Development and validation of a multivariable prediction model for all-cause cancer incidence based on health behaviours in the population setting

Courtney Maskerine

A thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the
MSc degree in Epidemiology

Epidemiology and Community Medicine
Faculty of Medicine
University of Ottawa

TABLE OF CONTENTS:

ABSTRACT:

Background: We examined if it was possible to use routinely available, self-reported data on health behaviours to predict incident cancer cases in the Ontario population.

Methods: This retrospective cohort study involved 43 696 female and 36 630 male respondents from Ontario, who were >20 years old and without a prior history of cancer, to the Canadian Community Health Survey (CCHS) cycles 2.1-4.1. The outcome of interest was malignant cancer from any site, termed all-cause cancer, determined from the Ontario Cancer Registry. Predictor variables in the risk algorithm were health behaviours including smoking status, pack-years of smoking, alcohol consumption, fruit and vegetable consumption and physical activity level. A competing-risk Cox proportional hazard model was utilized to determine hazard of incident cancer. The developed risk prediction tool was validated in the CCHS cycle 1.1 on 14 426 female and 11 970 male survey respondents.

Results: Incident cancer was predicted with a high degree of calibration (differences between observed and predicted values for females 2.97%, for males 4.23%) and discrimination (C-statistic: females 0.76, males 0.83). Similar results were obtained in the validation cohort.

Conclusions: Routinely collected self-reported information on health behaviours can be used to predict incident cancer in the Ontario population. This type of risk prediction tool is valuable for public health purposes of estimating population risk of incident cancer, as well as projection of future risk in the population over time.

CHAPTER 1: INTRODUCTION

Cancer risk prediction is an area of increasing interest among clinicians, policy-makers and the public. There are several available risk prediction models aimed at assessing the risk of future cancer development. The majority of these models are derived for site-specific cancers in a clinical rather than population-based setting with few, if any, modifiable behavioural factors included. Despite the paucity of risk prediction tools in the area of all-cause cancer development, the existing literature supports the feasibility of developing a risk prediction model using population-derived information on behavioural factors for determination of future risk of all-cause cancer development.

1.1 Cancer Risk Prediction

There is a vast array of risk prediction algorithms available for determining risk of specific cancers, including breast cancer, prostate cancer and colorectal cancer.[1–3] The majority of these have been developed using clinical determinants for individual risk of cancer. A review of prostate cancer prognostic tools reveals that they are exclusively developed for use in the clinical setting, with no modifiable lifestyle risk factors included in the models.[1] This is problematic from a population approach because of the lack of focus on modifiable exposures or behaviours, and because clinical variables are difficult to measure on a large population-based scale.[4] Furthermore, if a lifestyle-based index were as predictive or more predictive, at a population level, than a clinically-based index then it would not warrant the resource-intense collection of clinical data for a population risk algorithm.

There are some risk prediction models which include lifestyle factors, among them a study completed in Denmark which investigated the impact of adherence to lifestyle recommendations for physical activity, waist circumference, smoking, alcohol intake and diet on the risk of colorectal cancer.[3] This study involved 55 487 individuals aged 50-64 years old identified by the Civil Registration System in Denmark who responded to a detailed lifestyle questionnaire and completed anthropometrical measurements. Cancer diagnoses were identified from the Central Population Registry or the Danish Pathology Databank over a period of up to 13 years. Participants were categorized into groups with zero-one point on the lifestyle index, 2 points, 3 points, or 4-5 points, with a greater number of points meaning greater concordance with recommendations. Investigators determined that having 4-5 points on the lifestyle index was associated with a 30% reduction in risk of colorectal cancer (incidence rate ratio 0.70 (95% CI: 0.53-0.93)). It was estimated that if each participant had complied with just one additional recommendation, there would be a 13% reduction in incidence of colorectal cancer in the population. This study illustrates the important contribution of modifiable lifestyle factors in the incidence of specific causes of cancer, and their potential utility in predictive models at the population level.

Lifestyle factors, such as excess alcohol consumption, poor diet, smoking cigarettes and inadequate physical activity, are potentially modifiable risk factors for a vast array of health conditions, including cardiovascular disease (CVD) and many cancers.[5–7] There are widely used indices that assess the impact of lifestyle factors on CVD, such as the Framingham heart study which incorporates clinical and lifestyle variables to determine incidence of cardiovascular disease.[7] There are few similar algorithms to determine the

risk of all-cause cancer. This is despite evidence of a high attributable fraction of smoking, alcohol use, low fruit and vegetable intake, physical inactivity, and being obese or overweight, to incidence[6] and mortality[5,6] from many common cancers worldwide.

In a review of the literature, several studies were identified that assessed behavioural risk factors for cancer at any site, also termed all-cause cancer. These studies are reviewed in the following descriptions.

A recent publication assessed the incidence of cancer on the basis of meeting American Heart Association (AHA) guidelines for ideal cardiovascular (CV) health metrics.[8] The CV health metrics included baseline smoking status, body mass index, diet, physical activity level, cholesterol level, blood pressure, and fasting serum glucose. Participants were aged 45-64 years from four communities in the United States. The results indicated that in those with only one ideal health metric the hazard ratio (HR) of cancer was 0.79 (95% CI: 0.64-0.98) as compared to those with no ideal health metrics, while those with 6- 7 of the ideal health metrics had a significantly lower HR for cancer at 0.49 (95% CI: 0.35- 0.69). The trend for a decreased risk in cancer with greater number of health metrics was significant ($p<0.0001$) after adjusting for age, sex, race and study site. This study indicates that those with 6-7 ideal values for the studied health metrics had a greater than 50% reduction in risk of cancer with respect to those with none of these health metrics.[8]

The main limitation of this study was that it used a cohort initially designed to evaluate cardiovascular disease, and therefore included in the predictor variables both health behaviours and clinical health metrics. As the investigators attest, there was a pattern of

decreased incident cancer in those with more ideal health behaviours, but no consistent pattern for those with ideal clinical health factors.[8] This would be expected, as blood pressure, cholesterol and glucose measurement would be less likely to be associated with malignancy development. Furthermore, this study used only categorical predictor variables, again defined on the basis of cardiovascular risk. For example, participants were classified as non-smokers if they had never smoked or quit >12 months prior to study baseline. Given that smoking is a significant risk factor for several malignancies[5,6] it would be important to assess a continuous measure of smoking exposures or alternatively to use multiple categories that account for differences between never smoking, having previously quit and being a current smoker. Any participants who lacked information on all of the 7 ideal health metrics were excluded from the analysis, potentially introducing bias or reducing the statistical power. Finally, this study did not include any measures of model performance, nor was the model validated on either internal or external data sources.[8]

A study conducted by the European Prospective Investigation into Cancer and Nutrition (EPIC) developed a risk prediction score for incident cancer based on concordance with American Institute of Cancer Research (AICR)/World Cancer Research Fund (WCRF) recommendations on lifestyle factors.[9] The study was large, with 386 355 participants across 9 European countries. Prognostic factors included weight management, physical activity, consumption of foods and drinks that promote weight gain, consumption of plant foods, consumption of animal foods, consumption of alcoholic beverages and, in women, breastfeeding. Prognostic scores were developed by dichotomizing risk factors into categories of adherent or non-adherent to guidelines, then generating a score ranging

from 0-6 for men and 0-7 for women, with higher scores indicating greater concordance with recommendations. Cox regression models were developed to determine the association between risk scores and outcomes. This study showed that compared to those with the lowest scores (0-2 in men/0-3 in women), individuals with the highest scores (5-6 in men/6-7 in women) had an overall hazard ratio of 0.83 (95%CI: 0.75-0.90) for incident cancer. A one- point increase in predictor score (indicating better adherence to lifestyle recommendations) was associated with a 5% (95% CI: 3-7%) decrease in risk of cancer development.[9] Smoking status was not included in the risk score, but was examined and found not to be an effect modifier.

This study was very large and well powered to address the question of diet and activity effects on cancer development. Its limitations include that the predictive score was based on dichotomization of underlying variables- either adherent or non-adherent to lifestyle recommendations- without further categorization based on degree of adherence. The study was conducted in European countries, but was not population-based. It used a convenience sample consisting of a combination of participants from health insurance plans, blood donors, employees of several enterprises, civil servants, participants enrolled in mammogram screening, and "health conscious" subjects of whom many were vegetarian, as well as members of the general population.[9] Given the study population, the results may not be generalizable to the European population or the Canadian population as a whole, and may underestimate the association between lifestyle factors and cancer incidence due to the "healthy volunteer" effect. In addition, several groups of participants were excluded including those who did not complete all the diet or lifestyle questions, or those with missing data on weight, physical activity level or, in women,

breastfeeding status which may lead to misleading results. No measures of model performance or validation were included in the study.[9]

Similar to the two studies described above, other studies have investigated the relationship between adherence to cancer prevention guidelines (American Cancer Society[10,11] or a combination of French nutrition guidelines and WHO guidelines[12]) and incident cancer development. Two of the studies were based on populations of women only[10,12] and were conducted on a more limited age range than the general population (range of ages: 43-79).[10–12] Their common limitation, similar to the previously described studies, is in their creation of risk categories on a numeric scale based on the true underlying values of predictor variables, which reduces the information about the relationship between predictor variables and outcome. In addition, methodological challenges exist such as the exclusion of participants with any missing data on lifestyle factors,[10–12] definition of predictor categories (for example, categorizing physical activity based on frequency of performance rather than a measure of intensity[12]) and lack of reporting on model performance or validation.[10–12] The use of guidelines to create a score indicating adherence or non-adherence to lifestyle-related behaviours also limits generalizability; with changes or differences in guideline recommendations across time and geographic areas it is more difficult to utilize the score in future or in different locations. Nonetheless, all of the above studies found an association between greater adherence to lifestyle-related cancer prevention guidelines and reduction in all-cause cancer incidence (HR and 95% CI for highest vs. lowest scores ranged from HR=0.90 (95%CI=0.87-0.93)[11]; HR=0.83 (95%CI=0.75-0.92)[10] and HR=0.81 (95%CI=0.73-0.89)[12].

A study from Japan used an approach involving estimation of survival free from cancer or cardiovascular disease in individuals with varying lifestyle risk factors. This study investigated smoking status, alcohol intake and body mass index as predictor variables based on survey data collected in a cohort of 96 592 individuals aged 40-69 years old in the Japan Public Health Center-based Prospective Study. This study found that avoiding smoking, excess alcohol intake and elevated BMI increased disease-free survival to varying levels dependent upon age, but for example in 50-54 year olds the probability of disease-free survival in those ascribing to all healthy lifestyle factors increased by approximately 6% in women and 10% in men relative to those without healthy behaviours. This study was limited in that it only included three potential predictor variables, and did not include possible confounding variables such as sociodemographic factors. In addition, the methods for selection of candidate variables (including interaction terms) included backward selection, and authors compared 8 different models using cross validation techniques to find the model with the best fit based on minimizing differences between observed and predicted risks. No measures of model performance were reported and the model was not validated on internal or external data sets.[13]

Another tool for predicting cancer incidence was developed at Harvard using a method in which an expert consensus group identified risk factors for the ten most prevalent cancers in men and thirteen most prevalent cancers in women in the United States.[14] The causes of cancer were then categorized, based on opinion, into definite, probable or possible risk factors to give an indication of the strength of the evidence, and any that were considered only possible were excluded. Magnitude of risk was based on relative risk assessments

and the team developed five categories of relative risk (none=RR 0.9-1.19, weak=RR 1.2- <1.5, moderate= RR 1.5- <3.0, strong=RR 3.0- <7.0, very strong=RR ≥7). These relative risk categories were then translated into a scale and assigned numeric values (points). Subsequently, these risks were compared with the US population average, and through addition or subtraction each individual would be able to calculate their number of risk points compared with the US population average. This, in turn, would then be divided by the population average giving a relative score in 5 categories ranging from very low to very high risk over the subsequent 10 year period. To estimate total cancer risk, the risk points for an exposure were weighted based on the proportion of cancer contributed by each site.

The strength of this study was in the attempt to evaluate the risk factors that are implicated in cancer development in a quantitative manner. However, there are several limitations. Only the most common cancers accounting for 80% of cancer incidence were included in the study.[14] The study is based on the use of relative risks from the literature that are derived from a variety of populations and types of studies, as opposed to directly obtained from the population of interest, thereby possibly leading to over- or under-representation of risk. Furthermore, given the large number of steps to derive cancer risk there is likely measurement error in the estimates. Interactive effects were not examined, and as it is known that certain exposures can act synergistically[15], this may also alter the estimates of cancer risk. Other possible confounding factors were not included in the risk algorithm, such as ethnicity or income level, which may be a marker of increased vulnerability to exposures.[16] In summary, the Harvard study used literature on risk factors

present in the year 2000 to derive an estimate of cancer risk based on group consensus. It is therefore useful in estimating risk only at that time period given that the risk estimates are static and it would have limited generalizability in different populations as the risks are defined in the U.S. population alone. The reported validation component of this study assessed whether the risk index was useful in predicting colon cancer in an external dataset, and concluded that it had "good" performance despite there being a greater than 20% difference in the relative risk estimates in certain categories examined.[17] This risk algorithm is the most highly cited of the risk algorithms described for all-cause cancer incidence, clearly indicating the need for further more methodologically rigorous studies to be developed assessing concrete data within a specific population of interest.

1.2 Use of Risk Algorithms for Prediction:

Risk algorithms allow for the inclusion of important baseline characteristics, other than age and sex, which influence development of a given outcome. These algorithms are capable of improving predictive accuracy when compared to traditional methods of risk estimation.[4] Traditional methods include the use of Levin's method or the population attributable fraction (PAF) for estimating the proportion of disease cases that are preventable if a risk factor for that disease were eliminated.[18] The equation utilizes the relative risk of disease based on a specific risk factor, as determined in the literature, and the prevalence of the disease. One concern with the PAF is that for conditions with multiple risk factors, such as cancer or coronary artery disease, the PAF can be greater than 1, indicating that more than 100% of cases are preventable by the removal of risk factors.[18] A potential explanation for this phenomenon is that there are multiple risk

factors interacting to form effects in a more complex manner than each individual risk factor alone. The benefits of creating a predictive risk algorithm include the ability to assess the combined impact of multiple risk factors in the same model, as well as using the data to determine risk in the same population as it is being studied, rather than using estimates from literature on different populations to determine relative risks.

Investigations at the Institute for Clinical Evaluative Studies (ICES) have shown that risk prediction using health behaviours is possible with a high degree of accuracy. The DPoRT risk algorithm used health behaviours in the prediction of diabetes, and the life-expectancy algorithm used health behaviours in the prediction of mortality, both with high discrimination (c-statistic 0.77-0.80, and 0.87, respectively) and calibration (<20% difference between observed and predicted estimates).[19,20]

1.3 Optimizing Performance of Risk Prediction Models:

Literature suggests that the overall quality of reporting and methodological quality for multivariable prediction models is poor.[21–23] In a 2008 study reviewing predictive risk algorithms in six high-impact general medical journals, deficiencies were found in selection of candidate predictor variables, handling missing data, reporting of results and model performance in terms of internal and external validation.[23]

Selection of variables based on significant univariate associations between predictor variables and outcomes, or by using backward or forward selection in multivariable analyses, can lead to overfitting and unstable models.[23] Bouwmeester found that in

models using multivariable predictor selection, 19% of studies did not report the method of selection, and 23% of studies with the primary-aim being prediction used either backward or forward selection. A large proportion of these studies used a p-value cutoff of <0.05 as the level at which a predictor was selected into the model. In addition, many continuous predictors were dichotomized in studies, despite the loss of information that occurs with this process.[23]

Handling missing values is another important consideration, specifically addressing any missing values that are not completely at random. Exclusion of these participants can lead to a loss in statistical power as well as misleading or biased results.[23] Description of how missing values were handled was unavailable or unclear in 38% of studies from the literature review. The majority of studies included only those participants with complete data. Multiple imputation methods were only used in 8% of the included studies.[23]

In terms of assessing model performance, only 15% of studies reported a method of calibration, and 27% of studies reported the most common measure of discrimination, a C-statistic.[23] Measures of overall discrimination and calibration, including the Brier score and R^2 value, were reported in less than 10% of studies. External validation, in which a model's predictive performance is assessed in a data set other than the data used to derive the prognostic algorithm, was completed in only 6% of studies.[23] The authors conclude that the poor reporting and methods in many of the identified studies, which were drawn from high-impact journals presumably with more stringent quality guidelines than others, limits the reliability and applicability of the studies findings.

Similar issues were identified in a systematic review of prognostic models aimed at predicting patient outcomes in cancer research.[22] A particular consideration noted in this study was the practice of classifying predictor variables into dichotomous risk groups. Authors cite the importance of identifying individuals at intermediate risk, as these individuals are the ones in whom the benefit of a treatment decision or behavioural modification is less clear than in those at either high or low risk. The creation of risk groups occurred in the majority (76%) of included studies.[22] Options exist to improve the definition of predictor variables, including increased categorization (>3 risk groups) or the maintenance of continuous variables so as to maintain any information relating to how risk changes at different levels of a prognostic variable.

1.4 Reporting of Prediction Models:

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement was published in 2015 and outlines methods for clear reporting of information on prediction models. This statement calls for improved reporting, citing that reviews of multivariable prediction model studies "have shown that serious deficiencies in the statistical methods, use of small data sets, inappropriate handling of missing data, and lack of validation are common." Authors further contend that this may limit the utility of the prediction models being developed and account for the relatively few prediction models used in practice compared to the number of models published in the literature.[21] The TRIPOD statement established a 22-item checklist for

improved reporting of multivariable prediction models, including improved reporting of model development and specification, and reports on model performance.[21]

1.5 Methodological Approaches to Risk Prediction Models:

A methodological framework for developing multivariable prognostic models has been proposed by Harrell[24,25] and is the framework this study has followed to optimize prediction capabilities in the regression model. Harrell describes the necessity for pre-specification of predictor variables and interaction terms, the use of data reduction methods to simplify predictive models so that they predict more accurately in new (validation) sets of patients, the appropriate use of piecewise cubic polynomials (spline functions) to allow for modeling non-linear relationships between exposures and outcomes, and the importance of using methods to describe predictive accuracy of the model, particularly in validation cohorts.[24,25]

1.6 Study Objectives and General Approach:

The overall goal of this study was to examine the importance of a combination of self-reported behavioural factors, including smoking, activity level, dietary fruit and vegetable consumption and alcohol consumption, with respect to the risk of developing cancer at any site, among the Ontario population. Population-based risk algorithms such as this are useful in policy-making and health planning for three purposes: to describe the level of risk within a population and whether that risk is diffused among multiple individuals or concentrated among a select section of the population; projecting the number of new cases of disease over time; and evaluating public health interventions, particularly those

aimed at risk factor reduction.[4] In this study, we examined the hypothesis that the incidence of all-cause cancer can be accurately and discriminately estimated with the use of Canadian Community Health Survey (CCHS) data on health behaviours.

This study's strengths include the utilization of easily measured variables allowing for reproducibility in other populations, the use of a multivariable model adjusting for potential confounders in the relationship between exposures and outcome, and the use of a large cohort reflective of the population of interest in Ontario. Rigorous methodology employed in this study includes justification of predictor variable and model specification, as well as the use of multiple measures of model validation in an external cohort. In addition, this study ascribes to the TRIPOD statement for improved reporting in multivariable prognostic models. The focus on all-cause incident cancer allows for practical interpretation in public health of the overall role of health behaviours on risk of developing cancer.

CHAPTER 2: METHODS

2.1 Study Design:

This was a retrospective cohort study of all Ontario household respondents to the Canadian Community Health Survey (CCHS). The CCHS is a cross-sectional survey that collects information related to health status, health care use, and health determinants among Canadians. It uses a multistage stratified cluster design to sample the Canadian population (>12 years of age). Specifically, a sample is allocated among the provinces based on their size and the number of health regions within the province, after which each province's sample is proportionally allocated among the health regions within the province. Samples are obtained mainly from households selected within specified geographic areas and from lists of telephone numbers, with the remaining <1% obtained through random digit dialing. Interviews are conducted via telephone or in-person encounters, and all responses are self-reported.[26]

For the derivation cohort, we included respondents to CCHS cycle 2.1 (2003), 3.1 (2005) and 4.1 (2007); for the validation cohort we used CCHS cycle 1.1 (2001). The validation cohort was selected to ensure the algorithm was validated on an external cohort, rather than being internally validated using a split sample or bootstrapping. The follow-up period for detecting the outcome of interest, incident cancer, was 5 years following survey administration.

2.2 Participants:

2.2.1 Inclusion criteria:

Respondents from the CCHS survey 1.1, 2.1, 3.1 or 4.1 in Ontario were included if they were over 20 years of age, and had consented to have their survey responses linked to health care data.

2.2.2 Exclusion criteria:

Respondents with a pre-existing diagnosis of malignant cancer in the Ontario Cancer Registry (OCR) or self-reported prior cancer in the CCHS were excluded from the analysis. Any individuals that were not eligible for Ontario Universal Health Insurance (OHIP) were also excluded from the study. A further description of methods used to derive these exclusion criteria is presented in Appendix 1.

2.2.3 Study Population:

The final study population was derived through a stepwise process in which those with no OHIP insurance, those under 20 years of age, and those with a previous diagnosis of malignant cancer in either the CCHS or OCR were excluded. Individuals with identical encrypted health card numbers (IKNs) as well as having identical values for all other variables were excluded. Subsequently, the population was divided into the derivation cohort (CCHS cycle 2.1-4.1 respondents) and the validation cohort (CCHS cycle 1.1 respondents). Males and females were categorized as separate populations for analysis due to differences in the most predominant cancers between sexes.

2.3 Outcome:

The primary outcome measure of interest, cancer incidence from all sites, was determined from Ontario Cancer Registry (OCR) data. All-cause cancer incidence was defined as a primary malignant cancer diagnosis in an individual, regardless of cancer site. Primary malignancies were identified from the OCR, which uses the International Classification of Disease version 9 (ICD-9) code reported to be the primary site of cancer, based on the resolved ICD-9 code, with a range of ICD-9:140-208.9 (defining malignant neoplasms) and 238.6 (defining plasma cells of uncertain significance). Histologic behaviour code was used to identify malignant primary cancers, and differentiates them from in situ, benign or uncertain classification of malignancy. As per the Surveillance, Epidemiology and End Results (SEER) program manual, typically a histologic term "carries a clear indication of the likely behaviour of the tumor, whether malignant or benign" which is then reflected in the ICD histologic behavior. However, if the Pathologist codes the behavior differently, it is the Pathologists assessment of histologic behavior which is recorded in the behavior code.[27] The first primary malignancy was identified based on date of diagnosis. Therefore, a combination of ICD-9 code for primary site of cancer, histologic behaviour code and first diagnosis of primary malignancy were used to define the primary outcome of interest. The timing of this outcome was determined relative to the survey administration date. If the outcome occurred in the 5 years after survey administration, the respondent was classified as an incident cancer case.

Right censored observations were defined based on absence of an incident cancer diagnosis prior to 5 years of follow-up, death or the end of eligibility. The competing risk

event in this study, death, was defined by the date of death if it occurred prior to the minimum of 5 years of follow-up, incident cancer diagnosis or the end of eligibility.

2.4 Predictors:

The major predictor variables of interest in this study were pre-specified and represent lifestyle-type modifiable risk factors for cancer, specifically: intake of fruits and vegetables, activity level, smoking status and alcohol consumption. These variables were obtained from the self-reported data in the CCHS cycles 1.1, 2.1, 3.1 and 4.1. Several additional categories of risk factors were examined, namely socio-demographic factors, deprivation factors, and medical conditions. When possible, all variables were kept as continuous variables rather than categorizing variables, to minimize the loss of predictive information through categorization.[28] Table 1 lists the predictor variables and scale of measurement.

2.4.1 Behavioural Risk Factors:

Average daily intake of fruits and vegetables was derived from the combination of CCHS questions relating to number of times per day that fruit, salad or carrots were consumed, and number of servings of other vegetables (excluding carrots, potatoes and salad) that were consumed. Specifically, this variable did not include potatoes or fruit juices in the average of fruit and vegetable consumption. Average daily intake of fruits and vegetables was analyzed as a continuous variable.

Physical activity was defined by the average daily energy expenditure during leisure time activities. Energy expenditure in the CCHS was calculated as the product of the number of times an individual engaged in physical activity over the previous 12 months multiplied by the duration in hours of activity and the metabolic equivalents (METs) for the activity, divided by 365 (number of days per year). Metabolic equivalents are a method for quantifying activity level, where one MET reflects the amount of oxygen consumed while sitting quietly.[29] Average METs per activity were pre-specified in the CCHS and were based on the formula of number of kilocalories/kilogram/hour divided by 365 days per year.

Smoking status was captured using two variables, one categorical and one continuous, both derived from CCHS variables. In the CCHS, individuals were categorized as current daily smokers, occasional smokers who were former daily smokers, occasional smokers who were never daily smokers, non-smokers who were former daily smokers, non-smokers who were former occasional smokers and those who had never smoked. These CCHS categories were used to define in our study if an individual had been a former smoker at any time. The number of years prior to survey administration that an individual reported quitting smoking in the CCHS was then used to define if that individual had quit smoking within the last 5 years or greater than 5 years ago. Based on the CCHS information, this study's categorical smoking variable indicated smoking status as non-smoker, current smoker, former smoker who quit >5 years ago or former smoker who quit ≤5 years ago. Pack-years of smoking for individuals with a smoking history was determined from CCHS data indicating the difference between the individuals age at

survey administration and their age when initiating daily smoking, multiplied by the number packs of cigarettes smoked daily (given 20 cigarettes per pack). Although this classification had its limitations, such as the fact that time from quit date was not continuous, it did permit an understanding of the different effects of current, former and never smoking at specified exposure levels (pack-years).

Exposure to second-hand smoke was also included as a lifestyle risk factor. Second-hand smoke exposure in confined spaces, such as vehicles, can result in significantly higher exposure to harmful particulate matter than in more spacious areas, however time spent in these spaces may be less than in other areas (i.e. public spaces) in which exposure is less concentrated but over longer periods of time.[30,31] Given the lack of information in the CCHS regarding durations of exposure to second hand smoke in various locations, it was decided that all second hand smoke exposure identified by individuals regardless of location would be grouped into one category. Therefore, the CCHS variables of second-hand smoke exposure at home, second-hand smoke exposure in vehicles and second-hand smoke exposure in public places were combined such that if any one of these was present, overall second-hand smoke exposure was defined in the affirmative.

Alcohol consumption was measured using two different variables: number of alcoholic beverages in the previous week and former drinker status. Former drinker was based on the CCHS definition of having not drank alcohol in the past 12 months but having consumed alcohol at some point in the past. Number of alcoholic beverages in the prior

week was analyzed as a continuous variable and former drinker as a dichotomous variable.

2.4.2 Socio-demographic Factors:

Several socio-demographic factors were included, such as education level and ethnicity. Consideration was given to the appropriate method of grouping different ethnicities identified by the CCHS survey, in order to best identify those in whom similar propensities for malignancy development existed. The CCHS identified 12 different ethnicities, as well as a grouping for "other" and "don't know." Some of the groupings were anticipated to have small numbers of individuals, therefore a more limited number of groups was desired. The grouping of Asian ethnicities was determined based on literature indicating that cancer incidence among Asian Americans was lowest for Chinese and Korean individuals, higher for Filipino individuals and again higher for S.E. Asian and Japanese individuals. All individuals from this study had lower cancer incidences than those found in non-Hispanic Caucasian individuals.[32] Based on this information, it was decided to group as follows: Caucasian; Chinese or Korean; African American; Filipino, South East (S.E.) Asian or Japanese; Latin American; Aboriginal; all other ethnicities (including South Asian, Arab, West Asian) as well as "other" and "don't know."

Immigration status was reported in CCHS based on respondents indicating that they were or were not an immigrant. Percentage of time lived in Canada was defined based on the

length of time an individual reported being in Canada since his/her immigration, or the date difference between date of immigration and survey administration date.

Education was recorded in four categories: less than secondary school education, secondary school education with no post-secondary education, having completed some post-secondary education and having obtained a post-secondary degree/diploma.

2.4.3 Deprivation Factors:

Pampalon's index of material (lack of access to necessary goods and amenities) and social (fragility of social networks) deprivation derived in Canada was used to include a measure of socio-economic disparities that may affect health status.[33] This index was based on a spatial unit, rather than a measure for individuals, and each unit comprised one or more neighbouring blocks of houses with a population between 400- 700 individuals. The index was derived from six socio-economic indicators including lack of a high school diploma, the employment to population ratio, and personal income (factors reflective of material deprivation), as well as proportion of those living alone, proportion of individuals who are separated, divorced or widowed, and proportion of single-parent families (factors reflective of social deprivation). Subsequently, study authors assigned a value score for each level of material and social deprivation, and small areas were classified based on these scores into population-weighted quintiles (20^{th} percentiles) ranging from the least deprived (quintile 1) to most deprived (quintile 5).[33] In order to simplify this score for the purposes of our study, we created three categories of deprivation: low, moderate and high. Low deprivation was classified as both a material

and social deprivation below the third quintile, high deprivation was classified as both material and social deprivation above the third quintile and moderate deprivation encompassed all others.

2.4.4 Medical Conditions:

Health conditions that were recorded in the CCHS and were included in this study due to their possible relationship with an increased risk of certain cancers were ulcerative colitis, crohn's disease and chronic obstructive pulmonary disease (COPD). BMI was also included in this study and was defined as the weight in kilograms divided by height in meters squared.

Table 1: Pre-specified Predictor Variables and Scale of Measurement

Variable	**Scale**	**Min-Max/Levels in Females**	**Min-Max/Levels in Males**
Age	Continuous	20-101	20-98
Alcohol consumption (beverages per day in the prior week)	Continuous	0-24	0-51
Former alcohol drinker	Dichotomous	Yes, No	Yes, No
Dietary consumption of fruit and vegetables daily	Continuous	0-12.7	0-11.3
Pack-years of smoking	Continuous	0-76	0-108
Smoking status	Categorical	Non-smoker; current smoker; former smoker quit ≤5 years ago; former smoker quit >5 years ago	Non-smoker; current smoker; former smoker quit ≤5 years ago; former smoker quit >5 years ago
Second hand smoke exposure in home, public spaces or vehicles	Dichotomous	Yes, No	Yes, No

Leisure physical activity	Continuous	0-10.5	0-12.2
Body Mass Index	Continuous	9.9-47.17	10.10-43.30
Ethnicity	Categorical	Caucasian; Chinese or Korean; African American; Filipino, South East (S.E.) Asian or Japanese; Latin American; Aboriginal; all other ethnicities	Caucasian; Chinese or Korean; African American; Filipino, South East (S.E.) Asian or Japanese; Latin American; Aboriginal; all other ethnicities
History of Emphysema or COPD	Dichotomous	Yes, No	Yes, No
History of Inflammatory Bowel Disease (IBD)	Dichotomous	Yes, No	Yes, No
Percent of life lived in Canada	Continuous	0-100%	0-100%
Pampalon's Deprivation Index	Ordinal	Low, moderate, high	Low, moderate, high
Immigration Status	Dichotomous	Yes, No	Yes, No
Individual Education	Categorical	Secondary school graduation or less; Post-secondary education	Secondary school graduation or less; Post-secondary education
Survey Cycle	Categorical	2003, 2005, 2007	2003, 2005, 2007

2.5 Data Linkage:

Data linkage was conducted in conjunction with the Institute for Clinical Evaluative Studies (ICES) in Ottawa, a research community that enables access and linking of population-based health data as well as clinical and administrative data. CCHS cohort data and OCR registry information were deterministically linked using encrypted health card number (IKN).

2.6 Sample Size Calculation:

A commonly used sample size requirement for predictive algorithms with time-to-event outcomes is that there should be no more than 1 candidate predictor variable per 10 events.[23,25] The anticipated number of events for this study was calculated prior to accessing the Ontario Cancer Registry data by using data on incident cancer rates readily available from Statistics Canada. Based on Statistics Canada estimates from 2009 in Ontario, there was an annual rate of 496.6 incident cancer cases per 100 000 people for all age groups.[34] This would be an underestimate for our study, which included only those over age 20 years old at survey administration, in whom incident rates are higher than in younger individuals. In the CCHS datasets from 2001, 2003, 2005 and 2007 there are approximately 866 000 person-years of follow-up. Therefore, there would be an expected 4300 cases of incident cancer detected during the follow-up period. Thus, up to 430 candidate predictor variables could potentially be examined in the risk prediction model, which is considerably more than proposed in this study. Therefore, the sample size appears adequate for this investigation.

An additional method for defining the sample size was utilized to verify that power was adequate, and is presented in Appendix 2.

2.7 Missing Data:

Multiple imputation was used to account for missing values on predictor variables. We implemented the 'aregImpute' function in R to carry out the multiple imputations. This function uses predictive mean matching to replace missing values using observed values

from individuals with real data that most closely approximate the predicted values.[35] The imputation was performed using a two-step process. During the first step, all available predictor variables, regardless of whether they were to be used in the prediction model, were used to impute missing values in addition to interaction terms. In the second step, any variables that could not be imputed initially were then imputed by the same process. As an example, smoking status (i.e. current smoker, former smoker quitting ≤5 years ago, former smoker quitting >5 years ago or non-smoker) was imputed in the first iteration of imputation, while pack-years of smoking was imputed in the second iteration. This was done so that the imputation process utilized the imputed values for smoking status from the first iteration to generate appropriate values in the second iteration. Five multiple imputation data sets were created with the final results of these data sets being combined using rules from Rubin to account for imputation uncertainty.[36]

2.8 Ethical Approval Process:

Research ethics board approval was obtained from the Ottawa Hospital Research Ethics Board (REB).

CHAPTER 3: STATISTICAL METHODS

3.1 Data Cleaning and Coding:

Continuous variables were inspected using descriptive statistics including means and standard deviations. All continuous variables were also examined graphically using histograms. Number of pack-years, alcohol consumption, dietary fruit and vegetable consumption, energy expenditure and BMI were truncated at the 99.5th percentile. This was done to reduce the risk of measurement error and extreme variables being overly influential. All categorical predictors were examined using frequency distributions to identify categories with small frequencies that may cause instability in the model.

3.2 Model Specification:

Our general approach to model derivation was as described by Harrell.[25] Harrell recommends pre-specification of the model with all predictor variables included in the final model. He recommends that the degrees of freedom (df) allocated to each predictor, i.e. the degree of complexity with which each predictor is modeled, be determined by its predictive potential as measured by a partial chi-squared statistic. This approach is preferable to using scatter plots to examine the shape of the association with the outcome. It results in less bias due to the fact that all variables, regardless of strength of association with the outcome, are included in the prediction model. It avoids a data-driven approach in that the partial chi-squared statistics do not allow investigators to see the degree of non-linearity and thereby alter the functional form of the modeled association.[25]

Restricted cubic splines were used to allow continuous predictor variables to be included within the Cox proportional hazard model while allowing for any potential non-linear relationships. Restricted cubic splines are piecewise cubic functions that are smooth at the knots and restricted to be linear in the tails. The knots are placed at fixed quantiles of the distribution. For example, with 3 knots, the fixed quantiles are set at the 10^{th}, 50^{th} and 90^{th} centiles, and with 5 knots the 5^{th}, 27.5^{th}, 50^{th}, 72.5^{th} and 95^{th} centiles.[25]

Using this approach, all continuous predictor variables were represented as restricted cubic splines with a maximal number of knots allocated for each.[25] A priori, it was decided that for age, the maximal number of knots allocated was 5 knots (4 df) as it was most likely to be highly correlated with incident cancer. For all other continuous variables the maximal number of knots was set at 3 knots (2 df). All categorical predictor variables retained their original categories, with the degrees of freedom equal to the number of categories minus one. A partial chi-squared statistic was calculated to evaluate the association between each predictor variable and incident cancer, adjusted for all other predictor variables. To make corrections for any chance associations, the number of degrees of freedom was subtracted from the partial chi-square. The resulting partial associations were analyzed graphically as an ANOVA plot to determine the importance of each predictor variable and to determine the final degree of freedom allocation to each predictor.

The strength of the relationship between predictor variables and the outcome was used to decide whether the number of knots (and corresponding df) for continuous variables

could be reduced or number of categories collapsed for categorical variables. The allocation of degrees of freedom was conducted such that any variables with a higher correlation retained a greater number of degrees of freedom to permit modeling the complexity between predictor and outcome variables. In addition, the existing epidemiologic understanding of the relationship between predictor and outcome variables was taken into account (i.e. for BMI it was hypothesized that both low and high BMI would be correlated to cancer incidence, therefore it was allocated 2 df for both males and females despite a weaker correlation).

Interaction terms were analyzed only between age and smoking category, and between age and pack-years of smoking, to minimize complexity of the overall model while including interactions between variables identified as the most related to outcome in the ANOVA plots.

All predictor variables identified prior to study initiation were included in the model, and if they were deemed to be unimportant they remained in the model as simple linear terms if continuous, or with the minimum number of groups for categorical variables.

3.3 Multivariable Competing Risk Cox Proportional Hazard Model:

Survival analyses were conducted separately for males and females using a multivariable competing risk Cox proportional hazard model as described by Fine and Gray, with any non-cancer death considered a competing risk.[37] Hazard ratios were used to describe the relationship between each predictor variable and incidence of cancer for males and

females. Log hazard plots were generated to visualize the relationship between predictor variables (x-axis) and logarithmic hazard of incident cancer (y-axis).

Traditional Cox proportional hazard models were not used due to the fact that the competing risk event, death, would have been categorized as a right-censored observation. Right censoring occurs when an individual leaves the study or the study is completed prior to the event occurring. If an individual has died, he or she theoretically could have developed the event of interest, cancer, over the remaining study period and therefore should be categorized separately from those who completed the study without cancer development.

3.4 Maintaining Internal Validity with a Weighted Sample:

In terms of the clustered sampling design used in the CCHS, it is possible that subjects within a cluster are more similar than those in another cluster, even after consideration of baseline characteristics among subjects.[38] Therefore, it may be necessary to use multilevel regression models to account for the complex sampling strategy. In our study, neither multilevel modeling nor survey weights were used for derivation of the predictive model in order to maintain internal validity of the model. For example, if a small number of individuals in one category were heavily weighted, they may inappropriately represent the true values in the population. By selecting to use only the true values in the dataset, the internal validity of the study is upheld. However, to maximize external validity, the weights will be considered for use in any future application of the model within the Canadian population.

3.5 Proportional Hazard Assumption:

The proportional hazard assumption was assessed using plots of scaled Schoenfeld residuals versus time for each predictor. The proportional hazard assumption was tested for the overall model. If the proportional hazard assumption were to be violated, meaning that over time the hazard ratio changed, consideration would be given to the inclusion of time interactions.

3.6 Model Performance:

Model performance was assessed using measures of overall performance (Nagelkerke's R^2 and the Brier score which summarize model performance based on the distance between the observed and predicted outcomes[39]) and by calibration and discrimination.

Nagelkerke's R^2 is a logarithmic scoring rule[39] with value ranges from 0-1 and is described as the proportion of variation in the outcome variable which can be explained by the risk score, with higher numbers being optimal.[40] It is typically expressed as a percentage. It is calculated by comparing the −2 log likelihood of a model without predictors and a model with one or more predictors.[39] The Brier score is an overall performance measure in which the squared difference between the observed and predicted outcomes are calculated.[39] It has a range from 0 (for a perfect model) to 0.25 for an uninformative model with a 50% incidence of the outcome.[39]

Calibration, also known as accuracy, is a measure of how well an algorithm's predicted risk approximates the observed risk in the population.[4] Calibration was determined for the risk algorithm in the overall population and across pre-identified subgroups, including across age groups and socioeconomic groups. Poor calibration was defined as a difference of 20% or more between the observed and predicted estimates for categories with a prevalence of >5%, based on the expectation that greater than a 20% difference would be significant from both a clinical and health policy standard.

Discrimination, or the ability to differentiate between those with high versus low-risk of an outcome,[4] was evaluated using the C-statistic. Specifically, the C-statistic indicates the fraction of pairs (any two individuals from the available dataset) where the individual with an event has a higher predicted probability than an individual without the event.[41] The C-statistic was calculated using methodology published by W. Kremers.[41] A C-statistic over 0.7 was considered adequate discrimination, while a C- statistic over 0.8 was considered excellent discrimination.

Some concern exists in the interpretation of discrimination and calibration, specifically in the amount that is considered sufficient to suggest a model is useful in the clinical realm.[42] Steyerberg and Vickers state that the C-statistic for discrimination and an observed-to-predicted ratio for calibration are the necessary statistics for describing a predictive model and should be reported in all predictive studies.[39] A discrimination slope, or plot of the absolute difference in predictions for those with and without the outcome, can also be calculated but is not possible to use in survival analyses. Decision

analytic methods are also proposed by Steyerberg and Vickers, however these are methods to determine the clinical utility of a prediction tool. Since this algorithm will be targeted for use at the population level, an understanding of its utility for an individual patient is not necessary.[39]

3.7 Model Validation:

Risk prediction tools are becoming more common, and one key concern is the rapid development of tools without subsequent validation on a different data set than the one used to create the model.[43] The final risk algorithm was validated by applying it within the CCHS cycle 1.1 (2001) cohort. Within this cohort, predicted risks for incident cancer were generated using the risk algorithm. The algorithm's calibration and discrimination were again evaluated.

3.8 Statistical Software:

Statistical analyses were completed using SAS, version 6.1 (2013) and using R Studio, version 0.98.1091 (2009-2014).

