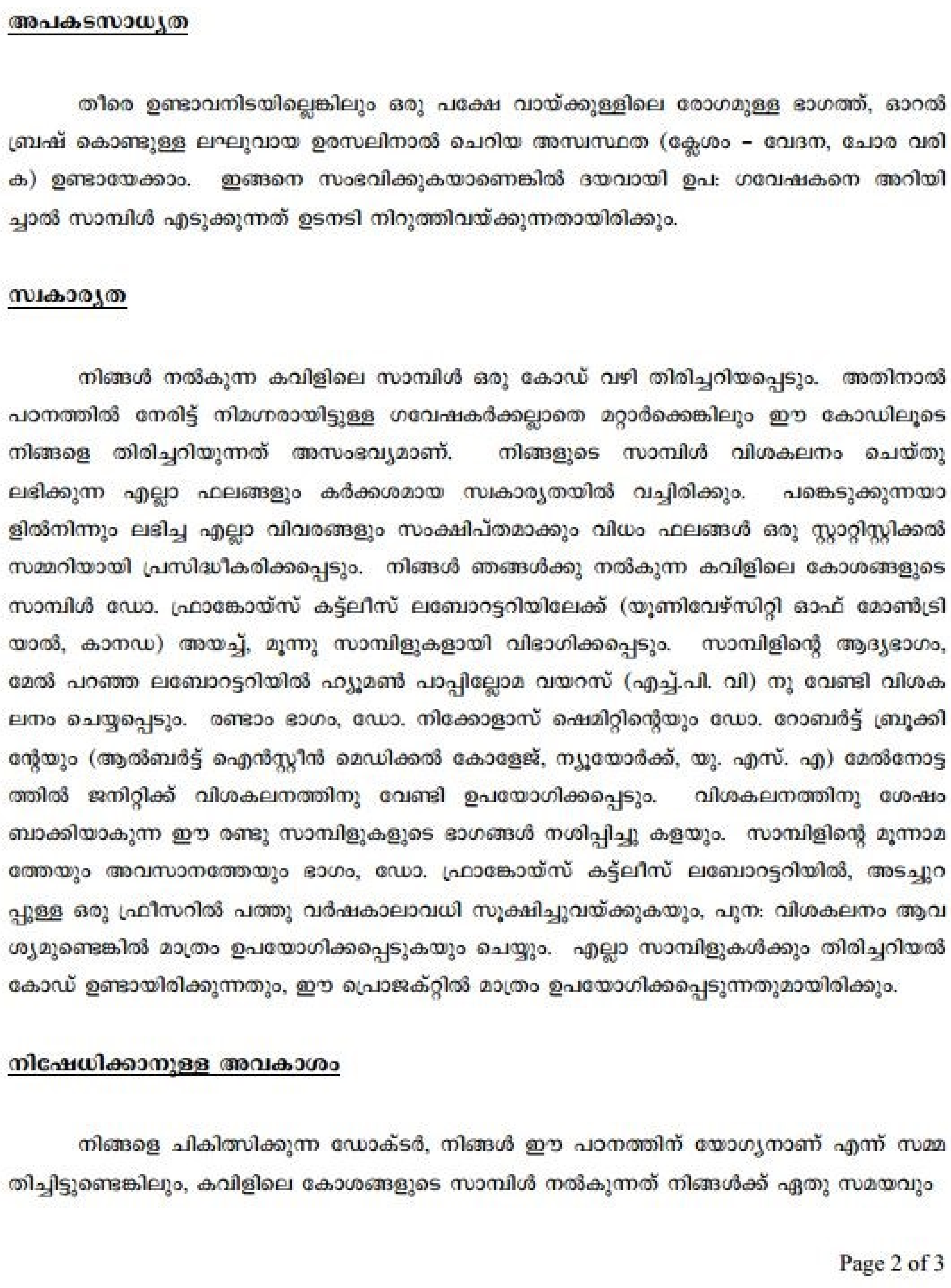
ഞങ്ങൾ കണ്ഠം, വായ, കണ്ഠനാളം, ഗ്രസനി തുടങ്ങിയവയിലെ രോഗങ്ങളുടെ സാധ്യമായ കാരണങ്ങളെ കുറിച്ചുള്ള ഒരു പ്രധാന പഠനം നടത്തിക്കൊണ്ടിരിക്കുകയാണ്. കനേഡിയൻ ഇൻസ്റ്റിറ്റ്യൂട്ട് ഓഫ് ഹെൽത്ത് റിസർച്ച് മുതൽ മുടക്കുന്ന ഈ പഠനം നടത്തുന്നത് വൈദ്യരംഗത്ത് പ്രവർത്തിക്കുന്ന ഗവൺമെന്റ് ദന്തൽ കോളേജിലെ ഗവേഷകരാണ്. ഡോ. ഐപ്പ് വർഗ്ഗീസാണ് ഈ പഠനത്തിന്റെ മുഖ്യ സൂക്ഷ്മ പരിശോധകൻ.

എന്താണ് ഞങ്ങൾ ചോദിക്കുന്നത്

ഈ പഠനവുമായി ബന്ധപ്പെട്ടുള്ള ഒരു അഭിമുഖ സംഭാഷണത്തിൽ പങ്കെടുക്കുവാനുള്ള നിങ്ങളുടെ സമ്മതം ഞങ്ങൾക്ക് മുൻപേ ലഭിച്ചുവല്ലോ. ഈ അഭിമുഖ സംഭാഷണത്തിന്റെ ഉദ്ദേശ്യം പൂർത്തീകരിക്കുവാൻ, നിങ്ങളുടെ കവിളിലെ കോശങ്ങളുടെ ഒരു സാമ്പിൾ കിട്ടിയാൽ കൊള്ളാം എന്നുണ്ട്. ടൂത്ത് ബ്രഷിനു സമാനമായ, മൃതുവായ ഒരു ബ്രഷ് കവിളിലൂടെ ഉരസിയാണ് ഇത് ചെയ്യപ്പെടുന്നത്. ചില ജീനുകൾക്ക് രൂപാന്തരം സംഭവിച്ചാൽ, കണ്ഠം, വായ, കണ്ഠനാളം, ഗ്രസനി തുടങ്ങിയവയിലെ രോഗങ്ങൾ ഉണ്ടാകുവാനുള്ള അപകട സാധ്യതയിൽ മാറ്റം വരുമോ എന്നതിനെപറ്റി, വ്യക്തമായി മനസ്സിലാക്കാൻ, കവിളിലെ കോശങ്ങൾ വിശകലനം ചെയ്യപ്പെടും.

കാര്യലാഭം

നിങ്ങൾ ഈ പഠനത്തിൽ പങ്കെടുക്കുന്നതുവഴി, വായിലെ രോഗകാരണങ്ങളെ കുറിച്ചു കൂടുതൽ വിവരങ്ങൾ ലഭിക്കുന്നതിനും, അതുവഴി ഭാവിയിൽ അവയെ നിവാരണം ചെയ്യുന്നതിലും ഉദകുന്നതായിരിക്കും.



അപകടസാധ്യത

തീരെ ഉണ്ടാവനിടയില്ലെങ്കിലും ഒരു പക്ഷേ വായ്ക്കുള്ളിലെ രോഗമുള്ള ഭാഗത്ത്, ഓറൽ ബ്രഷ് കൊണ്ടുള്ള ലഘുവായ ഉരസലിനാൽ ചെറിയ അസ്വസ്ഥത (ക്ലേശം - വേദന, ചോര വരിക) ഉണ്ടായേക്കാം. ഇങ്ങനെ സംഭവിക്കുകയാണെങ്കിൽ ദയവായി ഉപ: ഗവേഷകനെ അറിയിച്ചാൽ സാമ്പിൾ എടുക്കുന്നത് ഉടനടി നിറുത്തിവയ്ക്കുന്നതായിരിക്കും.

സ്വകാര്യത

നിങ്ങൾ നൽകുന്ന കവിളിലെ സാമ്പിൾ ഒരു കോഡ് വഴി തിരിച്ചറിയപ്പെടും. അതിനാൽ പഠനത്തിൽ നേരിട്ട് നിമഗ്നരായിട്ടുള്ള ഗവേഷകർക്കല്ലാതെ മറ്റാർക്കെങ്കിലും ഈ കോഡിലൂടെ നിങ്ങളെ തിരിച്ചറിയുന്നത് അസംഭവ്യമാണ്. നിങ്ങളുടെ സാമ്പിൾ വിശകലനം ചെയ്തു ലഭിക്കുന്ന എല്ലാ ഫലങ്ങളും കർക്കശമായ സ്വകാര്യതയിൽ വച്ചിരിക്കും. പങ്കെടുക്കുന്നയാളിൽനിന്നും ലഭിച്ച എല്ലാ വിവരങ്ങളും സംക്ഷിപ്തമാക്കും വിധം ഫലങ്ങൾ ഒരു സ്റ്റാറ്റിസ്റ്റിക്കൽ സമ്മറിയായി പ്രസിദ്ധീകരിക്കപ്പെടും. നിങ്ങൾ ഞങ്ങൾക്കു നൽകുന്ന കവിളിലെ കോശങ്ങളുടെ സാമ്പിൾ ഡോ. ഫ്രാങ്കോയ്സ് കട്ട്ലീസ് ലബോറട്ടറിയിലേക്ക് (യൂണിവേഴ്സിറ്റി ഓഫ് മോൺട്രിയാൽ, കാനഡ) അയച്ച്, മൂന്നു സാമ്പിളുകളായി വിഭാഗിക്കപ്പെടും. സാമ്പിളിന്റെ ആദ്യഭാഗം, മേൽ പറഞ്ഞ ലബോറട്ടറിയിൽ ഹ്യൂമൺ പാപ്പില്ലോമ വയറസ് (എച്ച്.പി. വി) നു വേണ്ടി വിശകലനം ചെയ്യപ്പെടും. രണ്ടാം ഭാഗം, ഡോ. നിക്കോളാസ് ഷെമിറ്റിന്റെയും ഡോ. റോബർട്ട് ബ്രൂക്കിന്റേയും (ആൽബർട്ട് ഐൻസ്റ്റീൻ മെഡിക്കൽ കോളേജ്, ന്യൂയോർക്ക്, യു. എസ്. എ) മേൽനോട്ടത്തിൽ ജനിറ്റിക്ക് വിശകലനത്തിനു വേണ്ടി ഉപയോഗിക്കപ്പെടും. വിശകലനത്തിനു ശേഷം ബാക്കിയാകുന്ന ഈ രണ്ടു സാമ്പിളുകളുടെ ഭാഗങ്ങൾ നശിപ്പിച്ചു കളയും. സാമ്പിളിന്റെ മൂന്നാമത്തേയും അവസാനത്തേയും ഭാഗം, ഡോ. ഫ്രാങ്കോയ്സ് കട്ട്ലീസ് ലബോറട്ടറിയിൽ, അടച്ചുറപ്പുള്ള ഒരു ഫ്രീസറിൽ പത്തു വർഷകാലാവധി സൂക്ഷിച്ചുവയ്ക്കുകയും, പുന: വിശകലനം ആവശ്യമുണ്ടെങ്കിൽ മാത്രം ഉപയോഗിക്കപ്പെടുകയും ചെയ്യും. എല്ലാ സാമ്പിളുകൾക്കും തിരിച്ചറിയൽ കോഡ് ഉണ്ടായിരിക്കുന്നതും, ഈ പ്രൊജക്റ്റിൽ മാത്രം ഉപയോഗിക്കപ്പെടുന്നതുമായിരിക്കും.

നിഷേധിക്കാനുള്ള അവകാശം

നിങ്ങളെ ചികിത്സിക്കുന്ന ഡോക്ടർ, നിങ്ങൾ ഈ പഠനത്തിന് യോഗ്യനാണ് എന്ന് സമ്മതിച്ചിട്ടുണ്ടെങ്കിലും, കവിളിലെ കോശങ്ങളുടെ സാമ്പിൾ നൽകുന്നത് നിങ്ങൾക്ക് ഏതു സമയവും

Page 2 of 3

പ്രതികൂലമായ പരിണിതഫലങ്ങളൊന്നും തന്നെയില്ലാതെ, നിഷേധിക്കുവാനുള്ള അവാകാശമുണ്ട്. ഈ പഠനത്തിൽ നിന്നും പിൻവാങ്ങുവാൻ നിങ്ങൾ തീരുമാനിക്കുകയാണെങ്കിൽ നിങ്ങൾ നൽകിയിരിക്കുന്ന കവിളിലെ കോശങ്ങളുടെ സാമ്പിൾ അപ്പോൾ തന്നെ നശിപ്പിക്കപ്പെടും.

കൂടുതൽ വിവരങ്ങൾക്ക്

ഈ പഠനത്തെക്കുറിച്ച് നിങ്ങൾക്ക് കൂടുതൽ വിവരങ്ങൾ അറിയണമെന്നുണ്ടെങ്കിൽ, അല്ലെങ്കിൽ എന്തെങ്കിലും ചോദ്യങ്ങളുണ്ടെങ്കിൽ ഈ പ്രൊജക്ടിന്റെ ഉപ:ഗവേഷകനുമായി ബന്ധപ്പെടാൻ മടിക്കരുത്.

സമ്മതം

ഞാൻ, ഈ പഠനത്തിലെ പൊതുവായ ഉദ്ദേശ്യങ്ങളെക്കുറിച്ചും, ഈ പഠനത്തിൽ പങ്കെടുക്കുന്നതു വഴി എനിക്കുള്ള അവകാശങ്ങൾ, ആവശ്യങ്ങൾ എന്നിവയെക്കുറിച്ചും ബോധവാനാണ്. ആയതിനാൽ ഞാൻ ഈ പഠനത്തിൽ പങ്കെടുക്കുവാനും, എന്റെ കവിളിലെ കോശങ്ങളുടെ ഒരു സാമ്പിൾ നൽകുവാനും, സ്വമനസ്സാൽ സമ്മതിക്കുന്നു. എന്റെ പങ്കുചേരൽ പൂർണ്ണമായും പരപ്രേരണ കൂടാതെയാണ്. എന്റെ വൈദ്യപരിശോധനയും അനുധാവനവും ബാധിക്കാതെ എനിക്ക് പിൻവാങ്ങാവുന്നതാണ്.

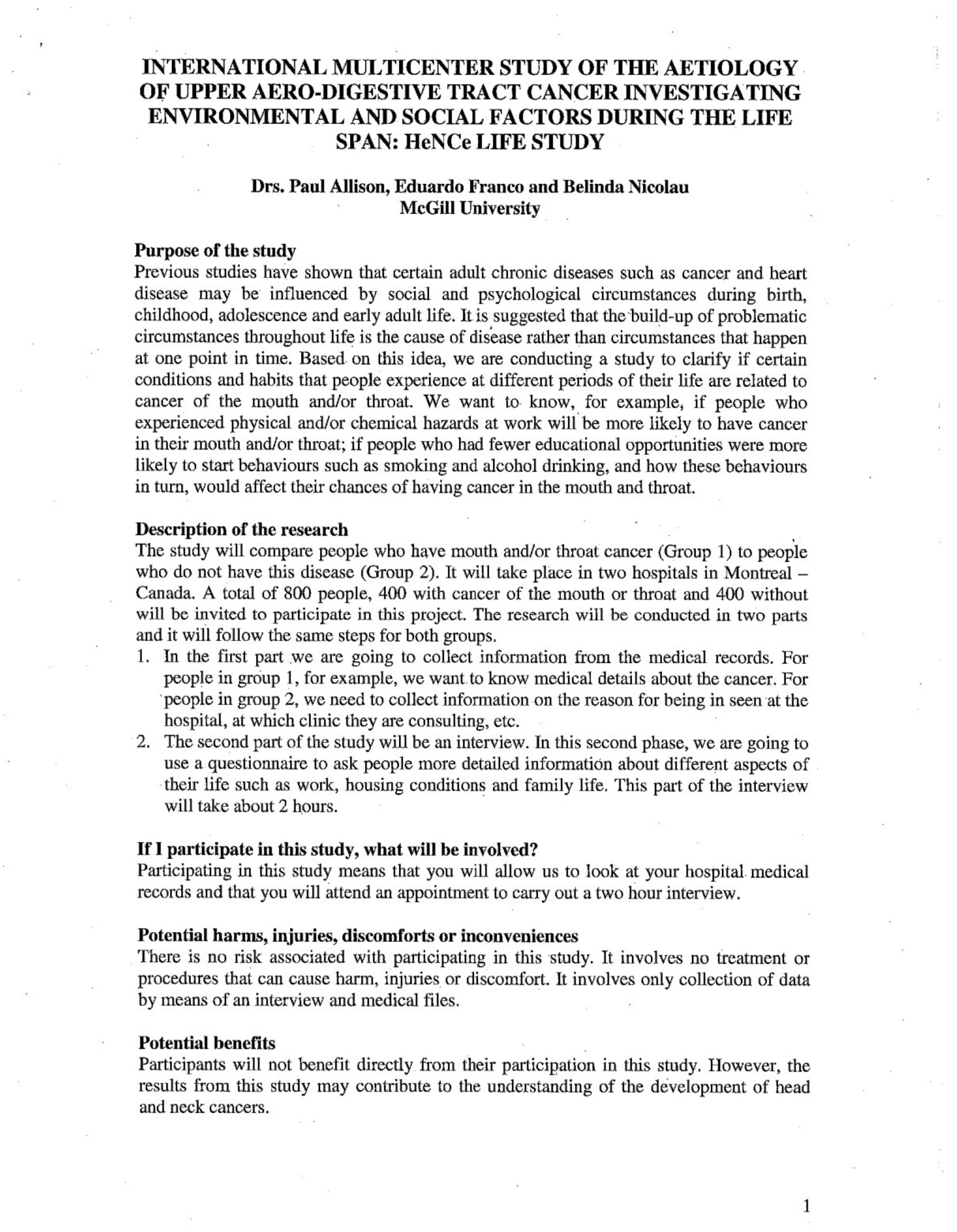
പങ്കെടുക്കുന്നയാളുടെ പേര് ______________________________ തിയ്യതി __________

പങ്കെടുക്കുന്നയാളുടെ ഒപ്പ് ______________________________ തിയ്യതി __________

സമ്മതപത്രം വിശദീകരിച്ചു തിയ്യതി
____
കൊടുത്തയാളുടെ പേര്

സമ്മതപത്രം വിശദീകരിച്ചു തിയ്യതി
____
കൊടുത്തയാളുടെ ഒപ്പ്

Page 3 of 3



INTERNATIONAL MULTICENTER STUDY OF THE AETIOLOGY OF UPPER AERO-DIGESTIVE TRACT CANCER INVESTIGATING ENVIRONMENTAL AND SOCIAL FACTORS DURING THE LIFE SPAN: HeNCe LIFE STUDY

Drs. Paul Allison, Eduardo Franco and Belinda Nicolau
McGill University

Purpose of the study

Previous studies have shown that certain adult chronic diseases such as cancer and heart disease may be influenced by social and psychological circumstances during birth, childhood, adolescence and early adult life. It is suggested that the build-up of problematic circumstances throughout life is the cause of disease rather than circumstances that happen at one point in time. Based on this idea, we are conducting a study to clarify if certain conditions and habits that people experience at different periods of their life are related to cancer of the mouth and/or throat. We want to know, for example, if people who experienced physical and/or chemical hazards at work will be more likely to have cancer in their mouth and/or throat; if people who had fewer educational opportunities were more likely to start behaviours such as smoking and alcohol drinking, and how these behaviours in turn, would affect their chances of having cancer in the mouth and throat.

Description of the research

The study will compare people who have mouth and/or throat cancer (Group 1) to people who do not have this disease (Group 2). It will take place in two hospitals in Montreal – Canada. A total of 800 people, 400 with cancer of the mouth or throat and 400 without will be invited to participate in this project. The research will be conducted in two parts and it will follow the same steps for both groups.

1. In the first part we are going to collect information from the medical records. For people in group 1, for example, we want to know medical details about the cancer. For people in group 2, we need to collect information on the reason for being in seen at the hospital, at which clinic they are consulting, etc.
2. The second part of the study will be an interview. In this second phase, we are going to use a questionnaire to ask people more detailed information about different aspects of their life such as work, housing conditions and family life. This part of the interview will take about 2 hours.

If I participate in this study, what will be involved?

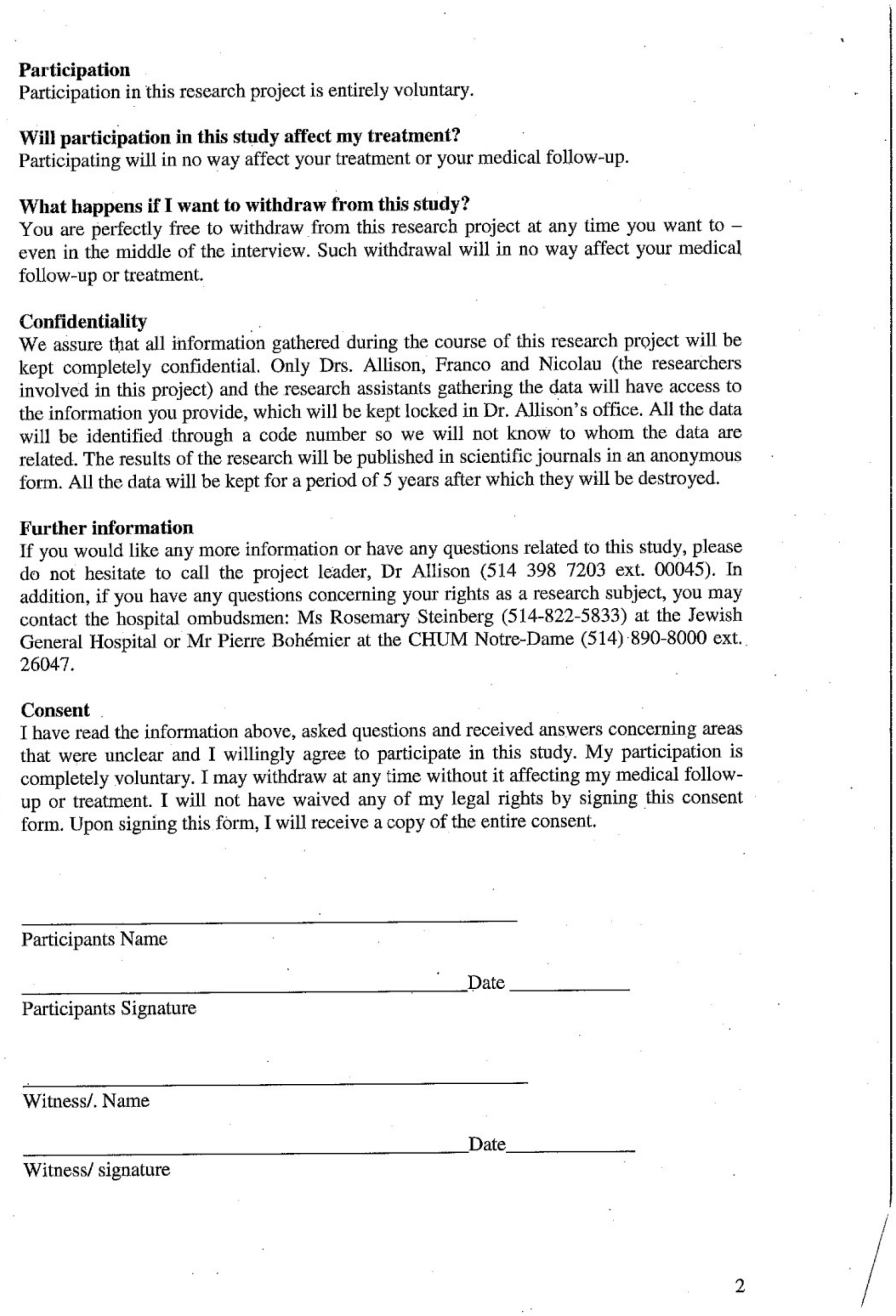
Participating in this study means that you will allow us to look at your hospital medical records and that you will attend an appointment to carry out a two hour interview.

Potential harms, injuries, discomforts or inconveniences

There is no risk associated with participating in this study. It involves no treatment or procedures that can cause harm, injuries or discomfort. It involves only collection of data by means of an interview and medical files.

Potential benefits

Participants will not benefit directly from their participation in this study. However, the results from this study may contribute to the understanding of the development of head and neck cancers.



Participation

Participation in this research project is entirely voluntary.

Will participation in this study affect my treatment?

Participating will in no way affect your treatment or your medical follow-up.

What happens if I want to withdraw from this study?

You are perfectly free to withdraw from this research project at any time you want to – even in the middle of the interview. Such withdrawal will in no way affect your medical follow-up or treatment.

Confidentiality

We assure that all information gathered during the course of this research project will be kept completely confidential. Only Drs. Allison, Franco and Nicolau (the researchers involved in this project) and the research assistants gathering the data will have access to the information you provide, which will be kept locked in Dr. Allison's office. All the data will be identified through a code number so we will not know to whom the data are related. The results of the research will be published in scientific journals in an anonymous form. All the data will be kept for a period of 5 years after which they will be destroyed.

Further information

If you would like any more information or have any questions related to this study, please do not hesitate to call the project leader, Dr Allison (514 398 7203 ext. 00045). In addition, if you have any questions concerning your rights as a research subject, you may contact the hospital ombudsmen: Ms Rosemary Steinberg (514-822-5833) at the Jewish General Hospital or Mr Pierre Bohémier at the CHUM Notre-Dame (514) 890-8000 ext. 26047.

Consent

I have read the information above, asked questions and received answers concerning areas that were unclear and I willingly agree to participate in this study. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment. I will not have waived any of my legal rights by signing this consent form. Upon signing this form, I will receive a copy of the entire consent.

____

Participants Name

____Date ___________

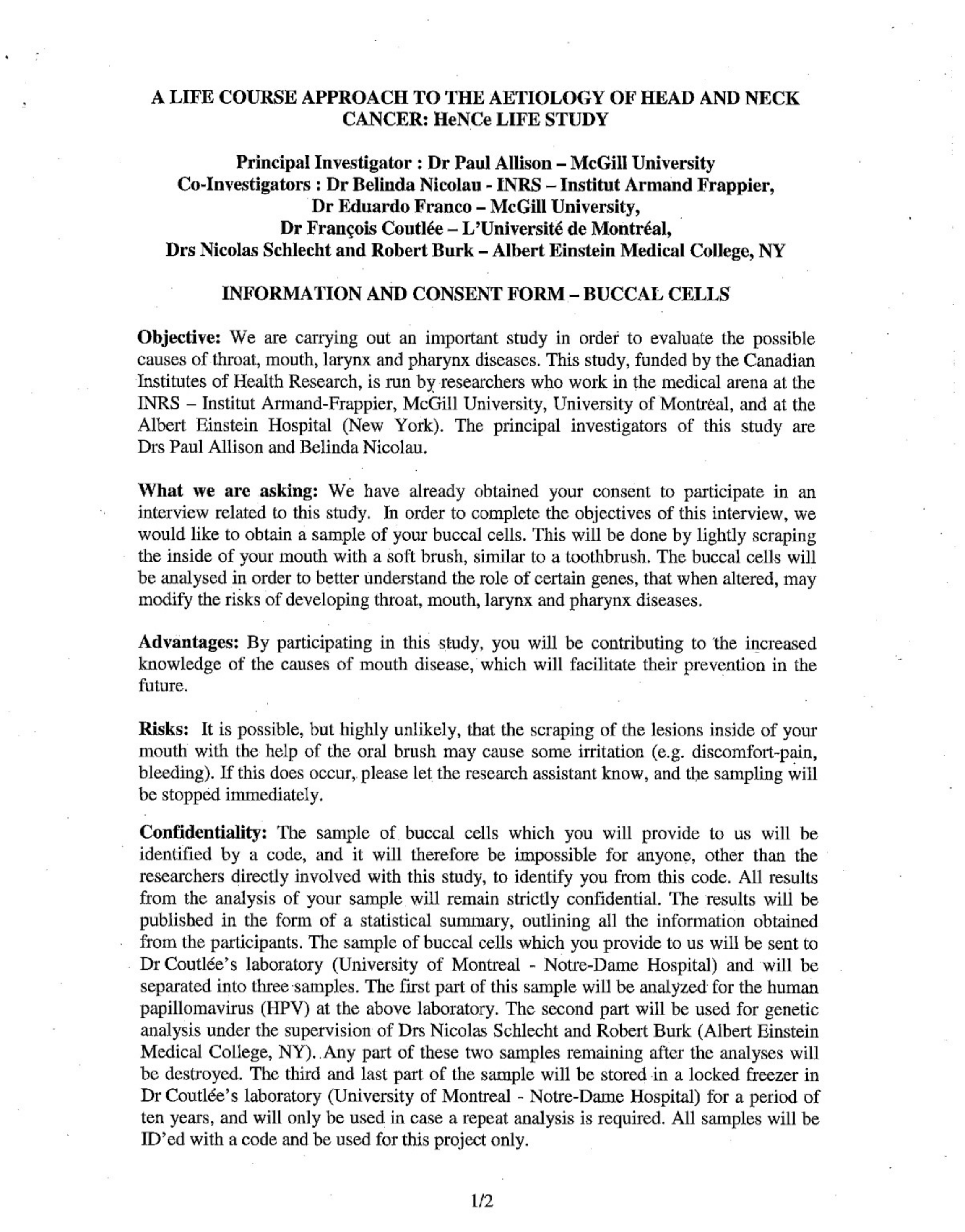
Participants Signature

____

Witness/. Name

____Date___________

Witness/ signature



A LIFE COURSE APPROACH TO THE AETIOLOGY OF HEAD AND NECK CANCER: HeNCe LIFE STUDY

Principal Investigator : Dr Paul Allison – McGill University
Co-Investigators : Dr Belinda Nicolau - INRS – Institut Armand Frappier,
Dr Eduardo Franco – McGill University,
Dr François Coutlée – L'Université de Montréal,
Drs Nicolas Schlecht and Robert Burk – Albert Einstein Medical College, NY

INFORMATION AND CONSENT FORM – BUCCAL CELLS

Objective: We are carrying out an important study in order to evaluate the possible causes of throat, mouth, larynx and pharynx diseases. This study, funded by the Canadian Institutes of Health Research, is run by researchers who work in the medical arena at the INRS – Institut Armand-Frappier, McGill University, University of Montreal, and at the Albert Einstein Hospital (New York). The principal investigators of this study are Drs Paul Allison and Belinda Nicolau.

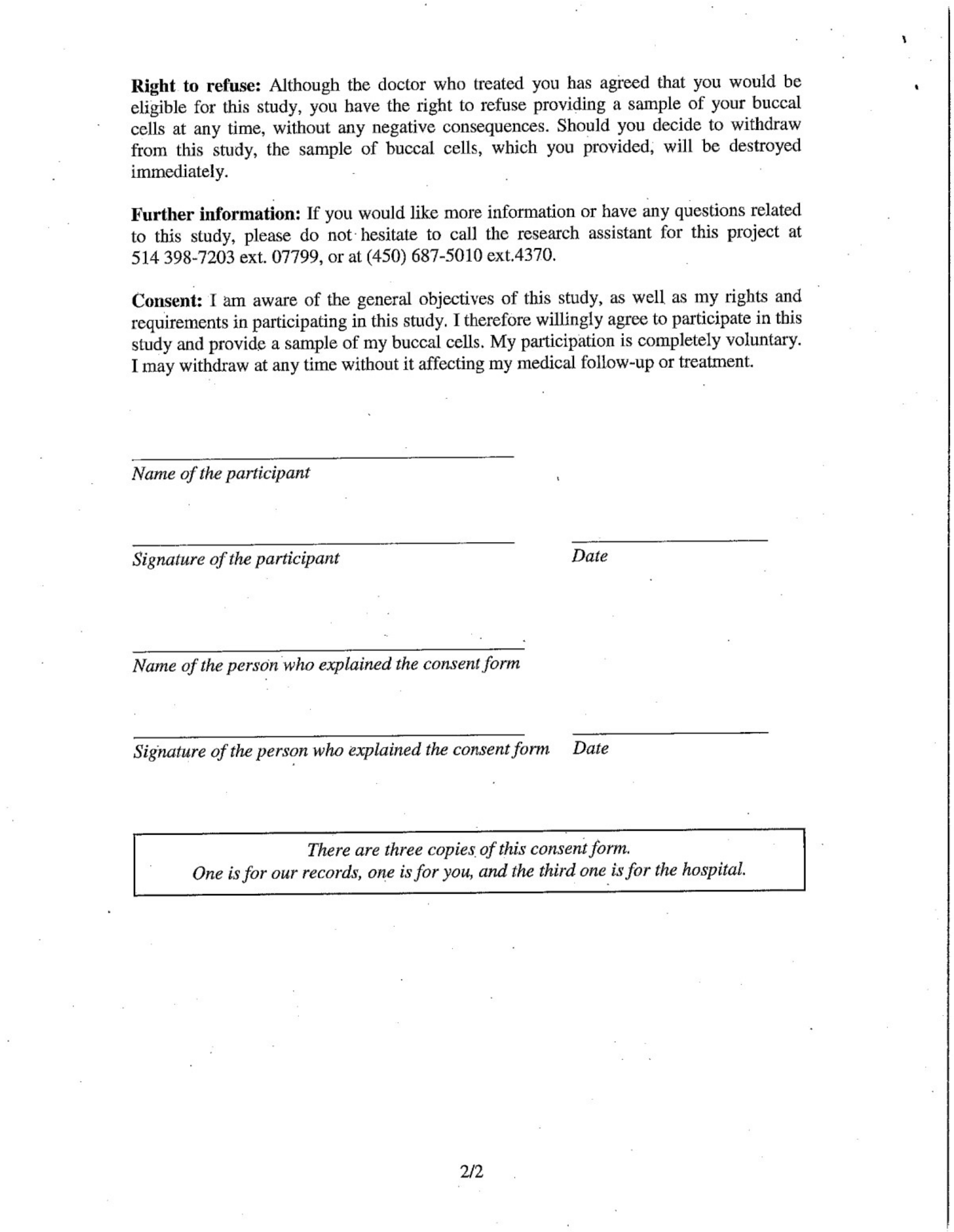
What we are asking: We have already obtained your consent to participate in an interview related to this study. In order to complete the objectives of this interview, we would like to obtain a sample of your buccal cells. This will be done by lightly scraping the inside of your mouth with a soft brush, similar to a toothbrush. The buccal cells will be analysed in order to better understand the role of certain genes, that when altered, may modify the risks of developing throat, mouth, larynx and pharynx diseases.

Advantages: By participating in this study, you will be contributing to the increased knowledge of the causes of mouth disease, which will facilitate their prevention in the future.

Risks: It is possible, but highly unlikely, that the scraping of the lesions inside of your mouth with the help of the oral brush may cause some irritation (e.g. discomfort-pain, bleeding). If this does occur, please let the research assistant know, and the sampling will be stopped immediately.

Confidentiality: The sample of buccal cells which you will provide to us will be identified by a code, and it will therefore be impossible for anyone, other than the researchers directly involved with this study, to identify you from this code. All results from the analysis of your sample will remain strictly confidential. The results will be published in the form of a statistical summary, outlining all the information obtained from the participants. The sample of buccal cells which you provide to us will be sent to Dr Coutlée's laboratory (University of Montreal - Notre-Dame Hospital) and will be separated into three samples. The first part of this sample will be analyzed for the human papillomavirus (HPV) at the above laboratory. The second part will be used for genetic analysis under the supervision of Drs Nicolas Schlecht and Robert Burk (Albert Einstein Medical College, NY). Any part of these two samples remaining after the analyses will be destroyed. The third and last part of the sample will be stored in a locked freezer in Dr Coutlée's laboratory (University of Montreal - Notre-Dame Hospital) for a period of ten years, and will only be used in case a repeat analysis is required. All samples will be ID'ed with a code and be used for this project only.

1/2



Right to refuse: Although the doctor who treated you has agreed that you would be eligible for this study, you have the right to refuse providing a sample of your buccal cells at any time, without any negative consequences. Should you decide to withdraw from this study, the sample of buccal cells, which you provided, will be destroyed immediately.

Further information: If you would like more information or have any questions related to this study, please do not hesitate to call the research assistant for this project at 514 398-7203 ext. 07799, or at (450) 687-5010 ext.4370.

Consent: I am aware of the general objectives of this study, as well as my rights and requirements in participating in this study. I therefore willingly agree to participate in this study and provide a sample of my buccal cells. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment.

Name of the participant

Signature of the participant *Date*

Name of the person who explained the consent form

Signature of the person who explained the consent form *Date*

There are three copies of this consent form.
One is for our records, one is for you, and the third one is for the hospital.

2/2

LE PARCOURS DE VIE COMME APPROCHE À L'ÉTUDE DE L'ÉTIOLOGIE DU CANCER DE LA BOUCHE ET DE LA GORGE

Drs Paul Allison, Eduardo Franco et Belinda Nicolau
Université McGill

Objectif de l'étude

Des études antérieures ont démontré que certaines maladies chroniques chez les adultes, telles que le cancer et les maladies du cœur, peuvent être influencées par des circonstances psychologiques et sociales lors de la naissance, l'enfance, l'adolescence et le début de l'âge adulte. Il est suggéré que ce soit le cumul de circonstances problématiques au cours de la vie qui cause la maladie plutôt que des circonstances présentent à un moment isolé dans le temps. En se basant sur cette idée, nous menons une étude qui vise à clarifier si certaines conditions et habitudes dont les gens font l'expérience à différentes périodes au cours de leur vie, sont en lien avec le cancer de la bouche et de la gorge. Nous cherchons à savoir si, par exemple, les personnes qui sont exposés à des dangers physiques et/ou chimiques dans leur milieu de travail auront une plus grande probabilité de développer un cancer dans leur bouche ou leur gorge, si les personnes qui ont eu moins d'opportunités pour se scolariser étaient plus susceptibles de développer des comportements comme fumer et boire de l'alcool, et comment à leur tour, ces comportements affecteraient leurs chances d'avoir un cancer dans la bouche ou la gorge.

Modalités et Procédures

L'étude comparera des personnes qui ont un cancer de la bouche et/ou de la gorge (groupe 1) à des personnes qui n'ont pas cette maladie (groupe 2). Elle se tiendra dans deux centres hospitaliers de Montréal, au Canada. Au total, 800 individus (400 ayant un cancer de la bouche ou de la gorge et 400 qui n'en ont pas) seront invités à participer à ce projet. La recherche se fera en deux parties et comportera les mêmes étapes pour les deux groupes.

1. Dans un premier temps, nous recueillerons des informations à partir du dossier médical. Pour les personnes dans le groupe 1 nous voulons, par exemple obtenir des détails scientifiques à propos du cancer. Pour les gens du groupe 2, nous avons besoin de connaître la raison de consultation, à quelle clinique ils sont suivis, etc.
2. La deuxième partie de l'étude sera une entrevue. Nous utiliserons un questionnaire afin de demander aux gens des renseignements plus détaillés sur différentes sphères de leur vie telles que leur travail, leur environnement résidentiel et leur vie familiale. Cette partie de l'entrevue prendra environ 2 heures.

Participer à cette étude, qu'est-ce que cela implique?

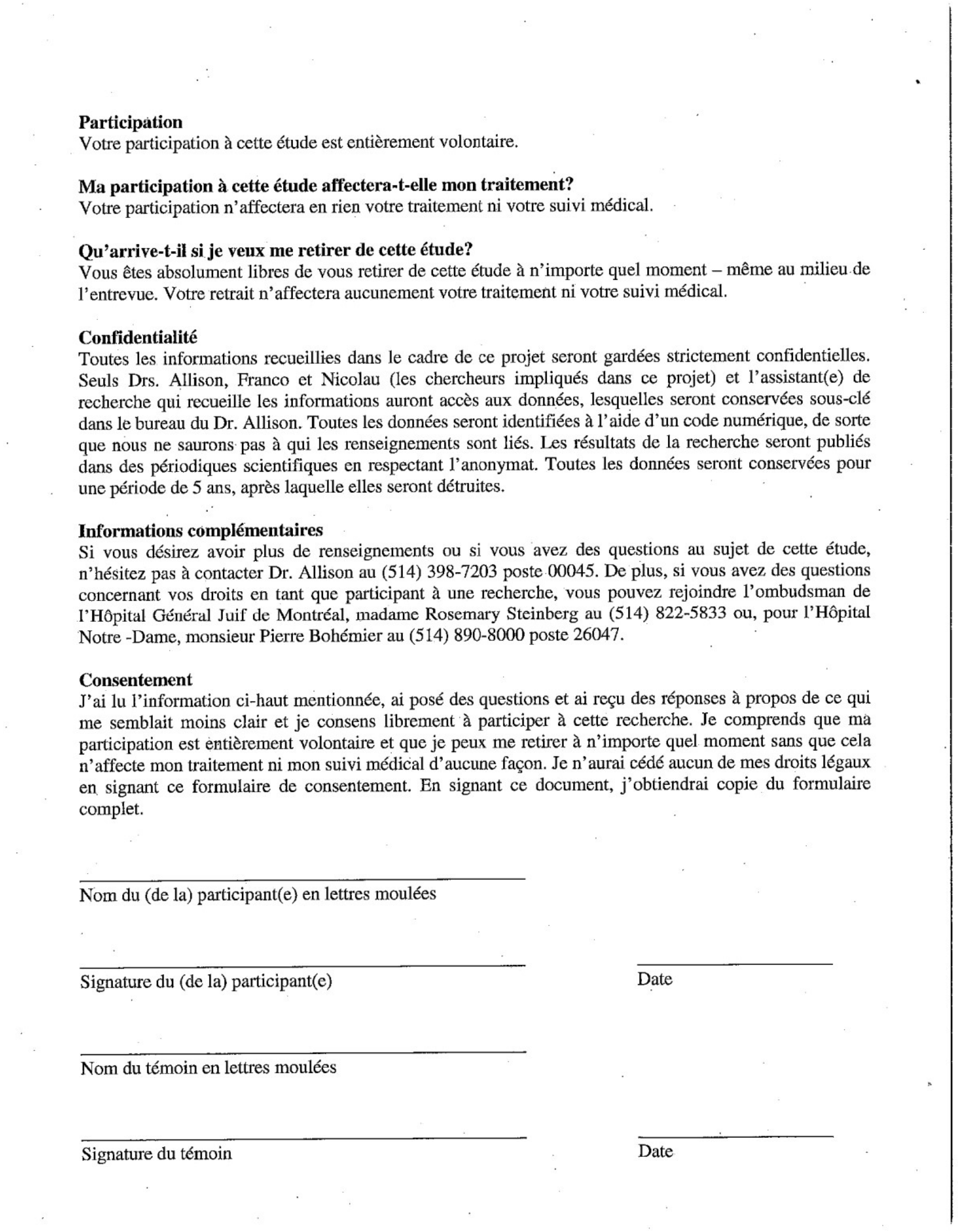
En participant à cette étude, vous nous permettez de consulter votre dossier médical et vous vous engagez à vous présenter à un rendez-vous à l'hôpital pour une entrevue d'environ deux heures.

Risques et inconforts

Il n'y a pas de risque associé à votre participation à cette étude. Elle n'implique aucun traitement ou procédure qui puisse vous causer du mal, des blessures ou de l'inconfort. Elle implique seulement la cueillette de données à partir de votre dossier médical et d'une entrevue.

Bénéfices

Les participants ne bénéficieront pas directement de cette étude. Cependant, les résultats de cette recherche pourraient contribuer à une meilleure compréhension du développement du cancer de la bouche et de la gorge.



Participation
Votre participation à cette étude est entièrement volontaire.

Ma participation à cette étude affectera-t-elle mon traitement?
Votre participation n'affectera en rien votre traitement ni votre suivi médical.

Qu'arrive-t-il si je veux me retirer de cette étude?
Vous êtes absolument libres de vous retirer de cette étude à n'importe quel moment – même au milieu de l'entrevue. Votre retrait n'affectera aucunement votre traitement ni votre suivi médical.

Confidentialité
Toutes les informations recueillies dans le cadre de ce projet seront gardées strictement confidentielles. Seuls Drs. Allison, Franco et Nicolau (les chercheurs impliqués dans ce projet) et l'assistant(e) de recherche qui recueille les informations auront accès aux données, lesquelles seront conservées sous-clé dans le bureau du Dr. Allison. Toutes les données seront identifiées à l'aide d'un code numérique, de sorte que nous ne saurons pas à qui les renseignements sont liés. Les résultats de la recherche seront publiés dans des périodiques scientifiques en respectant l'anonymat. Toutes les données seront conservées pour une période de 5 ans, après laquelle elles seront détruites.

Informations complémentaires
Si vous désirez avoir plus de renseignements ou si vous avez des questions au sujet de cette étude, n'hésitez pas à contacter Dr. Allison au (514) 398-7203 poste 00045. De plus, si vous avez des questions concernant vos droits en tant que participant à une recherche, vous pouvez rejoindre l'ombudsman de l'Hôpital Général Juif de Montréal, madame Rosemary Steinberg au (514) 822-5833 ou, pour l'Hôpital Notre -Dame, monsieur Pierre Bohémier au (514) 890-8000 poste 26047.

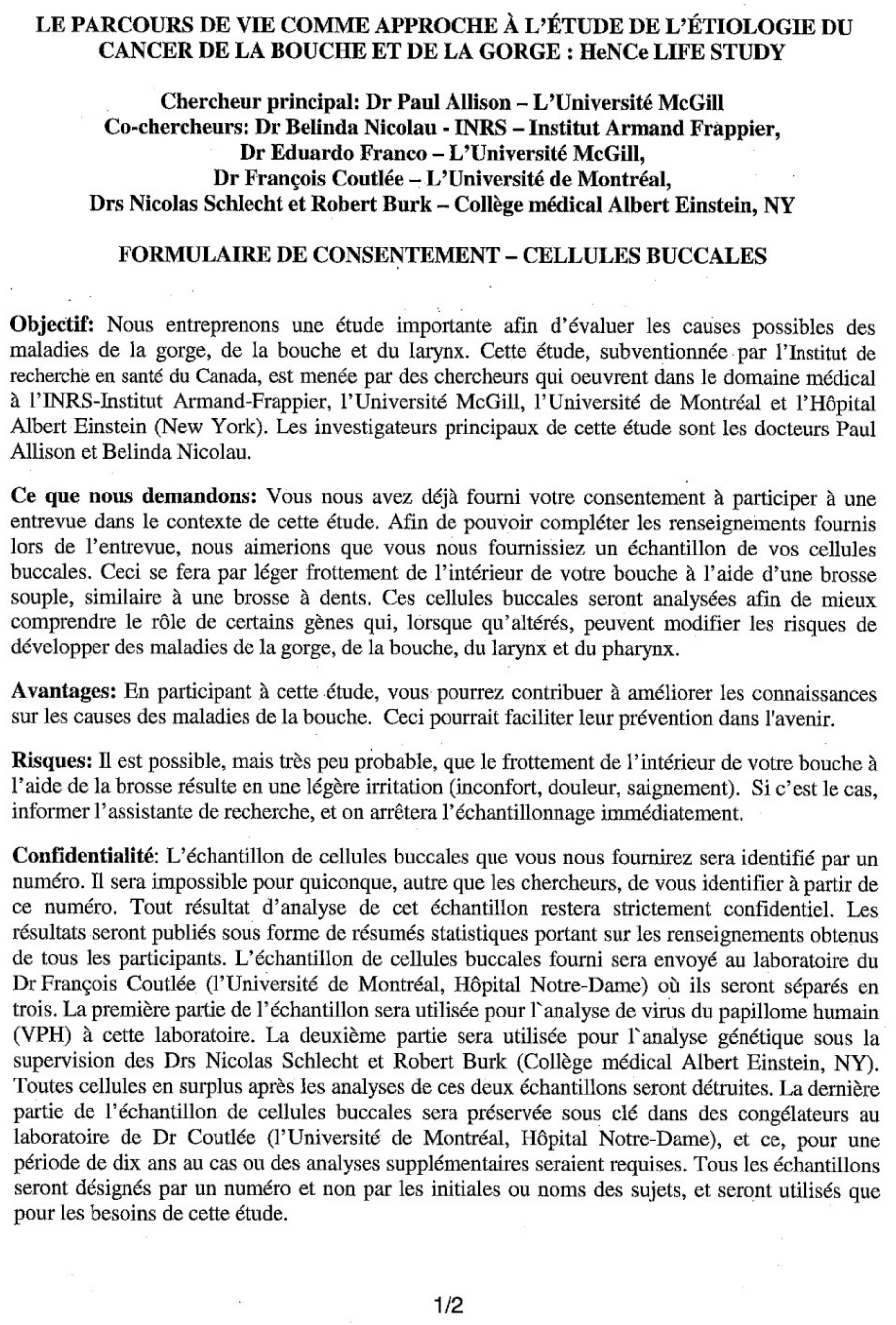
Consentement
J'ai lu l'information ci-haut mentionnée, ai posé des questions et ai reçu des réponses à propos de ce qui me semblait moins clair et je consens librement à participer à cette recherche. Je comprends que ma participation est entièrement volontaire et que je peux me retirer à n'importe quel moment sans que cela n'affecte mon traitement ni mon suivi médical d'aucune façon. Je n'aurai cédé aucun de mes droits légaux en signant ce formulaire de consentement. En signant ce document, j'obtiendrai copie du formulaire complet.