CHAPTER 4: RESULTS

4.1 Participants:

4.1.1 Derivation and Validation Cohort:

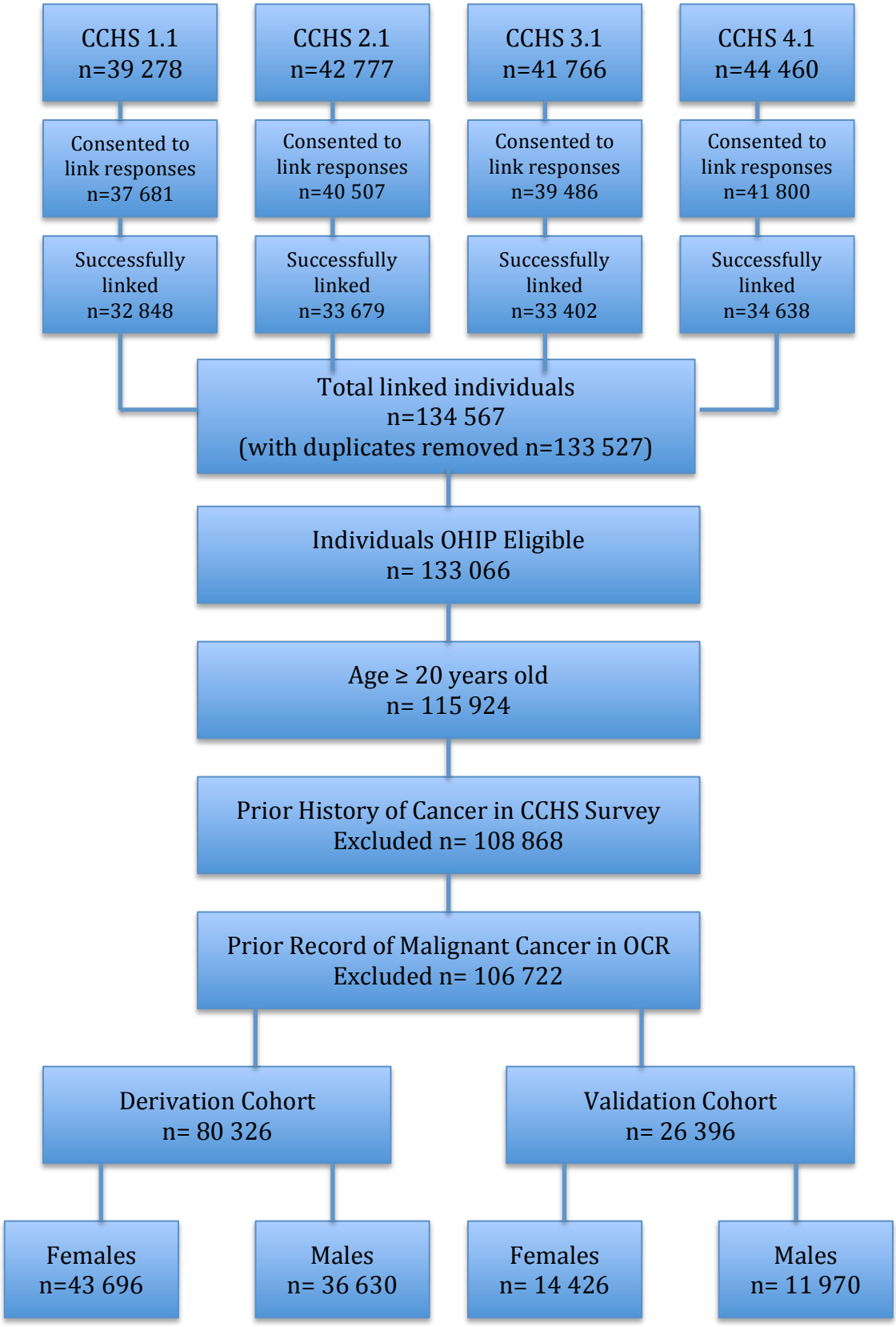
There were 106 722 eligible individuals from participants of CCHS cycles 1.1-4.1 after exclusion criteria were applied. From this population, the derivation cohort consisted of 80 326 individuals who had survey responses in CCHS cycles 2.1-4.1, of which 43 696 were female and 36 630 were male. The validation cohort consisted of 26 396 individuals who had survey responses in CCHS cycles 1.1, of which 14 426 were female and 11 970 were male. See Figure 1 for the cohort creation diagram.

4.1.2 Number of Incident Cancer Cases:

In the derivation cohort, there were 1559 (3.57%) incident cancer cases among females and 1544 (4.22%) incident cancer cases among males during the 5-year follow-up period post-survey administration. In terms of the competing risk event, there were 1224 (2.80%) deaths during that period among females and 1197 (3.27%) deaths among males.

In the validation cohort, there were 497 (3.45%) incident cancer cases among females and 498 (4.16%) cases among males during the 5-year follow-up period. Among females, there were 460 (3.19%) deaths, while among males there were 440 (3.68%) deaths during the 5-year follow-up period.

Figure 1: Cohort Creation Diagram



4.1.3 Demographic Characteristics on Raw Data for the Derivation Cohort:

The un-imputed data was analyzed and is presented in Table 2. The median (interquartile range, or IQR) age for females was 50 (35-64) years old and for males was 48 (35-61) years old. The median (IQR) dietary consumption of fruits and vegetable servings per day for females was 3.6 (2.4-5.3) with a range of 0-12.7 and for males was 2.7 (1.7-4.0) with a range of 0-11.3. Energy expenditure for females showed a median (IQR) of 1.3 (0.5-2.7) daily METS or kcal/kg/day and males had corresponding values of 1.6 (0.6-3.1). BMI of females had a median of 24.8 (IQR 22.0-28.7), while for males it was 26.3 (IQR 24.1-29.2). In terms of alcohol consumption, females consumed a median (IQR) of 0 (0-3) alcoholic beverages per week ranging from 0-24 drinks/week and males consumed a median (IQR) of 2 (0-8) alcoholic beverages per week ranging from 0-51 drinks/week.

The majority of females were nonsmokers (N=17 209, 39.4%), while 8340 (19.1%) were former smokers having quit >5 years ago, 3179 (7.3%) were former smokers having quit ≤5 years ago and 9493 (21.7%) were current smokers. Among males, there were 9246 (25.2%) non-smokers, 9462 (25.8%) former smokers having quit >5 years ago, 3 209 (8.8%) former smokers having quit ≤5 years ago and 9892 (27.0%) current smokers. The median pack-years of smoking among women was 0.007 (IQR 0.0-8.0) ranging from 0 to 76.25 and for males was 1.0 (0.0-18.8) with a range of 0 to 108.

The remainder of the descriptive statistics for the derivation cohort are outlined in Table 2.

4.1.4 Demographic Characteristics on Raw Data for the Validation Cohort:

In the validation cohort, the median (IQR) age for females was 46 (35-63) years and for males was 45 (35-59) years. The median (IQR) dietary consumption of fruits and vegetable servings per day for females was 3.1 (2.1-4.5) and for males was 2.4 (1.6-3.6). Female energy expenditure had a median (IQR) of 1.2 (0.4-2.5) daily METS or kcal/kg/day while males had a value of 1.4 (0.5-3.0). Median (IQR) BMI for females was 24.6 (21.8-28.1) and for males was 26.2 (23.8-28.8). Alcohol consumption among females was a median (IQR) of 0 (0-2) alcoholic beverages per week and among males was 2 (0-7) alcoholic beverages per week.

The number (%) of female non-smokers was 5272 (36.6), former smokers having quit >5 years ago was 2749 (19.1), former smokers having quit ≤5 years ago was 897 (6.2) and current smokers was 3696 (25.6). For males, number (%) of non-smokers was 2845 (23.8), former smokers having quit >5 years ago was 3185 (26.6), former smokers having quit ≤5 years ago was 841 (7.0) and current smokers was 3636 (30.4). The median (IQR) pack-years of smoking among women was 0.007 (0.0-6.5) and for males was 0.007 (0.0-16.5).

The remainder of the descriptive statistics for the validation cohort are outlined in Table 2.

Table 2: Baseline Characteristics Prior to Imputation According to Sex in the Derivation and Validation Cohorts

	Derivation Cohort		Validation Cohort	
Characteristic	**Female**	**Male**	**Female**	**Male**
Patient-related Factors				
Age	50 (35-64)	48 (35-61)	46 (35-63)	45 (35-59)
Body Mass Index (BMI)	24.8 (22.0-28.7)	26.3 (24.1-29.2)	24.6 (21.8-28.1)	26.2 (23.8-28.8)
Behavioural Factors				
Alcohol consumption weekly	0 (0-3)	2 (0-8)	0 (0-2)	2 (0-7)
Former alcohol drinker	7 757 (17.75)	4 369 (11.93)	2265 (15.70)	1324 (11.06)
Dietary consumption of fruit and vegetables daily	3.6 (2.4-5.3)	2.7 (1.7-4.0)	3.1 (2.1-4.5)	2.4 (1.6-3.6)
Pack years of smoking	0.007 (0.0-8.0)	1.0 (0.0-18.8)	0.007 (0.0-6.5)	0.007 (0.0-16.5)
Smoking status				
Non-smoker	17 209 (39.38)	9246 (25.24)	5272 (36.55)	2845 (23.77)
Former smoker quit >5 years ago	8340 (19.09)	9462 (25.83)	2749 (19.06)	3185 (26.61)
Former smoker quit ≤5 years ago	3179 (7.28)	3209 (8.76)	897 (6.22)	841 (7.03)
Current smoker	9493 (21.73)	9892 (27.01)	3696 (25.62)	3636 (30.38)
Second hand smoke exposure	8 661 (19.82)	8 659 (23.62)	2100 (14.56)	2142 (17.89)
Leisure physical activity (average daily METS or kcal/kg/day)	1.3 (0.5-2.7)	1.6 (0.6-3.1)	1.2 (0.4-2.5)	1.4 (0.5-3.0)
Socio-demographic Factors				
Ethnicity				
Caucasian	38 704 (88.58)	32 212 (87.94)	13 207 (91.55)	10 939 (91.39)
Chinese or Korean	899 (2.06)	726 (1.98)	247 (1.71)	171 (1.43)
African American	272 (0.62)	168 (0.46)	73 (0.51)	49 (0.41)
Filipino, SE Asian or Japanese	1028 (2.35)	1009 (2.75)	255 (1.77)	232 (1.94)
Latin American	669 (1.53)	614 (1.68)	179 (1.24)	181 (1.51)
Aboriginal	982 (2.25)	799 (2.18)	166 (1.15)	122 (1.02)
S Asian, Arab, W Asian, Other or Don't Know	1142 (2.61)	1102 (3.01)	299 (2.07)	276 (2.31)
Percent of life lived	100 (33.9-100)*	100 (34.7-	100 (34.8-	100 (34.1-

in Canada		100)*	100)*	100)*
Immigration	9 220 (21.10)	7704 (21.03)	2822 (19.56)	2342 (19.57)
Individual Education				
Less than high school level education	8117 (18.58)	6441 (17.58)	3369 (23.35)	2632 (21.99)
High school graduation	8466 (19.37)	6408 (17.49)	3148 (21.82)	2375 (19.84)
Some post-secondary education	2 971 (6.80)	2653 (7.24)	1109 (7.69)	887 (7.41)
Post-secondary degree/diploma	23 845 (54.57)	20 800 (56.78)	6705 (46.48)	5985 (50.00)
Deprivation Factors				
Deprivation index				
Low	8 193 (18.75)	7194 (19.64)	2700 (18.72)	2444 (20.42)
Moderate	26 981 (61.75)	22 689 (61.94)	8999 (62.38)	7425 (62.03)
High	7 223 (16.53)	5733 (15.65)	2429 (16.84)	1843 (15.40)
Medical Conditions				
History of Emphysema or COPD	828 (1.89)	737 (2.01)	172 (1.19%)	170 (1.42)
History of Inflammatory Bowel Disease (IBD)	3069 (7.02)	1022 (2.79)	626 (4.34)	209 (1.75)

Results are presented as N (%) for categorical variables or median (interquartile range) for continuous variables unless otherwise specified
*median (5^{th}-95^{th}%)

4.1.5 Percentage of Missing Data in the Derivation Cohort:

The prevalence of missing data for each variable was examined. For females in the derivation cohort, there were 1313 (3%) missing from the dietary consumption of fruits and vegetables, 472 (1.08%) missing from energy expenditure, 2333 (5.34%) missing from BMI, 490 (1.12%) from weekly alcohol consumption, 5475 (12.53%) from smoker category and 4394 (10.06%) from pack-years of smoking.

For males in the derivation cohort, there were 1549 (4.23%) missing from the dietary consumption of fruits and vegetables, 771 (2.10%) missing from energy expenditure, 423 (1.15%) missing from BMI, 911 (2.49%) from weekly alcohol consumption, 4812 (13.16%) from smoker category and 4680 (12.78%) from pack years.

All variables in the derivation cohort had <15% missing values prior to imputation for both males and females.

4.1.6 Percentage of Missing Data in the Validation Cohort:

There were similar numbers of missing data in the derivation and validation datasets for both males and females, with the exception that number of pack-years of smoking was missing to a greater extent in the validation cohort than in the derivation cohort. Specifically, for females there was 3250 (22.53%) missing from pack-years and for males there was 3642 (30.43%) missing from pack-years. The greatest number of missing among all variables for both males and females was pack-years of smoking. All variables had <31% missing, and when not including pack-years of smoking, all other variables had <15% missing values prior to imputation.

4.1.7 Demographic Characteristics on Imputed Data:

Visual examination of the medians and interquartile ranges for continuous variables, and numbers with percentages for categorical variables, revealed that there were no large observable differences after imputation was completed. See Table 3 for descriptive statistics for imputed data.

Table 3: Baseline Characteristics After Imputation According to Sex in the Derivation and Validation Cohorts

	Derivation Cohort		Validation Cohort	
Characteristic	**Female (N=43 696)**	**Male (N=36 630)**	**Female (N=14 426)**	**Male (N=11 970)**
Patient-related Factors				
Age	50 (35-64)	48 (35-61)	46 (35-63)	45 (35-59)
Body Mass Index (BMI)	24.9 (22.0-28.9)	26.3 (24.1-29.2)	24.7 (21.8-28.7)	26.2 (23.8-28.8)
Behavioural Factors				
Alcohol consumption weekly	0 (0-3)	2 (0-8)	0 (0-2)	2 (0-7)
Former alcohol drinker	7 774 (17.79)	4378 (11.95)	2267 (15.71)	1325 (11.07)
Dietary consumption of fruit and vegetables daily	3.6 (2.4-5.2)	2.6 (1.7-4.0)	3.1 (2.1-4.5)	2.4 (1.6-3.6)
Pack years of smoking	0.007 (0.0-7.5)	0.75 (0.0-18.0)	0.007 (0.0-6.0)	0.014 (0.0-15.0)
Smoking status				
Non-smoker	17 227 (39.42)	9 253 (25.26)	5279 (36.59)	2850 (23.81)
Former smoker quit >5 years ago	12 352 (28.27)	12 903 (35.23)	4117 (28.54)	4257 (35.56)
Former smoker quit ≤5 years ago	4 617 (10.57)	4 566 (12.47)	1331 (9.23)	1222 (10.21)
Current smoker	9 500 (21.74)	9 908 (27.05)	3699 (25.64)	3641 (30.42)
Second hand smoke exposure	8 661 (19.82)	8 659 (23.64)	2100 (14.56)	2142 (17.89)
Leisure physical activity (average daily METS or kcal/kg/day)	1.3 (0.5-2.7)	1.6 (0.6-3.1)	1.2 (0.4-2.5)	1.4 (0.5-2.9)
Socio-demographic Factors				
Ethnicity				
Caucasian	38 704 (88.58)	32 212 (87.94)	13 207 (91.55)	10 939 (91.39)
Chinese or Korean	899 (2.06)	726 (1.98)	247 (1.71)	171 (1.43)
African American	272 (0.62)	168 (0.46)	73 (0.51)	49 (0.41)
Filipino, SE Asian or Japanese	1 028 (2.35)	1 009 (2.75)	255 (1.77)	232 (1.94)
Latin American	669 (1.53)	614 (1.68)	179 (1.24)	181 (1.51)
Aboriginal	982 (2.25)	799 (2.18)	166 (1.15)	122 (1.02)
S Asian, Arab, W Asian, Other or Don't Know	1 142 (2.61)	1 102 (3.01)	299 (2.07)	276 (2.31)
Percent of life lived	100 (33.9-100)*	100 (34.7-	100 (34.8-	100 (34.1-

in Canada		100)*	100)*	100)*
Immigration	9229 (21.12)	7710 (21.05)	2828 (19.60)	2344 (19.58)
Individual Education				
Less than high school level education	8414 (19.26)	6769 (18.48)	3464 (24.01)	2723 (22.75)
High school graduation	8466 (19.37)	6408 (17.49)	3148 (21.82)	2375 (19.84)
Some post-secondary education	2971 (6.80)	2653 (7.24)	1109 (7.69)	887 (7.41)
Post-secondary degree/diploma	23 845 (54.57)	20 800 (56.78)	6705 (46.48)	5985 (50.00)
Deprivation Factors				
Deprivation index				
Low	8416 (19.26)	7401 (20.20)	2755 (19.10)	2500 (20.89)
Moderate	27 788 (63.59)	23 312 (63.64)	9189 (63.70)	7581 (63.33)
High	7492 (17.15)	5 917 (16.15)	2482 (17.21)	1889 (15.78)
Medical Conditions				
History of Emphysema or COPD	967 (2.21)	851 (2.32)	213 (1.48)	205 (1.71)
History of Inflammatory Bowel Disease (IBD)	3 071 (7.03)	1 023 (2.79)	627 (4.35)	209 (1.75)

Results are presented as N (%) for categorical variables or median (interquartile range) for continuous variables unless otherwise specified
*median (5^{th}-95^{th}%)

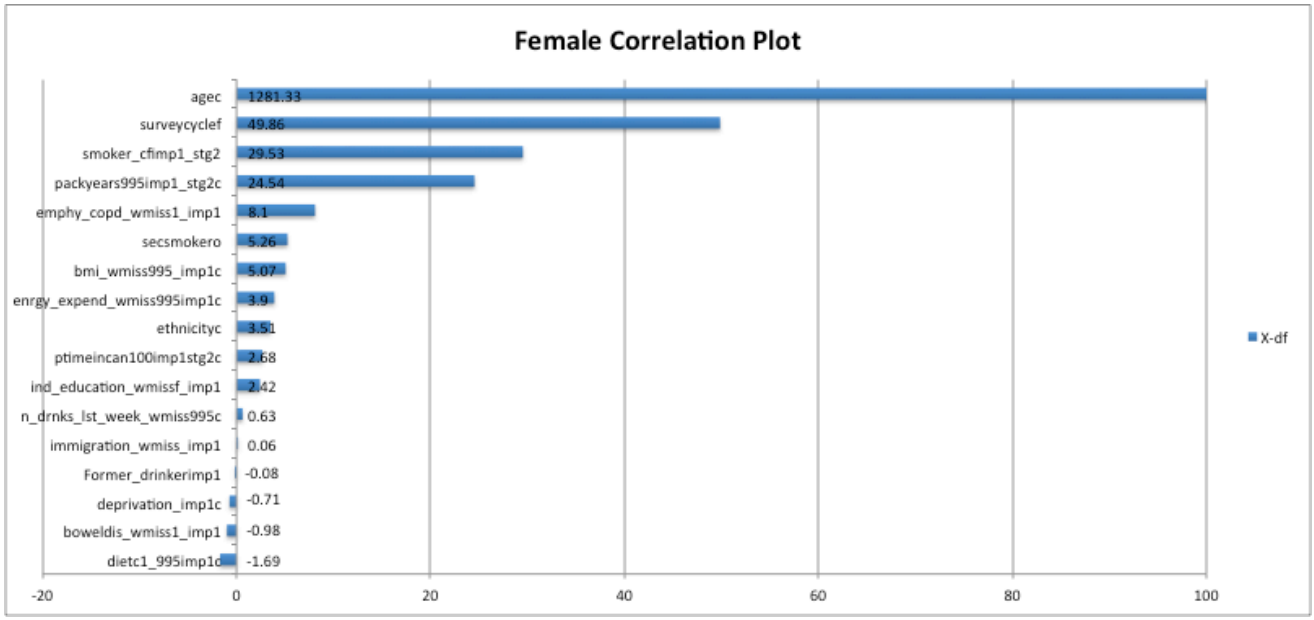
Histograms were created to examine the underlying distribution of predictor variables in the derivation cohort. For females, age showed a bimodal distribution with peaks in the mid-30's and mid-50's. The distribution for BMI and dietary fruit and vegetable consumption were both reasonably symmetrical, but with a slight rightward skew. Activity level, alcohol consumption and smoking pack-years were all right-skewed with the modal value being zero. Percent time lived in Canada was highly left-skewed with the modal value being 100%. For males, similar patterns were observed. (See Appendix 3)

4.2 Model Development:

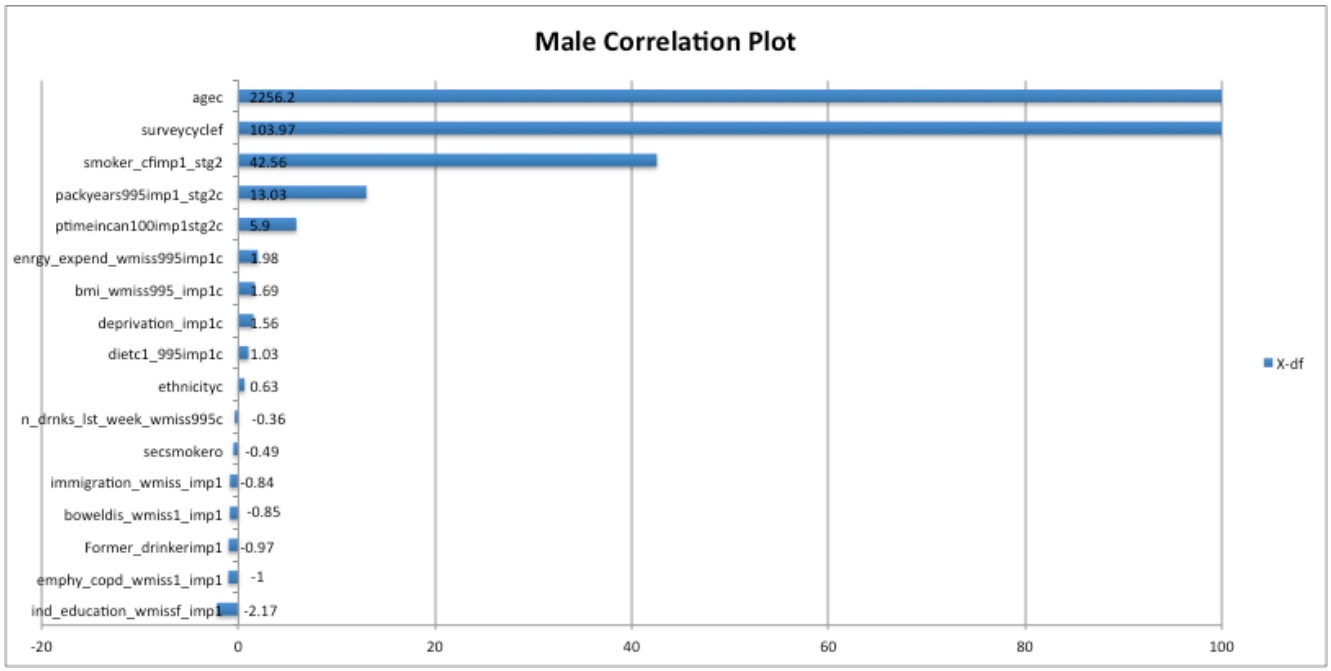
4.2.1 ANOVA Plots:

ANOVA plots showing the relationship between imputed predictor variables and the outcome of incident cancer were generated. The female ANOVA plot showed the highest association between age and outcome, with relationships also existing between survey cycle, smoking status, pack-years of smoking, emphysema or COPD, second hand smoking status and BMI with outcome, in descending order. (see Figure 2) The male ANOVA plot similarly showed the strongest association between age and the outcome, but with fewer other variables associated with the outcome. Survey cycle, smoking status, pack-years of smoking and percentage of time in Canada were also related to the outcome of interest in males.

Figure 2:
(A) Partial Model Likelihood Ratio Chi-square Statistic of Each Predictor Included in the Model for Females



(B) Partial Model Likelihood Ratio Chi-square Statistic of Each Predictor Included in the Model for Males



4.2.2 Determining Complexity of Predictor Variables:

The allocation of degrees of freedom is presented in Table 4. Age was allocated the greatest number of degrees of freedom (4 degrees of freedom) due to its strong correlation with the outcome. Survey cycle and smoking category were both strongly associated with incident cancer, therefore all categories within these variables were retained. Pack-years of smoking, BMI and for females, leisure physical activity, were allocated the next greatest number of degrees of freedom (2 degrees of freedom) for both females and males. Due to a weak correlation between both education and ethnicity with incident cancer, the categories for education and for ethnicity were collapsed to become dichotomous, but these variables were retained in the model. For education, the categories became secondary school graduation or less, versus some post-secondary education. For ethnicity, the categories became Caucasian or Other ethnicity. All other

variables were allocated 1 degree of freedom, with continuous variables being modeled as linear terms and categorical variables remaining dichotomous.

Table 4: Pre-specified Predictor Variables with Degrees of Freedom (DF) Allocation for Male and Female Models in the Derivation Cohort

	Degrees of Freedom (DF)	
Characteristic	**Female**	**Male**
Patient-related Factors		
Age	4	4
Body Mass Index (BMI)	2	2
Behavioural Factors		
Alcohol consumption weekly	1	1
Former alcohol drinker	1	1
Dietary consumption of fruit and vegetables daily	1	1
Pack years of smoking	2	2
Smoking status	3	3
Second hand smoke exposure	1	1
Leisure physical activity	2	1
Socio-demographic Factors		
Ethnicity	1	1
Percent of life lived in Canada	1	1
Immigration	1	1
Individual Education	1	1
Deprivation Factors		
Deprivation index	1	1
Medical Conditions		
History of Emphysema or COPD	1	1
History of Inflammatory Bowel Disease (IBD)	1	1
Survey Administration		
Survey Cycle	2	2

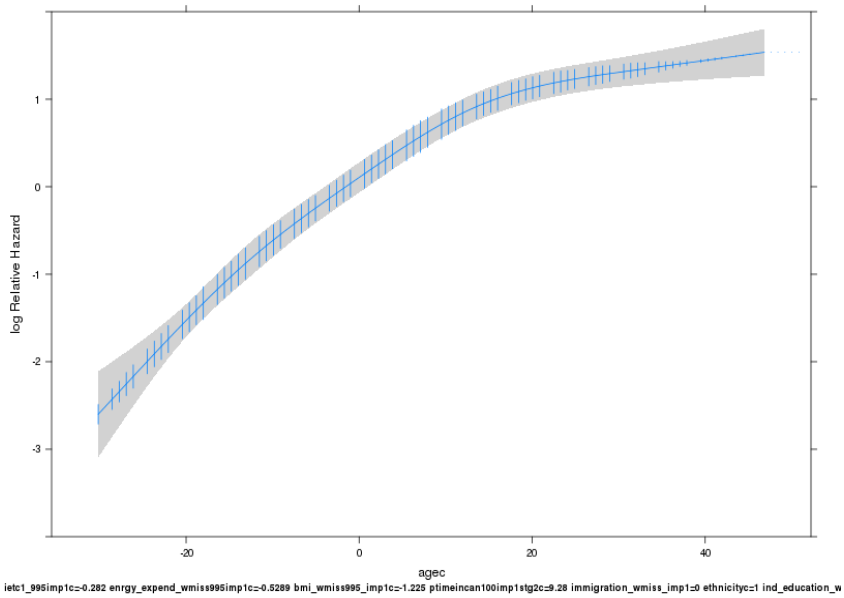
4.2.3 Examining Relationship Between Predictor Variables and Incident Cancer:

Log hazard plots were generated to visualize the relationship between predictor variables (x-axis) and logarithmic hazard of incident cancer (y-axis). As indicated by the log hazard plot, there was a positive association between age and the log hazard of incident cancer for females. The increase in log hazard was nearly linear, then plateaued somewhat at a higher age range. Among males, the plot of age versus log hazard of incident cancer reveals a near linear increase in log hazard of cancer with increasing age, again with a reduction in the slope of increase at the higher extremes of age.

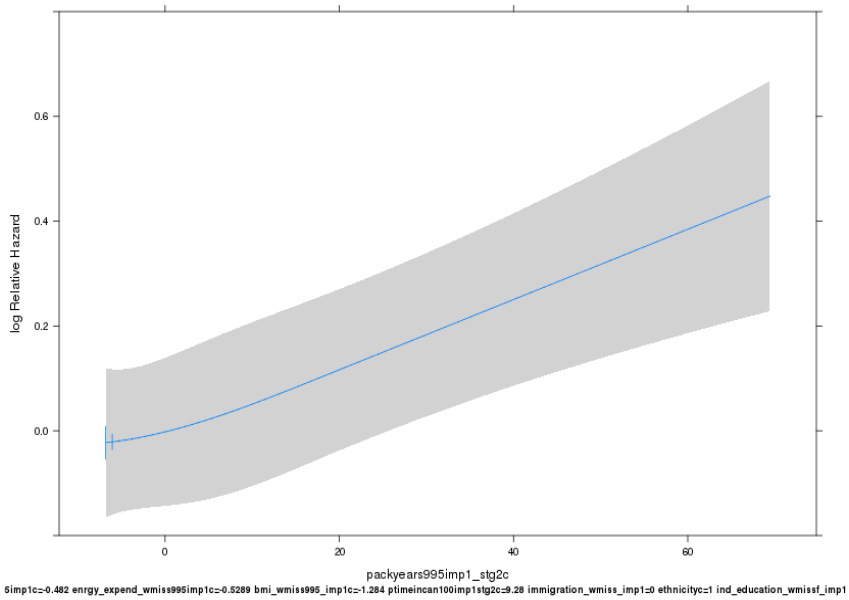
In women and men, a nearly linear relationship was observed between increasing number of pack-years of smoking and logarithmic hazard of incident cancer. Similarly, a trend toward increasing log hazard of cancer was seen with increasing number of alcoholic beverages per week, but with broadening confidence intervals around the estimate as number of alcoholic beverages increased. In females, low physical activity level was related to log hazard of incident cancer, while dietary fruit and vegetable consumption did not show any apparent relationship with log hazard of incident cancer. In males, a low and a high activity level were associated with an increased log hazard of incident cancer, however confidence intervals were wide particularly at higher activity levels. In males, increased dietary fruit and vegetable consumption was associated with a slightly lower log hazard of incident cancer. In males and females, both low BMI and high BMI were associated with increased log hazard of cancer. (see Figure 3)

Figure 3:
(A) Log Hazard Plots of Predictor Variables with Incident Cancer for Females in the Derivation Cohort (CCHS cycles 2-4) on Imputed Data

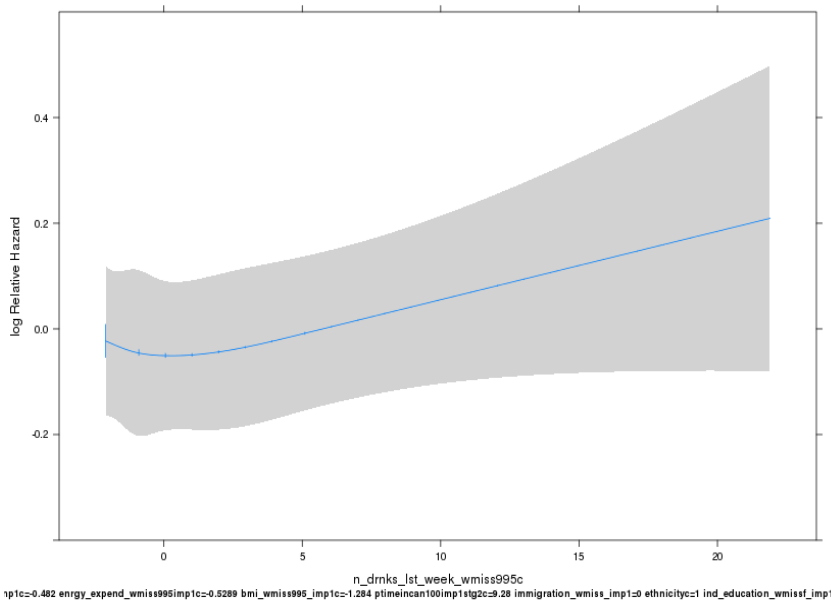
(i) Age:



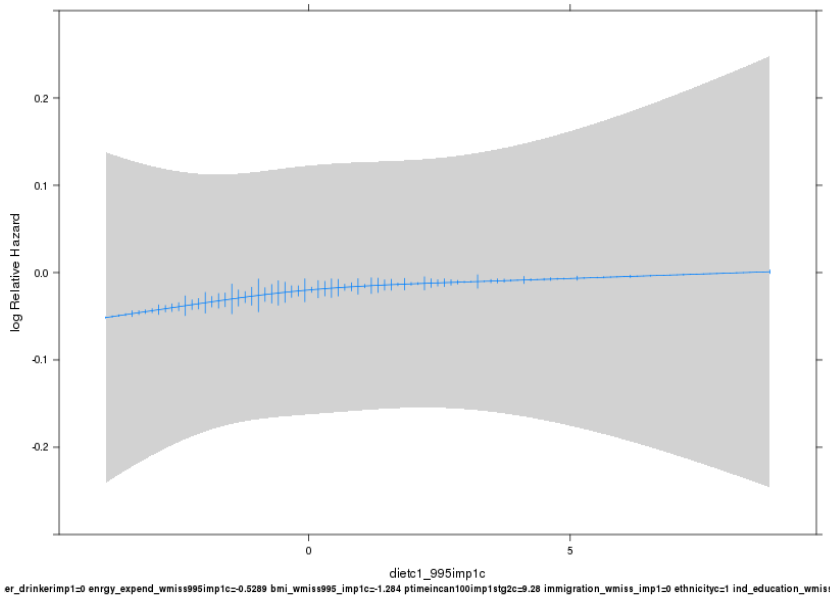
(ii) Pack-years Smoking:



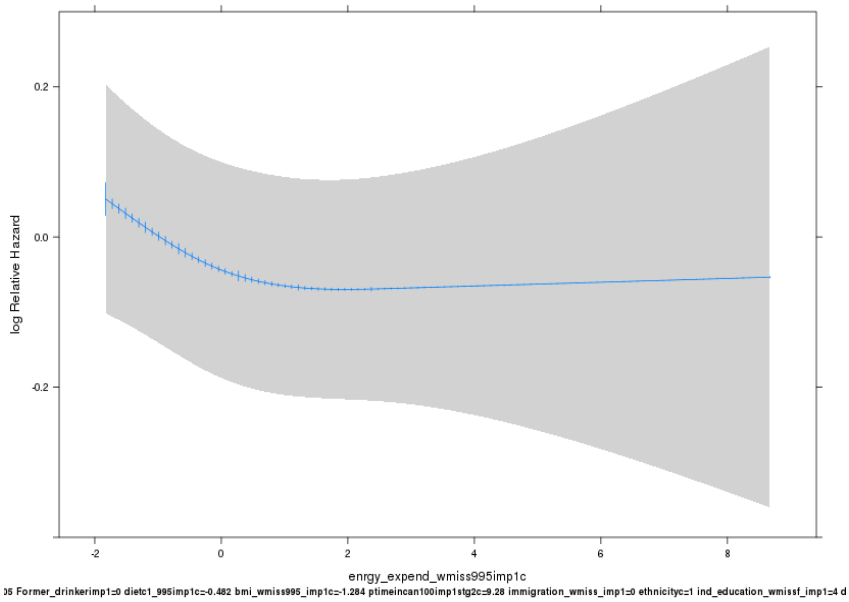
(iii) Alcohol Consumption (Number of Alcoholic Beverages Consumed Weekly):



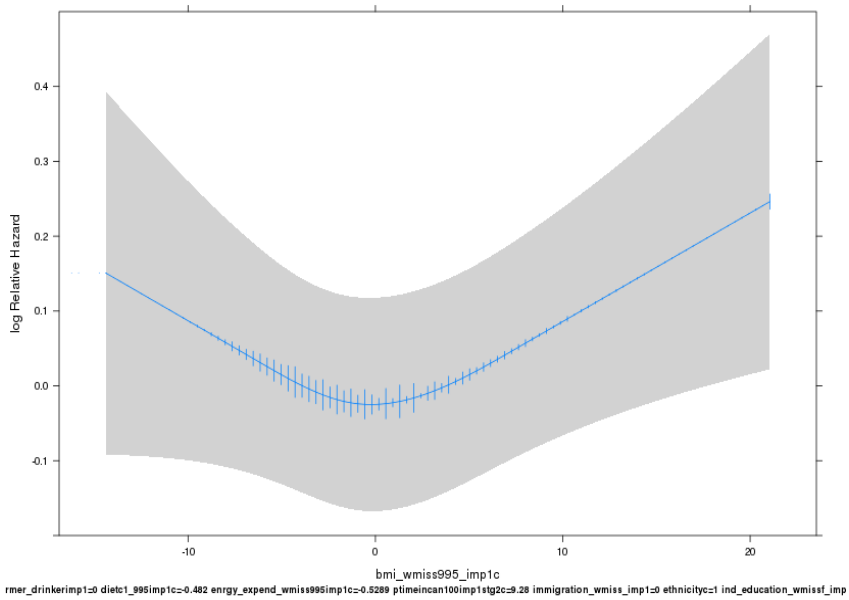
(iv) Dietary Fruit and Vegetable Consumption (Daily):



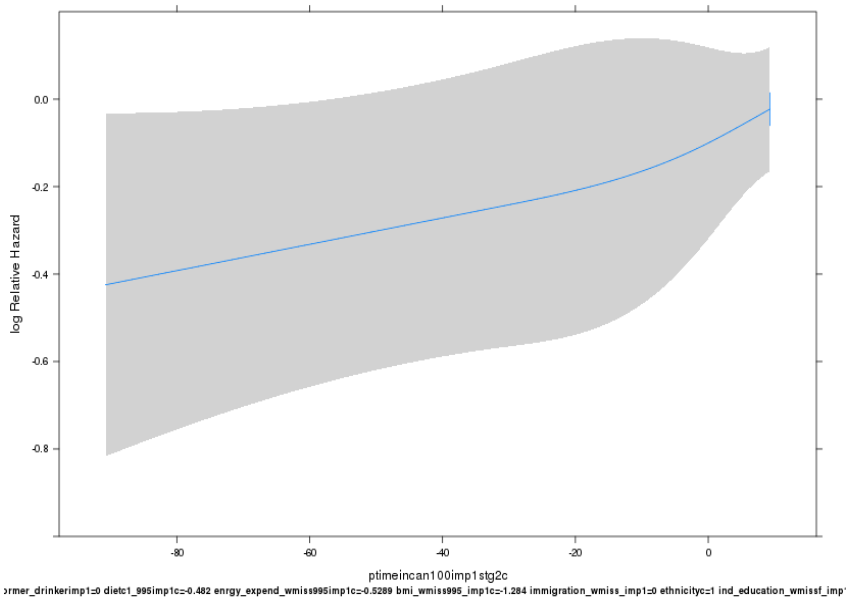
(v) Leisure Physical Activity:



(vi) BMI:

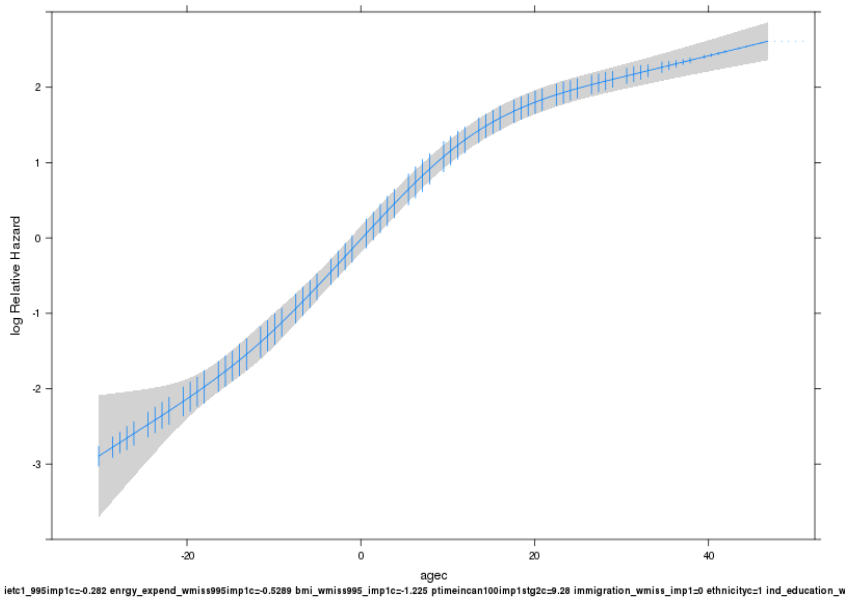


(vii) Percent Time Lived in Canada:

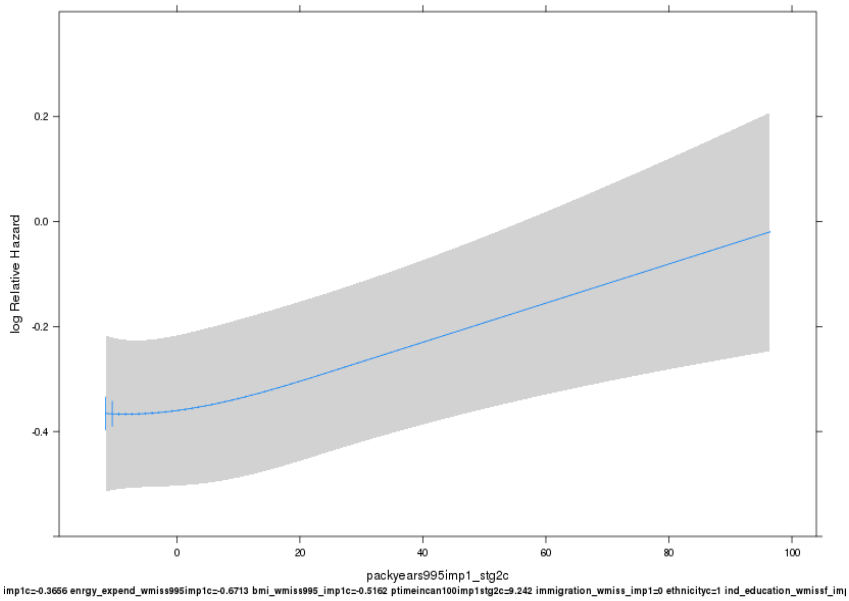


(B) Log Hazard Plots of Predictor Variables with Incident Cancer for Males in the Derivation Cohort (CCHS cycles 2-4) on Imputed Data

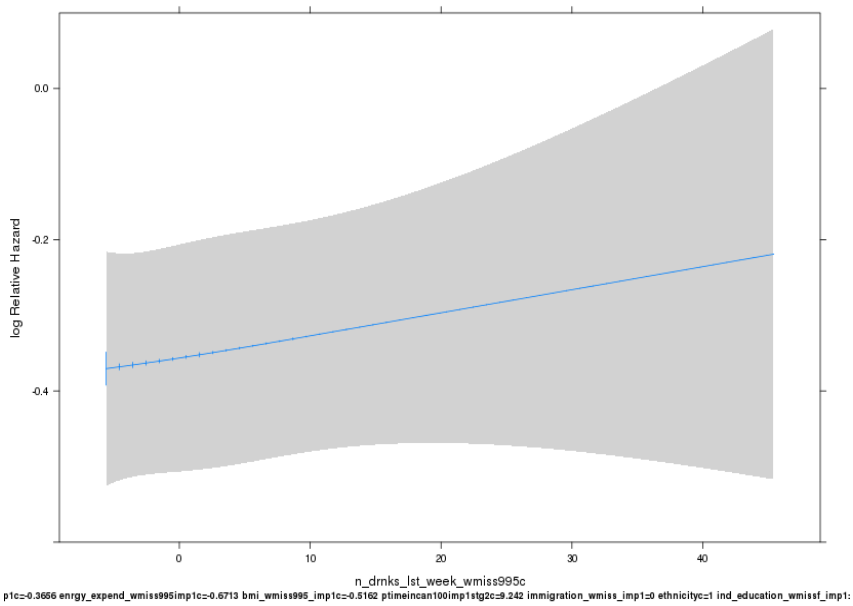
(i) Age:



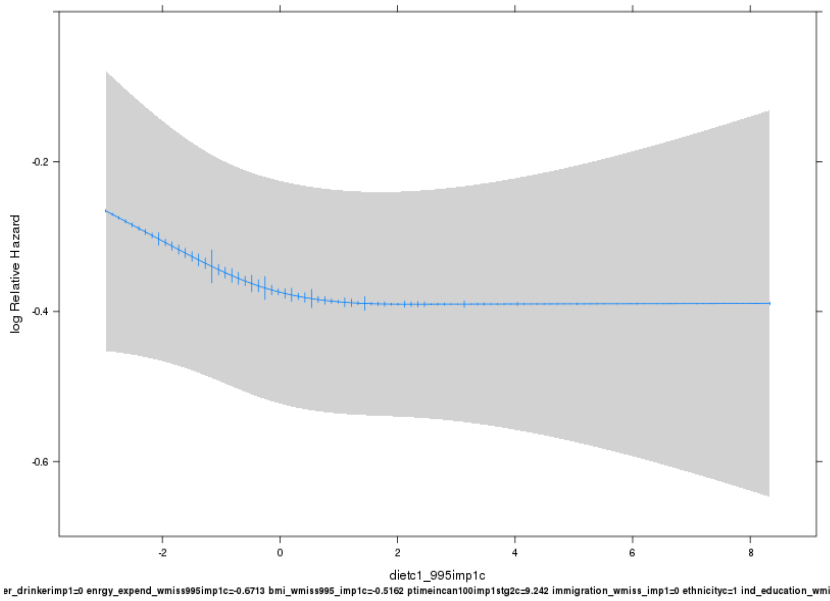
(ii) Pack-years Smoking:



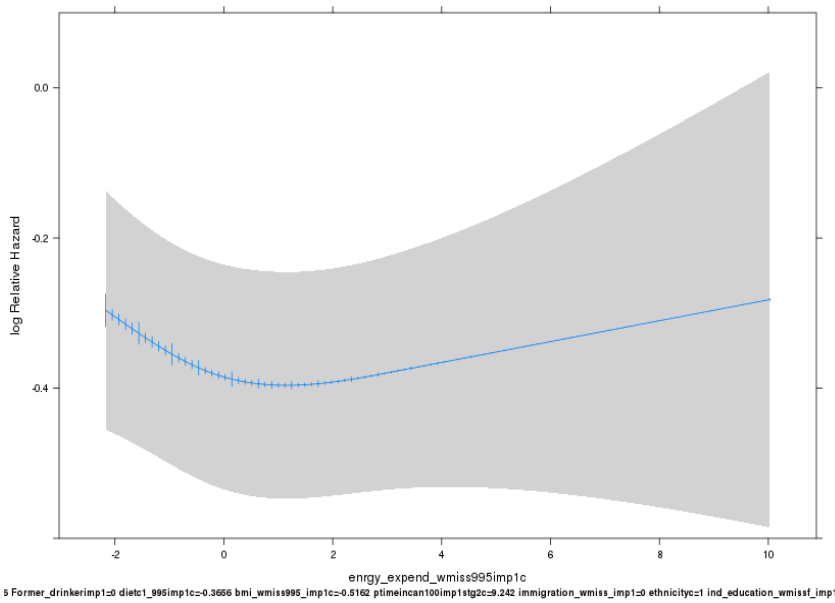
(iii)Alcohol Consumption (Number of Alcoholic Beverages Consumed Weekly):



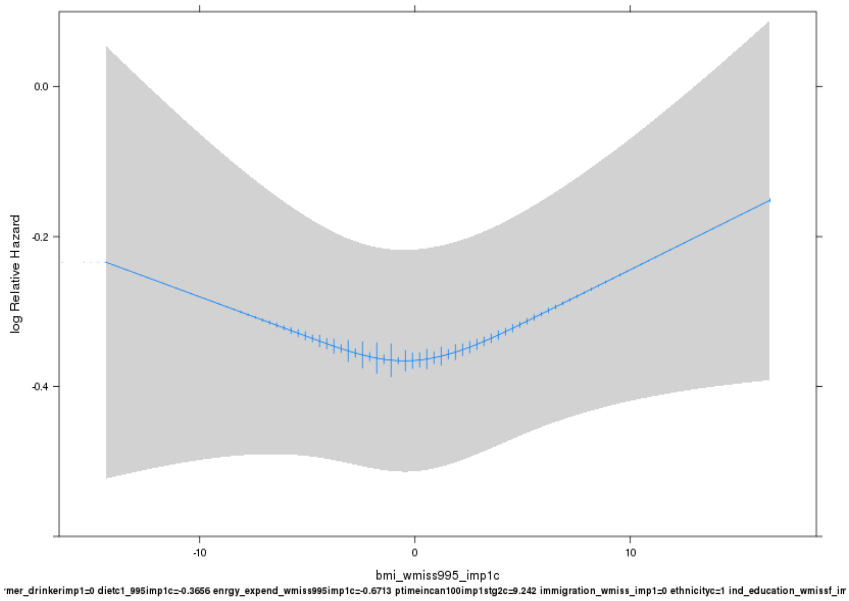
(iv) Dietary Fruit and Vegetable Consumption (Daily):



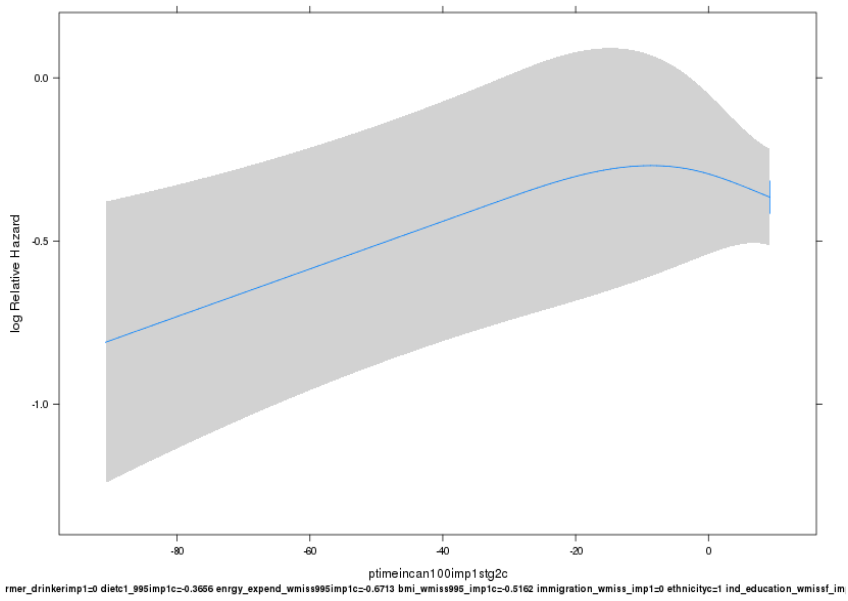
(v) Leisure Physical Activity:



(vi) BMI:



(vii) Percent Time Lived in Canada:



4.3 Model Specification:

4.3.1 Competing Risk Cox Proportional Hazard Regression Model:

The results of the competing risk regression models for females and for males over a 5-year follow-up period are presented in Table 5.

Interpretation of the relationship between each of age, pack-years of smoking, BMI, and in females leisure physical activity level, with incident cancer was limited from the Cox proportional hazard regression model due to the spline terms for these variables. Therefore, greater reliance was placed upon visual examination of the log hazard plots to assess the relationship between predictors and outcome for these variables. However, examination of other variables revealed that there was a 34% greater hazard of incident cancer among women who were immigrants relative to women who had not immigrated to Canada. There was a 1% increase in the hazard of incident cancer for every additional year lived in Canada. Those females with COPD had a 35% higher hazard of incident cancer than those without.

Among males, there was a 0.7% increased hazard of incident cancer for every one increase in number of alcoholic beverages consumed in the prior week. There was a borderline statistically significant 2% decreased hazard of incident cancer with each additional daily fruit and vegetable consumed, as well as a borderline 0.5% increased hazard of incident cancer for each additional year lived in Canada.

For both males and females, at least one category of smoking status had a significant interaction with age, while an interaction between age and pack-years of smoking was not observed for either sex.

The highly elevated hazard ratio observed for pack-years of smoking status in females prompted further investigation. Upon inspection, the large hazard ratio was the result of the close positioning of two knots in the restricted cubic spline due to the underlying highly skewed distribution of pack-years of smoking. The multivariate Cox proportional hazard model was conducted again for females, but with pack-years of smoking collapsed to a linear term. The hazard ratio for pack-years of smoking became HR=1.002 (1.00-1.01). The remainder of the multivariate-adjusted hazard ratios in the model remained largely unchanged, as did the measures of model performance including observed versus predicted risk across a variety of subgroups (data not shown).

Table 5: Multivariate Competing-Risk Cox Proportional Hazard Model for Incident Cancer Development over a 5-Year Follow-up Period in the Derivation Cohort

	Incident Malignancy			
	Female		**Male**	
	Multivariate-adjusted HR (95% CI)	**P value for trend**	**Multivariate-adjusted HR (95% CI)**	**P value for trend**
Age	1.13 (1.07-1.20)	*<0.0001*	1.09 (1.00-1.18)	0.049
Age'	0.88 (0.67-1.16)	0.37	1.06 (0.73-1.55)	0.75
Age''	1.21 (0.57-2.55)	0.62	1.04 (0.34-3.23)	0.94
Age'''	0.90 (0.44-1.85)	0.77	0.53 (0.18-1.60)	0.26
Pack-years smoking	0.99 (0.97-1.02)	0.59	1.00 (0.99-1.01)	0.98
Pack-years smoking'	>1000 (3.37 e-10 – 3.39 e20)	0.47	1.09 (0.82-1.43)	0.56
Former smoker, quit >5 years	1.16 (0.94-1.43)	0.15	0.87 (0.66-1.14)	0.31

ago				
Former smoker, quit ≤5 years ago	1.48 (1.13-1.93)	*0.0041*	0.96 (0.68-1.37)	0.85
Current Smoker	1.60 (1.24-2.07)	*0.0003*	1.09 (0.81-1.48)	0.54
Second Hand Smoke Exposure	1.04 (0.90-1.19)	0.61	1.08 (0.95-1.22)	0.24
Alcohol consumption weekly	1.00 (0.99-1.02)	0.50	1.007 (1.001-1.01)	*0.032*
Former alcohol drinker	1.00 (0.88-1.14)	0.98	1.06 (0.93-1.22)	0.38
Dietary consumption of fruit and vegetables daily	1.01 (0.98-1.03)	0.68	0.98 (0.95-1.00)	0.099
Leisure physical activity (daily METS)	0.99 (0.91-1.07)	0.75	1.00 (0.97-1.03)	0.99
Leisure physical activity (daily METS)'	1.02 (0.91-1.15)	0.71	N/A	N/A
Body Mass Index (BMI)	0.99 (0.96-1.01)	0.39	1.00 (0.97-1.04)	0.82
BMI'	1.03 (0.99-1.06)	0.14	1.00 (0.97-1.04)	0.86
Percent of life lived in Canada	1.01 (1.00-1.01)	*0.0053*	1.005 (1.00-1.01)	0.072
Immigrant	1.34 (1.08-1.67)	*0.0080*	1.17 (0.95-1.44)	0.14
Caucasian ethnicity	0.93 (0.75-1.15)	0.49	1.05 (0.85-1.30)	0.66
Post-secondary education	0.93 (0.84-1.03)	0.16	0.98 (0.88-1.08)	0.64
Deprivation	1.00 (0.92-1.09)	0.98	1.02 (0.94-1.11)	0.64
Emphysema or COPD	1.35 (1.08-1.69)	*0.0084*	0.91 (0.71-1.17)	0.46
Inflammatory bowel disease	1.07 (0.90-1.28)	0.42	0.93 (0.69-1.25)	0.64
Survey cycle 3	1.08 (0.95-1.22)	0.25	0.96 (0.85-1.09)	0.52
Survey cycle 4	1.11 (0.98-1.25)	0.11	0.97 (0.86-1.10)	0.65
Age x Pack-years Smoking	1.00 (1.00-1.00)	0.78	1.00 (1.00-1.00)	0.46
Age x Former smoker, quit >5 years ago	1.01 (1.00-1.02)	0.23	1.01 (1.00-1.02)	*0.047*
Age x Former smoker, quit ≤5	1.01 (1.00-1.02)	0.12	1.01 (1.00-1.03)	0.11

years ago				
Age x Current Smoker	1.01 (1.00-1.02)	*0.033*	1.01 (1.00-1.03)	0.058

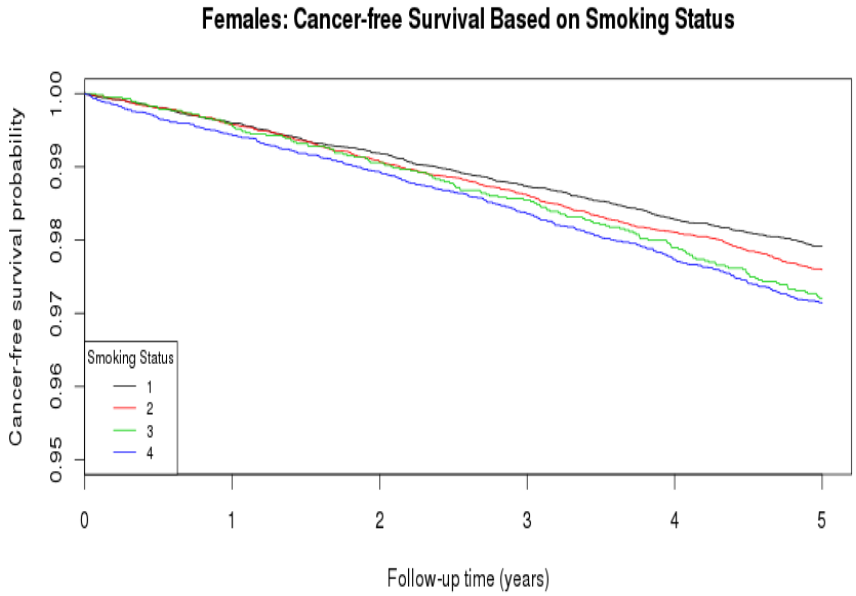
4.3.2 Proportional Hazard Assumption:

The proportional hazard assumption was assessed for each predictor using scaled Schoenfeld residuals over time. The proportional hazard assumption was not violated, as evidenced by the random pattern over time for each covariate examined. (Appendix 4)

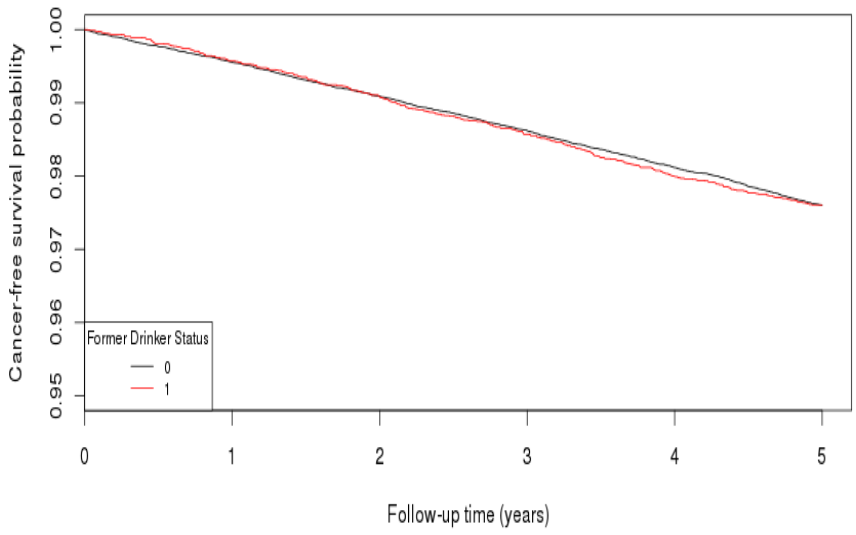
4.3.3 Cancer-free Survival Curves based on Categorical Predictor Variables:

Survival graphs were generated by creating a Cox proportional hazard model (eliminating the competing risk) for both males and females, for the purposes of graphical representation of risk based on different levels of categorical predictor variables. The survival curves based on various predictive factors are in Figure 4. At 5 years of follow-up time, cancer-free survival curves for males and females show the greatest survival was among non-smokers, then decreased in descending order for former smokers who had quit >5 years ago, former smokers who had quit ≤5 years ago and current smokers. Male and female cancer-free survival curves also showed differences with greater survival among non-immigrants relative to immigrants, and greater survival for those without COPD relative to those with COPD. Importantly, these graphs were generated using a non-competing risk model and the observed differences between groups may not be statistically significant.

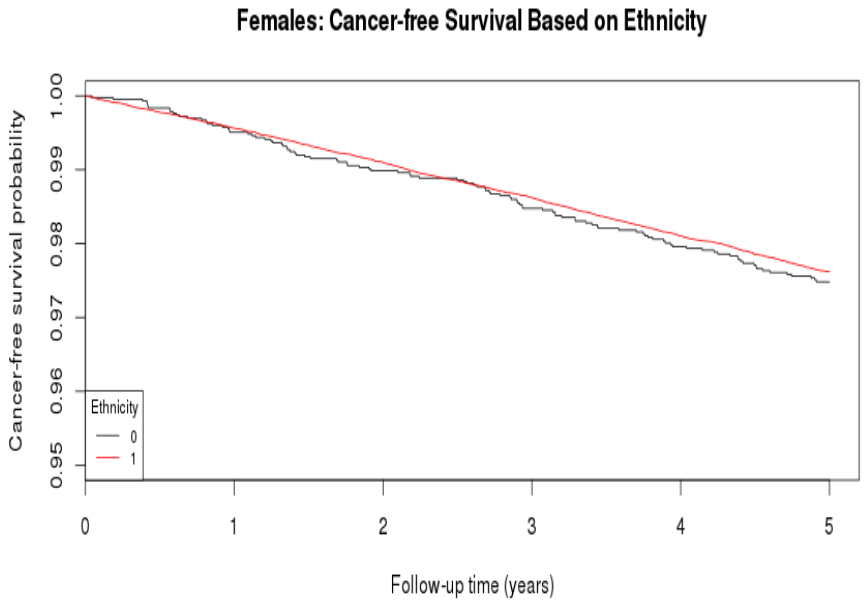
Figure 4: (A) Female Cancer-free Survival Curves in the Derivation Cohort:



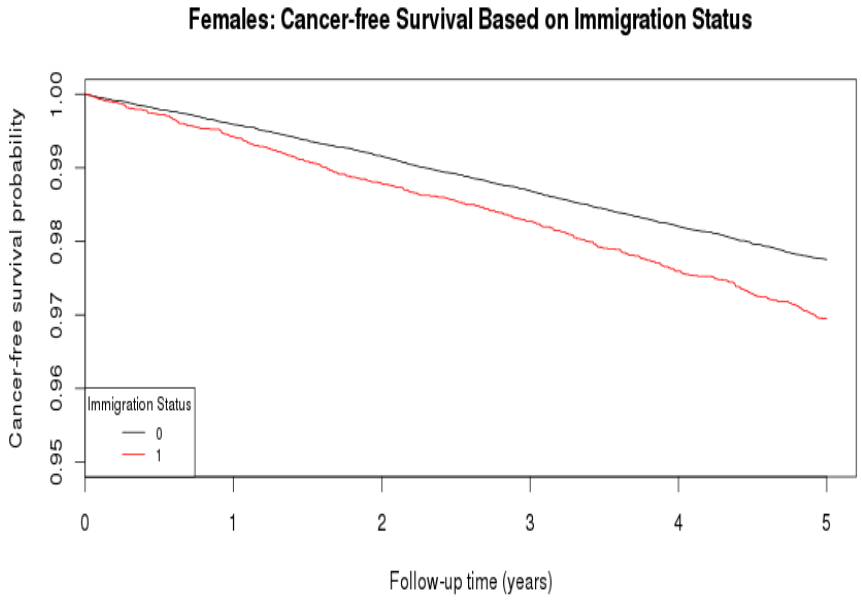
Legend- Smoking Status:
1=non-smoker
2=former smoker, quit >5 years ago
3=former smoker, quit ≤5 years ago
4=current smoker



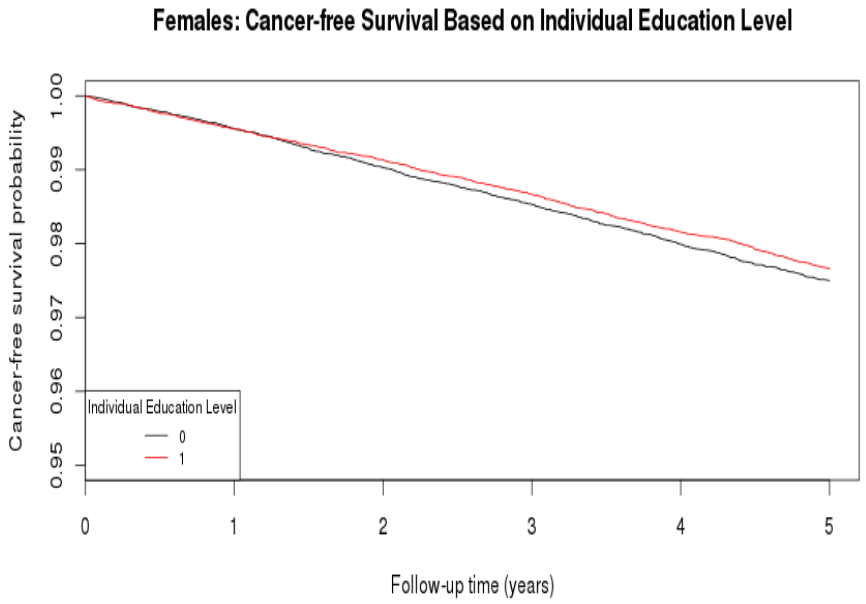
Legend- Former Drinker Status:
0=not a former drinker
1=former drinker defined as having not drank alcohol in the past 12 months but having consumed alcohol at some point in the past



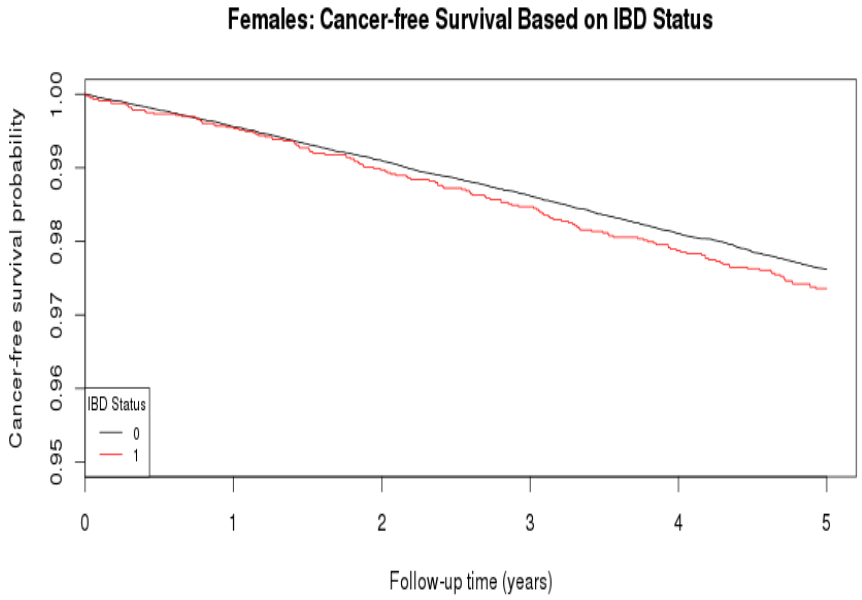
Legend- Ethnicity:
0=All ethnicities other than Caucasian
1=Caucasian



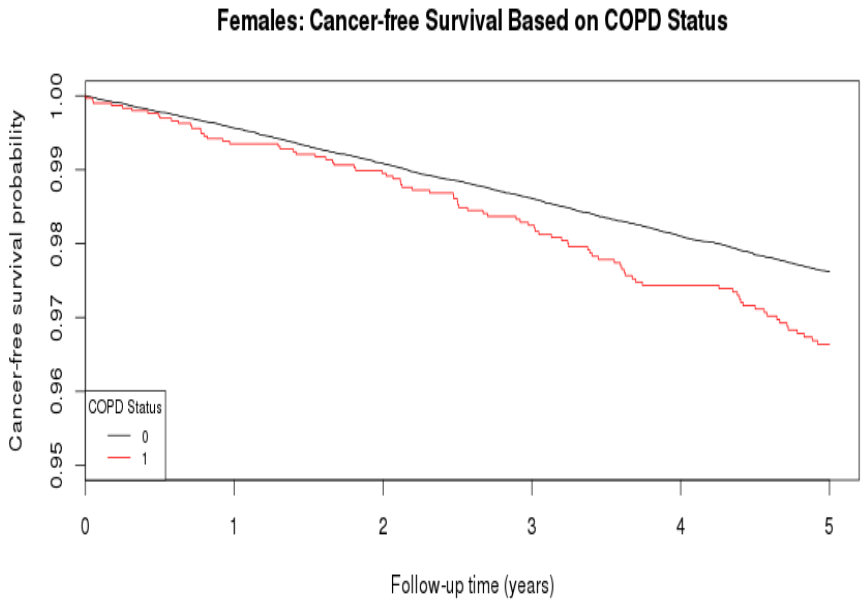
Legend-Immigration Status:
0=Non-Immigrant
1=Immigrant



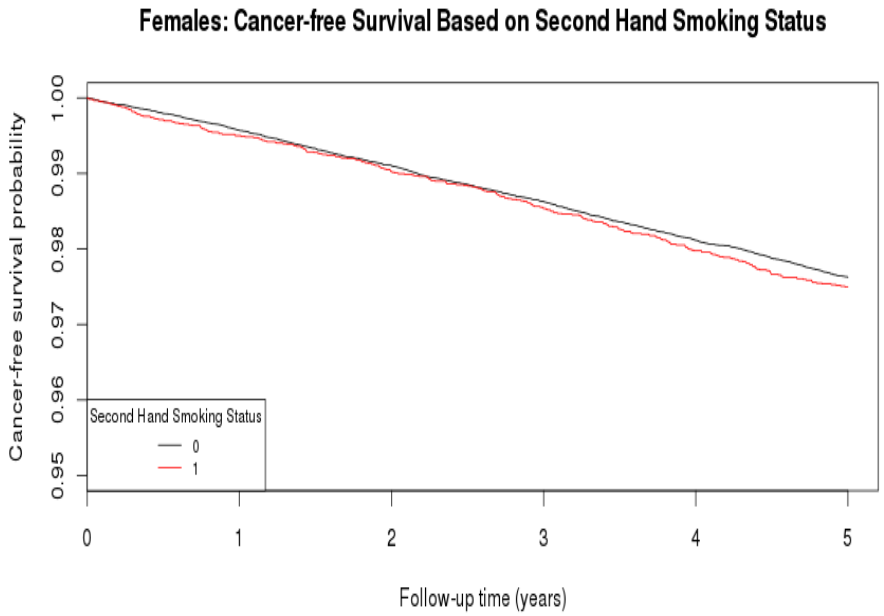
Legend- Education Level:
0= Secondary school graduation or less
1=Post-secondary education



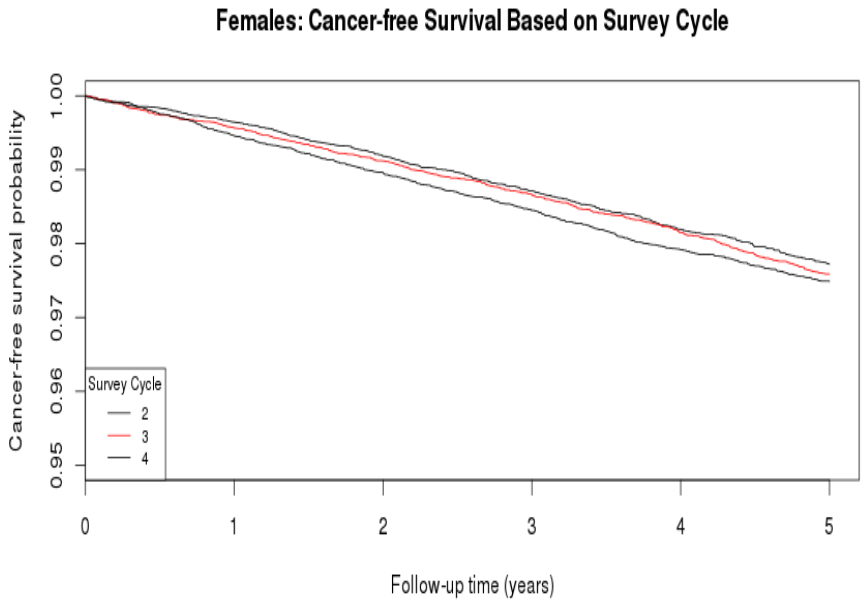
Legend- Inflammatory Bowel Disease (IBD) Status:
0=no history of IBD
1=history of IBD



Legend-COPD Status:
0=no history of COPD
1=history of COPD

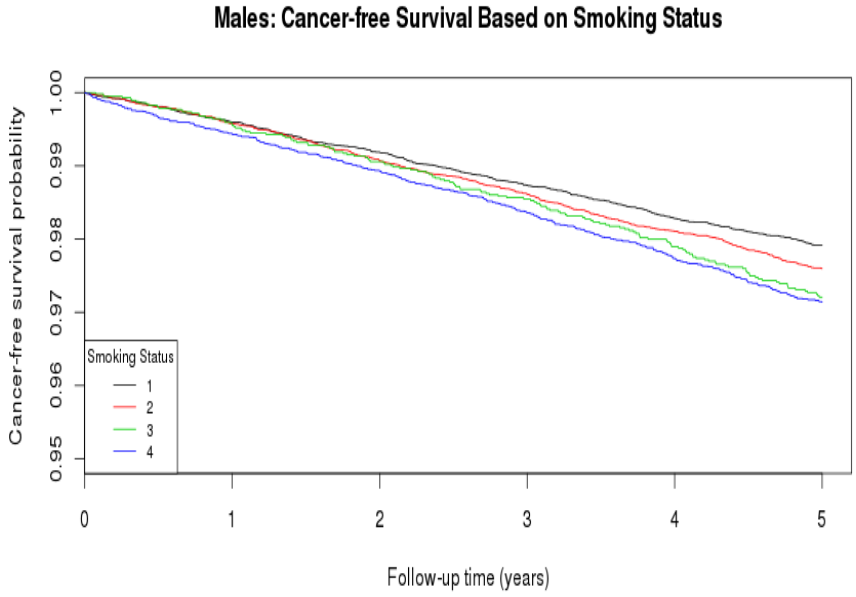


Legend- Second Hand Smoking Status:
0=No second hand smoke exposure
1=Second hand smoke exposure

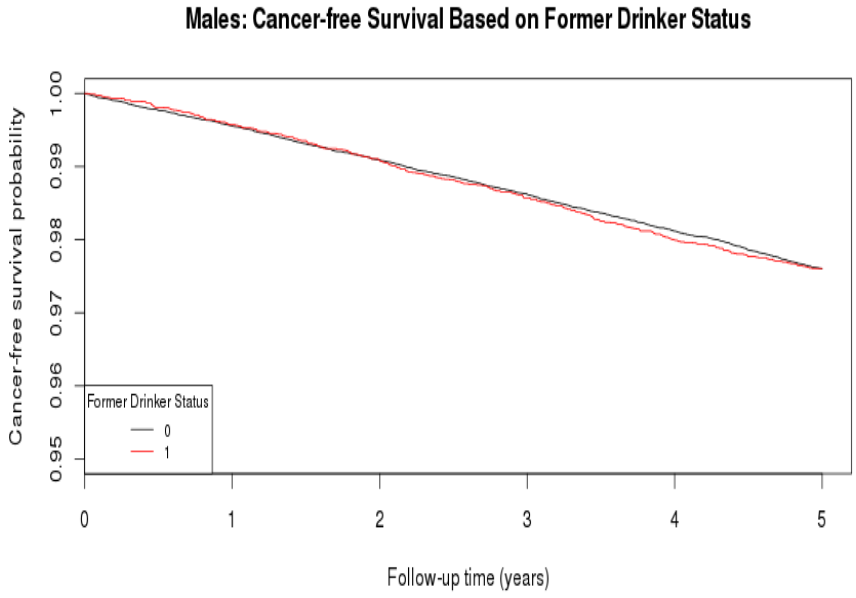


Legend-Survey Cycle:
2= CCHS Survey Cycle 2.1
3= CCHS Survey Cycle 3.1
4= CCHS Survey Cycle 4.1

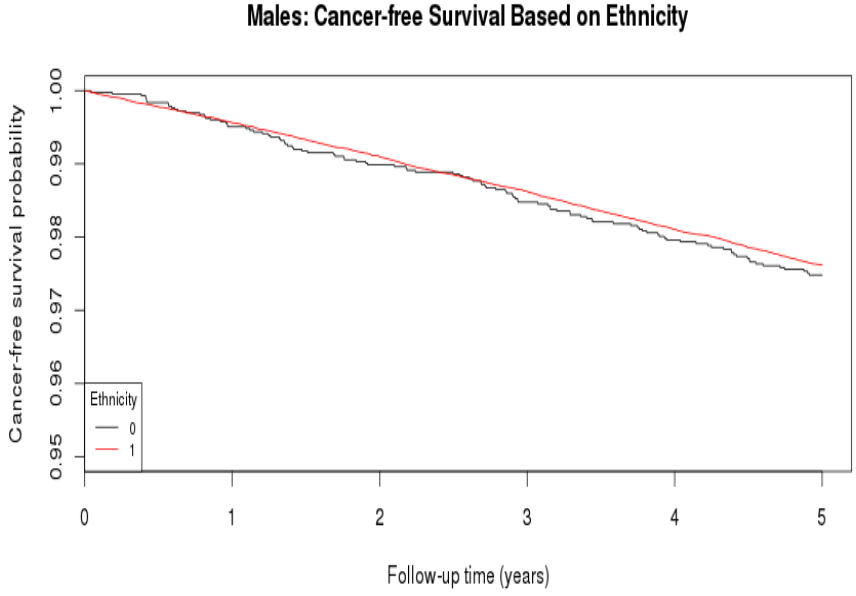
(B) Male Cancer-free Survival Curves in the Derivation Cohort:



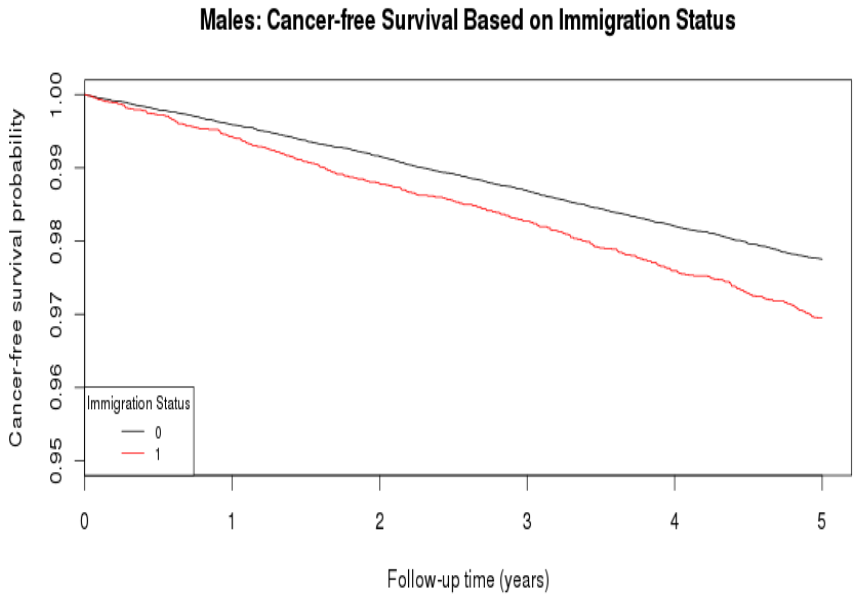
Legend- Smoking Status:
1=non-smoker
2=former smoker, quit >5 years ago
3=former smoker, quit ≤5 years ago
4=current smoker



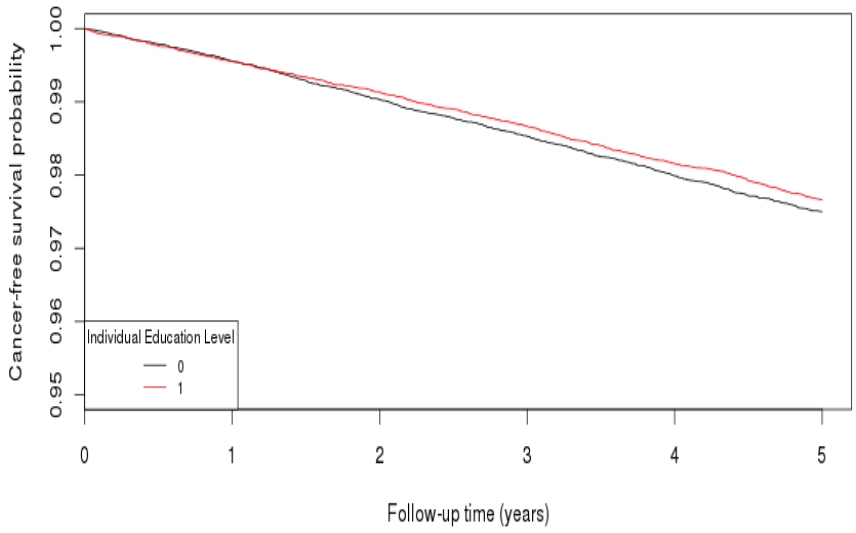
Legend- Former Drinker Status:
0=not a former drinker
1=former drinker defined as having not drank alcohol in the past 12 months but having consumed alcohol at some point in the past



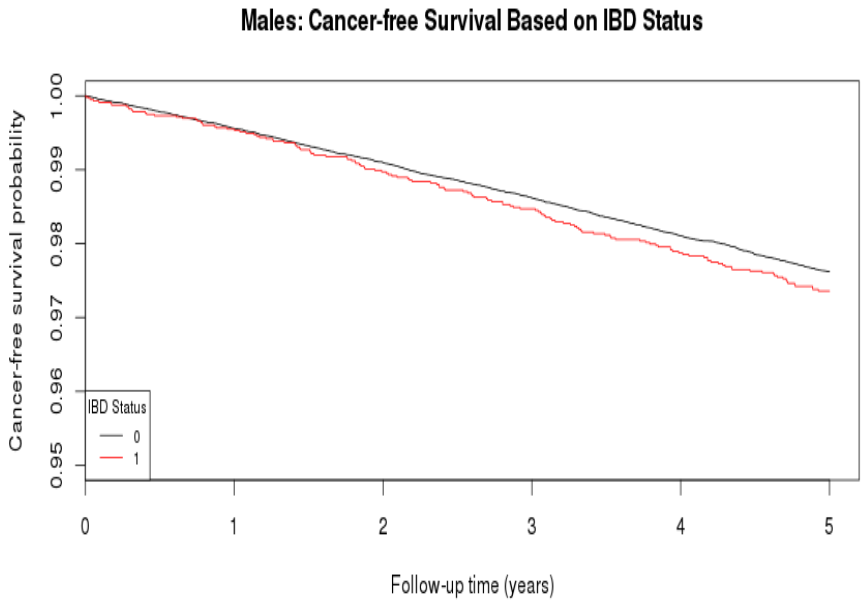
Legend- Ethnicity:
0=All ethnicities other than Caucasian
1=Caucasian