____
Nom du (de la) participant(e) en lettres moulées

____________________________________ ________________
Signature du (de la) participant(e) Date

____
Nom du témoin en lettres moulées

____________________________________ ________________
Signature du témoin Date



LE PARCOURS DE VIE COMME APPROCHE À L'ÉTUDE DE L'ÉTIOLOGIE DU CANCER DE LA BOUCHE ET DE LA GORGE : HeNCe LIFE STUDY

Chercheur principal: Dr Paul Allison – L'Université McGill
Co-chercheurs: Dr Belinda Nicolau - INRS – Institut Armand Frappier,
Dr Eduardo Franco – L'Université McGill,
Dr François Coutlée – L'Université de Montréal,
Drs Nicolas Schlecht et Robert Burk – Collège médical Albert Einstein, NY

FORMULAIRE DE CONSENTEMENT – CELLULES BUCCALES

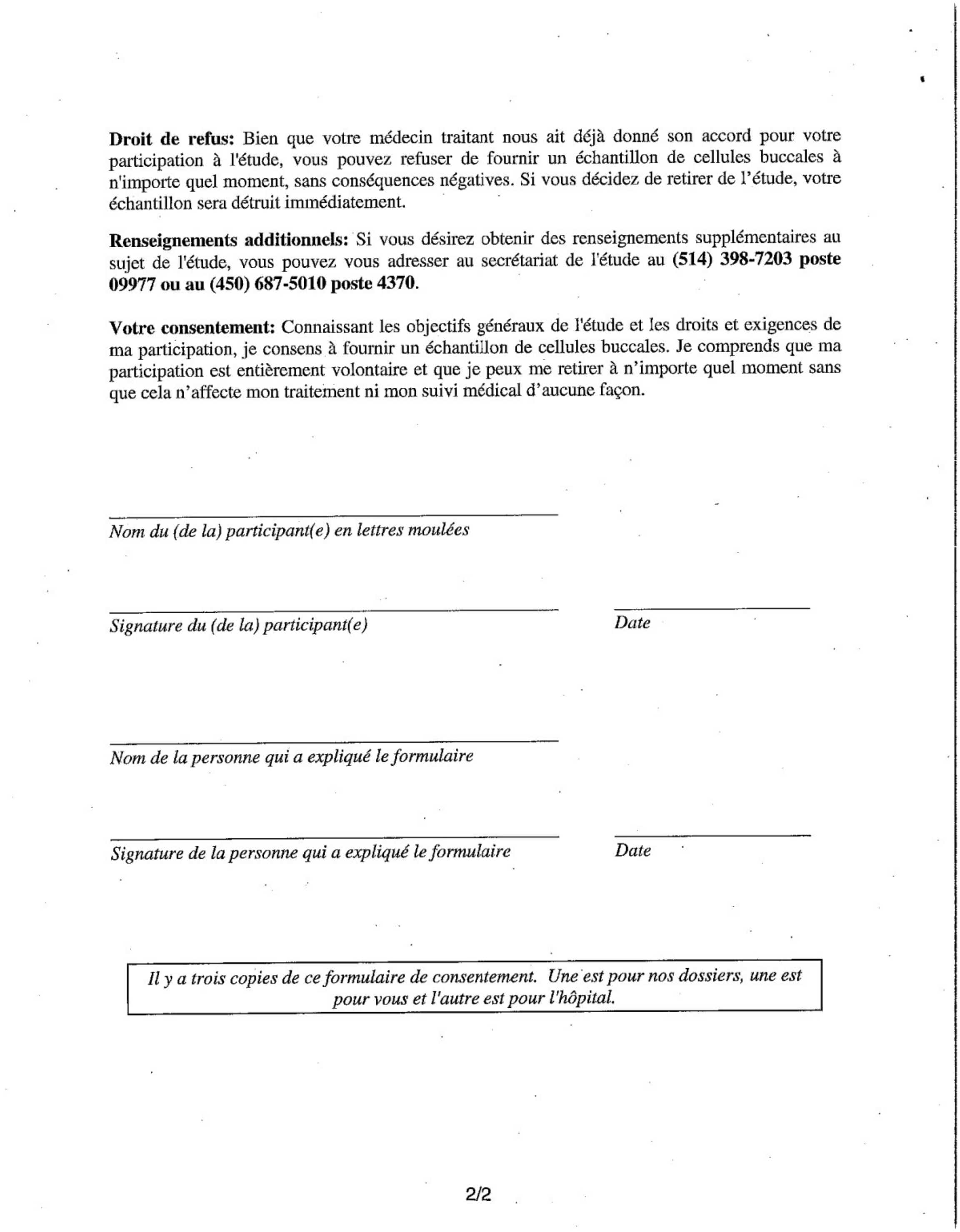
Objectif: Nous entreprenons une étude importante afin d'évaluer les causes possibles des maladies de la gorge, de la bouche et du larynx. Cette étude, subventionnée par l'Institut de recherche en santé du Canada, est menée par des chercheurs qui oeuvrent dans le domaine médical à l'INRS-Institut Armand-Frappier, l'Université McGill, l'Université de Montréal et l'Hôpital Albert Einstein (New York). Les investigateurs principaux de cette étude sont les docteurs Paul Allison et Belinda Nicolau.

Ce que nous demandons: Vous nous avez déjà fourni votre consentement à participer à une entrevue dans le contexte de cette étude. Afin de pouvoir compléter les renseignements fournis lors de l'entrevue, nous aimerions que vous nous fournissiez un échantillon de vos cellules buccales. Ceci se fera par léger frottement de l'intérieur de votre bouche à l'aide d'une brosse souple, similaire à une brosse à dents. Ces cellules buccales seront analysées afin de mieux comprendre le rôle de certains gènes qui, lorsque qu'altérés, peuvent modifier les risques de développer des maladies de la gorge, de la bouche, du larynx et du pharynx.

Avantages: En participant à cette étude, vous pourrez contribuer à améliorer les connaissances sur les causes des maladies de la bouche. Ceci pourrait faciliter leur prévention dans l'avenir.

Risques: Il est possible, mais très peu probable, que le frottement de l'intérieur de votre bouche à l'aide de la brosse résulte en une légère irritation (inconfort, douleur, saignement). Si c'est le cas, informer l'assistante de recherche, et on arrêtera l'échantillonnage immédiatement.

Confidentialité: L'échantillon de cellules buccales que vous nous fournirez sera identifié par un numéro. Il sera impossible pour quiconque, autre que les chercheurs, de vous identifier à partir de ce numéro. Tout résultat d'analyse de cet échantillon restera strictement confidentiel. Les résultats seront publiés sous forme de résumés statistiques portant sur les renseignements obtenus de tous les participants. L'échantillon de cellules buccales fourni sera envoyé au laboratoire du Dr François Coutlée (l'Université de Montréal, Hôpital Notre-Dame) où ils seront séparés en trois. La première partie de l'échantillon sera utilisée pour l`analyse de virus du papillome humain (VPH) à cette laboratoire. La deuxième partie sera utilisée pour l`analyse génétique sous la supervision des Drs Nicolas Schlecht et Robert Burk (Collège médical Albert Einstein, NY). Toutes cellules en surplus après les analyses de ces deux échantillons seront détruites. La dernière partie de l'échantillon de cellules buccales sera préservée sous clé dans des congélateurs au laboratoire de Dr Coutlée (l'Université de Montréal, Hôpital Notre-Dame), et ce, pour une période de dix ans au cas ou des analyses supplémentaires seraient requises. Tous les échantillons seront désignés par un numéro et non par les initiales ou noms des sujets, et seront utilisés que pour les besoins de cette étude.



Droit de refus: Bien que votre médecin traitant nous ait déjà donné son accord pour votre participation à l'étude, vous pouvez refuser de fournir un échantillon de cellules buccales à n'importe quel moment, sans conséquences négatives. Si vous décidez de retirer de l'étude, votre échantillon sera détruit immédiatement.

Renseignements additionnels: Si vous désirez obtenir des renseignements supplémentaires au sujet de l'étude, vous pouvez vous adresser au secrétariat de l'étude au **(514) 398-7203 poste 09977 ou au (450) 687-5010 poste 4370**.

Votre consentement: Connaissant les objectifs généraux de l'étude et les droits et exigences de ma participation, je consens à fournir un échantillon de cellules buccales. Je comprends que ma participation est entièrement volontaire et que je peux me retirer à n'importe quel moment sans que cela n'affecte mon traitement ni mon suivi médical d'aucune façon.

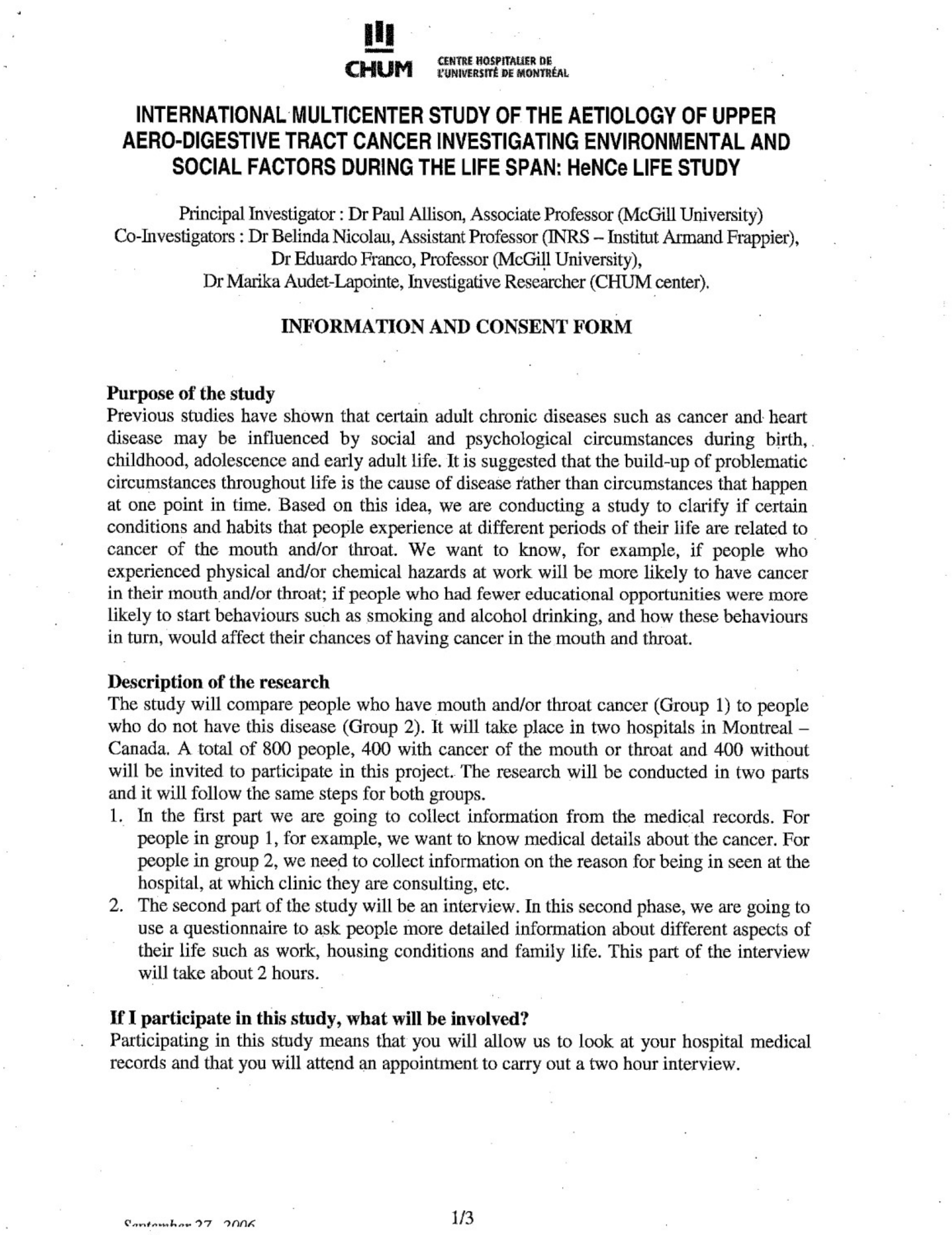
Nom du (de la) participant(e) en lettres moulées

Signature du (de la) participant(e) *Date*

Nom de la personne qui a expliqué le formulaire

Signature de la personne qui a expliqué le formulaire *Date*

Il y a trois copies de ce formulaire de consentement. Une est pour nos dossiers, une est pour vous et l'autre est pour l'hôpital.



CHUM CENTRE HOSPITALIER DE L'UNIVERSITÉ DE MONTRÉAL

INTERNATIONAL MULTICENTER STUDY OF THE AETIOLOGY OF UPPER AERO-DIGESTIVE TRACT CANCER INVESTIGATING ENVIRONMENTAL AND SOCIAL FACTORS DURING THE LIFE SPAN: HeNCe LIFE STUDY

Principal Investigator : Dr Paul Allison, Associate Professor (McGill University)
Co-Investigators : Dr Belinda Nicolau, Assistant Professor (INRS – Institut Armand Frappier),
Dr Eduardo Franco, Professor (McGill University),
Dr Marika Audet-Lapointe, Investigative Researcher (CHUM center).

INFORMATION AND CONSENT FORM

Purpose of the study

Previous studies have shown that certain adult chronic diseases such as cancer and heart disease may be influenced by social and psychological circumstances during birth, childhood, adolescence and early adult life. It is suggested that the build-up of problematic circumstances throughout life is the cause of disease rather than circumstances that happen at one point in time. Based on this idea, we are conducting a study to clarify if certain conditions and habits that people experience at different periods of their life are related to cancer of the mouth and/or throat. We want to know, for example, if people who experienced physical and/or chemical hazards at work will be more likely to have cancer in their mouth and/or throat; if people who had fewer educational opportunities were more likely to start behaviours such as smoking and alcohol drinking, and how these behaviours in turn, would affect their chances of having cancer in the mouth and throat.

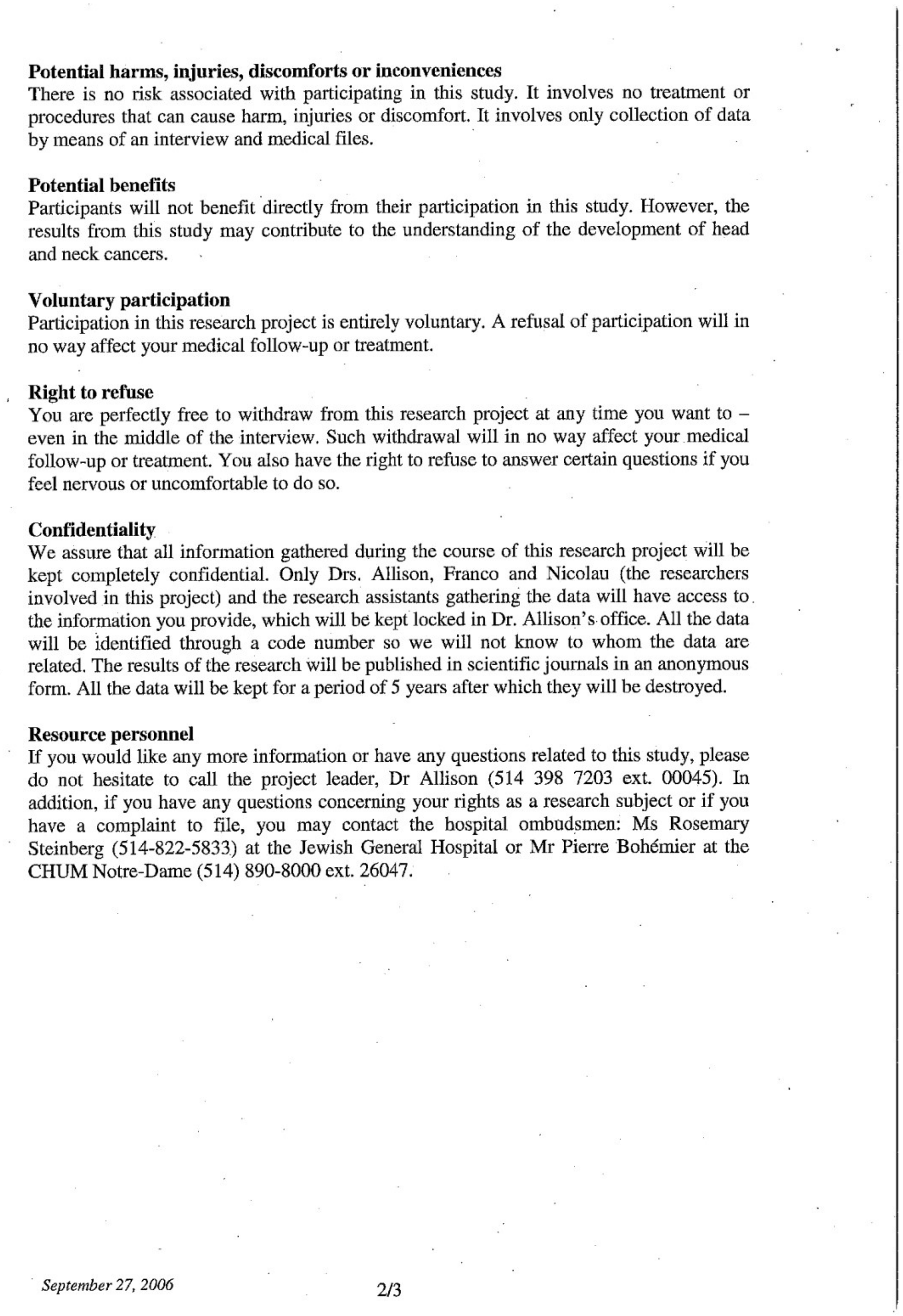
Description of the research

The study will compare people who have mouth and/or throat cancer (Group 1) to people who do not have this disease (Group 2). It will take place in two hospitals in Montreal – Canada. A total of 800 people, 400 with cancer of the mouth or throat and 400 without will be invited to participate in this project. The research will be conducted in two parts and it will follow the same steps for both groups.

1. In the first part we are going to collect information from the medical records. For people in group 1, for example, we want to know medical details about the cancer. For people in group 2, we need to collect information on the reason for being in seen at the hospital, at which clinic they are consulting, etc.
2. The second part of the study will be an interview. In this second phase, we are going to use a questionnaire to ask people more detailed information about different aspects of their life such as work, housing conditions and family life. This part of the interview will take about 2 hours.

If I participate in this study, what will be involved?

Participating in this study means that you will allow us to look at your hospital medical records and that you will attend an appointment to carry out a two hour interview.



Potential harms, injuries, discomforts or inconveniences
There is no risk associated with participating in this study. It involves no treatment or procedures that can cause harm, injuries or discomfort. It involves only collection of data by means of an interview and medical files.

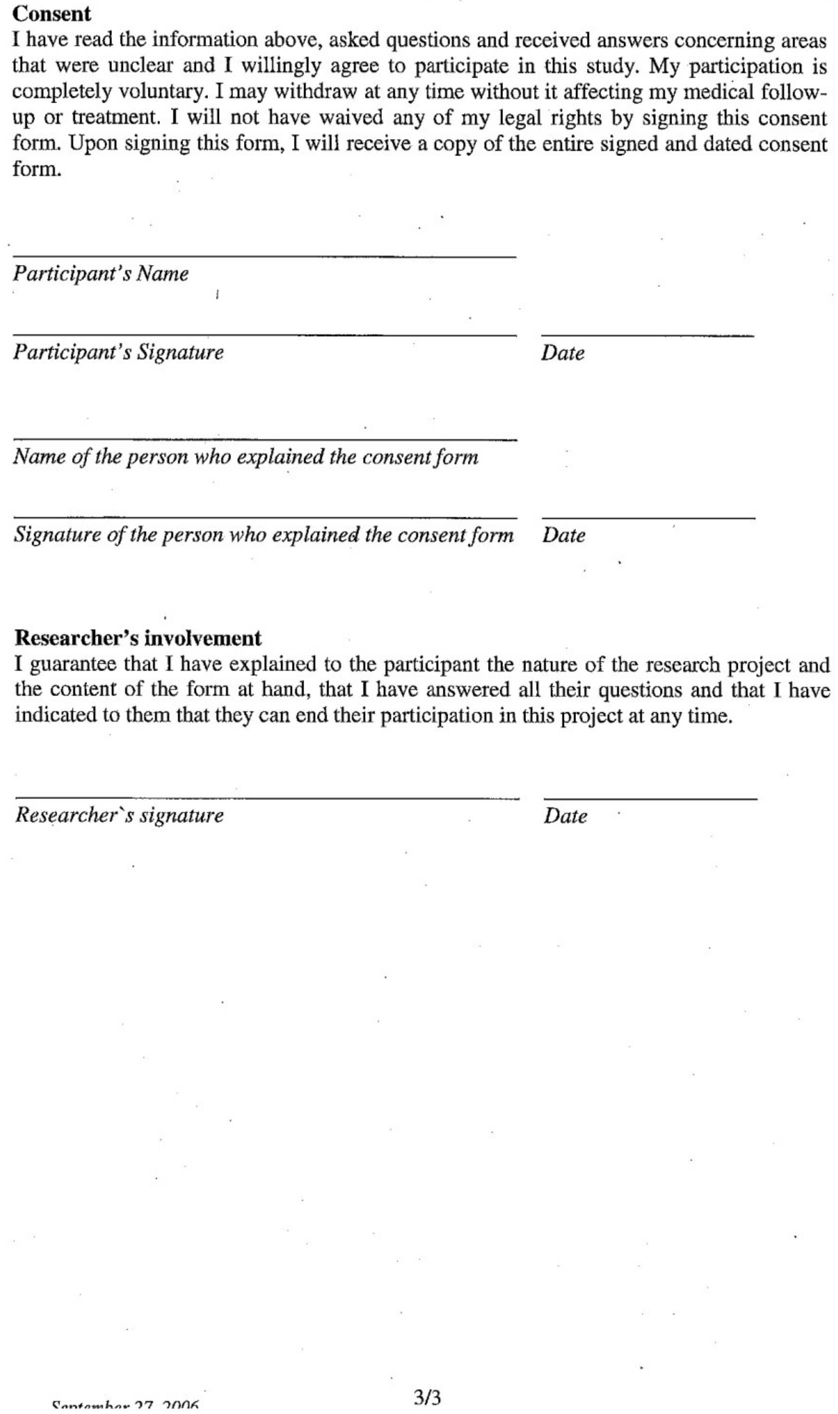
Potential benefits
Participants will not benefit directly from their participation in this study. However, the results from this study may contribute to the understanding of the development of head and neck cancers.

Voluntary participation
Participation in this research project is entirely voluntary. A refusal of participation will in no way affect your medical follow-up or treatment.

Right to refuse
You are perfectly free to withdraw from this research project at any time you want to – even in the middle of the interview. Such withdrawal will in no way affect your medical follow-up or treatment. You also have the right to refuse to answer certain questions if you feel nervous or uncomfortable to do so.

Confidentiality
We assure that all information gathered during the course of this research project will be kept completely confidential. Only Drs. Allison, Franco and Nicolau (the researchers involved in this project) and the research assistants gathering the data will have access to the information you provide, which will be kept locked in Dr. Allison's office. All the data will be identified through a code number so we will not know to whom the data are related. The results of the research will be published in scientific journals in an anonymous form. All the data will be kept for a period of 5 years after which they will be destroyed.

Resource personnel
If you would like any more information or have any questions related to this study, please do not hesitate to call the project leader, Dr Allison (514 398 7203 ext. 00045). In addition, if you have any questions concerning your rights as a research subject or if you have a complaint to file, you may contact the hospital ombudsmen: Ms Rosemary Steinberg (514-822-5833) at the Jewish General Hospital or Mr Pierre Bohémier at the CHUM Notre-Dame (514) 890-8000 ext. 26047.



Consent

I have read the information above, asked questions and received answers concerning areas that were unclear and I willingly agree to participate in this study. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment. I will not have waived any of my legal rights by signing this consent form. Upon signing this form, I will receive a copy of the entire signed and dated consent form.

Participant's Name

Participant's Signature *Date*

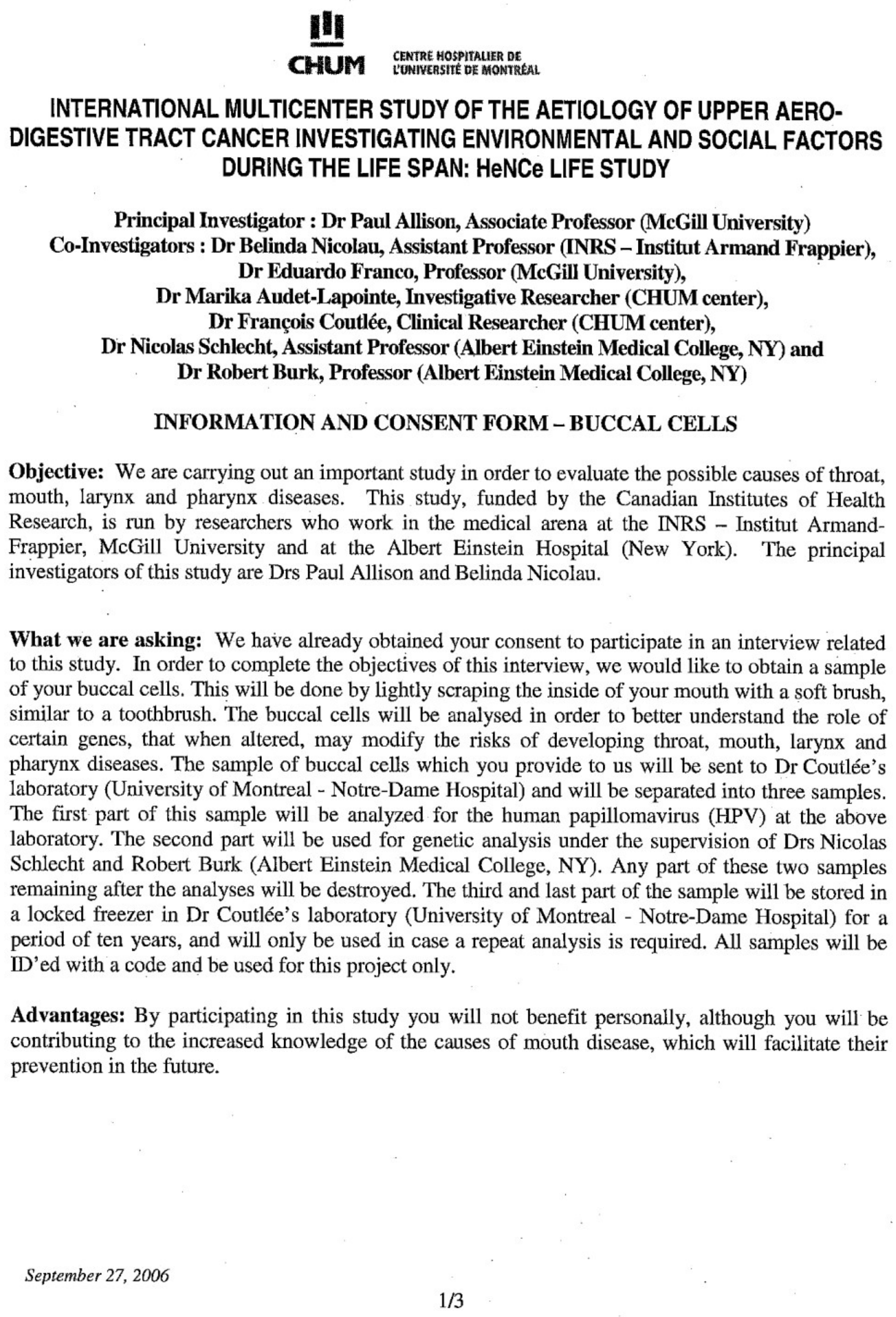
Name of the person who explained the consent form

Signature of the person who explained the consent form *Date*

Researcher's involvement

I guarantee that I have explained to the participant the nature of the research project and the content of the form at hand, that I have answered all their questions and that I have indicated to them that they can end their participation in this project at any time.

Researcher`s signature *Date*



CHUM CENTRE HOSPITALIER DE L'UNIVERSITÉ DE MONTRÉAL

INTERNATIONAL MULTICENTER STUDY OF THE AETIOLOGY OF UPPER AERO-DIGESTIVE TRACT CANCER INVESTIGATING ENVIRONMENTAL AND SOCIAL FACTORS DURING THE LIFE SPAN: HeNCe LIFE STUDY

Principal Investigator : Dr Paul Allison, Associate Professor (McGill University)
Co-Investigators : Dr Belinda Nicolau, Assistant Professor (INRS – Institut Armand Frappier),
Dr Eduardo Franco, Professor (McGill University),
Dr Marika Audet-Lapointe, Investigative Researcher (CHUM center),
Dr François Coutlée, Clinical Researcher (CHUM center),
Dr Nicolas Schlecht, Assistant Professor (Albert Einstein Medical College, NY) and
Dr Robert Burk, Professor (Albert Einstein Medical College, NY)

INFORMATION AND CONSENT FORM – BUCCAL CELLS

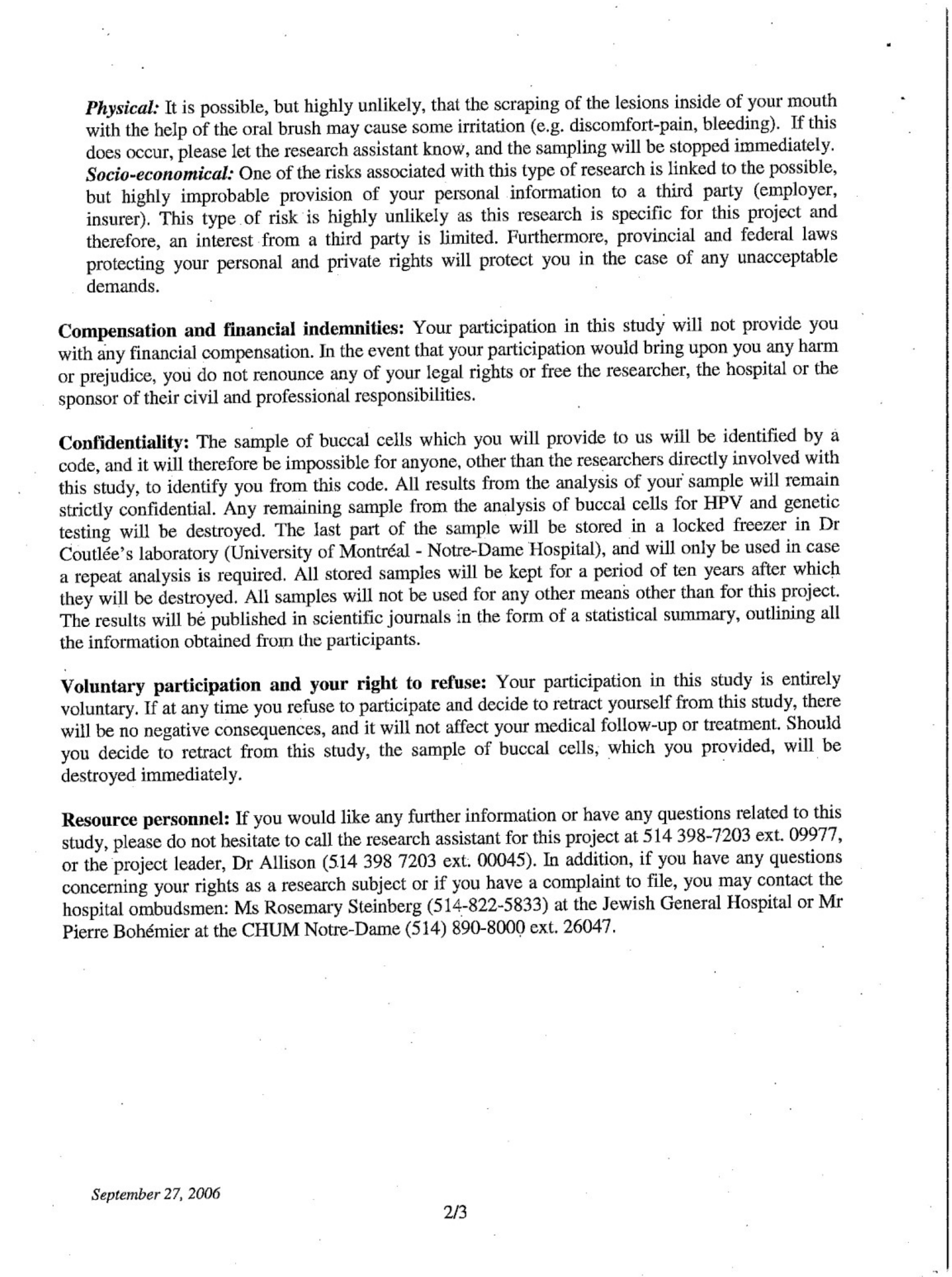
Objective: We are carrying out an important study in order to evaluate the possible causes of throat, mouth, larynx and pharynx diseases. This study, funded by the Canadian Institutes of Health Research, is run by researchers who work in the medical arena at the INRS – Institut Armand-Frappier, McGill University and at the Albert Einstein Hospital (New York). The principal investigators of this study are Drs Paul Allison and Belinda Nicolau.

What we are asking: We have already obtained your consent to participate in an interview related to this study. In order to complete the objectives of this interview, we would like to obtain a sample of your buccal cells. This will be done by lightly scraping the inside of your mouth with a soft brush, similar to a toothbrush. The buccal cells will be analysed in order to better understand the role of certain genes, that when altered, may modify the risks of developing throat, mouth, larynx and pharynx diseases. The sample of buccal cells which you provide to us will be sent to Dr Coutlée's laboratory (University of Montreal - Notre-Dame Hospital) and will be separated into three samples. The first part of this sample will be analyzed for the human papillomavirus (HPV) at the above laboratory. The second part will be used for genetic analysis under the supervision of Drs Nicolas Schlecht and Robert Burk (Albert Einstein Medical College, NY). Any part of these two samples remaining after the analyses will be destroyed. The third and last part of the sample will be stored in a locked freezer in Dr Coutlée's laboratory (University of Montreal - Notre-Dame Hospital) for a period of ten years, and will only be used in case a repeat analysis is required. All samples will be ID'ed with a code and be used for this project only.

Advantages: By participating in this study you will not benefit personally, although you will be contributing to the increased knowledge of the causes of mouth disease, which will facilitate their prevention in the future.

September 27, 2006

1/3



Physical: It is possible, but highly unlikely, that the scraping of the lesions inside of your mouth with the help of the oral brush may cause some irritation (e.g. discomfort-pain, bleeding). If this does occur, please let the research assistant know, and the sampling will be stopped immediately.
Socio-economical: One of the risks associated with this type of research is linked to the possible, but highly improbable provision of your personal information to a third party (employer, insurer). This type of risk is highly unlikely as this research is specific for this project and therefore, an interest from a third party is limited. Furthermore, provincial and federal laws protecting your personal and private rights will protect you in the case of any unacceptable demands.

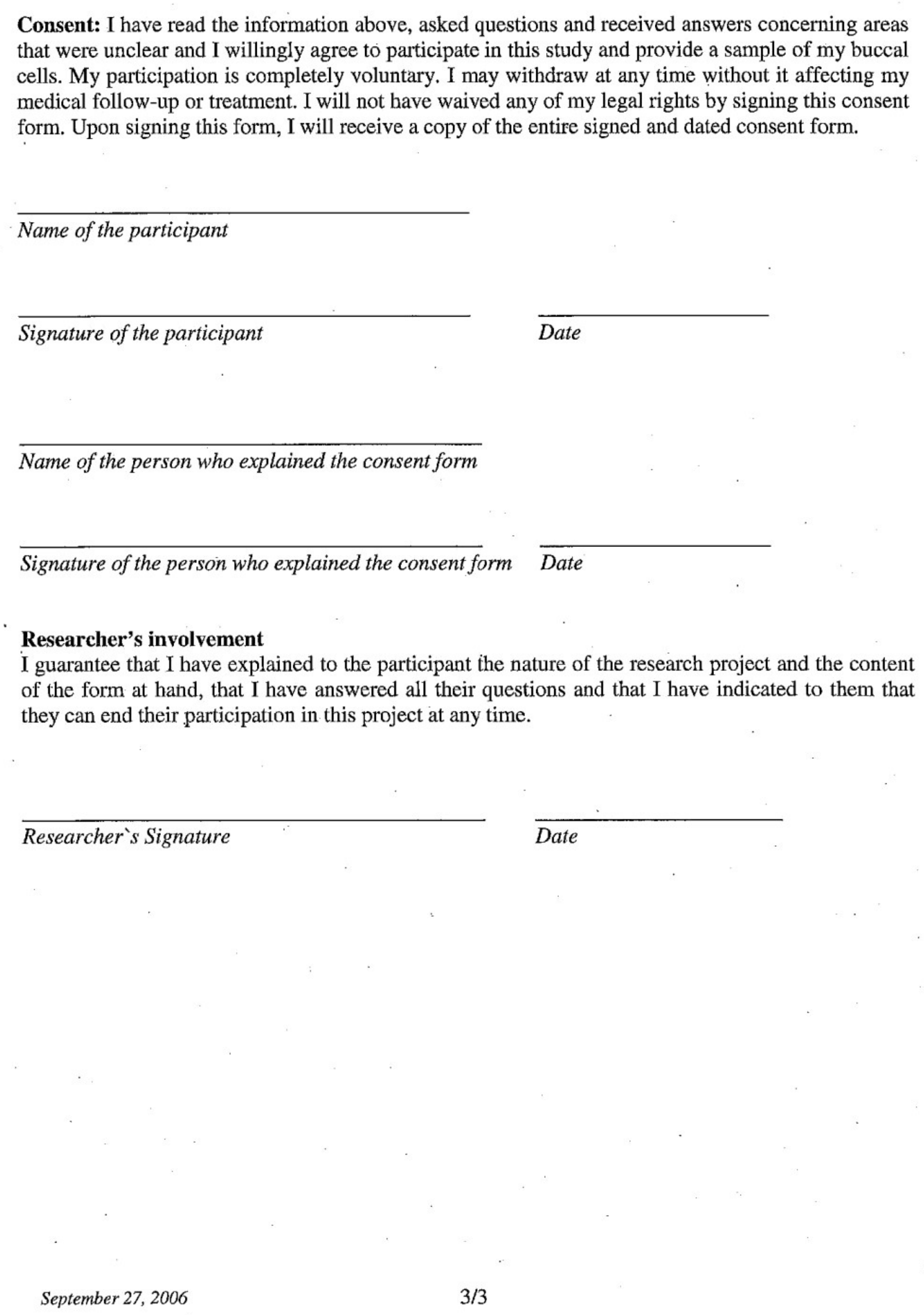
Compensation and financial indemnities: Your participation in this study will not provide you with any financial compensation. In the event that your participation would bring upon you any harm or prejudice, you do not renounce any of your legal rights or free the researcher, the hospital or the sponsor of their civil and professional responsibilities.

Confidentiality: The sample of buccal cells which you will provide to us will be identified by a code, and it will therefore be impossible for anyone, other than the researchers directly involved with this study, to identify you from this code. All results from the analysis of your sample will remain strictly confidential. Any remaining sample from the analysis of buccal cells for HPV and genetic testing will be destroyed. The last part of the sample will be stored in a locked freezer in Dr Coutlée's laboratory (University of Montréal - Notre-Dame Hospital), and will only be used in case a repeat analysis is required. All stored samples will be kept for a period of ten years after which they will be destroyed. All samples will not be used for any other means other than for this project. The results will be published in scientific journals in the form of a statistical summary, outlining all the information obtained from the participants.

Voluntary participation and your right to refuse: Your participation in this study is entirely voluntary. If at any time you refuse to participate and decide to retract yourself from this study, there will be no negative consequences, and it will not affect your medical follow-up or treatment. Should you decide to retract from this study, the sample of buccal cells, which you provided, will be destroyed immediately.

Resource personnel: If you would like any further information or have any questions related to this study, please do not hesitate to call the research assistant for this project at 514 398-7203 ext. 09977, or the project leader, Dr Allison (514 398 7203 ext. 00045). In addition, if you have any questions concerning your rights as a research subject or if you have a complaint to file, you may contact the hospital ombudsmen: Ms Rosemary Steinberg (514-822-5833) at the Jewish General Hospital or Mr Pierre Bohémier at the CHUM Notre-Dame (514) 890-8000 ext. 26047.

September 27, 2006



Consent: I have read the information above, asked questions and received answers concerning areas that were unclear and I willingly agree to participate in this study and provide a sample of my buccal cells. My participation is completely voluntary. I may withdraw at any time without it affecting my medical follow-up or treatment. I will not have waived any of my legal rights by signing this consent form. Upon signing this form, I will receive a copy of the entire signed and dated consent form.

Name of the participant

Signature of the participant *Date*

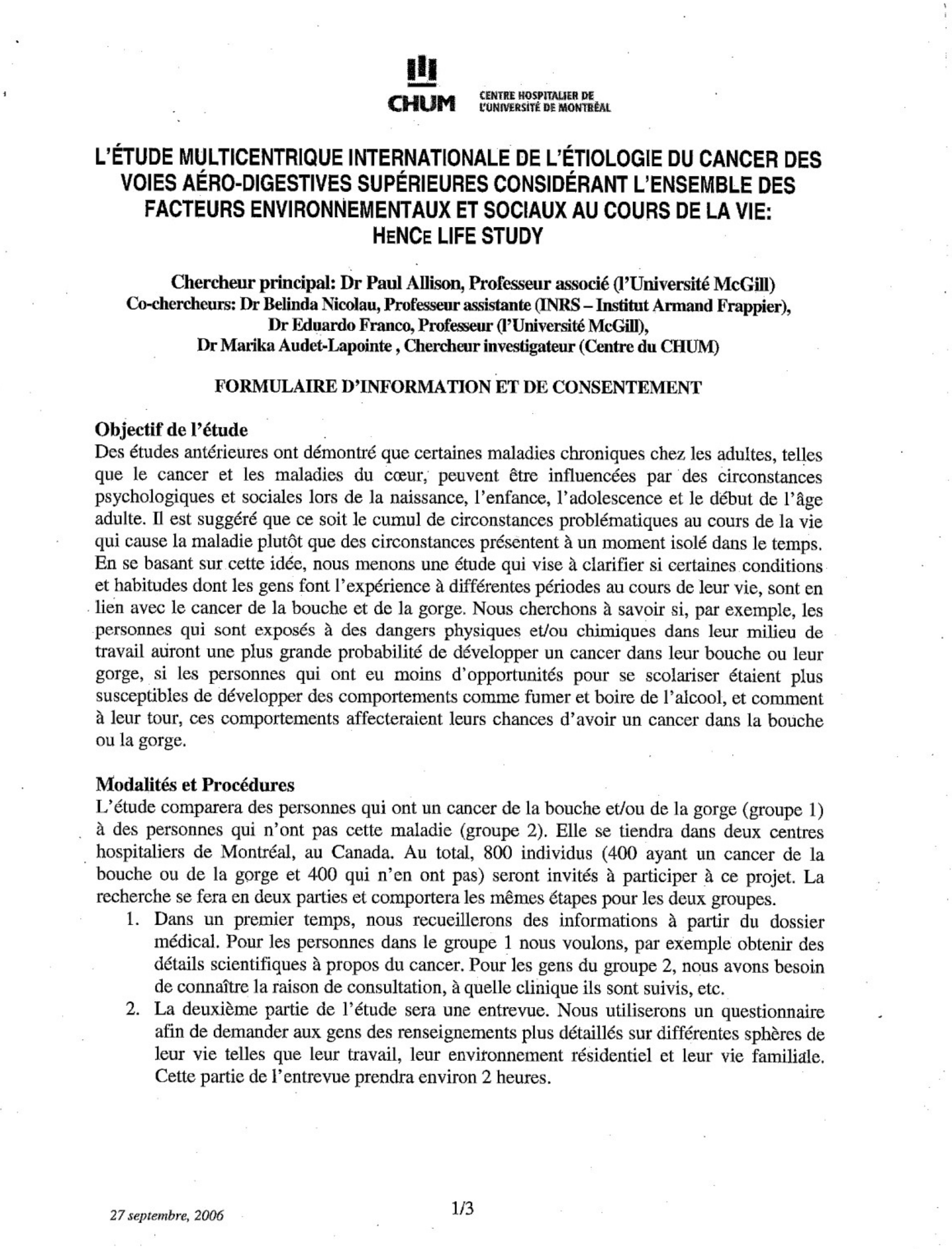
Name of the person who explained the consent form

Signature of the person who explained the consent form *Date*

Researcher's involvement

I guarantee that I have explained to the participant the nature of the research project and the content of the form at hand, that I have answered all their questions and that I have indicated to them that they can end their participation in this project at any time.

Researcher`s Signature *Date*



CHUM CENTRE HOSPITALIER DE L'UNIVERSITÉ DE MONTRÉAL

L'ÉTUDE MULTICENTRIQUE INTERNATIONALE DE L'ÉTIOLOGIE DU CANCER DES VOIES AÉRO-DIGESTIVES SUPÉRIEURES CONSIDÉRANT L'ENSEMBLE DES FACTEURS ENVIRONNEMENTAUX ET SOCIAUX AU COURS DE LA VIE: HENCE LIFE STUDY

Chercheur principal: Dr Paul Allison, Professeur associé (l'Université McGill)
Co-chercheurs: Dr Belinda Nicolau, Professeur assistante (INRS – Institut Armand Frappier),
Dr Eduardo Franco, Professeur (l'Université McGill),
Dr Marika Audet-Lapointe , Chercheur investigateur (Centre du CHUM)

FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

Objectif de l'étude

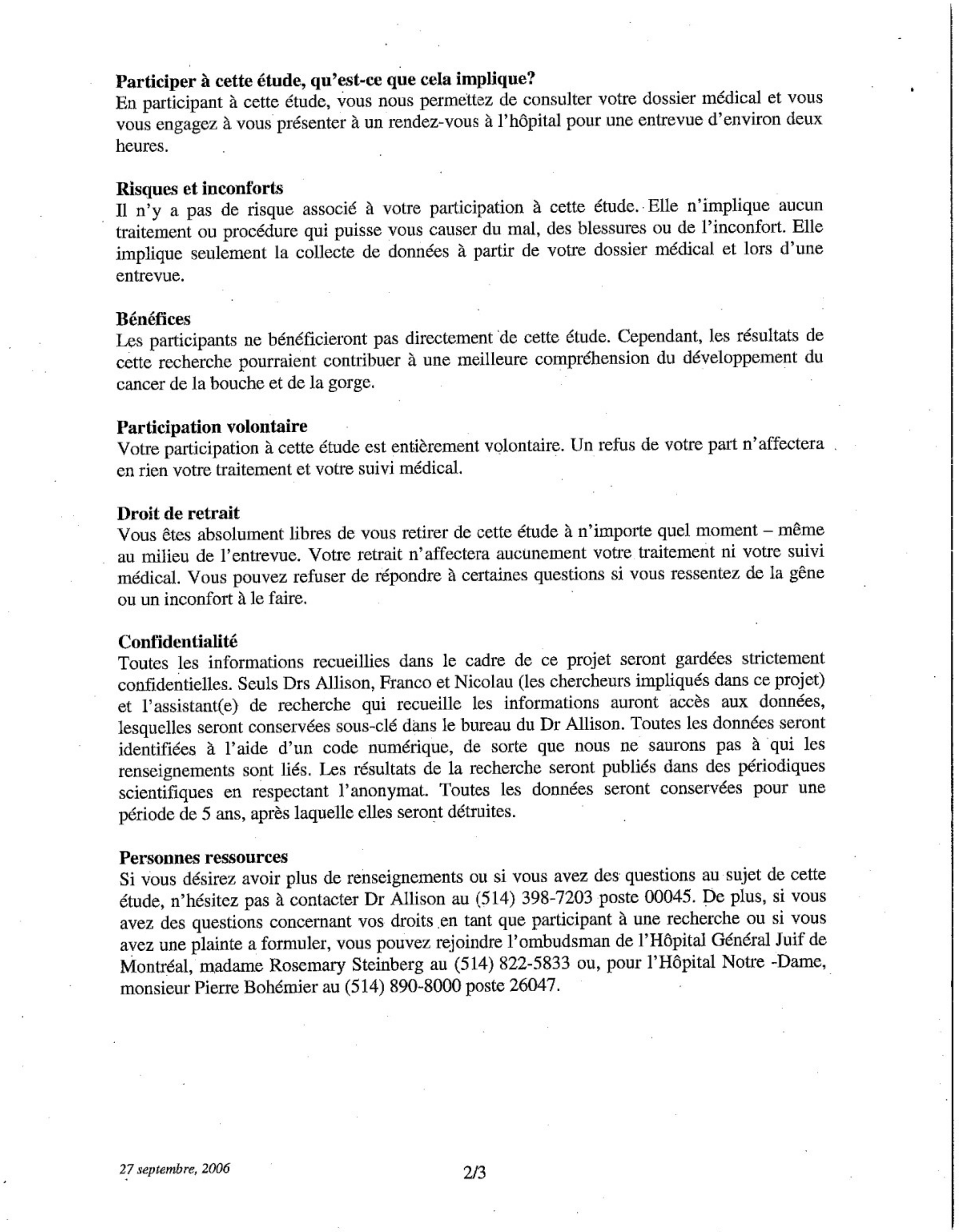
Des études antérieures ont démontré que certaines maladies chroniques chez les adultes, telles que le cancer et les maladies du cœur, peuvent être influencées par des circonstances psychologiques et sociales lors de la naissance, l'enfance, l'adolescence et le début de l'âge adulte. Il est suggéré que ce soit le cumul de circonstances problématiques au cours de la vie qui cause la maladie plutôt que des circonstances présentent à un moment isolé dans le temps. En se basant sur cette idée, nous menons une étude qui vise à clarifier si certaines conditions et habitudes dont les gens font l'expérience à différentes périodes au cours de leur vie, sont en lien avec le cancer de la bouche et de la gorge. Nous cherchons à savoir si, par exemple, les personnes qui sont exposés à des dangers physiques et/ou chimiques dans leur milieu de travail auront une plus grande probabilité de développer un cancer dans leur bouche ou leur gorge, si les personnes qui ont eu moins d'opportunités pour se scolariser étaient plus susceptibles de développer des comportements comme fumer et boire de l'alcool, et comment à leur tour, ces comportements affecteraient leurs chances d'avoir un cancer dans la bouche ou la gorge.

Modalités et Procédures

L'étude comparera des personnes qui ont un cancer de la bouche et/ou de la gorge (groupe 1) à des personnes qui n'ont pas cette maladie (groupe 2). Elle se tiendra dans deux centres hospitaliers de Montréal, au Canada. Au total, 800 individus (400 ayant un cancer de la bouche ou de la gorge et 400 qui n'en ont pas) seront invités à participer à ce projet. La recherche se fera en deux parties et comportera les mêmes étapes pour les deux groupes.

1. Dans un premier temps, nous recueillerons des informations à partir du dossier médical. Pour les personnes dans le groupe 1 nous voulons, par exemple obtenir des détails scientifiques à propos du cancer. Pour les gens du groupe 2, nous avons besoin de connaître la raison de consultation, à quelle clinique ils sont suivis, etc.
2. La deuxième partie de l'étude sera une entrevue. Nous utiliserons un questionnaire afin de demander aux gens des renseignements plus détaillés sur différentes sphères de leur vie telles que leur travail, leur environnement résidentiel et leur vie familiale. Cette partie de l'entrevue prendra environ 2 heures.

27 septembre, 2006 1/3



Participer à cette étude, qu'est-ce que cela implique?
En participant à cette étude, vous nous permettez de consulter votre dossier médical et vous vous engagez à vous présenter à un rendez-vous à l'hôpital pour une entrevue d'environ deux heures.

Risques et inconforts
Il n'y a pas de risque associé à votre participation à cette étude. Elle n'implique aucun traitement ou procédure qui puisse vous causer du mal, des blessures ou de l'inconfort. Elle implique seulement la collecte de données à partir de votre dossier médical et lors d'une entrevue.

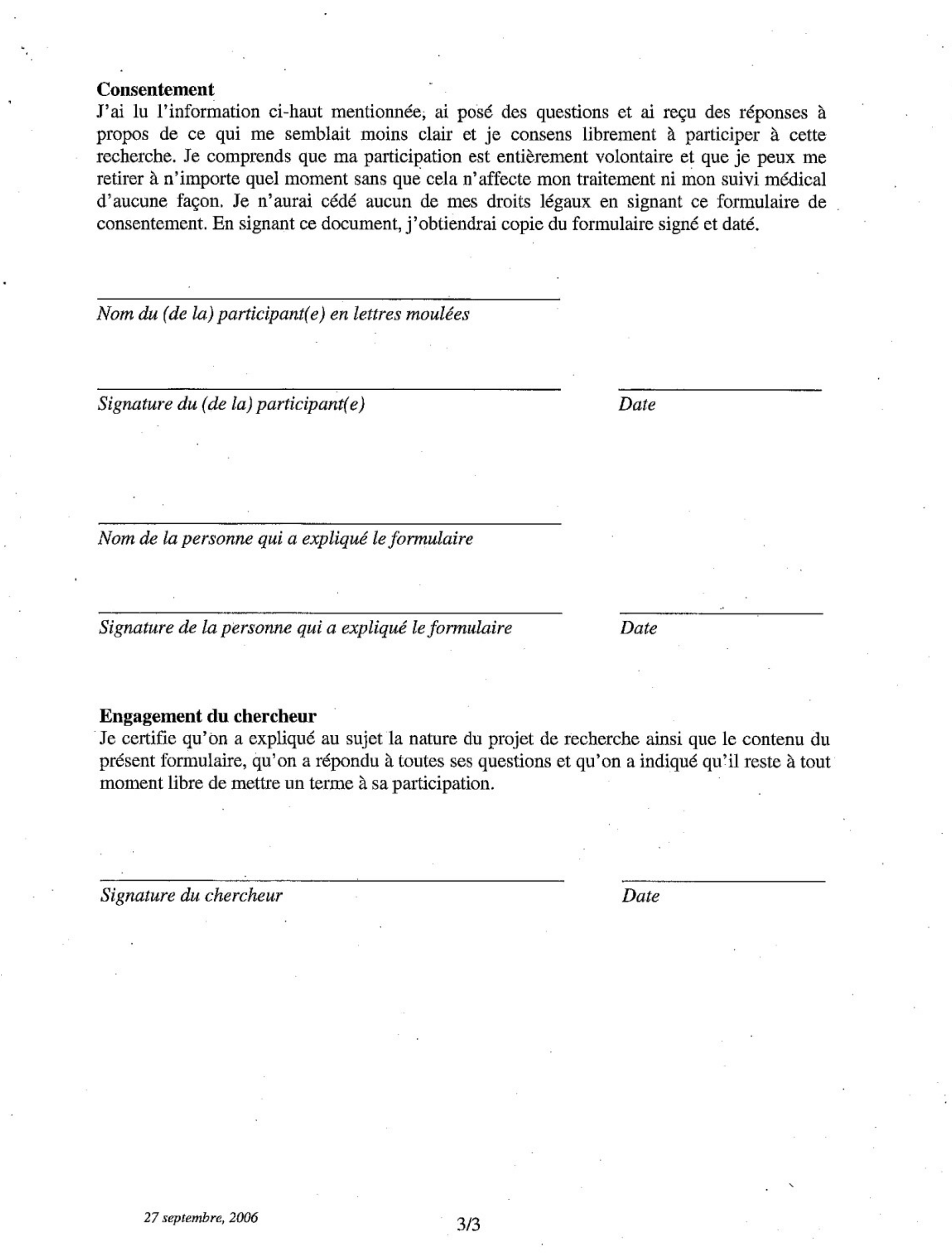
Bénéfices
Les participants ne bénéficieront pas directement de cette étude. Cependant, les résultats de cette recherche pourraient contribuer à une meilleure compréhension du développement du cancer de la bouche et de la gorge.

Participation volontaire
Votre participation à cette étude est entièrement volontaire. Un refus de votre part n'affectera en rien votre traitement et votre suivi médical.

Droit de retrait
Vous êtes absolument libres de vous retirer de cette étude à n'importe quel moment – même au milieu de l'entrevue. Votre retrait n'affectera aucunement votre traitement ni votre suivi médical. Vous pouvez refuser de répondre à certaines questions si vous ressentez de la gêne ou un inconfort à le faire.

Confidentialité
Toutes les informations recueillies dans le cadre de ce projet seront gardées strictement confidentielles. Seuls Drs Allison, Franco et Nicolau (les chercheurs impliqués dans ce projet) et l'assistant(e) de recherche qui recueille les informations auront accès aux données, lesquelles seront conservées sous-clé dans le bureau du Dr Allison. Toutes les données seront identifiées à l'aide d'un code numérique, de sorte que nous ne saurons pas à qui les renseignements sont liés. Les résultats de la recherche seront publiés dans des périodiques scientifiques en respectant l'anonymat. Toutes les données seront conservées pour une période de 5 ans, après laquelle elles seront détruites.

Personnes ressources
Si vous désirez avoir plus de renseignements ou si vous avez des questions au sujet de cette étude, n'hésitez pas à contacter Dr Allison au (514) 398-7203 poste 00045. De plus, si vous avez des questions concernant vos droits en tant que participant à une recherche ou si vous avez une plainte a formuler, vous pouvez rejoindre l'ombudsman de l'Hôpital Général Juif de Montréal, madame Rosemary Steinberg au (514) 822-5833 ou, pour l'Hôpital Notre -Dame, monsieur Pierre Bohémier au (514) 890-8000 poste 26047.



Consentement

J'ai lu l'information ci-haut mentionnée, ai posé des questions et ai reçu des réponses à propos de ce qui me semblait moins clair et je consens librement à participer à cette recherche. Je comprends que ma participation est entièrement volontaire et que je peux me retirer à n'importe quel moment sans que cela n'affecte mon traitement ni mon suivi médical d'aucune façon. Je n'aurai cédé aucun de mes droits légaux en signant ce formulaire de consentement. En signant ce document, j'obtiendrai copie du formulaire signé et daté.

____

Nom du (de la) participant(e) en lettres moulées

______________________________ ______________

Signature du (de la) participant(e) *Date*

____

Nom de la personne qui a expliqué le formulaire

______________________________ ______________

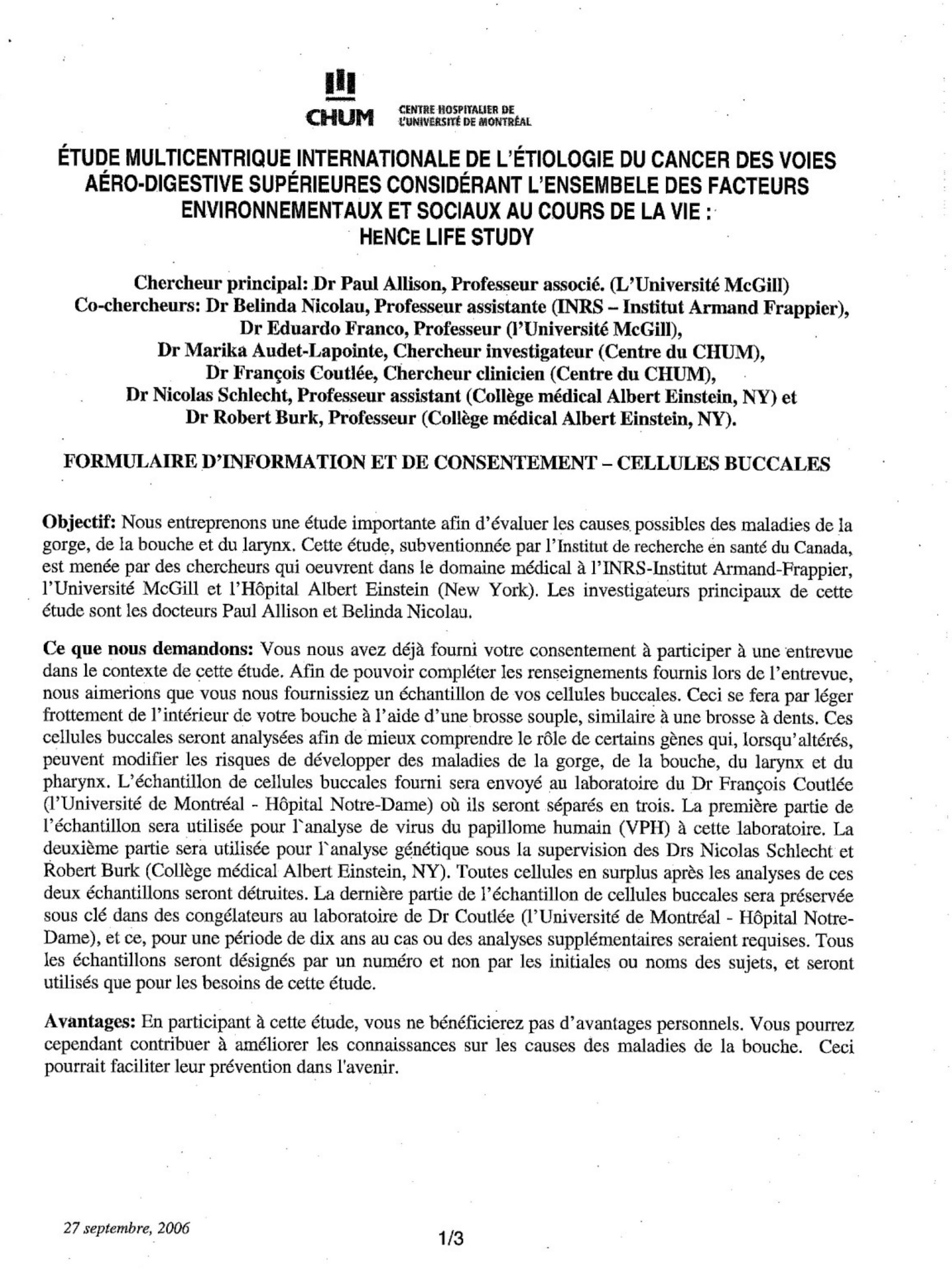
Signature de la personne qui a expliqué le formulaire *Date*

Engagement du chercheur

Je certifie qu'on a expliqué au sujet la nature du projet de recherche ainsi que le contenu du présent formulaire, qu'on a répondu à toutes ses questions et qu'on a indiqué qu'il reste à tout moment libre de mettre un terme à sa participation.

______________________________ ______________

Signature du chercheur *Date*



CHUM CENTRE HOSPITALIER DE L'UNIVERSITÉ DE MONTRÉAL

ÉTUDE MULTICENTRIQUE INTERNATIONALE DE L'ÉTIOLOGIE DU CANCER DES VOIES AÉRO-DIGESTIVE SUPÉRIEURES CONSIDÉRANT L'ENSEMBELE DES FACTEURS ENVIRONNEMENTAUX ET SOCIAUX AU COURS DE LA VIE : HENCE LIFE STUDY

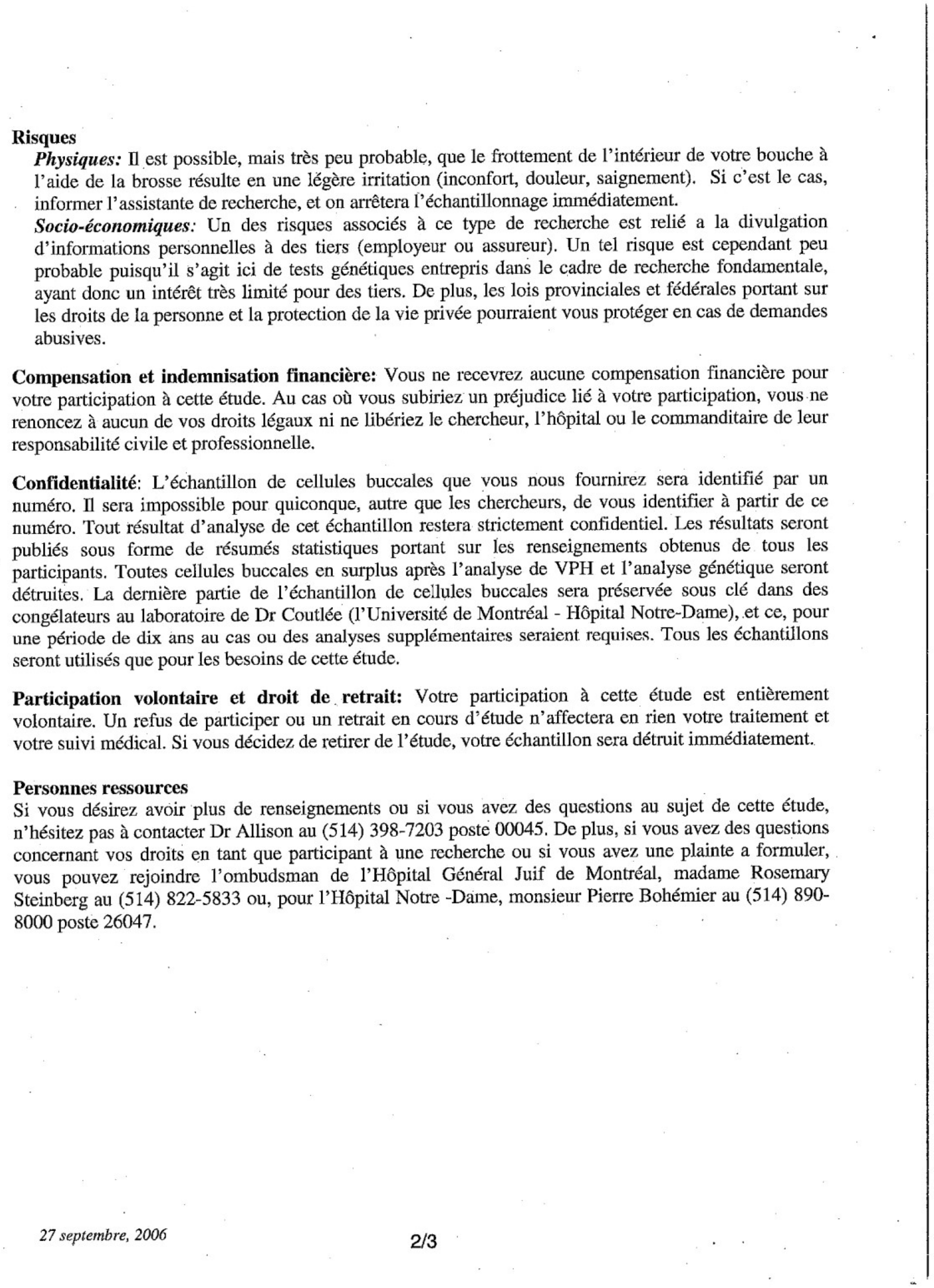
Chercheur principal: Dr Paul Allison, Professeur associé. (L'Université McGill)
Co-chercheurs: Dr Belinda Nicolau, Professeur assistante (INRS – Institut Armand Frappier),
Dr Eduardo Franco, Professeur (l'Université McGill),
Dr Marika Audet-Lapointe, Chercheur investigateur (Centre du CHUM),
Dr François Coutlée, Chercheur clinicien (Centre du CHUM),
Dr Nicolas Schlecht, Professeur assistant (Collège médical Albert Einstein, NY) et
Dr Robert Burk, Professeur (Collège médical Albert Einstein, NY).

FORMULAIRE D'INFORMATION ET DE CONSENTEMENT – CELLULES BUCCALES

Objectif: Nous entreprenons une étude importante afin d'évaluer les causes possibles des maladies de la gorge, de la bouche et du larynx. Cette étude, subventionnée par l'Institut de recherche en santé du Canada, est menée par des chercheurs qui oeuvrent dans le domaine médical à l'INRS-Institut Armand-Frappier, l'Université McGill et l'Hôpital Albert Einstein (New York). Les investigateurs principaux de cette étude sont les docteurs Paul Allison et Belinda Nicolau.

Ce que nous demandons: Vous nous avez déjà fourni votre consentement à participer à une entrevue dans le contexte de cette étude. Afin de pouvoir compléter les renseignements fournis lors de l'entrevue, nous aimerions que vous nous fournissiez un échantillon de vos cellules buccales. Ceci se fera par léger frottement de l'intérieur de votre bouche à l'aide d'une brosse souple, similaire à une brosse à dents. Ces cellules buccales seront analysées afin de mieux comprendre le rôle de certains gènes qui, lorsqu'altérés, peuvent modifier les risques de développer des maladies de la gorge, de la bouche, du larynx et du pharynx. L'échantillon de cellules buccales fourni sera envoyé au laboratoire du Dr François Coutlée (l'Université de Montréal - Hôpital Notre-Dame) où ils seront séparés en trois. La première partie de l'échantillon sera utilisée pour l`analyse de virus du papillome humain (VPH) à cette laboratoire. La deuxième partie sera utilisée pour l`analyse génétique sous la supervision des Drs Nicolas Schlecht et Robert Burk (Collège médical Albert Einstein, NY). Toutes cellules en surplus après les analyses de ces deux échantillons seront détruites. La dernière partie de l'échantillon de cellules buccales sera préservée sous clé dans des congélateurs au laboratoire de Dr Coutlée (l'Université de Montréal - Hôpital Notre-Dame), et ce, pour une période de dix ans au cas ou des analyses supplémentaires seraient requises. Tous les échantillons seront désignés par un numéro et non par les initiales ou noms des sujets, et seront utilisés que pour les besoins de cette étude.

Avantages: En participant à cette étude, vous ne bénéficierez pas d'avantages personnels. Vous pourrez cependant contribuer à améliorer les connaissances sur les causes des maladies de la bouche. Ceci pourrait faciliter leur prévention dans l'avenir.



Risques

Physiques: Il est possible, mais très peu probable, que le frottement de l'intérieur de votre bouche à l'aide de la brosse résulte en une légère irritation (inconfort, douleur, saignement). Si c'est le cas, informer l'assistante de recherche, et on arrêtera l'échantillonnage immédiatement.

Socio-économiques: Un des risques associés à ce type de recherche est relié a la divulgation d'informations personnelles à des tiers (employeur ou assureur). Un tel risque est cependant peu probable puisqu'il s'agit ici de tests génétiques entrepris dans le cadre de recherche fondamentale, ayant donc un intérêt très limité pour des tiers. De plus, les lois provinciales et fédérales portant sur les droits de la personne et la protection de la vie privée pourraient vous protéger en cas de demandes abusives.

Compensation et indemnisation financière: Vous ne recevrez aucune compensation financière pour votre participation à cette étude. Au cas où vous subiriez un préjudice lié à votre participation, vous ne renoncez à aucun de vos droits légaux ni ne libériez le chercheur, l'hôpital ou le commanditaire de leur responsabilité civile et professionnelle.

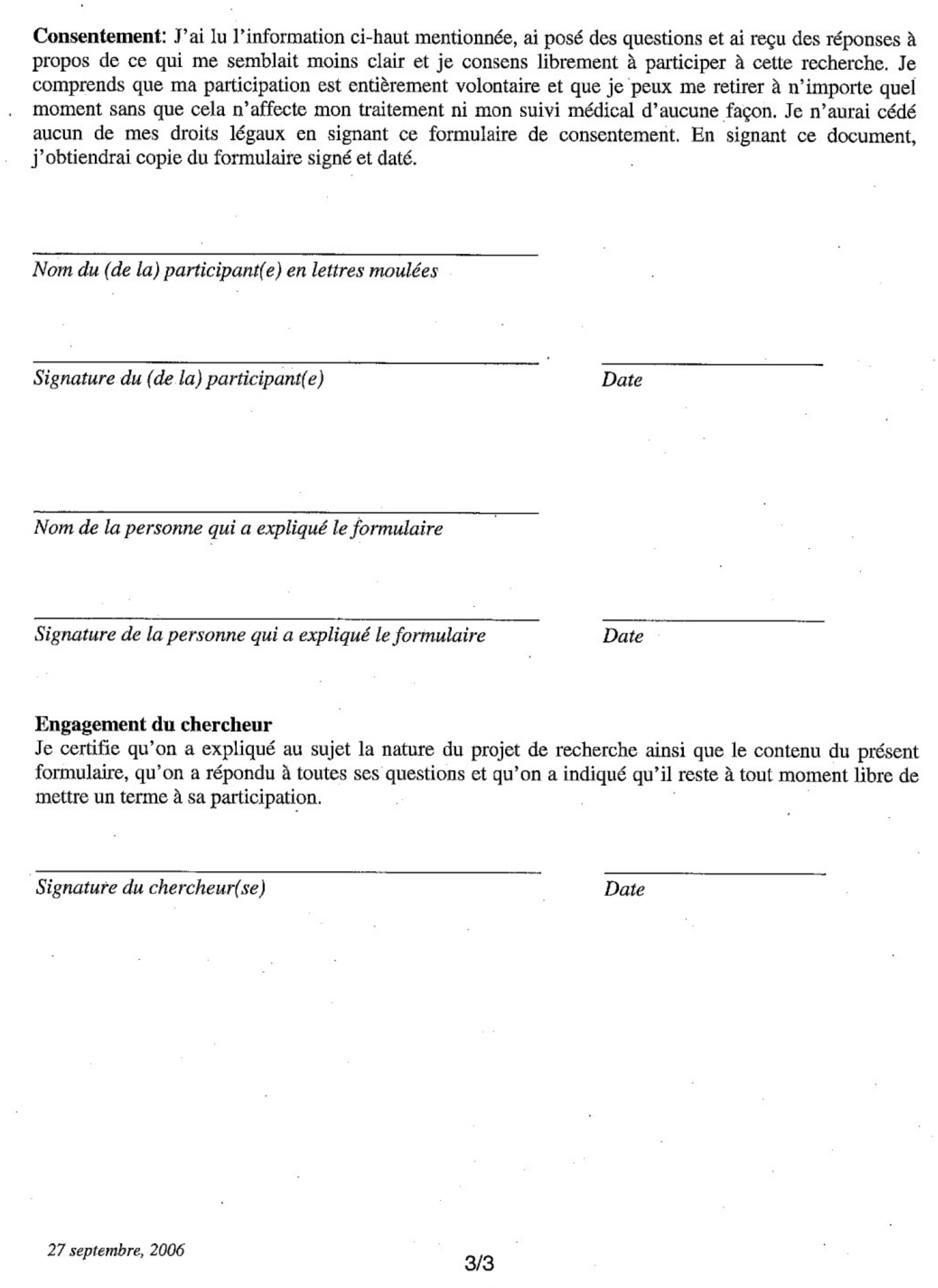
Confidentialité: L'échantillon de cellules buccales que vous nous fournirez sera identifié par un numéro. Il sera impossible pour quiconque, autre que les chercheurs, de vous identifier à partir de ce numéro. Tout résultat d'analyse de cet échantillon restera strictement confidentiel. Les résultats seront publiés sous forme de résumés statistiques portant sur les renseignements obtenus de tous les participants. Toutes cellules buccales en surplus après l'analyse de VPH et l'analyse génétique seront détruites. La dernière partie de l'échantillon de cellules buccales sera préservée sous clé dans des congélateurs au laboratoire de Dr Coutlée (l'Université de Montréal - Hôpital Notre-Dame), et ce, pour une période de dix ans au cas ou des analyses supplémentaires seraient requises. Tous les échantillons seront utilisés que pour les besoins de cette étude.

Participation volontaire et droit de retrait: Votre participation à cette étude est entièrement volontaire. Un refus de participer ou un retrait en cours d'étude n'affectera en rien votre traitement et votre suivi médical. Si vous décidez de retirer de l'étude, votre échantillon sera détruit immédiatement.

Personnes ressources

Si vous désirez avoir plus de renseignements ou si vous avez des questions au sujet de cette étude, n'hésitez pas à contacter Dr Allison au (514) 398-7203 poste 00045. De plus, si vous avez des questions concernant vos droits en tant que participant à une recherche ou si vous avez une plainte a formuler, vous pouvez rejoindre l'ombudsman de l'Hôpital Général Juif de Montréal, madame Rosemary Steinberg au (514) 822-5833 ou, pour l'Hôpital Notre -Dame, monsieur Pierre Bohémier au (514) 890-8000 poste 26047.

27 septembre, 2006 2/3



Consentement: J'ai lu l'information ci-haut mentionnée, ai posé des questions et ai reçu des réponses à propos de ce qui me semblait moins clair et je consens librement à participer à cette recherche. Je comprends que ma participation est entièrement volontaire et que je peux me retirer à n'importe quel moment sans que cela n'affecte mon traitement ni mon suivi médical d'aucune façon. Je n'aurai cédé aucun de mes droits légaux en signant ce formulaire de consentement. En signant ce document, j'obtiendrai copie du formulaire signé et daté.

____
Nom du (de la) participant(e) en lettres moulées

______________________________________ ____________________
Signature du (de la) participant(e) *Date*

____
Nom de la personne qui a expliqué le formulaire

______________________________________ ____________________
Signature de la personne qui a expliqué le formulaire *Date*

Engagement du chercheur
Je certifie qu'on a expliqué au sujet la nature du projet de recherche ainsi que le contenu du présent formulaire, qu'on a répondu à toutes ses questions et qu'on a indiqué qu'il reste à tout moment libre de mettre un terme à sa participation.

______________________________________ ____________________
Signature du chercheur(se) *Date*

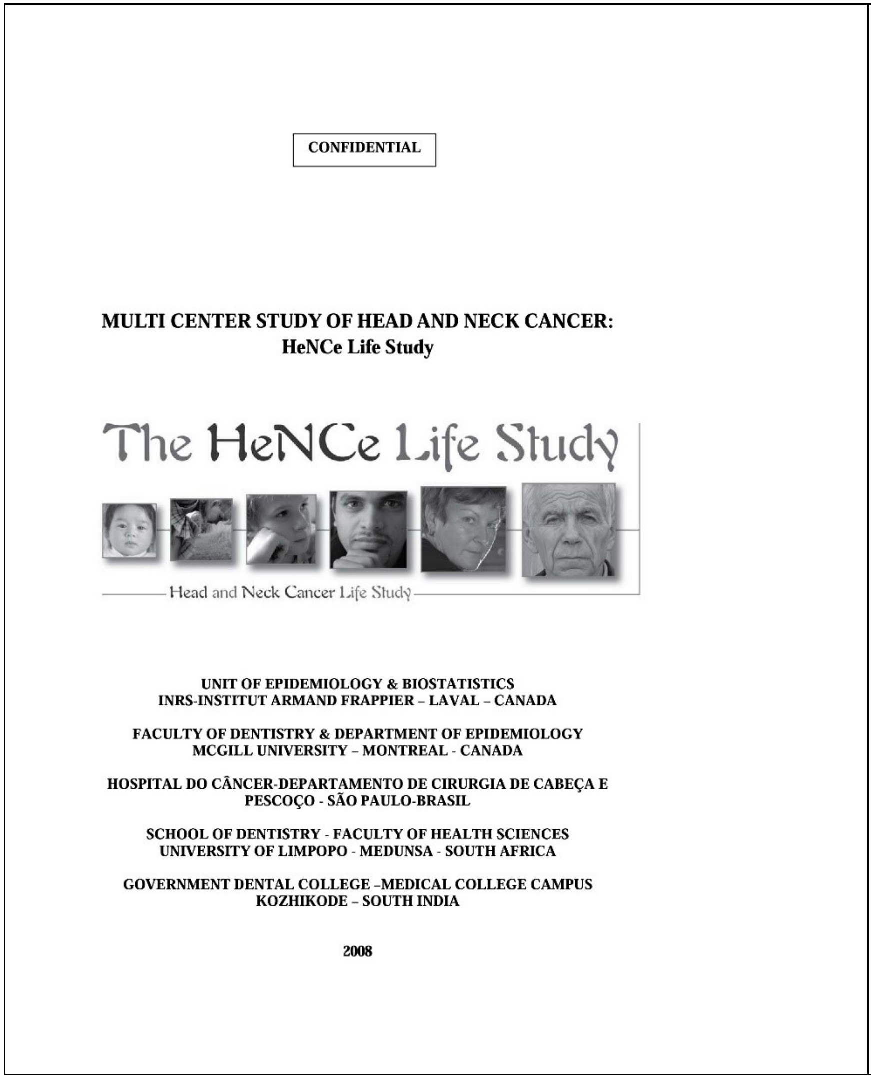
Appendix II

HeNCe life study questionnaire, India version

(For questionnaire used at the Canadian site, please visit

http://digitool.Library.McGill.CA:80/R/-?func=dbin-jump-full&object_id=130445&silo_library=GEN01)[9]

[9] Farsi N. Epidemiology of human papilloma virus related head and neck cancers [Manuscript-based]. Montreal, Canada: McGill University; 2014.



CONFIDENTIAL

MULTI CENTER STUDY OF HEAD AND NECK CANCER:
HeNCe Life Study

The HeNCe Life Study

Head and Neck Cancer Life Study

UNIT OF EPIDEMIOLOGY & BIOSTATISTICS
INRS-INSTITUT ARMAND FRAPPIER – LAVAL – CANADA

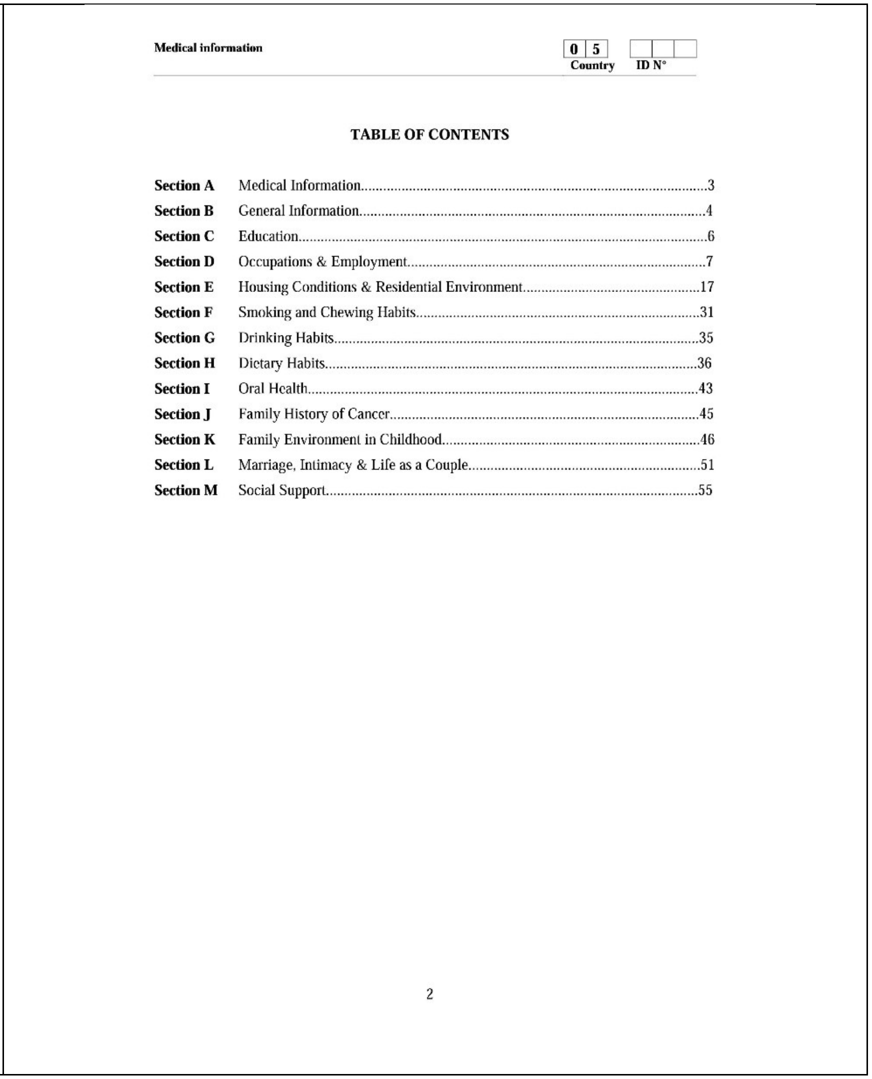
FACULTY OF DENTISTRY & DEPARTMENT OF EPIDEMIOLOGY
MCGILL UNIVERSITY – MONTREAL - CANADA

HOSPITAL DO CÂNCER-DEPARTAMENTO DE CIRURGIA DE CABEÇA E
PESCOÇO - SÃO PAULO-BRASIL

SCHOOL OF DENTISTRY - FACULTY OF HEALTH SCIENCES
UNIVERSITY OF LIMPOPO - MEDUNSA - SOUTH AFRICA

GOVERNMENT DENTAL COLLEGE –MEDICAL COLLEGE CAMPUS
KOZHIKODE – SOUTH INDIA

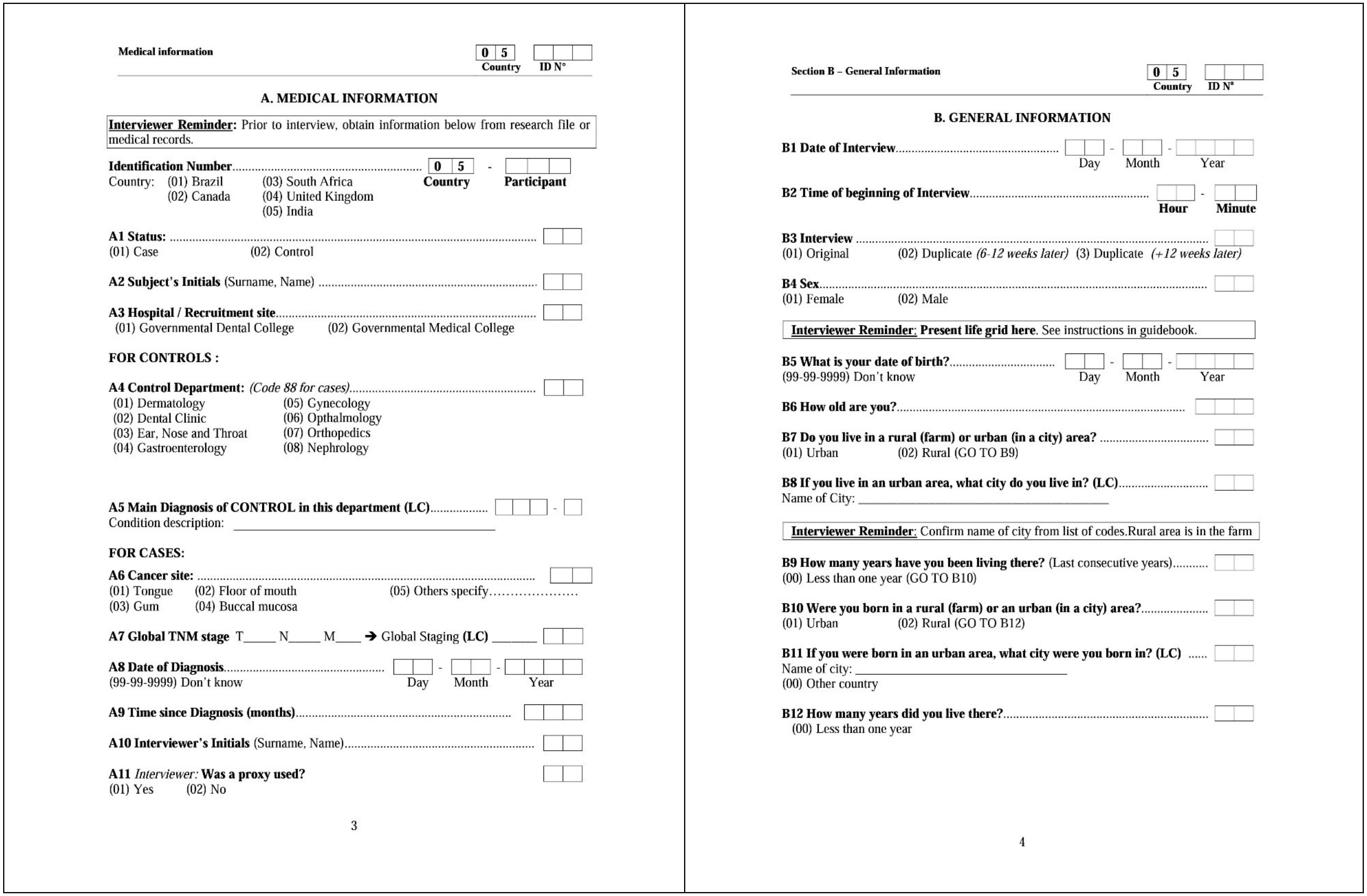
2008



Medical information 0 5 Country □□□ ID N°

TABLE OF CONTENTS

2



Medical information 0 5 □□□ Country ID N°

A. MEDICAL INFORMATION

Interviewer Reminder: Prior to interview, obtain information below from research file or medical records.

Identification Number.. 0 5 - □□□ **Country** **Participant**
Country: (01) Brazil (03) South Africa
(02) Canada (04) United Kingdom
(05) India

A1 Status: .... □□
(01) Case (02) Control

A2 Subject's Initials (Surname, Name) .... □□

A3 Hospital / Recruitment site.. □□
(01) Governmental Dental College (02) Governmental Medical College

FOR CONTROLS :

A4 Control Department: *(Code 88 for cases)*.. □□
(01) Dermatology (05) Gynecology
(02) Dental Clinic (06) Opthalmology
(03) Ear, Nose and Throat (07) Orthopedics
(04) Gastroenterology (08) Nephrology

A5 Main Diagnosis of CONTROL in this department (LC)................. □□□ - □
Condition description: ________________________________

FOR CASES:

A6 Cancer site: .... □□
(01) Tongue (02) Floor of mouth (05) Others specify…………………
(03) Gum (04) Buccal mucosa

A7 Global TNM stage T_____ N_____ M____ ➔ Global Staging **(LC)** _______ □□

A8 Date of Diagnosis.. □□ - □□ - □□□□
(99-99-9999) Don't know Day Month Year

A9 Time since Diagnosis (months)... □□□

A10 Interviewer's Initials (Surname, Name)... □□

A11 *Interviewer:* **Was a proxy used?** □□
(01) Yes (02) No

3

Section B – General Information 0 5 □□□ Country ID N°

B. GENERAL INFORMATION

B1 Date of Interview... □□ - □□ - □□□□
Day Month Year

B2 Time of beginning of Interview.. □□ - □□
Hour **Minute**

B3 Interview .... □□
(01) Original (02) Duplicate *(6-12 weeks later)* (3) Duplicate *(+12 weeks later)*

B4 Sex.. □□
(01) Female (02) Male

Interviewer Reminder: Present life grid here. See instructions in guidebook.

B5 What is your date of birth?............................... □□ - □□ - □□□□
(99-99-9999) Don't know Day Month Year

B6 How old are you?... □□□

B7 Do you live in a rural (farm) or urban (in a city) area? .... □□
(01) Urban (02) Rural (GO TO B9)

B8 If you live in an urban area, what city do you live in? (LC)........................... □□
Name of City: ______________________________________

Interviewer Reminder: Confirm name of city from list of codes.Rural area is in the farm

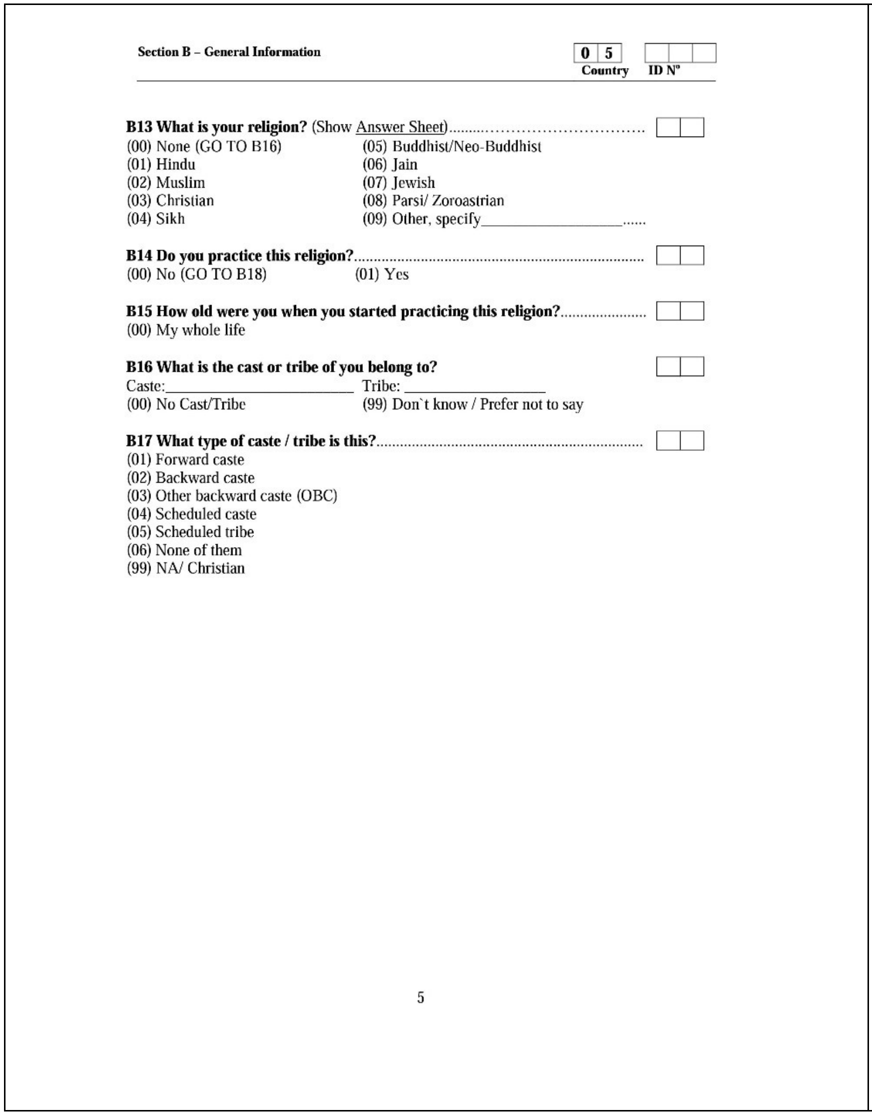
B9 How many years have you been living there? (Last consecutive years)........... □□
(00) Less than one year (GO TO B10)

B10 Were you born in a rural (farm) or an urban (in a city) area?..................... □□
(01) Urban (02) Rural (GO TO B12)

B11 If you were born in an urban area, what city were you born in? (LC) ...... □□
Name of city: ______________________________
(00) Other country

B12 How many years did you live there?... □□
(00) Less than one year

4



Section B – General Information 0 5 Country ID N° □□□

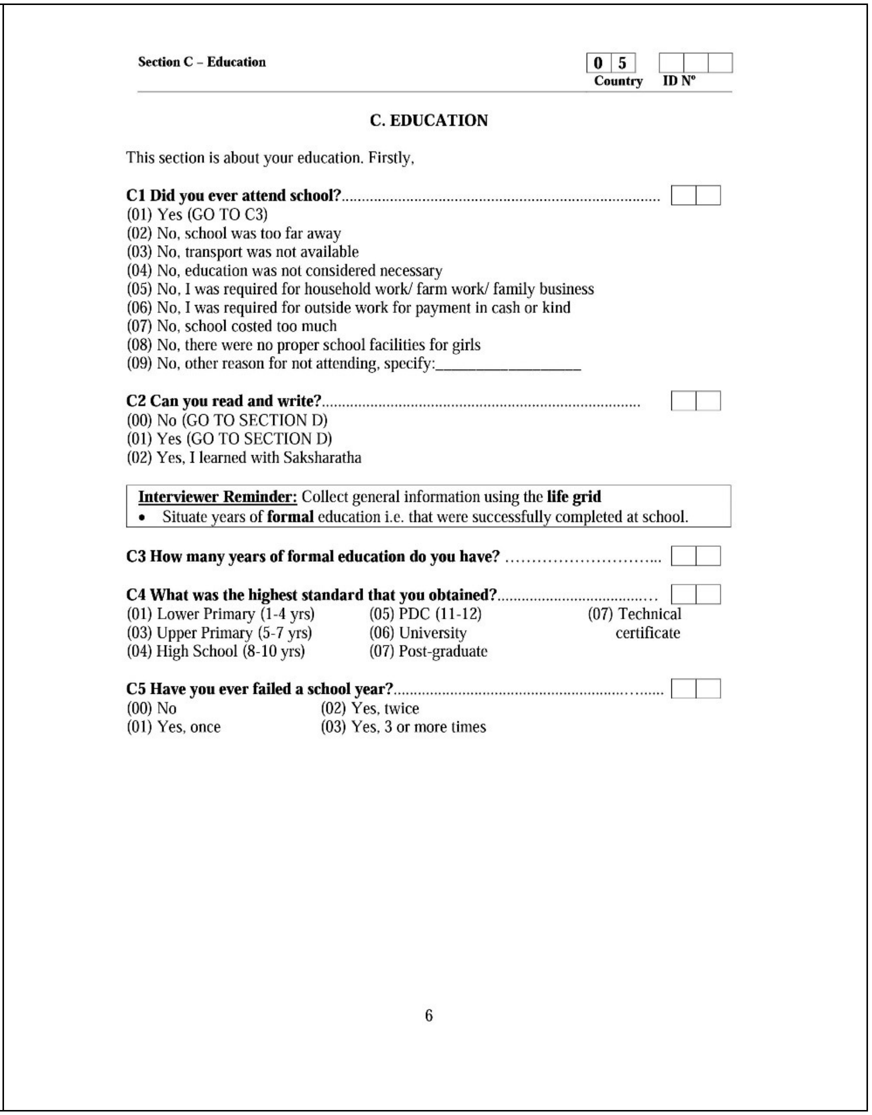
B13 What is your religion? (Show Answer Sheet)……………………………… □□
(00) None (GO TO B16)
(01) Hindu
(02) Muslim
(03) Christian
(04) Sikh
(05) Buddhist/Neo-Buddhist
(06) Jain
(07) Jewish
(08) Parsi/ Zoroastrian
(09) Other, specify__________________......

B14 Do you practice this religion?…………………………………………………… □□
(00) No (GO TO B18) (01) Yes

B15 How old were you when you started practicing this religion?…………………… □□
(00) My whole life

B16 What is the cast or tribe of you belong to? □□
Caste:________________________ Tribe: __________________
(00) No Cast/Tribe (99) Don`t know / Prefer not to say

B17 What type of caste / tribe is this?……………………………………………… □□
(01) Forward caste
(02) Backward caste
(03) Other backward caste (OBC)
(04) Scheduled caste
(05) Scheduled tribe
(06) None of them
(99) NA/ Christian



Section C – Education 0 5 Country ID N° □□□

C. EDUCATION

This section is about your education. Firstly,

C1 Did you ever attend school?…………………………………………………………… □□
(01) Yes (GO TO C3)
(02) No, school was too far away
(03) No, transport was not available
(04) No, education was not considered necessary
(05) No, I was required for household work/ farm work/ family business
(06) No, I was required for outside work for payment in cash or kind
(07) No, school costed too much
(08) No, there were no proper school facilities for girls
(09) No, other reason for not attending, specify:__________________

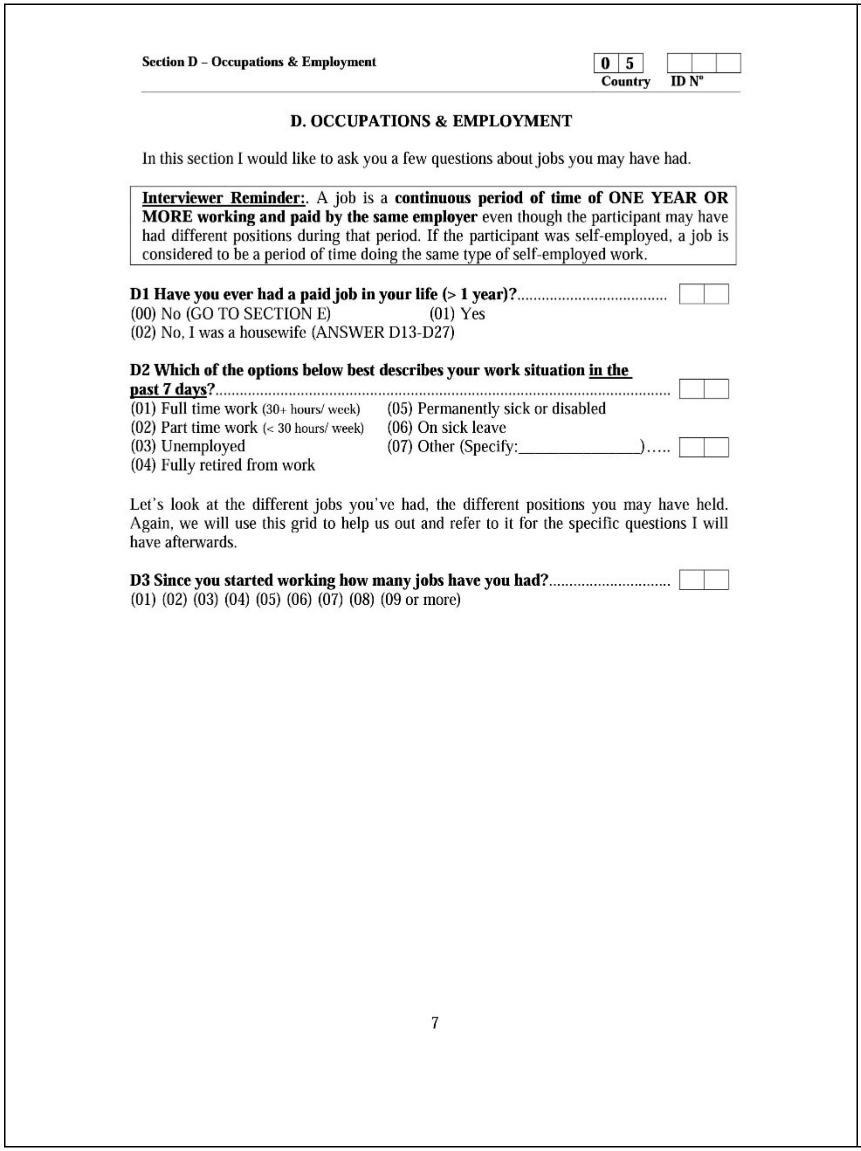
C2 Can you read and write?…………………………………………………………… □□
(00) No (GO TO SECTION D)
(01) Yes (GO TO SECTION D)
(02) Yes, I learned with Saksharatha

Interviewer Reminder: Collect general information using the **life grid**
- Situate years of **formal** education i.e. that were successfully completed at school.