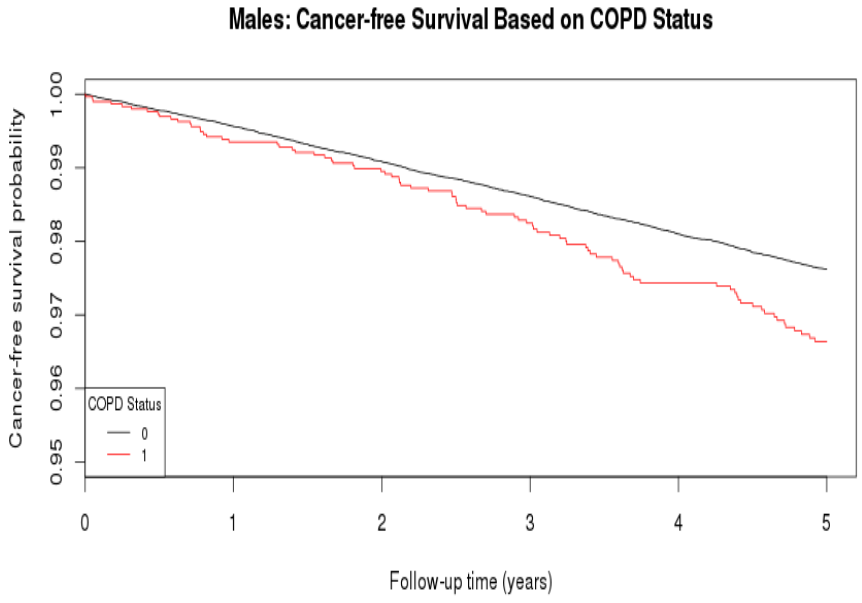
Legend-Immigration Status:
0=Non-Immigrant
1=Immigrant



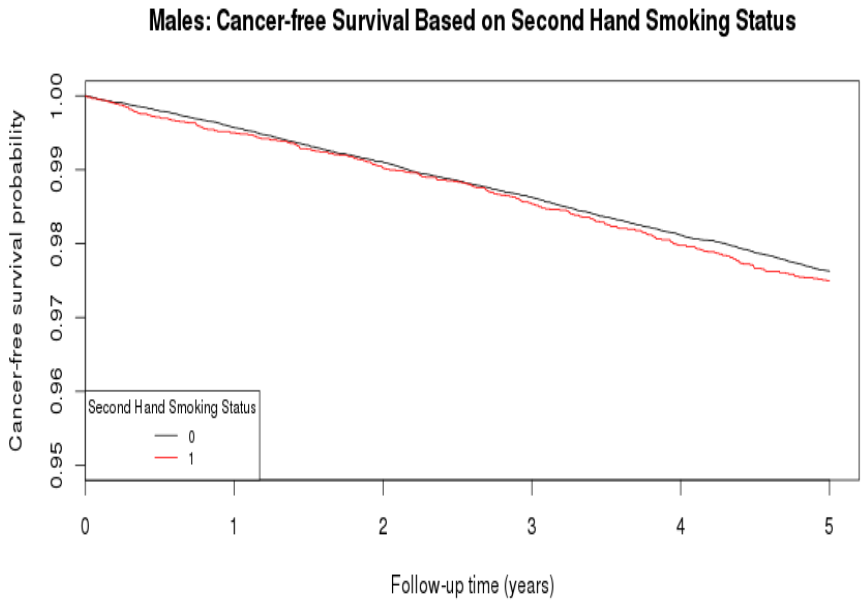
Legend- Education Level:
0= Secondary school graduation or less
1=Post-secondary education



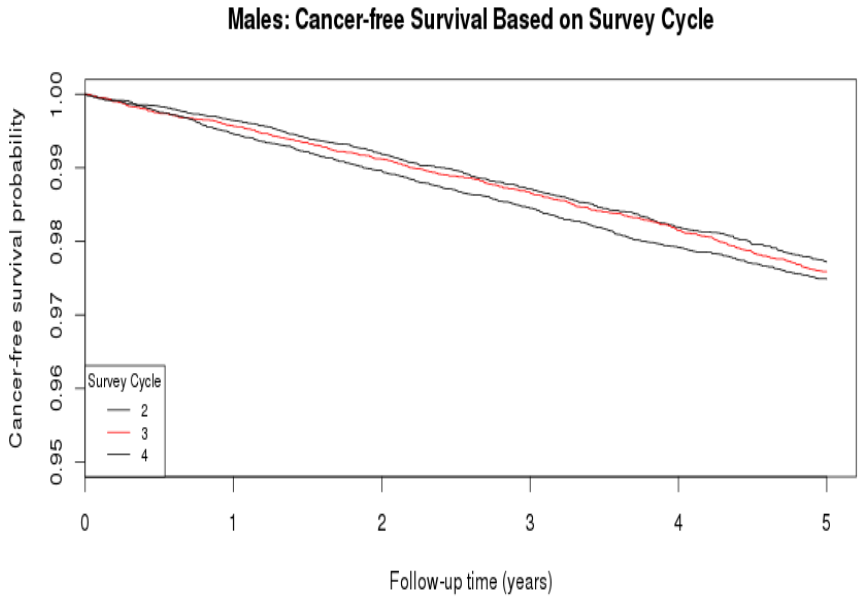
Legend- Inflammatory Bowel Disease (IBD) Status:
0=no history of IBD
1=history of IBD



Legend-COPD Status:
0=no history of COPD
1=history of COPD



Legend- Second Hand Smoking Status:
0=No second hand smoke exposure
1=Second hand smoke exposure

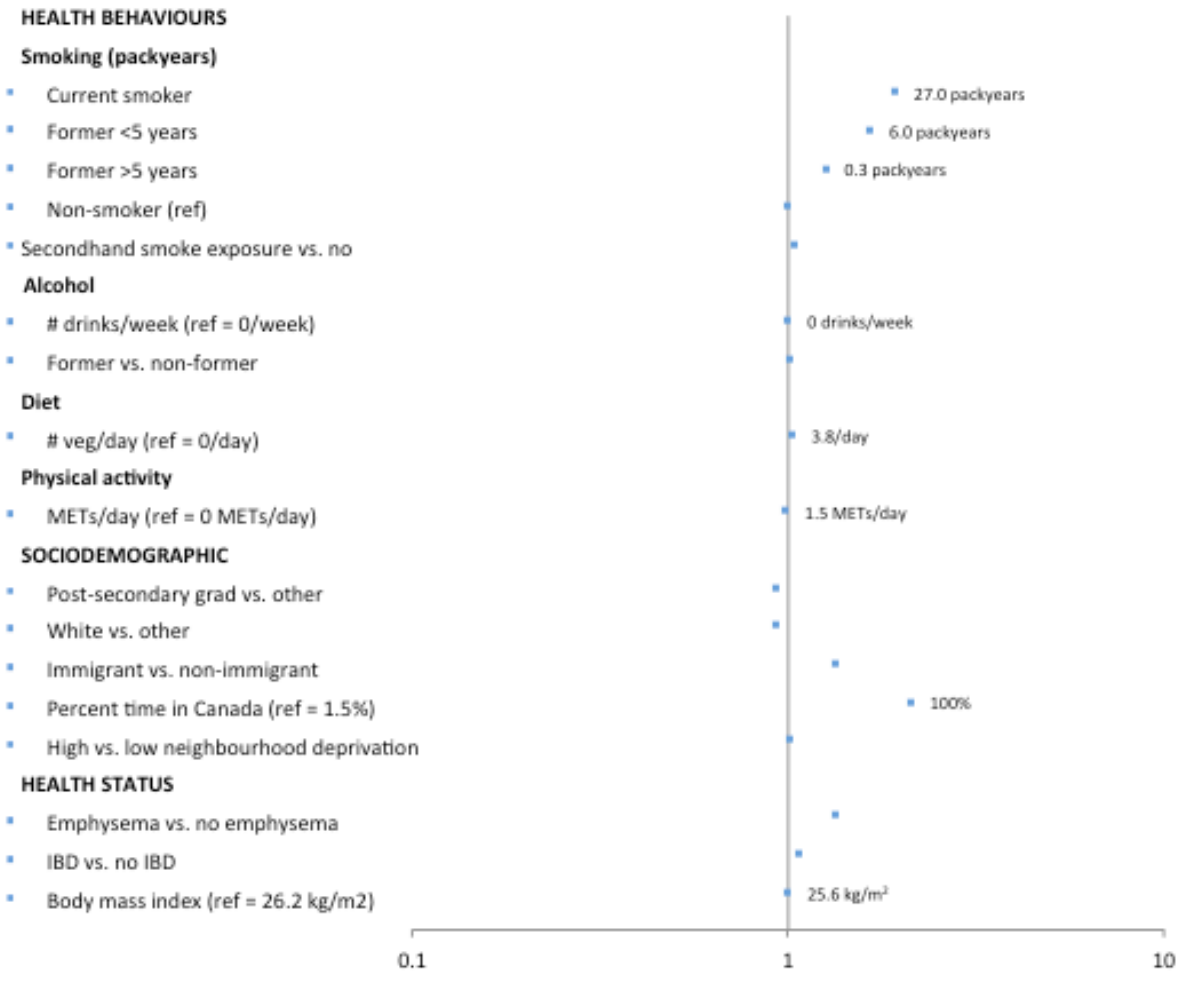


Legend-Survey Cycle:
2= CCHS Survey Cycle 2.1
3= CCHS Survey Cycle 3.1
4= CCHS Survey Cycle 4.1

4.3.4 Illustration of Risk of Incident Cancer Based on Median Exposure Level:

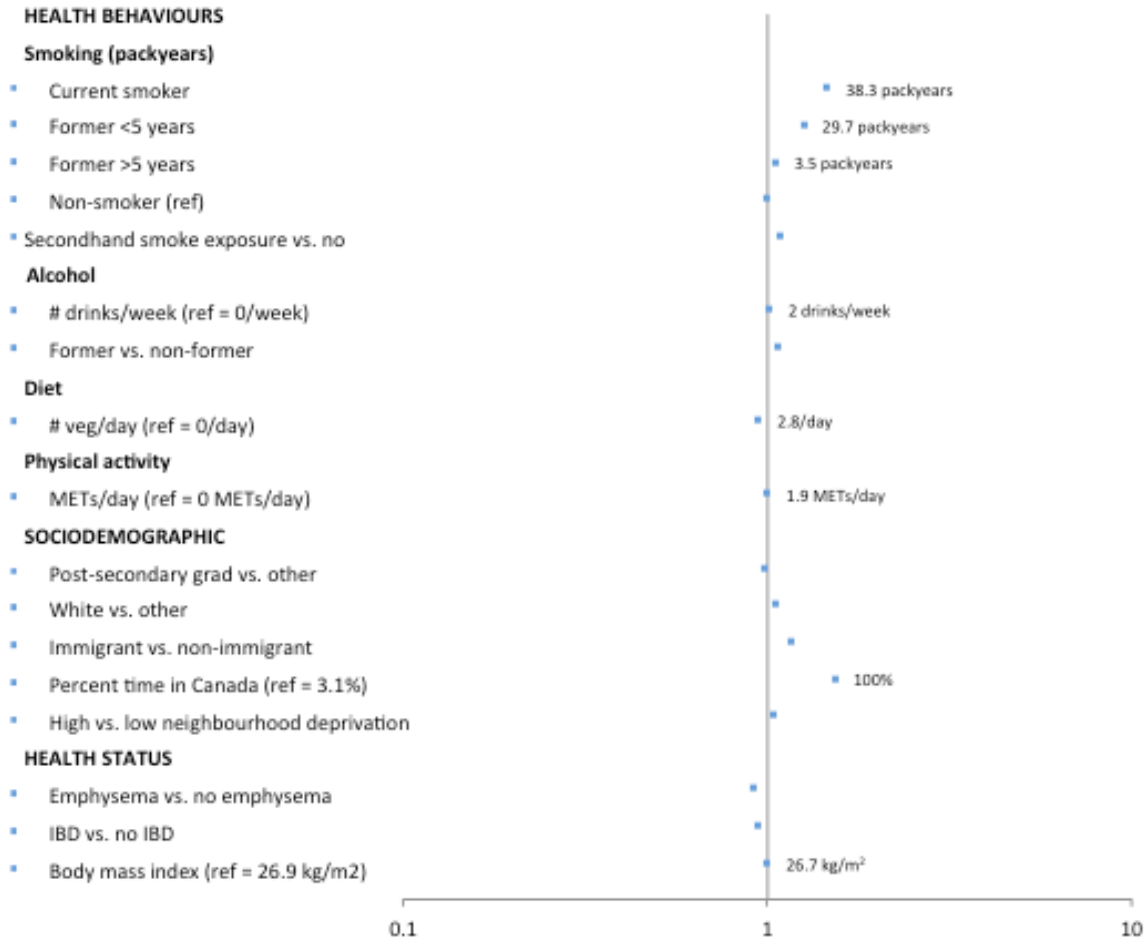
The relative predicted risk of incident cancer is portrayed in the following example, indicating the risk based on median exposure level for each predictor variable in a 65 year old female and in a 65 year old male. (see Figure 5)

Figure 5:
(A) Relative Risk of Cancer Based on Median Exposure Level for 65 year old Female



Legend:
X-axis indicates relative predicted risk of incident cancer
Notations adjacent to plotted points indicate the median level of exposure for each predictor variable.

(B) Relative Risk of Cancer Based on Median Exposure Level for 65 year old Male



Legend:
X-axis indicates relative predicted risk of incident cancer
Notations adjacent to plotted points indicate the median level of exposure for each predictor variable.

4.4 Model Performance:

4.4.1 Model Performance in the Derivation Cohort:

The Nagelkerke R^2 score using the Cox proportional hazard model for females was 0.056 and for males was 0.097. The Brier Score for females was 0.034 and for males was 0.038.

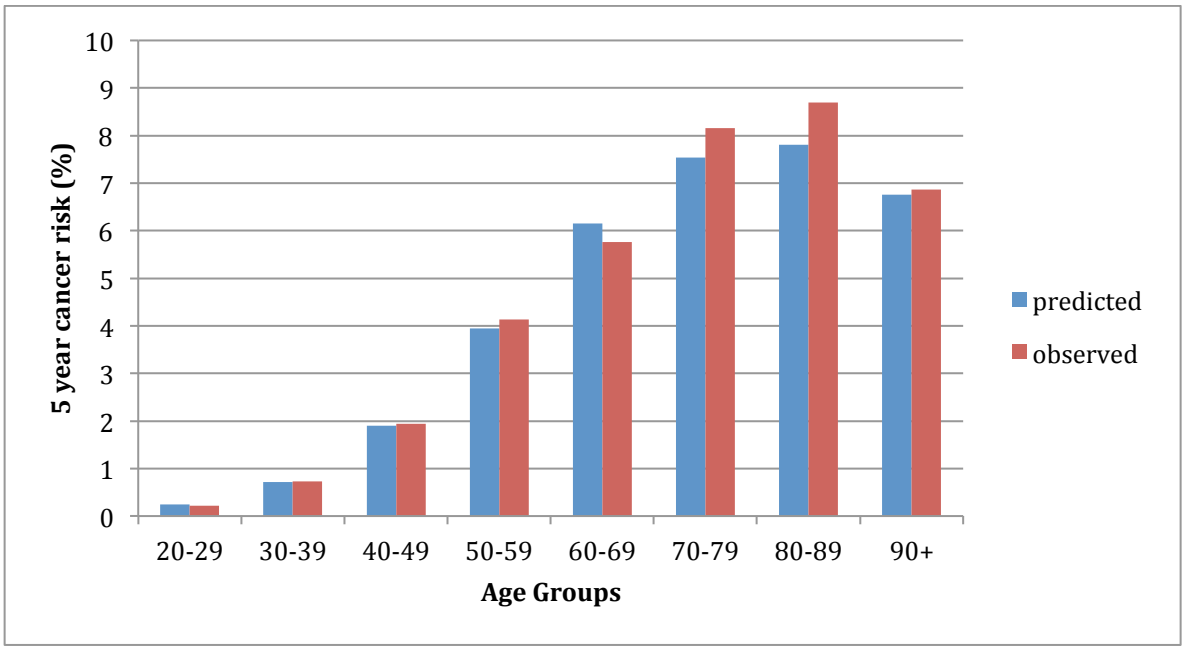
Calibration in the overall population of females was based on an observed number of 1559 incident cancer cases and a predicted number of 1514.02 at 5 years of follow-up, therefore the algorithm slightly under-predicted risk of cancer in the population. The difference between overall observed and predicted incident cancer for females was 2.97%. Among males, the observed number of incident cancer cases was 1544 and the predicted risk was 1481.41 at 5 years of follow-up. Again, the algorithm under-predicted cancer incidence by a difference of 4.23%.

The observed versus predicted incident cancer cases for both males and females across the subgroups of age (10 year increments), predicted risk deciles, smoking category, pack-year smoking tertiles, dietary fruit and vegetable consumption quartiles, energy expenditure quartiles, alcohol tertiles, former alcohol drinker status, BMI quartiles, emphysema/COPD status, IBD status, ethnicity, immigration status and survey cycle are graphically depicted using histograms in Figure 6.

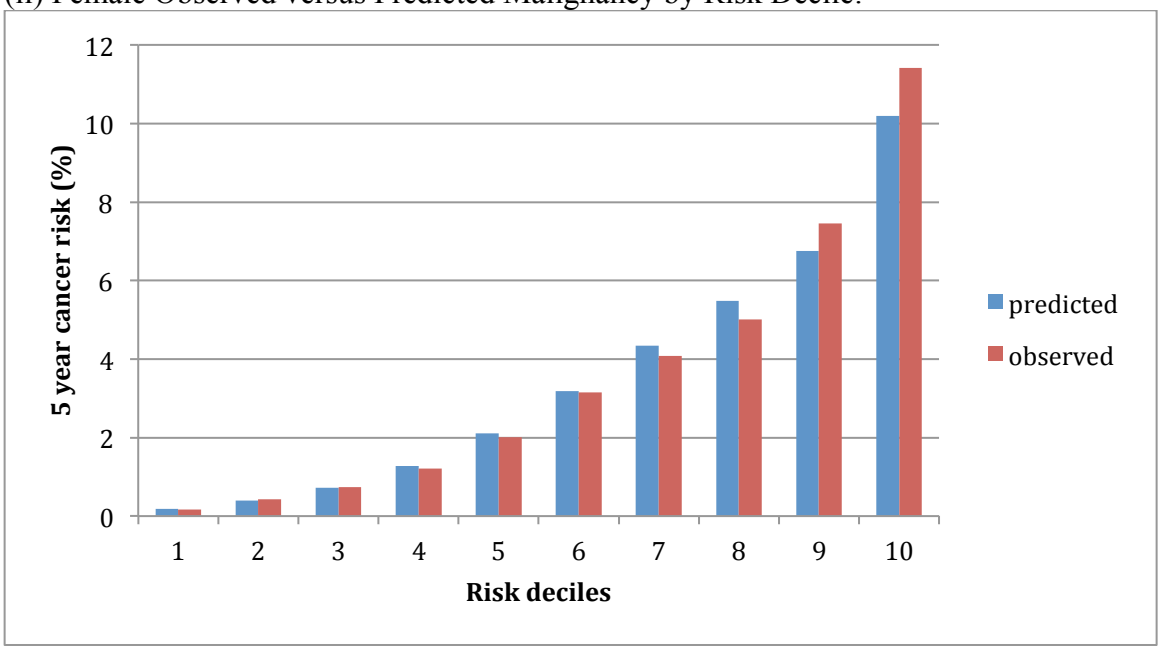
Across 10-year age increments, the percentage difference between observed versus predicted risk ranged from 1.69% to 11.34% for females. Across predicted risk deciles for females, the percentage difference between observed versus predicted risk ranged from 0.57% to 13.37%. For all other subgroups analyzed among females, the difference in predicted versus observed risk was always <20% which was defined as the cut-off for adequate calibration.

Figure 6:
(A) Female Observed versus Predicted Risk of Incident Cancer

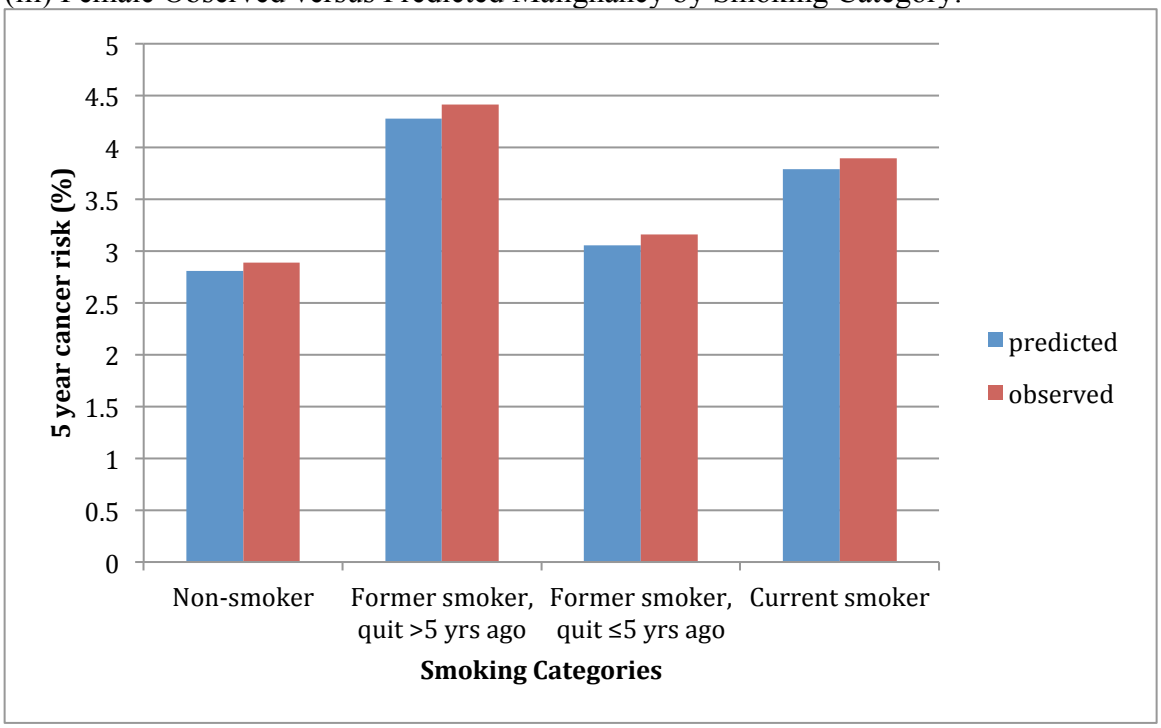
(i) Female Observed versus Predicted Malignancy by Age Group:



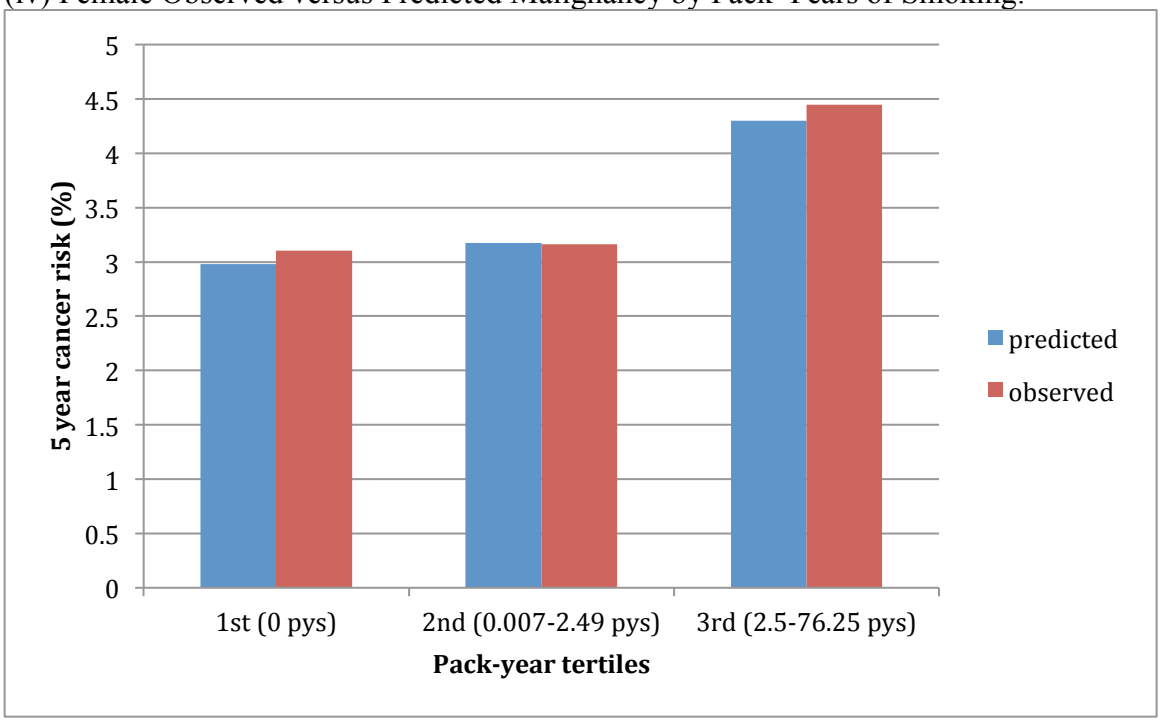
(ii) Female Observed versus Predicted Malignancy by Risk Decile:



(iii) Female Observed versus Predicted Malignancy by Smoking Category:

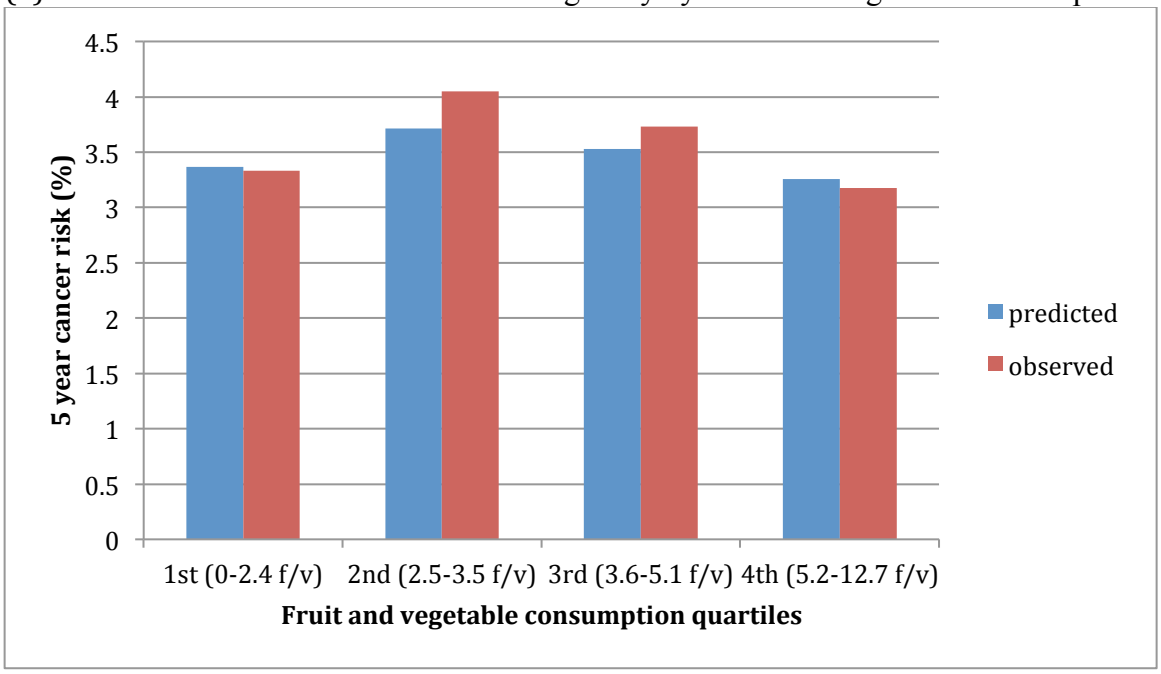


(iv) Female Observed versus Predicted Malignancy by Pack-Years of Smoking:



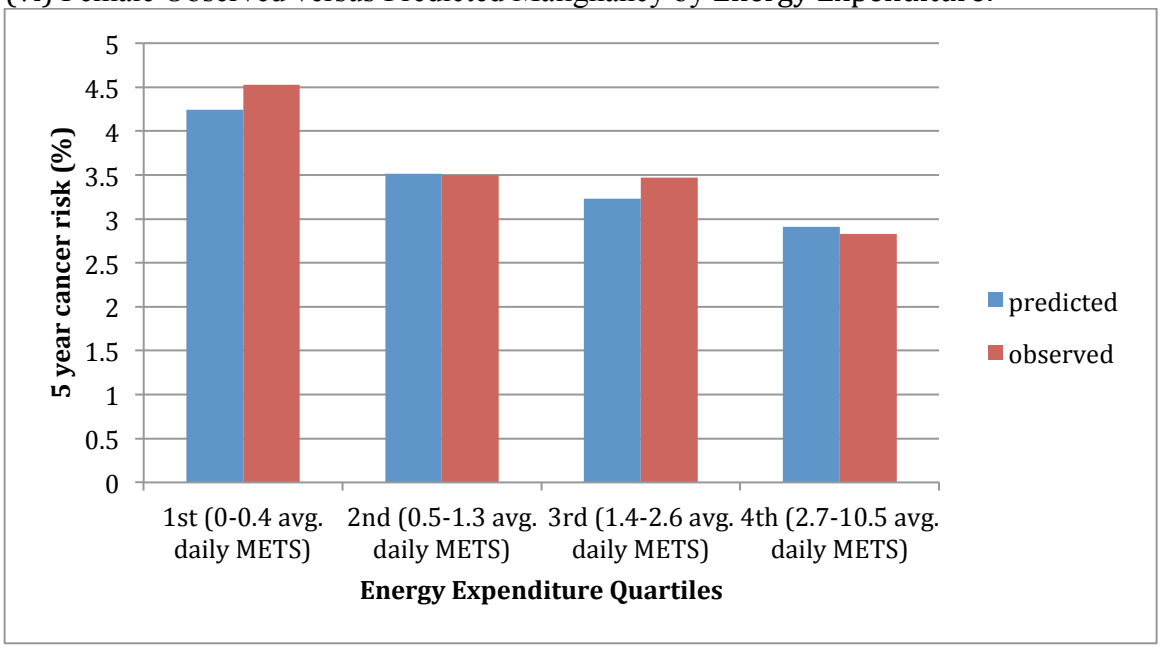
Legend: pys = pack-years of smoking

(v) Female Observed versus Predicted Malignancy by Fruit and Vegetable Consumption:

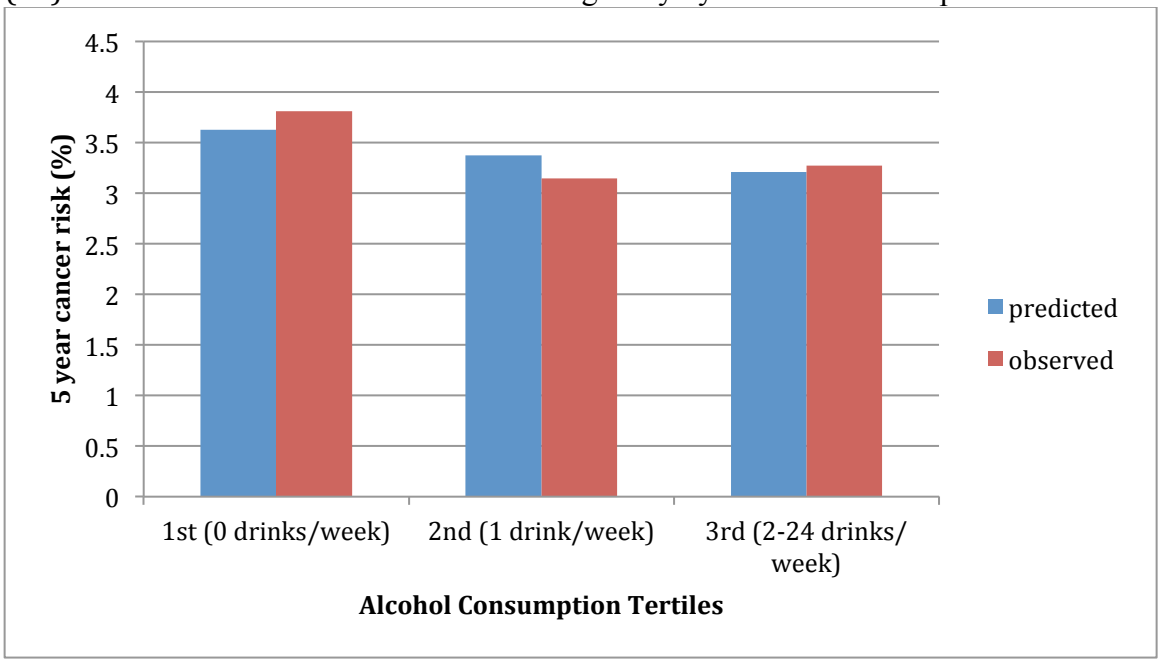


Legend: f/v = fruit and vegetable consumption

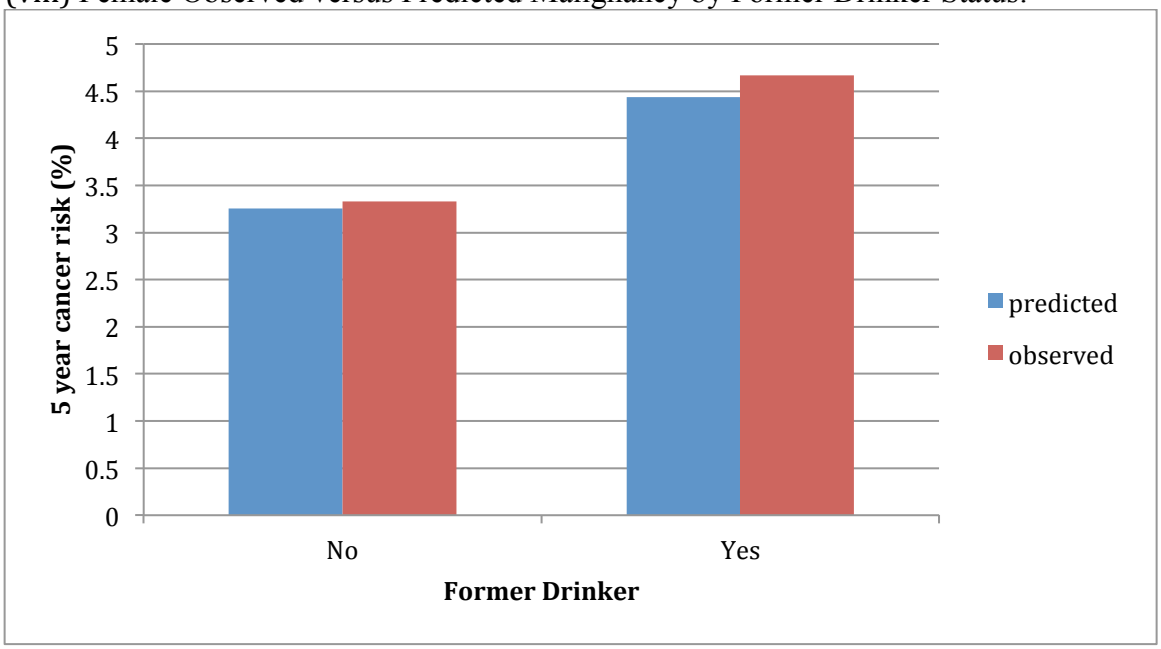
(vi) Female Observed versus Predicted Malignancy by Energy Expenditure:



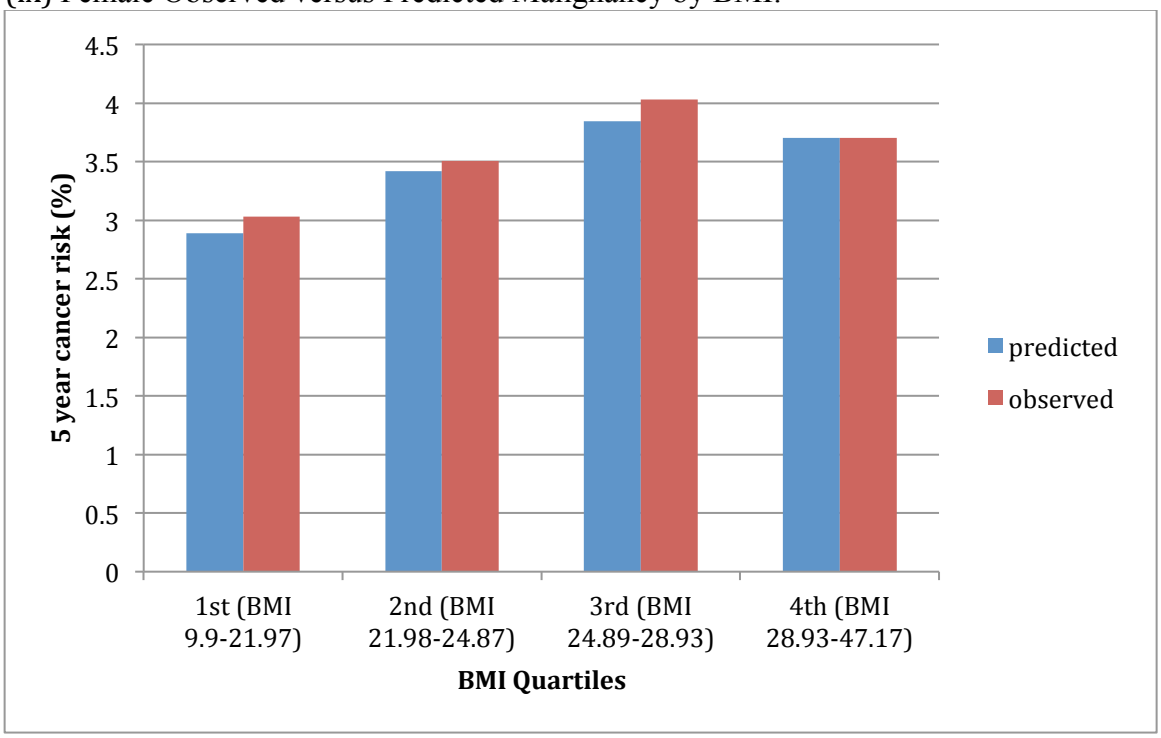
(vii) Female Observed versus Predicted Malignancy by Alcohol Consumption:



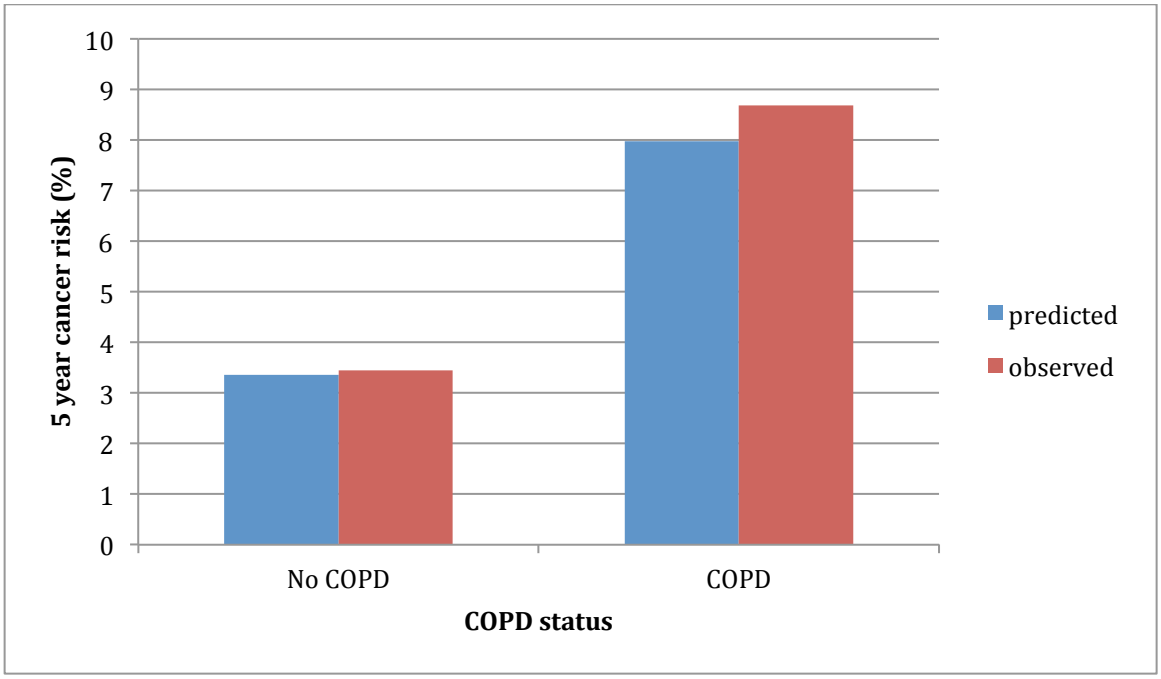
(viii) Female Observed versus Predicted Malignancy by Former Drinker Status:



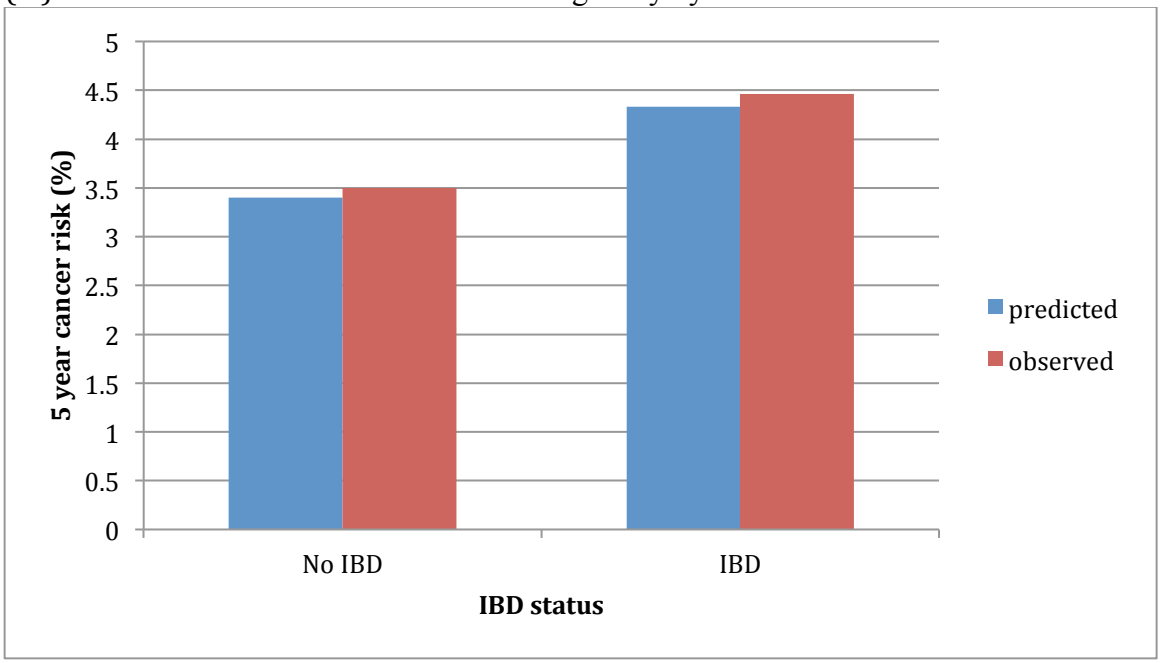
(ix) Female Observed versus Predicted Malignancy by BMI:



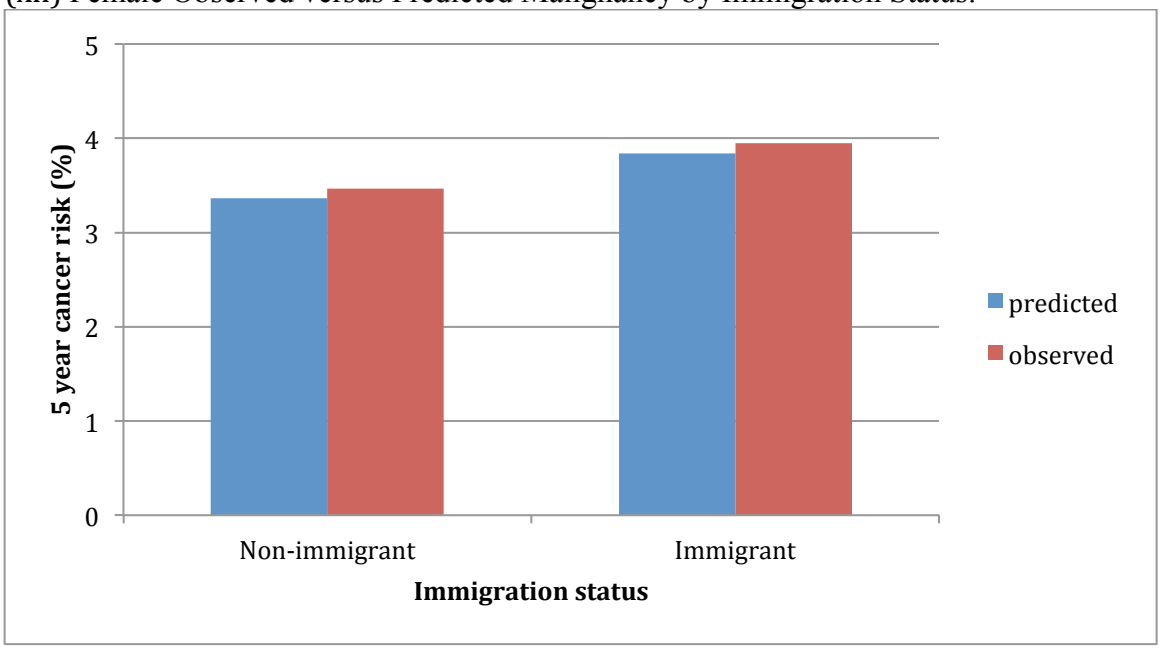
(x) Female Observed versus Predicted Malignancy by COPD Status:



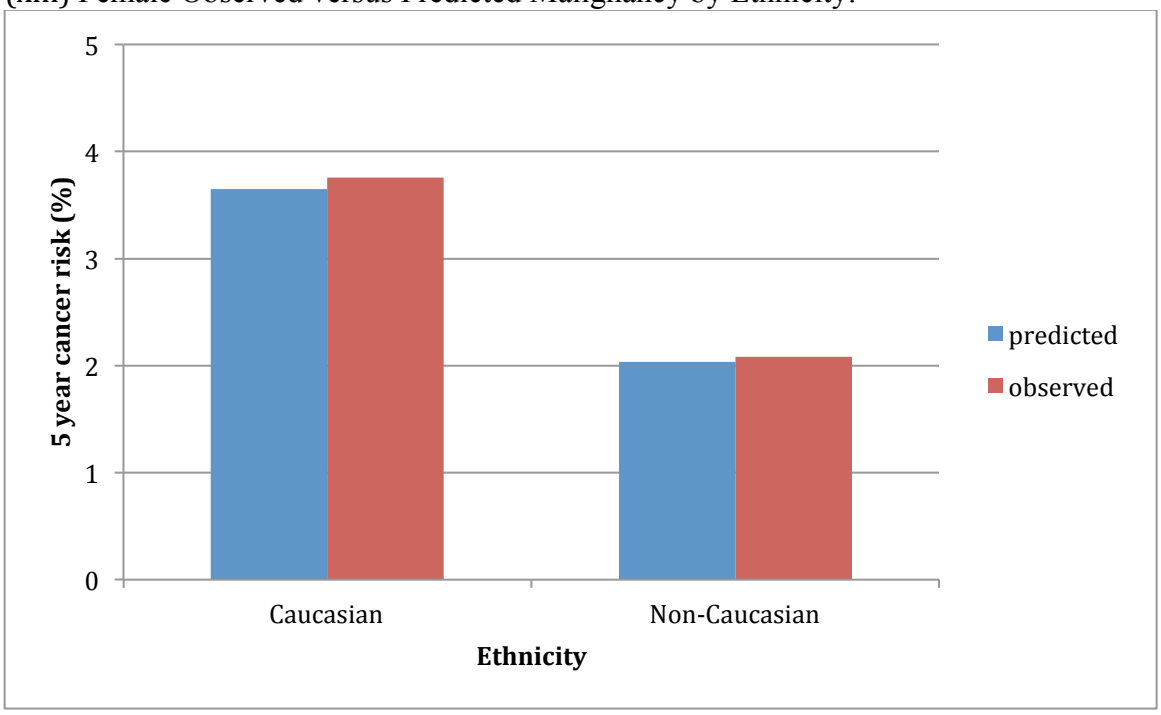
(xi) Female Observed versus Predicted Malignancy by IBD Status:



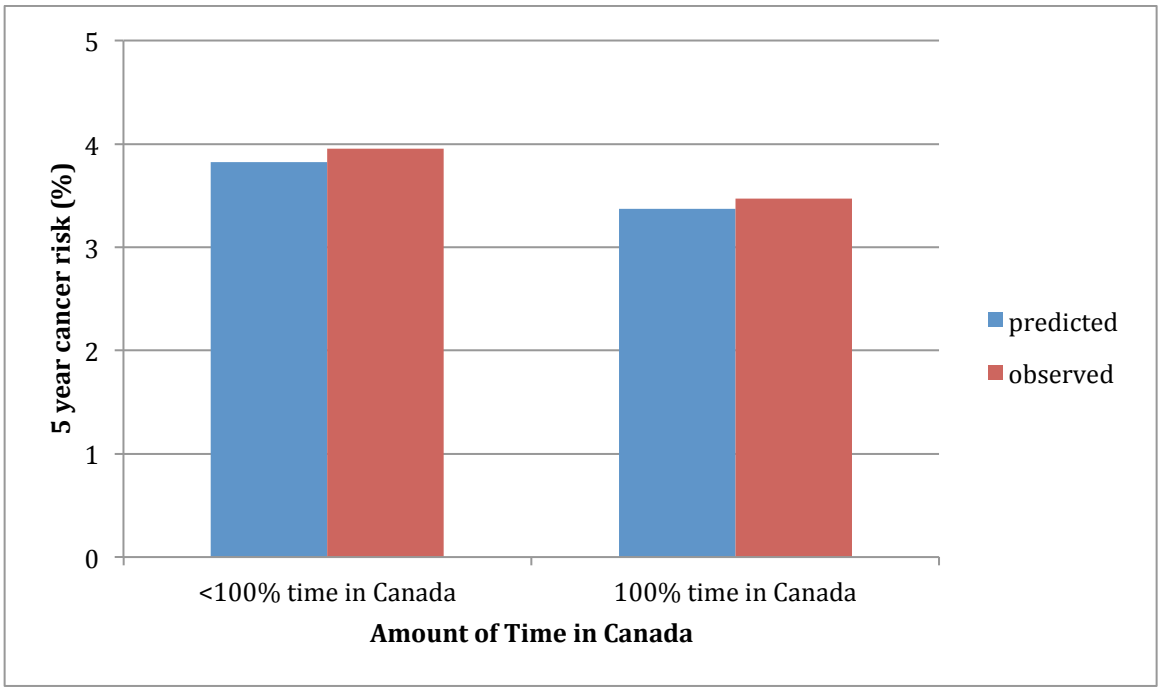
(xii) Female Observed versus Predicted Malignancy by Immigration Status:



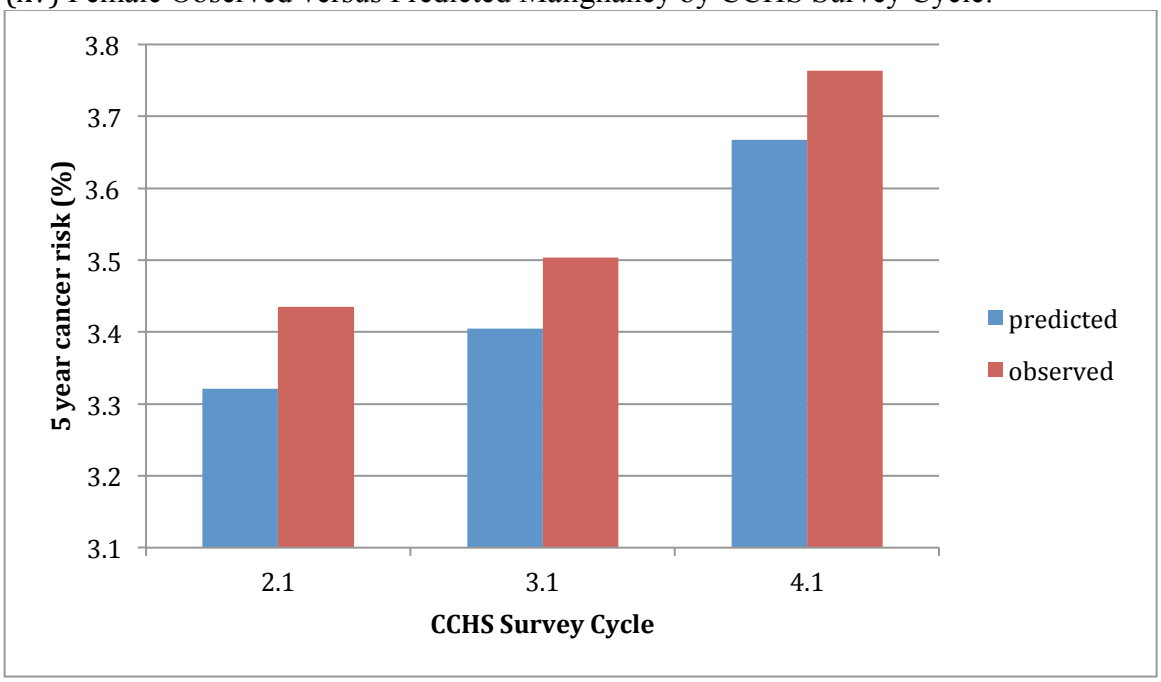
(xiii) Female Observed versus Predicted Malignancy by Ethnicity:



(xiv) Female Observed versus Predicted Malignancy by Percent of Life Lived in Canada:

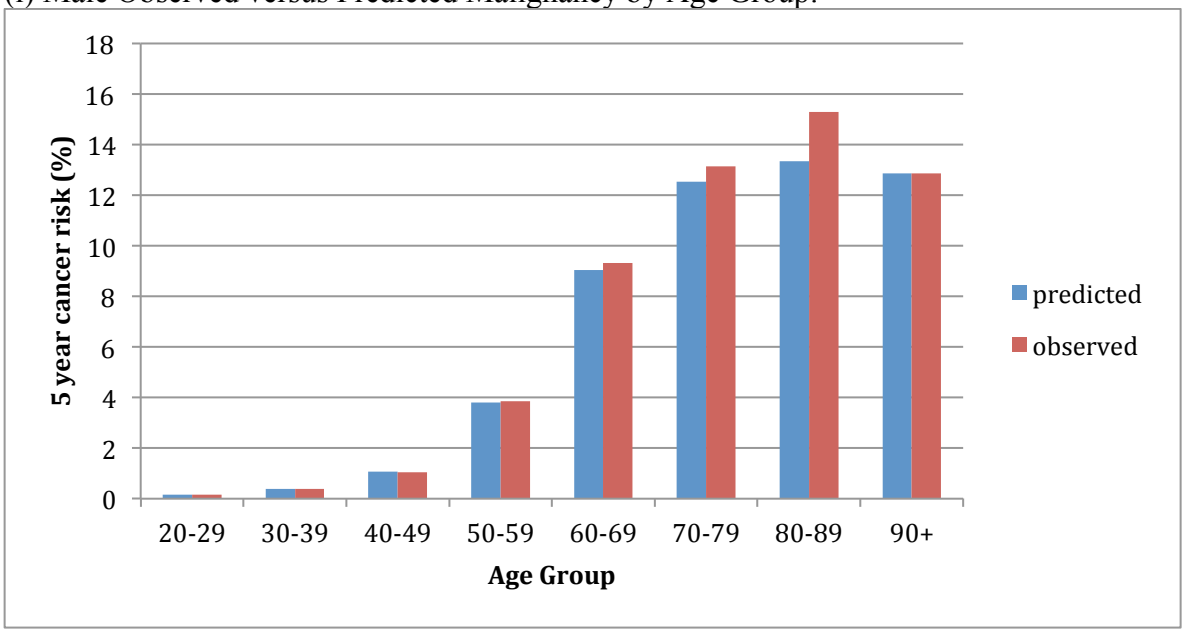


(xv) Female Observed versus Predicted Malignancy by CCHS Survey Cycle:

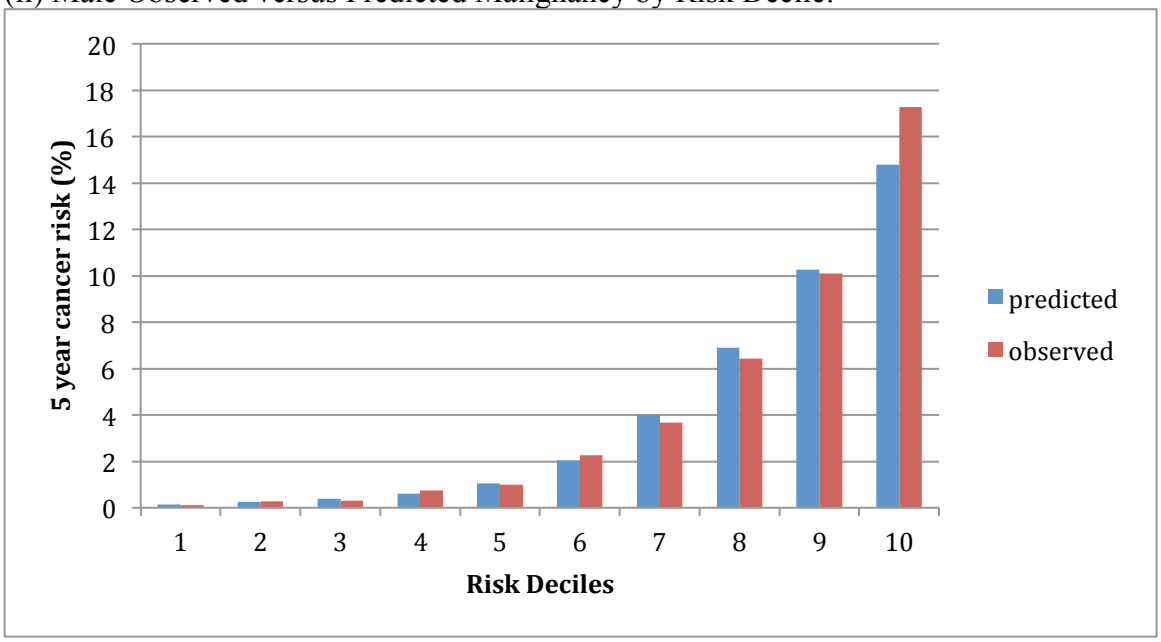


(B) Male Observed versus Predicted Risk of Incident Cancer

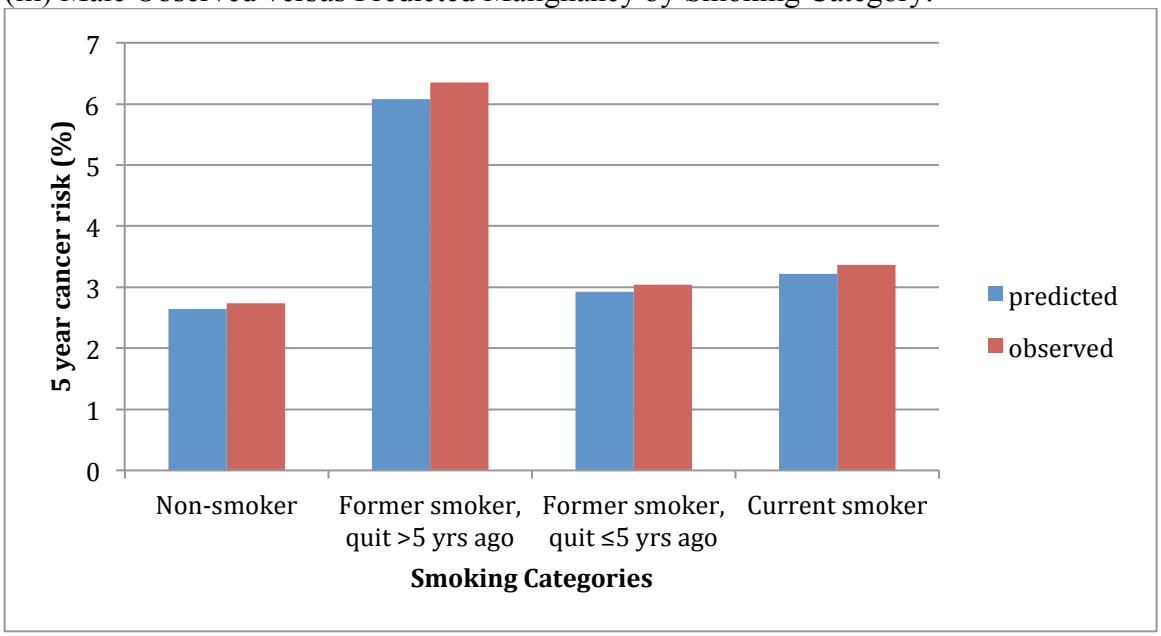
(i) Male Observed versus Predicted Malignancy by Age Group:



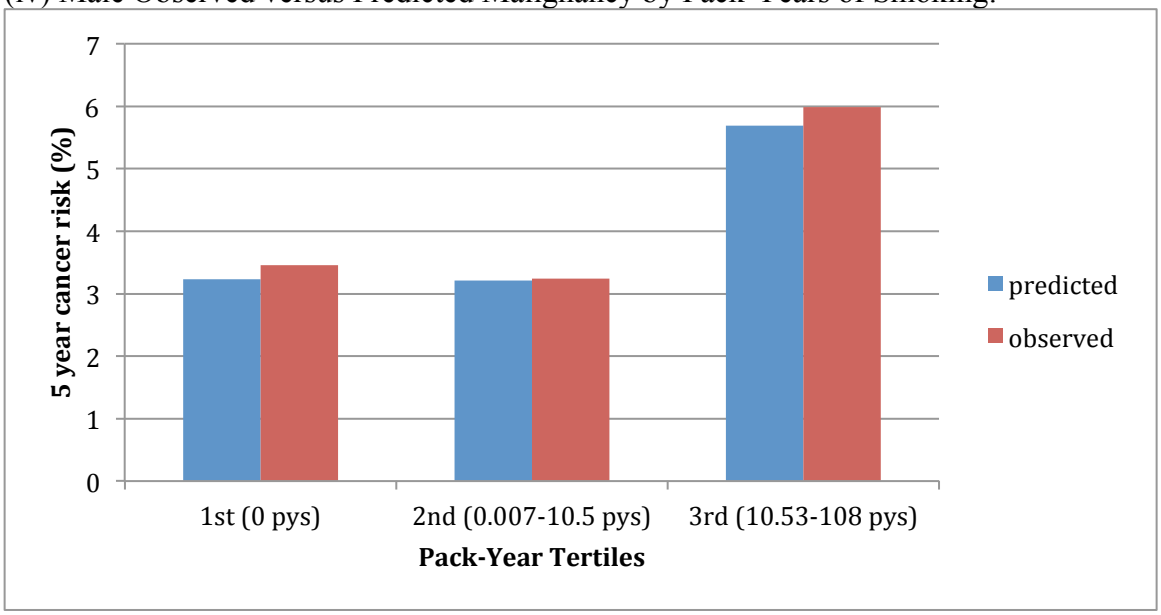
(ii) Male Observed versus Predicted Malignancy by Risk Decile:



(iii) Male Observed versus Predicted Malignancy by Smoking Category:

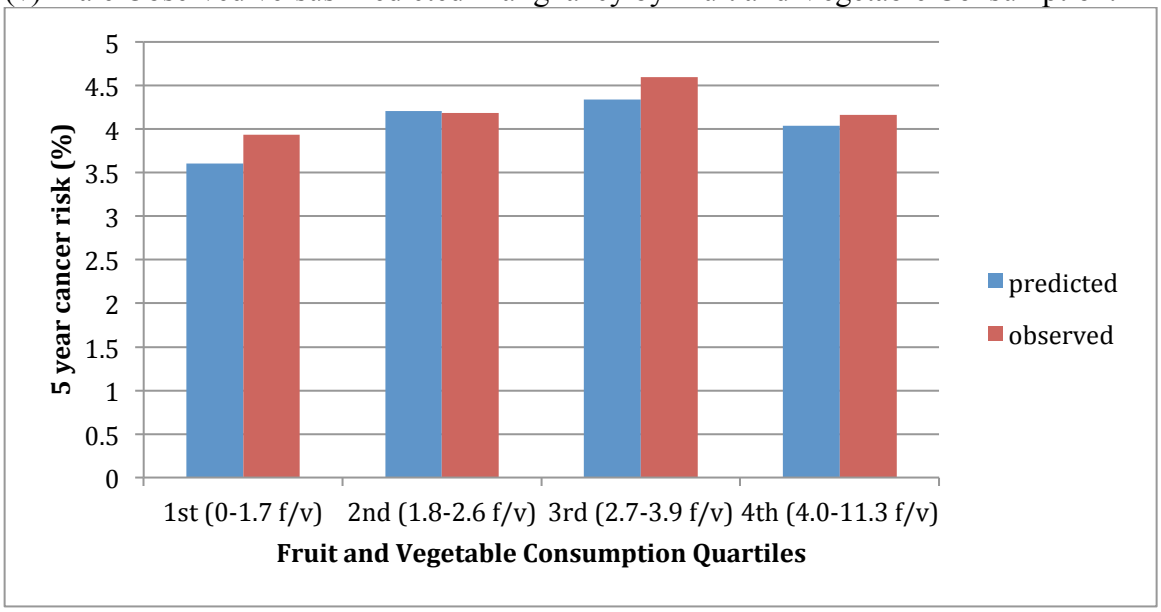


(iv) Male Observed versus Predicted Malignancy by Pack-Years of Smoking:



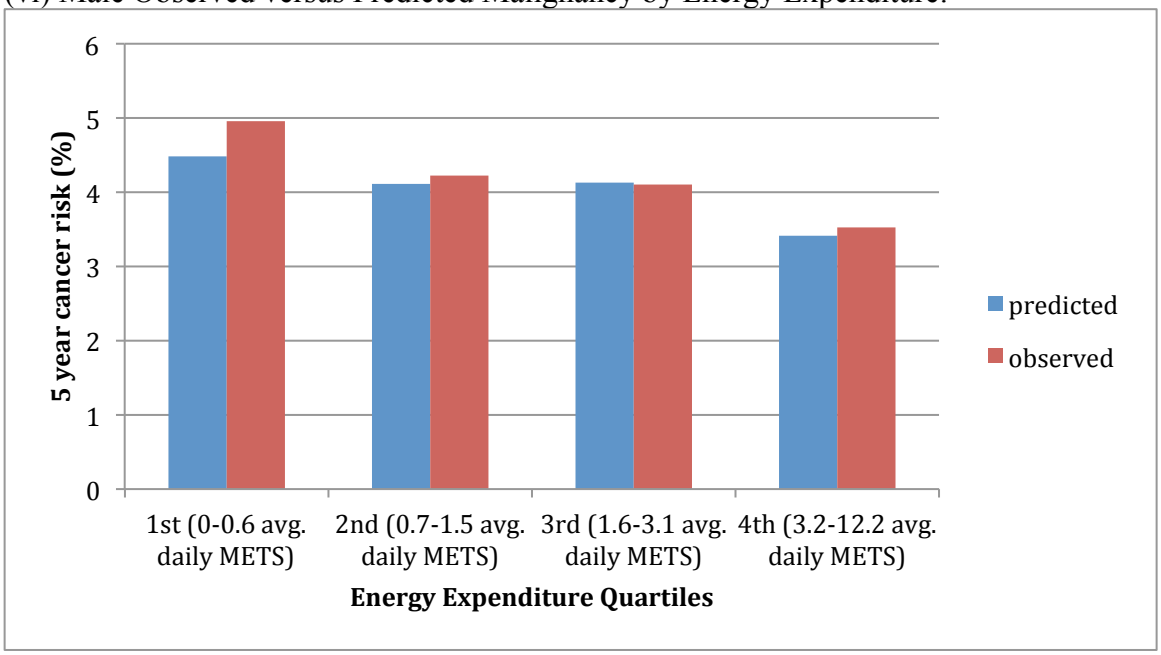
Legend: pys= pack-years of smoking

(v) Male Observed versus Predicted Malignancy by Fruit and Vegetable Consumption:

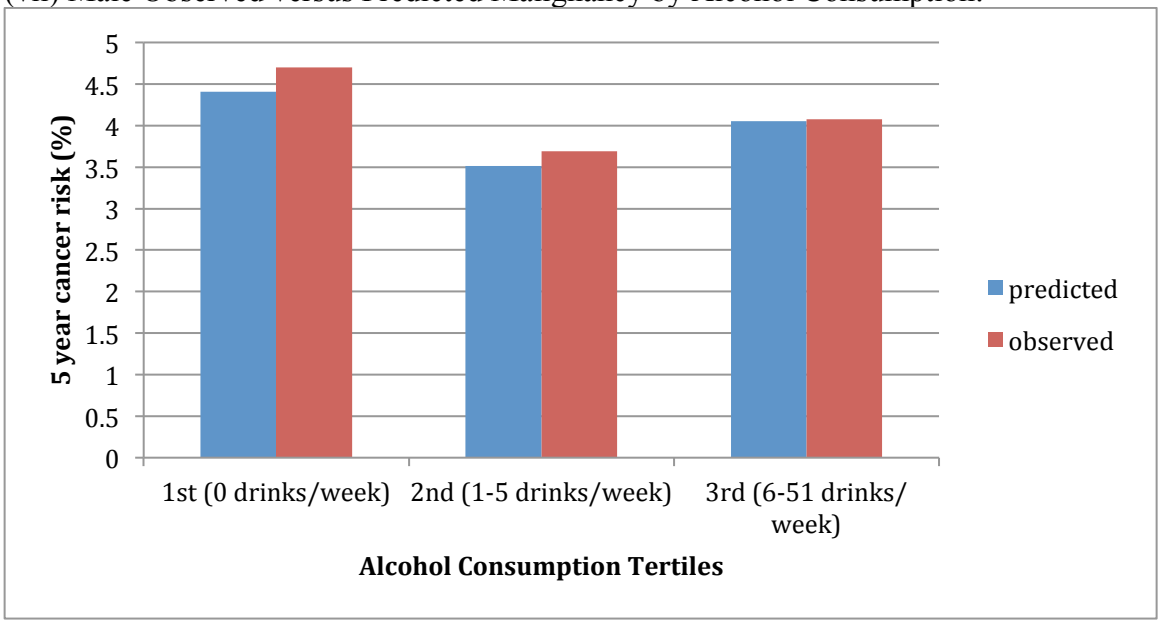


Legend: f/v = fruit and vegetable consumption

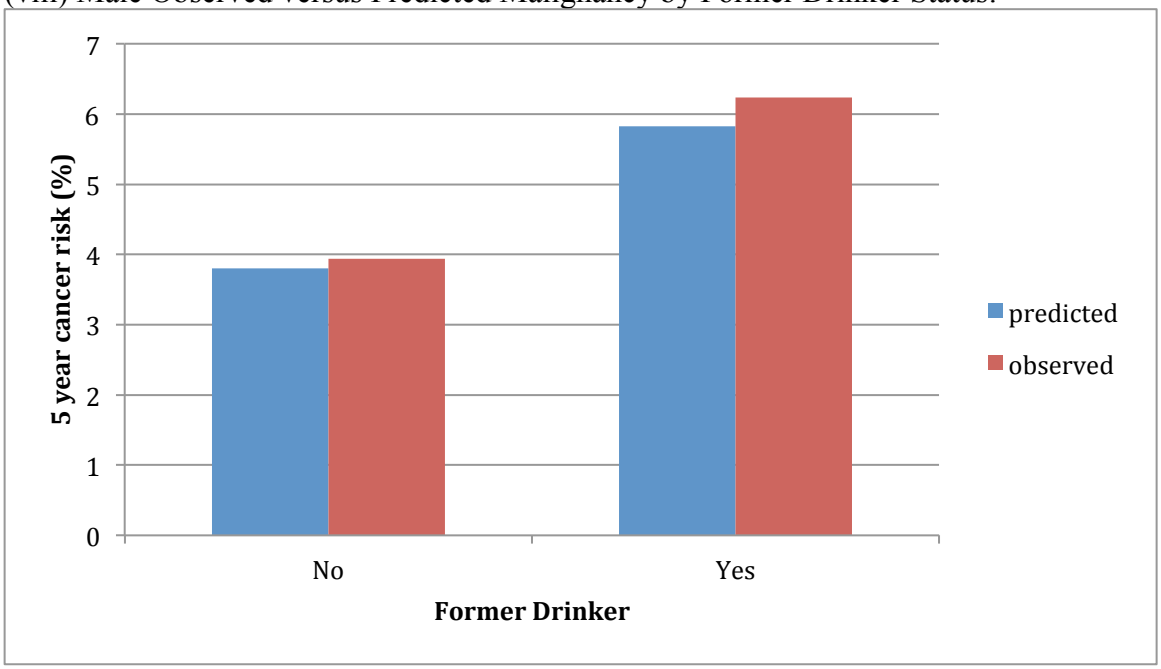
(vi) Male Observed versus Predicted Malignancy by Energy Expenditure:



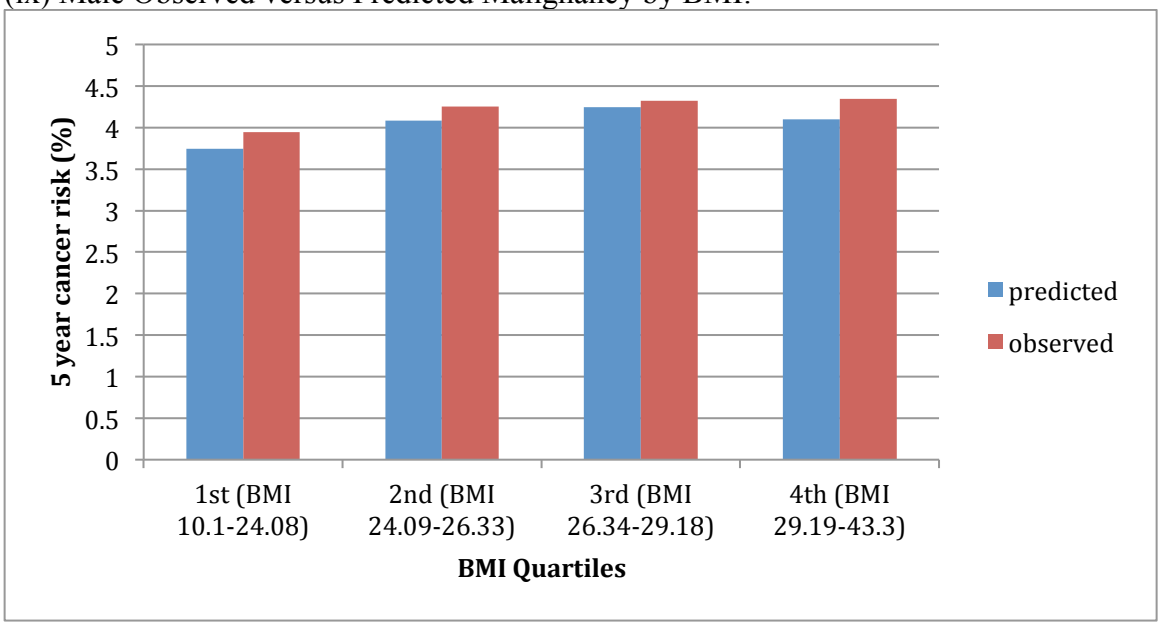
(vii) Male Observed versus Predicted Malignancy by Alcohol Consumption:



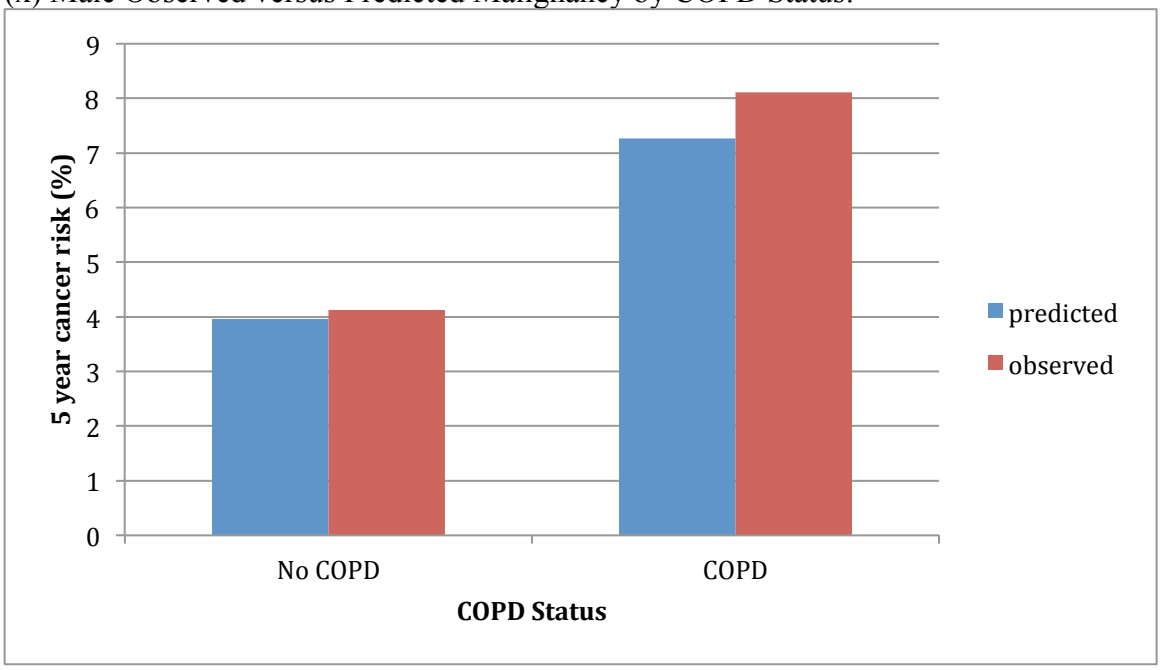
(viii) Male Observed versus Predicted Malignancy by Former Drinker Status:



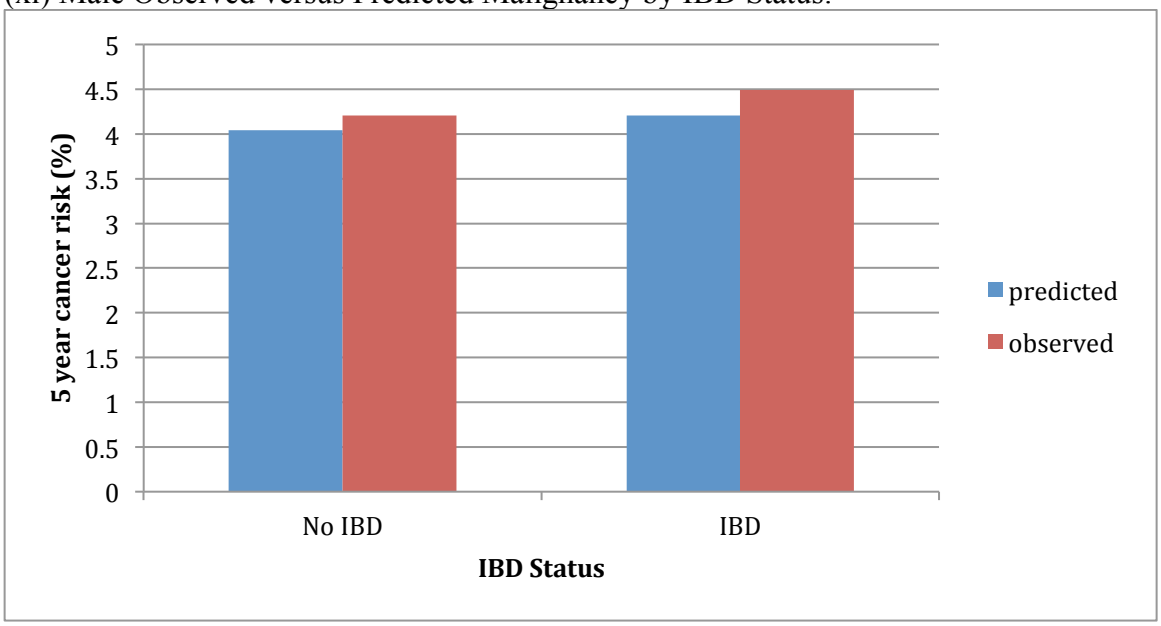
(ix) Male Observed versus Predicted Malignancy by BMI:



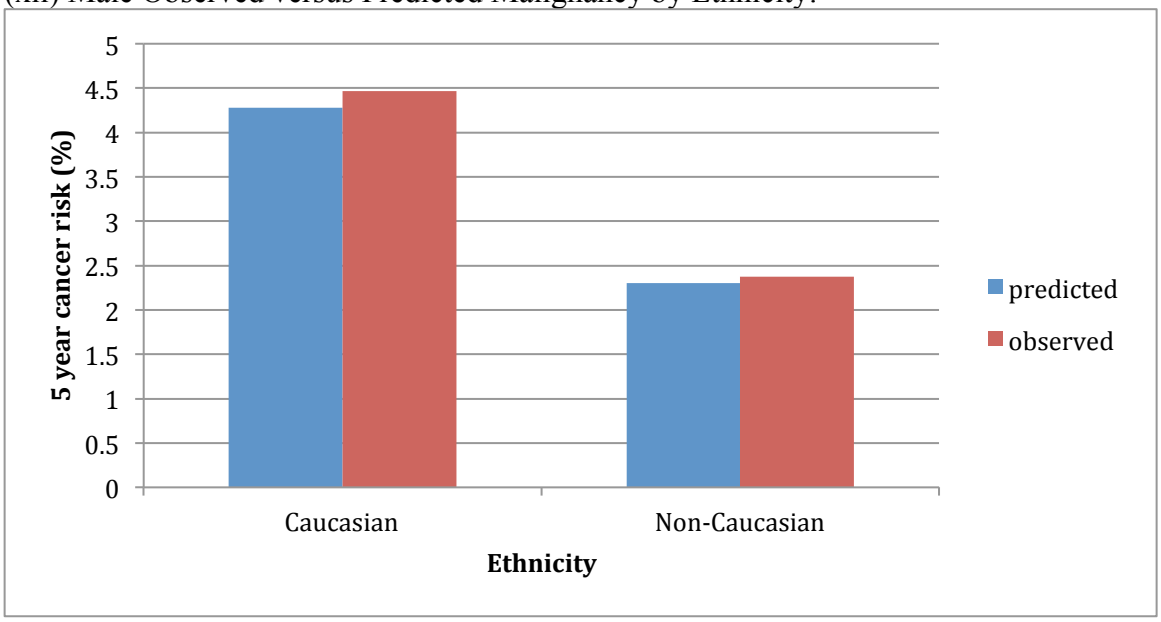
(x) Male Observed versus Predicted Malignancy by COPD Status:



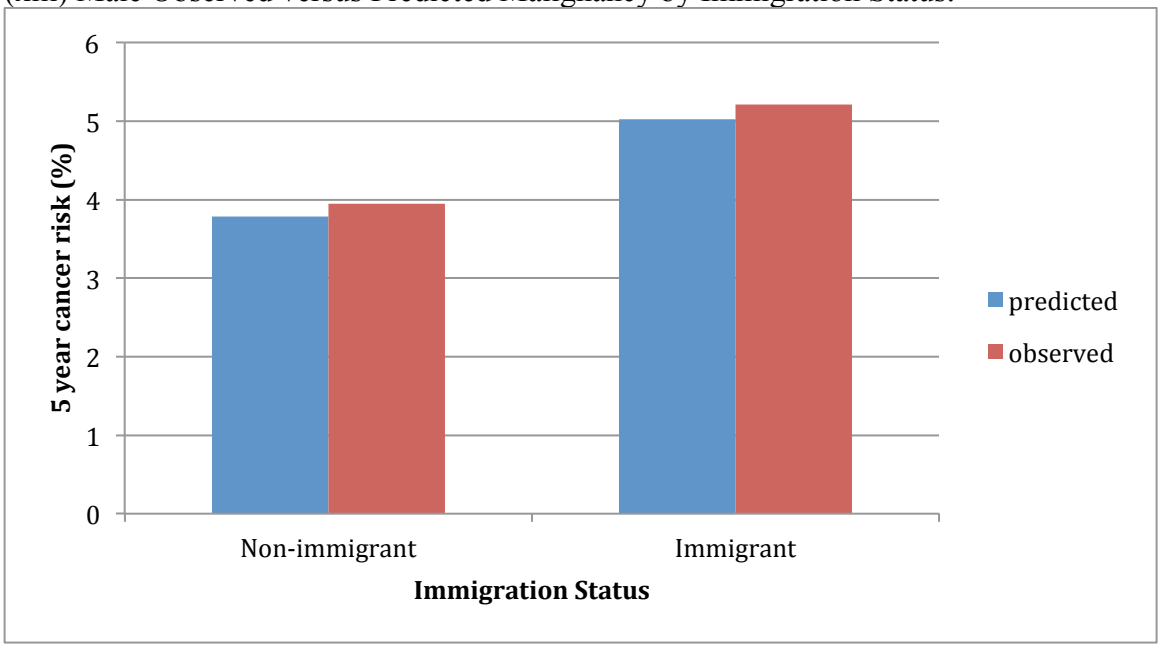
(xi) Male Observed versus Predicted Malignancy by IBD Status:



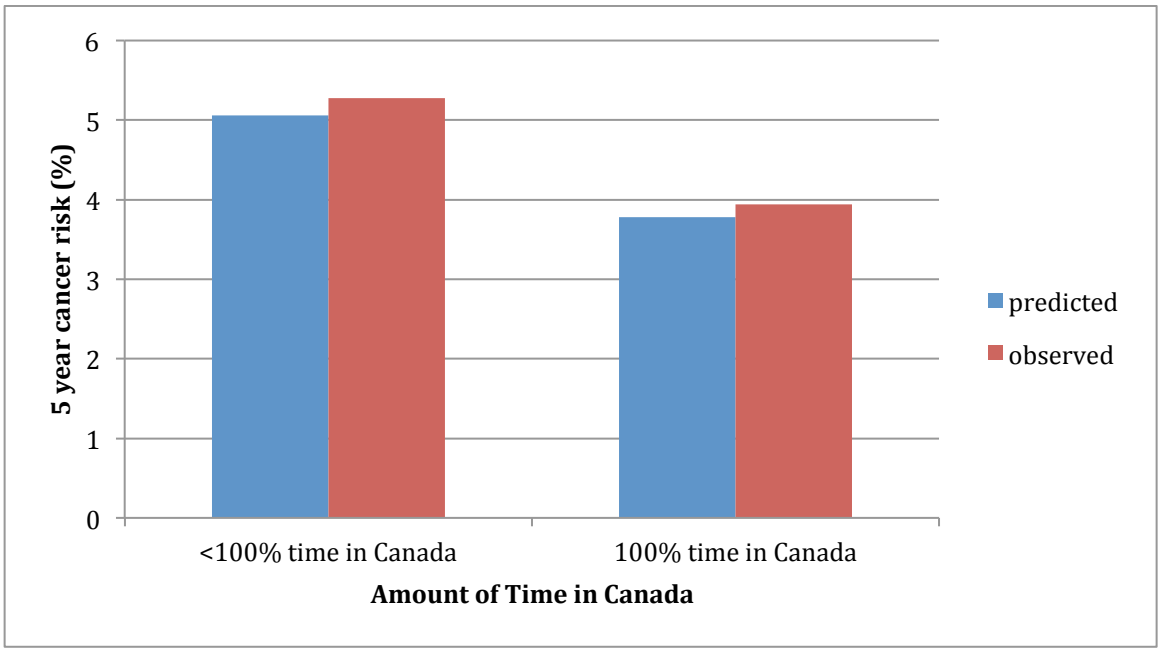
(xii) Male Observed versus Predicted Malignancy by Ethnicity:



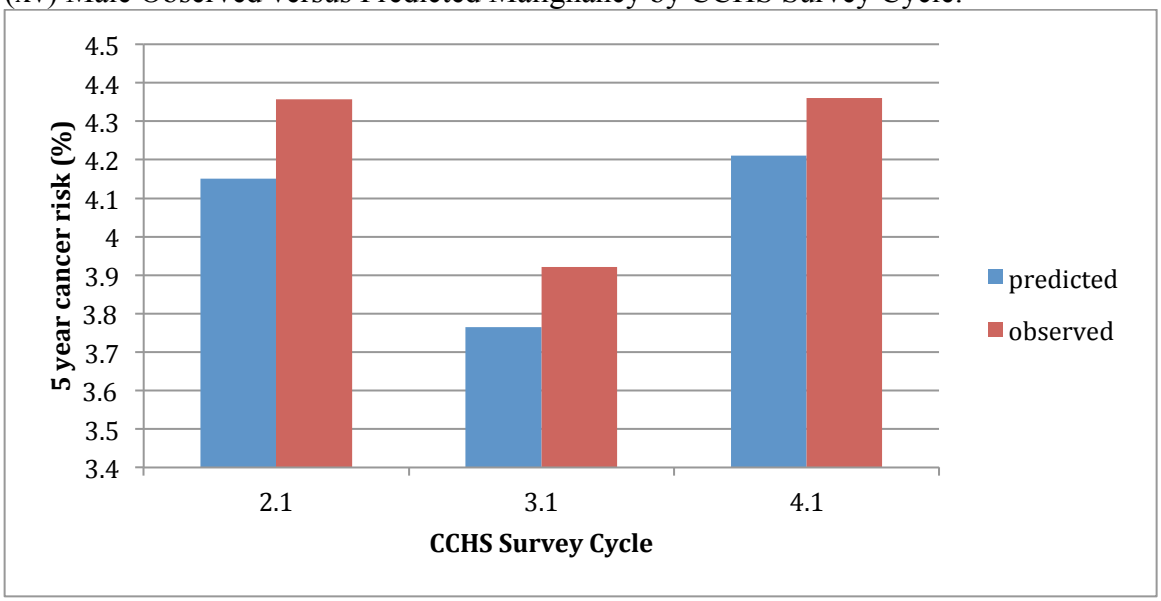
(xiii) Male Observed versus Predicted Malignancy by Immigration Status:



(xiv) Male Observed versus Predicted Malignancy by Percentage of Life Lived in Canada:



(xv) Male Observed versus Predicted Malignancy by CCHS Survey Cycle:



For males, across 10-year age increments, the percentage difference between observed and predicted cancer risk ranged from 0.08% to 14.55%. For males across predicted risk deciles, the percentage difference between observed and predicted values ranged from 1.64% to 21.99%. The percentage difference exceeding 20% occurred in the third risk decile. When considering only those risk deciles with >5% of the incident cancer cases among males, the greatest difference in predicted versus observed risk occurred in the 10^{th} decile at 16.85% difference. In all additional subgroup analyses as specified above, the difference between observed and predicted risk of incident cancer among males was always <20%.

Discrimination, or the ability to determine who is at high versus low-risk of an outcome[4], was evaluated using a C-statistic. The C-statistic for females was 0.76 showing adequate discrimination and for males was 0.83 indicating excellent discrimination.

4.4.2 Model Performance in the Validation Cohort:

The Nagelkerke R^2 value for females in the validation cohort was 0.058 and for males was 0.117. The Brier Score for females was 0.033 and for males was 0.038.

Among females, the observed number of incident cancer cases was 497 and the predicted risk was 477.19 cases, indicating a difference between observed and predicted risk 4.15%. The observed number of incident cancer cases among males was 498 and the predicted risk was 453.63 cases, indicating a difference between observed and predicted

risk of 9.78%. For both females and males, the risk of incident cancer was under-predicted by the algorithm.

For females across 10-year age increments, the difference between observed and predicted incident cancer risk ranged from 2.74% to 14.71%. The highest risk differences were observed in the highest age groups (80-89 years old and ≥90 years old). Across predicted risk deciles among females, the percentage difference between observed and predicted values ranged from <0.001% to 49.63%. The highest difference was found in the third risk decile, while the 2nd risk decile was also high at 25.60% difference.

Among males, the observed versus predicted risk across 10-year age increments ranged from 1.35% to 54.55%. The highest difference was observed in the >90 year old group. Across risk deciles, there was a range of 2.99% to 82.37% difference between observed and predicted values for males, with the highest risk differences in the 1st, 2nd and 4th risk deciles.

Discrimination in the validation cohort was indicated by a C-statistic of 0.75 among females indicating adequate discrimination and 0.84 among males, indicating excellent discrimination.

Table 6: Performance Measures for the Derivation and Validation Cohort Risk Algorithms by Sex

	Derivation Cohort		**Validation Cohort**	
	Female (N=43 696)	**Male (N=36 630)**	**Female (N=14 426)**	**Male (N=11 970)**
Overall Performance				
Nagelkerke R^2	0.056	0.0965	0.058	0.117
Brier Score	0.034	0.038	0.033	0.038
Calibration				
Overall Calibration	2.97% (1514.02: 1559)	4.23% (1481.41: 1544)	4.15% (477.19: 497)	9.78% (453.63: 498)
Discrimination				
C-statistic	0.76 (95%CI: 0.75-0.77)	0.83 (95% CI: 0.82-0.83)	0.75 (95% CI: 0.73-0.77)	0.84 (95% CI: 0.82-0.85)

*shows the (predicted risk of cancer: observed incidence of cancer)

CHAPTER 5: DISCUSSION

5.1 Interpretation:

Through the use of self-reported behavioural factors from CCHS data it was possible to develop a risk algorithm that accurately and discriminately predicted incident cancer in both males and females.

The overall calibration of the model was excellent, with differences between observed and predicted cancer incidence being less than 5% in both male and female populations. Furthermore, calibration was maintained across several subsets of interest including different age categories and different levels of risk, consistently displaying a difference of <20% between observed and predicted estimates of cancer incidence except in instances in which a category had a very low incidence of cancer. The algorithm discrimination was also very good for females and excellent for males. Therefore, using routinely available self-reported data, it was possible to determine individuals at high risk and at low risk of incident cancer, with predicted risk estimates approximating the observed risk in the population.

The risk algorithm was validated in an earlier cycle of the CCHS (1.1) and results were similar to those obtained in the derivation cohort. There was a <10% difference between observed risk and predicted risk of incident cancer in both males and females, indicating a high degree of accuracy. Across some age categories and risk categories, the difference between observed and predicted risk values was elevated beyond a threshold of 20% defined for adequate calibration. However, on further examination, this occurred almost

exclusively in categories with a very limited number of incident cancer cases, resulting in a higher percentage difference between observed and predicted risk despite only small variations in predicted risk estimates. Discrimination between those at high and low risk of incident cancer was similar to that found in the derivation cohort, being very good among females and excellent for males.

Similar to this study, several other studies have assessed the utility of lifestyle or behavioural risk factors in prediction of cancer.[8–13] These studies uniformly found that poorer health behaviours were associated with a higher cancer incidence. The main study design feature that limits the utility of interpreting these studies is the use of categorical predictor variables that dichotomize underlying values, thereby leading to a loss of information on the continuous relationship between predictors and outcomes.[8–12] In addition, most of these studies classified individuals as adherent or non-adherent to health behaviours and then created a numeric scale to assess the impact of having a greater number of healthful behaviours on incident cancer.[8–12]

In contrast, our study assessed behavioural factors in a continuous fashion whenever possible to model the relationship between predictors and incident cancer, allowing for more accurate risk assessment, particularly when considering those at moderate risk.[22]

The main predictor variables identified through the ANOVA plots as important predictors of incident cancer were age, smoking status and pack-years of smoking among females and males in our study. The results from the Cox proportional hazard model also

indicated that age and smoking category, but not pack-years of smoking, were important predictors for females. For males, the Cox model indicated that age, but not smoking status or pack-years of smoking, was a predictor of incident cancer.

The finding that smoking status was an important behavioural factor in the prediction of incident cancer among females is supported by the literature. There is an association between smoking and several types of cancer, including lung cancer, oropharyngeal cancer, esophageal cancer, stomach cancer, liver cancer, pancreatic cancer, cervical cancer, bladder cancer and leukemia. In the majority of these cancers, smoking has been identified as the most important behavioural risk factor.[5,6] The degree to which smoking impacts cancer development may be influenced by prior duration of smoking, intensity of smoking, age at initiation and age at cessation.[6]

Further examination of the lack of statistically significant association between pack-years of smoking status and incident cancer in both males and females was undertaken by exploring the possibility of multicollinearity. A variable clustering algorithm to identify correlated variables was implemented and the results showed that pack-years of smoking and smoking category were collinear for both males and females, with one variable explaining more than 70% of the variation in the other. Therefore, the regression coefficients for pack-years of smoking and smoking status cannot be interpreted as separate predictors in the presence of the other, as these variables covary. We decided to retain these variables in the prediction tool, as the overall goal of our model is prediction

of cancer from a range of variables, rather than estimating the magnitude of any individual associations with incident cancer.

Among females, being an immigrant was a predictor of all-cause cancer in the multivariate analysis. The unadjusted relationship for females between immigrant status and incident cancer also revealed a positive association (HR=1.14, 95% CI: 1.01-1.28). This was in contrast to what was expected based on the literature, specifically that immigrants would have a lower incidence of cancer. In one review of 37 studies on immigrants to Europe, findings suggested that there was an overall lower incidence of cancer among immigrants from a variety of countries including those in the Middle East, Asia and South/Central America. However, migrants from non-western countries were more likely to develop cancers related to infectious diseases than those in the European population, specifically cancers of the oral cavity, nasopharynx, stomach, liver, gallbladder, cervix, prostate and lymphomas.[44] In a research initiative linking Canadian immigration databases to health administrative databases, preliminary results also indicated that immigrants have lower incidences of all-cause cancer.[45] Our results indicating immigration among females was a predictor of all-cause cancer may reflect a higher number of cancers that are more common to migrant populations, however further analysis would be required to elucidate the specific cancer types that compose all-cause cancer in our study. Both males and females showed that a longer duration of time lived in Canada was associated with a higher incidence of cancer, which would be consistent with studies indicating that risk of cancer in migrant populations is inbetween that of their home and host countries.[44]

COPD was independently associated with a greater hazard of incident cancer among females, but not among males, even after adjustment for smoking status. Previous literature indicates that COPD is a risk factor for certain cancers, particularly lung cancer, after adjustment is made for smoking status.[46,47] The timing of COPD diagnosis may be important, with one study indicating that a more recent diagnosis of COPD (in the 6 months prior to cancer diagnosis) was associated with a three-fold higher odds ratio of lung cancer than a diagnosis of COPD made in the 10 years prior to cancer diagnosis.[46] The variations in types of cancer composing all-cause cancer between males and females again may account for the differences observed in the associations of COPD with cancer incidence between sexes. Additionally, variations in timing of COPD diagnosis may be implicated in differences in the association of COPD with cancer incidence among females and males, however this is beyond the capability of analysis with the available survey data.

Among males, an increased number of alcoholic beverages consumed per week were associated with incident cancer. This is consistent with literature indicating that an increased hazard of incident cancer occurs with increasing number of alcoholic beverages consumed.[48] In addition, there was a borderline association between increased fruit and vegetable consumption daily with a decreased hazard of incident cancer among males.

Hazard of incident cancer was not impacted by leisure physical activity level, BMI level, being a former alcohol drinker or being exposed to second-hand smoke when accounting

for other variables, among both males and females. In addition, hazard of incident cancer was not different for those of different ethnicities, for those with different levels of education or for those at different levels of deprivation. Therefore, it may have been possible to estimate cancer risk using a more simplified algorithm, however despite the variables being relatively non-contributory in the risk algorithm they were maintained according to the pre-specified protocol for this study.

5.2 Study Strengths:

A primary strength of this study is the ability to assess the combined impact of multiple, modifiable behavioural risk factors in the Ontario population to predict risk of incident cancer. This study utilized a large cohort representative of the Ontario population to generate risk estimates. Prediction of risk was conducted without the use of clinical variables that require resource-intensive data collection, but rather used routinely available survey data to both accurately and discriminately predict cancer incidence in the population. This facilities the utilization of the risk algorithm in other settings where similar routinely available population health data is available, particularly given that the algorithm has been validated in an external cohort thereby indicating its utility outside of the setting in which it was developed. The focus on modifiable health behaviours is important from a population approach, in which individuals and public health professionals are interested in the potential to decrease incident cancer risk with changes to health behaviours. These health behaviours are “upstream” to many clinical variables, such as COPD or diabetes mellitus, and are in contrast to non-modifiable risk factors,

such as age at menarche or menopause, thereby allowing them to be targeted through interventions to improve population health.

A working group on cancer prediction models sponsored by the National Cancer Institute in the United States, and encompassing a panel of experts in the field, made several recommendations on the development and application of future models.[42] They suggested that the ability to estimate the population burden of disease is a key application of risk algorithms. They identified that most algorithms have been developed among Caucasian populations, and suggested that risk algorithms involving other ethnicities be a priority. They also identified that cancer risk prediction tools may be of low value in the clinical realm unless the relative risk of an outcome based on risk factors is highly elevated (i.e. 20 or more). This is due to the fact that the positive predictive value for most cancers, even if they are common in the population, is still low given that the incidence of cancer over a lifetime is low.[42] Most individuals who develop cancer have risk profiles that are typically close to the average risk, therefore it is suggested that population strategies rather than strategies targeted at high risk individuals may have the greatest effect on reducing cancer incidence. Our study targeted all of the above recommendations, with a large cohort reflective of the diversity within the population used to develop the population-based risk algorithm which can be used to estimate risk of developing cancer.

Our study is also unique in the methodology employed to generate the predictive risk algorithm. The use of pre-specified predictor variables and interaction terms that are included in the model regardless of strength of association with the outcome results in a

less data-driven approach to defining the predictive risk algorithm. This methodology, alongside the methods used to simplify the predictive model, may result in an improved predictive ability in different populations to that in which the model was derived. In addition, restricted cubic splines were used to allow for complex modeling of non-linear relationships between exposures and incident cancer. Measures of predictive accuracy and discrimination were described for the risk prediction algorithm, and the algorithm was validated in an external cohort.

5.3 Study Limitations:

5.3.1 Limitations Associated with a Composite Primary Outcome:

The primary outcome of this study, all-cause cancer incidence, is a composite of multiple different types of cancers arising from various sites and with varying pathophysiology. Given that it is a heterogeneous outcome measure, not all predictor variables may be associated with each type of cancer that is incorporated into the outcome of all-cause cancer. Furthermore, some predictor variables may be associated positively with certain cancers and negatively with other cancers incorporated in the same outcome variable, thereby biasing the association between predictor variables and outcome toward a lack of association.[23] Despite this limitation, there are several reasons that all-cause cancer is an important outcome to consider. First, despite the heterogeneity in type and location of each cancer, there are mechanistic similarities underlying all cancer development[49] and the conditions that predispose to these cellular changes may be similar. Second, if each clinical condition were taken individually for policy decision-making, there would be difficulty in establishing healthy guidelines for the population. As an example, moderate

alcohol consumption is associated with decreased cardiovascular risk but with increased risk of many cancers.[50,51] By assessing a composite outcome, such as all-cause cancer incidence, it is possible to have an overall determination of the types of policies that will encourage health behaviours most associated with decreasing a broader range of adverse outcomes. This is reiterated by the working group on cancer prediction, which suggests that risk models should be developed with multiple cancer outcomes, to increase benefit-risk indices for different interventions and to facilitate decision-making.[42]

5.3.2 Duration of Follow-up:

Outcomes were measured from date of CCHS study enrollment (earliest year is 2001) for a period of 5 years. Despite this length of follow-up, 5 years may be an insufficient period of analysis given the length of time over which malignancies may develop. This would result in an underestimation of the association between exposure and outcome variables. However, there is evidence that current health behaviours tend to be indicative of prior health behaviours. A study assessing stability of lifestyle behaviours over a 4 year period based on self-reported data showed that cigarette smoking and alcohol consumption were relatively stable over time, but that there was slight variability in diet and physical activity.[52] Despite some variability found across behaviours, in the overall study population 48% of individuals made no change in behaviour, 41% made one change in behaviour (in one domain, such as smoking, alcohol consumption, diet or physical activity) and only 11% made more than one change in behaviour. Of those who changed behaviour, it was approximately evenly divided between improved behaviour or declined behaviour. Therefore, although this study utilized specific data from each cycle

of the CCHS to evaluate the role of lifestyle behaviours in the development of cancer, some of the association between variables and outcomes may be due to the persistence of lifestyle behaviours over time.

5.3.3 Limitations of the Data for Outcome Variables:

There are limitations of using administrative data that depend on ICD-9 codes from the Ontario Cancer Registry. Sources of error in ICD-9 codes include the quality of information recorded by clinicians or administrators, variations in precision of terms used to identify conditions, transcribing and conveyance of test or procedure results and variations introduced in the "paper trail" from written and electronic records at the physician level to the administrative staff.[38,53]

At this time, it is not feasible to link the CCHS data with the Canadian Cancer Registry (CCR) given time and organizational constraints. Therefore, there is a possibility that some outcomes may be missing due to individuals moving to other provinces and being lost to follow-up. These outcomes would be considered missing completely at random, as they would not be related to any intervention or to a specific outcome as similar services are available across Canada for cancer diagnosis and management.

5.3.4 Limitations of the Data for Predictor Variables:

The predictor variables in this study were derived from self-reported data in the Canadian Community Health Survey. Self-reported data may be flawed, namely due to

misclassification error. All self-reported data were gathered at a single time point at which predictor variables were measured without knowledge of past behaviour.

With survey data it is not possible to obtain additional information on risk factors which may be relevant in the prediction of cancer incidence. Unfortunately data in the CCHS are limited on some potentially important risk factors, specifically personal reproductive history and family history of cancer. For example, maternal breastfeeding is recorded in the CCHS but it was not included in the analysis, as it was only reported for those women who breastfed an infant in the past 5 years.

Many screening examinations are used to identify disease at an early stage, such as mammography for detection of breast cancer. Other tests, such as fecal occult blood test or colonoscopy/sigmoidoscopy screening for colorectal cancer and pap smear screening for cervical cancer, may detect premalignant lesions. The screening tests that detect cancer at an early stage may inflate estimates of cancer incidence,[54] while those that detect and treat premalignant lesions may decrease estimates of cancer incidence. In this study, we chose not to include screening examinations in the algorithm, but rather to focus on modifiable health behaviours that impact on cancer development. In particular, it would have been challenging to elucidate which tests were intended as screening examinations and which were diagnostic, therefore impacting on the interpretability of the results.

5.4 Implications:

The purpose of this study is to develop a risk algorithm for both males and females that predicts incident cancer with a high degree of calibration and discrimination. This prediction is based upon knowledge of important behavioural risk factors for development of malignancy, as well as the availability of survey data on these risk factors in a large database. The intent to predict incident cancer superseded the possibility of creating an explanatory model, in which the individual risk factors would have been included in the model only if they were causally related to the outcome of malignant cancer at any site. Part of the rationale for this approach is that the study was observational in nature and not intended to delineate causal relationships. In addition, this risk algorithm is intended to be used at the population level by policy-makers and public health professionals for the prediction of malignancy using routinely available data within that population. Such data may include the CCHS data for Canadian policy-makers or similar types of administrative data in other jurisdictions. As such, despite the causal underpinnings for the selection of many variables within our model, certain variables (such as smoking status and pack-years of smoking) were retained in the model for their predictive value despite collinearity between these categories. This makes interpretation of any hazard ratios for incident malignancy, based on smoking status and pack-years of smoking, impossible and therefore the model is not as useful for strictly explanatory purposes. More complex modeling strategies were also used with the intent of improved prediction, including pre-specification of predictor variables regardless of their association with the outcome and the use of restricted cubic splines. Although this results in hazard ratios which are more difficult to interpret from a causal perspective, it

allows for improved prediction when relationships between predictor variables and the outcome variable are non-linear.

Our risk prediction model is ideally designed for providing information to policy-makers on the risk of malignancy in a given population, among males and females, based on health behaviours within that population. It is also useful for predicting future cases of malignancy based on the array of predictive risk factors included in the model. It improves upon existing models that derive relative risks from other populations to infer risk within the population of interest. It also improves modeling capability by including multiple risk factors to predict the outcome of interest, rather than simpler models based on age and sex alone.[4] Finally, it allows for an understanding of how risk is dispersed within a population, whether it is diffused among many members of population or concentrated among only a few.[4]

The issue of risk dispersion within a population is important to policy-makers and public health professionals when determining the targeting and allocation of limited public health resources. If risk is diffused within the population, then a strategy used to "shift the curve" of risk factors among the population as a whole is an effective strategy to reduce incident malignancy. However, in populations in which the risk is more concentrated, targeting of resources to these populations at higher risk may result in improved outcomes.[4,55] Our risk prediction model helps to understand the dispersion of risk within a population and therefore permits improved allocation of finite resources toward risk factor prevention and improved screening uptake.[55] This is an important

consideration given time and resource constraints for distributing preventive care services.

Once resources are allocated for particular public health interventions, this multivariable risk algorithm may be used to evaluate the impact of the public health interventions over time.[4] For example, the risk algorithm can be used to determine if a change in incident cancers within the population has occurred following alterations in lifestyle behaviours among members of the population.

Although the risk of malignancy associated with individual behavioural factors is difficult to interpret due to the complex modeling strategy used in this risk algorithm, it can still be used to understand the relationship between a mosaic of risk factors and incident cancer. This understanding is useful to provide evidence of the importance of behaviour change for the reduction of adverse health outcomes, such as malignancy. This evidence can be presented to funding bodies for procurement of resources to target reductions in risk factors for malignancy, as well as in messaging to the general public on the importance of health behaviours in the development of disease.

5.5 Summary:

This study has developed a risk algorithm that is able to predict incident cancer at the population level based on modifiable health behaviours. It has been developed and validated on external data, with the potential for implementation in public health strategies for the estimation of cancer risk in the Ontario population and more broadly to

other populations that obtain routinely collected survey data on health behaviours. The utility of this algorithm is in the accurate estimation of cancer risk, which improves upon current methods of estimation and may identify populations at high risk of cancer, thereby permitting public health interventions aimed at cancer prevention and early identification to be focused for their greatest impact.

Bibliography:

1. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. *Cancer*. 2008;113(11):3075-3099. doi:10.1002/cncr.23908.

2. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat*. 2012;133(1):1-10. doi:10.1007/s10549-011-1853-z.

3. Kirkegaard H, Johnsen NF, Christensen J, Frederiksen K, Overvad K, Tjønneland A. Association of adherence to lifestyle recommendations and risk of colorectal cancer: a prospective Danish cohort study. *BMJ*. 2010;341:c5504. doi:10.1136/bmj.c5504.

4. Manuel DG, Rosella LC, Hennessy D, Sanmartin C, Wilson K. Predictive risk algorithms in a population setting: an overview. *J Epidemiol Community Health*. 2012;66(10):859-865. doi:10.1136/jech-2012-200971.

5. Danaei G, Vander Hoorn S, Lopez AD, Murray CJL, Ezzati M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet*. 2005;366(9499):1784-1793. doi:10.1016/S0140-6736(05)67725-2.

6. Schottenfeld D, Beebe-Dimmer JL, Buffler P a, Omenn GS. Current perspective on the global and United States cancer burden attributable to lifestyle and environmental risk factors. *Annu Rev Public Health*. 2013;34(February):97-117. doi:10.1146/annurev-publhealth-031912-114350.

7. Pencina MJ, D'Agostino RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. *Circulation*. 2009;119(24):3078-3084. doi:10.1161/CIRCULATIONAHA.108.816694.

8. Rasmussen-Torvik LJ1, Shay CM, Abramson JG, Friedrich CA, Nettleton JA, Prizment AE FA. Cancer : The Atherosclerosis Risk in Communities Study. *Circulation*. 2013;127(12):1270-1275. doi:10.1161/CIRCULATIONAHA.112.001183.Ideal.

9. Romaguera D, Vergnaud A-C, Peeters PH, et al. Is concordance with World Cancer Research Fund/American Institute for Cancer Research guidelines for cancer prevention related to subsequent risk of cancer? Results from the EPIC study. *Am J Clin Nutr*. 2012;96(1):150-163. doi:10.3945/ajcn.111.031674.

10. Thomson CA, Mccullough ML, Wertheim BC, et al. Nutrition and Physical Activity Cancer Prevention Guidelines , Cancer Risk , and Mortality in the

Women's Health Initiative. 2014;7(January):42-54. doi:10.1158/1940-6207.CAPR-13-0258.

11. Kabat GC, Matthews CE, Kamensky V, Hollenbeck AR, Rohan TE. Adherence to cancer prevention guidelines and cancer incidence , cancer mortality , and total mortality : a prospective cohort study 1 – 4. *Am J Clin Nutr*. 2015;(1):558-569. doi:10.3945/ajcn.114.094854.INTRODUCTION.

12. Dartois L, Fagherazzi G, Mesrine S. Association between Five Lifestyle Habits and Cancer Risk : Results from the E3N Cohort. *Cancer Prev Res*. 2014;7(May):26-28. doi:10.1158/1940-6207.CAPR-13-0325.

13. Tanaka S, Yamamoto S, Inoue M, et al. Projecting the probability of survival free from cancer and cardiovascular incidence through lifestyle modification in Japan. *Prev Med (Baltim)*. 2009;48(2):128-133. doi:10.1016/j.ypmed.2008.11.006.

14. Colditz GA, Atwood KA, Emmons K, et al. Harvard Report on Cancer Prevention Volume 4 : Harvard Cancer Risk Index. 2000;4.

15. Sethi TK, El-ghamry MN, Kloecker GH. Radon and Lung Cancer. *Clin Adv Hematol Oncol*. 2012;10(3):157-164.

16. Sidorchuk A, Agardh EE, Aremu O, Hallqvist J, Allebeck P, Moradi T. Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis. *Cancer Causes Control*. 2009;20(4):459-471. doi:10.1007/s10552-009-9300-8.

17. Colditz GA, Atwood KA, Emmons K, et al. Harvard Report on Cancer Prevention Volume 4 : Harvard Cancer Risk Index. *Cancer Causes Control*. 2000;4:477-488.

18. Rowe AK, Powell KE, Flanders WD. Why population attributable fractions can sum to more than one. *Am J Prev Med*. 2004;26(3):243-249. doi:10.1016/j.amepre.2003.12.007.

19. Rosella LC, Manuel DG, Burchill C, Stukel T a. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *J Epidemiol Community Health*. 2011;65(7):613-620. doi:10.1136/jech.2009.102244.

20. Manuel DG, Roberts M, Manson H. *SEVEN MORE YEARS : The Impact of Smoking , Alcohol , Diet , Physical Activity and Stress on Health and Life Expectancy in Ontario*.; 2012. http://www.ices.on.ca/Publications/Atlases-and-Reports/2012/Seven-More-Years.

21. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015;162(1). doi:10.7326/M14-0697.

22. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer : a review. *BMC Med*. 2010;8(21).

23. Bouwmeester W, Zuithoff NP a, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9(5):1-12. doi:10.1371/journal.pmed.1001221.

24. Harrell Jr FE, Lee KL, Mark DB. Tutorial in Biostatistics Multivariable Prognostic Models: Issues in dEVELOPING models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Stat Med*. 1996;15(4):361-387.

25. Harrell Jr FE. *Regression Modeling Strategies*. NY, USA: Springer-Verlag New York, Inc.; 2001.

26. Canadian Community Health Survey - Annual Component (CCHS). January 2015. http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&Item_Id=144171&lang=en. Accessed September 9, 2015.

27. Seer NCI. SEER Program Coding and Staging Manual 2011. 2011;(14):1-167. papers2://publication/uuid/512EBCE8-D635-4348-A67D-22DD52988F4C.

28. Taljaard M, Tuna M, Bennett C, et al. Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open*. 2014;4(10):e006701-e006701. doi:10.1136/bmjopen-2014-006701.

29. Blumchent G. Metabolic Equivalents (METS) in Exercise Testing , Exercise Prescription , and Evaluation of Functional Capacity. *Clin Cardiol*. 1990;13(8):555-565.

30. Desapriya E, Turcotte K, Subzwari S, Pike I. Smoking inside vehicles should be banned globally. *Am J Public Health*. 2009;99(7):1158-1159; author reply 1159. doi:10.2105/AJPH.2009.160127.

31. Sendzik T, Fong GT, Travers MJ, Hyland A. An experimental investigation of tobacco smoke pollution in cars. *Nicotine Tob Res*. 2009;11(6):627-634. doi:10.1093/ntr/ntp019.

32. McCracken M, Olsen M, Chen MS, et al. Cancer Incidence, Mortality, and Associated Risk Factors Among Asian Americans of Chinese, Filipino, Vietnamese, Korean, and Japanese Ethnicities. *CA Cancer J Clin*. 2007;57(4):190-205. doi:10.3322/canjclin.57.4.190.

33. Pampalon R, Hamel D, Gamache P, Philibert MD, Raymond G, Simpson A. An Area-based Material and Social Deprivation Index for Public Health in Québec and Canada. *Can J Public Heal*. 2012;(October):17-22.

34. Canada S. Tables. 2012. http://www.statcan.gc.ca/pub/82-231-x/2010001/t007-eng.pdf. Accessed March 11, 2016.

35. Harrell F. aregImpute {Hmisc}. http://www.inside-r.org/packages/cran/hmisc/docs/aregImpute. Accessed May 9, 2016.

36. Little R & Rubin D. *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons Inc; 2002.

37. Fine, J & Gray R. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc*. 1999;94:496-509.

38. Van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol*. 2012;65(2):126-131. doi:10.1016/j.jclinepi.2011.08.002.

39. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2.Assessing.

40. Pace NL, Eberhart LHJ, Kranke PR. Quantifying prognosis with risk predictions. *Eur J Anaesthesiol*. 2012;29(1):7-16. doi:10.1097/EJA.0b013e32834d9474.

41. Kremers W. *Concordance for Survival Time Data: Fixed and Time-Dependent Covariates and Possible Ties in Predictor and Time*.; 2007. http://www.mayo.edu/research/documents/biostat-80pdf/doc-10027891.

42. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst*. 2005;97(10):715-723. doi:10.1093/jnci/dji128.

43. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol*. 2011;29(22):2951-2952. doi:10.1200/JCO.2011.36.1329.

44. Arnold M, Razum O, Coebergh J. Cancer risk diversity in non-western migrants to Europe : An overview of the literature. *Eur J Cancer*. 2010;46(14):2647-2659. doi:10.1016/j.ejca.2010.07.050.

45. Desmeules M, Gold J, Kazanjian A, et al. New Approaches to Immigrant Health Assessment. *Can J Public Heal*. 2004;95(3):I22-I26.

46. Powell HA, Iyen-omofoman B, Baldwin DR, Hubbard RB, Tata LJ. Chronic obstructive pulmonary disease and risk of lung cancer: the importance of smoking and timing of diagnosis. *J Thorac Oncol*. 2013;8(1):6-11.

47. Brenner DR, Mclaughlin JR, Hung RJ. Previous Lung Diseases and Lung Cancer Risk : A Systematic Review and Meta-Analysis Abstract Background Materials and Methods Literature Review. *PLoS One*. 2011;6(3):e17479.

48. Schütze M1, Boeing H, Pischon T, Rehm J, Kehoe T, Gmel G, Olsen A, Tjønneland AM, Dahm CC, Overvad K, Clavel-Chapelon F, Boutron-Ruault MC, Trichopoulou A, Benetou V, Zylis D, Kaaks R, Rohrmann S, Palli D, Berrino F, Tumino R, Vineis P, Rodríguez L, Agudo BM. Alcohol attributable burden of incidence of cancer in eight European countries based on results from prospective cohort study. *BMJ*. 2011;342:d1584.

49. Hanahan D, Weinberg R a. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013.

50. Hamajima N, Hirose K, Tajima K, et al. Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br J Cancer*. 2002;87(11):1234-1245. doi:10.1038/sj.bjc.6600596.

51. Thompson PL. J-curve revisited: cardiovascular benefits of moderate alcohol use cannot be dismissed. *Med J Aust*. 2013;198(8):419-422. doi:10.5694/mja12.10922.

52. Mulder M, Ranchor A V, Sanderman R, Bouma J, Ja W, Heuvel V Den. The stability of lifestyle behaviour. 1998:199-207.

53. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40(5 Pt 2):1620-1639. doi:10.1111/j.1475-6773.2005.00444.x.

54. Cary KC, Cooperberg MR. Biomarkers in prostate cancer surveillance and screening: past, present, and future. *Ther Adv Urol*. 2013;5(6):318-329. doi:10.1177/1756287213495915.

55. Usher-smith J, Emery J, Hamilton W, Griffin SJ, Walter FM. Risk prediction tools for cancer in primary care. *Br J Cancer*. 2015;(October):1645-1650. doi:10.1038/bjc.2015.409.

56. Muggah E, Graves E, Bennett C, Manuel DG. Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health*. 2013;13:16. doi:10.1186/1471-2458-13-16.

57. Hsieh FY, Lavori PW. Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates. *Control Clin Trials*. 2000;(6):552-560.

APPENDIX 1: Defining Pre-Existing Diagnosis of Malignant Cancer for Exclusion Criteria

This study required a population in which individuals with previous malignant cancer had been excluded, to be able to identify only primary malignancies as the outcome of interest. During the data preparation phase for this thesis, two possible methods of identifying individuals with previous cancer were examined, based on either self-reported prior malignancy in the CCHS survey, or based on previous malignant cancers captured in the Ontario Cancer Registry prior to CCHS survey administration date.

The issue of whether population health surveys are adequate to identify individuals with disease has been raised in other contexts, such as for those with diabetes, congestive heart failure and stroke.[56] First, health surveys are subject to recall bias. Second, self-report of malignancies is likely subject to misclassification error due to under-reporting incidence of cancer that is in remission following surgical or chemotherapy/radiation therapy intervention. This is particularly an issue in the earlier cycles of the CCHS, in which the question regarding cancer is whether an individual currently has cancer, not if they have ever had cancer.[26] (see Table A1) Third, incidences of typically non-malignant cancers such as non-melanoma skin cancers can also result in misclassification but as false positives.

Table A1: Question posed in each CCHS cycle regarding any cancer diagnosis
CCHS Cycle 1.1 (2000-2001): CC_Q131: "Do you have cancer?"

CCHS Cycle 2.1 (2003):
CCC_Q131: "Do you have cancer?"

CCHS Cycle 3.1 (2005):
CCC_Q131: "Do you have cancer?"
CCC_Q132: "Have you ever been diagnosed with cancer?"

CCHS Cycle 4.1 (2007):
CCC_Q131: "Do you have cancer?"
CCC_Q132: "Have you ever been diagnosed with cancer?"

*CCHS questionnaires from Statistics Canada[26]

Misclassification error may be dependent on the length of time since diagnosis (i.e. a longer time period may offer more opportunity for remission and self-identification as having no malignancy) and recall bias will also depend on time, therefore it is important to assess if the utility of self-reported data varies over different periods of analysis or look-back windows.

After exclusion of those individuals not eligible for OHIP insurance at CCHS survey administration and those individuals less than 20 years old, 115 924 individuals remained in the study population. Using this population, the accuracy of CCHS self-reported malignancy was assessed against the "gold standard" of OCR-identified malignancy prior to CCHS survey administration using measures of sensitivity and specificity.

The results indicate that there was a high rate of false positive CCHS self-reported cancer even when examining a >20 year period prior to CCHS survey administration. The high false positive rate may have been a result of individuals having moved from out of province (therefore malignancy data is in another province's/country's cancer registry)

and those who had a non-malignant cancer (such as a non-melanoma skin cancer) who still self-report having had a malignancy. The high false positive rate reduced both the positive predictive value (PPV) and the specificity, although specificity was consistently fairly high due to the large number of individuals who self-reported no malignancy in the CCHS and were found not to have a malignancy in the OCR (i.e. true negatives).