C3 How many years of formal education do you have? ……………………….. □□

C4 What was the highest standard that you obtained?…………………………… □□
(01) Lower Primary (1-4 yrs)
(03) Upper Primary (5-7 yrs)
(04) High School (8-10 yrs)
(05) PDC (11-12)
(06) University
(07) Post-graduate
(07) Technical certificate

C5 Have you ever failed a school year?……………………………………………… □□
(00) No
(01) Yes, once
(02) Yes, twice
(03) Yes, 3 or more times



D. OCCUPATIONS & EMPLOYMENT

In this section I would like to ask you a few questions about jobs you may have had.

Interviewer Reminder:. A job is a **continuous period of time of ONE YEAR OR MORE working and paid by the same employer** even though the participant may have had different positions during that period. If the participant was self-employed, a job is considered to be a period of time doing the same type of self-employed work.

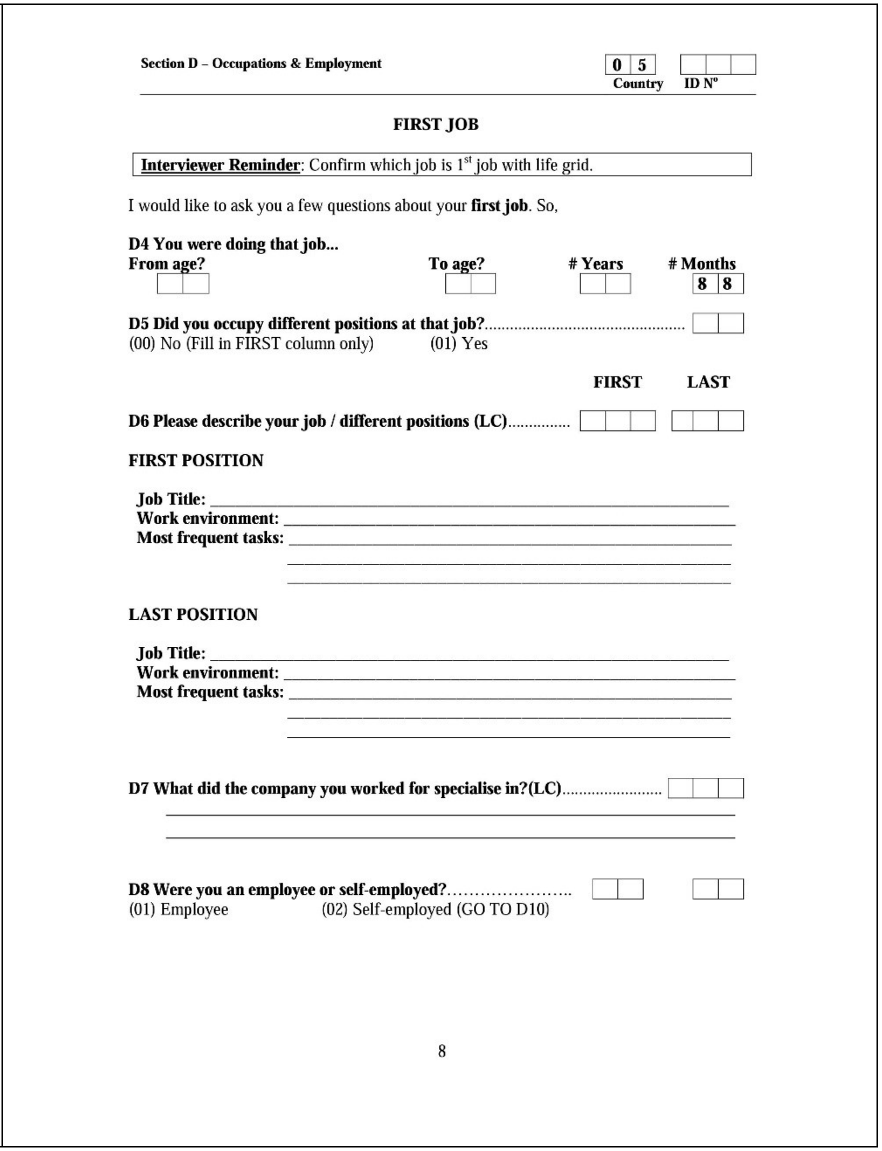
D1 Have you ever had a paid job in your life (> 1 year)?.................................... □□
(00) No (GO TO SECTION E) (01) Yes
(02) No, I was a housewife (ANSWER D13-D27)

D2 Which of the options below best describes your work situation in the past 7 days?.. □□

(01) Full time work (30+ hours/ week)	(05) Permanently sick or disabled
(02) Part time work (< 30 hours/ week)	(06) On sick leave
(03) Unemployed	(07) Other (Specify:_______________)….. □□
(04) Fully retired from work	

Let's look at the different jobs you've had, the different positions you may have held. Again, we will use this grid to help us out and refer to it for the specific questions I will have afterwards.

D3 Since you started working how many jobs have you had?.............................. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09 or more)



FIRST JOB

Interviewer Reminder: Confirm which job is 1st job with life grid.

I would like to ask you a few questions about your **first job**. So,

D4 You were doing that job...

From age?	To age?	# Years	# Months
□□	□□	□□	8 8

D5 Did you occupy different positions at that job?... □□
(00) No (Fill in FIRST column only) (01) Yes

	FIRST	LAST
D6 Please describe your job / different positions (LC)..............	□□□	□□□

FIRST POSITION

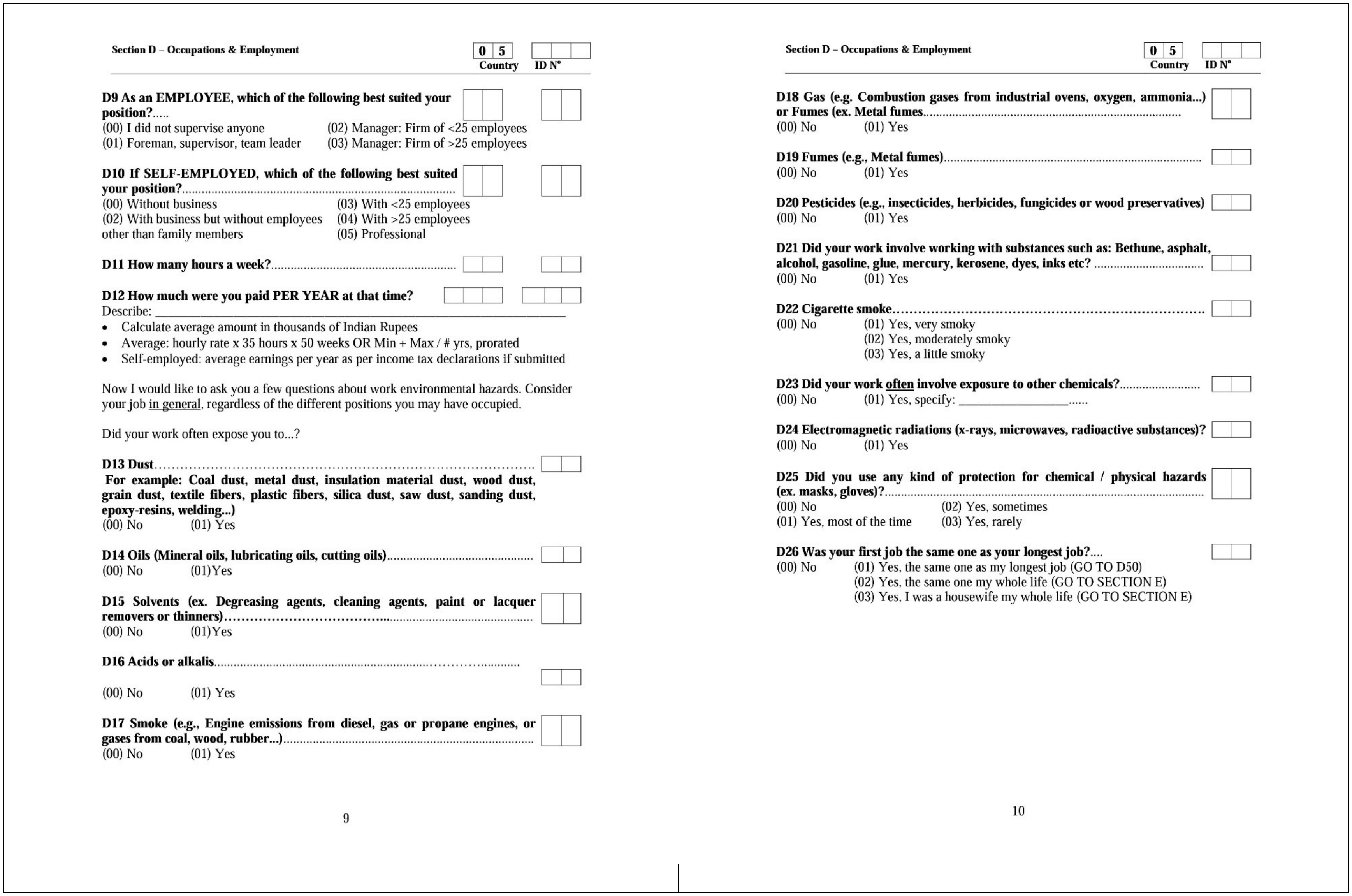
Job Title: ____
Work environment: ____
Most frequent tasks: ____
____
____

LAST POSITION

Job Title: ____
Work environment: ____
Most frequent tasks: ____
____
____

D7 What did the company you worked for specialise in?(LC)........................ □□□
____
____

D8 Were you an employee or self-employed?………………….. □□ □□
(01) Employee (02) Self-employed (GO TO D10)



Section D – Occupations & Employment 0 5 Country □□□ ID N°

D9 As an EMPLOYEE, which of the following best suited your position?..... □□ □□
(00) I did not supervise anyone (02) Manager: Firm of <25 employees
(01) Foreman, supervisor, team leader (03) Manager: Firm of >25 employees

D10 If SELF-EMPLOYED, which of the following best suited your position?.. □□ □□
(00) Without business (03) With <25 employees
(02) With business but without employees other than family members (04) With >25 employees
(05) Professional

D11 How many hours a week?... □□ □□

D12 How much were you paid PER YEAR at that time? □□□ □□□
Describe: ____

- Calculate average amount in thousands of Indian Rupees
- Average: hourly rate x 35 hours x 50 weeks OR Min + Max / # yrs, prorated
- Self-employed: average earnings per year as per income tax declarations if submitted

Now I would like to ask you a few questions about work environmental hazards. Consider your job in general, regardless of the different positions you may have occupied.

Did your work often expose you to...?

D13 Dust………………………………………………………………………….. □□
For example: Coal dust, metal dust, insulation material dust, wood dust, grain dust, textile fibers, plastic fibers, silica dust, saw dust, sanding dust, epoxy-resins, welding...)
(00) No (01) Yes

D14 Oils (Mineral oils, lubricating oils, cutting oils)... □□
(00) No (01)Yes

D15 Solvents (ex. Degreasing agents, cleaning agents, paint or lacquer removers or thinners)……………………………..……................................. □□
(00) No (01)Yes

D16 Acids or alkalis...……….……….......... □□
(00) No (01) Yes

D17 Smoke (e.g., Engine emissions from diesel, gas or propane engines, or gases from coal, wood, rubber...).. □□
(00) No (01) Yes

9

Section D – Occupations & Employment 0 5 Country □□□ ID N°

D18 Gas (e.g. Combustion gases from industrial ovens, oxygen, ammonia...) or Fumes (ex. Metal fumes.. □□
(00) No (01) Yes

D19 Fumes (e.g., Metal fumes).. □□
(00) No (01) Yes

D20 Pesticides (e.g., insecticides, herbicides, fungicides or wood preservatives) □□
(00) No (01) Yes

D21 Did your work involve working with substances such as: Bethune, asphalt, alcohol, gasoline, glue, mercury, kerosene, dyes, inks etc? .... □□
(00) No (01) Yes

D22 Cigarette smoke……………………………………………………………… □□
(00) No (01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

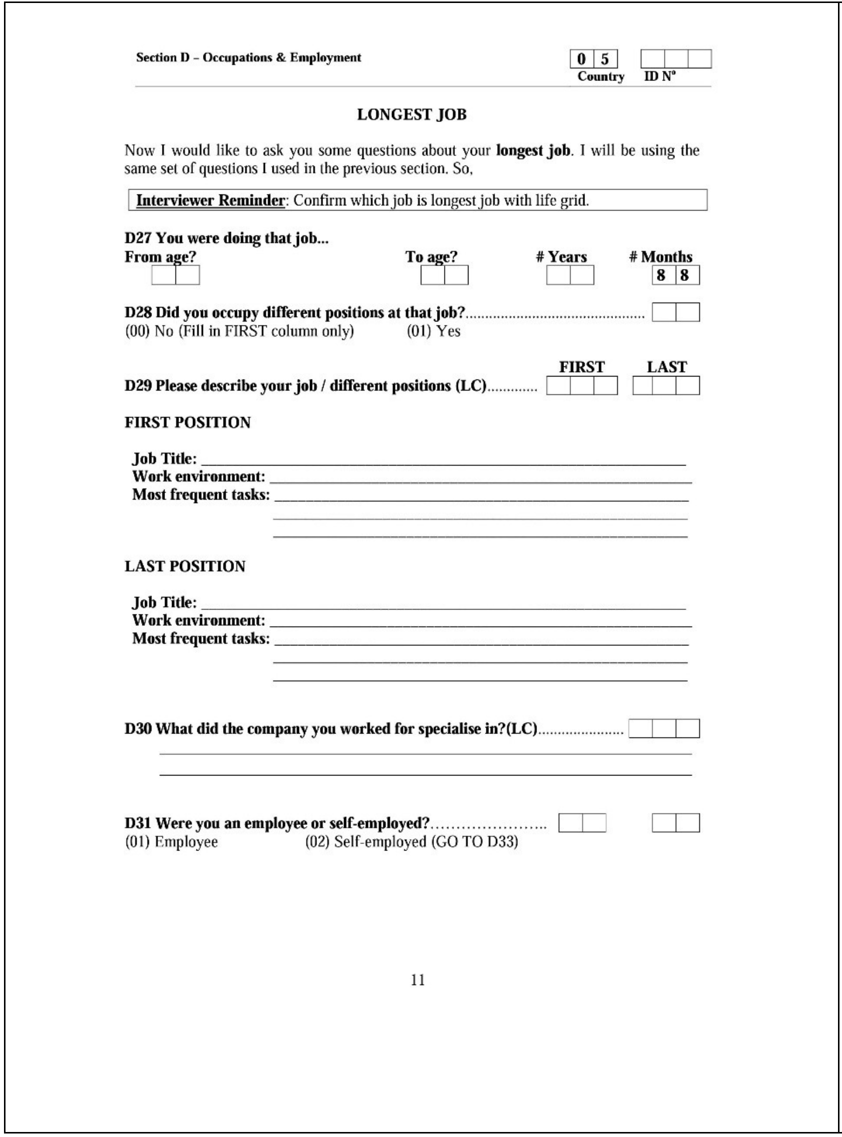
D23 Did your work often involve exposure to other chemicals?........................ □□
(00) No (01) Yes, specify: _______________......

D24 Electromagnetic radiations (x-rays, microwaves, radioactive substances)? □□
(00) No (01) Yes

D25 Did you use any kind of protection for chemical / physical hazards (ex. masks, gloves)?... □□
(00) No (02) Yes, sometimes
(01) Yes, most of the time (03) Yes, rarely

D26 Was your first job the same one as your longest job?.... □□
(00) No (01) Yes, the same one as my longest job (GO TO D50)
(02) Yes, the same one my whole life (GO TO SECTION E)
(03) Yes, I was a housewife my whole life (GO TO SECTION E)

10



Section D – Occupations & Employment

0 5 Country □□□ ID N°

LONGEST JOB

Now I would like to ask you some questions about your **longest job**. I will be using the same set of questions I used in the previous section. So,

Interviewer Reminder: Confirm which job is longest job with life grid.

D27 You were doing that job...

From age?	To age?	# Years	# Months
□□	□□	□□	8 8

D28 Did you occupy different positions at that job?.. □□
(00) No (Fill in FIRST column only) (01) Yes

	FIRST	LAST
D29 Please describe your job / different positions (LC)............	□□□	□□□

FIRST POSITION

Job Title: ______________________________
Work environment: ______________________________
Most frequent tasks: ______________________________
____
____

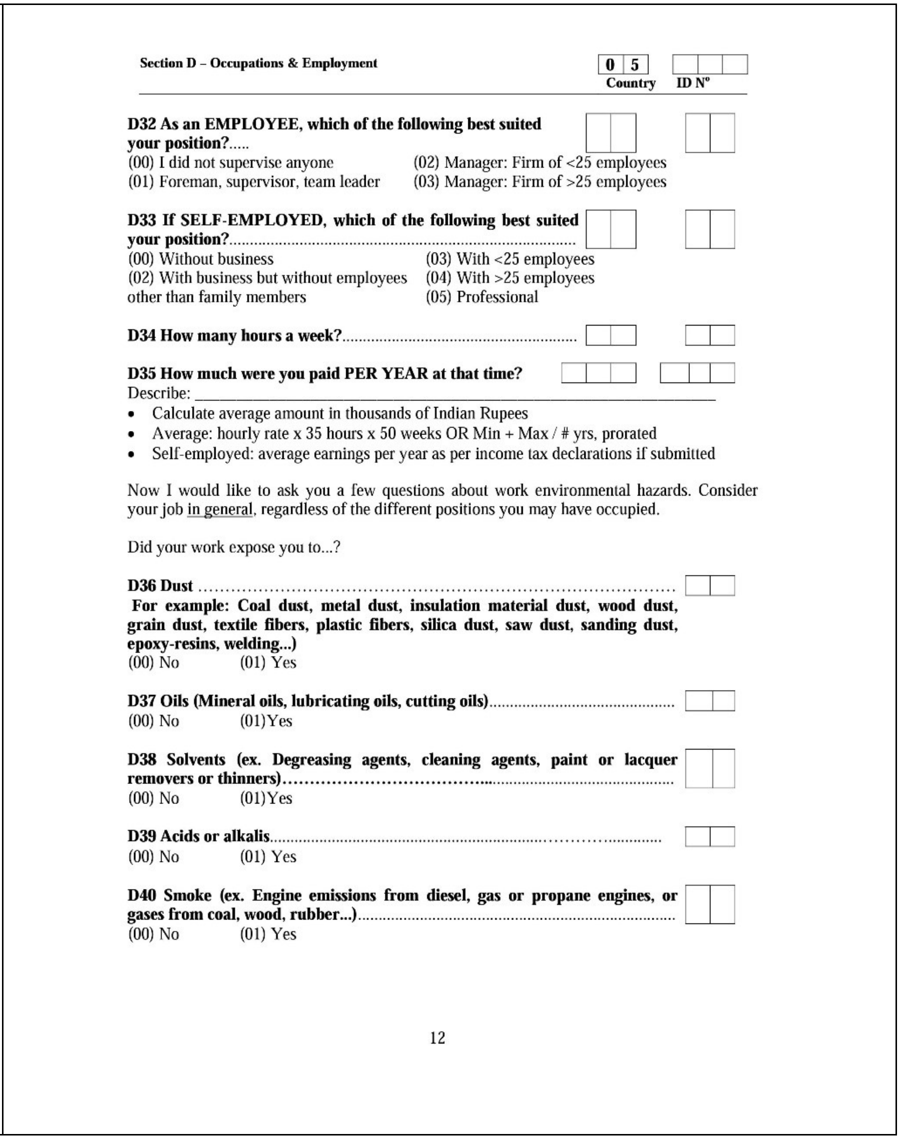
LAST POSITION

Job Title: ______________________________
Work environment: ______________________________
Most frequent tasks: ______________________________
____
____

D30 What did the company you worked for specialise in?(LC)...................... □□□
____
____

D31 Were you an employee or self-employed?…………………... □□ □□
(01) Employee (02) Self-employed (GO TO D33)

11



Section D – Occupations & Employment

0 5 Country □□□ ID N°

D32 As an EMPLOYEE, which of the following best suited your position?..... □□ □□
(00) I did not supervise anyone (02) Manager: Firm of <25 employees
(01) Foreman, supervisor, team leader (03) Manager: Firm of >25 employees

D33 If SELF-EMPLOYED, which of the following best suited your position?... □□ □□
(00) Without business (03) With <25 employees
(02) With business but without employees other than family members (04) With >25 employees
(05) Professional

D34 How many hours a week?.. □□ □□

D35 How much were you paid PER YEAR at that time? □□□ □□□
Describe: ______________________________

- Calculate average amount in thousands of Indian Rupees
- Average: hourly rate x 35 hours x 50 weeks OR Min + Max / # yrs, prorated
- Self-employed: average earnings per year as per income tax declarations if submitted

Now I would like to ask you a few questions about work environmental hazards. Consider your job in general, regardless of the different positions you may have occupied.

Did your work expose you to...?

D36 Dust ……………………………………………………………………………… □□
For example: Coal dust, metal dust, insulation material dust, wood dust, grain dust, textile fibers, plastic fibers, silica dust, saw dust, sanding dust, epoxy-resins, welding...)
(00) No (01) Yes

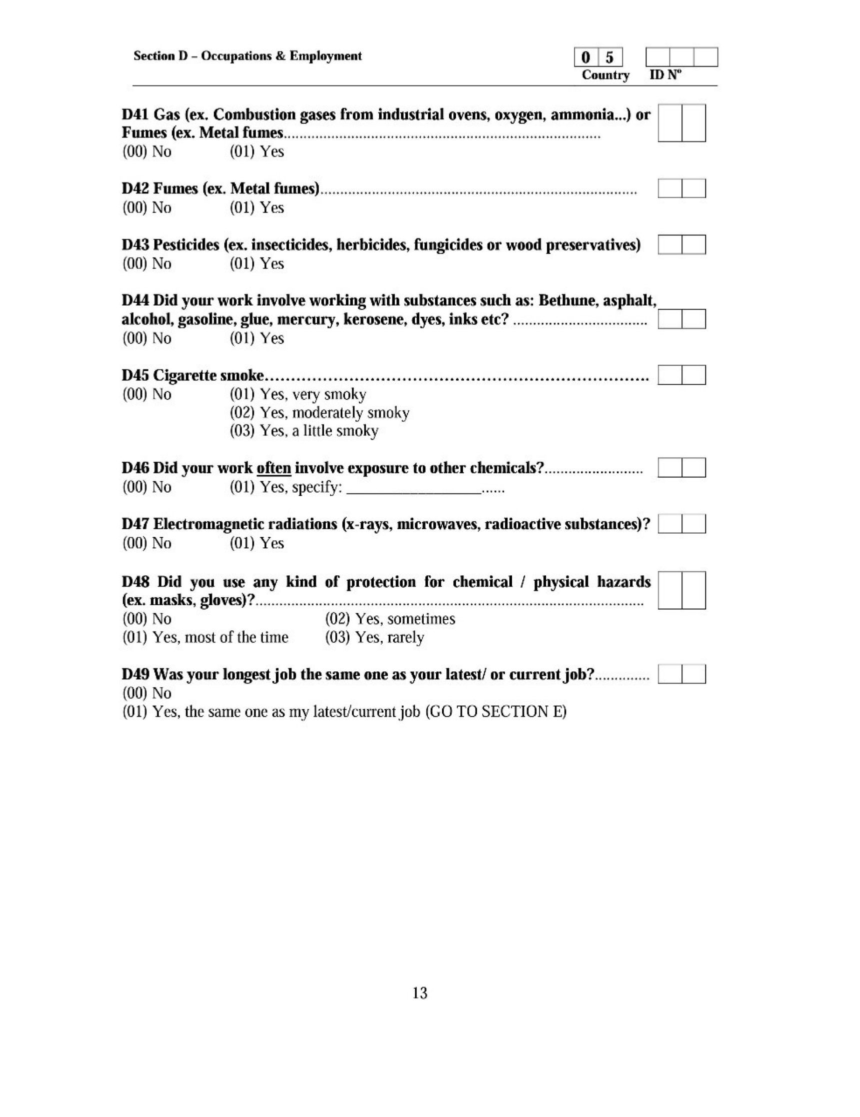
D37 Oils (Mineral oils, lubricating oils, cutting oils).. □□
(00) No (01)Yes

D38 Solvents (ex. Degreasing agents, cleaning agents, paint or lacquer removers or thinners)………………………………... □□
(00) No (01)Yes

D39 Acids or alkalis...…………............. □□
(00) No (01) Yes

D40 Smoke (ex. Engine emissions from diesel, gas or propane engines, or gases from coal, wood, rubber...)... □□
(00) No (01) Yes

12



Section D – Occupations & Employment | 0 | 5 | Country | □□□ ID N°

D41 Gas (ex. Combustion gases from industrial ovens, oxygen, ammonia...) or Fumes (ex. Metal fumes.. □□
(00) No (01) Yes

D42 Fumes (ex. Metal fumes).. □□
(00) No (01) Yes

D43 Pesticides (ex. insecticides, herbicides, fungicides or wood preservatives) □□
(00) No (01) Yes

D44 Did your work involve working with substances such as: Bethune, asphalt, alcohol, gasoline, glue, mercury, kerosene, dyes, inks etc? .... □□
(00) No (01) Yes

D45 Cigarette smoke………………………………………………………………. □□
(00) No (01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

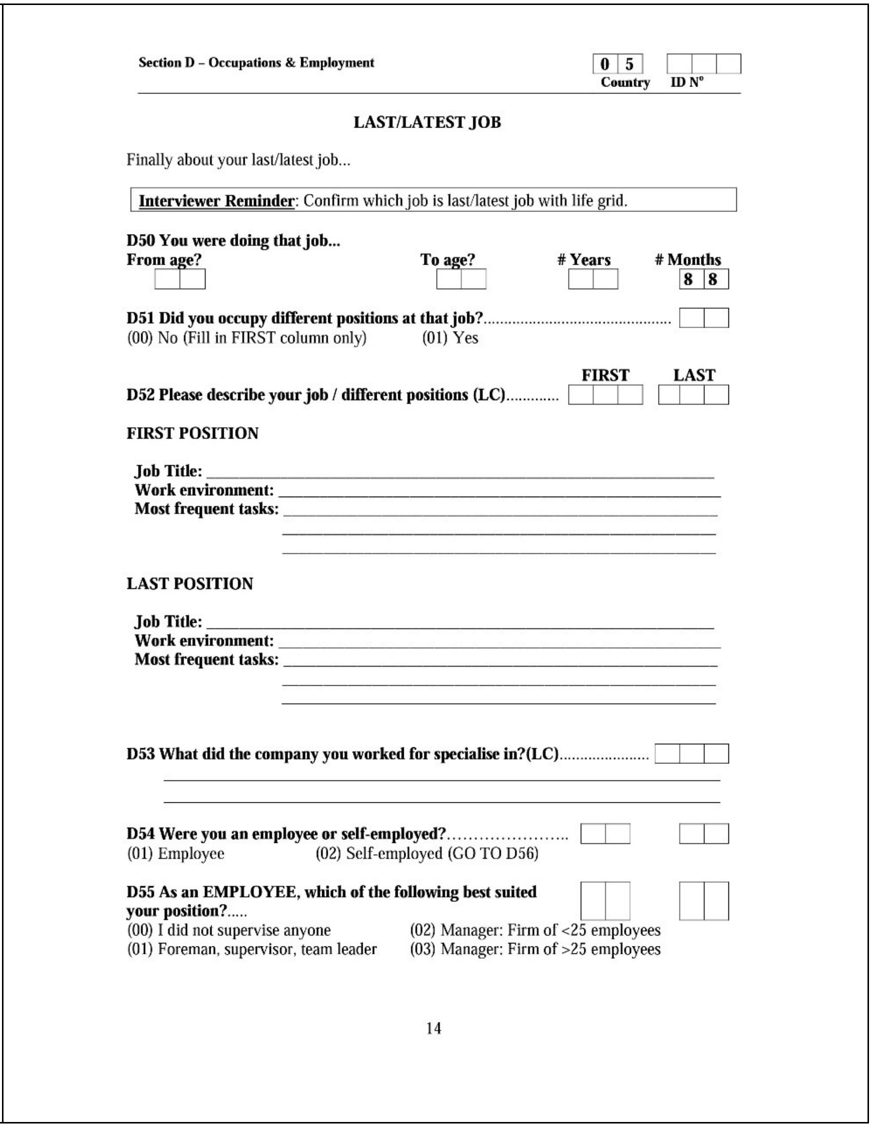
D46 Did your work <u>often</u> involve exposure to other chemicals?........................ □□
(00) No (01) Yes, specify: _________________......

D47 Electromagnetic radiations (x-rays, microwaves, radioactive substances)? □□
(00) No (01) Yes

D48 Did you use any kind of protection for chemical / physical hazards (ex. masks, gloves)?.. □□
(00) No (02) Yes, sometimes
(01) Yes, most of the time (03) Yes, rarely

D49 Was your longest job the same one as your latest/ or current job?............. □□
(00) No
(01) Yes, the same one as my latest/current job (GO TO SECTION E)

13



Section D – Occupations & Employment | 0 | 5 | Country | □□□ ID N°

LAST/LATEST JOB

Finally about your last/latest job...

<u>Interviewer Reminder</u>: Confirm which job is last/latest job with life grid.

D50 You were doing that job...

From <u>age</u>?	**To <u>age</u>?**	**# Years**	**# Months**
□□	□□	□□	8 8

D51 Did you occupy different positions at that job?.. □□
(00) No (Fill in FIRST column only) (01) Yes

D52 Please describe your job / different positions (LC)............ **FIRST** □□□ **LAST** □□□

FIRST POSITION

Job Title: ____
Work environment: ____
Most frequent tasks: ____
____
____

LAST POSITION

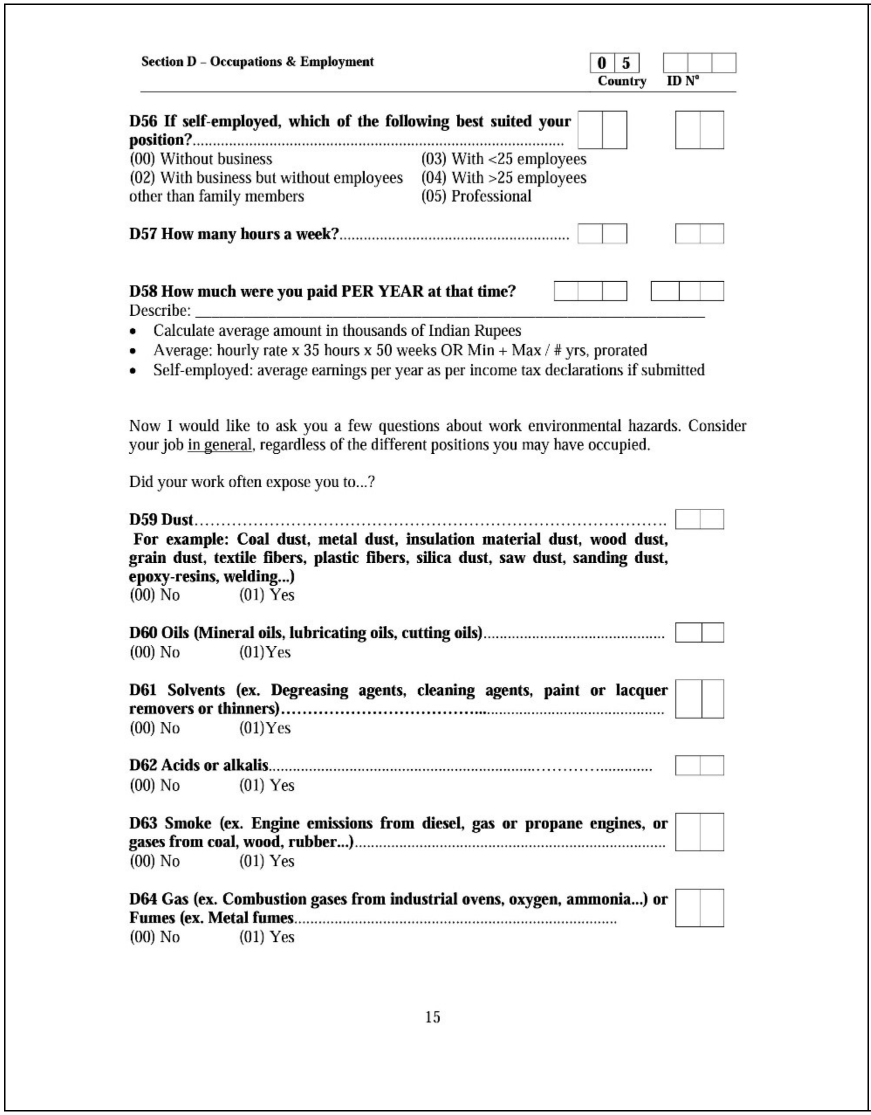
Job Title: ____
Work environment: ____
Most frequent tasks: ____
____
____

D53 What did the company you worked for specialise in?(LC)..................... □□□
____
____

D54 Were you an employee or self-employed?………………….. □□ □□
(01) Employee (02) Self-employed (GO TO D56)

D55 As an EMPLOYEE, which of the following best suited your position?..... □□ □□
(00) I did not supervise anyone (02) Manager: Firm of <25 employees
(01) Foreman, supervisor, team leader (03) Manager: Firm of >25 employees

14



Section D – Occupations & Employment 0 5 Country □□□ ID N°

D56 If self-employed, which of the following best suited your position?.. □□ □□

(00) Without business	(03) With <25 employees
(02) With business but without employees other than family members	(04) With >25 employees (05) Professional

D57 How many hours a week?.. □□ □□

D58 How much were you paid PER YEAR at that time? □□□ □□□

Describe: ____

- Calculate average amount in thousands of Indian Rupees
- Average: hourly rate x 35 hours x 50 weeks OR Min + Max / # yrs, prorated
- Self-employed: average earnings per year as per income tax declarations if submitted

Now I would like to ask you a few questions about work environmental hazards. Consider your job in general, regardless of the different positions you may have occupied.

Did your work often expose you to...?

D59 Dust…………………………………………………………………………… □□
For example: Coal dust, metal dust, insulation material dust, wood dust, grain dust, textile fibers, plastic fibers, silica dust, saw dust, sanding dust, epoxy-resins, welding...)
(00) No (01) Yes

D60 Oils (Mineral oils, lubricating oils, cutting oils).. □□
(00) No (01)Yes

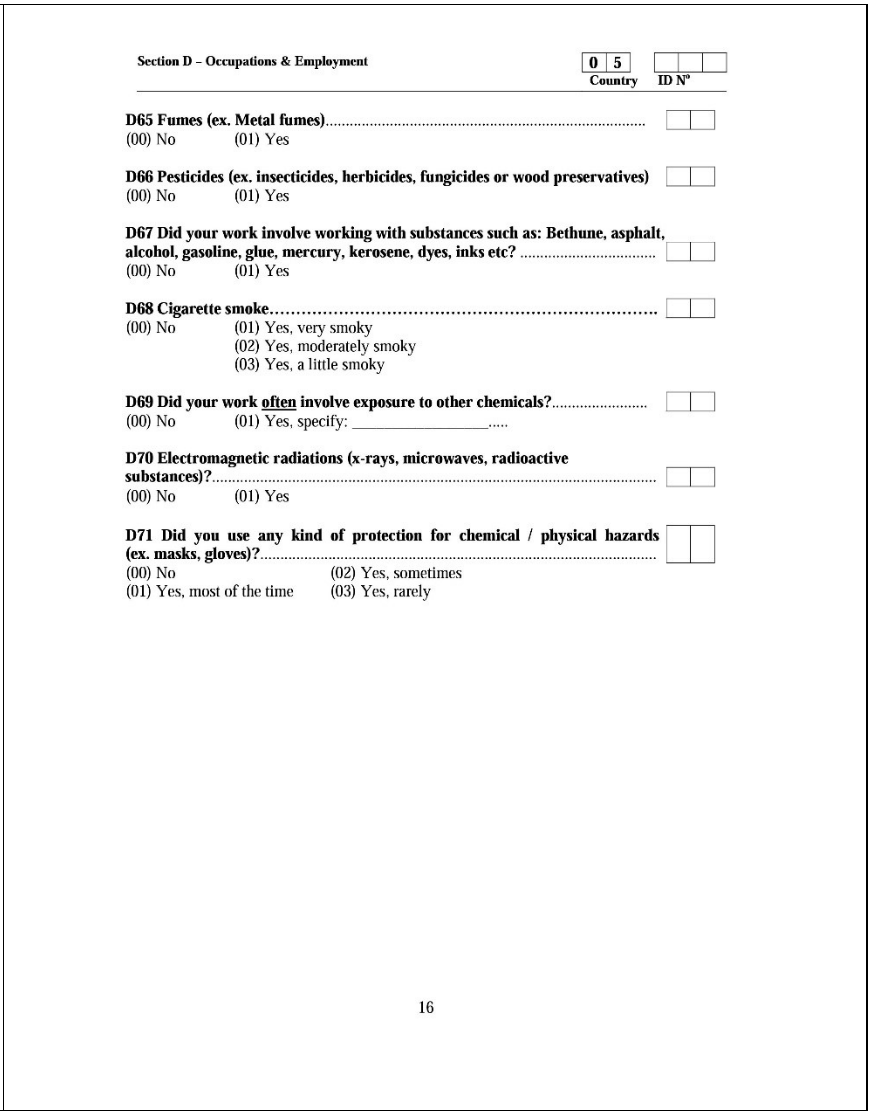
D61 Solvents (ex. Degreasing agents, cleaning agents, paint or lacquer removers or thinners)……………………………...………................................ □□
(00) No (01)Yes

D62 Acids or alkalis...………….............. □□
(00) No (01) Yes

D63 Smoke (ex. Engine emissions from diesel, gas or propane engines, or gases from coal, wood, rubber...)... □□
(00) No (01) Yes

D64 Gas (ex. Combustion gases from industrial ovens, oxygen, ammonia...) or Fumes (ex. Metal fumes... □□
(00) No (01) Yes

15



Section D – Occupations & Employment 0 5 Country □□□ ID N°

D65 Fumes (ex. Metal fumes).. □□
(00) No (01) Yes

D66 Pesticides (ex. insecticides, herbicides, fungicides or wood preservatives) □□
(00) No (01) Yes

D67 Did your work involve working with substances such as: Bethune, asphalt, alcohol, gasoline, glue, mercury, kerosene, dyes, inks etc? .... □□
(00) No (01) Yes

D68 Cigarette smoke………………………………………………………………. □□
(00) No (01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

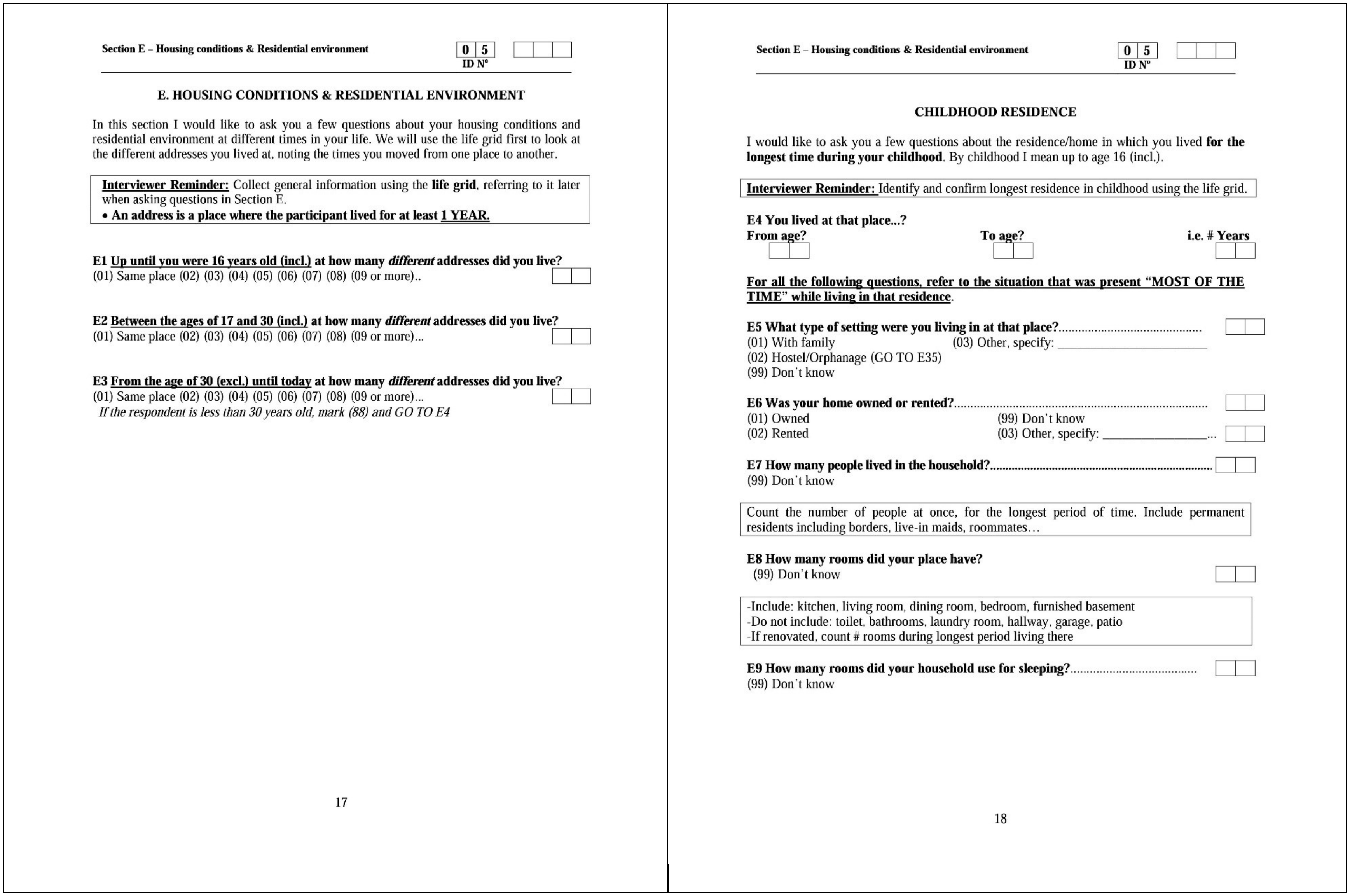
D69 Did your work often involve exposure to other chemicals?........................ □□
(00) No (01) Yes, specify: ________________.....

D70 Electromagnetic radiations (x-rays, microwaves, radioactive substances)?.. □□
(00) No (01) Yes

D71 Did you use any kind of protection for chemical / physical hazards (ex. masks, gloves)?... □□

(00) No	(02) Yes, sometimes
(01) Yes, most of the time	(03) Yes, rarely

16



E. HOUSING CONDITIONS & RESIDENTIAL ENVIRONMENT

In this section I would like to ask you a few questions about your housing conditions and residential environment at different times in your life. We will use the life grid first to look at the different addresses you lived at, noting the times you moved from one place to another.

Interviewer Reminder: Collect general information using the **life grid**, referring to it later when asking questions in Section E.
- **An address is a place where the participant lived for at least 1 YEAR.**

E1 Up until you were 16 years old (incl.) at how many *different* addresses did you live? □□
(01) Same place (02) (03) (04) (05) (06) (07) (08) (09 or more)..

E2 Between the ages of 17 and 30 (incl.) at how many *different* addresses did you live? □□
(01) Same place (02) (03) (04) (05) (06) (07) (08) (09 or more)...

E3 From the age of 30 (excl.) until today at how many *different* addresses did you live? □□
(01) Same place (02) (03) (04) (05) (06) (07) (08) (09 or more)...
If the respondent is less than 30 years old, mark (88) and GO TO E4

CHILDHOOD RESIDENCE

I would like to ask you a few questions about the residence/home in which you lived **for the longest time during your childhood**. By childhood I mean up to age 16 (incl.).

Interviewer Reminder: Identify and confirm longest residence in childhood using the life grid.

E4 You lived at that place...?
From age? □□ **To age?** □□ **i.e. # Years** □□

For all the following questions, refer to the situation that was present "MOST OF THE TIME" while living in that residence.

E5 What type of setting were you living in at that place?... □□
(01) With family (03) Other, specify: ______________________
(02) Hostel/Orphanage (GO TO E35)
(99) Don't know

E6 Was your home owned or rented?.. □□
(01) Owned (99) Don't know
(02) Rented (03) Other, specify: _______________... □□

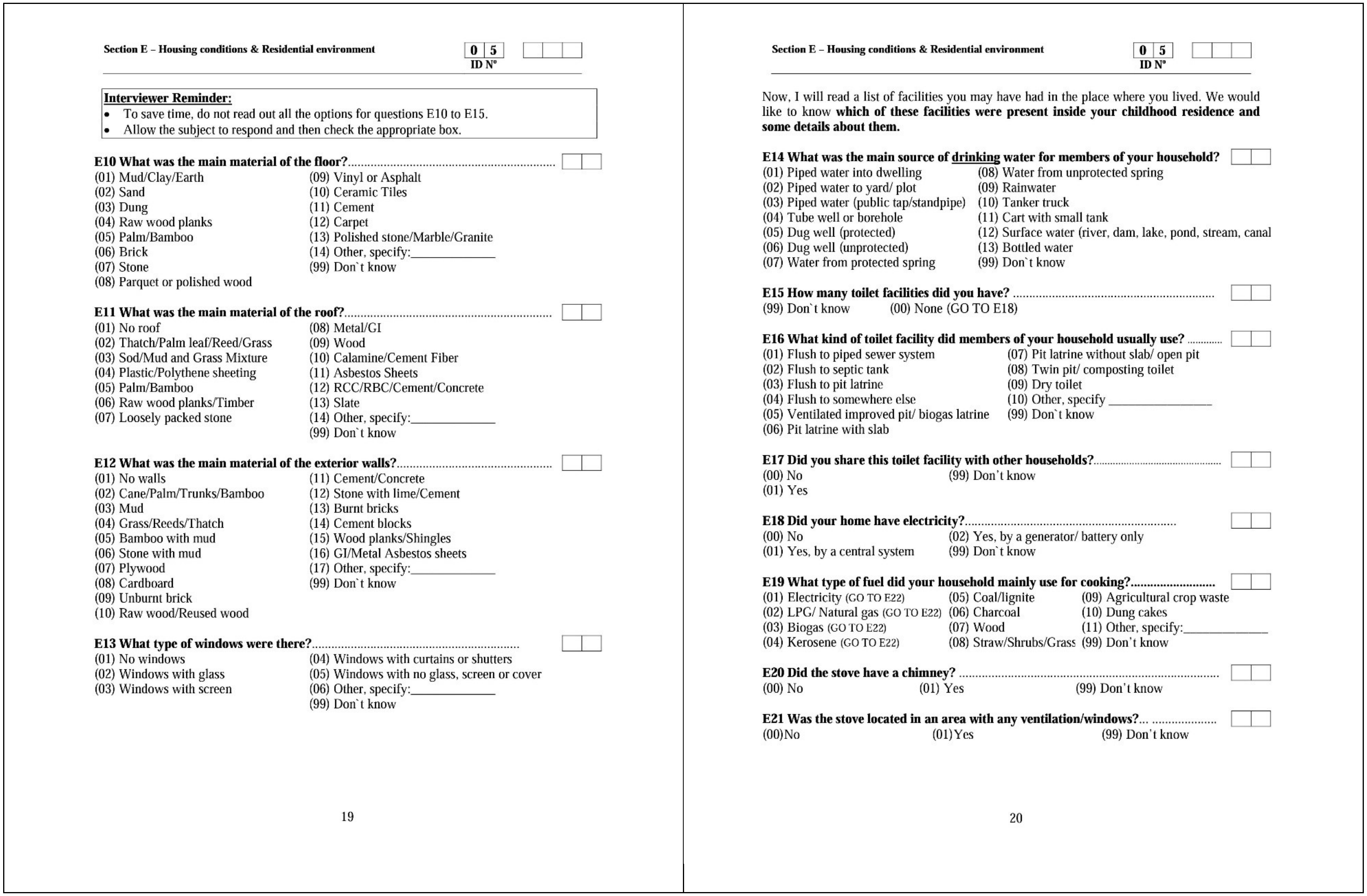
E7 How many people lived in the household?.. □□
(99) Don't know

Count the number of people at once, for the longest period of time. Include permanent residents including borders, live-in maids, roommates…

E8 How many rooms did your place have? □□
(99) Don't know

-Include: kitchen, living room, dining room, bedroom, furnished basement
-Do not include: toilet, bathrooms, laundry room, hallway, garage, patio
-If renovated, count # rooms during longest period living there

E9 How many rooms did your household use for sleeping?...................................... □□
(99) Don't know



Section E – Housing conditions & Residential environment 0 5 □□□
ID N°

Interviewer Reminder:
- To save time, do not read out all the options for questions E10 to E15.
- Allow the subject to respond and then check the appropriate box.

E10 What was the main material of the floor?... □□

(01) Mud/Clay/Earth
(02) Sand
(03) Dung
(04) Raw wood planks
(05) Palm/Bamboo
(06) Brick
(07) Stone
(08) Parquet or polished wood
(09) Vinyl or Asphalt
(10) Ceramic Tiles
(11) Cement
(12) Carpet
(13) Polished stone/Marble/Granite
(14) Other, specify:____________
(99) Don`t know

E11 What was the main material of the roof?.. □□

(01) No roof
(02) Thatch/Palm leaf/Reed/Grass
(03) Sod/Mud and Grass Mixture
(04) Plastic/Polythene sheeting
(05) Palm/Bamboo
(06) Raw wood planks/Timber
(07) Loosely packed stone
(08) Metal/GI
(09) Wood
(10) Calamine/Cement Fiber
(11) Asbestos Sheets
(12) RCC/RBC/Cement/Concrete
(13) Slate
(14) Other, specify:____________
(99) Don`t know

E12 What was the main material of the exterior walls?... □□

(01) No walls
(02) Cane/Palm/Trunks/Bamboo
(03) Mud
(04) Grass/Reeds/Thatch
(05) Bamboo with mud
(06) Stone with mud
(07) Plywood
(08) Cardboard
(09) Unburnt brick
(10) Raw wood/Reused wood
(11) Cement/Concrete
(12) Stone with lime/Cement
(13) Burnt bricks
(14) Cement blocks
(15) Wood planks/Shingles
(16) GI/Metal Asbestos sheets
(17) Other, specify:____________
(99) Don`t know

E13 What type of windows were there?.. □□

(01) No windows
(02) Windows with glass
(03) Windows with screen
(04) Windows with curtains or shutters
(05) Windows with no glass, screen or cover
(06) Other, specify:____________
(99) Don`t know

Section E – Housing conditions & Residential environment 0 5 □□□
ID N°

Now, I will read a list of facilities you may have had in the place where you lived. We would like to know **which of these facilities were present inside your childhood residence and some details about them.**

E14 What was the main source of drinking water for members of your household? □□

(01) Piped water into dwelling
(02) Piped water to yard/ plot
(03) Piped water (public tap/standpipe)
(04) Tube well or borehole
(05) Dug well (protected)
(06) Dug well (unprotected)
(07) Water from protected spring
(08) Water from unprotected spring
(09) Rainwater
(10) Tanker truck
(11) Cart with small tank
(12) Surface water (river, dam, lake, pond, stream, canal
(13) Bottled water
(99) Don`t know

E15 How many toilet facilities did you have? .... □□

(99) Don`t know (00) None (GO TO E18)

E16 What kind of toilet facility did members of your household usually use? .... □□

(01) Flush to piped sewer system
(02) Flush to septic tank
(03) Flush to pit latrine
(04) Flush to somewhere else
(05) Ventilated improved pit/ biogas latrine
(06) Pit latrine with slab
(07) Pit latrine without slab/ open pit
(08) Twin pit/ composting toilet
(09) Dry toilet
(10) Other, specify _______________
(99) Don`t know

E17 Did you share this toilet facility with other households?.. □□

(00) No
(01) Yes
(99) Don't know

E18 Did your home have electricity?.. □□

(00) No
(01) Yes, by a central system
(02) Yes, by a generator/ battery only
(99) Don`t know

E19 What type of fuel did your household mainly use for cooking?......................... □□

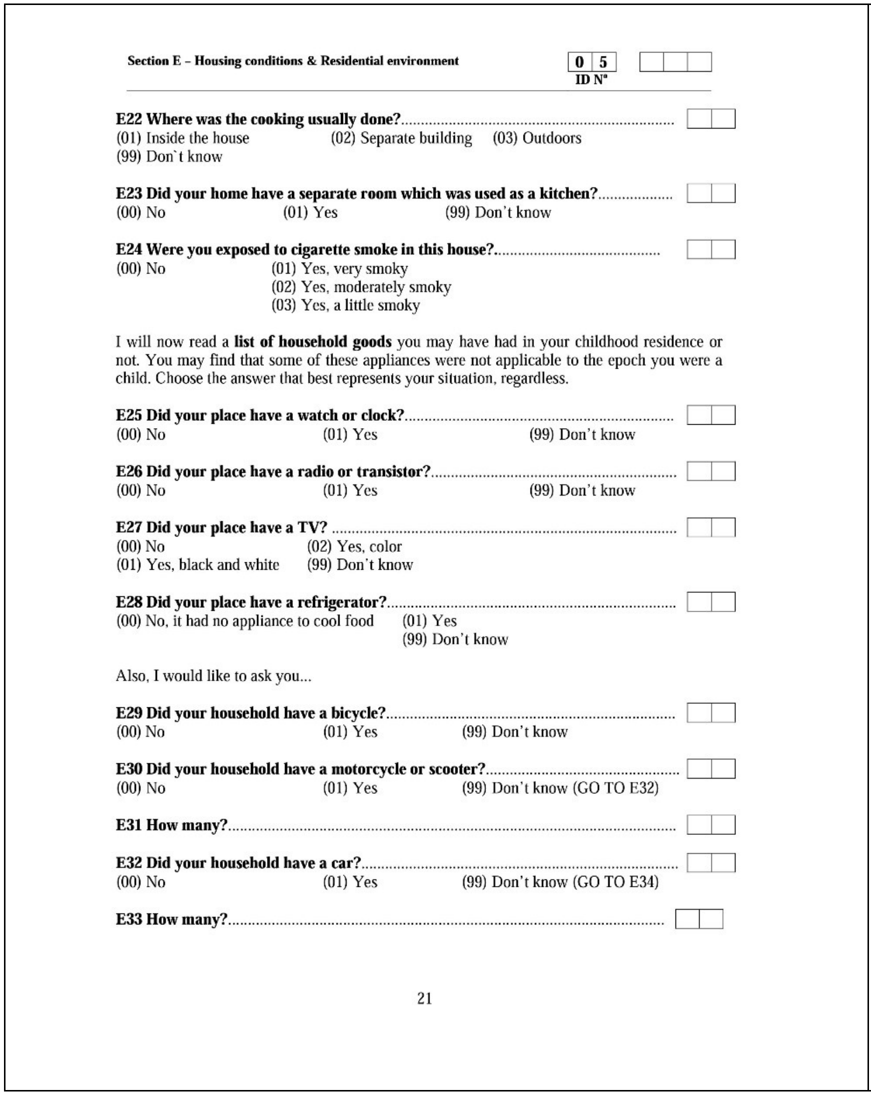
(01) Electricity (GO TO E22)
(02) LPG/ Natural gas (GO TO E22)
(03) Biogas (GO TO E22)
(04) Kerosene (GO TO E22)
(05) Coal/lignite
(06) Charcoal
(07) Wood
(08) Straw/Shrubs/Grass
(09) Agricultural crop waste
(10) Dung cakes
(11) Other, specify:____________
(99) Don`t know

E20 Did the stove have a chimney? .... □□

(00) No (01) Yes (99) Don't know

E21 Was the stove located in an area with any ventilation/windows?... .... □□

(00)No (01)Yes (99) Don`t know



Section E – Housing conditions & Residential environment 0 5 ID N° □□□

E22 Where was the cooking usually done?.. □□
(01) Inside the house (02) Separate building (03) Outdoors
(99) Don`t know

E23 Did your home have a separate room which was used as a kitchen?.................. □□
(00) No (01) Yes (99) Don't know

E24 Were you exposed to cigarette smoke in this house?... □□
(00) No (01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

I will now read a **list of household goods** you may have had in your childhood residence or not. You may find that some of these appliances were not applicable to the epoch you were a child. Choose the answer that best represents your situation, regardless.

E25 Did your place have a watch or clock?.. □□
(00) No (01) Yes (99) Don't know

E26 Did your place have a radio or transistor?... □□
(00) No (01) Yes (99) Don't know

E27 Did your place have a TV? .... □□
(00) No (02) Yes, color
(01) Yes, black and white (99) Don't know

E28 Did your place have a refrigerator?.. □□
(00) No, it had no appliance to cool food (01) Yes
(99) Don't know

Also, I would like to ask you...

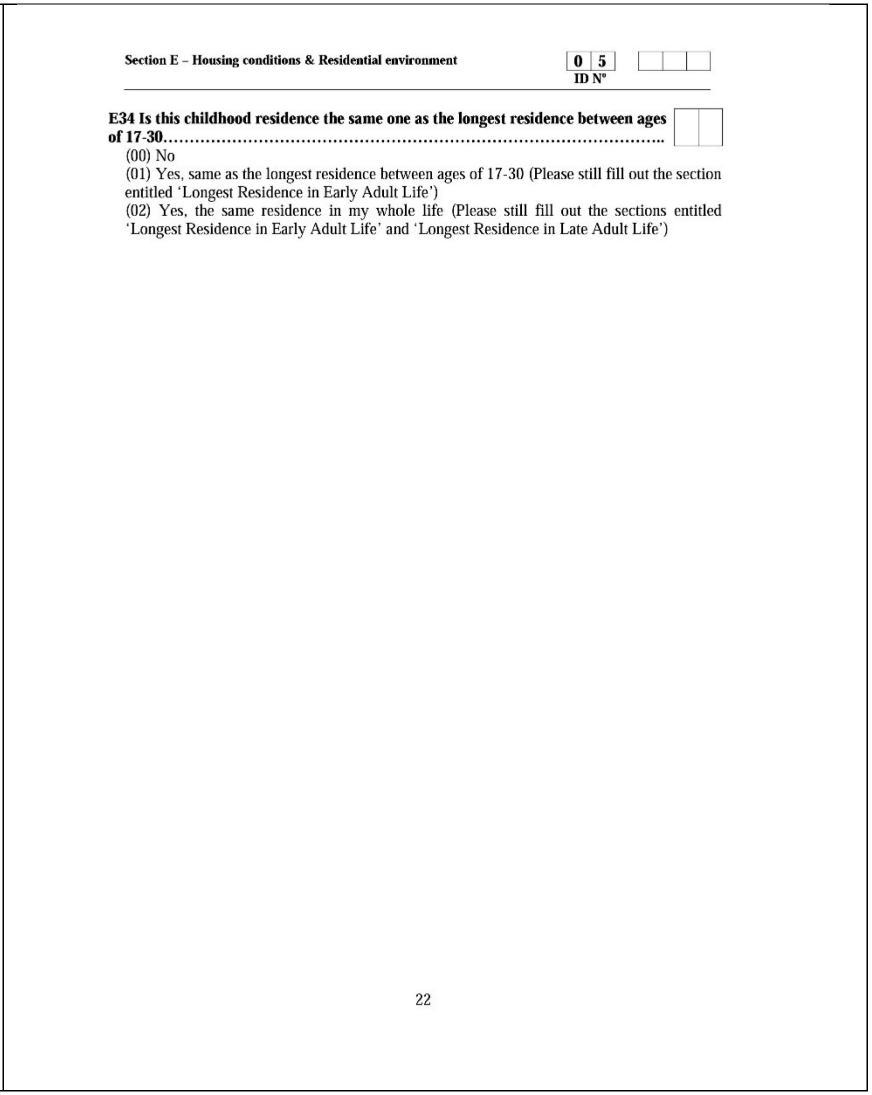
E29 Did your household have a bicycle?.. □□
(00) No (01) Yes (99) Don't know

E30 Did your household have a motorcycle or scooter?.. □□
(00) No (01) Yes (99) Don't know (GO TO E32)

E31 How many?... □□

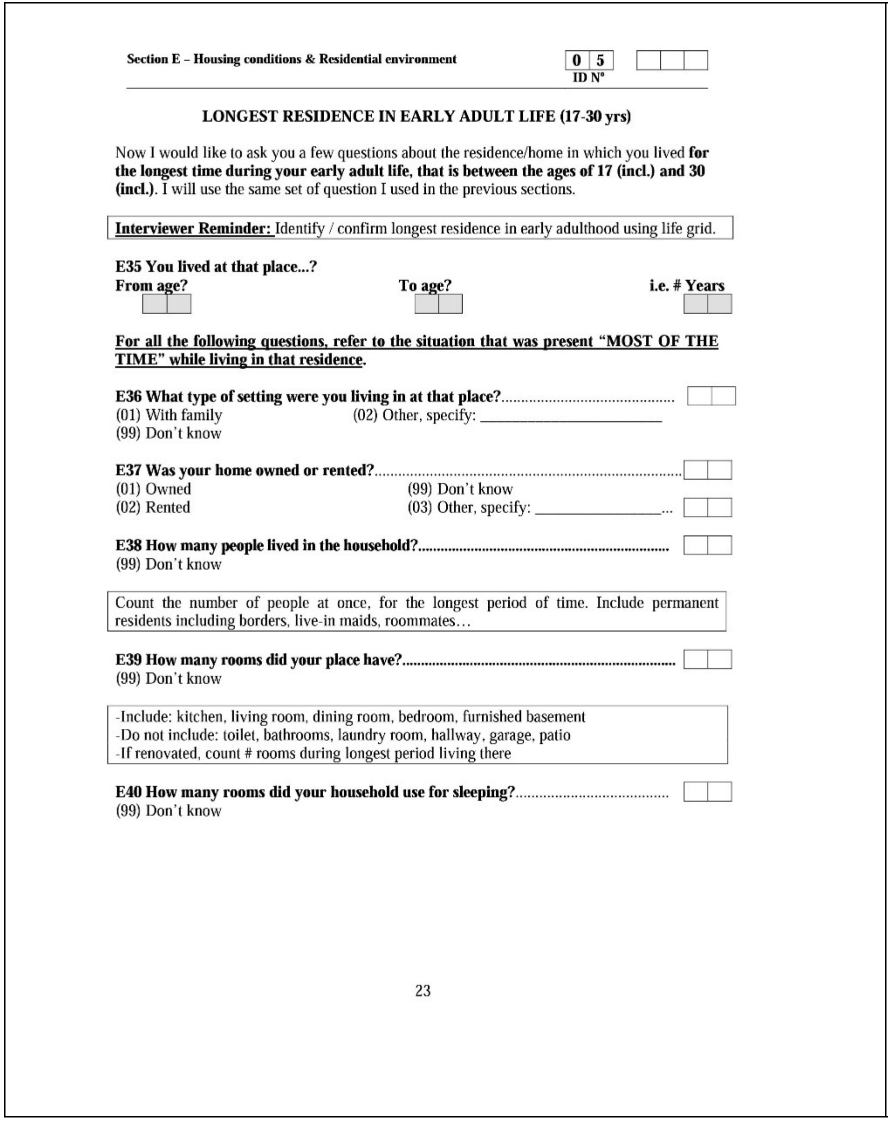
E32 Did your household have a car?.. □□
(00) No (01) Yes (99) Don't know (GO TO E34)

E33 How many?... □□



Section E – Housing conditions & Residential environment 0 5 ID N° □□□

E34 Is this childhood residence the same one as the longest residence between ages of 17-30………………………………………………………………………………….. □□
(00) No
(01) Yes, same as the longest residence between ages of 17-30 (Please still fill out the section entitled 'Longest Residence in Early Adult Life')
(02) Yes, the same residence in my whole life (Please still fill out the sections entitled 'Longest Residence in Early Adult Life' and 'Longest Residence in Late Adult Life')



Section E – Housing conditions & Residential environment 0 5 ID N° □□□

LONGEST RESIDENCE IN EARLY ADULT LIFE (17-30 yrs)

Now I would like to ask you a few questions about the residence/home in which you lived **for the longest time during your early adult life, that is between the ages of 17 (incl.) and 30 (incl.)**. I will use the same set of question I used in the previous sections.

Interviewer Reminder: Identify / confirm longest residence in early adulthood using life grid.

E35 You lived at that place...?

From age? □□ **To age?** □□ **i.e. # Years** □□

For all the following questions, refer to the situation that was present "MOST OF THE TIME" while living in that residence.

E36 What type of setting were you living in at that place?... □□

(01) With family (02) Other, specify: ______________________
(99) Don't know

E37 Was your home owned or rented?.. □□

(01) Owned (99) Don't know
(02) Rented (03) Other, specify: ________________... □□

E38 How many people lived in the household?.. □□

(99) Don't know

Count the number of people at once, for the longest period of time. Include permanent residents including borders, live-in maids, roommates…

E39 How many rooms did your place have?.. □□

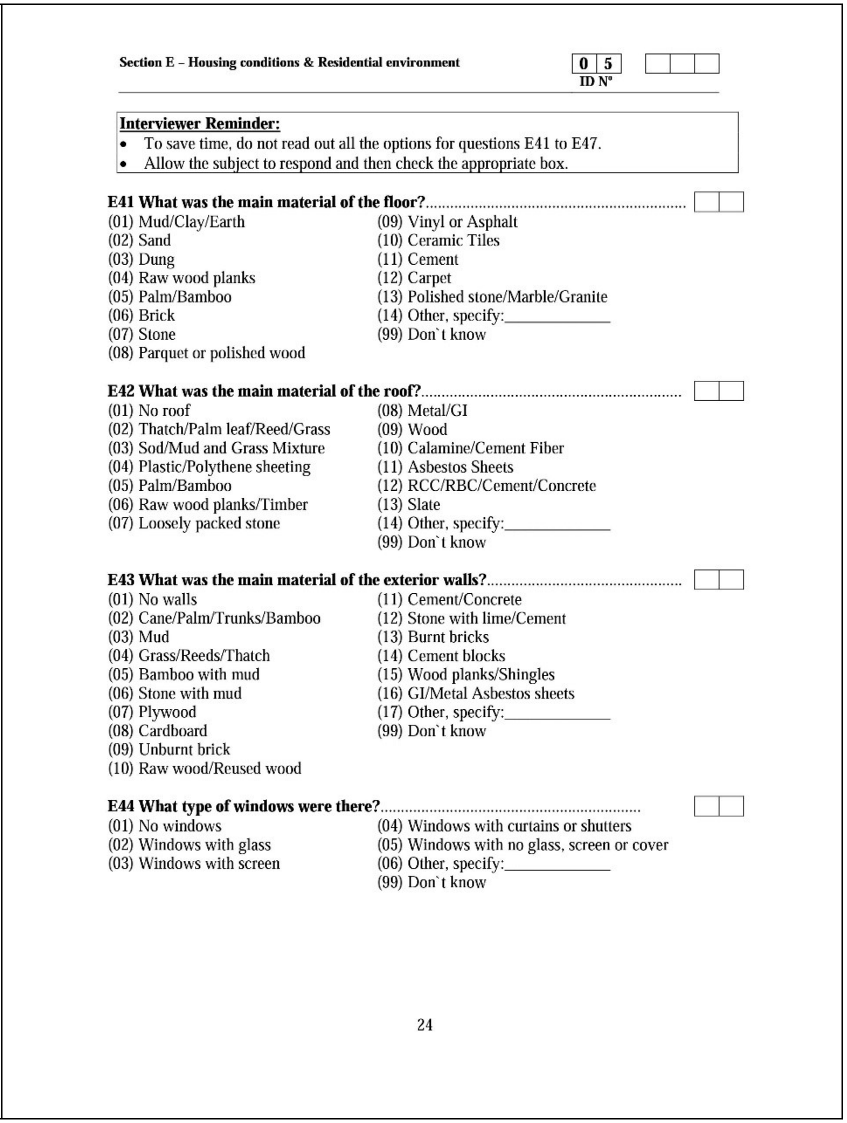
(99) Don't know

-Include: kitchen, living room, dining room, bedroom, furnished basement
-Do not include: toilet, bathrooms, laundry room, hallway, garage, patio
-If renovated, count # rooms during longest period living there

E40 How many rooms did your household use for sleeping?...................................... □□

(99) Don't know

23



Section E – Housing conditions & Residential environment 0 5 ID N° □□□

Interviewer Reminder:

- To save time, do not read out all the options for questions E41 to E47.
- Allow the subject to respond and then check the appropriate box.

E41 What was the main material of the floor?... □□

(01) Mud/Clay/Earth
(02) Sand
(03) Dung
(04) Raw wood planks
(05) Palm/Bamboo
(06) Brick
(07) Stone
(08) Parquet or polished wood
(09) Vinyl or Asphalt
(10) Ceramic Tiles
(11) Cement
(12) Carpet
(13) Polished stone/Marble/Granite
(14) Other, specify:____________
(99) Don`t know

E42 What was the main material of the roof?... □□

(01) No roof
(02) Thatch/Palm leaf/Reed/Grass
(03) Sod/Mud and Grass Mixture
(04) Plastic/Polythene sheeting
(05) Palm/Bamboo
(06) Raw wood planks/Timber
(07) Loosely packed stone
(08) Metal/GI
(09) Wood
(10) Calamine/Cement Fiber
(11) Asbestos Sheets
(12) RCC/RBC/Cement/Concrete
(13) Slate
(14) Other, specify:____________
(99) Don`t know

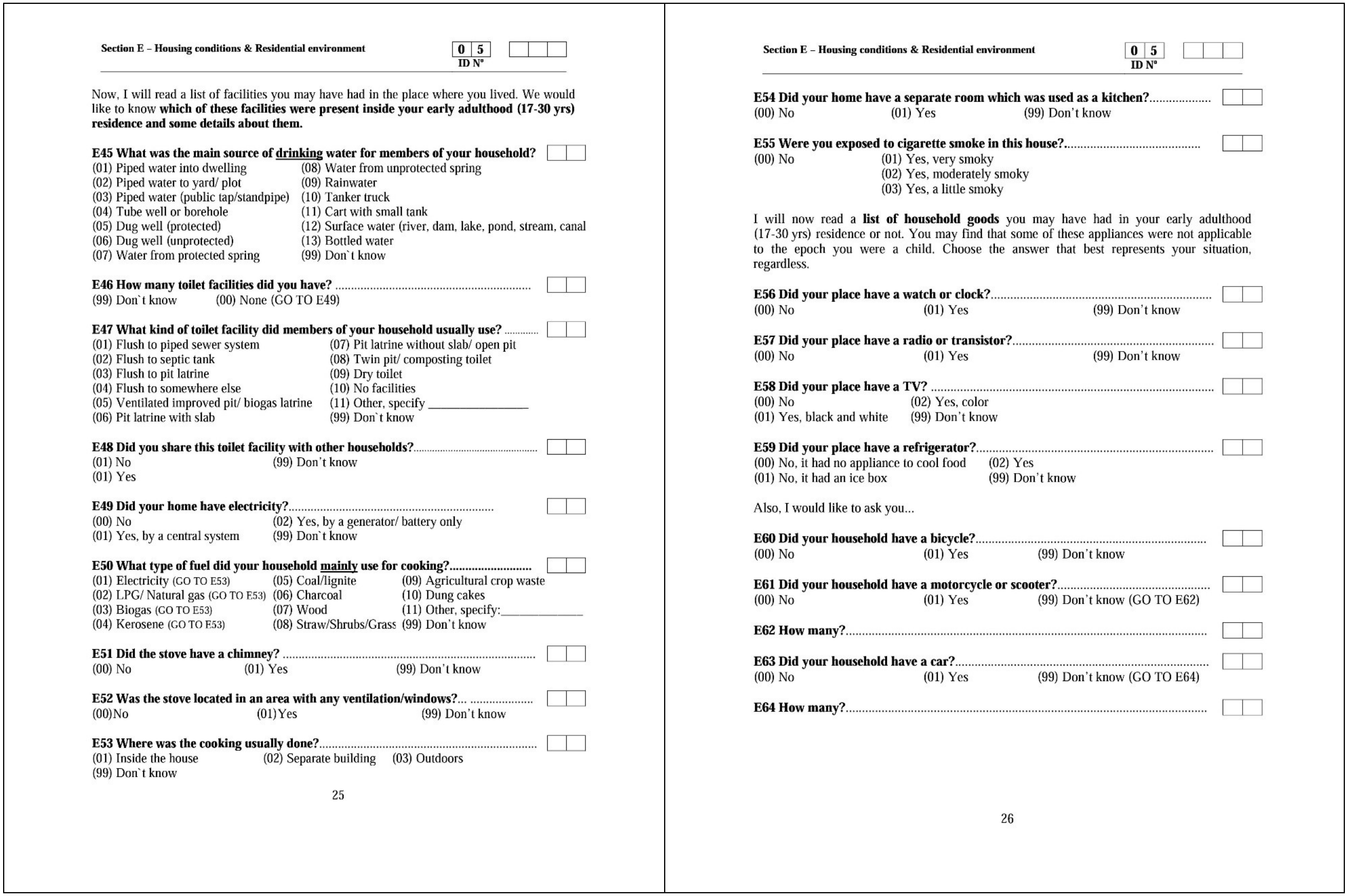
E43 What was the main material of the exterior walls?... □□

(01) No walls
(02) Cane/Palm/Trunks/Bamboo
(03) Mud
(04) Grass/Reeds/Thatch
(05) Bamboo with mud
(06) Stone with mud
(07) Plywood
(08) Cardboard
(09) Unburnt brick
(10) Raw wood/Reused wood
(11) Cement/Concrete
(12) Stone with lime/Cement
(13) Burnt bricks
(14) Cement blocks
(15) Wood planks/Shingles
(16) GI/Metal Asbestos sheets
(17) Other, specify:____________
(99) Don`t know

E44 What type of windows were there?... □□

(01) No windows
(02) Windows with glass
(03) Windows with screen
(04) Windows with curtains or shutters
(05) Windows with no glass, screen or cover
(06) Other, specify:____________
(99) Don`t know

24



Section E – Housing conditions & Residential environment 0 5 □□□ **ID N°**

Now, I will read a list of facilities you may have had in the place where you lived. We would like to know **which of these facilities were present inside your early adulthood (17-30 yrs) residence and some details about them.**

E45 What was the main source of <u>drinking</u> water for members of your household? □□

(01) Piped water into dwelling
(02) Piped water to yard/ plot
(03) Piped water (public tap/standpipe)
(04) Tube well or borehole
(05) Dug well (protected)
(06) Dug well (unprotected)
(07) Water from protected spring
(08) Water from unprotected spring
(09) Rainwater
(10) Tanker truck
(11) Cart with small tank
(12) Surface water (river, dam, lake, pond, stream, canal
(13) Bottled water
(99) Don`t know

E46 How many toilet facilities did you have? .... □□

(99) Don`t know (00) None (GO TO E49)

E47 What kind of toilet facility did members of your household usually use? .... □□

(01) Flush to piped sewer system
(02) Flush to septic tank
(03) Flush to pit latrine
(04) Flush to somewhere else
(05) Ventilated improved pit/ biogas latrine
(06) Pit latrine with slab
(07) Pit latrine without slab/ open pit
(08) Twin pit/ composting toilet
(09) Dry toilet
(10) No facilities
(11) Other, specify ________________
(99) Don`t know

E48 Did you share this toilet facility with other households?.. □□

(01) No (99) Don't know
(01) Yes

E49 Did your home have electricity?.. □□

(00) No (02) Yes, by a generator/ battery only
(01) Yes, by a central system (99) Don`t know

E50 What type of fuel did your household <u>mainly</u> use for cooking?......................... □□

(01) Electricity (GO TO E53)
(02) LPG/ Natural gas (GO TO E53)
(03) Biogas (GO TO E53)
(04) Kerosene (GO TO E53)
(05) Coal/lignite
(06) Charcoal
(07) Wood
(08) Straw/Shrubs/Grass
(09) Agricultural crop waste
(10) Dung cakes
(11) Other, specify:_____________
(99) Don't know

E51 Did the stove have a chimney? .... □□

(00) No (01) Yes (99) Don't know

E52 Was the stove located in an area with any ventilation/windows?... .... □□

(00)No (01)Yes (99) Don't know

E53 Where was the cooking usually done?... □□

(01) Inside the house (02) Separate building (03) Outdoors
(99) Don`t know

25

Section E – Housing conditions & Residential environment 0 5 □□□ **ID N°**

E54 Did your home have a separate room which was used as a kitchen?.................. □□

(00) No (01) Yes (99) Don't know

E55 Were you exposed to cigarette smoke in this house?........................................ □□

(00) No (01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

I will now read a **list of household goods** you may have had in your early adulthood (17-30 yrs) residence or not. You may find that some of these appliances were not applicable to the epoch you were a child. Choose the answer that best represents your situation, regardless.

E56 Did your place have a watch or clock?.. □□

(00) No (01) Yes (99) Don't know

E57 Did your place have a radio or transistor?... □□

(00) No (01) Yes (99) Don't know

E58 Did your place have a TV? .... □□

(00) No (02) Yes, color
(01) Yes, black and white (99) Don't know

E59 Did your place have a refrigerator?.. □□

(00) No, it had no appliance to cool food (02) Yes
(01) No, it had an ice box (99) Don't know

Also, I would like to ask you...

E60 Did your household have a bicycle?.. □□

(00) No (01) Yes (99) Don't know

E61 Did your household have a motorcycle or scooter?.. □□

(00) No (01) Yes (99) Don't know (GO TO E62)

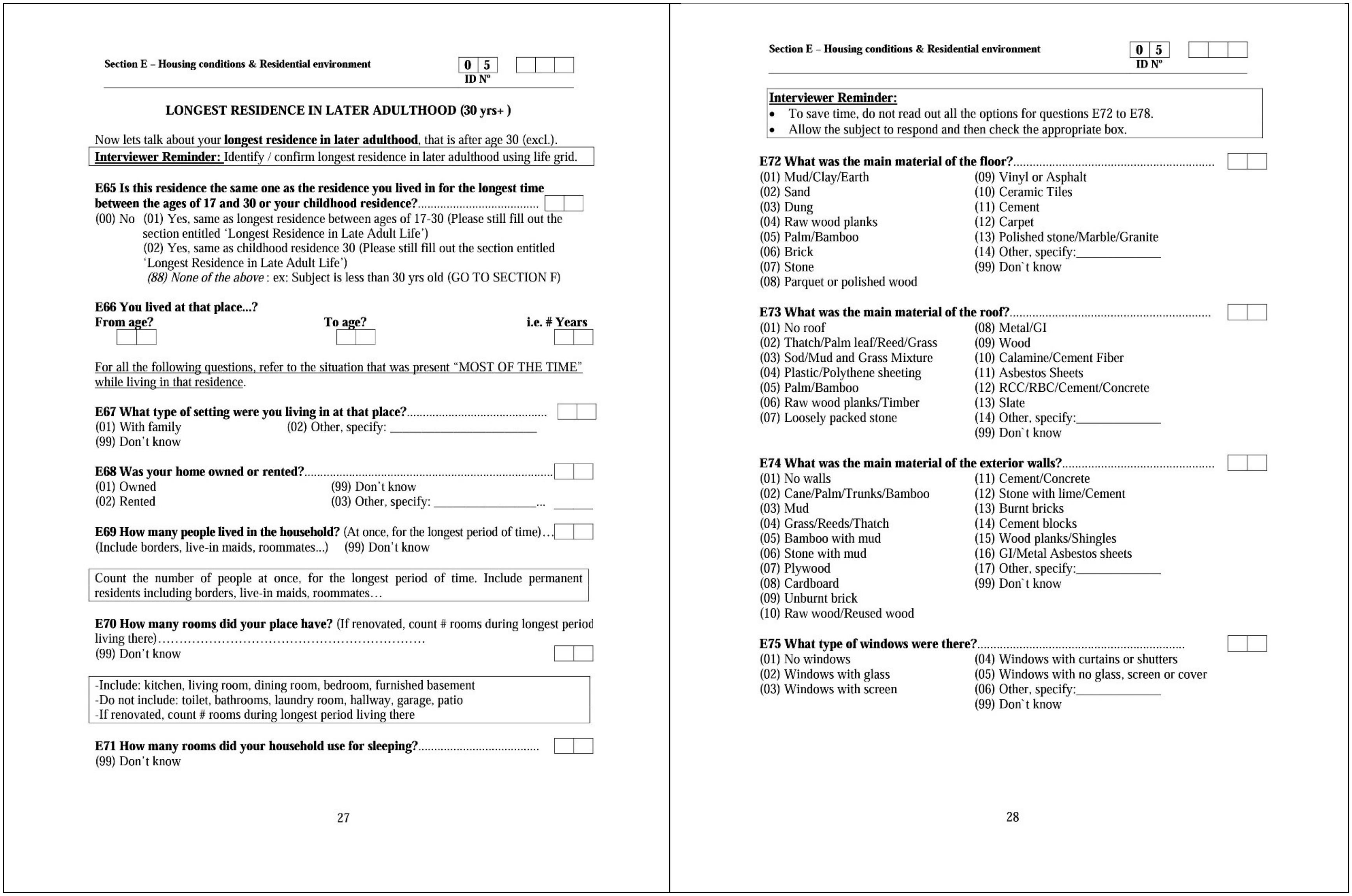
E62 How many?.. □□

E63 Did your household have a car?... □□

(00) No (01) Yes (99) Don't know (GO TO E64)

E64 How many?.. □□

26



Section E – Housing conditions & Residential environment | 0 | 5 | ID N° □□□

LONGEST RESIDENCE IN LATER ADULTHOOD (30 yrs+)

Now lets talk about your **longest residence in later adulthood**, that is after age 30 (excl.).

Interviewer Reminder: Identify / confirm longest residence in later adulthood using life grid.

E65 Is this residence the same one as the residence you lived in for the longest time between the ages of 17 and 30 or your childhood residence?...................................... □□

(00) No (01) Yes, same as longest residence between ages of 17-30 (Please still fill out the section entitled 'Longest Residence in Late Adult Life')

(02) Yes, same as childhood residence 30 (Please still fill out the section entitled 'Longest Residence in Late Adult Life')

(88) None of the above : ex: Subject is less than 30 yrs old (GO TO SECTION F)

E66 You lived at that place...?

From age? □□ **To age?** □□ **i.e. # Years** □□

For all the following questions, refer to the situation that was present "MOST OF THE TIME" while living in that residence.

E67 What type of setting were you living in at that place?... □□

(01) With family (02) Other, specify: _______________________

(99) Don't know

E68 Was your home owned or rented?.. □□

(01) Owned (99) Don't know

(02) Rented (03) Other, specify: _______________... ______

E69 How many people lived in the household? (At once, for the longest period of time)… □□

(Include borders, live-in maids, roommates...) (99) Don't know

Count the number of people at once, for the longest period of time. Include permanent residents including borders, live-in maids, roommates…

E70 How many rooms did your place have? (If renovated, count # rooms during longest period living there)……………………………………………………

(99) Don't know □□

-Include: kitchen, living room, dining room, bedroom, furnished basement
-Do not include: toilet, bathrooms, laundry room, hallway, garage, patio
-If renovated, count # rooms during longest period living there

E71 How many rooms did your household use for sleeping?..................................... □□

(99) Don't know

Section E – Housing conditions & Residential environment | 0 | 5 | ID N° □□□

Interviewer Reminder:
- To save time, do not read out all the options for questions E72 to E78.
- Allow the subject to respond and then check the appropriate box.

E72 What was the main material of the floor?.. □□

(01) Mud/Clay/Earth
(02) Sand
(03) Dung
(04) Raw wood planks
(05) Palm/Bamboo
(06) Brick
(07) Stone
(08) Parquet or polished wood
(09) Vinyl or Asphalt
(10) Ceramic Tiles
(11) Cement
(12) Carpet
(13) Polished stone/Marble/Granite
(14) Other, specify:____________
(99) Don`t know

E73 What was the main material of the roof?.. □□

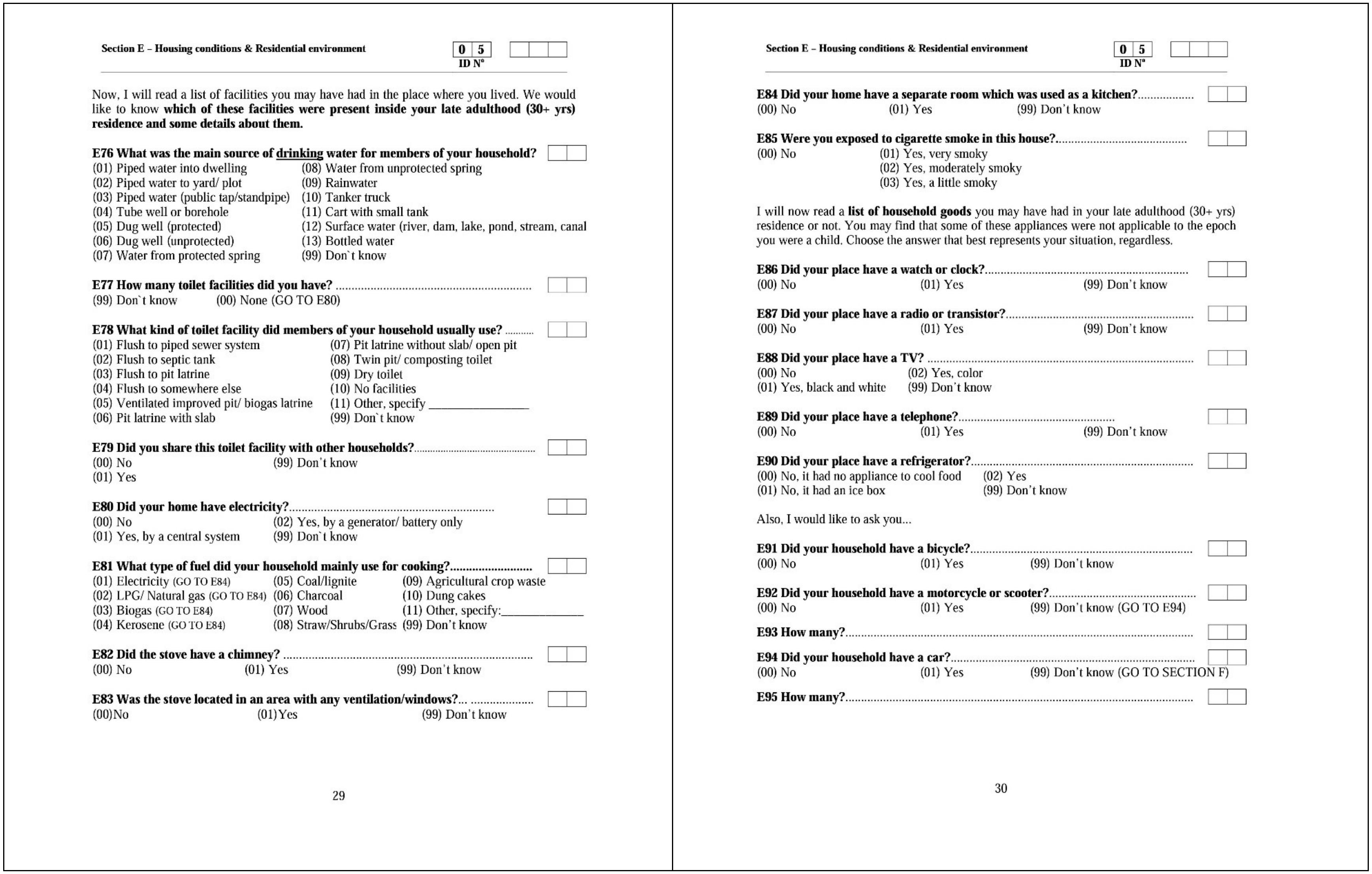
(01) No roof
(02) Thatch/Palm leaf/Reed/Grass
(03) Sod/Mud and Grass Mixture
(04) Plastic/Polythene sheeting
(05) Palm/Bamboo
(06) Raw wood planks/Timber
(07) Loosely packed stone
(08) Metal/GI
(09) Wood
(10) Calamine/Cement Fiber
(11) Asbestos Sheets
(12) RCC/RBC/Cement/Concrete
(13) Slate
(14) Other, specify:____________
(99) Don`t know

E74 What was the main material of the exterior walls?... □□

(01) No walls
(02) Cane/Palm/Trunks/Bamboo
(03) Mud
(04) Grass/Reeds/Thatch
(05) Bamboo with mud
(06) Stone with mud
(07) Plywood
(08) Cardboard
(09) Unburnt brick
(10) Raw wood/Reused wood
(11) Cement/Concrete
(12) Stone with lime/Cement
(13) Burnt bricks
(14) Cement blocks
(15) Wood planks/Shingles
(16) GI/Metal Asbestos sheets
(17) Other, specify:____________
(99) Don`t know

E75 What type of windows were there?.. □□

(01) No windows
(02) Windows with glass
(03) Windows with screen
(04) Windows with curtains or shutters
(05) Windows with no glass, screen or cover
(06) Other, specify:____________
(99) Don`t know



Section E – Housing conditions & Residential environment 0 5 □□□ **ID N°**

Now, I will read a list of facilities you may have had in the place where you lived. We would like to know **which of these facilities were present inside your late adulthood (30+ yrs) residence and some details about them.**

E76 What was the main source of <u>drinking</u> water for members of your household? □□

(01) Piped water into dwelling
(02) Piped water to yard/ plot
(03) Piped water (public tap/standpipe)
(04) Tube well or borehole
(05) Dug well (protected)
(06) Dug well (unprotected)
(07) Water from protected spring
(08) Water from unprotected spring
(09) Rainwater
(10) Tanker truck
(11) Cart with small tank
(12) Surface water (river, dam, lake, pond, stream, canal
(13) Bottled water
(99) Don`t know

E77 How many toilet facilities did you have? .... □□
(99) Don`t know (00) None (GO TO E80)

E78 What kind of toilet facility did members of your household usually use? .... □□

(01) Flush to piped sewer system
(02) Flush to septic tank
(03) Flush to pit latrine
(04) Flush to somewhere else
(05) Ventilated improved pit/ biogas latrine
(06) Pit latrine with slab
(07) Pit latrine without slab/ open pit
(08) Twin pit/ composting toilet
(09) Dry toilet
(10) No facilities
(11) Other, specify _______________
(99) Don`t know

E79 Did you share this toilet facility with other households?.. □□
(00) No
(01) Yes
(99) Don't know

E80 Did your home have electricity?... □□
(00) No
(01) Yes, by a central system
(02) Yes, by a generator/ battery only
(99) Don`t know

E81 What type of fuel did your household mainly use for cooking?......................... □□

(01) Electricity (GO TO E84)
(02) LPG/ Natural gas (GO TO E84)
(03) Biogas (GO TO E84)
(04) Kerosene (GO TO E84)
(05) Coal/lignite
(06) Charcoal
(07) Wood
(08) Straw/Shrubs/Grass
(09) Agricultural crop waste
(10) Dung cakes
(11) Other, specify:_____________
(99) Don't know

E82 Did the stove have a chimney? .... □□
(00) No (01) Yes (99) Don't know

E83 Was the stove located in an area with any ventilation/windows?... .... □□
(00)No (01)Yes (99) Don't know

29

Section E – Housing conditions & Residential environment 0 5 □□□ **ID N°**

E84 Did your home have a separate room which was used as a kitchen?................. □□
(00) No (01) Yes (99) Don't know

E85 Were you exposed to cigarette smoke in this house?... □□
(00) No
(01) Yes, very smoky
(02) Yes, moderately smoky
(03) Yes, a little smoky

I will now read a **list of household goods** you may have had in your late adulthood (30+ yrs) residence or not. You may find that some of these appliances were not applicable to the epoch you were a child. Choose the answer that best represents your situation, regardless.

E86 Did your place have a watch or clock?... □□
(00) No (01) Yes (99) Don't know

E87 Did your place have a radio or transistor?... □□
(00) No (01) Yes (99) Don't know

E88 Did your place have a TV? .... □□
(00) No
(01) Yes, black and white
(02) Yes, color
(99) Don`t know

E89 Did your place have a telephone?... □□
(00) No (01) Yes (99) Don't know

E90 Did your place have a refrigerator?.. □□
(00) No, it had no appliance to cool food
(01) No, it had an ice box
(02) Yes
(99) Don't know

Also, I would like to ask you...

E91 Did your household have a bicycle?.. □□
(00) No (01) Yes (99) Don't know

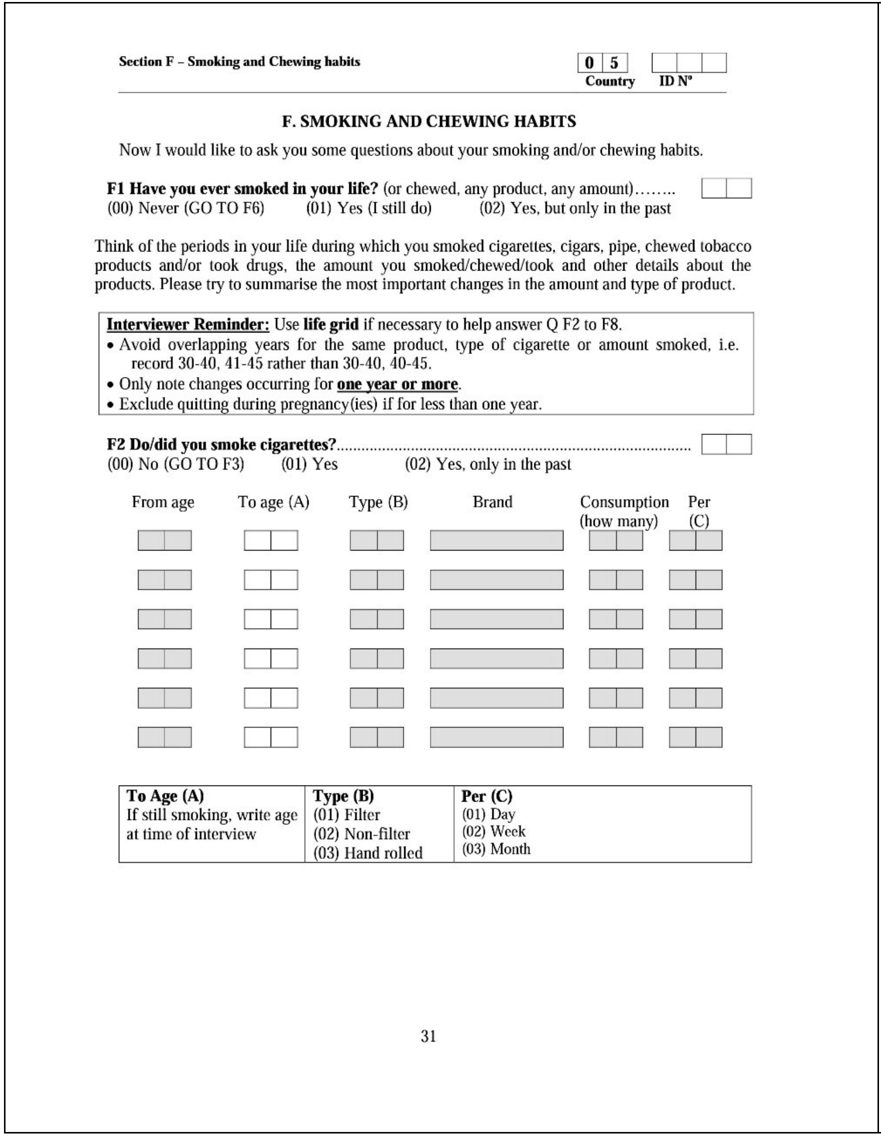
E92 Did your household have a motorcycle or scooter?.. □□
(00) No (01) Yes (99) Don't know (GO TO E94)

E93 How many?.. □□

E94 Did your household have a car?.. □□
(00) No (01) Yes (99) Don't know (GO TO SECTION F)

E95 How many?.. □□

30



Section F – Smoking and Chewing habits

0 5	□□□
Country	ID N°

F. SMOKING AND CHEWING HABITS

Now I would like to ask you some questions about your smoking and/or chewing habits.

F1 Have you ever smoked in your life? (or chewed, any product, any amount)…….. □□

(00) Never (GO TO F6) (01) Yes (I still do) (02) Yes, but only in the past

Think of the periods in your life during which you smoked cigarettes, cigars, pipe, chewed tobacco products and/or took drugs, the amount you smoked/chewed/took and other details about the products. Please try to summarise the most important changes in the amount and type of product.

Interviewer Reminder: Use **life grid** if necessary to help answer Q F2 to F8.
- Avoid overlapping years for the same product, type of cigarette or amount smoked, i.e. record 30-40, 41-45 rather than 30-40, 40-45.
- Only note changes occurring for **one year or more**.
- Exclude quitting during pregnancy(ies) if for less than one year.

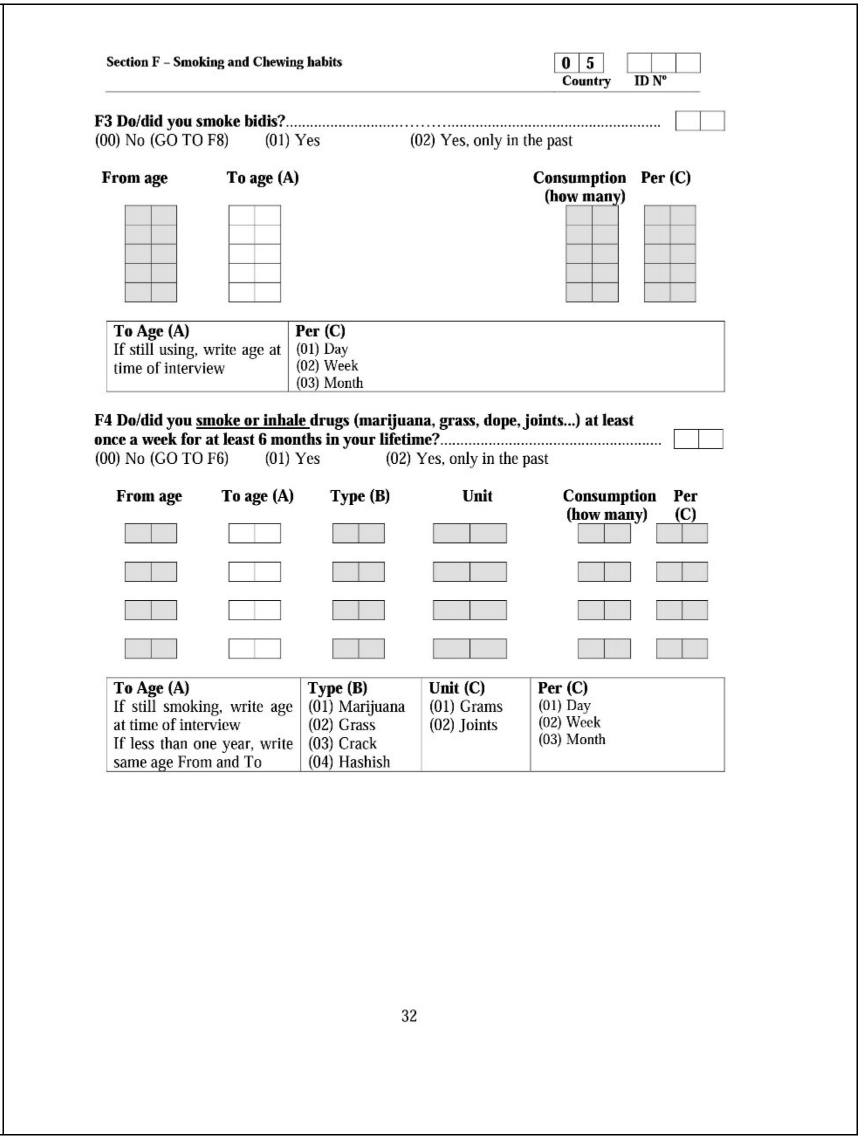
F2 Do/did you smoke cigarettes?.. □□

(00) No (GO TO F3) (01) Yes (02) Yes, only in the past

From age	To age (A)	Type (B)	Brand	Consumption (how many)	Per (C)
□□	□□	□□		□□	□□
□□	□□	□□		□□	□□
□□	□□	□□		□□	□□
□□	□□	□□		□□	□□
□□	□□	□□		□□	□□
□□	□□	□□		□□	□□

To Age (A)	**Type (B)**	**Per (C)**
If still smoking, write age at time of interview	(01) Filter (02) Non-filter (03) Hand rolled	(01) Day (02) Week (03) Month

31



Section F – Smoking and Chewing habits

0 5	□□□
Country	ID N°

F3 Do/did you smoke bidis?..............................……….. □□

(00) No (GO TO F8) (01) Yes (02) Yes, only in the past

From age	To age (A)	Consumption (how many)	Per (C)
□□	□□	□□	□□
□□	□□	□□	□□
□□	□□	□□	□□
□□	□□	□□	□□
□□	□□	□□	□□

To Age (A)	**Per (C)**
If still using, write age at time of interview	(01) Day (02) Week (03) Month

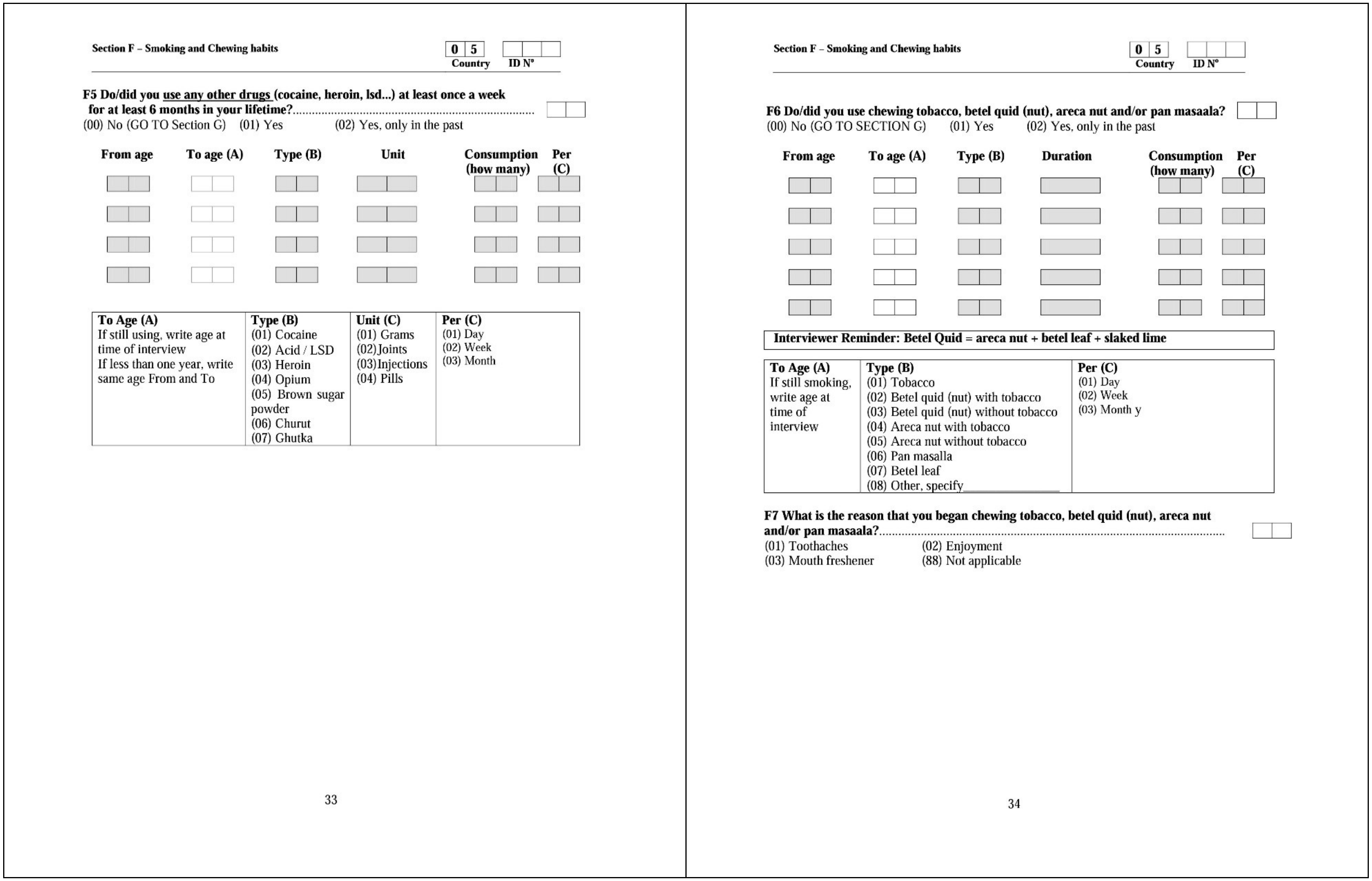
F4 Do/did you smoke or inhale drugs (marijuana, grass, dope, joints...) at least once a week for at least 6 months in your lifetime?.. □□

(00) No (GO TO F6) (01) Yes (02) Yes, only in the past

From age	To age (A)	Type (B)	Unit	Consumption (how many)	Per (C)
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□

To Age (A)	**Type (B)**	**Unit (C)**	**Per (C)**
If still smoking, write age at time of interview If less than one year, write same age From and To	(01) Marijuana (02) Grass (03) Crack (04) Hashish	(01) Grams (02) Joints	(01) Day (02) Week (03) Month

32



Section F – Smoking and Chewing habits 0 5 Country □□□ ID N°

F5 Do/did you use any other drugs (cocaine, heroin, lsd...) at least once a week for at least 6 months in your lifetime?.. □□

(00) No (GO TO Section G) (01) Yes (02) Yes, only in the past

From age	To age (A)	Type (B)	Unit	Consumption (how many)	Per (C)
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□

To Age (A)	Type (B)	Unit (C)	Per (C)
If still using, write age at time of interview If less than one year, write same age From and To	(01) Cocaine (02) Acid / LSD (03) Heroin (04) Opium (05) Brown sugar powder (06) Churut (07) Ghutka	(01) Grams (02)Joints (03)Injections (04) Pills	(01) Day (02) Week (03) Month

33

Section F – Smoking and Chewing habits 0 5 Country □□□ ID N°

F6 Do/did you use chewing tobacco, betel quid (nut), areca nut and/or pan masaala? □□

(00) No (GO TO SECTION G) (01) Yes (02) Yes, only in the past

From age	To age (A)	Type (B)	Duration	Consumption (how many)	Per (C)
□□	□□	□□	□	□□	□□
□□	□□	□□	□	□□	□□
□□	□□	□□	□	□□	□□
□□	□□	□□	□	□□	□□
□□	□□	□□	□	□□	□□

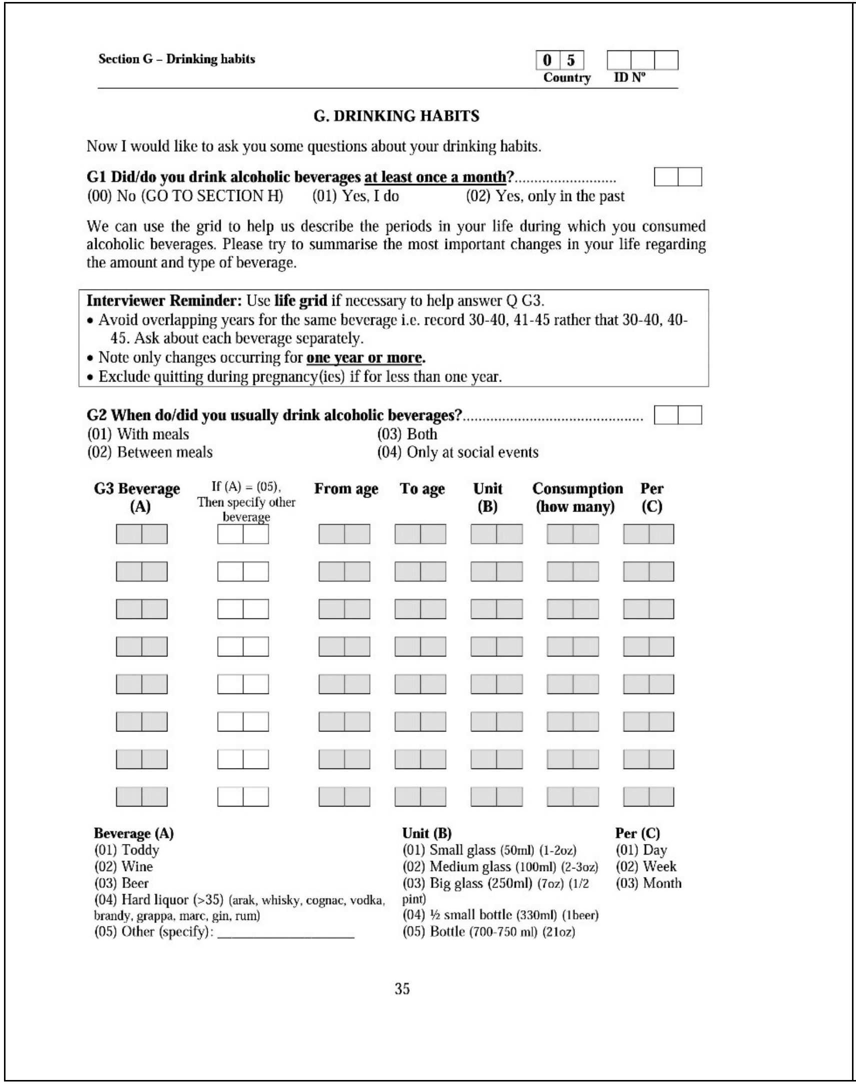
Interviewer Reminder: Betel Quid = areca nut + betel leaf + slaked lime

To Age (A)	Type (B)	Per (C)
If still smoking, write age at time of interview	(01) Tobacco (02) Betel quid (nut) with tobacco (03) Betel quid (nut) without tobacco (04) Areca nut with tobacco (05) Areca nut without tobacco (06) Pan masalla (07) Betel leaf (08) Other, specify_______________	(01) Day (02) Week (03) Month y

F7 What is the reason that you began chewing tobacco, betel quid (nut), areca nut and/or pan masaala?.. □□

(01) Toothaches (02) Enjoyment
(03) Mouth freshener (88) Not applicable

34



Section G – Drinking habits 0 5 □□□ Country ID N°

G. DRINKING HABITS

Now I would like to ask you some questions about your drinking habits.

G1 Did/do you drink alcoholic beverages <u>at least once a month</u>?.......................... □□
(00) No (GO TO SECTION H) (01) Yes, I do (02) Yes, only in the past

We can use the grid to help us describe the periods in your life during which you consumed alcoholic beverages. Please try to summarise the most important changes in your life regarding the amount and type of beverage.

> **Interviewer Reminder:** Use **life grid** if necessary to help answer Q G3.
> - Avoid overlapping years for the same beverage i.e. record 30-40, 41-45 rather that 30-40, 40-45. Ask about each beverage separately.
> - Note only changes occurring for **<u>one year or more</u>.**
> - Exclude quitting during pregnancy(ies) if for less than one year.

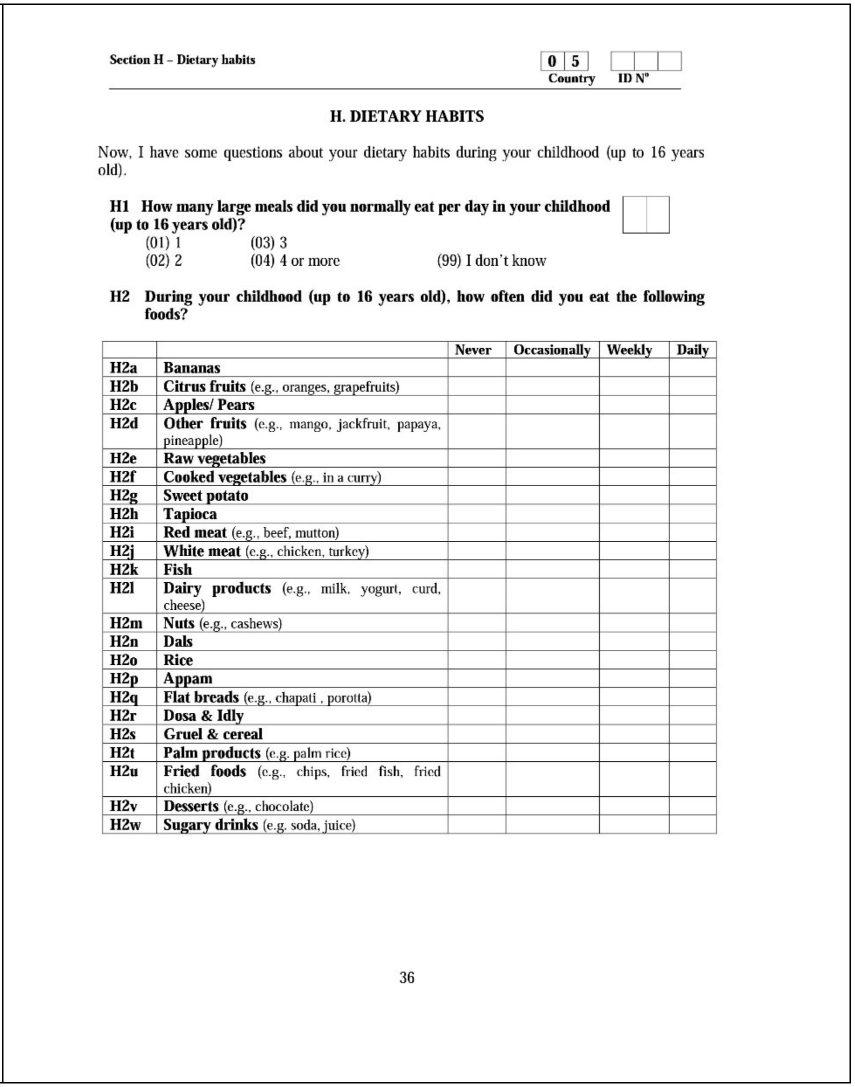
G2 When do/did you usually drink alcoholic beverages?... □□
(01) With meals (03) Both
(02) Between meals (04) Only at social events

G3 Beverage (A)	If (A) = (05), Then specify other beverage	**From age**	**To age**	**Unit (B)**	**Consumption (how many)**	**Per (C)**
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□
□□	□□	□□	□□	□□	□□	□□

Beverage (A)
(01) Toddy
(02) Wine
(03) Beer
(04) Hard liquor (>35) (arak, whisky, cognac, vodka, brandy, grappa, marc, gin, rum)
(05) Other (specify): ____________________

Unit (B)
(01) Small glass (50ml) (1-2oz)
(02) Medium glass (100ml) (2-3oz)
(03) Big glass (250ml) (7oz) (1/2 pint)
(04) ½ small bottle (330ml) (1beer)
(05) Bottle (700-750 ml) (21oz)

Per (C)
(01) Day
(02) Week
(03) Month



Section H – Dietary habits 0 5 □□□ Country ID N°

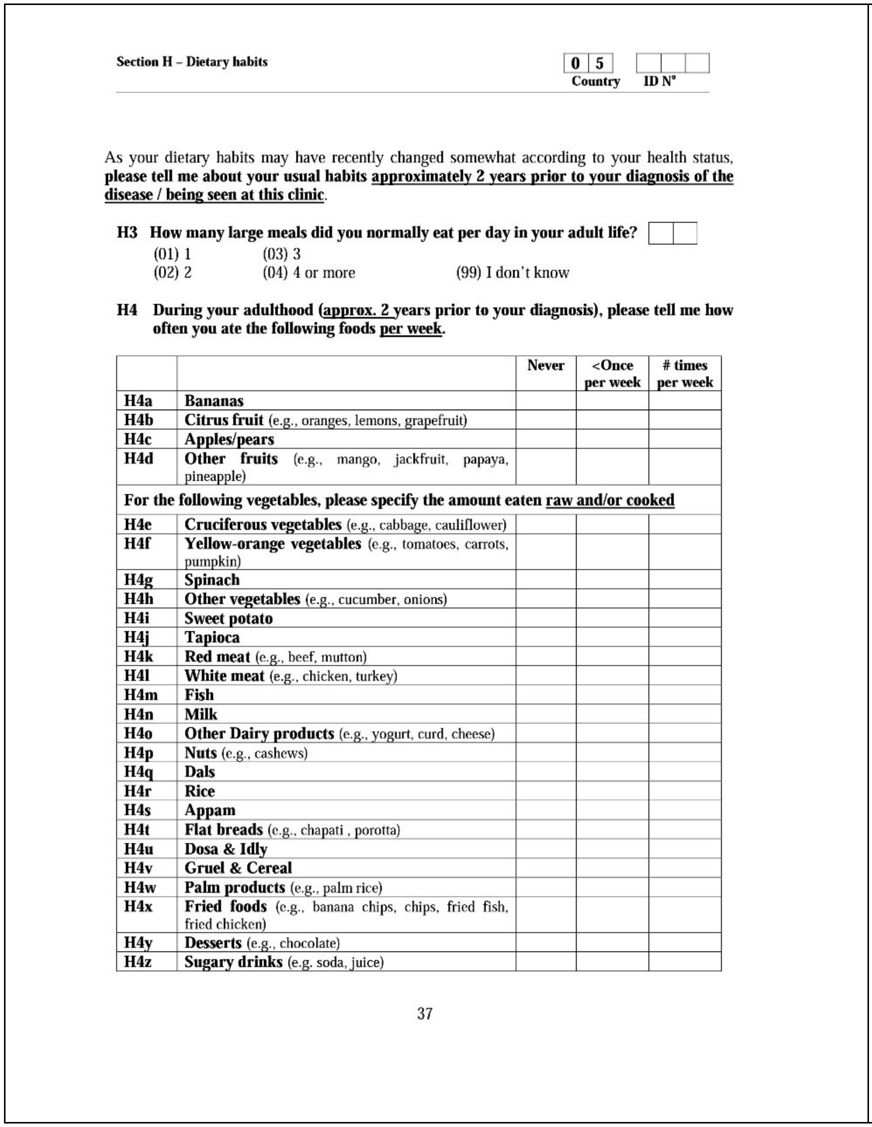
H. DIETARY HABITS

Now, I have some questions about your dietary habits during your childhood (up to 16 years old).

H1 How many large meals did you normally eat per day in your childhood (up to 16 years old)? □□
(01) 1 (03) 3
(02) 2 (04) 4 or more (99) I don't know

H2 During your childhood (up to 16 years old), how often did you eat the following foods?

		Never	**Occasionally**	**Weekly**	**Daily**
H2a	**Bananas**				
H2b	**Citrus fruits** (e.g., oranges, grapefruits)				
H2c	**Apples/ Pears**				
H2d	**Other fruits** (e.g., mango, jackfruit, papaya, pineapple)				
H2e	**Raw vegetables**				
H2f	**Cooked vegetables** (e.g., in a curry)				
H2g	**Sweet potato**				
H2h	**Tapioca**				
H2i	**Red meat** (e.g., beef, mutton)				
H2j	**White meat** (e.g., chicken, turkey)				
H2k	**Fish**				
H2l	**Dairy products** (e.g., milk, yogurt, curd, cheese)				
H2m	**Nuts** (e.g., cashews)				
H2n	**Dals**				
H2o	**Rice**				
H2p	**Appam**				
H2q	**Flat breads** (e.g., chapati , porotta)				
H2r	**Dosa & Idly**				
H2s	**Gruel & cereal**				
H2t	**Palm products** (e.g. palm rice)				
H2u	**Fried foods** (e.g., chips, fried fish, fried chicken)				
H2v	**Desserts** (e.g., chocolate)				
H2w	**Sugary drinks** (e.g. soda, juice)				



Section H – Dietary habits

0 5 Country □□□ ID N°

As your dietary habits may have recently changed somewhat according to your health status, **please tell me about your usual habits <u>approximately 2 years prior to your diagnosis of the disease / being seen at this clinic</u>**.

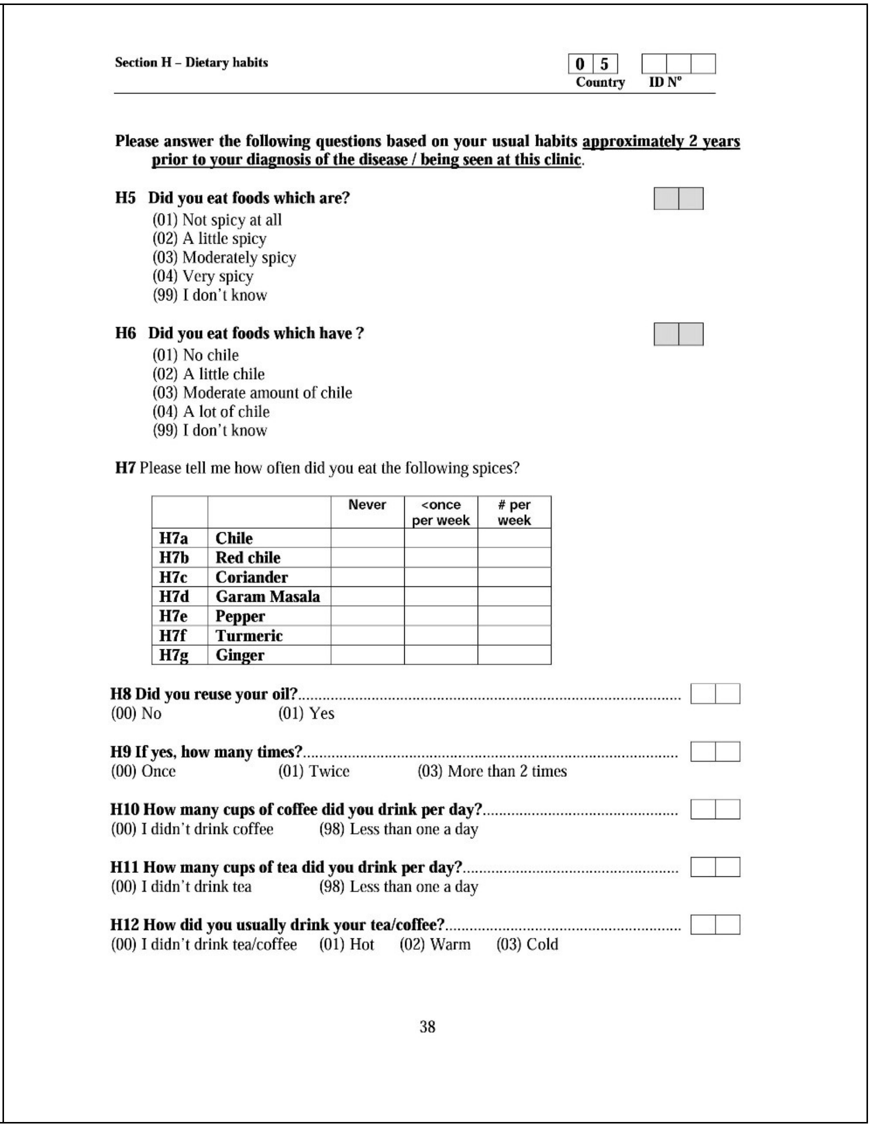
H3 How many large meals did you normally eat per day in your adult life? □□

(01) 1 (03) 3
(02) 2 (04) 4 or more (99) I don't know

H4 During your adulthood (<u>approx. 2 years</u> prior to your diagnosis), please tell me how often you ate the following foods <u>per week</u>.

		Never	<Once per week	# times per week
H4a	**Bananas**			
H4b	**Citrus fruit** (e.g., oranges, lemons, grapefruit)			
H4c	**Apples/pears**			
H4d	**Other fruits** (e.g., mango, jackfruit, papaya, pineapple)			
For the following vegetables, please specify the amount eaten <u>raw and/or cooked</u>				
H4e	**Cruciferous vegetables** (e.g., cabbage, cauliflower)			
H4f	**Yellow-orange vegetables** (e.g., tomatoes, carrots, pumpkin)			
H4g	**Spinach**			
H4h	**Other vegetables** (e.g., cucumber, onions)			
H4i	**Sweet potato**			
H4j	**Tapioca**			
H4k	**Red meat** (e.g., beef, mutton)			
H4l	**White meat** (e.g., chicken, turkey)			
H4m	**Fish**			
H4n	**Milk**			
H4o	**Other Dairy products** (e.g., yogurt, curd, cheese)			
H4p	**Nuts** (e.g., cashews)			
H4q	**Dals**			
H4r	**Rice**			
H4s	**Appam**			
H4t	**Flat breads** (e.g., chapati , porotta)			
H4u	**Dosa & Idly**			
H4v	**Gruel & Cereal**			
H4w	**Palm products** (e.g., palm rice)			
H4x	**Fried foods** (e.g., banana chips, chips, fried fish, fried chicken)			
H4y	**Desserts** (e.g., chocolate)			
H4z	**Sugary drinks** (e.g. soda, juice)			

37



Section H – Dietary habits

0 5 Country □□□ ID N°

Please answer the following questions based on your usual habits <u>approximately 2 years prior to your diagnosis of the disease / being seen at this clinic</u>.

H5 Did you eat foods which are? □□

(01) Not spicy at all
(02) A little spicy
(03) Moderately spicy
(04) Very spicy
(99) I don't know

H6 Did you eat foods which have ? □□

(01) No chile
(02) A little chile
(03) Moderate amount of chile
(04) A lot of chile
(99) I don't know

H7 Please tell me how often did you eat the following spices?

		Never	<once per week	# per week
H7a	**Chile**			
H7b	**Red chile**			
H7c	**Coriander**			
H7d	**Garam Masala**			
H7e	**Pepper**			
H7f	**Turmeric**			
H7g	**Ginger**			

H8 Did you reuse your oil?.. □□
(00) No (01) Yes

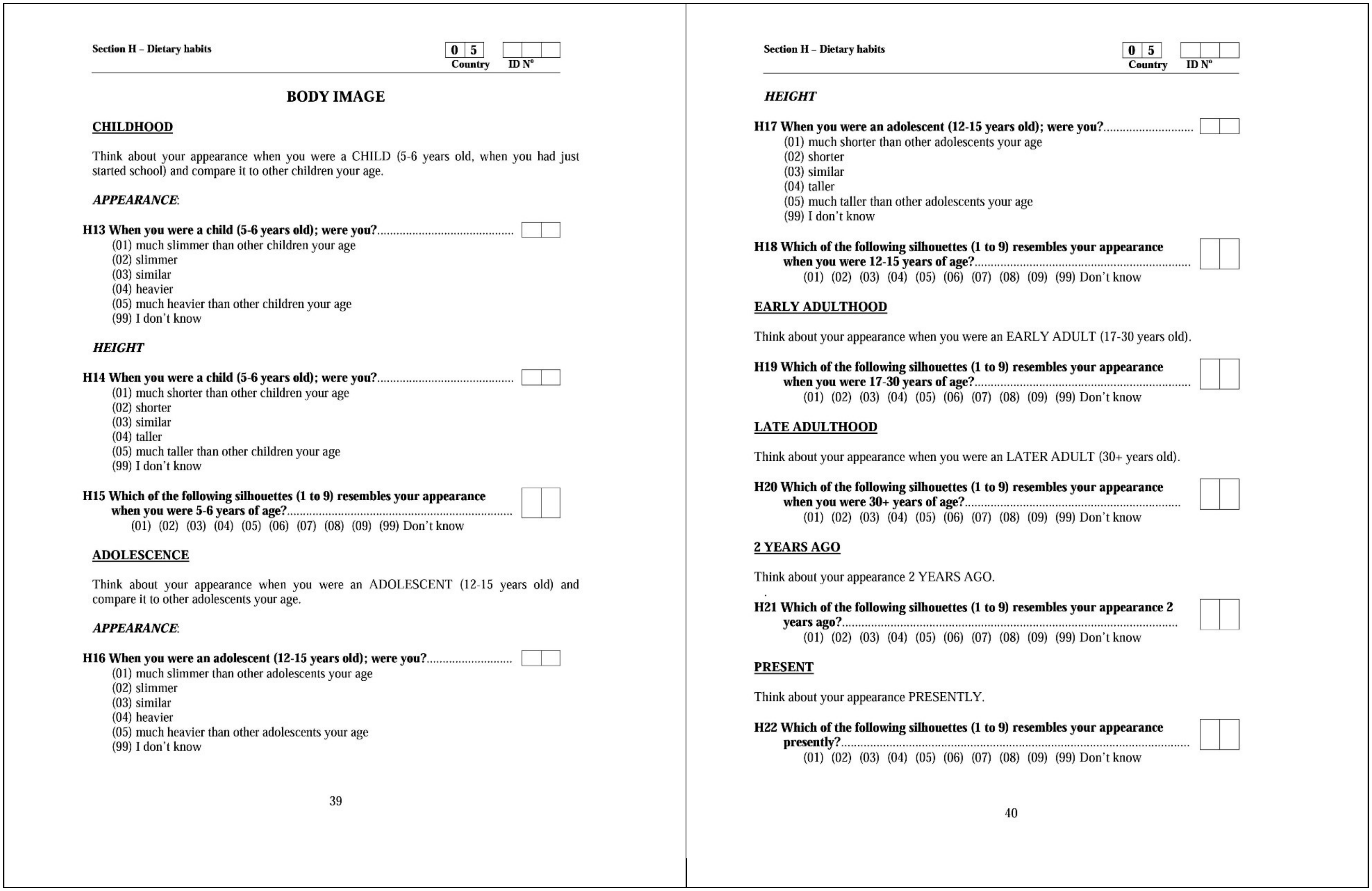
H9 If yes, how many times?.. □□
(00) Once (01) Twice (03) More than 2 times

H10 How many cups of coffee did you drink per day?... □□
(00) I didn't drink coffee (98) Less than one a day

H11 How many cups of tea did you drink per day?... □□
(00) I didn't drink tea (98) Less than one a day

H12 How did you usually drink your tea/coffee?... □□
(00) I didn't drink tea/coffee (01) Hot (02) Warm (03) Cold

38



Section H – Dietary habits

0	5		□	□	□
Country			ID N°		

BODY IMAGE

CHILDHOOD

Think about your appearance when you were a CHILD (5-6 years old, when you had just started school) and compare it to other children your age.

APPEARANCE:

H13 When you were a child (5-6 years old); were you?... □□
(01) much slimmer than other children your age
(02) slimmer
(03) similar
(04) heavier
(05) much heavier than other children your age
(99) I don't know

HEIGHT

H14 When you were a child (5-6 years old); were you?... □□
(01) much shorter than other children your age
(02) shorter
(03) similar
(04) taller
(05) much taller than other children your age
(99) I don't know

H15 Which of the following silhouettes (1 to 9) resembles your appearance when you were 5-6 years of age?.. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

ADOLESCENCE

Think about your appearance when you were an ADOLESCENT (12-15 years old) and compare it to other adolescents your age.

APPEARANCE:

H16 When you were an adolescent (12-15 years old); were you?........................... □□
(01) much slimmer than other adolescents your age
(02) slimmer
(03) similar
(04) heavier
(05) much heavier than other adolescents your age
(99) I don't know

39

Section H – Dietary habits

0	5		□	□	□
Country			ID N°		

HEIGHT

H17 When you were an adolescent (12-15 years old); were you?............................ □□
(01) much shorter than other adolescents your age
(02) shorter
(03) similar
(04) taller
(05) much taller than other adolescents your age
(99) I don't know

H18 Which of the following silhouettes (1 to 9) resembles your appearance when you were 12-15 years of age?.. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

EARLY ADULTHOOD

Think about your appearance when you were an EARLY ADULT (17-30 years old).

H19 Which of the following silhouettes (1 to 9) resembles your appearance when you were 17-30 years of age?.. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

LATE ADULTHOOD

Think about your appearance when you were an LATER ADULT (30+ years old).

H20 Which of the following silhouettes (1 to 9) resembles your appearance when you were 30+ years of age?.. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

2 YEARS AGO

Think about your appearance 2 YEARS AGO.

H21 Which of the following silhouettes (1 to 9) resembles your appearance 2 years ago?... □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

PRESENT

Think about your appearance PRESENTLY.

H22 Which of the following silhouettes (1 to 9) resembles your appearance presently?.. □□
(01) (02) (03) (04) (05) (06) (07) (08) (09) (99) Don't know

40

Section H – Dietary habits

0 5 □□□
Country ID Nº

INTERVIEWER REMINDER: Weight measurement
Weigh the participant using the HeNCe Life study scale provided. Measurement should be done in **kgs**, but if it is done in **lbs**, convert the measurement to **kgs** as specified below.

H23 Weight measurement? □□□
(__________ lbs) ÷ 2.2042 = ________**kgs**

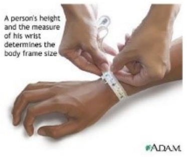
INTERVIEWER REMINDER: Height measurement
Measure the participant using the measuring tape provided in the HeNCe Life study package. The participant must be positioned with their
-feet together and flat on the ground
-heels touching the wall
-legs straight
-buttocks against the wall
-arms loosely at their side
-ensure that their feet and heels do not raise up off the ground
Measurement should be done in **cms**, but if it is done in **inches**, convert the measurement to **cms** as specified below. You may find it easier to ask someone to help you take the measurement.

H24 Height measurement? .... □□□
(___feet ___ inches) = _____inches x 2.54 = _______**cm**

INTERVIEWER REMINDER: Finger measurements
Please ask the participant to lay their **right hand** on the table palm-up, with the fingers fully extended. Measure the lengths of both the index finger (2nd finger) and the ring finger (4th finger). ***Note: the thumb is considered the 1st finger.*** The measurement should be taken from the tip of the finger to the lowest (most proximal) crease using the ruler provided in the HeNCe Life package. The index finger usually has only one proximal crease, whereas the ring finger sometimes has two.

H25 Finger measurements (right hand)? .... □□□
Index (2^{nd} finger):______ cm (1 decimal)
Ring finger (4^{th} finger): _____ cm (1 decimal)

H26 Wrist measurement? .... □□□
Around the small of the right wrist:______ inches (1 decimal)


A person's height and the measure of his wrist determines the body frame size
ADAM.

41

Section H – Dietary habits

0 5 □□□
Country ID Nº

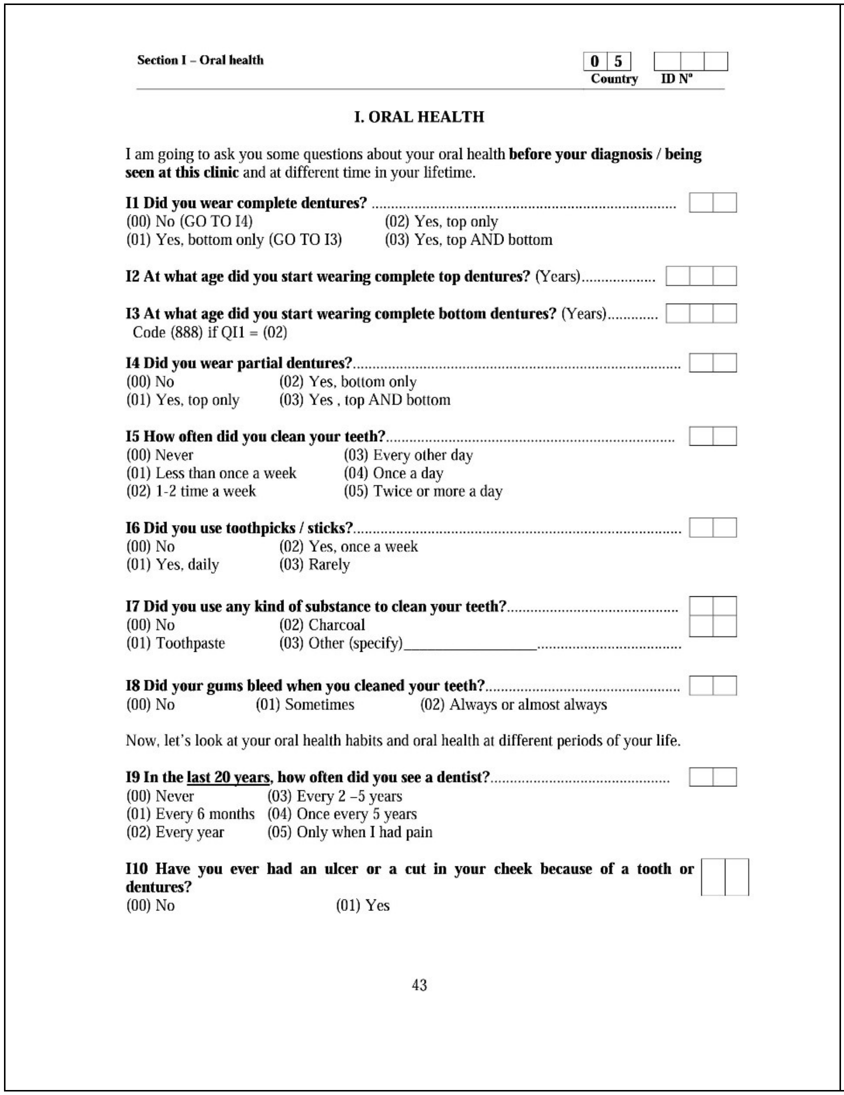
H27 What sized body frame does the subject have? □□
Please refer to the Interviewer's Guide for coding
(01) small body frame
(02) medium body frame
(03) large body frame
(04) man under 165 cm height

H28 When you're AT WORK (include work as a housewife), which of the following best describes your level of activity?
(01) Very active (e.g., farmer, labourer, athlete)
(02) Moderately active
(03) Sedentary (e.g., desk job)
(04) I don't work
(99) I don't know

H29 When you're AT HOME, which of the following best describes your level of activity?
(01) Very active
(02) Moderately active
(03) Sedentary
(99) I don't know

H30 DURING LEISURE TIME, which of the following best describes your level of activity?
(01) Very active
(02) Moderately active
(03) Sedentary
(99) I don't know

42



Section I – Oral health 0 5 □□□
Country ID N°

I. ORAL HEALTH

I am going to ask you some questions about your oral health **before your diagnosis / being seen at this clinic** and at different time in your lifetime.

I1 Did you wear complete dentures? .... □□
(00) No (GO TO I4) (02) Yes, top only
(01) Yes, bottom only (GO TO I3) (03) Yes, top AND bottom

I2 At what age did you start wearing complete top dentures? (Years)................... □□□

I3 At what age did you start wearing complete bottom dentures? (Years)............. □□□
Code (888) if QI1 = (02)

I4 Did you wear partial dentures?.. □□
(00) No (02) Yes, bottom only
(01) Yes, top only (03) Yes , top AND bottom

I5 How often did you clean your teeth?.. □□
(00) Never (03) Every other day
(01) Less than once a week (04) Once a day
(02) 1-2 time a week (05) Twice or more a day

I6 Did you use toothpicks / sticks?... □□
(00) No (02) Yes, once a week
(01) Yes, daily (03) Rarely

I7 Did you use any kind of substance to clean your teeth?... □□
(00) No (02) Charcoal □□
(01) Toothpaste (03) Other (specify)_________________....................................

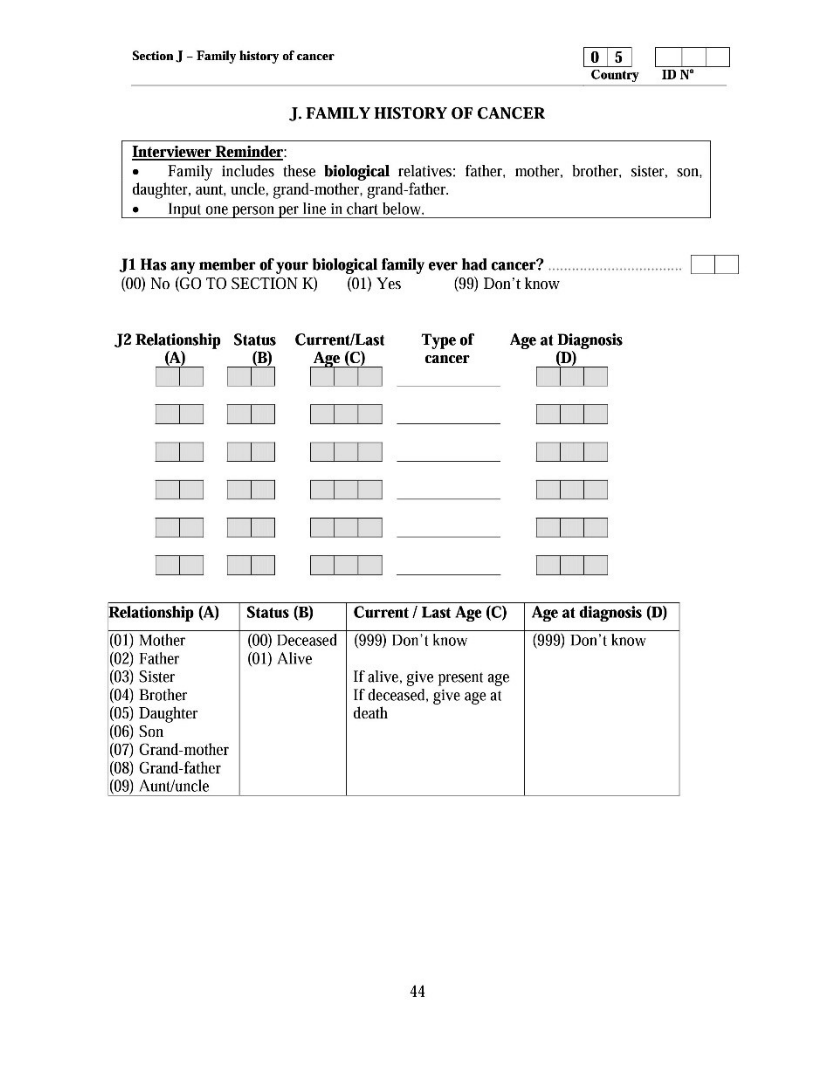
I8 Did your gums bleed when you cleaned your teeth?... □□
(00) No (01) Sometimes (02) Always or almost always

Now, let's look at your oral health habits and oral health at different periods of your life.

I9 In the <u>last 20 years</u>, how often did you see a dentist?.. □□
(00) Never (03) Every 2 –5 years
(01) Every 6 months (04) Once every 5 years
(02) Every year (05) Only when I had pain

I10 Have you ever had an ulcer or a cut in your cheek because of a tooth or dentures? □□
(00) No (01) Yes

43



Section J – Family history of cancer 0 5 □□□
Country ID N°

J. FAMILY HISTORY OF CANCER

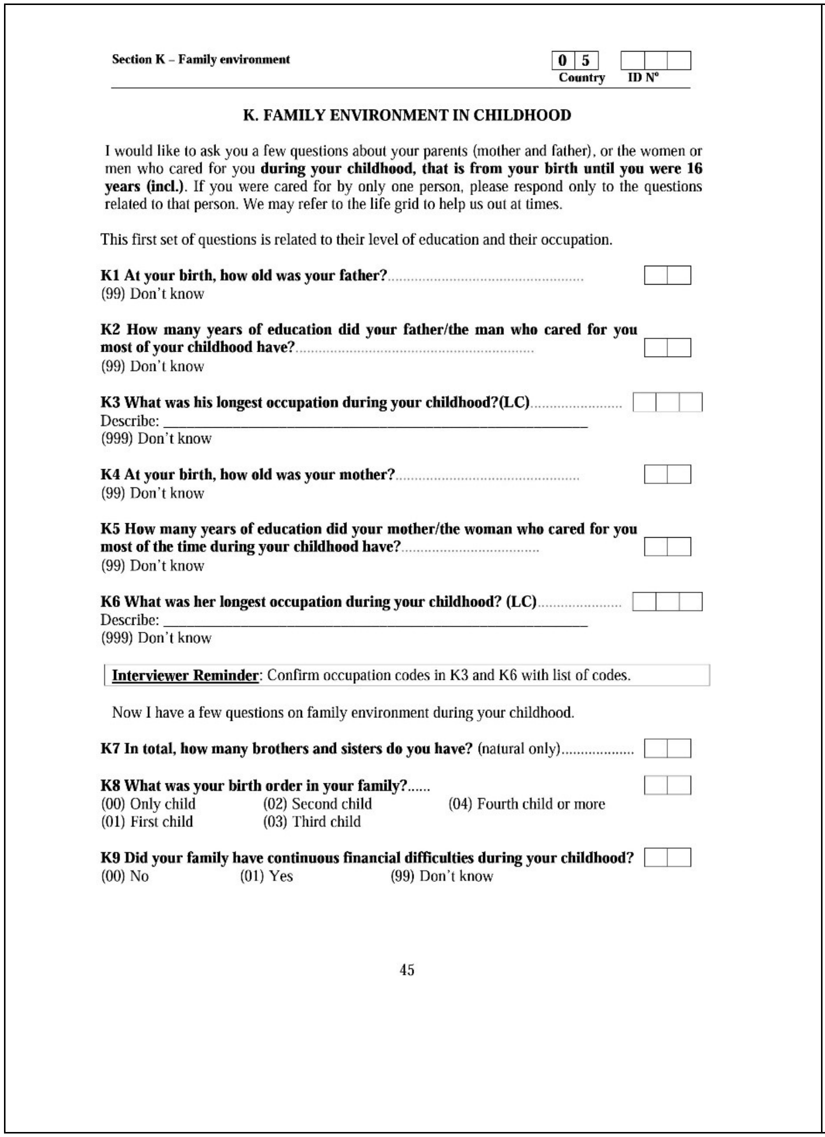
<u>Interviewer Reminder</u>:
- Family includes these **biological** relatives: father, mother, brother, sister, son, daughter, aunt, uncle, grand-mother, grand-father.
- Input one person per line in chart below.

J1 Has any member of your biological family ever had cancer? .... □□
(00) No (GO TO SECTION K) (01) Yes (99) Don't know

J2 Relationship (A)	**Status (B)**	**Current/Last Age (C)**	**Type of cancer**	**Age at Diagnosis (D)**
□□	□□	□□□	__________	□□□
□□	□□	□□□	__________	□□□
□□	□□	□□□	__________	□□□
□□	□□	□□□	__________	□□□
□□	□□	□□□	__________	□□□
□□	□□	□□□	__________	□□□

Relationship (A)	Status (B)	Current / Last Age (C)	Age at diagnosis (D)
(01) Mother (02) Father (03) Sister (04) Brother (05) Daughter (06) Son (07) Grand-mother (08) Grand-father (09) Aunt/uncle	(00) Deceased (01) Alive	(999) Don't know If alive, give present age If deceased, give age at death	(999) Don't know

44



Section K – Family environment

0	5	□□□
Country		ID N°

K. FAMILY ENVIRONMENT IN CHILDHOOD

I would like to ask you a few questions about your parents (mother and father), or the women or men who cared for you **during your childhood, that is from your birth until you were 16 years (incl.)**. If you were cared for by only one person, please respond only to the questions related to that person. We may refer to the life grid to help us out at times.

This first set of questions is related to their level of education and their occupation.

K1 At your birth, how old was your father?.. □□
(99) Don't know

K2 How many years of education did your father/the man who cared for you most of your childhood have?.. □□
(99) Don't know

K3 What was his longest occupation during your childhood?(LC)........................ □□□
Describe: ____
(999) Don't know

K4 At your birth, how old was your mother?... □□
(99) Don't know

K5 How many years of education did your mother/the woman who cared for you most of the time during your childhood have?................................... □□
(99) Don't know

K6 What was her longest occupation during your childhood? (LC)...................... □□□
Describe: ____
(999) Don't know

Interviewer Reminder: Confirm occupation codes in K3 and K6 with list of codes.

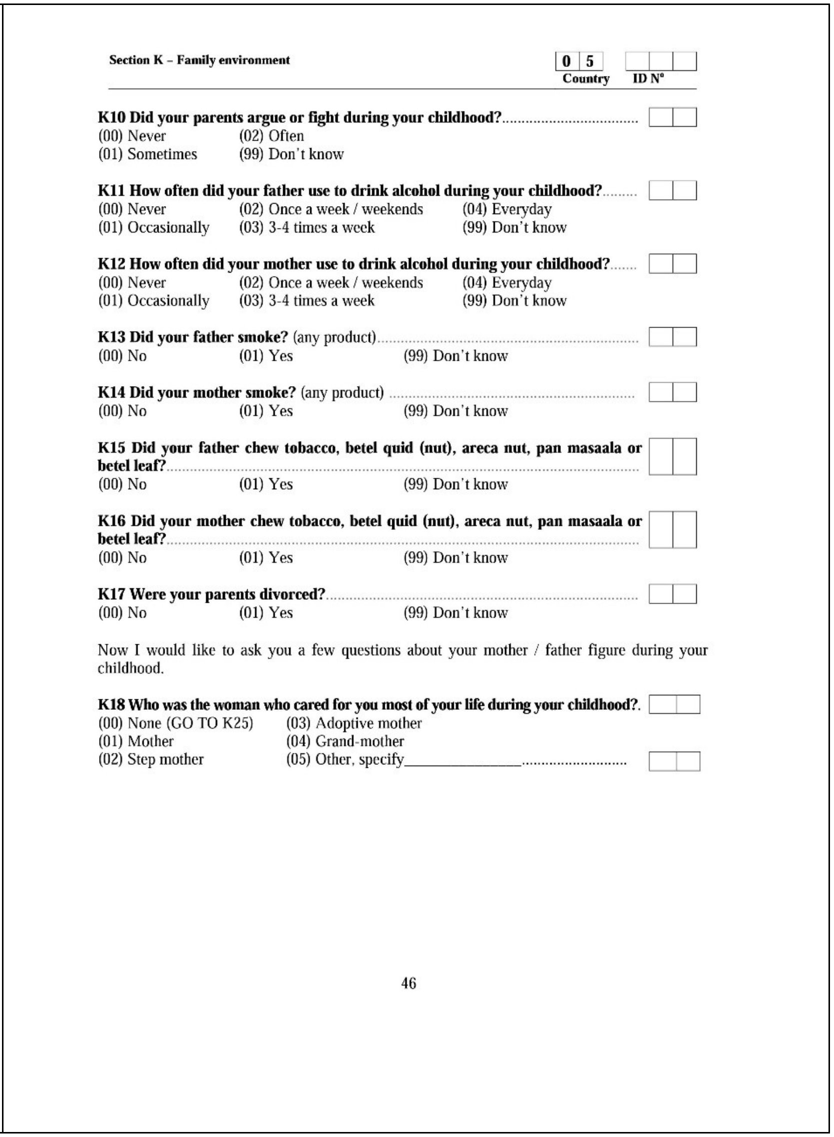
Now I have a few questions on family environment during your childhood.

K7 In total, how many brothers and sisters do you have? (natural only).................. □□

K8 What was your birth order in your family?...... □□
(00) Only child (02) Second child (04) Fourth child or more
(01) First child (03) Third child

K9 Did your family have continuous financial difficulties during your childhood? □□
(00) No (01) Yes (99) Don't know

45



Section K – Family environment

0	5	□□□
Country		ID N°

K10 Did your parents argue or fight during your childhood?.................................. □□
(00) Never (02) Often
(01) Sometimes (99) Don't know

K11 How often did your father use to drink alcohol during your childhood?......... □□
(00) Never (02) Once a week / weekends (04) Everyday
(01) Occasionally (03) 3-4 times a week (99) Don't know

K12 How often did your mother use to drink alcohol during your childhood?....... □□
(00) Never (02) Once a week / weekends (04) Everyday
(01) Occasionally (03) 3-4 times a week (99) Don't know

K13 Did your father smoke? (any product)... □□
(00) No (01) Yes (99) Don't know

K14 Did your mother smoke? (any product) .... □□
(00) No (01) Yes (99) Don't know

K15 Did your father chew tobacco, betel quid (nut), areca nut, pan masaala or betel leaf?.. □□
(00) No (01) Yes (99) Don't know

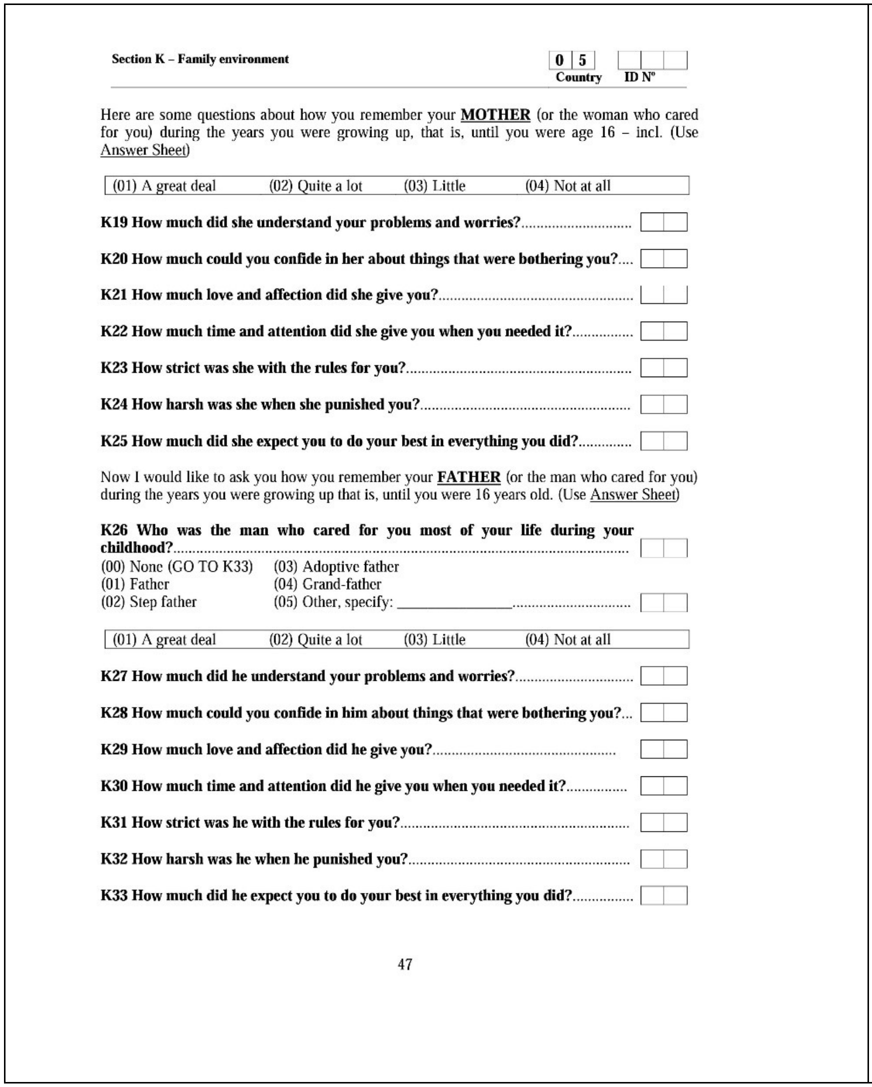
K16 Did your mother chew tobacco, betel quid (nut), areca nut, pan masaala or betel leaf?.. □□
(00) No (01) Yes (99) Don't know

K17 Were your parents divorced?... □□
(00) No (01) Yes (99) Don't know

Now I would like to ask you a few questions about your mother / father figure during your childhood.

K18 Who was the woman who cared for you most of your life during your childhood?. □□
(00) None (GO TO K25) (03) Adoptive mother
(01) Mother (04) Grand-mother
(02) Step mother (05) Other, specify_______________.......................... □□

46



Section K – Family environment

0 5 Country □□□ ID N°

Here are some questions about how you remember your **MOTHER** (or the woman who cared for you) during the years you were growing up, that is, until you were age 16 – incl. (Use Answer Sheet)

(01) A great deal	(02) Quite a lot	(03) Little	(04) Not at all

K19 How much did she understand your problems and worries?............................ □□

K20 How much could you confide in her about things that were bothering you?.... □□

K21 How much love and affection did she give you?... □□

K22 How much time and attention did she give you when you needed it?............... □□

K23 How strict was she with the rules for you?... □□

K24 How harsh was she when she punished you?.. □□

K25 How much did she expect you to do your best in everything you did?............. □□

Now I would like to ask you how you remember your **FATHER** (or the man who cared for you) during the years you were growing up that is, until you were 16 years old. (Use Answer Sheet)

K26 Who was the man who cared for you most of your life during your childhood?.. □□

(00) None (GO TO K33) (03) Adoptive father
(01) Father (04) Grand-father
(02) Step father (05) Other, specify: _______________.............................. □□

(01) A great deal	(02) Quite a lot	(03) Little	(04) Not at all

K27 How much did he understand your problems and worries?.............................. □□

K28 How much could you confide in him about things that were bothering you?... □□

K29 How much love and affection did he give you?... □□

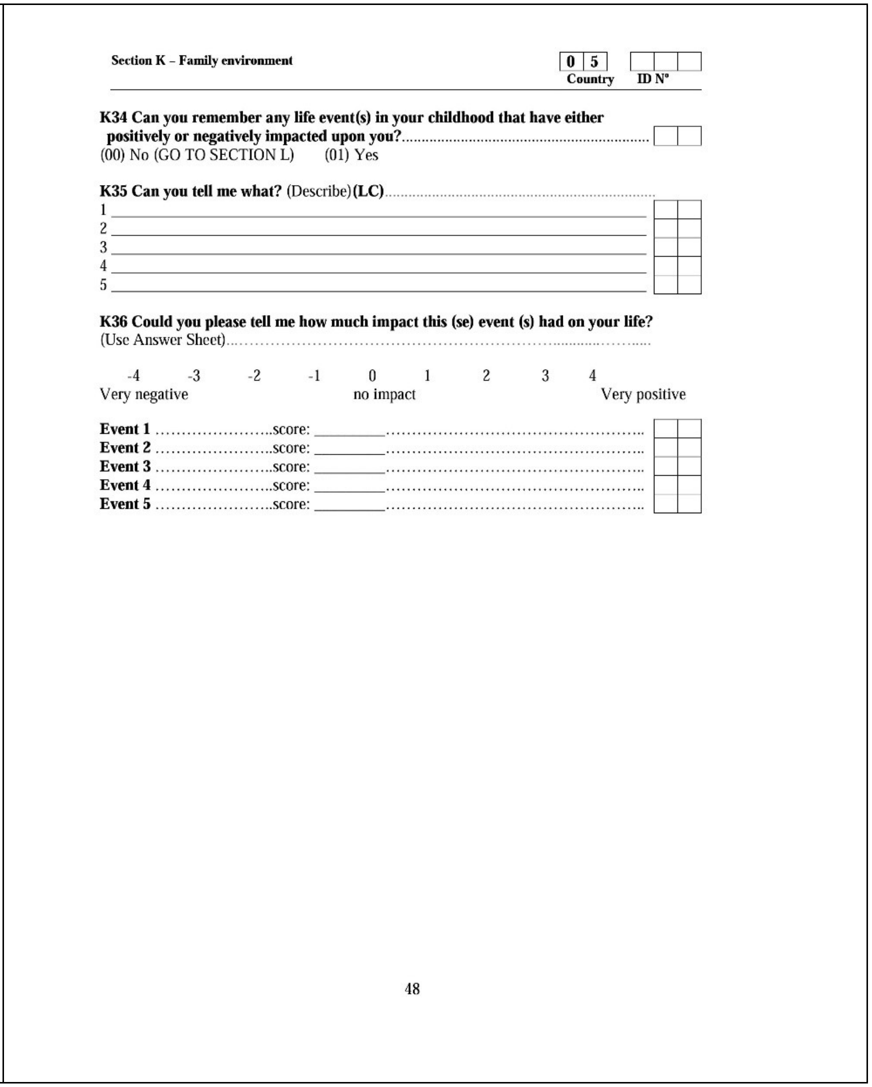
K30 How much time and attention did he give you when you needed it?............... □□

K31 How strict was he with the rules for you?.. □□

K32 How harsh was he when he punished you?... □□

K33 How much did he expect you to do your best in everything you did?............... □□

47



Section K – Family environment

0 5 Country □□□ ID N°

K34 Can you remember any life event(s) in your childhood that have either positively or negatively impacted upon you?... □□
(00) No (GO TO SECTION L) (01) Yes

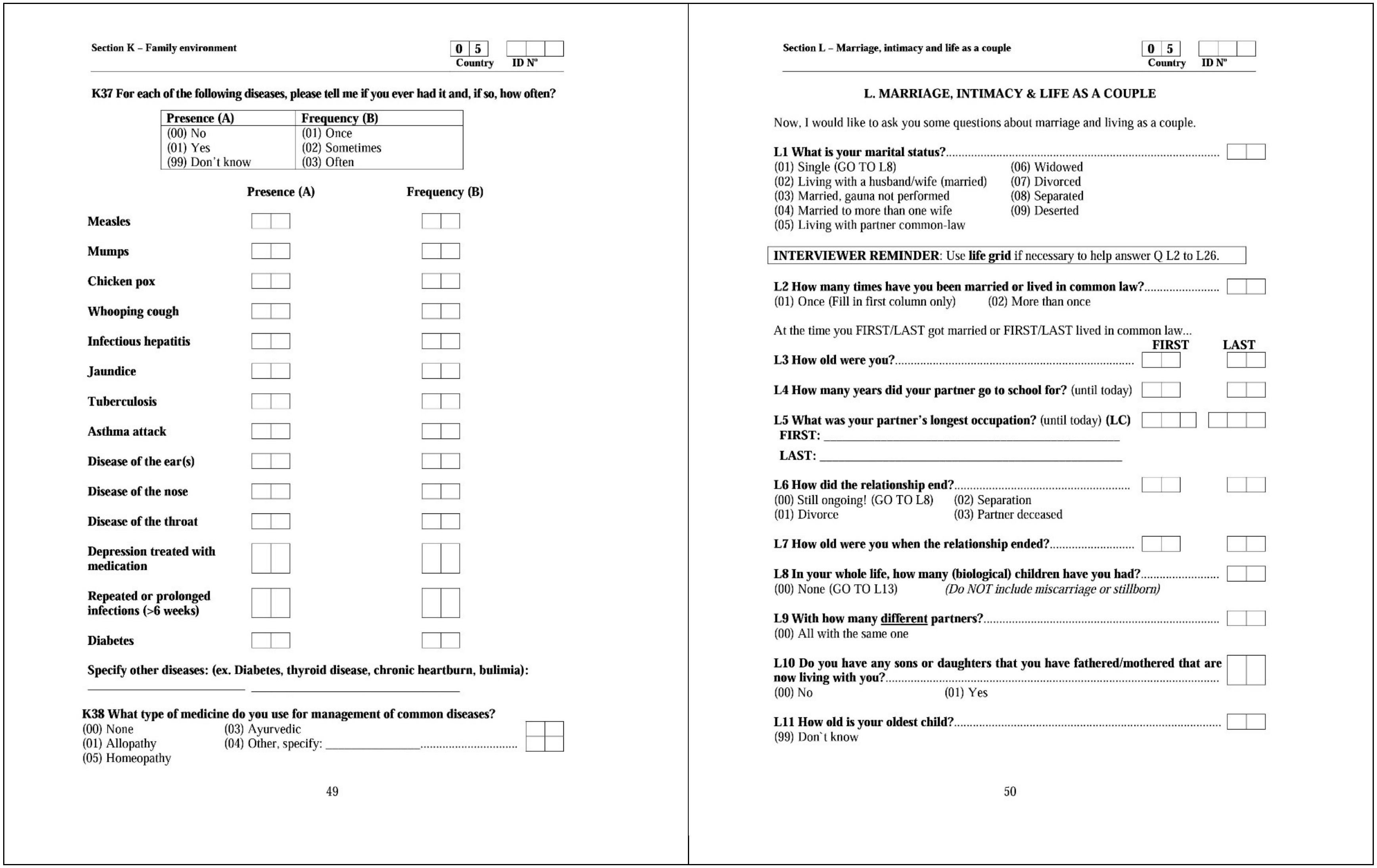
K35 Can you tell me what? (Describe)**(LC)**...
1 ____ □□
2 ____ □□
3 ____ □□
4 ____ □□
5 ____ □□

K36 Could you please tell me how much impact this (se) event (s) had on your life?
(Use Answer Sheet)...

-4	-3	-2	-1	0	1	2	3	4
Very negative				no impact				Very positive

Event 1 …………………..score: _________…………………………………………. □□
Event 2 …………………..score: _________…………………………………………. □□
Event 3 …………………..score: _________…………………………………………. □□
Event 4 …………………..score: _________…………………………………………. □□
Event 5 …………………..score: _________…………………………………………. □□

48



Section K – Family environment

0 5	□□□
Country	ID N°

K37 For each of the following diseases, please tell me if you ever had it and, if so, how often?

Presence (A)	Frequency (B)
(00) No	(01) Once
(01) Yes	(02) Sometimes
(99) Don't know	(03) Often

	Presence (A)	Frequency (B)
Measles	□□	□□
Mumps	□□	□□
Chicken pox	□□	□□
Whooping cough	□□	□□
Infectious hepatitis	□□	□□
Jaundice	□□	□□
Tuberculosis	□□	□□
Asthma attack	□□	□□
Disease of the ear(s)	□□	□□
Disease of the nose	□□	□□
Disease of the throat	□□	□□
Depression treated with medication	□□	□□
Repeated or prolonged infections (>6 weeks)	□□	□□
Diabetes	□□	□□

Specify other diseases: (ex. Diabetes, thyroid disease, chronic heartburn, bulimia):

________________ ________________

K38 What type of medicine do you use for management of common diseases? □□ □□

(00) None (03) Ayurvedic
(01) Allopathy (04) Other, specify: ______________..............................
(05) Homeopathy

Section L – Marriage, intimacy and life as a couple

0 5	□□□
Country	ID N°

L. MARRIAGE, INTIMACY & LIFE AS A COUPLE

Now, I would like to ask you some questions about marriage and living as a couple.

L1 What is your marital status?.. □□

(01) Single (GO TO L8) (06) Widowed
(02) Living with a husband/wife (married) (07) Divorced
(03) Married, gauna not performed (08) Separated
(04) Married to more than one wife (09) Deserted
(05) Living with partner common-law

INTERVIEWER REMINDER: Use **life grid** if necessary to help answer Q L2 to L26.

L2 How many times have you been married or lived in common law?........................ □□

(01) Once (Fill in first column only) (02) More than once

At the time you FIRST/LAST got married or FIRST/LAST lived in common law...

	FIRST	LAST
L3 How old were you?..	□□	□□
L4 How many years did your partner go to school for? (until today)	□□	□□
L5 What was your partner's longest occupation? (until today) **(LC)**	□□□	□□□

FIRST: ____

LAST: ____

	FIRST	LAST
L6 How did the relationship end?...	□□	□□

(00) Still ongoing! (GO TO L8) (02) Separation
(01) Divorce (03) Partner deceased

	FIRST	LAST
L7 How old were you when the relationship ended?..........................	□□	□□

L8 In your whole life, how many (biological) children have you had?......................... □□

(00) None (GO TO L13) *(Do NOT include miscarriage or stillborn)*

L9 With how many <u>different</u> partners?.. □□

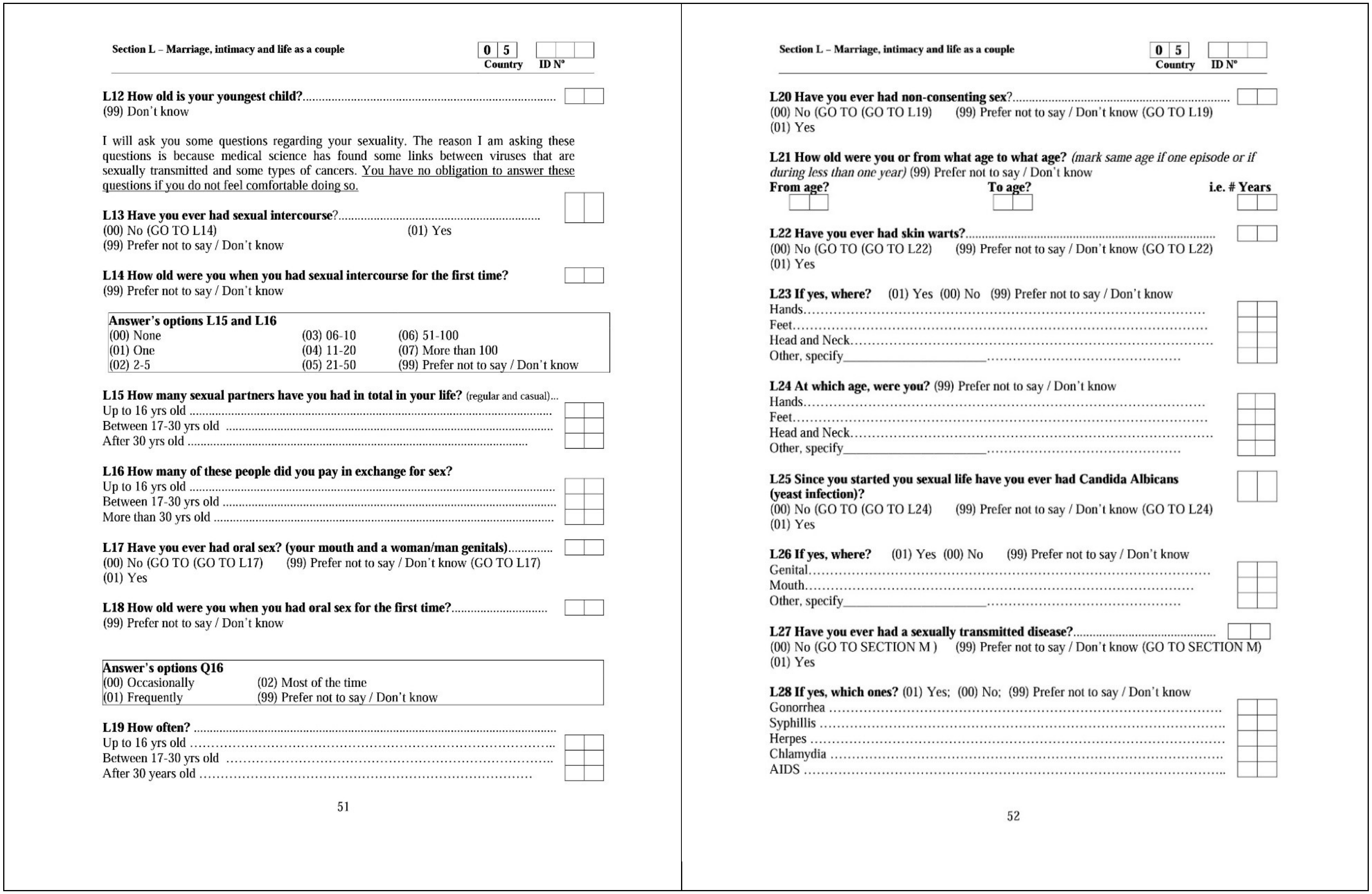
(00) All with the same one

L10 Do you have any sons or daughters that you have fathered/mothered that are now living with you?.. □□

(00) No (01) Yes

L11 How old is your oldest child?... □□

(99) Don`t know



Section L – Marriage, intimacy and life as a couple 0 5 Country ID N°

L12 How old is your youngest child?.. □□
(99) Don't know

I will ask you some questions regarding your sexuality. The reason I am asking these questions is because medical science has found some links between viruses that are sexually transmitted and some types of cancers. You have no obligation to answer these questions if you do not feel comfortable doing so.

L13 Have you ever had sexual intercourse?.. □□
(00) No (GO TO L14) (01) Yes
(99) Prefer not to say / Don't know

L14 How old were you when you had sexual intercourse for the first time? □□
(99) Prefer not to say / Don't know

Answer's options L15 and L16		
(00) None	(03) 06-10	(06) 51-100
(01) One	(04) 11-20	(07) More than 100
(02) 2-5	(05) 21-50	(99) Prefer not to say / Don't know

L15 How many sexual partners have you had in total in your life? (regular and casual)...
Up to 16 yrs old .... □□
Between 17-30 yrs old .... □□
After 30 yrs old .... □□

L16 How many of these people did you pay in exchange for sex?
Up to 16 yrs old .... □□
Between 17-30 yrs old .... □□
More than 30 yrs old .... □□

L17 Have you ever had oral sex? (your mouth and a woman/man genitals)............. □□
(00) No (GO TO (GO TO L17) (99) Prefer not to say / Don't know (GO TO L17)
(01) Yes

L18 How old were you when you had oral sex for the first time?............................ □□
(99) Prefer not to say / Don't know

Answer's options Q16	
(00) Occasionally	(02) Most of the time
(01) Frequently	(99) Prefer not to say / Don't know

L19 How often? ....
Up to 16 yrs old .... □□
Between 17-30 yrs old .... □□
After 30 years old .... □□

51

Section L – Marriage, intimacy and life as a couple 0 5 Country ID N°

L20 Have you ever had non-consenting sex?.. □□
(00) No (GO TO (GO TO L19) (99) Prefer not to say / Don't know (GO TO L19)
(01) Yes

L21 How old were you or from what age to what age? *(mark same age if one episode or if during less than one year)* (99) Prefer not to say / Don't know
From age? □□ **To age?** □□ **i.e. # Years** □□

L22 Have you ever had skin warts?.. □□
(00) No (GO TO (GO TO L22) (99) Prefer not to say / Don't know (GO TO L22)
(01) Yes

L23 If yes, where? (01) Yes (00) No (99) Prefer not to say / Don't know
Hands.. □□
Feet.. □□
Head and Neck.. □□
Other, specify______________________... □□

L24 At which age, were you? (99) Prefer not to say / Don't know
Hands.. □□
Feet.. □□
Head and Neck.. □□
Other, specify______________________... □□

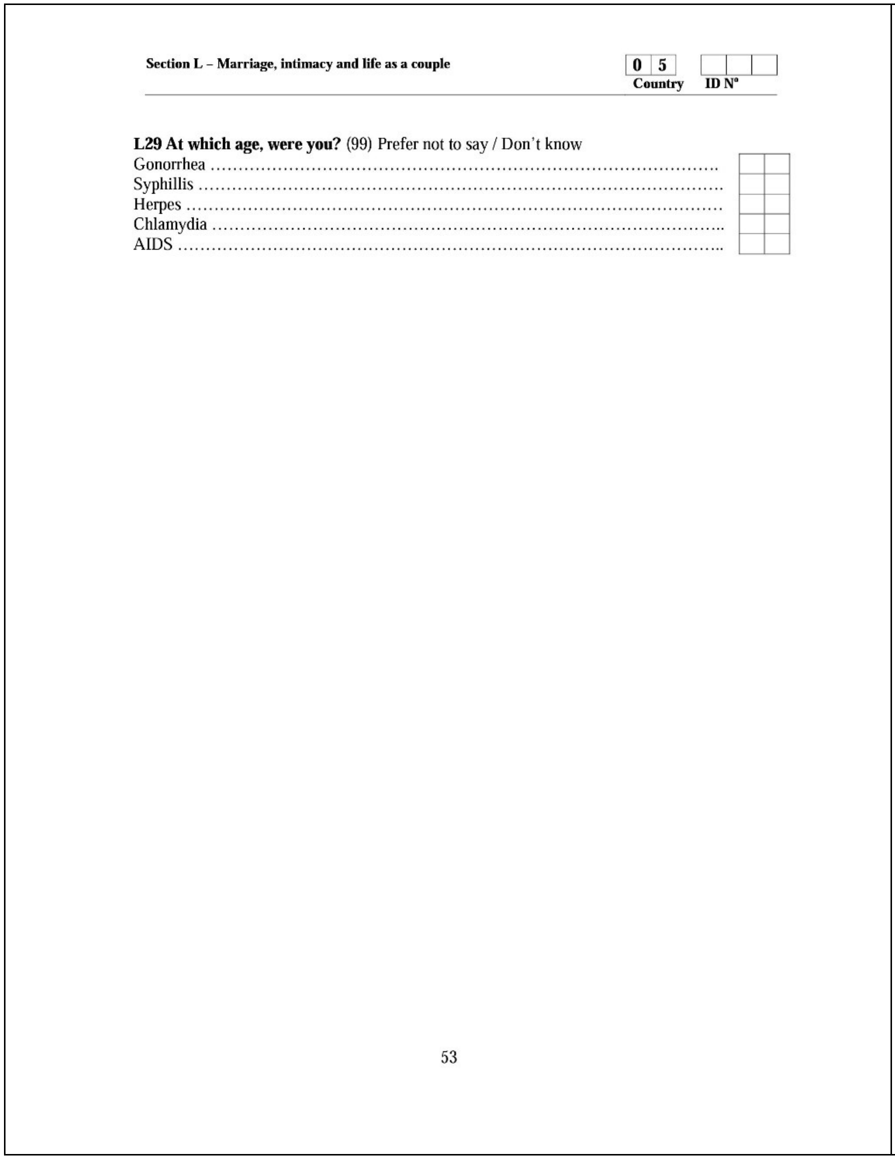
L25 Since you started you sexual life have you ever had Candida Albicans (yeast infection)? □□
(00) No (GO TO (GO TO L24) (99) Prefer not to say / Don't know (GO TO L24)
(01) Yes

L26 If yes, where? (01) Yes (00) No (99) Prefer not to say / Don't know
Genital.. □□
Mouth.. □□
Other, specify______________________... □□

L27 Have you ever had a sexually transmitted disease?... □□
(00) No (GO TO SECTION M) (99) Prefer not to say / Don't know (GO TO SECTION M)
(01) Yes

L28 If yes, which ones? (01) Yes; (00) No; (99) Prefer not to say / Don't know
Gonorrhea .... □□
Syphillis .... □□
Herpes .... □□
Chlamydia .... □□
AIDS .... □□

52

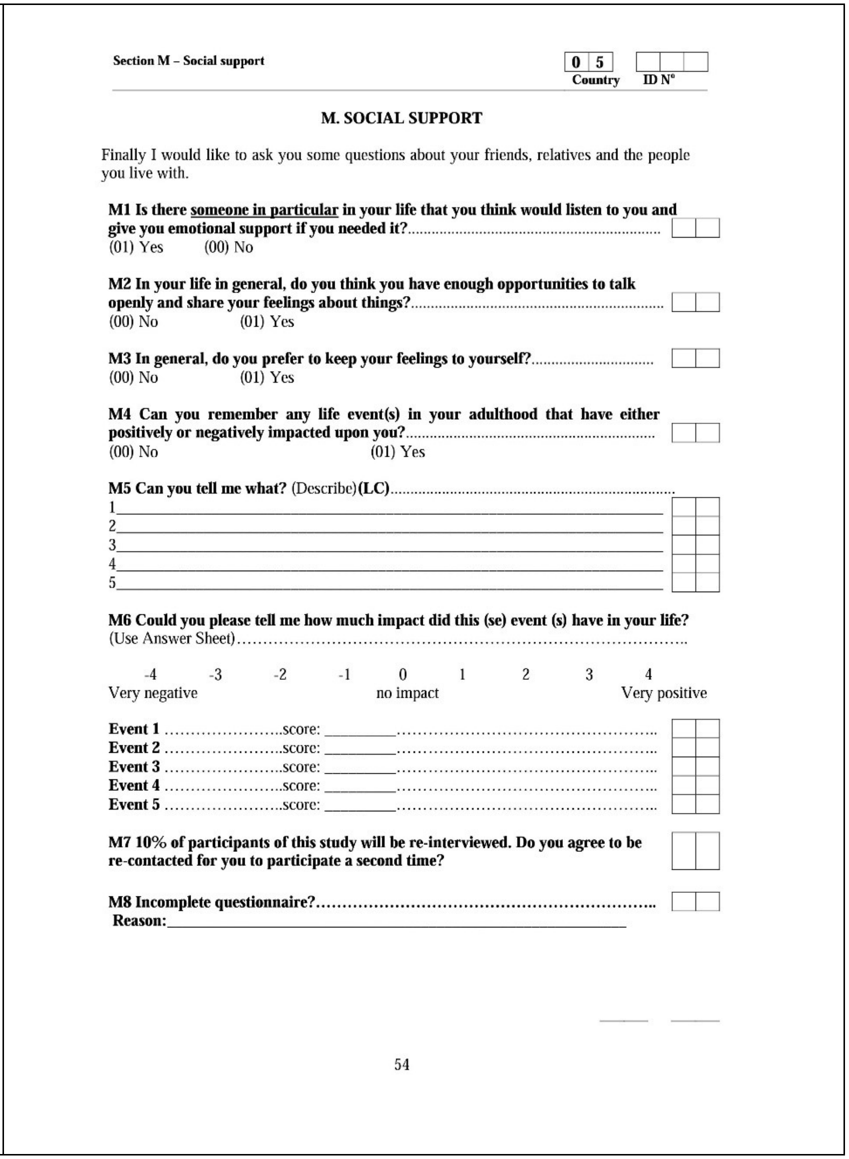


Section L – Marriage, intimacy and life as a couple 0 5 □□□
Country **ID N°**

L29 At which age, were you? (99) Prefer not to say / Don't know

Gonorrhea ………………………………………………………………………………. □□
Syphillis …………………………………………………………………………………. □□
Herpes …………………………………………………………………………………. □□
Chlamydia ……………………………………………………………………………….. □□
AIDS ……………………………………………………………………………………….. □□

53



Section M – Social support 0 5 □□□
Country **ID N°**

M. SOCIAL SUPPORT

Finally I would like to ask you some questions about your friends, relatives and the people you live with.

M1 Is there someone in particular in your life that you think would listen to you and give you emotional support if you needed it?... □□
(01) Yes (00) No

M2 In your life in general, do you think you have enough opportunities to talk openly and share your feelings about things?.. □□
(00) No (01) Yes

M3 In general, do you prefer to keep your feelings to yourself?.............................. □□
(00) No (01) Yes

M4 Can you remember any life event(s) in your adulthood that have either positively or negatively impacted upon you?.. □□
(00) No (01) Yes

M5 Can you tell me what? (Describe)**(LC)**..

1____ □□
2____ □□
3____ □□
4____ □□
5____ □□

M6 Could you please tell me how much impact did this (se) event (s) have in your life?
(Use Answer Sheet)…………………………………………………………………………..

-4	-3	-2	-1	0	1	2	3	4
Very negative				no impact				Very positive

Event 1 …………………..score: _________…………………………………………….. □□
Event 2 …………………..score: _________…………………………………………….. □□
Event 3 …………………..score: _________…………………………………………….. □□
Event 4 …………………..score: _________…………………………………………….. □□
Event 5 …………………..score: _________…………………………………………….. □□

M7 10% of participants of this study will be re-interviewed. Do you agree to be re-contacted for you to participate a second time? □□

M8 Incomplete questionnaire?……………………………………………………….. □□
Reason:____

54

Section M – Social support 0 5 □□□
Country ID N°

M9 Time of end of interview.. □□ - □□
Hour Minute

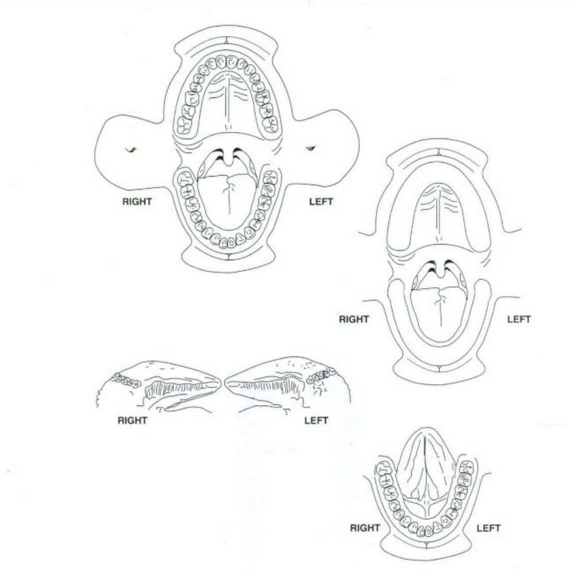
M10 Data enterer's initials? .... □□

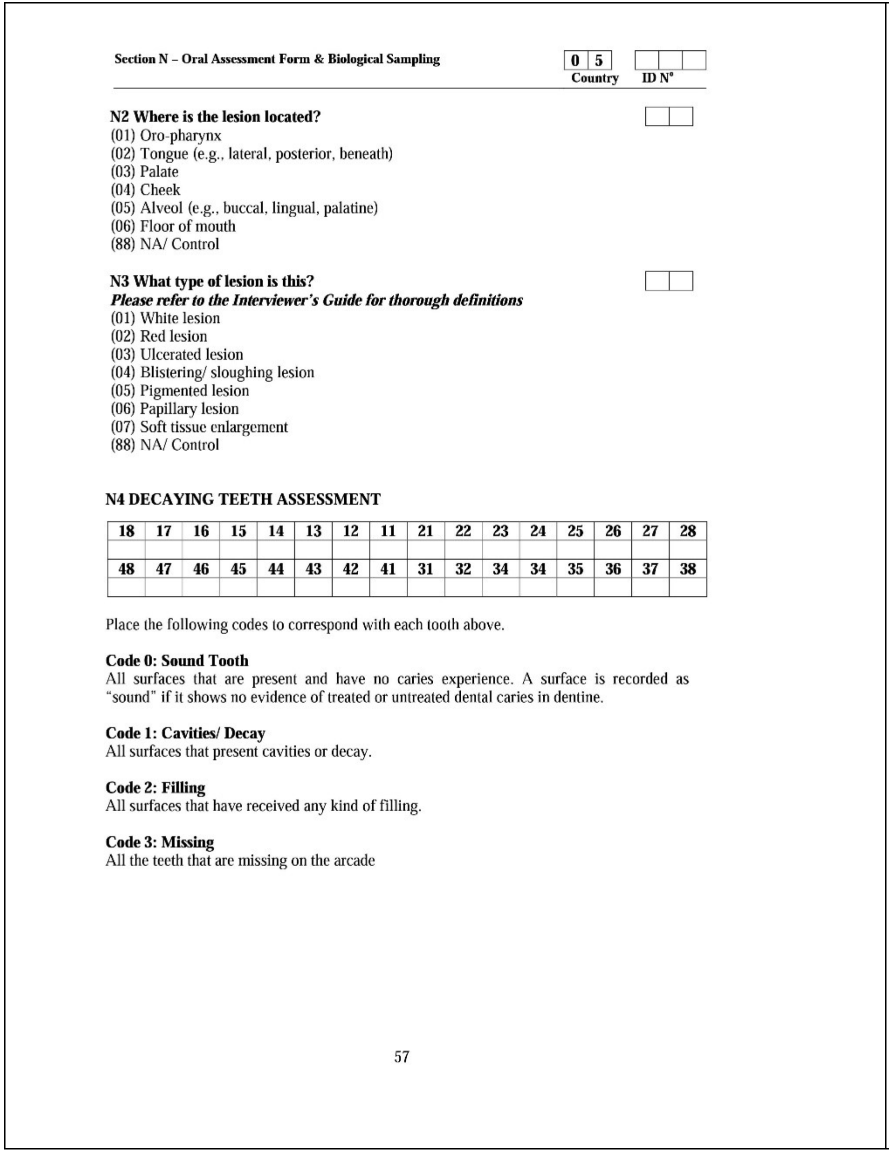
Participant's comments:

Section N – Oral Assessment Form & Biological Sampling 0 5 □□□
Country ID N°

N. ORAL ASSESSMENT FORM & BIOLOGICAL SAMPLING

N1. VISIBLE LESIONS AND IRRITATIONS
Circle the place in the mouth where you see the lesion.





Section N – Oral Assessment Form & Biological Sampling

0	5				
Country		ID N°			

N2 Where is the lesion located? □□

(01) Oro-pharynx
(02) Tongue (e.g., lateral, posterior, beneath)
(03) Palate
(04) Cheek
(05) Alveol (e.g., buccal, lingual, palatine)
(06) Floor of mouth
(88) NA/ Control

N3 What type of lesion is this? □□
Please refer to the Interviewer's Guide for thorough definitions

(01) White lesion
(02) Red lesion
(03) Ulcerated lesion
(04) Blistering/ sloughing lesion
(05) Pigmented lesion
(06) Papillary lesion
(07) Soft tissue enlargement
(88) NA/ Control

N4 DECAYING TEETH ASSESSMENT

18	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
48	**47**	**46**	**45**	**44**	**43**	**42**	**41**	**31**	**32**	**34**	**34**	**35**	**36**	**37**	**38**

Place the following codes to correspond with each tooth above.

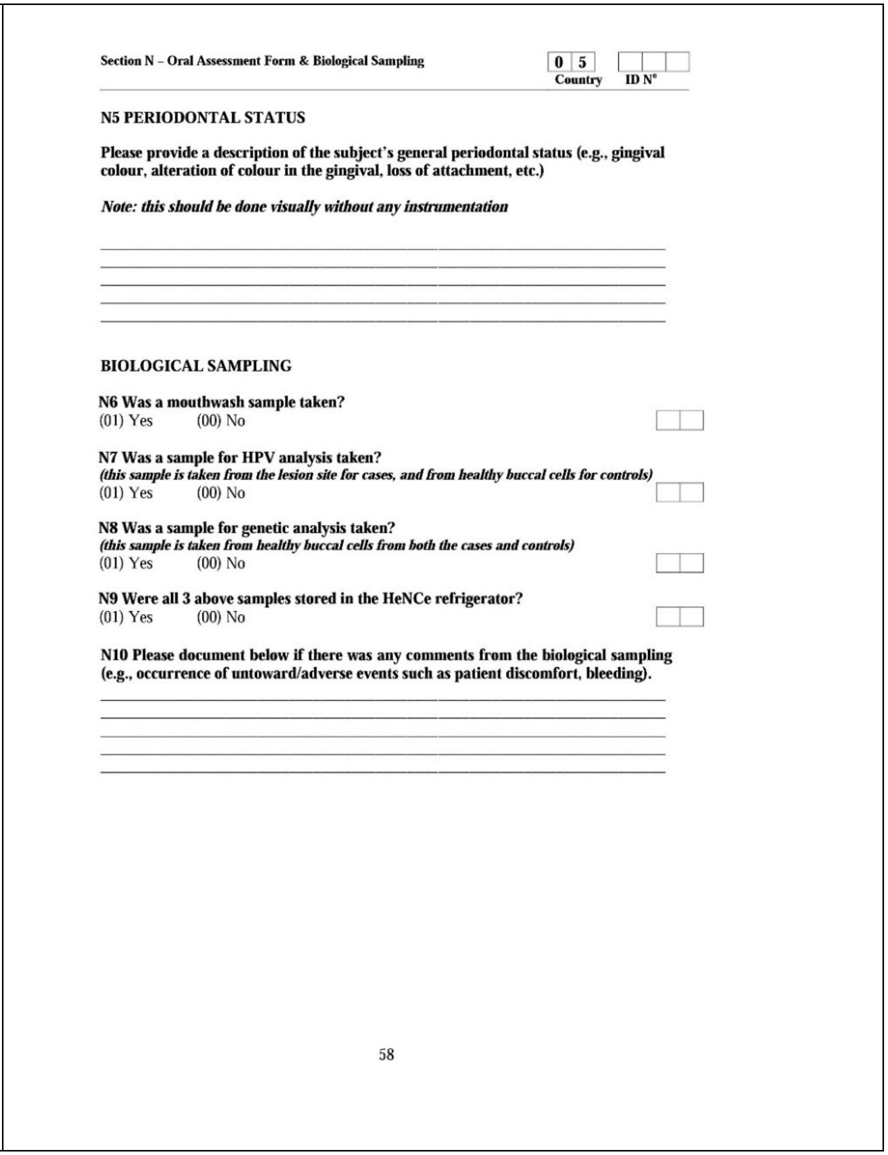
Code 0: Sound Tooth
All surfaces that are present and have no caries experience. A surface is recorded as "sound" if it shows no evidence of treated or untreated dental caries in dentine.

Code 1: Cavities/ Decay
All surfaces that present cavities or decay.

Code 2: Filling
All surfaces that have received any kind of filling.

Code 3: Missing
All the teeth that are missing on the arcade

57



Section N – Oral Assessment Form & Biological Sampling

0	5				
Country		ID N°			

N5 PERIODONTAL STATUS

Please provide a description of the subject's general periodontal status (e.g., gingival colour, alteration of colour in the gingival, loss of attachment, etc.)

Note: this should be done visually without any instrumentation

____
____
____
____
____

BIOLOGICAL SAMPLING

N6 Was a mouthwash sample taken?
(01) Yes (00) No □□

N7 Was a sample for HPV analysis taken?
(this sample is taken from the lesion site for cases, and from healthy buccal cells for controls)
(01) Yes (00) No □□

N8 Was a sample for genetic analysis taken?
(this sample is taken from healthy buccal cells from both the cases and controls)
(01) Yes (00) No □□

N9 Were all 3 above samples stored in the HeNCe refrigerator?
(01) Yes (00) No □□

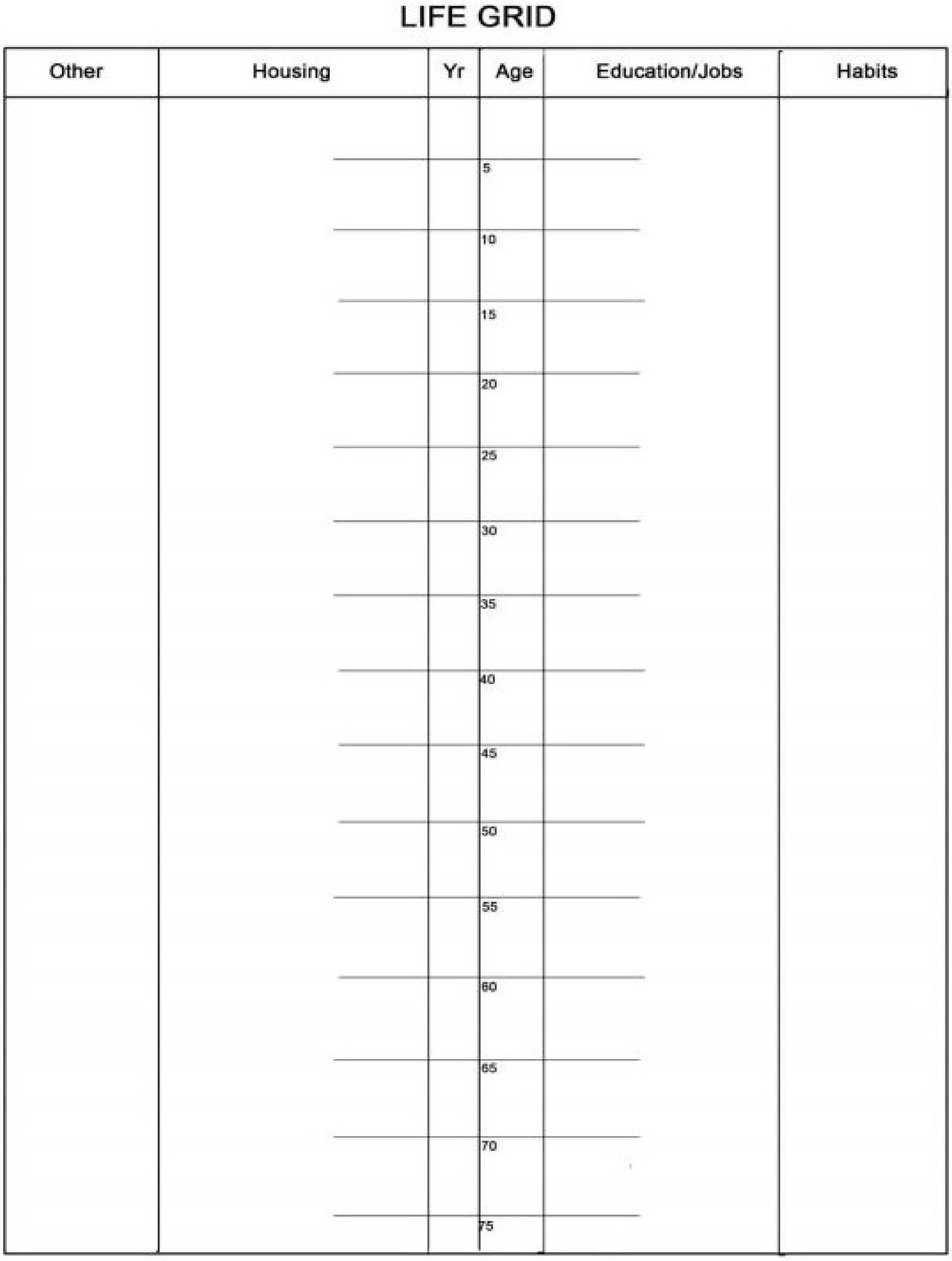
N10 Please document below if there was any comments from the biological sampling (e.g., occurrence of untoward/adverse events such as patient discomfort, bleeding).

____
____
____
____
____

58

Appendix III

Life-grid



LIFE GRID

Other	Housing	Yr	Age	Education/Jobs	Habits
			5		
			10		
			15		
			20		
			25		
			30		
			35		
			40		
			45		
			50		
			55		
			60		
			65		
			70		
			75		

Appendix IV

Biological sample collection procedures

BIOLOGICAL SPECIMEN COLLECTION

- **Collection of biological specimens will be done by the Clinician treating the patient or a Research Assistant in the presence of the clinician treating the patient.**
- **Samples should be taken before the patient's treatment.**
- **Samples should preferably be taken in the morning hours to ensure an adequate amount of cells for sampling.**

1. Prepare kit with supplies:

a) 20 mL wide mouth cylindrical **collection vial** with a label containing the patient code and date on both the side and top. Vials should be cryogenic polypropylene vials and polypropylene screw tops with a hermetic seal.

b) One single use bottle (~20 mL) of **mouthwash** (non-alcoholic solution provided)

c) Two individually packaged **Oral CDx® brushes** kept sealed until ready for use

d) Two single use vials (~20 mL) of **PreservCyt® (Cytyc Inc.)** buffer bottles kept sealed until ready to use. These vials should be **pre-labelled and numbered** prior to sampling to facilitate tracking. One vial should be labelled for genetic analysis (**GEN**) and the second for HPV analysis (**HPV**).

e) **Surgical gloves** to be worn when collecting the sample

For CONTROLS:

1. Explain the procedures to the subject ensuring they understand all aspects and have signed the consent form.
2. Instruct patient to remove dentures if worn.
3. Instruct patient to rinse mouth vigorously and gargle for 15-30 seconds with mouthwash from the container. **Watch the clock while you do this**. The patient should gargle but should not clear their throat.
4. Ask the patient to spit all of the solution into the empty pre-labeled collection vial while holding the container close to their mouth.
5. Unwrap **Oral CDx® brush** and ask the patient open their mouth.
6. Collect oral sample by brushing the oral cavity with the **Oral CDx® brush** with several (5-10) gentle strokes on each of the following areas:

   a. Right buccal mucosa (from high to low position)

   b. Left buccal mucosa (from high to low position)

   c. Right side of the tongue

   d. Dorsal side of the tongue near to the base (note: may cause subject to experience a gag reflex - STOP if patient is too uncomfortable to complete the brushing at the site)

e. Left side of the tongue

1. After performing the brushing, carefully insert the brush into the vial containing **PreservCyt®** labeled **GEN** so as to avoid scraping any sample on the edges of the tube opening and so that the brush is suspended in the solution. Twirl brush vigorously to release as much of the sample into the solution. If sample remains on brush, repeat previous steps until most of sample is in solution. A (white) layer of cellular material should be visible settling on the bottom of the vial.
2. Repeat steps 5 to 7 with the second Oral CDx® brush and the second PreservCyt® vial (labeled **HPV**).
3. Place cap tightly on vials and place samples in the laboratory fridge for storage **as soon as possible.**
4. Document the occurrence of untoward/adverse events (e.g., excess bleeding, patient discomfort) in the HeNCe log book.
5. Record sampling and the deposit of the sample in the laboratory fridge in the HeNCe log book.

For CASES:

1. Explain the procedures to the subject ensuring they understand all aspects and have signed the consent form.
2. Instruct patient to remove dentures if worn.
3. Instruct patient to rinse mouth vigorously and gargle for 15-30 seconds with mouthwash from the container. **Watch the clock while you do this**. The patient should gargle but should not clear their throat. This step may be skipped if the patient has an overt oral lesion that would make it too painful to use an alcohol-based oral mouthwash.
4. Ask the patient to spit all of the solution into the empty pre-labeled collection vial while holding the container close to their mouth.
5. Unwrap one **Oral CDx®** brush and ask the patient open their mouth.
6. Collect a normal oral sample by brushing the healthy buccal mucosa with the **Oral CDx® brush** with several (5-10) gentle strokes on either of the following areas (**whichever is furthest away from the lesion**):
   a. Right buccal mucosa (from high to low position)
   b. Left buccal mucosa (from high to low position)

1. After performing the brushing, carefully insert the brush into the vial containing **PreservCyt®** (labeled **GEN**) so as to avoid scraping any sample on the edges of the tube opening and so that the brush is suspended in the solution. Twirl brush vigorously to release as much of the sample into the solution. If sample remains on brush, repeat previous steps until most of sample is in solution. A (white) layer of cellular material should be visible settling on the bottom of the vial.
2. For cases, an additional brushing (for HPV analysis) is necessary with a **second** **Oral CDx® brush**. Unwrap the second Oral CDx® brush, and brush the visible lesion site of the cancer in several (5-10) gentle strokes trying to avoid any necrotic areas. Ensure that an adequate sample of tissue is collected on the brush. Note location of lesion/tumor brushing in oral exam form.
3. After performing the brushing of tumor or lesion, carefully insert the brush into the prefilled vial containing **PreservCyt®** (labeled **HPV**) so as to avoid scraping any sample on the edges of the tube opening and so that the brush is suspended in the solution. Twirl brush vigorously to release as much of the sample into the solution. If sample remains on brush, repeat previous steps until most

of sample is in solution. A (white) layer of cellular material should be visible settling on the bottom of the vial.

4. Place cap tightly on mouthwash vial as well as the two Preservcyt® vials and place in the laboratory fridge for storage **as soon as possible**.
5. Document the occurrence of untoward/adverse events (e.g., excess bleeding, patient discomfort) in the HeNCe log book.
6. Record sampling and the deposit of the sample in the laboratory fridge in the HeNCe log book.

Figure 1

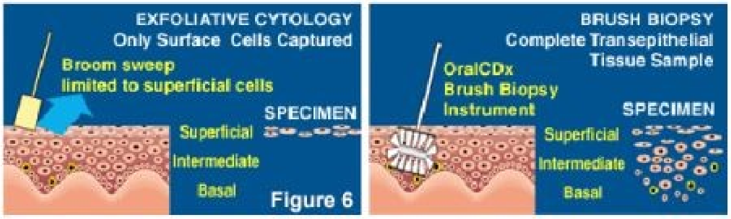
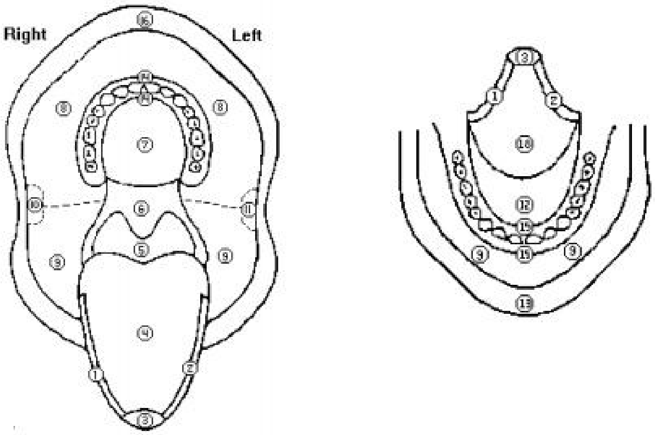


Figure 2

Sampling locations:

- Any epithelial area within the oral cavity including: buccal mucosa, inner and outer gums, tongue, inside of lips, soft and hard palate
- Oropharynx including: tonsils, oropharyngeal wall, base of tongue

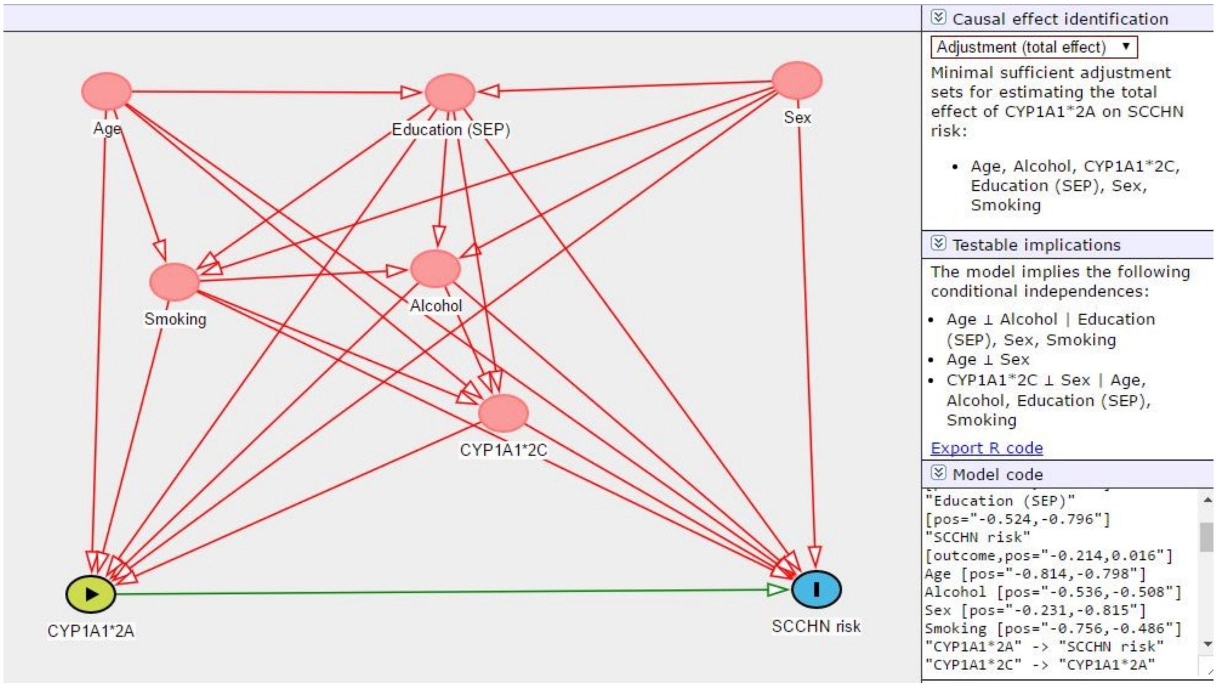


Appendix V

DAGs for manuscripts II and III

DAGs constructed in DAGitty version 2.3 for identifying minimal sufficient sets of potential confounders for estimating the total effect of each genetic variant on SCCHN risk with the available data at the Canadian site, HeNCe Life study, 2005 to 2013

Figure A1: Total effect of CYP1A1*2A on SCCHN risk



Causal effect identification

Adjustment (total effect) ▾

Minimal sufficient adjustment sets for estimating the total effect of CYP1A1*2A on SCCHN risk:

- Age, Alcohol, CYP1A1*2C, Education (SEP), Sex, Smoking

Testable implications

The model implies the following conditional independences:

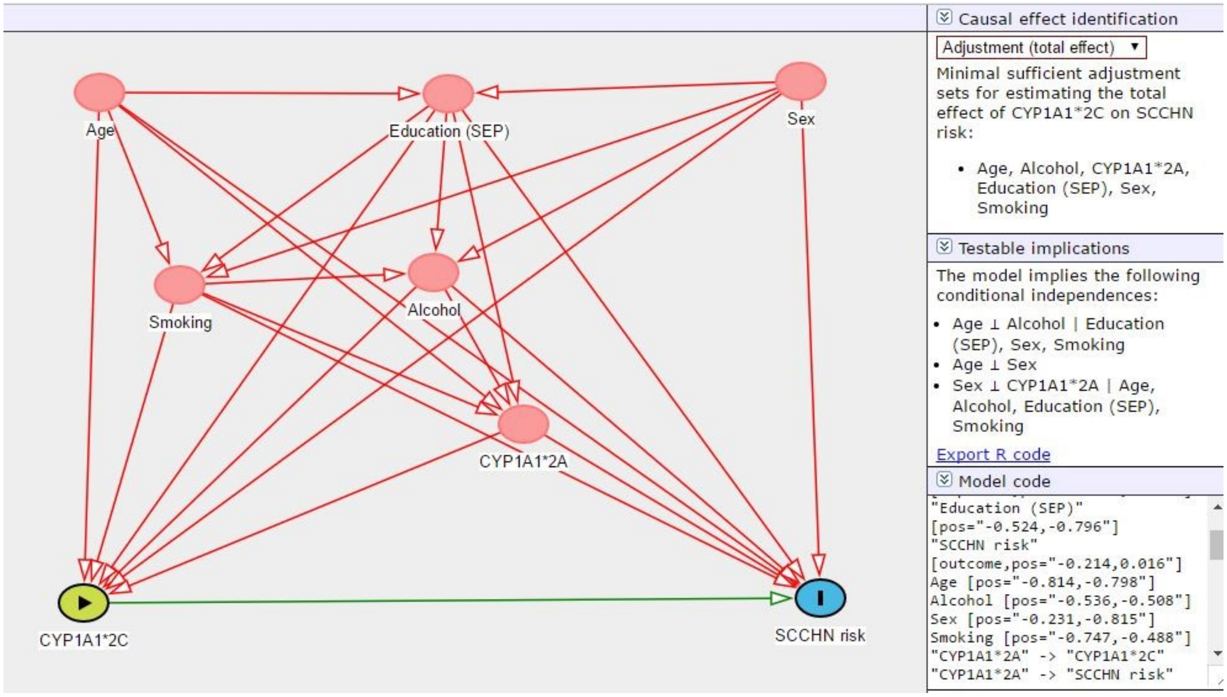
- Age ⊥ Alcohol | Education (SEP), Sex, Smoking
- Age ⊥ Sex
- CYP1A1*2C ⊥ Sex | Age, Alcohol, Education (SEP), Smoking

Export R code

Model code

```
"Education (SEP)"
[pos="-0.524,-0.796"]
"SCCHN risk"
[outcome,pos="-0.214,0.016"]
Age [pos="-0.814,-0.798"]
Alcohol [pos="-0.536,-0.508"]
Sex [pos="-0.231,-0.815"]
Smoking [pos="-0.756,-0.486"]
"CYP1A1*2A" -> "SCCHN risk"
"CYP1A1*2C" -> "CYP1A1*2A"
```

Figure A2: Total effect of CYP1A1*2C on SCCHN risk



Causal effect identification

Adjustment (total effect) ▾

Minimal sufficient adjustment sets for estimating the total effect of CYP1A1*2C on SCCHN risk:

- Age, Alcohol, CYP1A1*2A, Education (SEP), Sex, Smoking

Testable implications

The model implies the following conditional independences:

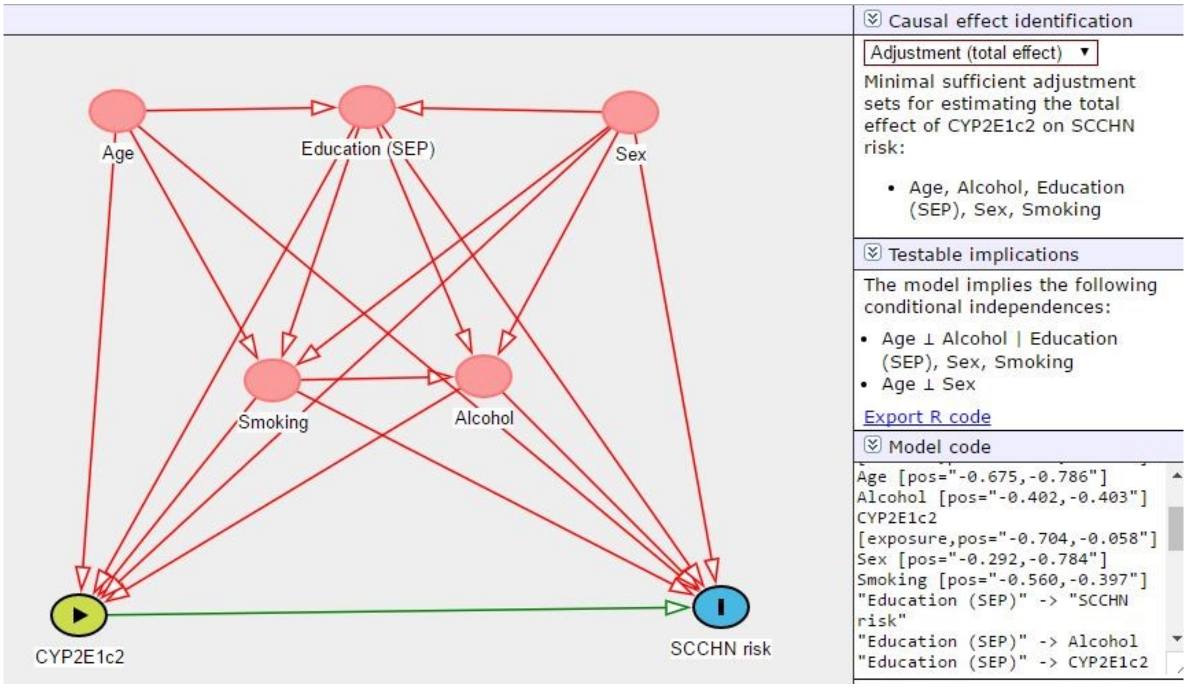
- Age ⊥ Alcohol | Education (SEP), Sex, Smoking
- Age ⊥ Sex
- Sex ⊥ CYP1A1*2A | Age, Alcohol, Education (SEP), Smoking

Export R code

Model code

```
"Education (SEP)"
[pos="-0.524,-0.796"]
"SCCHN risk"
[outcome,pos="-0.214,0.016"]
Age [pos="-0.814,-0.798"]
Alcohol [pos="-0.536,-0.508"]
Sex [pos="-0.231,-0.815"]
Smoking [pos="-0.747,-0.488"]
"CYP1A1*2A" -> "CYP1A1*2C"
"CYP1A1*2A" -> "SCCHN risk"
```

Figure A3: Total effect of CYP2E1c2 on SCCHN risk



Minimal sufficient adjustment sets for estimating the total effect of CYP2E1c2 on SCCHN risk:

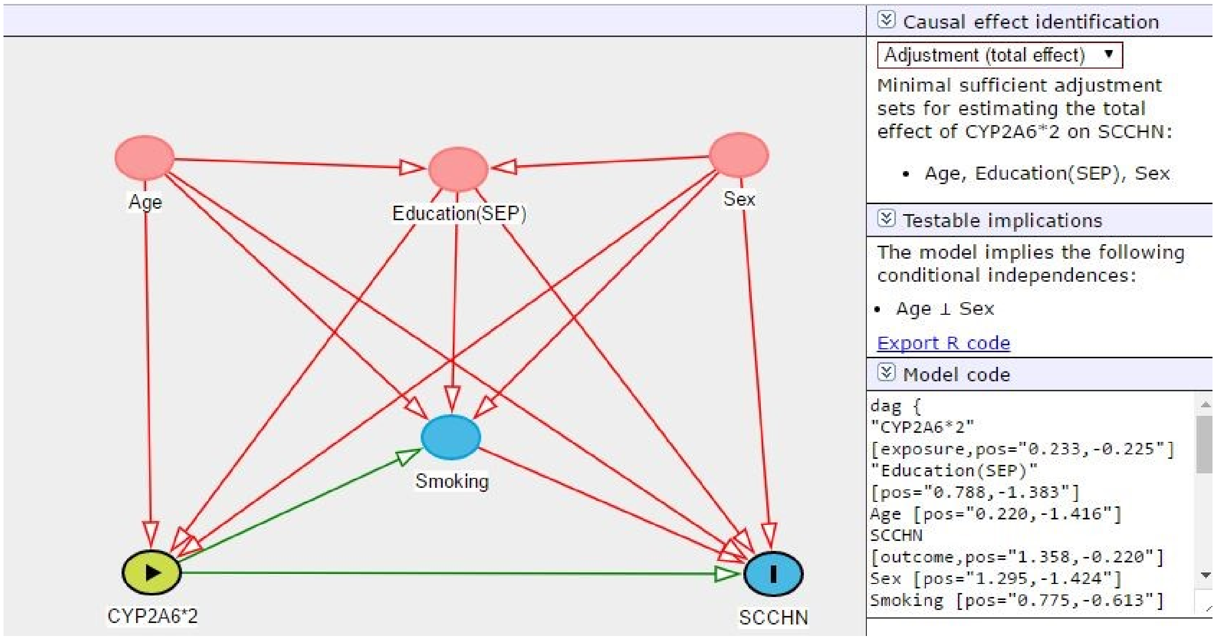
- Age, Alcohol, Education (SEP), Sex, Smoking

The model implies the following conditional independences:

- Age ⊥ Alcohol | Education (SEP), Sex, Smoking
- Age ⊥ Sex

```
Age [pos="-0.675,-0.786"]
Alcohol [pos="-0.402,-0.403"]
CYP2E1c2
[exposure,pos="-0.704,-0.058"]
Sex [pos="-0.292,-0.784"]
Smoking [pos="-0.560,-0.397"]
"Education (SEP)" -> "SCCHN
risk"
"Education (SEP)" -> Alcohol
"Education (SEP)" -> CYP2E1c2
```

Figure A4: Total effect of CYP2A6*2 on SCCHN risk



Minimal sufficient adjustment sets for estimating the total effect of CYP2A6*2 on SCCHN:

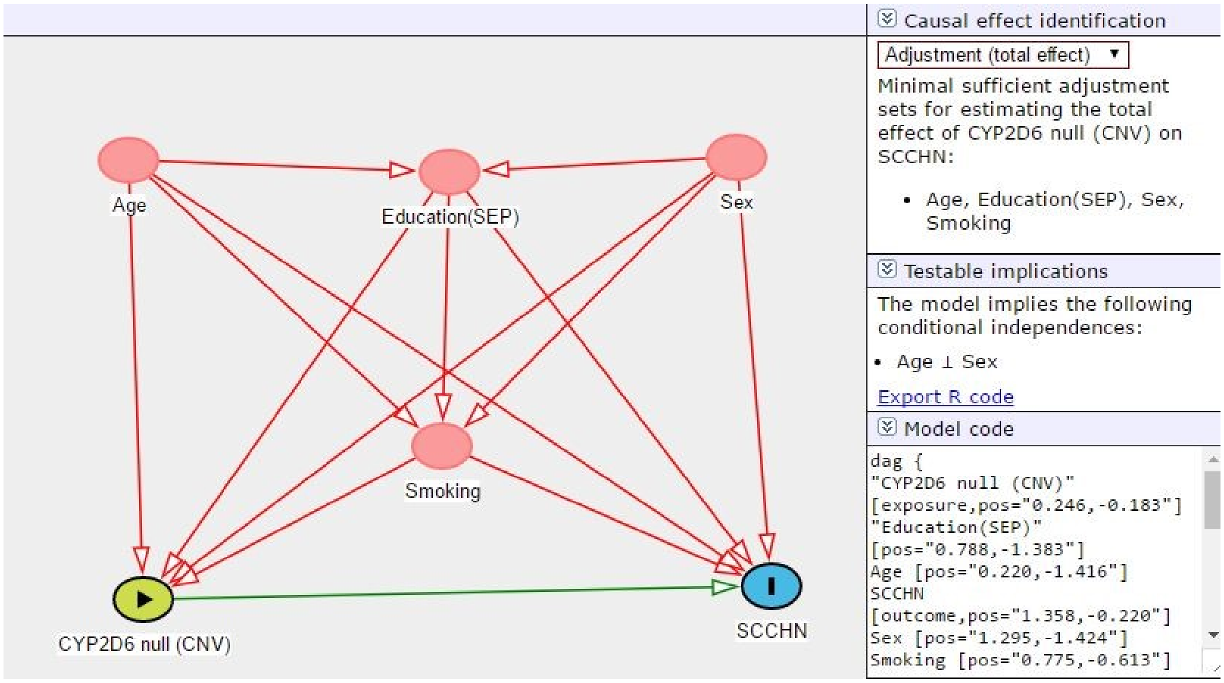
- Age, Education(SEP), Sex

The model implies the following conditional independences:

- Age ⊥ Sex

```
dag {
"CYP2A6*2"
[exposure,pos="0.233,-0.225"]
"Education(SEP)"
[pos="0.788,-1.383"]
Age [pos="0.220,-1.416"]
SCCHN
[outcome,pos="1.358,-0.220"]
Sex [pos="1.295,-1.424"]
Smoking [pos="0.775,-0.613"]
```

Figure A5: Total effect of CYP2D6 null (CNV) on SCCHN risk



Causal effect identification

Adjustment (total effect) ▼

Minimal sufficient adjustment sets for estimating the total effect of CYP2D6 null (CNV) on SCCHN:

- Age, Education(SEP), Sex, Smoking

Testable implications

The model implies the following conditional independences:

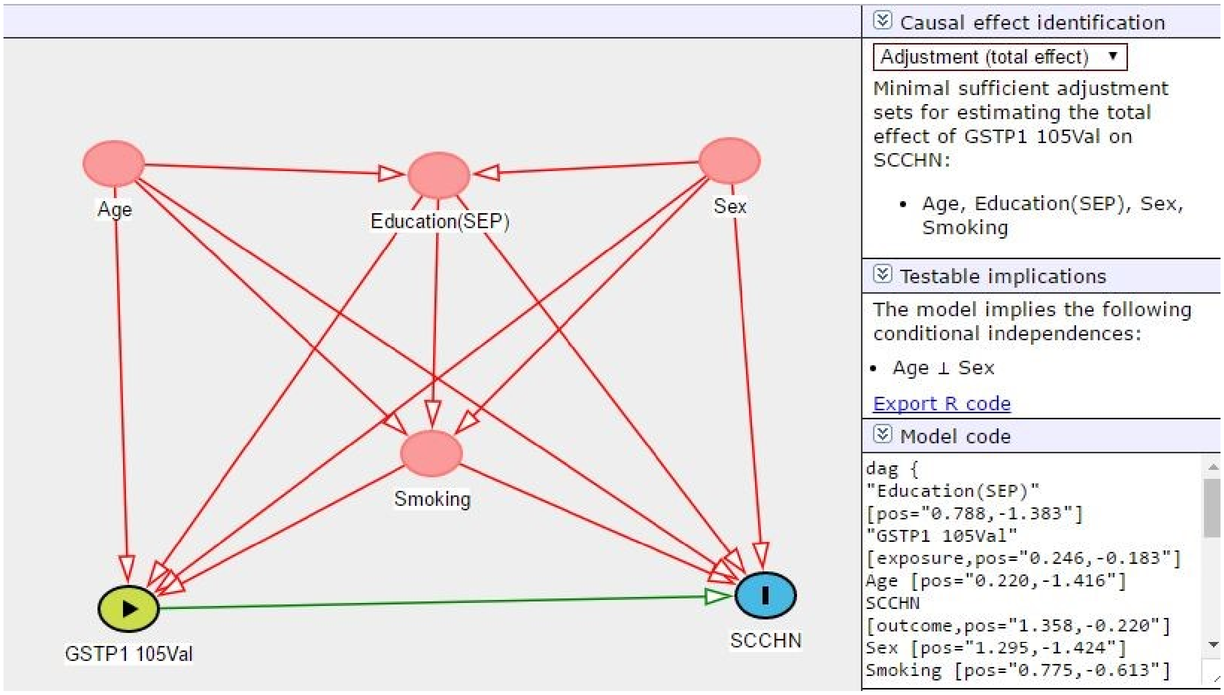
- Age ⊥ Sex

Export R code

Model code

```
dag {
"CYP2D6 null (CNV)"
[exposure,pos="0.246,-0.183"]
"Education(SEP)"
[pos="0.788,-1.383"]
Age [pos="0.220,-1.416"]
SCCHN
[outcome,pos="1.358,-0.220"]
Sex [pos="1.295,-1.424"]
Smoking [pos="0.775,-0.613"]
```

Figure A6: Total effect of GSTP1 105Val on SCCHN risk



Causal effect identification

Adjustment (total effect) ▼

Minimal sufficient adjustment sets for estimating the total effect of GSTP1 105Val on SCCHN:

- Age, Education(SEP), Sex, Smoking

Testable implications

The model implies the following conditional independences:

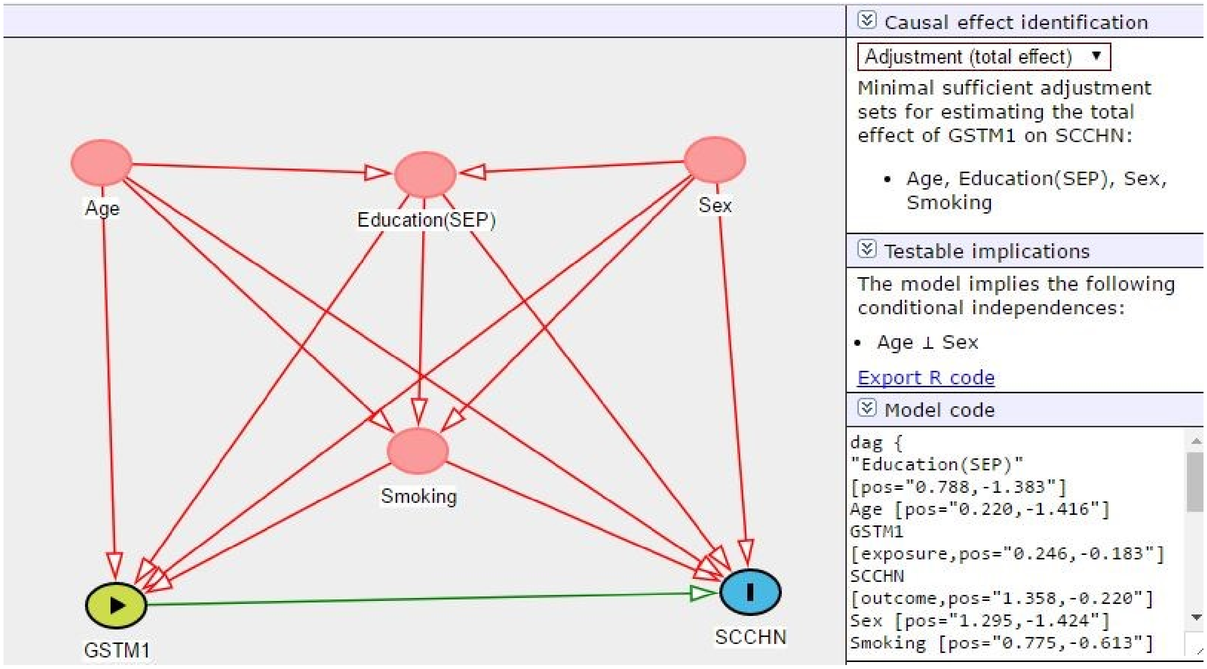
- Age ⊥ Sex

Export R code

Model code

```
dag {
"Education(SEP)"
[pos="0.788,-1.383"]
"GSTP1 105Val"
[exposure,pos="0.246,-0.183"]
Age [pos="0.220,-1.416"]
SCCHN
[outcome,pos="1.358,-0.220"]
Sex [pos="1.295,-1.424"]
Smoking [pos="0.775,-0.613"]
```

Figure A7: Total effect of GSTM1 null on SCCHN risk



Causal effect identification

Adjustment (total effect) ▼

Minimal sufficient adjustment sets for estimating the total effect of GSTM1 on SCCHN:

- Age, Education(SEP), Sex, Smoking

Testable implications

The model implies the following conditional independences:

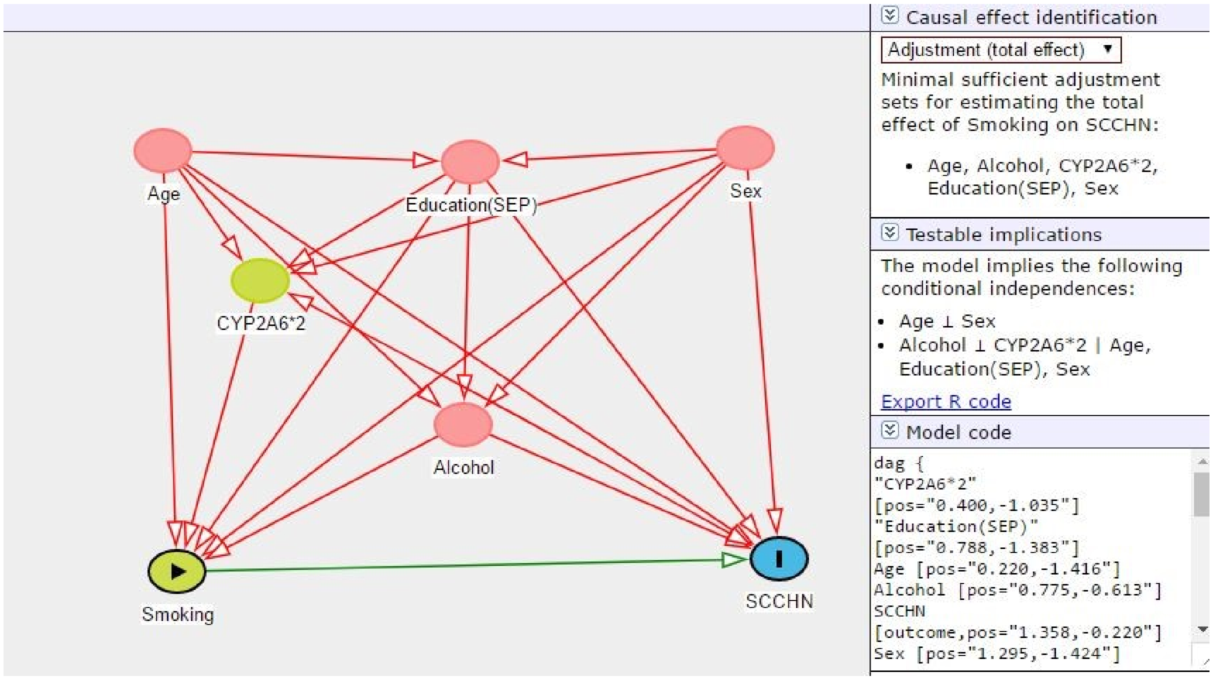
- Age ⊥ Sex

Export R code

Model code

```
dag {
"Education(SEP)"
[pos="0.788,-1.383"]
Age [pos="0.220,-1.416"]
GSTM1
[exposure,pos="0.246,-0.183"]
SCCHN
[outcome,pos="1.358,-0.220"]
Sex [pos="1.295,-1.424"]
Smoking [pos="0.775,-0.613"]
```

Figure A8: Total effect of Smoking on SCCHN risk



Causal effect identification

Adjustment (total effect) ▼

Minimal sufficient adjustment sets for estimating the total effect of Smoking on SCCHN:

- Age, Alcohol, CYP2A6*2, Education(SEP), Sex

Testable implications

The model implies the following conditional independences:

- Age ⊥ Sex
- Alcohol ⊥ CYP2A6*2 | Age, Education(SEP), Sex

Export R code

Model code

```
dag {
"CYP2A6*2"
[pos="0.400,-1.035"]
"Education(SEP)"
[pos="0.788,-1.383"]
Age [pos="0.220,-1.416"]
Alcohol [pos="0.775,-0.613"]
SCCHN
[outcome,pos="1.358,-0.220"]
Sex [pos="1.295,-1.424"]
```

Appendix VI

HeNCe India data vs Census of India 2011: Comparison of assets

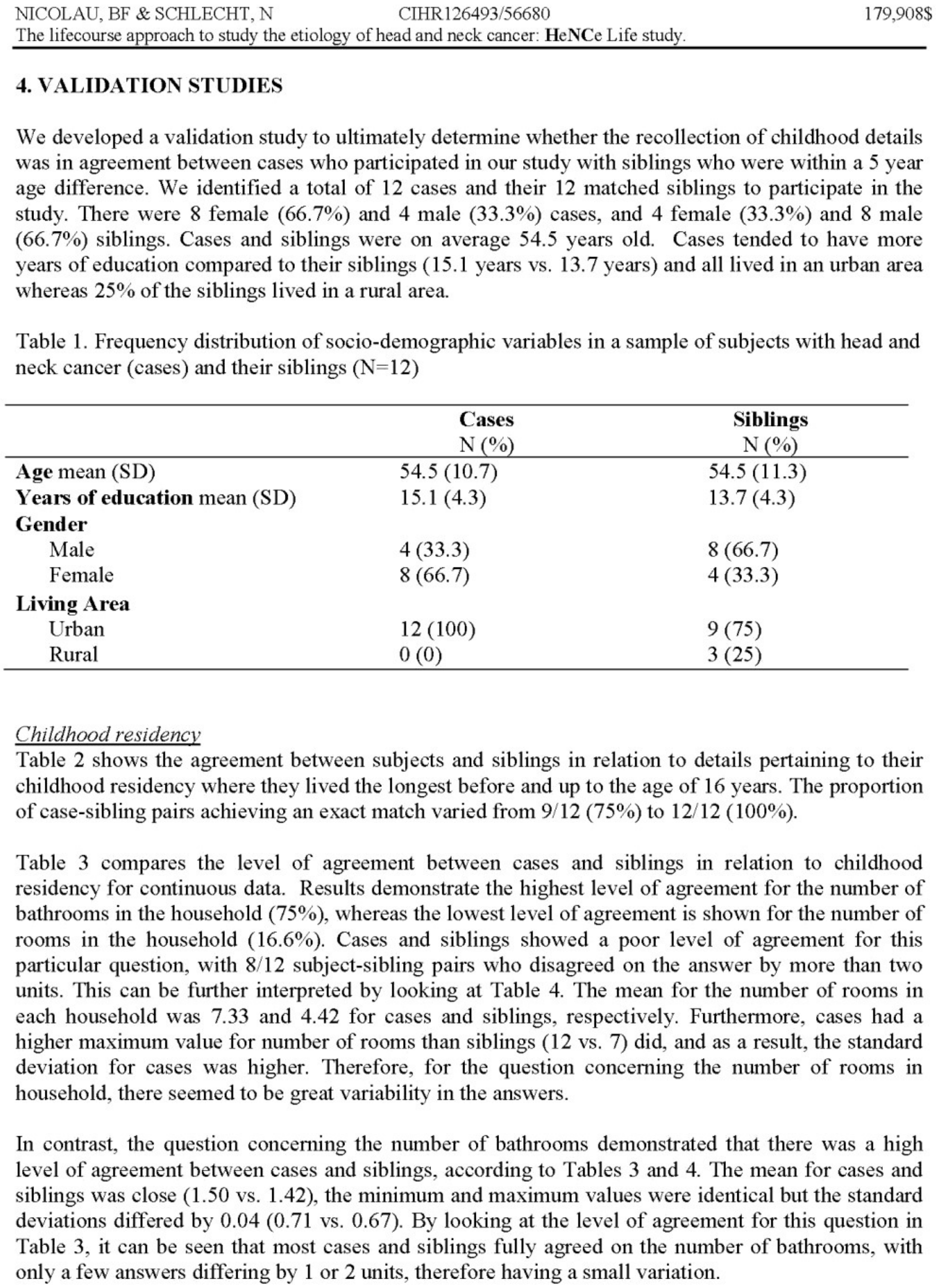
A comparison of percentage distribution of housing assets (longest residence in late adulthood) of controls recruited in HeNCe life study Calicut, India site and available data from the Census of India 2011, Calicut district, Kerala

Housing Assets	**Census of India 2001 Calicut district, Kerala %***	**HeNCe Life study, India site %***
Number of rooms		
1	1.1	2.0
2	7.0	9.0
3	28.0	29.0
3+	34.0	30.0
Water system		
Tap water	21.0	22.0
Well	72.8	74.0
Spring/River/Canal/Tank/Pond	1.8	1.4
Electricity		
Yes	93.8	94.0
No	6.2	6.0
Sanitation		
Septic tank/latrine/slab covered	83.0	90.0
others	17.0	10.0
Kitchen facility		
Yes	97.1	96.0
No	2.7	3.4
No self cooking	0.2	0.5
TV (Present)	71.76	73
Telephone (Present)	78	74
Scooter/motorbike (Present)	25	33
Car/Jeep (Present)	8	11

*Cumulative percentage may not add to 100%. The comparison is for assets whose information was available in both data sets.

APPENDIX VII

Validation study - Canada Site



NICOLAU, BF & SCHLECHT, N CIHR126493/56680 179,908$
The lifecourse approach to study the etiology of head and neck cancer: **HeNCe** Life study.

4. VALIDATION STUDIES

We developed a validation study to ultimately determine whether the recollection of childhood details was in agreement between cases who participated in our study with siblings who were within a 5 year age difference. We identified a total of 12 cases and their 12 matched siblings to participate in the study. There were 8 female (66.7%) and 4 male (33.3%) cases, and 4 female (33.3%) and 8 male (66.7%) siblings. Cases and siblings were on average 54.5 years old. Cases tended to have more years of education compared to their siblings (15.1 years vs. 13.7 years) and all lived in an urban area whereas 25% of the siblings lived in a rural area.

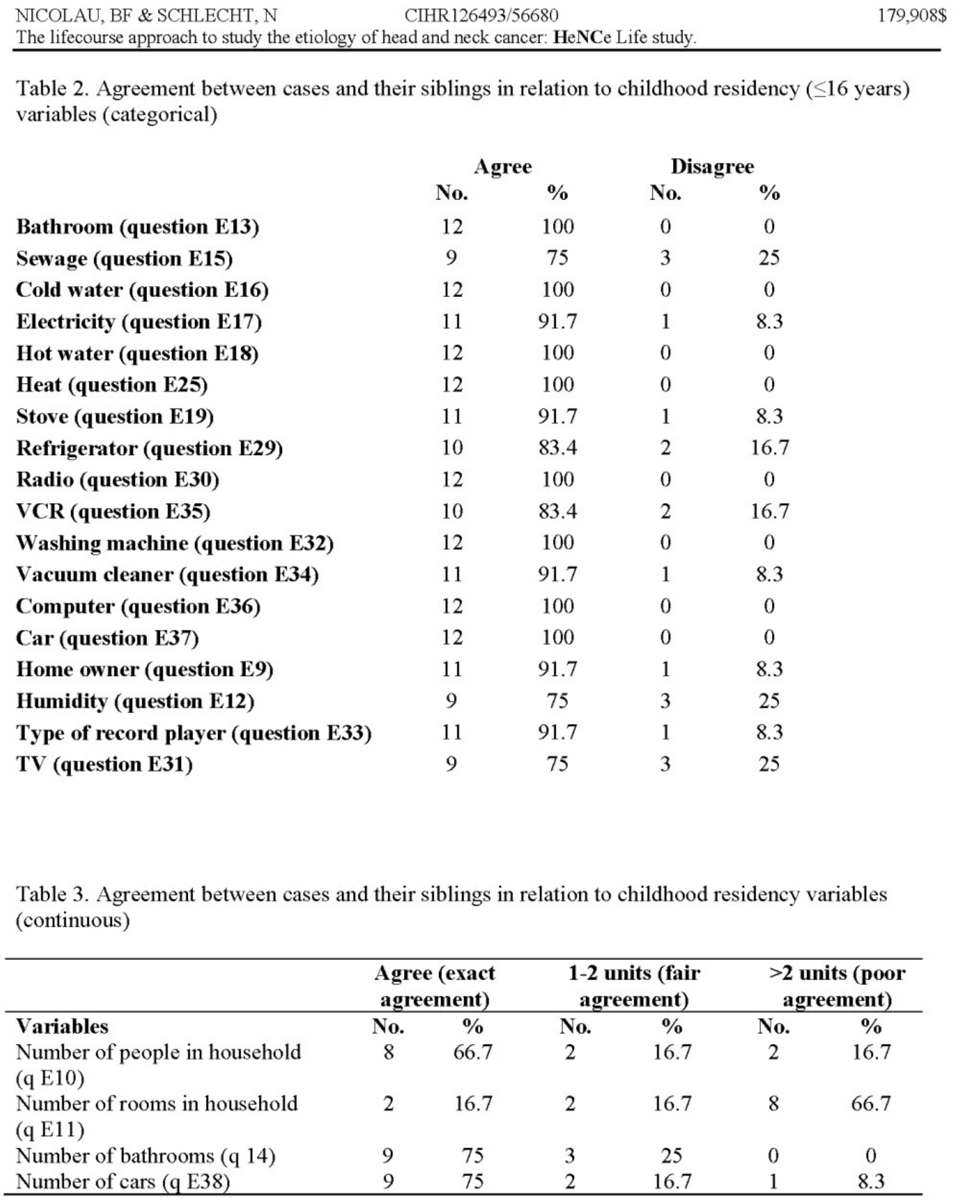
Table 1. Frequency distribution of socio-demographic variables in a sample of subjects with head and neck cancer (cases) and their siblings (N=12)

	Cases N (%)	**Siblings** N (%)
Age mean (SD)	54.5 (10.7)	54.5 (11.3)
Years of education mean (SD)	15.1 (4.3)	13.7 (4.3)
Gender		
Male	4 (33.3)	8 (66.7)
Female	8 (66.7)	4 (33.3)
Living Area		
Urban	12 (100)	9 (75)
Rural	0 (0)	3 (25)

Childhood residency
Table 2 shows the agreement between subjects and siblings in relation to details pertaining to their childhood residency where they lived the longest before and up to the age of 16 years. The proportion of case-sibling pairs achieving an exact match varied from 9/12 (75%) to 12/12 (100%).

Table 3 compares the level of agreement between cases and siblings in relation to childhood residency for continuous data. Results demonstrate the highest level of agreement for the number of bathrooms in the household (75%), whereas the lowest level of agreement is shown for the number of rooms in the household (16.6%). Cases and siblings showed a poor level of agreement for this particular question, with 8/12 subject-sibling pairs who disagreed on the answer by more than two units. This can be further interpreted by looking at Table 4. The mean for the number of rooms in each household was 7.33 and 4.42 for cases and siblings, respectively. Furthermore, cases had a higher maximum value for number of rooms than siblings (12 vs. 7) did, and as a result, the standard deviation for cases was higher. Therefore, for the question concerning the number of rooms in household, there seemed to be great variability in the answers.

In contrast, the question concerning the number of bathrooms demonstrated that there was a high level of agreement between cases and siblings, according to Tables 3 and 4. The mean for cases and siblings was close (1.50 vs. 1.42), the minimum and maximum values were identical but the standard deviations differed by 0.04 (0.71 vs. 0.67). By looking at the level of agreement for this question in Table 3, it can be seen that most cases and siblings fully agreed on the number of bathrooms, with only a few answers differing by 1 or 2 units, therefore having a small variation.



NICOLAU, BF & SCHLECHT, N CIHR126493/56680 179,908$
The lifecourse approach to study the etiology of head and neck cancer: **HeNC**e Life study.

Table 2. Agreement between cases and their siblings in relation to childhood residency (≤16 years) variables (categorical)

	Agree		Disagree	
	No.	**%**	**No.**	**%**
Bathroom (question E13)	12	100	0	0
Sewage (question E15)	9	75	3	25
Cold water (question E16)	12	100	0	0
Electricity (question E17)	11	91.7	1	8.3
Hot water (question E18)	12	100	0	0
Heat (question E25)	12	100	0	0
Stove (question E19)	11	91.7	1	8.3
Refrigerator (question E29)	10	83.4	2	16.7
Radio (question E30)	12	100	0	0
VCR (question E35)	10	83.4	2	16.7
Washing machine (question E32)	12	100	0	0
Vacuum cleaner (question E34)	11	91.7	1	8.3
Computer (question E36)	12	100	0	0
Car (question E37)	12	100	0	0
Home owner (question E9)	11	91.7	1	8.3
Humidity (question E12)	9	75	3	25
Type of record player (question E33)	11	91.7	1	8.3
TV (question E31)	9	75	3	25

Table 3. Agreement between cases and their siblings in relation to childhood residency variables (continuous)

	Agree (exact agreement)		1-2 units (fair agreement)		>2 units (poor agreement)	
Variables	**No.**	**%**	**No.**	**%**	**No.**	**%**
Number of people in household (q E10)	8	66.7	2	16.7	2	16.7
Number of rooms in household (q E11)	2	16.7	2	16.7	8	66.7
Number of bathrooms (q 14)	9	75	3	25	0	0
Number of cars (q E38)	9	75	2	16.7	1	8.3

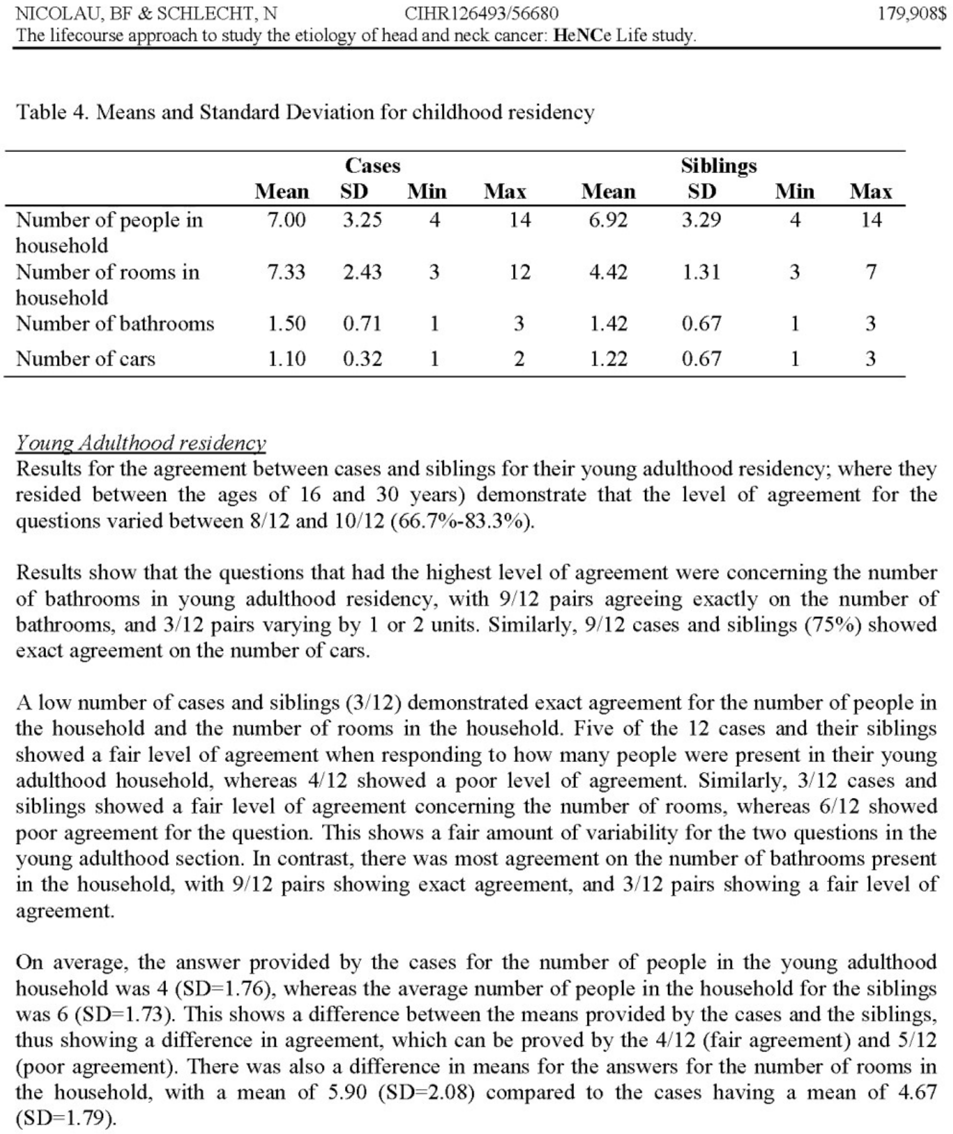


Table 4. Means and Standard Deviation for childhood residency

	Cases				Siblings			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Number of people in household	7.00	3.25	4	14	6.92	3.29	4	14
Number of rooms in household	7.33	2.43	3	12	4.42	1.31	3	7
Number of bathrooms	1.50	0.71	1	3	1.42	0.67	1	3
Number of cars	1.10	0.32	1	2	1.22	0.67	1	3

Young Adulthood residency

Results for the agreement between cases and siblings for their young adulthood residency; where they resided between the ages of 16 and 30 years) demonstrate that the level of agreement for the questions varied between 8/12 and 10/12 (66.7%-83.3%).

Results show that the questions that had the highest level of agreement were concerning the number of bathrooms in young adulthood residency, with 9/12 pairs agreeing exactly on the number of bathrooms, and 3/12 pairs varying by 1 or 2 units. Similarly, 9/12 cases and siblings (75%) showed exact agreement on the number of cars.

A low number of cases and siblings (3/12) demonstrated exact agreement for the number of people in the household and the number of rooms in the household. Five of the 12 cases and their siblings showed a fair level of agreement when responding to how many people were present in their young adulthood household, whereas 4/12 showed a poor level of agreement. Similarly, 3/12 cases and siblings showed a fair level of agreement concerning the number of rooms, whereas 6/12 showed poor agreement for the question. This shows a fair amount of variability for the two questions in the young adulthood section. In contrast, there was most agreement on the number of bathrooms present in the household, with 9/12 pairs showing exact agreement, and 3/12 pairs showing a fair level of agreement.

On average, the answer provided by the cases for the number of people in the young adulthood household was 4 (SD=1.76), whereas the average number of people in the household for the siblings was 6 (SD=1.73). This shows a difference between the means provided by the cases and the siblings, thus showing a difference in agreement, which can be proved by the 4/12 (fair agreement) and 5/12 (poor agreement). There was also a difference in means for the answers for the number of rooms in the household, with a mean of 5.90 (SD=2.08) compared to the cases having a mean of 4.67 (SD=1.79).