There were also a number of individuals who did not report a history of malignancy in the CCHS but were found to have a malignancy in the OCR, and this number increased with longer lookback windows (or periods of time prior to survey administration). These false negative results led to a reduced sensitivity, meaning a reduced ability to detect true cancer diagnoses using the CCHS survey results, among all those with cancer identified in the OCR. The false negative rate may have been a result of those who had a cancer in remission or cured due to surgical or medical intervention and therefore did not self-report having a malignancy and may be due to the question in earlier cycles of the CCHS that only asked about current malignancy. (see Table A2)

Table A2: Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Likelihood Ratio (LR) of Self-Reported malignancy compared to "Gold Standard" Ontario Cancer Registry data based on varying periods prior to CCHS survey administration

	Period Prior to CCHS Survey Administration					
Summary Measure	**1 year**	**5 years**	**10 years**	**15 years**	**20 years**	**>20 years**
Sensitivity	0.761	0.711	0.688	0.677	0.670	0.665
Specificity	0.943	0.953	0.961	0.965	0.967	0.970
PPV	0.076	0.242	0.380	0.455	0.489	0.529
NPV	0.998	0.994	0.989	0.986	0.984	0.982
LR	13.35	15.13	17.64	19.34	20.30	22.17

Overall Incidence (%) malignancy reported in CCHS vs. recorded in OCR:

CCHS previous cancer incidence= 7 056/115 962 (100%)= 6.1%
OCR previous cancer incidence= 5 617/115 962 (100%) = 4.8%

Based on these results, it was decided to exclude individuals who had a prior diagnosis of cancer either as self-reported in the CCHS or in the Ontario Cancer Registry prior to CCHS survey administration date. This decision was reached due to the low sensitivity when comparing CCHS self-report to "gold standard" OCR cancer data, as well as the fact that once an individual was flagged in the OCR as having cancer prior to the survey administration date they were subsequently incapable of developing the primary outcome of interest (first primary malignancy at any site).

APPENDIX 2: Sample Size Calculation

A second method of calculating sample size was utilized to verify the study power. This calculation is based on the article by Hseih et. al. (2000) in which the equation to determine number of deaths required in a survival analysis study to achieve a certain power is[57]:

$$D = (Z_{1-\alpha} + Z_{1-\beta})^2 [\sigma^2 (\log\Delta)^2] - 1$$

Where D= deaths α = type 1 error ß= type 2 error σ2= variance of a covariate, X1 logΔ = the log hazard ratio associated with a one unit change in X1.

In this study, rather than being interested in the outcome of death, we are interested in the incidence of cancer and therefore incidence of cancer (I) will be substituted for deaths (D). The variables of interest in our study are smoking, diet, physical activity and alcohol consumption. There is minimal literature on the association of each of these factors with combined cancer incidence. Some of the highest quality data available is from the EPIC studies, one of which assess the burden of incident cancer based on alcohol consumption.[48] In addition, it would be expected that alcohol consumption may have a lower impact on cancer incidence than such factors as smoking, therefore an estimate based on alcohol consumption would likely be more conservative. From the EPIC study, HR for cancer based on 12 g/day (or one standard alcoholic drink) for women (in whom the confidence interval around the HR is wider) is 1.03 (95% CI: 1.01 to 1.05). With alpha=0.05 and beta=0.80, and given a mean consumption of 1.32 drinks per day and SD of 1.64 from the study article, the equation is as follows:

Based on the hazard ratio, the number of incident cancer cases required would be:

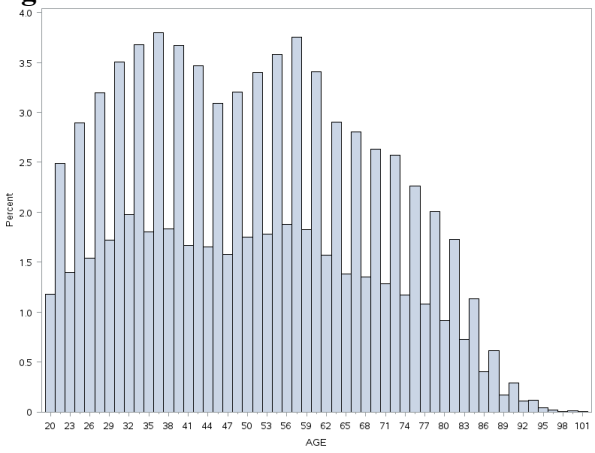
$I=(Z\ 1-\alpha + Z\ 1-ß)2\ [\sigma 2(\log\Delta)2]\ -1$
$I=(1.645 + 0.842)2\ [1.642(\ln(1.03))2]\ -1$
$I=(6.185)[1.642(0.0296)2]-1$
$I=(6.185)[2.69(0.000874)]-1$
$I=(6.185)(0.0024)-1$
$I=(6.185)(416.70)$
$I=2577.08$

Using Statistics Canada estimates, the event rate of cancer in Ontario is 496.6/100 000 = 0.0049 or approximately 0.5 percent per year.[34] Multiplying the number of incident cancer cases by the inverse of the event rate gives the sample size required of 515 400. The 95% confidence interval around this estimate is wide, indicating an imprecise measure. However, there are 866 000 person-years of follow-up in this study, therefore given the estimated sample size it is an adequately powered study.

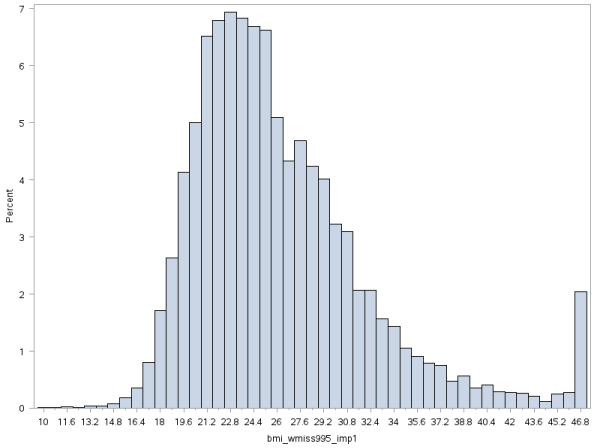
APPENDIX 3: Histograms of Continuous Predictor Variables in the Derivation Cohort (CCHS Cycles 2-4)

(A) Females, Unimputed Data

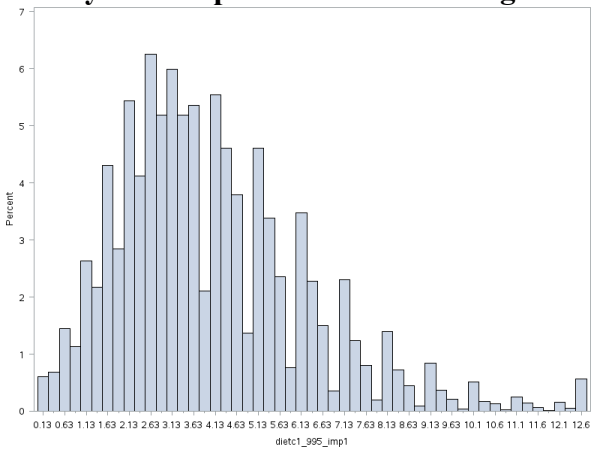
Age:



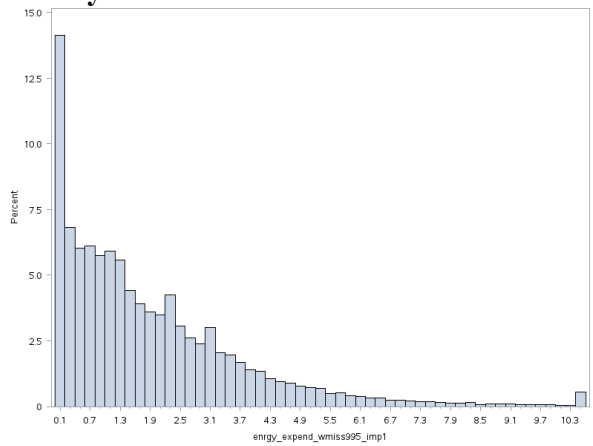
BMI:



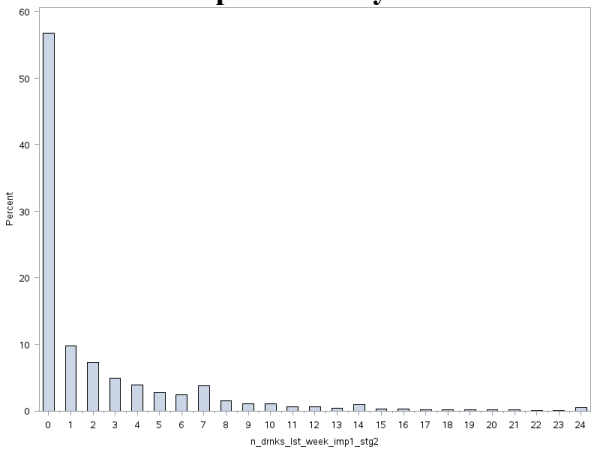
Dietary consumption of fruits and vegetables:



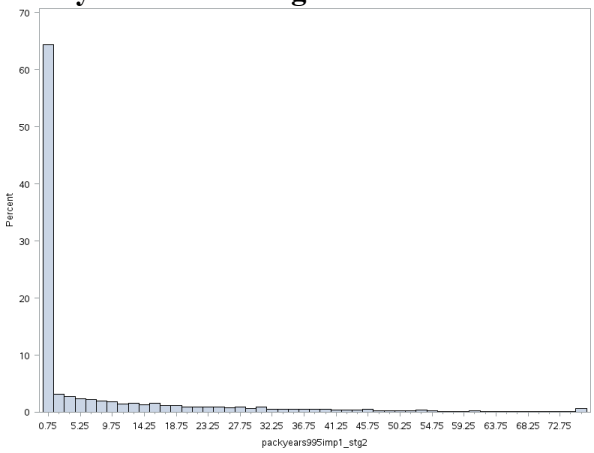
Activity level:



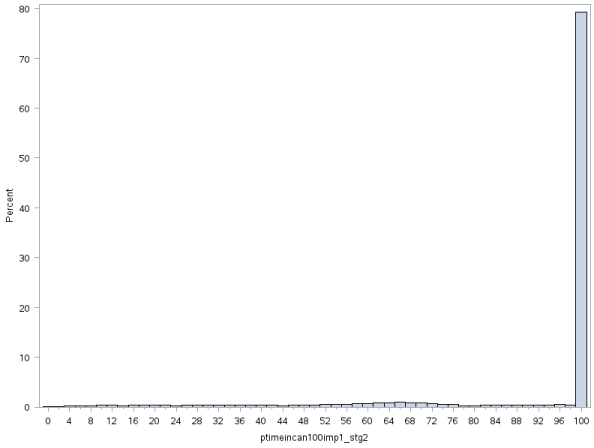
Alcohol consumption weekly:



Pack-years of smoking:

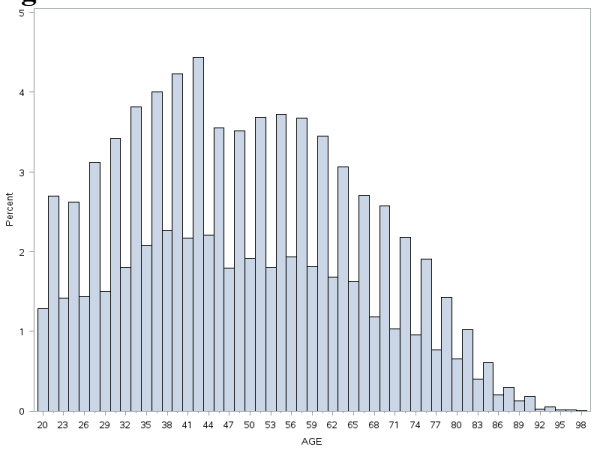


Percent time lived in Canada:

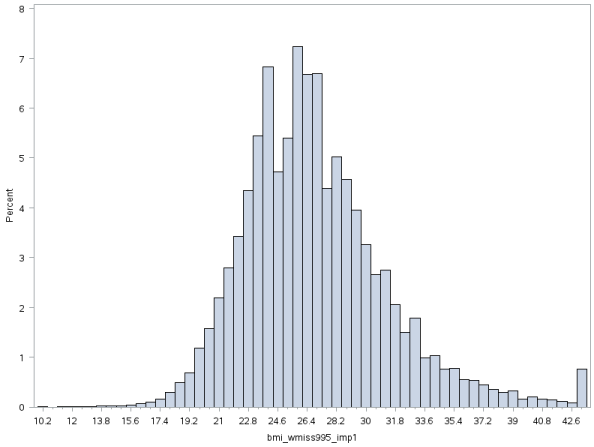


(B) Males, Unimputed Data

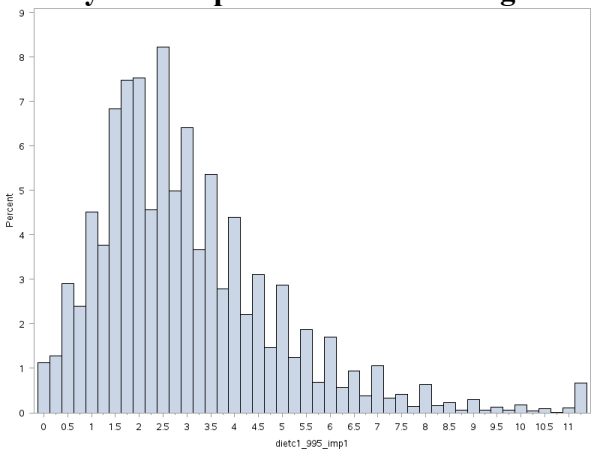
Age:



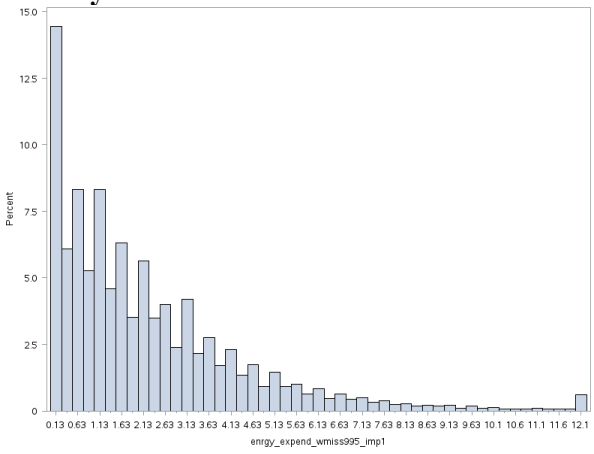
BMI:



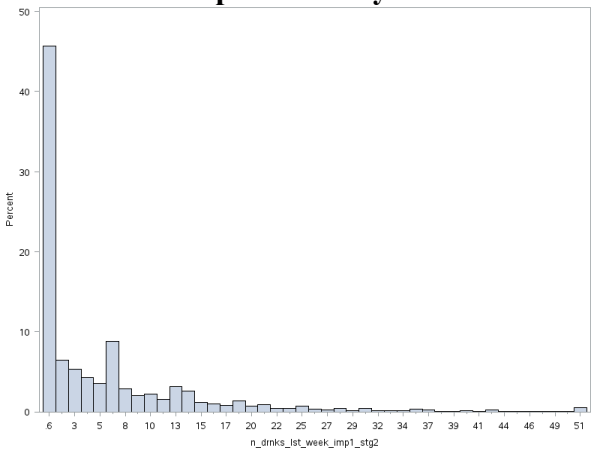
Dietary consumption of fruits and vegetables:



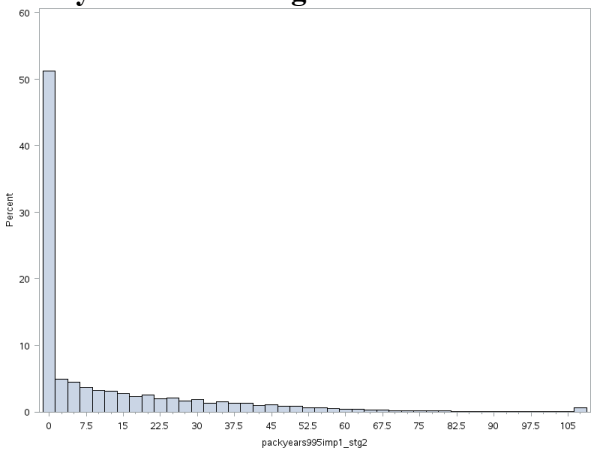
Activity level:



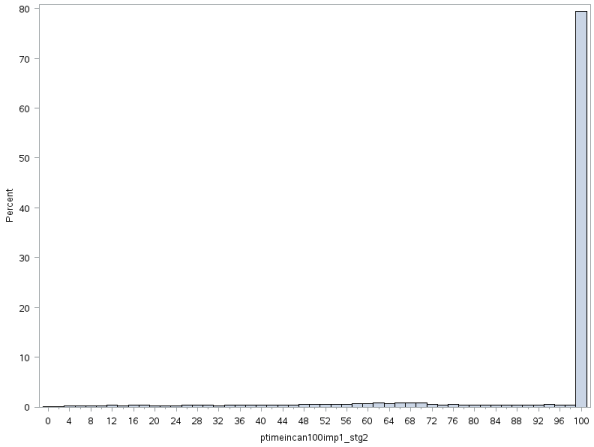
Alcohol consumption weekly:



Pack-years of smoking:



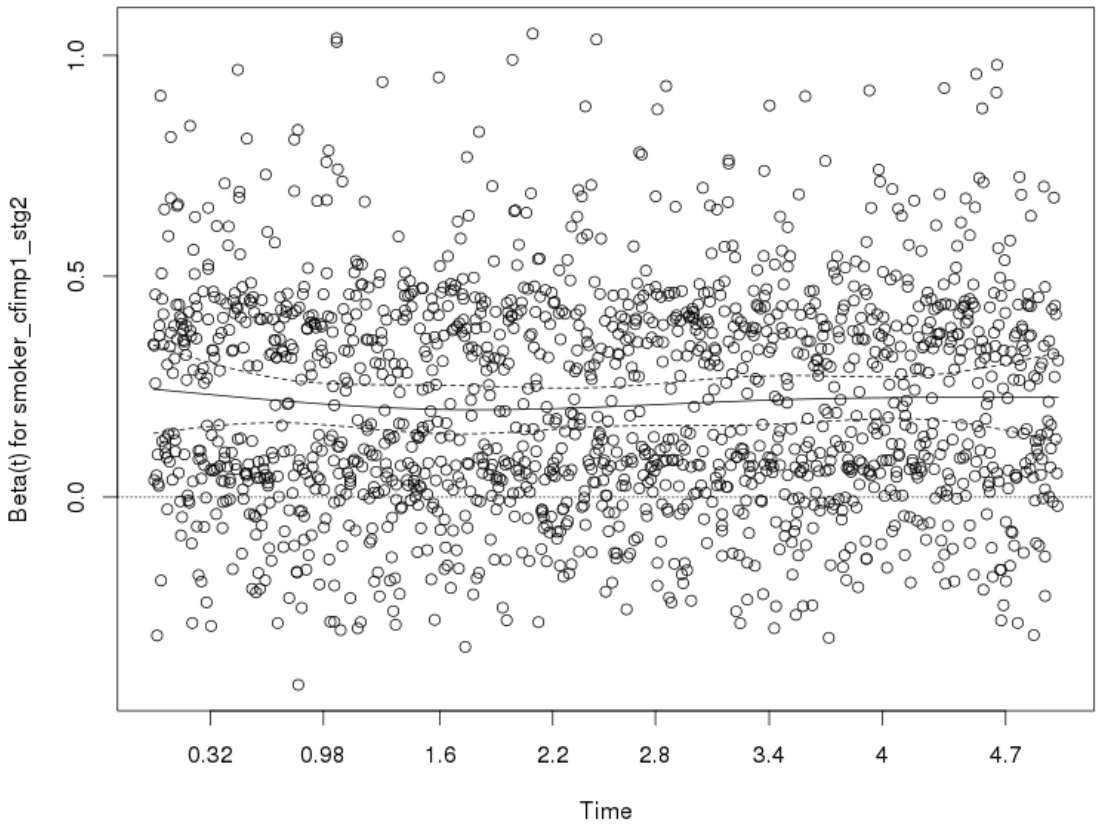
Percent time lived in Canada:



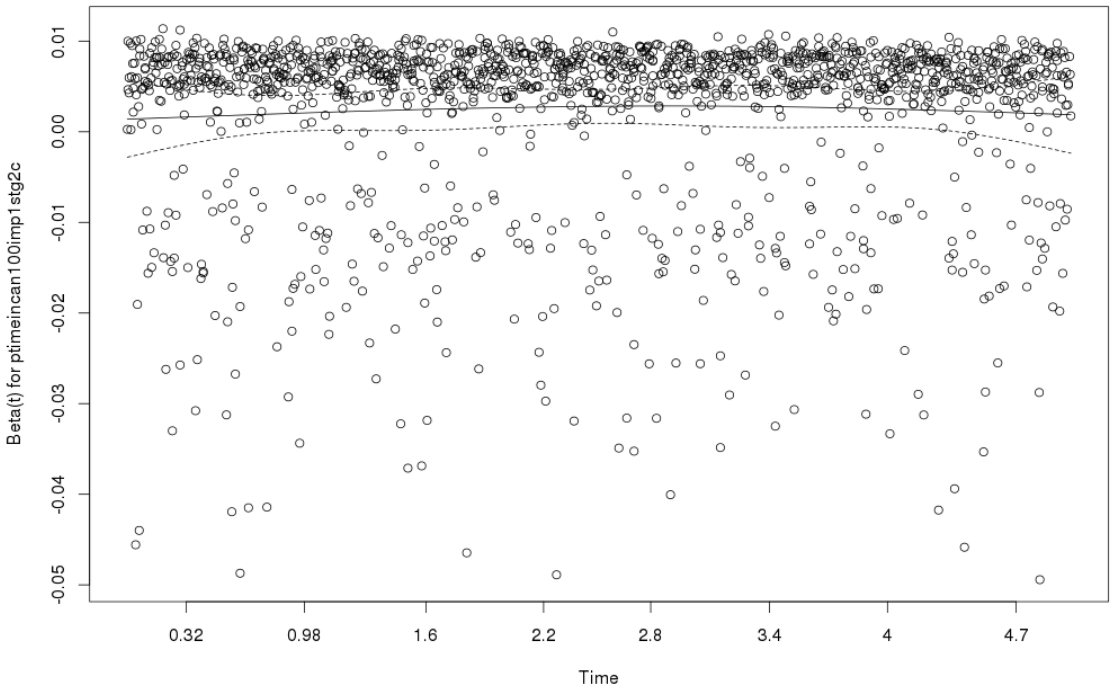
APPENDIX 4: Schoenfeld Residuals for Individual Predictor Variables Across Time

(A) Schoenfeld Residuals for Females

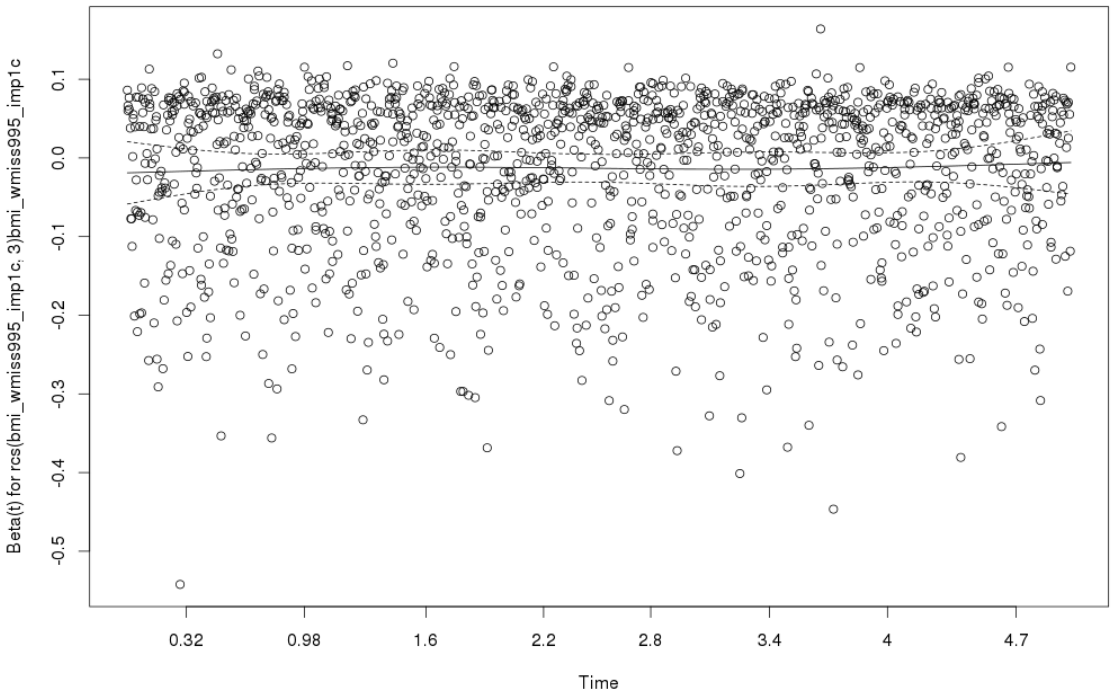
Schoenfeld Residuals Based on Smoker Category:



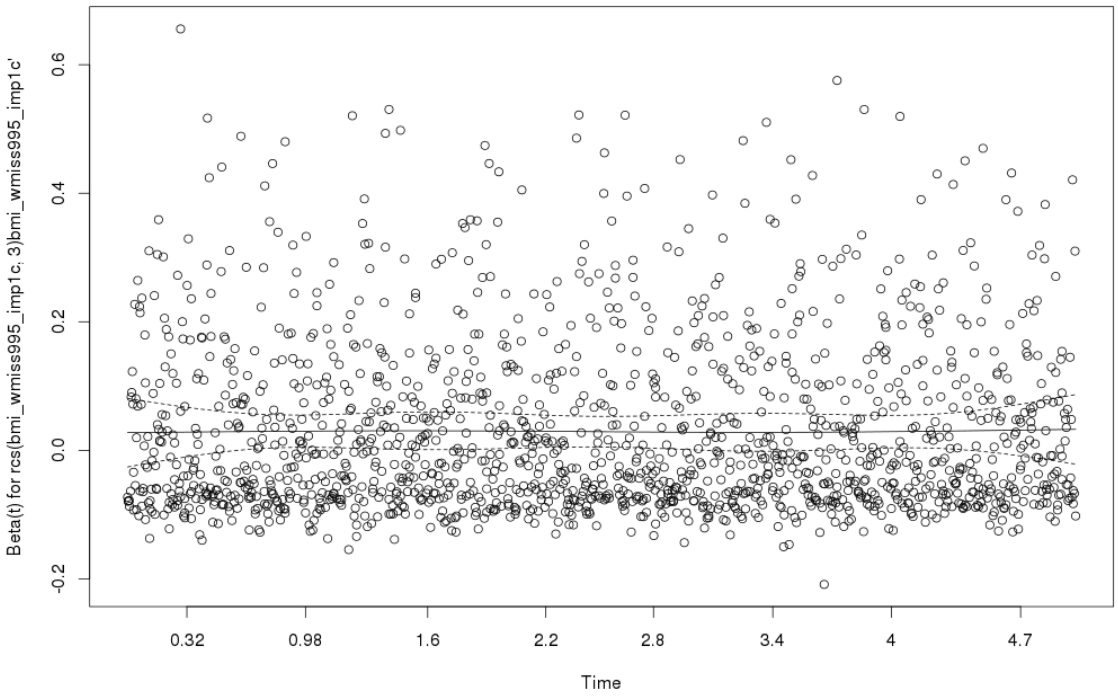
Schoenfeld Residuals Based on Percentage of Life Lived in Canada:



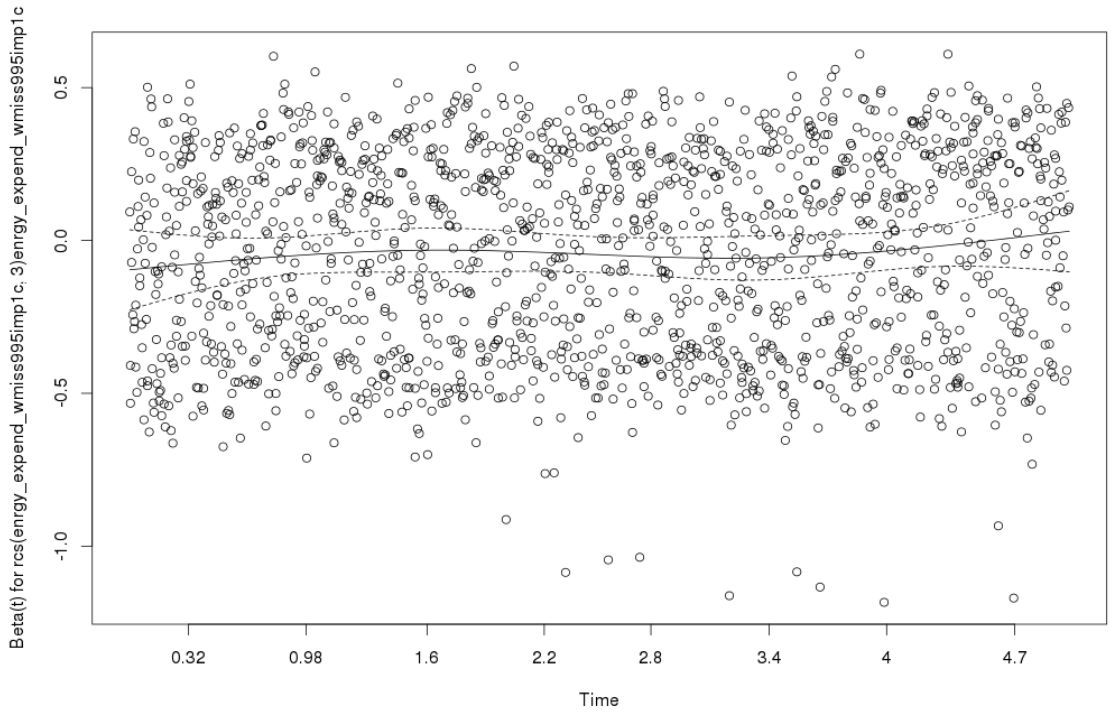
Schoenfeld Residuals Based on BMI- 1ST Spline:



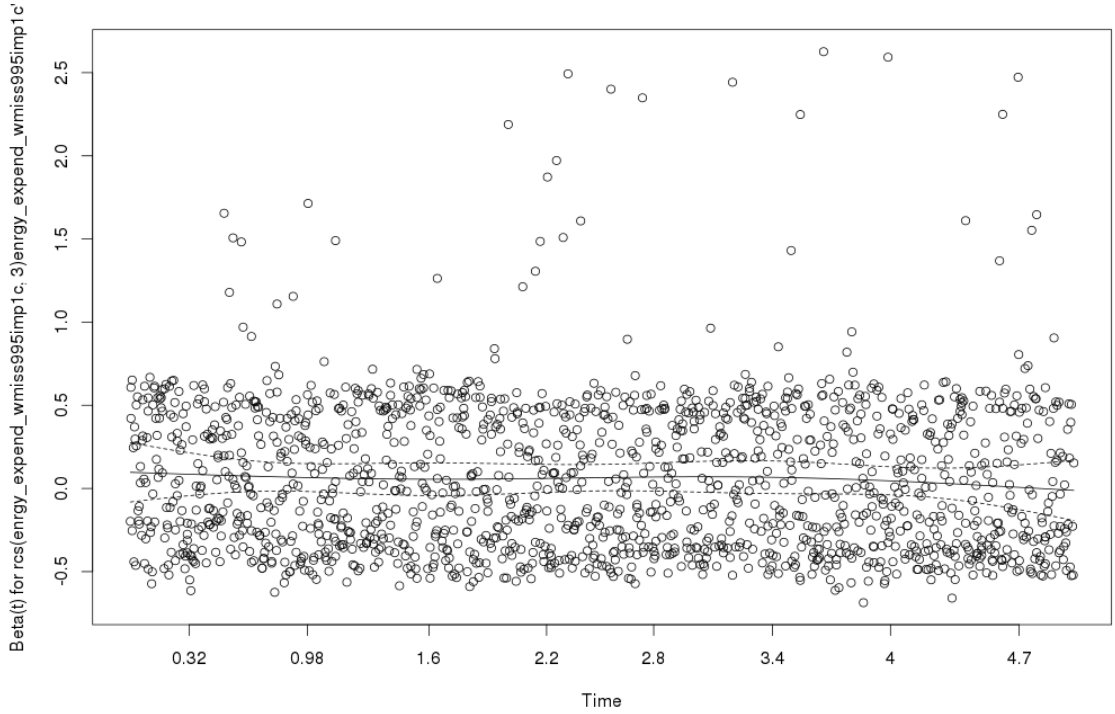
Schoenfeld Residuals Based on BMI- 2nd Spline:



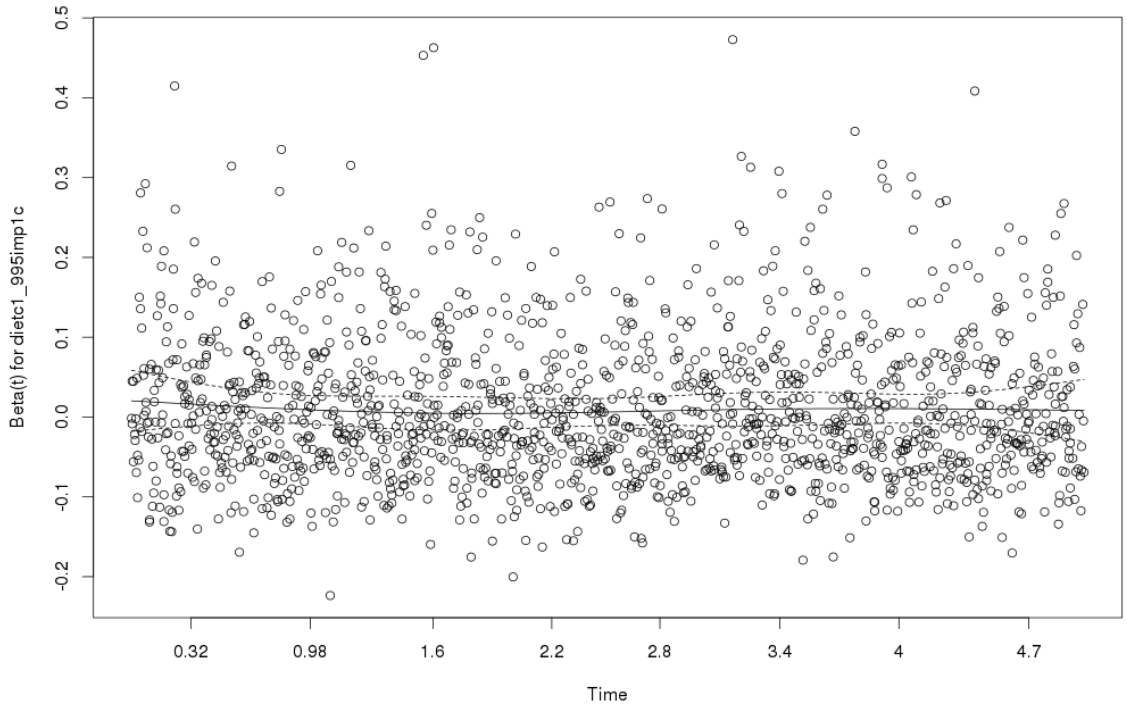
Schoenfeld Residuals Based on Leisure Physical Activity Level- 1st Spline:



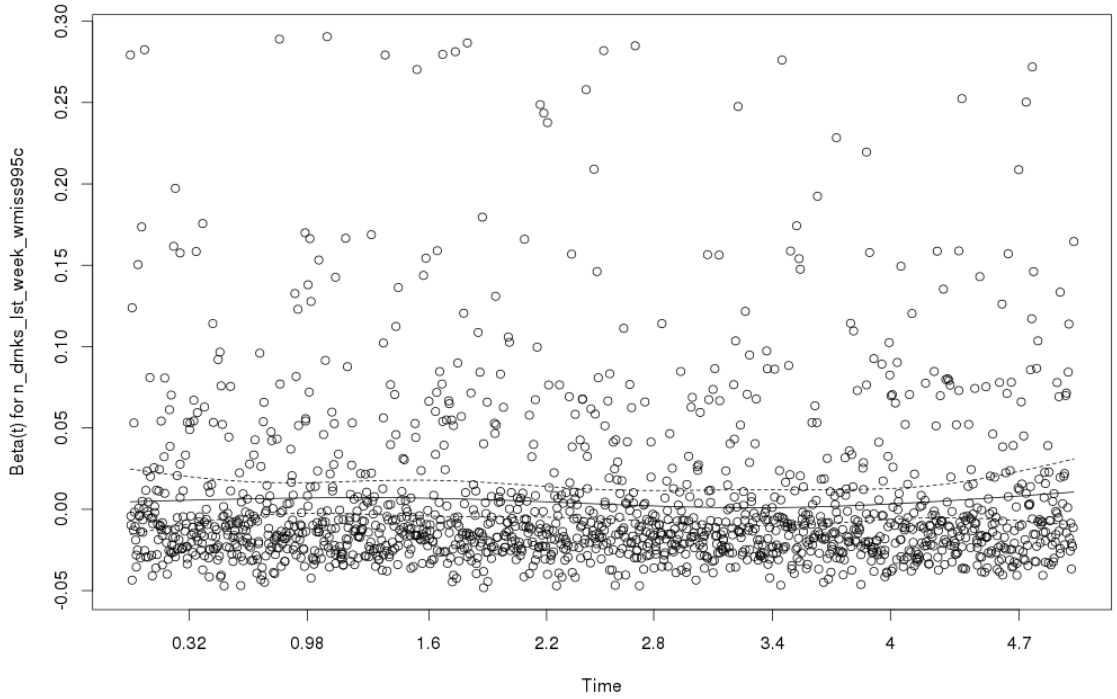
Schoenfeld Residuals Based on Leisure Physical Activity Level- 2nd Spline:



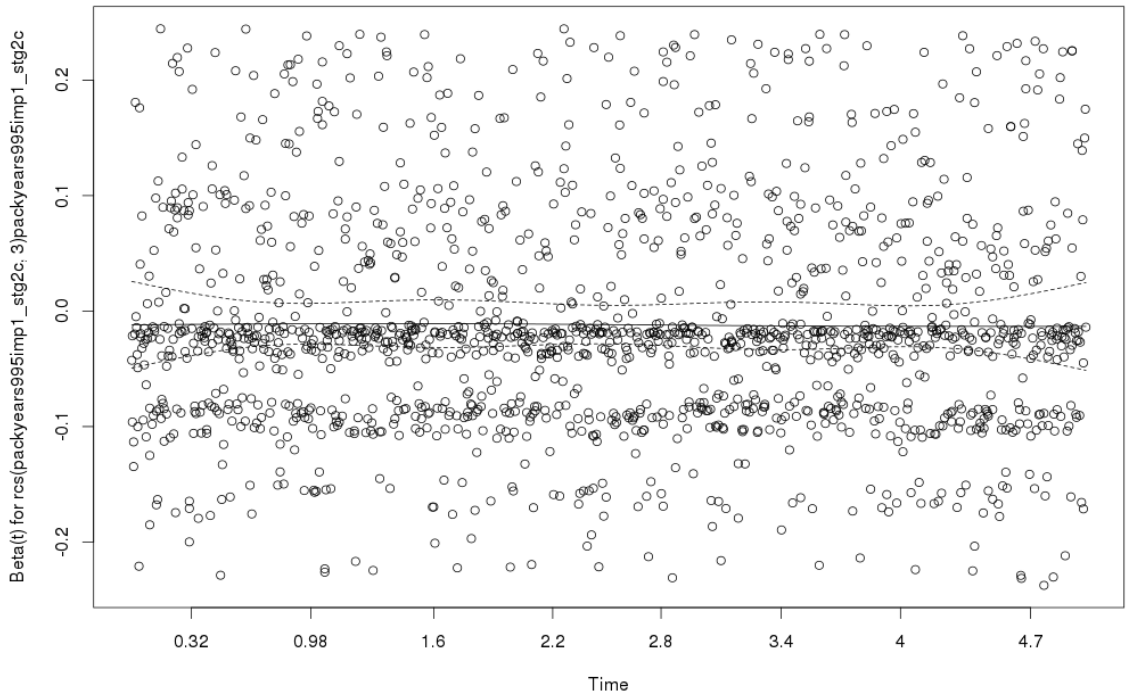
Schoenfeld Residuals Based on Dietary Consumption of Fruits and Vegetables:



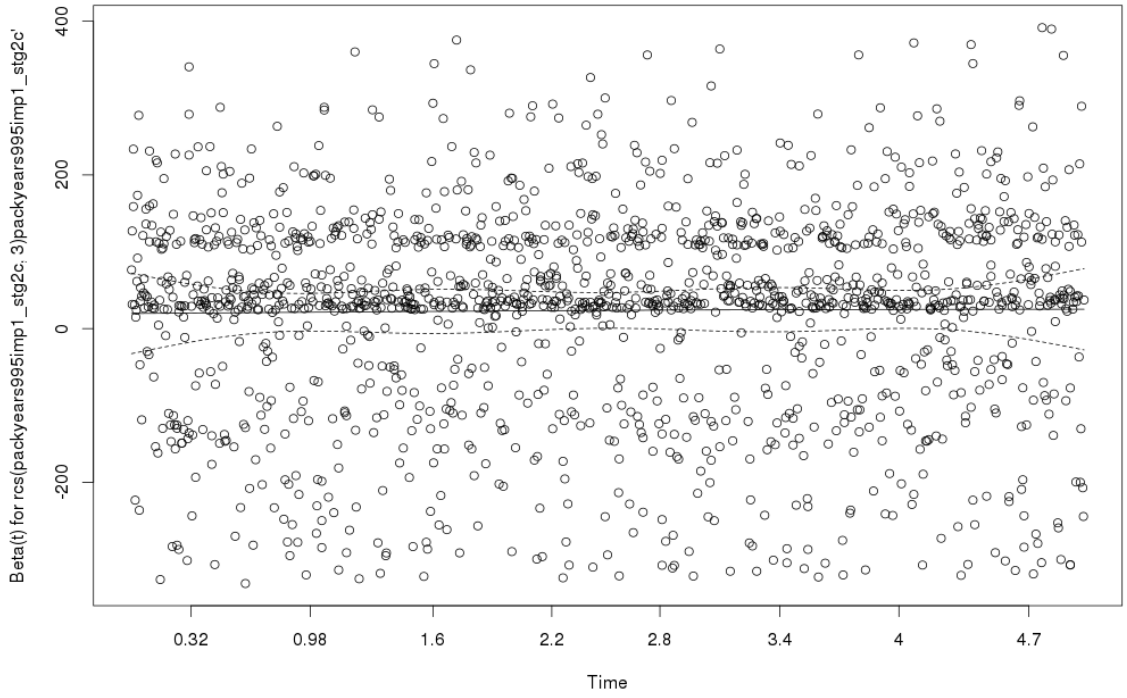
Schoenfeld Residuals Based on Alcohol Consumption:



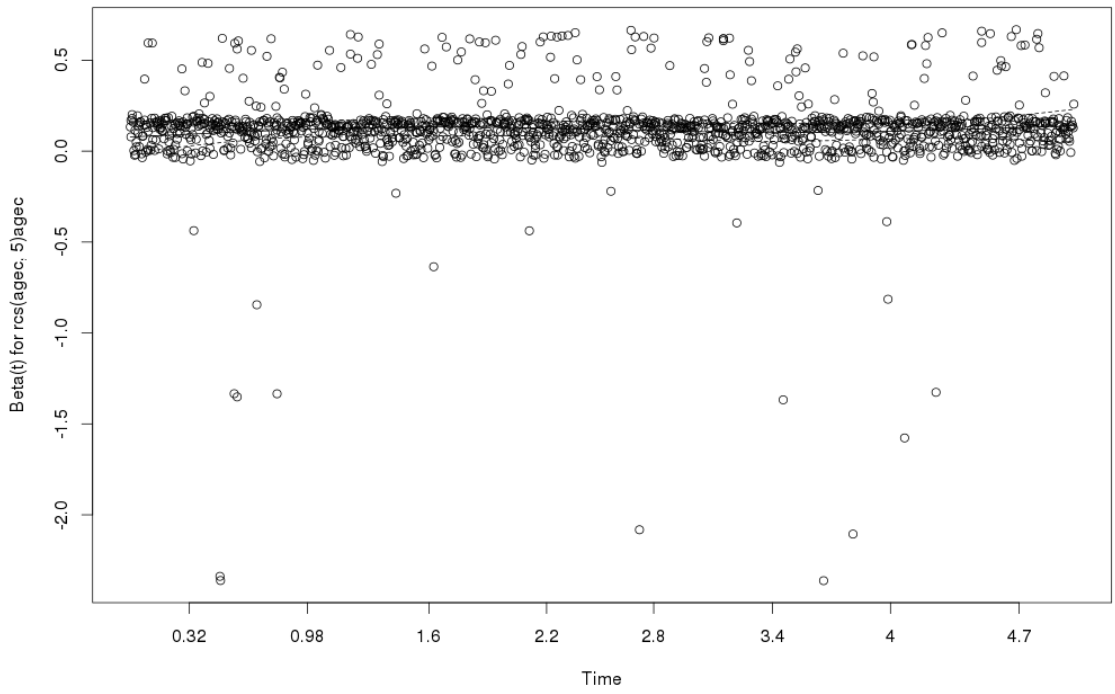
Schoenfeld Residuals Based on Pack-Years of Smoking- 1st Spline:



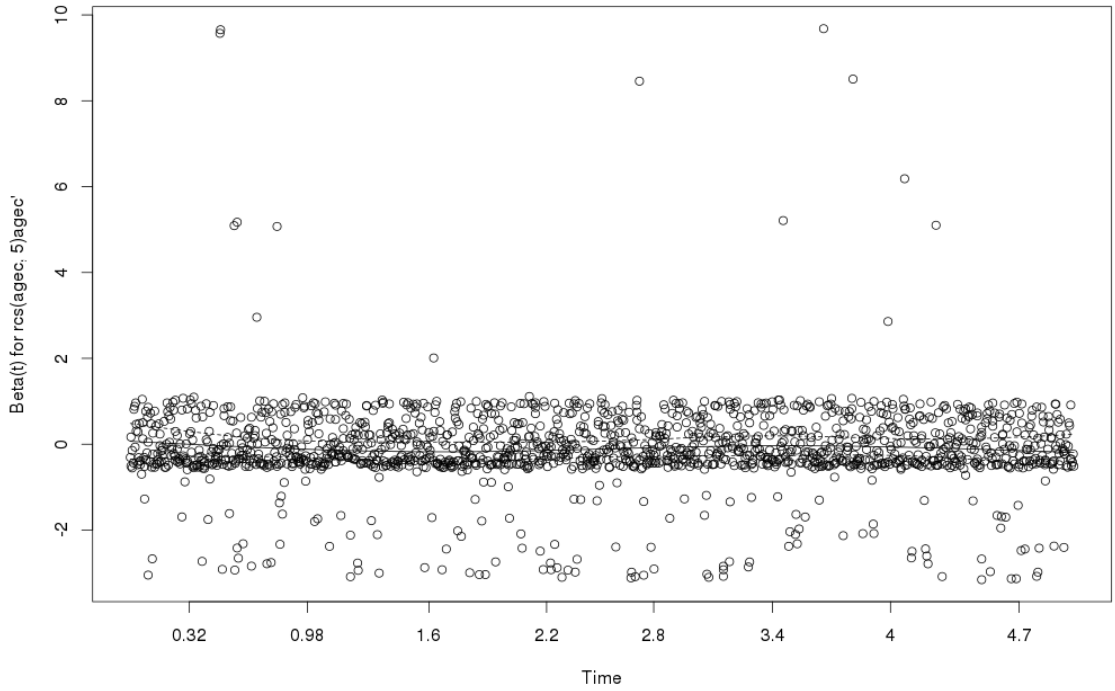
Schoenfeld Residuals Based on Pack-Years of Smoking- 2ND Spline:



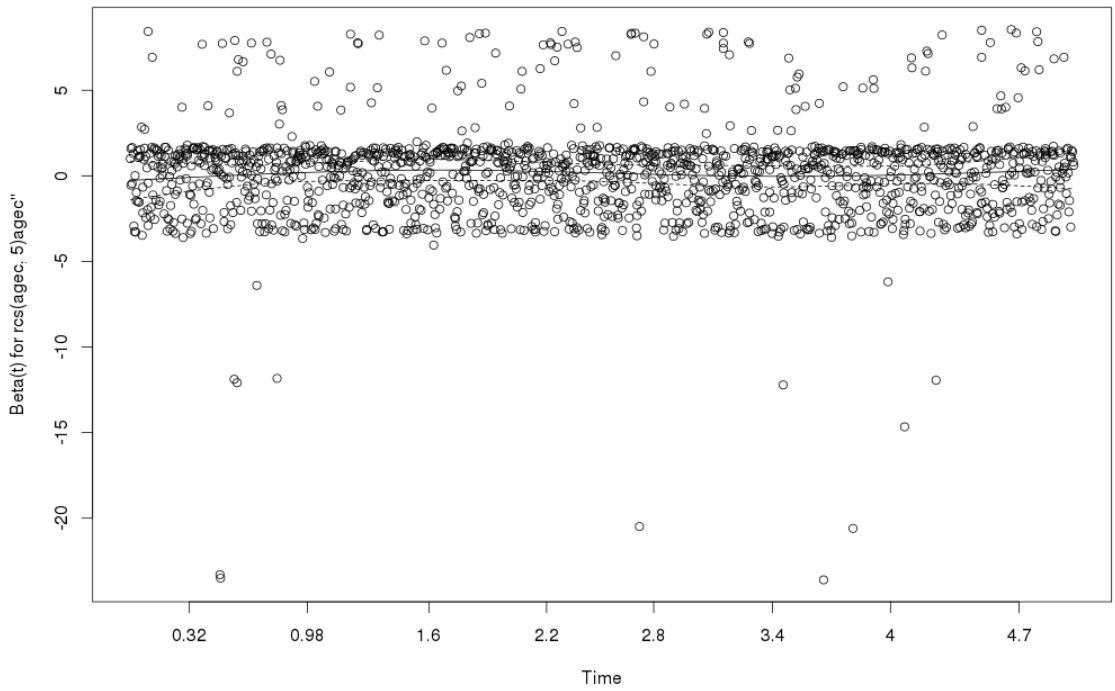
Schoenfeld Residuals Based on Age- 1ST Spline:



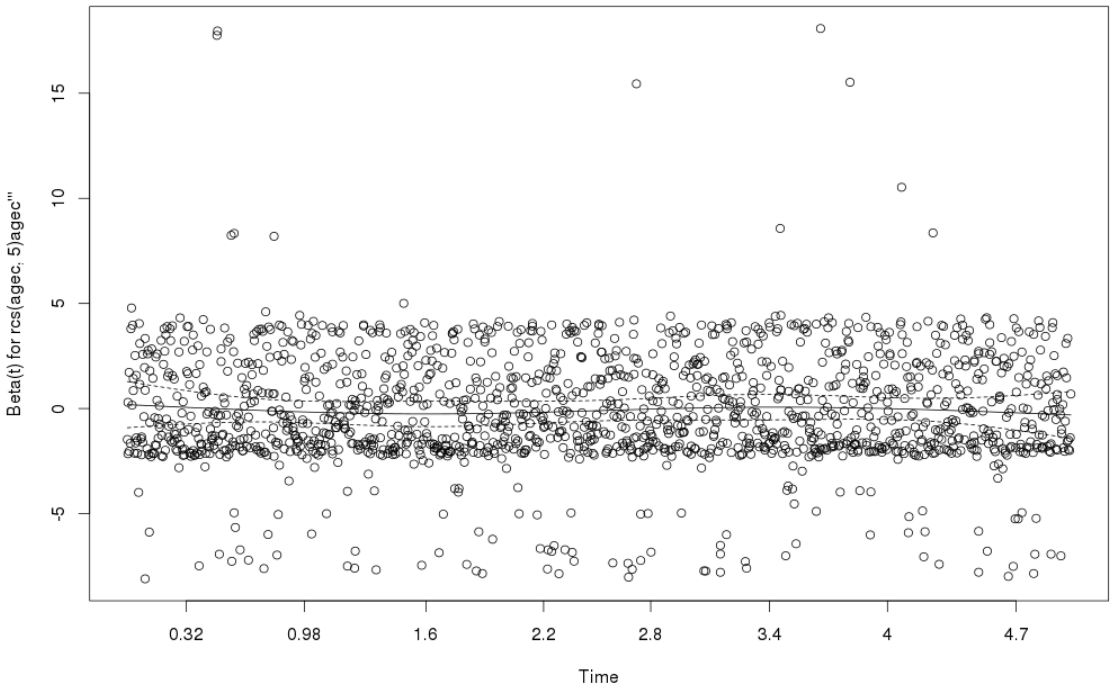
Schoenfeld Residuals Based on Age- 2nd Spline:



Schoenfeld Residuals Based on Age- 3rd Spline:

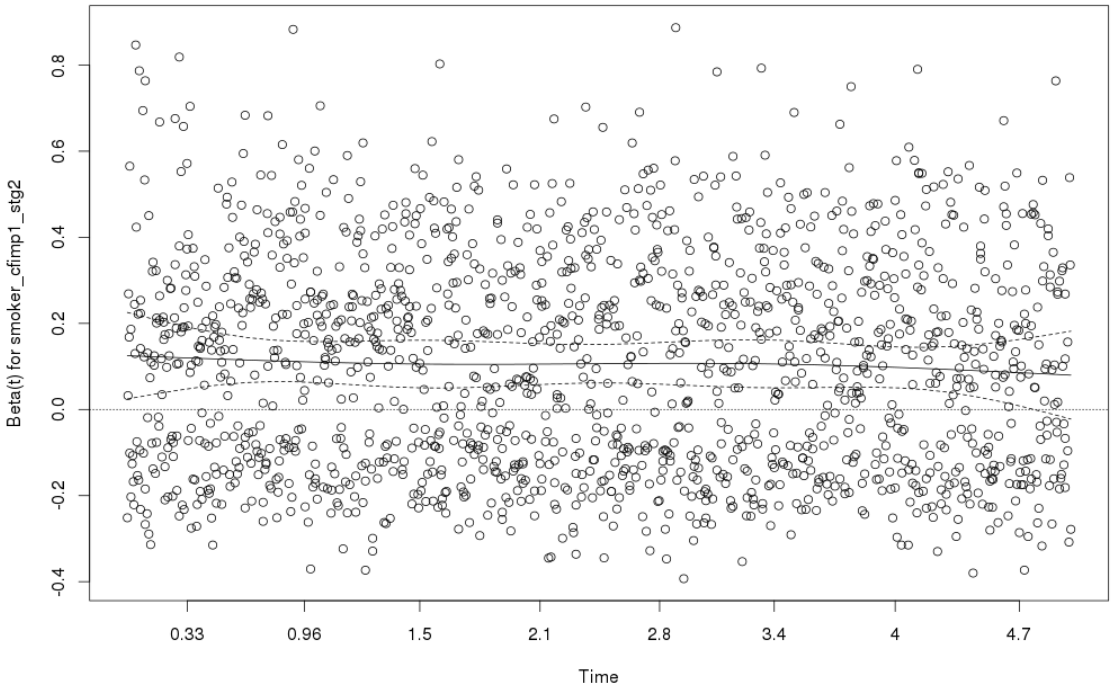


Schoenfeld Residuals Based on Age- 4th Spline:

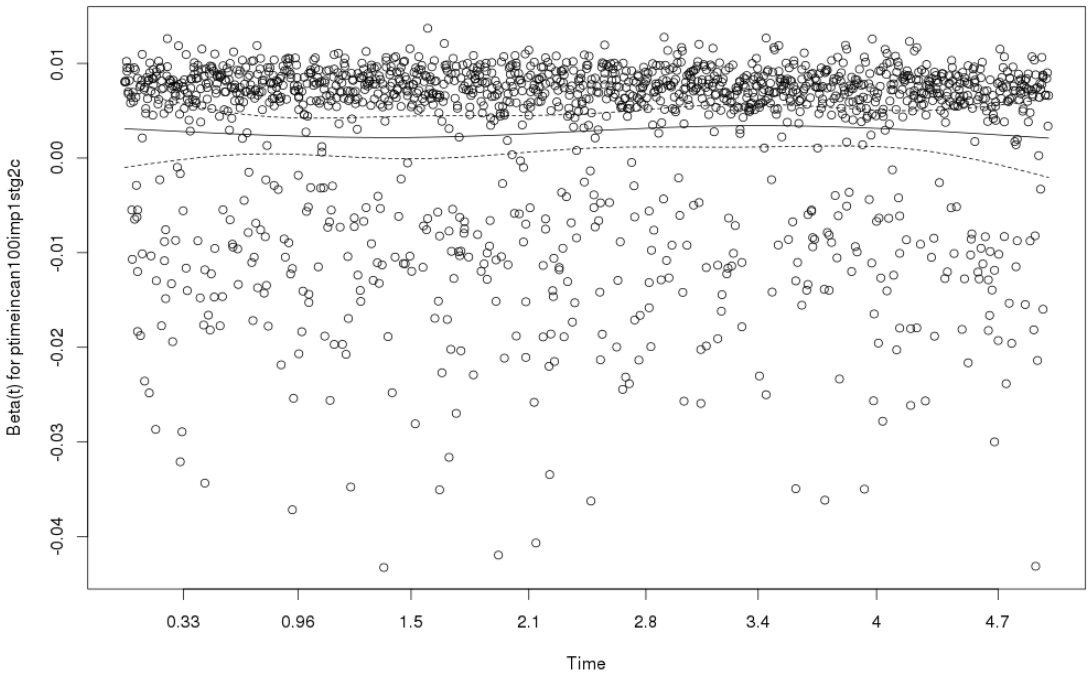


(B) Schoenfeld Residuals for Males

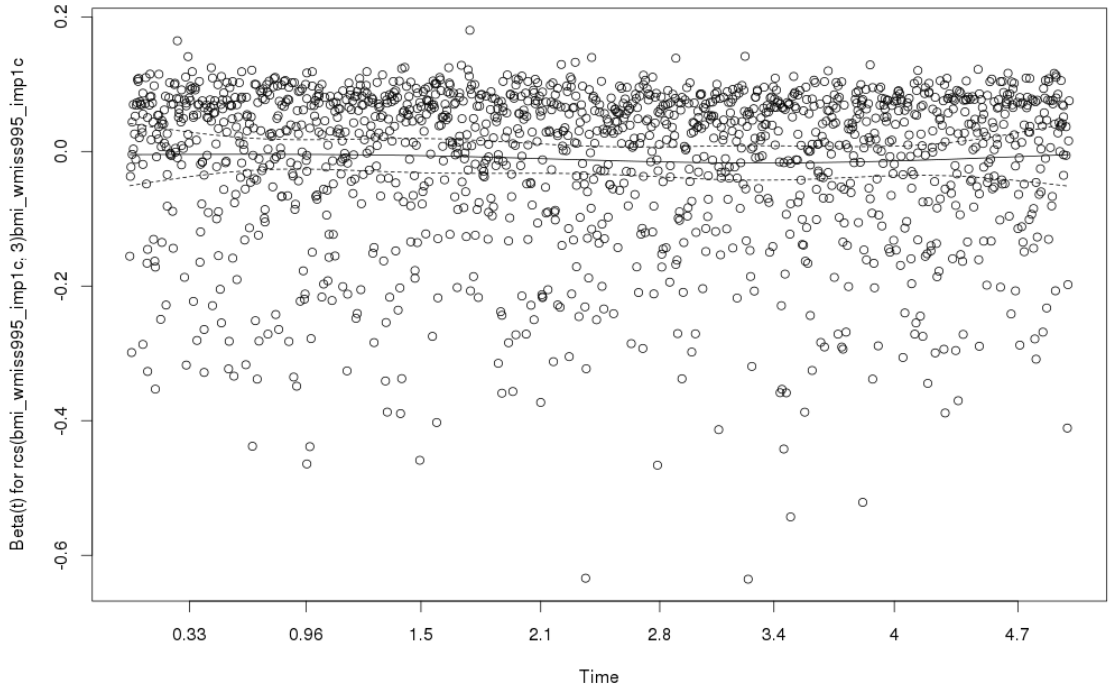
Schoenfeld Residuals Based on Smoker Category:



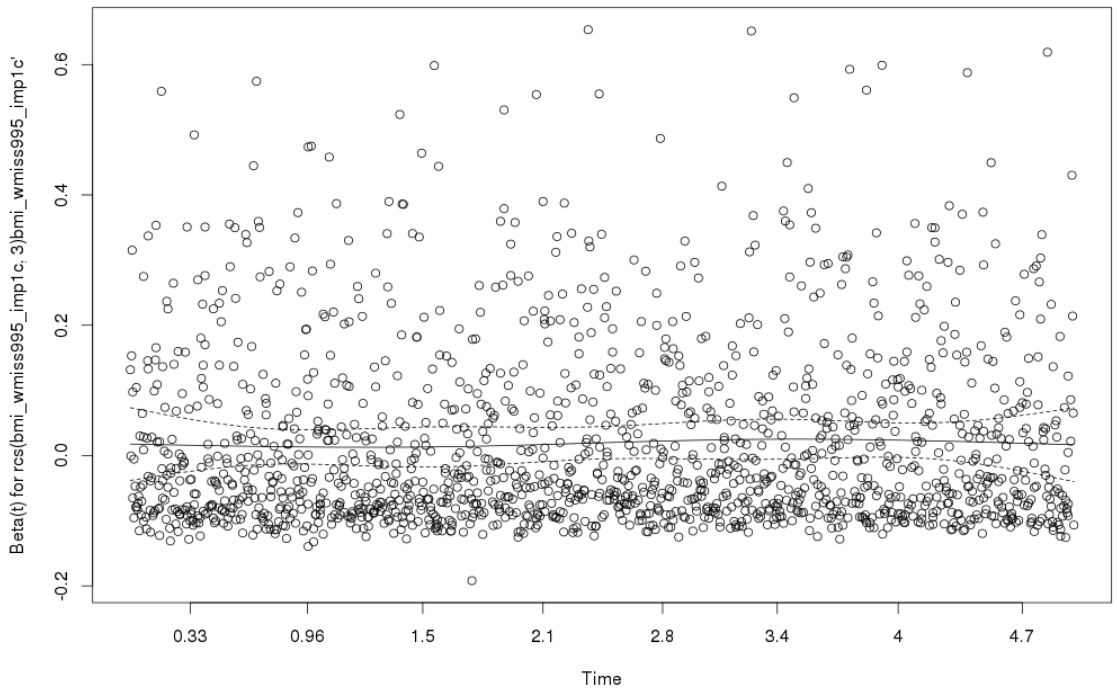
Schoenfeld Residuals Based on Percentage of Life Lived in Canada:



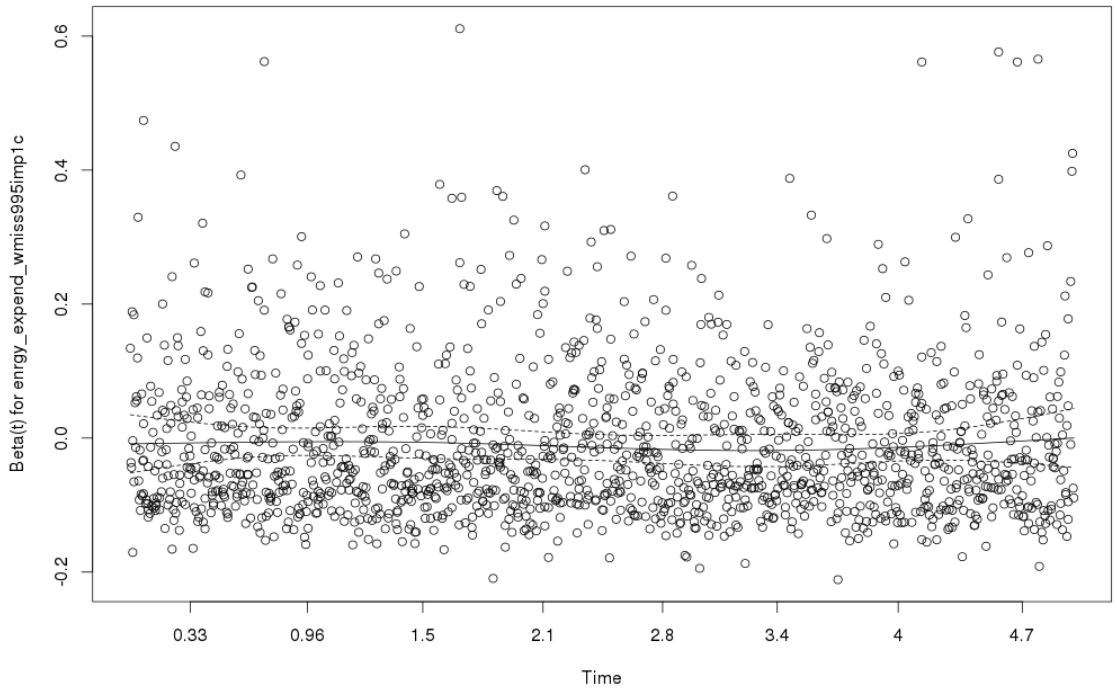
Schoenfeld Residuals Based on BMI- 1ST Spline:



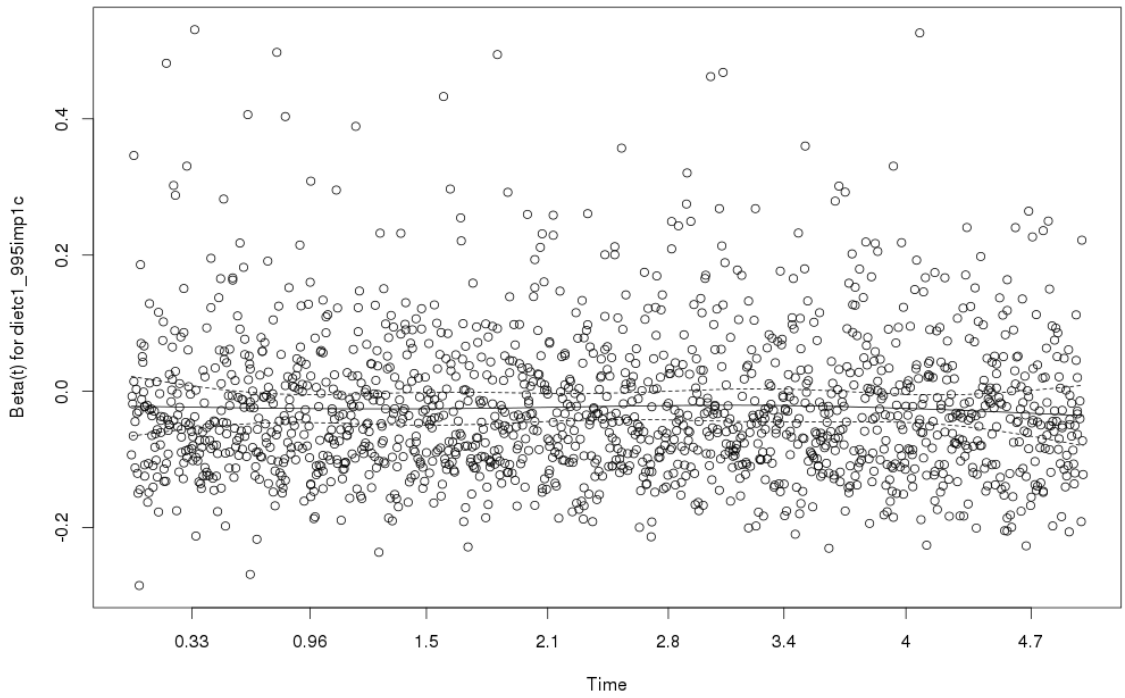
Schoenfeld Residuals Based on BMI- 2nd Spline:



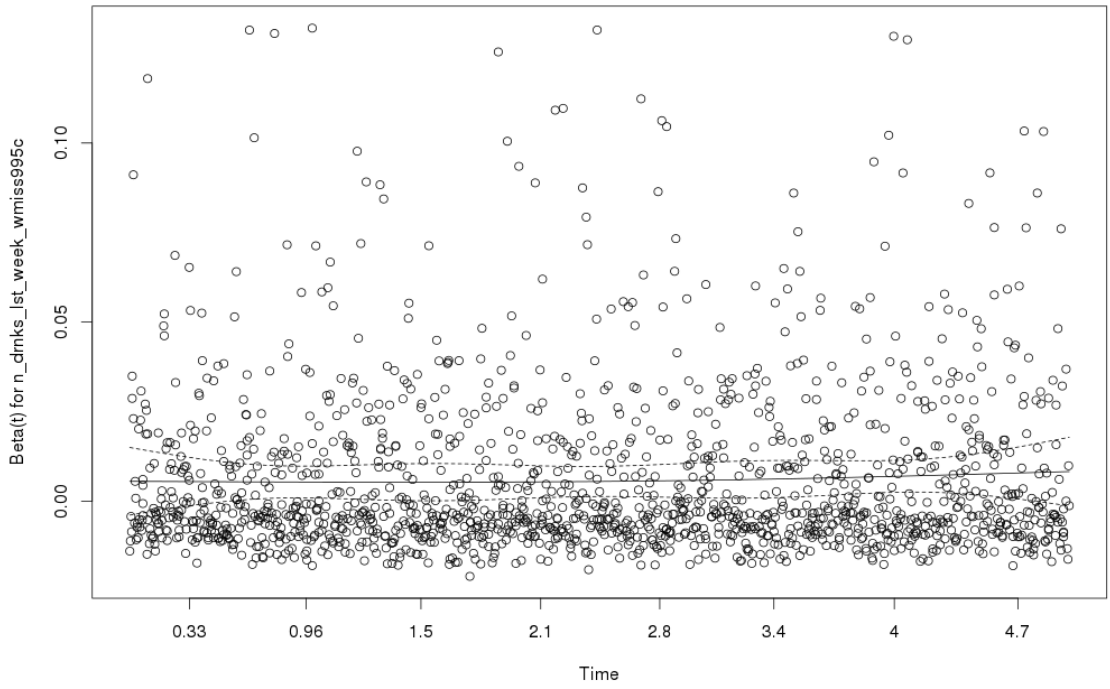
Schoenfeld Residuals Based on Leisure Physical Activity Level:



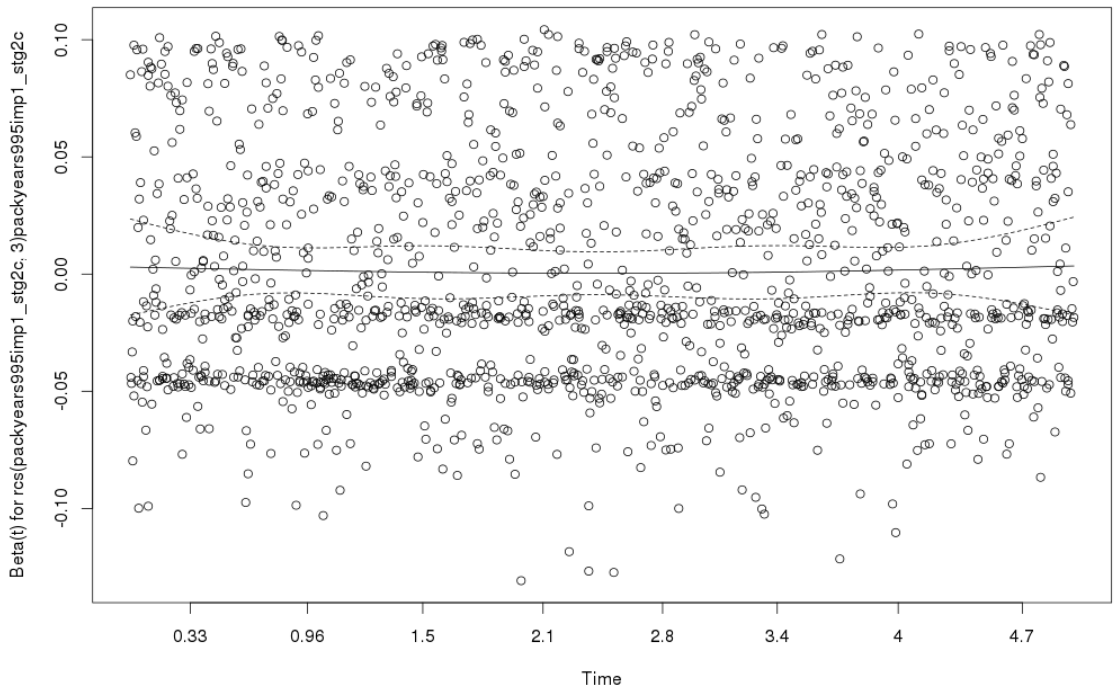
Schoenfeld Residuals Based on Dietary Consumption of Fruits and Vegetables:



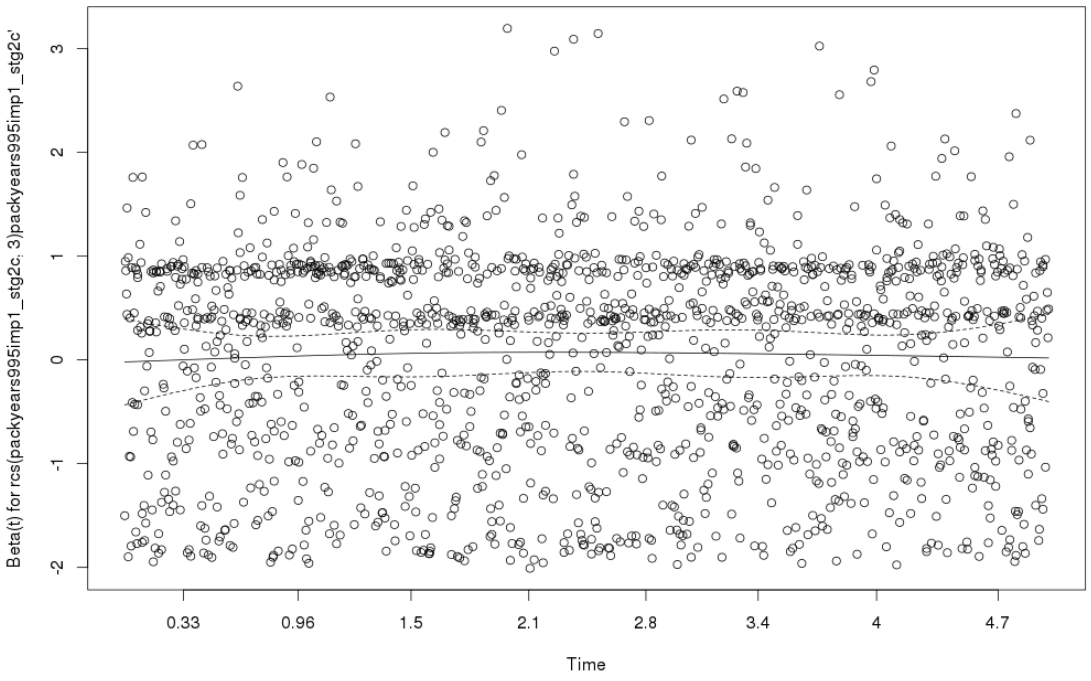
Schoenfeld Residuals Based on Alcohol Consumption:



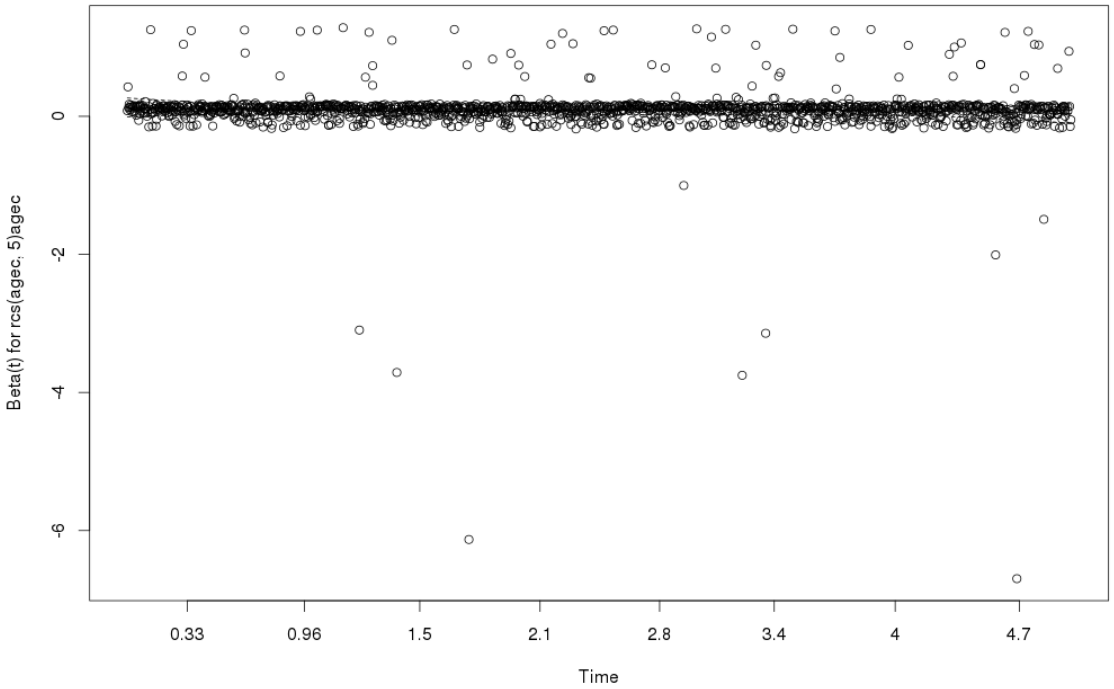
Schoenfeld Residuals Based on Pack-Years of Smoking- 1st Spline:



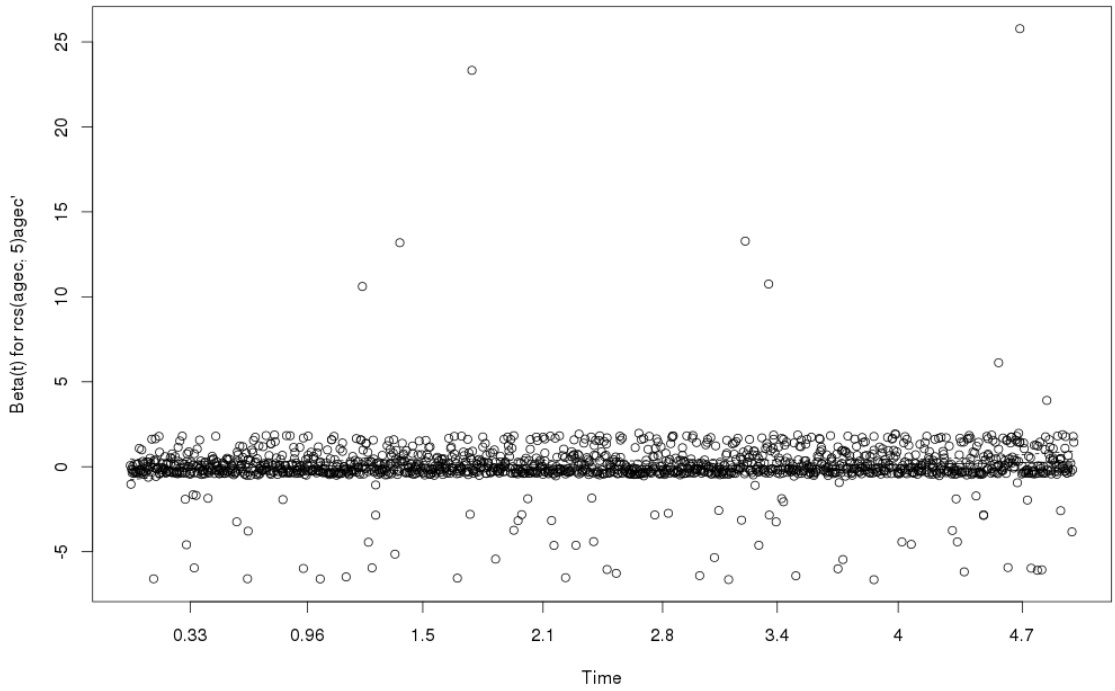
Schoenfeld Residuals Based on Pack-Years of Smoking- 2nd Spline:



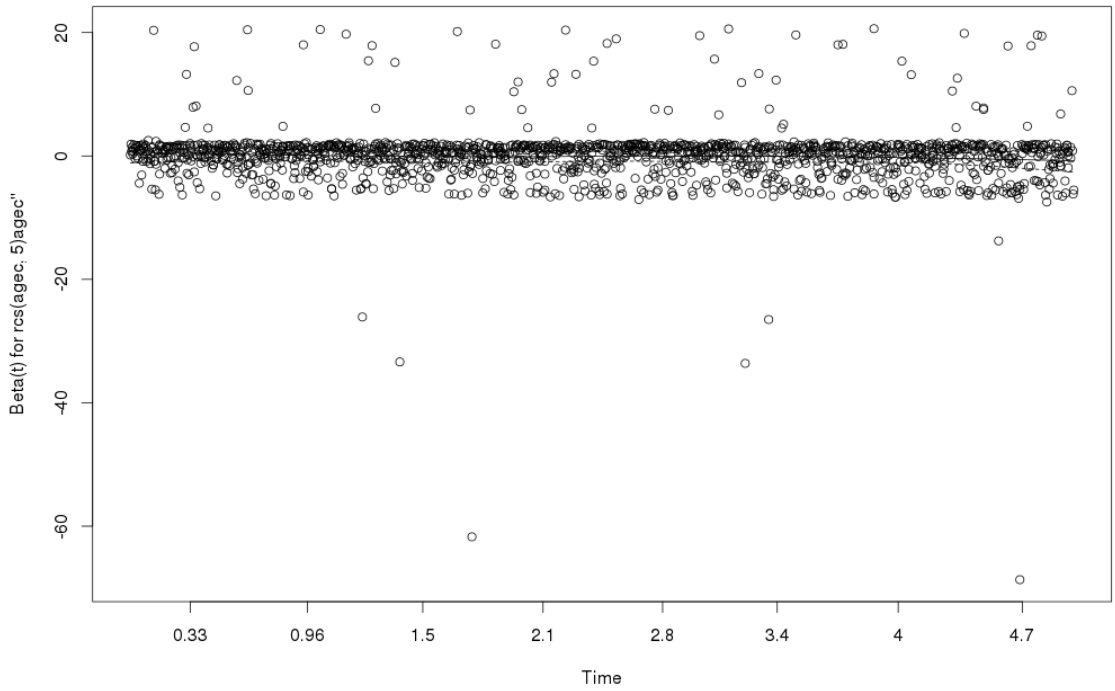
Schoenfeld Residuals Based on Age- 1st Spline:



Schoenfeld Residuals Based on Age- 2nd Spline:



Schoenfeld Residuals Based on Age- 3rd Spline:



Schoenfeld Residuals Based on Age- 4th Spline: