An algorithm for augmenting cancer registry data for epidemiological research applied to oesophageal cancers

Paul Patrick Fahey

Translational Health Research Institute

Western Sydney University

A thesis submitted for the degree of

Doctor of Philosophy

Dedications

To Josephine and Leslie Fahey who always hoped this would happen.

Acknowledgements

I wish to express my appreciation to my supervisors Prof Andrew Page, Prof Glenn Stone, Prof David Whiteman, and Prof Thomas Astell-Burt for their long-term support.

I am particularly grateful for wife Dr Kylie-Ann Mallitt who provided constant and loving encouragement to keep me moving forward.

I also wish to thank Prof David Zucker for providing his software for corrected score misclassification corrections in Cox regression.

Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

(Paul Fahey)

Declarations

I declare no conflict of interest.

I declare no funding other than Candidature Support Funds.

(Paul Fahey)

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Full Term
BRFSS	Behavioural Risk Factor Surveillance System
BMI	Body Mass Index
CI	Confidence interval
I2C2	The impute, impute, calibrate, correct algorithm investigated in this thesis
OC	Oesophageal cancer
OAC	Oesophageal adenocarcinoma
OSCC	Oesophageal squamous cell carcinoma
HR	Hazard ratio
SEER	Surveillance, Epidemiology, and End Results Program
RR	Relative risk

Thesis Abstract

Background

Oesophageal cancer is an important cancer with short survival, but the relationship between pre-diagnosis health behaviour and post-diagnosis survival remains poorly understood. Cancer registries can provide a high quality census of cancer cases but do not record pre-diagnosis exposures. The aim of this thesis is to document relationships between pre-diagnosis health behaviours on post-diagnosis survival times in oesophageal cancer, developing new methods as required.

Methods

A systematic review and meta-analysis conducted in 2014, and updated in 2021, to investigate the association between pre-diagnosis health behaviours and oesophageal cancer.

Visualising health behaviour variables as part of the cancer registry data set, with 100% missing data, led to the development of new approaches for augmenting US oesophageal cancer registry data with health behaviour data from a US national health survey Firstly, the health survey data were used to create logistic regression models of the probability of each behaviour relative to demographic characteristics and then these models were applied to cancer cases to estimate their probability of each behaviour. Secondly, cold-deck imputation such that two randomly selected but demographically similar health survey respondents both donated their health behaviour to the matching cancer case. The agreement between these two imputed values was used as an estimate of the misclassification and corrected for during the analyses.

Results

The logistic regression imputation-based analyses returned accurate point estimates, with wide confidence intervals, if the behaviour occurred in more than approximately 5% of cases.

Our reviews and analyses confirmed that pre-diagnosis smoking decreased survival in oesophageal cancer (hazard ratio (HR) 1.08, 95% confidence interval (CI) 1.00-1.17) particularly squamous cell carcinoma when comparing highest to lowest lifetime exposure (and HR 1.55, 95%CI 1.25-1.94); with similar associations for alcohol consumption. Pre-diagnosis leisure time physical activity was found to be associated with reduced hazard (HR 0.25, 95%CI 0.03,0.81) overall.

Discussion

Findings from these analyses can assist in modelling the impact of current changes in community health behaviour, as well as informing prognosis and treatment decisions at the individual level. This novel method of augmenting cancer registry data with pre-diagnosis variables appears to be effective and will benefit from further validation. This thesis has significantly progressed both issues and identified future opportunities for research and development.

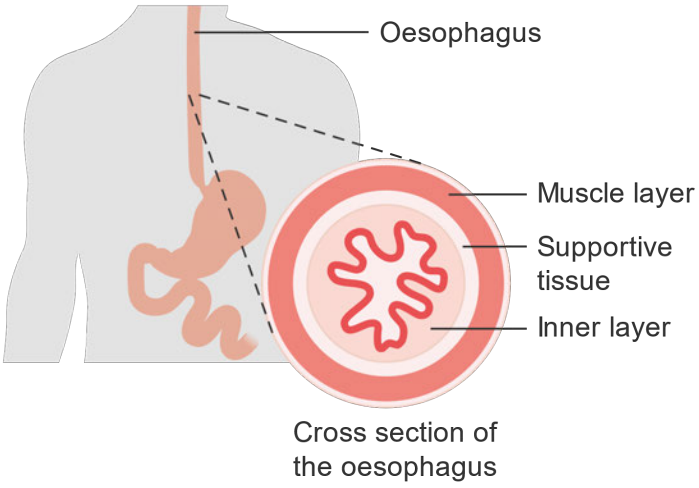
Chapter 1 Introduction

1.1 Oesophageal cancer

Oesophageal cancer (OC) is an important cancer. In 2018, it was estimated there were 572,034 new cases worldwide (or 3.2% of all new cancers) and 508,585 deaths (or 5.3% of all cancer deaths) (Bray et al., 2018). It has been estimated that nearly 4 million (disability adjusted) life years were lost to OC in 2008 (Di Pardo et al., 2016). However, as with most diseases, OC is still relatively rare with an estimated cumulative risk of diagnosis from birth to 75 years of 1.15% of males and 0.43% of females (Bray et al., 2018).

Figure 1.1 shows the location of the oesophagus and a cross-section through it. OC occurs when malignant cells develop in the innermost layer (mucosa) of the oesophagus. From there cancer may spread outwards to involve the submucosa, muscular layer and eventually structures outside the oesophagus (Thrumurthy, Chaudry, Thrumurthy, & Mughal, 2019).

Figure 1.1 Cross-section of the layers of the oesophagus



The two main histological subtypes of OC are oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC) which together account for more than 90% of OC (Enzinger & Mayer, 2003). Squamous cells are the flat, thin cells that line the surface of the oesophagus. Squamous cell carcinomas are more frequent in the upper and middle portions of the oesophagus. Adenocarcinoma begin in the cells of the mucus excreting glands in the oesophagus and is more frequent in the lower portion of the oesophagus. Other less common OC include small cell carcinomas, sarcoma, lymphoma, melanoma and choriocarcinoma (Thrumurthy et al., 2019).

OC incidence varies considerably between genders and across geographic regions ranging from a high of 17.9 per 100,000 in males and 6.8 per 100,000 in females in Eastern Asia to 1.6 and 0.8 per 100,000 in males and females in Western Africa (Bray et al., 2018). Nearly 80% of cases occur in low and middle income countries (Abbas & Krasna, 2017). OSCC is the most prevalent subtype accounting for 87% of OC worldwide but OAC is more prevalent in high income countries (Abbas & Krasna, 2017). This suggests a very different distribution of risk factors across the world plus a difference in risks by histological subtype.

1.2 Risk factors

OC is more common in older age groups. In the UK for example, only 26% of all OC cases are less than 65 years of age at diagnosis, 48% are between 65 and 74 years of age and 26% are 75 or more years of age (Cancer Research UK).

OSCC is thought to be associated with chronic irritation and inflation of the oesophageal mucosa (Enzinger & Mayer, 2003). Common risk factors include achalasia (leading to retention of food within the oesophagus) (approximately 10-fold risk), smoking (9 fold risk), heavy alcohol use (≥3 drinks per day has a 3 to 5 fold risk), black race (3 fold risk)

and insufficient fruit and vegetables (Clara Castro, Peleteiro, & Lunet, 2018; Short, Burgers, & Fry, 2017).

OAC is thought to be associated with gastro-oesophageal acid reflux. Common risk factors include male sex (8 fold risk), white race (5 fold risk), gastroesophageal reflux disease (5 to 7 fold risk), obesity (2.4 fold risk) and smoking (2 fold risk) (Clara Castro et al., 2018; Short et al., 2017) .

Physical activity has also been shown to be protective of most cancers (Friedenreich, Neilson, & Lynch, 2010). Two meta analyses from 2014 both found moderate protective effects of physical activity on OC: RR=0.71, 95% CI 0.57,0.89, I^2 47% (Singh, Devanna, Varayil, Murad, & Iyer, 2014); RR=0.73, 95% CI 0.56,0.97, I^2 58% (Y. Chen, Yu, & Li, 2014).

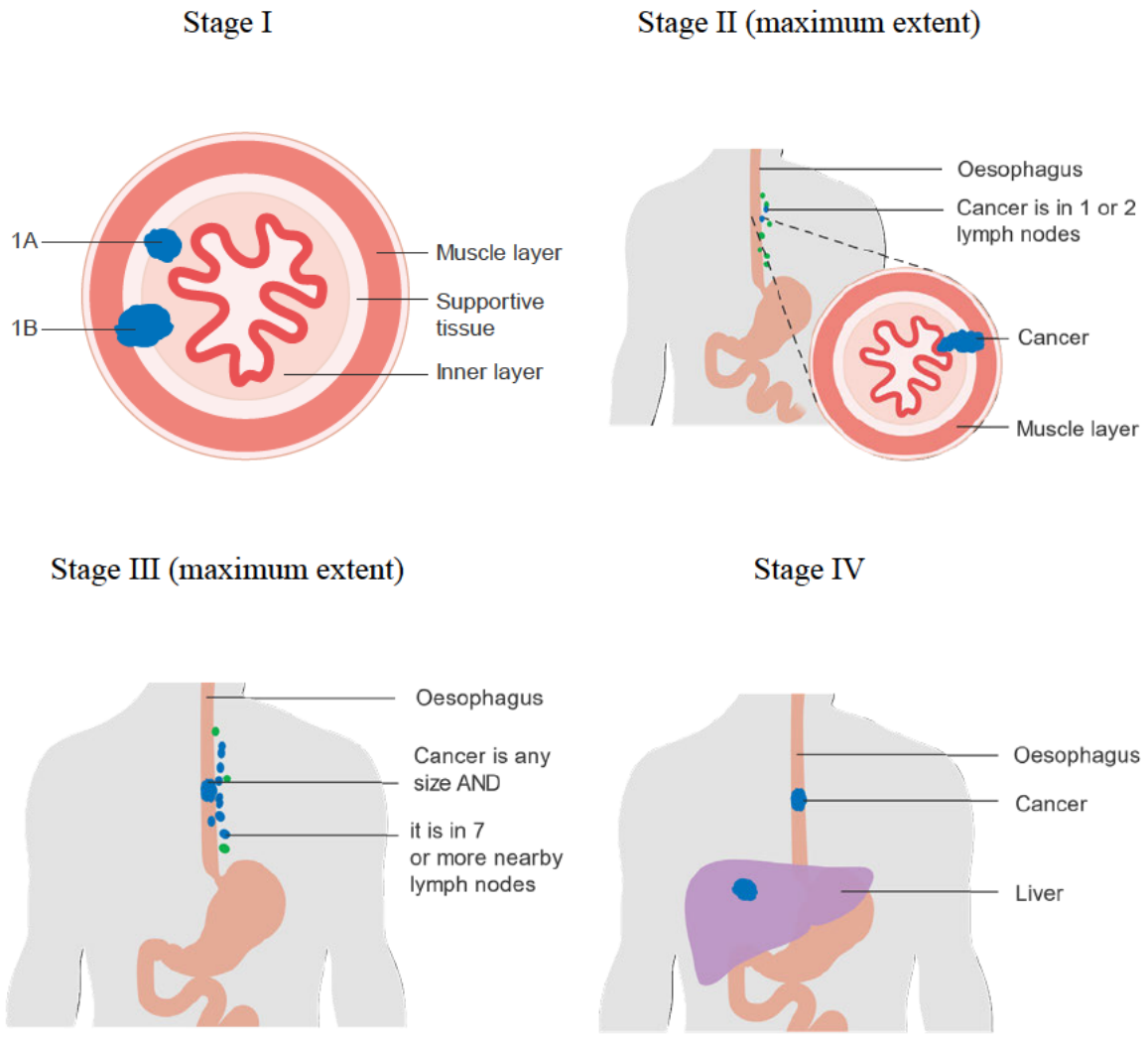
Several of the main risk factors for OC are behavioural and modifiable (Clara Castro et al., 2018), including tobacco smoking, alcohol consumption, obesity and physical activity. Indeed, in the US in 2014 it was estimated that 74.7% (95% CI 72.3%, 77.1%) of male and 67.5% (95% CI 63.2%, 72.0%) of female oesophageal cancer diagnoses arose from modifiable lifestyle factors and that cigarette smoking, alcohol consumption and excess body weight could account for up to 50.0% (95% CI 48.5%, 51.2%), 21.0% (95% CI 18.5%, 23.3%) and 32.2% (95% CI 29.6%, 34.7%) of cases respectively (Islami et al., 2018).

1.3 Cancer stage at diagnosis

Cancer is detected at different stages of growth in different individuals. Cancer staging is used to describe how much cancer is in the body and where it is located. The cancer stage at diagnosis has a large impact on prognosis and treatment for that individual.

The widely used American Joint Committee on Cancer (AJCC) staging system starts with separate codes for the extent of the tumour (the T code), whether or not the cancer has reached nearby lymph nodes (the N code) and whether or not there are distance metastasis (the M code). Once the TNM codes are known, they can be combined into a single overall stage with categories denoted 0 (mucosa layer only, no spread to nearby tissues), I (least spread), II, III and IV (advanced or metastatic cancer) (American Joint Committee on Cancer). The stage 0, or in situ cancers, are generally excluded from administrative and research data bases. As shown in Figure 1.2, stage I cancers are contained within the oesophagus, stage II cancers are also contained within the oesophagus except for the involvement of just 1 or 2 lymph nodes, stage III cancers could have through the muscle wall of the oesophagus and/or spread the lymph nodes and stage IV cancer have spread to distant organs.

Figure 1.2 Oesophageal cancer disease stages defined by the American Joint Committee on Cancer



Attribution: Cancer Research UK uploader, CC BY-SA 4.0, via Wikimedia Commons

For example, in England in 2013-2017, there were 37,169 persons diagnosed with OC with 3,651 in Stage I (9.9%), 4,705 (12.7%) at Stage II, 10,952 (29.5%) at Stage III and 11,093 (29.8%) at Stage IV with the remaining either unknown/missing (n=6,676, 18.0%) or unstageable (n=92, 0.2%) (Office for National Statistics).

Alternatively, in the US, the widely used SEER (Surveillance, Epidemiology and End Results) cancer registry codes stage at diagnosis into three categories: localised (confined to

primary site), regional (spread to regional lymph nodes) or distant (cancer has metastasised) (National Cancer Institute, 2021c). In 1998-2009, 22% of OC was classified as localised at diagnosis, 30% regional and 35% distant with the remaining 13% unknown/missing (Zhang, 2013).

As the oesophagus is elastic, tumours tend to be quite large and advanced before they begin to produce symptoms or blockages. Hence, most tumours are advanced (stage III and IV or metastasised) at diagnosis, contributing to the relatively poor prognosis of OC.

1.4 Prognosis

One of the more important characteristics of OC is its relatively short survival times. The measurement of survival time differs according to the choice of start time, choice of end time and any quality of life discounting applied.

One possible start time is the first onset of cancer within the body, but this is as yet unknown and unknowable. Survival is often measured from the start of treatment, particularly the date of first surgery. But the literature on post-surgical survival times is omitted because not all OC patients are treated with surgery (for example, 52.3% in South Korea from 2005 to 2017 (Jung et al., 2020) and 27.1% of patients over 65 years in the US from 1998 to 2013 (Tramontano et al., 2019)) and the clinical decision of who to accept for surgery has the potential to alter post-surgery survival times. Survival from diagnosis is used in this thesis because all cancer patients have a date of diagnosis recorded and diagnosis is the day everything changes for the patients, their families, and the treating clinicians.

Common end dates include all cause death and disease specific death. The relatively short survival time of OC means people don’t have a lot of time to die from other causes. The

majority of OC patients die from OC (for example, an estimated 79.5% in Sweden in 1961 to 2014 (Xie, Wahlin, & Lagergren, 2017)). All cause death is easier to measure and more commonly reported. Most statistics reported below are for survival to death from all causes, with exceptions noted.

While quality adjusted life years have been used to compare treatment outcomes (Pennathur, Zhang, Chen, & Luketich, 2010) they are not routinely available in official statistics of OC survival. The relatively short survival times also seem to decrease concern over the quality of life during survival. It has been estimated that 96.8% of the loss of quality of life caused by OC is lost years of life with just 3.2% of the total loss ascribed to disability (Di Pardo et al., 2016).

Finally, survival times are also more difficult to obtain from developing countries where most OC cases arise. The examples below are mainly derived from developed countries.

OC survival time varies strongly with age and disease stage at diagnosis. For example, in England between 2013 and 2017 the estimated 5-year net survival rates were 52.8%, 29.9%, 16.3% and 0% for those diagnosed at disease stages I, II, III and IV respectively. Among those who were diagnosed with Stage I cancer the estimated 5 year net survival rates were 66.8% (95% CI, 57.5%,76.1%) for those who were 15-54 years at diagnosis, and 36.5% (95% CI 31.8%,41.3%) for those who were 75 or more years of age at diagnosis respectively (Office for National Statistics).

In the US in 2010 to 2016, it was estimated that 47% of patients with localised disease at diagnosis survived for at least 5 years. This decreased to 25% for those with regional spread and just 5% for those with distant spread (American Cancer Society, 2021b)..

Overall, OC seemed to have missed the large improvements in survival times over the past 30 years which have been observed for other cancers such as colon, rectal and breast cancer (Allemani et al., 2015). In the US for example, median survival time for localised OC increased from 11 months to 35 months between 1970 and 2000. But the corresponding increases in survival times for regional disease and metastatic disease were 10 to 15 months and 4 to 6 months respectively (Dubecz et al., 2012). While a 50% increase in median survival months over a 30-year period is technically important, from the individual's perspective the extra 5 or 2 months survival for regional or metastatic disease may be somewhat disappointing.

1.5 Predictors of survival time

The most important predictors of OC post-diagnosis survival times are the cancer stage at diagnosis and patient age at diagnosis.

Obviously receiving the right treatment is also important for survival, and this may not always happen. There are disparities between high income and low and middle income countries. Even within high income countries there have been reports of disparities between, for example, race and ethnicity (Dong, Gu, El-Serag, & Thrift, 2019; Tramontano et al., 2018) and between education levels, marital status and place of residence (Ljung, Drefahl, Andersson, & Lagergren, 2013).

Somewhat surprisingly, there is relatively little information about the impact of pre-diagnosis health behaviours on post-diagnosis survival times in OC (see systematic review of this evidence in Chapter 2). Many of the contributing factors to OC are health behaviours (smoking, alcohol consumption and obesity) and it is not unrealistic to expect that these behaviours will have some carry-over effect on prognosis. Indeed, the relatively short

survival times for OC, give less time for any such carry-over effect to dissipate. It is proposed that that health behaviour prior to diagnosis could be an important determinant of post-diagnosis survival time.

It is important to differentiate between pre-diagnosis health behaviours and post-diagnosis behaviours. A diagnosis of OC is a life changing event and life changing events often lead to changes in behaviour (Demark-Wahnefried, Aziz, Rowland, & Pinto, 2005). Also, symptoms are usually appearing and worsening around the time of diagnosis, and health behaviour is likely to change in response to such symptoms. Finally, from an intervention point of view the process of changing behaviour prior to diagnosis is a health promotion exercise, but interventions to change behaviour after diagnosis are more likely to be treatments prescribed by the treating clinicians.

1.6 Investigating pre-diagnosis factors

Investigation of the relationship between pre-diagnosis health behaviour and post-diagnosis survival time in OC requires longitudinal data: the health behaviour occurs prior to the diagnosis which occurs prior to death. Complicating issues are that the disease is relatively rare with just 1.15% of males and 0.43% of females ever diagnosed with OC before age 75 years (Bray et al., 2018) and some health behaviours are also relatively rare (for example, in the US an estimated 8.3% of men and 4.5% of women aged 18 and older engaged in ‘heavy’ alcohol use in the past month (National Institute on Alcohol Abuse and Alcoholism)).

One possible approach would be a prospective cohort study. Health behaviours could be measured when participants are enrolled and then participants could be monitored over time to see if they are ever diagnosed with OC and then monitored further to record how long

they survived. But with a rare disease like OC, and sometimes uncommon pre-diagnosis health behaviours, the cohort would need to be very large and/or followed up for a long period to generate sufficient data for analysis. At least one such prospective study is currently underway. In China, The National Cohort of Esophageal Cancer is currently recruiting 100,000 40-69 year old men and women in regions with high rates of OC (R. Chen et al., 2019). The challenges of the prospective cohort include the costs, the time delay of waiting for cases to occur and the potential for bias arising from participant loss to follow-up over time.

Costs could be reduced by incorporating data items about OC into existing large cohort studies measuring health behaviour and date of death. Of course, as data collection is not specific to OC, the inclusion and exclusion criteria will not be optimised for OC and even larger cohorts will be required to return an adequate sample size of OC. Further the choice of health behaviour measures will not be specific, to or optimised for, OC.

Given the costs, time delays and loss to follow-up bias arising from prospective cohort studies, would retrospective data collection be more effective? The Australian Cancer Study Clinical Follow-up Study targeted all patients 18-79 years of age with OC or cancer of the gastro-oesophageal junction on mainland Australia in 2002 to 2005. Of 3,273 potentially eligible patients 1,007 (30.8%) had died before they could be invited to participate, 134 (4.1%) were too ill, etc. Eventually baseline questionnaire data was obtained from 1,056 (32.3%) of those thought potentially eligible. The questionnaire asked participants to recall and report their health behaviour one year prior to diagnosis. As illustrated in this example, most retrospective data collection is subject to survivor bias and recall bias.

1.7 Cancer registry data

When initiating a study in cancer epidemiology, an obvious starting place is the cancer registry. A cancer registry is an information system for the collection, storage and management of data from people with cancer (National Cancer Institute Surveillance Epidemiology and End Results Program). In many countries it is mandatory for diagnosing clinicians to report cancer cases to a centralised cancer registry (for example, in NSW Australia, mandatory reporting is supported by the Public Health Act 1991 and the Public Health General Regulation 2002 (NSW Health, 2009)). There is usually regulatory clearance for using these data for research purposes without having to seek informed consent from individual cancer cases.

However, cancer registries can be challenging to establish (Jedy-Agba et al., 2015) and expensive to maintain (Tangka et al., 2016) and as a consequence the scope and quality of available data can vary considerably between countries (Forsea, 2016). Based on 2006 data it was estimated cancer registry coverage varied from 99% of the population in North America to 8% of the population in Asia (excluding Japan) (Valsecchi & Steliarova-Foucher, 2008).

In high income countries, cancer registries provide high quality, census data for cancers and often record patient outcomes such as survival. The typical data items collated within a cancer registry (National Cancer Institute Surveillance Epidemiology and End Results Program) include:

- patient demographics such as patient's name, age, gender, race, ethnicity, and birthplace;
- tumour characteristics such as tumour cell type(s), body organ where the cancer started, and, increasingly, genomic information on the tumour;

- stage of disease; and
- outcomes such as survival status, cause of death, and survival time.

Pre-diagnosis exposures aren't routinely collected, and so cancer registry data often needs to be augmented with additional data to investigate the role of pre-diagnosis exposures.

Additional data can be added through supplementary data collections. This may involve drawing a random sample of patients from cancer registry lists and then contacting those people requesting they participate in further data collection. For example, a prospective study of 2,383 women with breast cancer obtained from the US Surveillance, Epidemiology, and End Result (SEER) cancer registries achieved a response rate of 78.4% with the median time between cancer diagnosis and survey response being 7.9 months (Lantz et al., 2005).

However, this approach may be less effective for studying diseases with short survival times or for studying pre-diagnosis behaviours. The processing time for registering new cancer cases onto the cancer registry and then contacting these people, could lead to considerable survivor bias. The need to obtain patient's informed consent creates the potential for non-participation bias and, in the case of pre-diagnosis behaviours, the retrospective nature of the data collection creates potential for recall bias.

Data linkage provides an alternate approach. Data (or record) linkage involves identifying, matching, and merging data records from the same individual from different data sources (Antoni & Sakshaug, 2020). With the proliferation of health-related data sets, it is possible that a cancer patient's pre-diagnosis behaviour may already be on record in some other data collection. If that record exists and can be identified, it could be merged with that patient's cancer registry data to provide all required data to establish a retrospective cohort study.

In South Korea, for example, the National Health Insurance Corporation, regularly surveys the health behaviours of their clients. Of 901,979 men surveyed in this program in 1996, 14,996 were later identified in the Korean Central Cancer registry of 1996 to 2004 and 7,271 of these were later identified in mortality data from the National Statistical Office to 2004 (S. M. Park, Lim, Shin, & Yun, 2006). Merging these data records led to the creation of a data set containing pre-diagnosis health behaviour, cancer characteristics at diagnosis and post-diagnosis survival status and survival time for these 14,996 patients.

In most developed countries an important source of health behaviour data are national health surveys. In Australia for example, the National Health Survey (Australian Bureau of Statistics) collects data on the health behaviours of a representative national sample of Australian adults each year. In the US there are the National Health and Nutrition Examination Survey (National Center for Health Statistics), the National Health Interview Survey (National Center for Health Statistics) and the Behavioural Risk Factor Surveillance System (Centers for Disease Control and Prevention, 2013a).

A major advantage of data linkage is that there is no primary data collection, avoiding associated costs and time delays. Also, as the cancer registries and national health surveys have large infrastructure and specialist expertise, response rates and data quality are usually maximised. But one significant difficulty with record linkage is that it increases the risk of breaches in patient confidentiality and the consequences of any such breach. Efforts to protect confidentiality add considerable cost and complexity to the linkage process (Harron et al., 2017). Also, missed matches and false matches occur (Harron et al., 2017). The matching error rate will depend on the accuracy, completeness and informativeness of the matching variables, and the strengths and weaknesses of the matching algorithm and can be difficult to quantify.

For rare diseases, sample sizes may also be insufficient and matching errors may dominate. While the cancer registry is generally a census of all cancer patients, most national health surveys only enrol a sample from the population. The Australian National Health Survey for example, records data from about 1 in every 1000 adult Australians (Australian Bureau of Statistics), which crudely translates to having the health behaviours of just 1 in 1,000 oesophageal cases being measured each year. (In 2020, the estimated number of new cases of oesophageal cancer in Australia was 1,587 (Cancer Australia)). Aside from insufficient sample size, this '1 in 1000' may be well below the linkage error rate for these data sets, allowing linkage errors to dominate in resulting files.

So even though data linkage is a widely used, cost effective approach, there are a range of research questions which cannot be addressed in this way; particularly those involving relatively rare diseases.

1.8 Is a different method possible?

If linking an individual cancer case to their own pre-diagnosis behaviour is not feasible, what would happen if a 'similar' person was substituted instead?

It is known that particular health behaviours are more common in some demographic groups. For example, alcohol consumption in the US differs by age, sex and race and this can be further stratified into smaller demographic subgroups. For example, younger people are more likely to drink, but young college students are more likely to drink than other young people (Delker, Brown, & Hasin, 2016). Also, people of Asian descent are less likely to drink, but Koreans, Japanese, Taiwanese and Chinese are more likely to drink than other Asian subgroups (Delker et al., 2016). In Italy, not only does smoking prevalence differ by

age, sex and education, but rates also differ between residents of northern Italy and those of southern Italy (Tramacere et al., 2009).

Given the health behaviours of people from the same demographic background tend to be more similar than the behaviour of people with different demographic backgrounds, matching a cancer case to the earlier behaviour of a demographically similar person may convey some information about the cancer case's own pre-diagnosis health behaviour. Obviously, individuals' behaviour varies considerably even within the same demographic cluster. Whether an individual lives in the north or south of Italy is an extremely poor proxy measure of their smoking status, although it does convey some small quantum of information. Any information retained through substituting the behaviour of a demographically similar person would be very small and so large data sets would be required to detect potentially weak signals.

Large, computerised data sets are routinely available to test this idea. Data sets grow over time and currently, when combined across collection years, both cancer registries and national health surveys provide very large sample sizes. For example, in the US an estimated 19,260 new OC cases will be added to cancer registries in 2021 (National Cancer Institute). Combining just 3 years data could produce a sample size of more than 50,000 for analysis.

But other than sample size and statistical power, the other factor which will determine the feasibility of detecting a signal of pre-diagnosis behaviour based on specific demographic strata is whether that signal exists and how strong it is. The greater the variability in behaviour across demographic groups, the more information a patient's demographic group conveys about their likely pre-diagnosis health behaviour. That is, the more informative the demographic variables used to describe the cancer patient, the more information which may be retained about their pre-diagnosis health behaviour.

Fortunately, most human data sets contain a range of demographic descriptors of their participants (age, sex, education, place of residence, etc) the combinations of which, allow many thousands of demographic sub-groups to be defined.

1.9 Research aim

This study will focus on those modifiable health behaviours which are associated with the onset of oesophageal cancer. The first research aim is to describe the carry-over effect, if any, of these pre-diagnosis health behaviours on post-diagnosis survival times in patients diagnosed with malignant OC.

The second research aim is to develop, describe and evaluate a novel algorithm for addressing the first aim: using existing data only, when one-to-one data linkage is not feasible.

1.10 Importance of this research

This research, if successful, could produce a range of clinical and methodological benefits. Clinical benefits include better informed health service planning, and health promotion activities as well as better information for individual patients and clinicians for decision making.

The health behaviours of a community can change quite rapidly. For example, in Australia the proportion people aged 14 and over who were daily smokers was estimated to be 11% in 2019. This rate has more than halved in less than 30 years (from 24% in 1991) (Australian Institute of Health and Welfare, 2020b). Also, between 2001 and 2019 the proportion of 18-24 year old Australian abstaining from alcohol increased from 10% to 21%

but the proportion of those aged 70 or more abstaining from alcohol decreased from 32% to 28% (Australian Institute of Health and Welfare, 2020b). The prevalence of discretionary physical activity among Canadian adults rose from 20.6% to 41.1% between 1981 and 2000 (Craig, Russell, Cameron, & Bauman, 2004).

Such changes will likely lead to future changes in OC incidence and survival. An improved understanding of the association between community health behaviour and OC incidence and survival can anticipate the future health needs of this group. Importantly, predictions of the number of new cases alone is insufficient for planning. Treatment stage, a function of survival time, is a major determinant of health care costs (Tramontano et al., 2019).

Many of these changes in health behaviour are being driven or encouraged by public health programs. Being able to quantify the association between health behaviours and cancer outcomes, will help to inform the implementation and evaluation of public health programs in the community. That is, this research could lead to improvements in the targeting and evaluation of public health policy.

Documenting the association between pre-diagnosis health behaviour and OC survival also provides information which could be included in the content of public health programs to provide individuals with better understanding of possible consequences of their health behaviours.

At the level of the individual patient and individual clinician, quantifying the link between pre-diagnosis behaviour and survival time may allow needed improvements in prognostic tools (Gupta et al., 2018) and a more accurate understanding of disease prognosis will aid in treatment decision making and patients' personal future planning.

Methodologically a new algorithm for augmenting cancer registry data sets with information on pre-diagnosis behaviour, could also allow a greater range of research questions to be addressed with existing cancer registry data sources and health survey data sets, circumventing logistical issues associated with data linkage and patient confidentiality.

In 2011 it was estimated that the cost of maintaining the US National Program of Cancer Registries was $US60.77 per case (Tangka et al., 2016). But the estimated number of new US cancer cases has increased from 1,596,670 in 2011 to 1,898,160 in 2021 (American Cancer Society) an increase of 19% in 11 years. The increasing costs of collecting and maintain data systems provide strong incentives to continually find new benefits, including new research applications, from these existing data sets.

If successful, the new algorithm could potentially replace the need for a much more expensive and time-consuming prospective cohort study, freeing up research resources for other projects. Even if only partially successful, the algorithm could be used as a pilot investigation to confirm the need for the more accurate and expensive cohort study.

Finally, as the new algorithm does not seek to find any individual's true health survey data record, it does not increase privacy issues (Price & Cohen, 2019) beyond the original cancer registry record. That is, the data added to the cancer registry record contains almost no additional information about that individual.

1.12 Thesis Outline

This thesis contains 7 chapters.

Chapter 2 summarises current knowledge of the relationship between pre-diagnosis health behaviour and post-diagnosis survival time for oesophageal cancer using systematic review and meta-analysis.

Chapter 3 describes and critiques the availability and suitability of data sets and describes the data which were selected for use in this thesis. It also considers how missing variables could added.

Chapters 4 to 6 document approaches to develop a novel algorithm to research the relationship between pre-diagnosis behaviour and post-diagnosis survival times using only existing data when data linkage is not feasible.

Chapter 4 discusses using logistic models on national health survey data to summarise individual's probability of displaying particular health behaviour based on their demographic characteristics, and then applies this model to the cancer registry cases to estimate their probability of having had that behaviour pre-diagnosis.

Chapter 5 introduces an imputation-based approach to assign pre-diagnosis behaviours to cancer registry cases, estimate the level of misclassification in the imputed values, and adjust for this misclassification during the analyses. In this Chapter, the solution is developed for estimating the relative risk of pre-diagnosis behaviour on 12 months survival post-diagnosis.

Chapter 6 formalises the methods in Chapter 5 as the I2C2 (Impute, Impute, Calibrate, Correct) algorithm and extends this algorithm to Cox proportional hazard models.

Chapter 7 discusses the overall findings, strengths and weaknesses of the research, future directions for research, and the potential importance of both the clinical findings and the I2C2 method.

Chapter 2

Current knowledge of the relationship between pre-diagnosis health behaviours and post-diagnosis survival times in oesophageal cancer

2.1 Background

The first aim of this research project is to describe the association between pre-diagnosis health behaviours and post-diagnosis survival times for OC. This chapter reviews and collates current knowledge in this area. This provides a reference point against which to compare and critique the results of the new methods and analyses presented in subsequent chapters.

To maximise the scope and validity of the information structured approaches to data collection were used and analyses and presented findings according to MOOSE (Meta-Analysis Of Observational Studies in Epidemiology) guidelines (Stroup et al., 2000) and the more recent PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) guidelines(Page et al., 2021) for systematic reviews and meta-analyses.

Our first systematic review was based on a database search conducted on July 29, 2014 with findings published in a peer reviewed journal in 2015. Given the data base search was conducted 6 years before the presentation of the thesis, a (less formal) update was undertaken, based on a database search conducted on April 27, 2021.

The updated systematic review and meta-analysis is presented in Section 2.2. An example of the STATA code uses to conduct the meta-analysis component of the systematic

review is presented in the thesis Appendix. Section 2.4 discusses the strengths, weaknesses of the review. Section 2.5 overviews the role of this systematic review in the context of the thesis.

2.2 Impact of pre-diagnosis behaviour on risk of death from oesophageal cancer: A systematic review and meta-analysis.

In this thesis, research commenced with a systematic review and meta-analysis, based on a database search conducted on July 29, 2014. The findings of this review were published as a peer reviewed journal article. The citation is: Fahey, P. P., Mallitt, K. A., Astell-Burt, T., Stone, G., & Whiteman, D. C. (2015). Impact of pre-diagnosis behavior on risk of death from esophageal cancer: a systematic review and meta-analysis. *Cancer Causes & Control*, 26(10), 1365-1373.

In 2015 the journal *Cancer, Causes and Control* has Q1 rating in Oncology and Q2 rating in Cancer Research on Scimago rating scheme and a 2-year impact factor of 3.261. The article has 24 citations (Google Scholar) of which 19 are peer reviewed journal articles (including 2 self-citations).

Following the CRediT Taxonomy (National Information Standards Organization, 2021), author contributions were:

- Paul Fahey, David Whiteman and Thomas Astell-Burt contributed to Conceptualisation
- Paul Fahey contributed to Data curation, Formal analysis, Methodology and Writing the original draft.
- Paul Fahey and Kylie-Ann Mallitt contributed to Investigation

- David Whiteman, Glenn Stone and Thomas Astell-Burt contributed to Supervision
- All contributed to Review and Editing of the draft paper.

In this thesis, the published paper has been updated to incorporate the findings a repeat of the data-base search which was conducted on April 27, 2021. So the material presented below differs substantially from the published paper, which is incorporated as Appendix 1 of this thesis.

2.2.1 **Abstract**

Purpose: Most people diagnosed with oesophageal cancer will die from their disease, but it is not known whether risk of death is influenced by pre-morbid behaviour. A systematic review and meta-analysis was undertaken to investigate the impact of pre-diagnosis behaviour on risk of death for oesophageal cancer.

Methods: A systematic review was performed of studies reporting on the relationship between pre-diagnosis smoking, alcohol consumption, overweight and obesity, physical activity and regular consumption of non-steroidal anti-inflammatory drugs (NSAID) and risk of death from oesophageal squamous cell carcinoma (OSCC) and adenocarcinomas (OAC) was performed in July, 2014 and updated in April, 2021. Study characteristics are presented, and aggregate results are compiled using meta-analysis.

Results: From an initial pool of 1452 non-duplicate records, 27 articles arising from 26 studies met the inclusion criteria. Considerable variation was observed between studies in location, measurement categories, adjustment for other risks and results. Pooled estimates suggested that pre-diagnosis smoking was associated with a 1.08 times (95% confidence

interval (CI) 1.00-1.17) increased risk for death and pre-diagnosis alcohol consumption with a 1.30 times increased risk of death (95% CI 1.09-1.56). Evidence of effect was stronger for OSCC than OAC. A lower risk of death was observed for high pre-diagnosis body mass index (BMI) ≥25 kg/m2 (OC hazard ratio 0.84, 95% CI 0.72,0.96), although there was significant heterogeneity across studies.

Conclusions: The findings of this review suggest that a number of modifiable pre-diagnosis risk factors have a carryover effect on the risk of death from oesophageal cancer. These include smoking, drinking alcohol and BMI.

2.2.2 **Introduction**

Certain behaviours are known to increase the risk of developing cancers (World Cancer Research Fund & American Institute for Cancer Research, 2007), but the influence of pre-diagnosis behaviour on cancer prognosis is poorly understood. Cancers of the oesophagus are relatively common cancers characterized by late presentation and poor survival. Much is known about the risk factors for these cancers, making them suitable candidates for studying how pre-diagnosis risk factors might influence risk of death.

There are two main histological types of oesophageal cancer, namely adenocarcinomas (hereafter OAC) and squamous cell carcinomas (OSCC), which have similar presentations but quite different risk factors. Behaviours which increase the risk of OSCC include tobacco smoking, excess alcohol consumption, and poor diet (Hongo, Nagasaki, & Shoji, 2009), whereas OAC is associated with obesity, gastro-oesophageal acid reflux, male sex and smoking(American Cancer Society, 2011). Behaviours suspected to decrease the risk of both cancers include regular consumption of non-steroidal anti-

inflammatory drugs (NSAIDs) (Corley, Kerlikowske, Verma, & Buffler, 2003) and physical activity (Friedenreich et al., 2010).

Behaviour prior to diagnosis and behaviour after diagnosis are very different issues. After diagnosis, behavioural modification could be viewed as a component of patient treatment (Demark-Wahnefried et al., 2005). Prior to diagnosis public health programs and community education dominate.

Given population behaviour can change quite rapidly, it is important to understand the likely impact of such changes on future cancer survival. In the USA, for example, it is estimated that the percentage of adults who were smokers declined from 23.2% in 2000 to 18.1% in 2012, the percentage of adults consuming 5 or more drinks of alcohol at least 12 times per year decreased from 9.8% in 1997 to 9.2% in 2012 and the percentage who met both aerobic activity and muscle strengthening guidelines for physical activity increased from 14.5% in 1998 to 20.3% in 2012 (National Center for Health Statistics, 2014). In contrast, between 1999-2002 and 2009-12 the percentage of adult Americans overweight or obese increased from 65.2% to 68.8% (National Center for Health Statistics, 2014).

The question of whether pre-diagnosis risk factors influence survival from oesophageal cancer has been addressed in a number of small epidemiologic studies, however there have been no systematic reviews investigating this question. Therefore, the aim in this paper was to systematically review and quantify the risk of death from oesophageal cancers associated with pre-diagnosis tobacco smoking, alcohol consumption, physical activity, obesity and long-term use of NSAIDs for individuals with oesophageal cancer.

2.2.4 **Methods**

This research was conducted in compliance with MOOSE guidelines (Stroup et al., 2000).

Studies were included if they reported the relationship between pre-diagnosis behaviours and post-diagnosis risk of death for OSCC and/or OAC. The outcome of interest was time from diagnosis to all cause death, although time to disease specific death was accepted as a reasonable proxy where all cause death was not available. Included behaviours were smoking, alcohol consumption, obesity, physical activity and/or regular consumption of NSAIDs prior to or proximal to diagnosis, measured either prospectively or retrospectively. The authors' definitions of how behaviours were measured were recorded and reviewed.

Reviews, commentaries were excluded as were reports of post-diagnosis behaviour as this is likely to differ considerably from pre-diagnosis behaviour (Demark-Wahnefried et al., 2005). Reports restricted to specific disease or treatment groups (such as patients with Barrett's oesophagus or those treated surgically) were excluded as behaviour and/or risk of death may not generalize. Studies of relative mortality were excluded as mortality combines risk of incidence with risk of death. Articles from overlapping samples, conference abstracts, publications prior to 1980, and articles printed in languages other than English were excluded.

Pubmed, Medline and Embase electronic data sets were searched on July 29, 2014 and repeated on April 27, 2021 excluding publications prior to 2014. With librarian input, MeSH headings and titles and abstracts were searched for a) oesophagus and b) cancer, carcinoma or neoplasm and c) survival or prognosis and d) each of the individual behaviours. Both English and American spellings and wildcards for alternate word endings were

included. The full syntax for the Pubmed search is provided as Figure 2.1. Reference lists of eligible articles were reviewed to identify any studies missed during the electronic search.

Figure 2.1 PubMed search criteria

("Esophagus"[Mesh] OR “Esophagus"[Title/Abstract] OR "Oesophagus"[Title/Abstract] OR "Esophageal"[Title/Abstract] OR "Oesophageal"[Title/Abstract]) AND ("Neoplasms"[Mesh] OR "Neoplasm"[Title/Abstract] OR "Cancer*"[Title/Abstract] OR "Carcinoma"[Title/Abstract] OR "Adenocarcinoma"[Title/Abstract]) AND(("Survival" [Mesh]) OR ("Prognosis" [Mesh]) OR ("survival" [Title/Abstract]) OR ("prognosis" [Title/Abstract]) OR ("prognostic" [Title/Abstract])) AND (("smoking" [Title/Abstract]) OR ("tobacco" [Title/Abstract]) OR ("alcohol" [Title/Abstract]) OR ("physical activity" [Title/Abstract]) OR ("exercise" [Title/Abstract]) OR ("sedentary lifestyle"[Title/Abstract]) OR ("body mass index"[Title/Abstract]) OR ("BMI" [Title/Abstract]) OR ("obesity" [Title/Abstract]) OR ("Aspirin"[Title/Abstract]) OR ("Non-Steroidal Anti-Inflammatory"[Title/Abstract]) OR ("NSAID"[Title/Abstract]) OR ("health behavior"[Title/Abstract]) OR ("health behaviour"[Title/Abstract]) OR ("life style"[Title/Abstract]) OR ("lifestyle"[Title/Abstract]) OR ("life-style"[Title/Abstract]) OR ("Smoking"[Mesh]) OR ("Alcohol Drinking"[Mesh]) OR ("Motor Activity"[Mesh]) OR ("Exercise"[Mesh]) OR ("Sedentary Lifestyle"[Mesh]) OR ("Body Mass Index"[Mesh]) OR ("Obesity"[Mesh]) OR ("Health Behavior"[Mesh]) OR ("Life Style"[Mesh]) OR ("Aspirin"[Mesh]) OR ("Anti-Inflammatory Agents, Non-Steroidal"[Mesh]))

For the initial 2014 review, one reviewer (PF) identified potentially relevant studies by screening titles and/or abstracts of all citations identified through the database search. Final selection of articles and data collection was performed independently by two reviewers (PF and KM). Discrepancies were resolved through discussion with all authors. The 2021 update was completed by a single reviewer (PF).

Data collection forms were developed in consultation with all authors. Information extracted from each paper includes authors, country and cohort, cancer type, scope of the collection (data sources, collection years and eligibility criteria), whether prospective or retrospective, the behaviours reported, adjustment for confounders, sample size, proportion of eligible cases and the statistical results.

Risk of bias

Methodological features were assessed individually in preference to an overall quality score (Jüni, Witschi, Bloch, & Egger, 1999). Review items were taken from the Joanna Briggs Institute's (JBI) Checklist for Cohort Studies (Moola et al., 2017). Assessments were conducted by a single reviewer (PF).

Statistical Methods

Most studies reported the relative risk of death between behaviour categories as hazard ratios (HRs) from proportional hazards models. Meta-analyses were conducted on HRs and omitted studies which reported other results. Priority was given to HRs adjusted for the greatest number of other covariates. Where studies reported different numbers of categories for behaviour measures, excess categories were combined following the methods of Borenstein et al, chapter 25 (Borenstein, Hedges, Higgins, & Rothstein, 2011). Sub-groups within the same study (such as males and females reported separately) were combined using fixed effect meta-analysis prior to the main analysis.

HRs were pooled across studies using random-effect meta analysis models (Borenstein et al., 2011). For each analysis, heterogeneity in HRs was assessed using the I^2 statistic and the χ^2 test of goodness of fit (Higgins, Thompson, Deeks, & Altman, 2003). I^2 values greater than 40% were interpreted as substantial heterogeneity (Higgins et al., 2003).

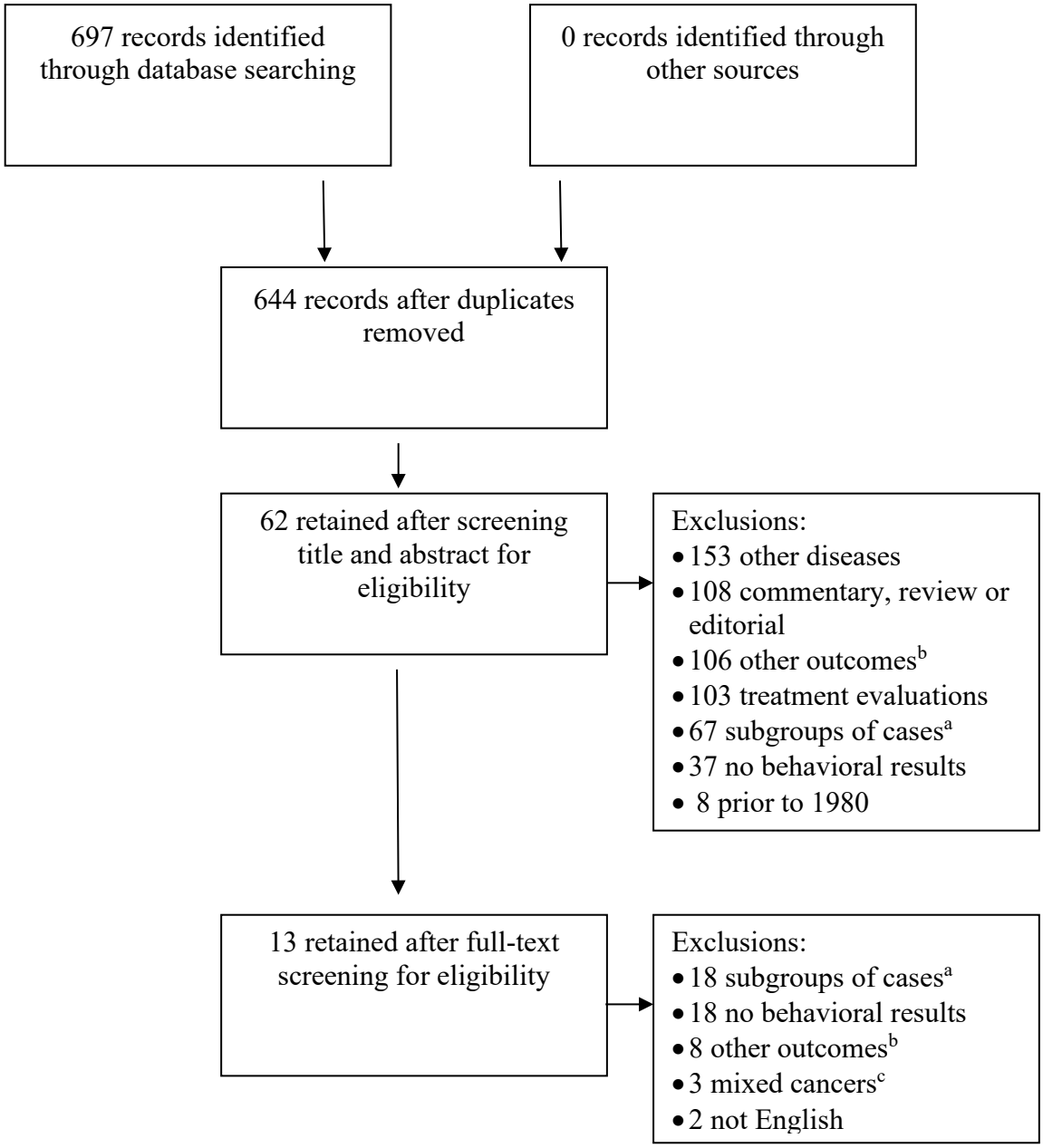
Sensitivity analyses were used to explore potential sources of heterogeneity and funnel plots and Egger's test to explore potential publication biases. Analyses were performed using Stata v14 (StataCorp, College Station, TX, USA).

2.2.5 **Results**

Figure 2.2 summarizes the results of the search and application of the exclusion criteria. The searches identified a combined total of 1452 citations after duplicates were removed. Of these, 27 were retained. The reasons for exclusions during the 2014 and 2021 searches are listed separately in Figure 2.2. In two instances, two papers reported results from separate subgroups of the same cohort (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012) and (Loehrer, Giovannucci, Betensky, Shafer, & Christiani, 2020; A. Spreafico et al., 2017) and all 4 papers are retained. However, where research articles reported results from the same individual OC cases (Y.-P. Pan et al., 2020; Y. P. Pan et al., 2018) and (L. Wang et al., 2020; R.-M. Zhou et al., 2020; R. M. Zhou, Li, Wang, Huang, & Cao, 2018), only the earliest published article was retained.

Figure 2.2 Flow chart summarising the identification of studies included for review
2014 systematic review

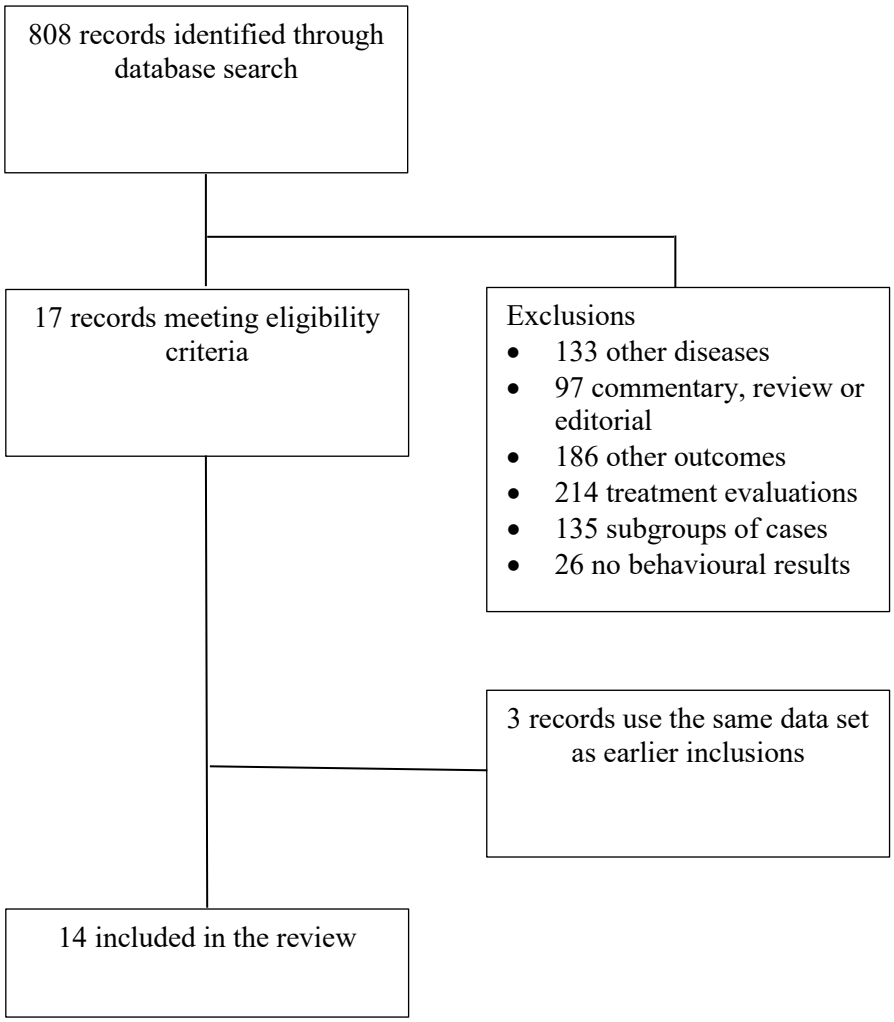
a) 2014 original review



[a] subgroups of cases include surgical cohorts, Barrett's esophagus, late stage cohorts, etc;
[b] other outcomes include relative mortality, time to progression and quality of life, etc;
[c] mixed cancers are where the esophageal cancers were combined with stomach cancer, etc.

Figure 2.2 (continued) Flow chart summarising the identification of studies included for review

b) 2021 update



The 27 articles report results from 26 studies described in Table 2.1, which were conducted in the Asia (15), North America (6), Europe (3), Australia (1) or South Africa (1). Although sometimes embedded in larger studies, all results used in this review came from case-series of oesophageal cancer patients. Only two studies was truly prospective (S. M. Park et al., 2006; Spence et al., 2018) collecting behavioural data well before diagnosis. Seventeen studies required histological confirmation of the cancer diagnosis, three relied on registry reports, two used non-histological criteria and the remaining four did not specify the

basis for diagnosis. Nine studies commenced enrolling cases during the 1990s, fifteen commenced in the 2000s and two in the 2010s. Follow-up ended in the 2000s seven studies, in the 2010s for six studies and was not clear for 13 studies.

Table 2.1 Study design and sample characteristics

Study	a) Sample size b) Diagnosis	a) Country b) Scope c) Response rate	Year of a) Enrolment b) Final follow-up	Behavioural risk factors	Timing of a) Data collection b) Exposure measurement	a) Outcome measure b) Results expressed as c) Results adjusted for
(Aghcheli et al., 2011)	a) 426 OSCC b) histologically confirmed	a) Iran b) single center c) not stated	a) 2002-2007 b) 2008	smoking	a) at diagnosis (presumed) b) lifetime, to 1 year prior to diagnosis	a) diagnosis to all causes death b) hazard ratios c) demographics, stage / comorbidity, treatment
(Cescon et al., 2009)	a) 300 OAC 63 OSCC 8 Poorly differentiated b) histologically confirmed	a) USA b) two centers c) not stated	a) 1999-2004 b) 2007	smoking	a) after diagnosis (presumed) b) not stated	a) diagnosis to all causes death b) hazard ratios c) nil
(Jing et al., 2012)	a) 168 OSCC b) histologically confirmed	a) China b) single center c) 90%	a) 2006-2008 b) 2011	smoking alcohol	a) not stated b) 1 year before diagnosis	a) diagnosis to disease specific death b) hazard ratios c) nil (presumed)
(S. M. Park et al., 2006)	a) 272 Oesophageal b) Notified to a registry	a) South Korea b) all male government employees c) not stated	a) 1996 behaviour, 1996-2002 diagnosis b) 2004	smoking alcohol BMI	a) prior to diagnosis b) 0-6 years before to diagnosis	a) diagnosis to all causes death b) demographics, stage / comorbidity, other behaviour
(Salek et al., 2009)	a) 552 Oesophageal b) histologically confirmed	a) Iran b) two centers c) 33%	a) 1997-2004 b) 2006	BMI	a) not stated b) not stated	a) diagnosis to all causes death b) median survival time c) nil
(Samadi et al., 2007)	a) 124 Oesophageal b) histologically confirmed	a) Iran b) single center c) 88%	a) 2000-2004 b) not stated	smoking alcohol	a) at diagnosis b) not stated	a) not stated b) hazard ratios c) demographics, stage / comorbidity, treatment, other behaviour

Table 2.1 (continued) Study design and sample characteristics

Study	a) Sample size b) Diagnosis	a) Country b) Scope c) Response rate	Year of a) Enrolment b) Final follow-up	Behavioural risk factors	Timing of a) Data collection b) Exposure measurement	a) Outcome measure b) Results expressed as c) Results adjusted for
(Sehgal, Kaul, Gupta, & Dhar, 2012)	a) 200 OSCC b) histologically confirmed	a) India b) single center c) not stated	a) 2007-2011 b) 3 years after baseline	smoking	a) at diagnosis (presumed) b) not stated	a) not stated b) Kaplan-Meier curves c) nil
(Shitara et al., 2010)	a) 363 OSCC b) not stated	a) Japan b) single center c) 97%	a) 2001-2005 b) not stated	smoking alcohol	a) prior to diagnosis b) prior to symptoms	a) diagnosis to all causes death b) hazard ratios c) demographics, stage / comorbidity, treatment, other behaviour
(Sundelöf, Lagergren, & Ye, 2008)	a) 159 OSCC 177 OAC b) histologically confirmed	a) Sweden b) all OAC, half OSCC in Sweden c) 77%	a) 1994-1997 b) 2004	smoking alcohol BMI phys activity	a) after diagnosis b) lifetime & 20 years prior to interview	a) diagnosis to all causes death b) hazard ratios c) demographics, stage / comorbidity, treatment, other behaviour
(Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012)	a) 301 OSCC 362 OAC b) histologically confirmed	a) Australia b) Australian mainland c) 70% of those detected	a) 2001-2005 b) 2010	smoking alcohol BMI phys activity (OAC) NSAID	a) after diagnosis b) 1 year prior to diagnosis & lifetime	a) diagnosis to all causes death b) hazard ratios c) demographics, stage / comorbidity, treatment, other behaviour
(Trivers et al., 2005)	a) 220 OSCC 298 OAC b) histologically confirmed	a) USA b) 2.5 US states c) 78%	a) 1993-1995 b) 2000	smoking alcohol BMI NSAID	a) mean 3.7 months after diagnosis b) 1 year prior to interview	a) diagnosis to all causes death b) hazard ratios c) nil
(Wu, Chen, Andrews, Ruiz, & Correa, 2007)	a) 718 OSCC b) histologically confirmed	a) Taiwan b) 3 centers c) 87%	a) 2000-2008 b) 2008	smoking alcohol	a) within 1 week of diagnosis b) lifetime	a) diagnosis to all causes death b) hazard ratios c) demographics, stage / comorbidity

Table 2.1 (continued) Study design and sample characteristics

Study	a) Sample size b) Diagnosis	a) Country b) Scope c) Response rate	Year of a) Enrolment b) Final follow-up	Behavioural risk factors	Timing of a) Data collection b) Exposure measurement	a) Outcome measure b) Results expressed as c) Results adjusted for
(Araujo et al., 2016)	a) 130 (105 OAC) b) histologically confirmed	a) USA b) 2 hospitals c) not stated	a) 2009 b) not stated (median 21.3 months)	NSAID	a) after diagnosis (median 48 days after) b) number of years taking aspirin	a) diagnosis to all causes death and disease specific death b) hazard ratios c) age, non-aspirin antiplatelet use, tumour subsite, nodal status, surgery
(Dandara, Robertson, Dzobo, Moodley, & Parker, 2016)	a) 1685 OSCC b) not stated	a) South Africa b) single hospital c) 100% (medical record review)	a) 1997-2007 b) max 5 years for each patient	smoking alcohol	a) not stated b) lifetime	a) diagnosis to all causes death b) hazard ratios c) nil
(He et al., 2020)	a) 5,283 (4,506 ESCC) b) National Comprehensive Cancer Network guidelines, 2010	a) China b) 18 hospitals c) 100% (medical record review)	a) 2011-2013 b) 2017	smoking alcohol BMI	a) not stated b) lifetime (smoking), at admission (BMI), not stated (alcohol)	a) diagnosis to all causes death b) hazard ratios c) nil
(Korpanty et al., 2017)	a) 270 OAC b) histologically confirmed	a) Canada b) single centre c) 72%	a) 2006-2013 b) not stated (median follow-up 58.5 months)	smoking alcohol	a) after diagnosis b) not stated smoking), lifetime (alcohol)	a) diagnosis to all causes death b) hazard ratios c) nil
(Loehrer et al., 2020)	a) 285 OAC b) histologically confirmed	a) USA b) single hospital c) not stated	a) 2004-2016 b) 2018	BMI	a) after diagnosis (median 4.8 weeks) b) lifetime	a) diagnosis to all causes death b) hazard ratios c) age, sex, age, smoking status, treatment, year of diagnosis

Table 2.1 (continued) Study design and sample characteristics

(Macfarlane, Murchie, & Watson, 2015)	a) 1197 b) Notified to a registry	a) Scotland b) cohort c) not stated	a) 1996-2010 b) not stated (max follow-up 181 months)	smoking alcohol NSAID	a) not stated b) lifetime (smoking), current (alcohol), not stated (NSAID)	a) diagnosis to all causes death b) hazard ratios c) age, gender, socio-economic status
(Okada et al., 2017)	a) 365 OSCC b) histologically confirmed	a) Japan b) 66 hospitals c) not stated	a) 2003-2007 b) 2010-2014	smoking alcohol BMI phys activity	a) not stated b) lifetime (smoking, alcohol, not stated (BMI, physical activity)	a) diagnosis to all causes death b) hazard ratios c) sex, age, year of diagnosis, BMI, smoking history, alcohol history, physical exercise, stage
(Y. P. Pan et al., 2018)	a) 205 OSCC b) histologically confirmed	a) Taiwan b) single hospital c) 100% (medical record review)	a) 2007-2012 b) not stated	smoking alcohol	a) not stated b) not stated	a) 5-year survival (y/n) b) odds ratios c) nil
(Y.-P. Pan et al., 2020)	This paper presents results for the same participants as Y.P. Pan et al, 2018 above					
(Spence et al., 2018)	a) 1,606 OAC 828 OSCC b) Notified to a registry	a) England & Scotland b) cancer registries c) not stated	a) 1998-2012 b) 2015	NSAID	a) not stated b) prior to diagnosis	a) diagnosis to disease specific death b) hazard ratios c) sex, age, year of diagnosis, deprivation, cancer treatment, comorbidities, other prescription medications

Table 2.1 (continued) Study design and sample characteristics

(A. Spreafico et al., 2017)	a) 564 OAC b) not stated	a) USA & Canada b) 2 cancer centres c) USA 82%, Canada 77%	a) USA 1999-2004 Canada 2006-2011 b) not stated	smoking BMI	a) after diagnosis b) lifetime	a) diagnosis to all causes death b) hazard ratios c) stage, Eastern Cooperative Oncology Group Performance Status, treatment, study site
(Talukdar et al., 2021)	a) 76 OSCC b) endoscopy & biopsy	a) India b) 2 hospitals c) not stated	a) 2016-2019 b) not stated	smoking alcohol	a) not stated b) not stated	a) diagnosis to all causes death b) hazard ratios c) nil
(X. P. Wang et al., 2016)	a) 210 OSCC b) not stated	a) China b) single centre c) not stated	a) 2007-2009 b) 2015	smoking alcohol	a) not stated b) lifetime (smoking), not stated (alcohol)	a) diagnosis to all causes death b) hazard ratios c) nil
(Yao et al., 2014)	a) 136 OSCC b) histologically confirmed	a) China b) single hospital c) 53%	a) 2005-2007 b) not stated (follow-up 3 to 7.5 years)	smoking	a) not stated b) not stated	a) diagnosis to disease specific death b) hazard ratios c) tumour size, age, differentiation, tumour site, gender, p53 overexpression
(R. M. Zhou et al., 2018)	a) 207 OSCC b) histologically confirmed	a) China b) cohort c) not stated	a) 2008-2012 b) not stated	smoking	a) not stated b) at least 2 year duration	a) diagnosis to all causes death b) hazard ratios c) nil
(R.-M. Zhou et al., 2020)	This paper presents results for the same participants as R.M. Zhou et al, 2018 above					
(L. Wang et al., 2020)	This paper presents results for the same participants as R.M. Zhou et al, 2018 above					

Four studies presented results for OSCC and OAC separately, four presented results for OAC alone and twelve for OSCC alone. The remaining six studies did not differentiate histological type. Of these, one from the US reported 81% of participants were OAC (Cescon et al., 2009) and has been grouped with the OAC cohorts in this review. In the United Kingdom, OAC is increasingly more common than OSCC (C Castro et al., 2014). Therefore, one study from England and Scotland (Macfarlane et al., 2015) was also grouped with OAC. Three studies from Asia reported 85% (Samadi et al., 2007), 93% (Salek et al., 2009) and 85% (He et al., 2020) of participants were OSCC and one was from South Korea (S. M. Park et al., 2006) where an estimated 95% of oesophageal cancers are OSCC (Hongo et al., 2009). These four studies were grouped with the OSCC cohorts.

Twelve studies reported collecting data from at least 70% of the target sample with one collecting BMI enrolling 33% of the target sample (Salek et al., 2009), and one collecting smoking enrolling 53% of the target sample (Yao et al., 2014). Response rates were not adequately stated in twelve articles. Seven studies presented HRs after adjustment for demographic characteristics, disease progression and/or comorbidities, treatment and other behaviours, seven adjusted for some but not all of these factors, nine presented unadjusted HRs only, one presented unadjusted odds ratio for 5-year survival (Y. P. Pan et al., 2018), one presented Kaplan-Meier curves rather than HRs (Sehgal et al., 2012) and one median survival times (Salek et al., 2009). No studies adjusted for post-diagnosis behaviour.

Risk of bias

Only 11 of the 26 studies focussed exclusively on relationships between pre-diagnosis health behaviour and post-diagnosis survival time in OC (Aghcheli et al., 2011; Araujo et al., 2016; Loehrer et al., 2020; Macfarlane et al., 2015; Y. P. Pan et al., 2018; S. M. Park et al.,

2006; Sehgal et al., 2012; Spence et al., 2018; A. Spreafico et al., 2017; Sundelöf et al., 2008; Wu et al., 2007). Of the remaining studies, eight investigated a wider range of demographic and clinical predictors of risk of death than just health behaviours (Dandara et al., 2016; He et al., 2020; Okada et al., 2017; Salek et al., 2009; Samadi et al., 2007; Shitara et al., 2010; Thrift, Nagle, Fahey, Russell, et al., 2012; Trivers et al., 2005) and seven focused mainly on laboratory markers (Cescon et al., 2009; Jing et al., 2012; Korpanty et al., 2017; Talukdar et al., 2021; X. P. Wang et al., 2016; Yao et al., 2014; R. M. Zhou et al., 2018). Where other predictors were the primary focus, the study designs and measurement tools specific to health behaviours were often poorly described.

Table 2.2 summarises each study's adherence to the JBI review items for cohort studies.

Table 2.2 Adherence to items listed in the JBI Checklist for Cohort Studies

	(Aghcheli et al., 2011)	(Cescon et al., 2009)	(Jing et al., 2012)	(S. M. Park et al., 2006)	(Salek et al., 2009)	(Samadi et al., 2007)	(Sehgal et al., 2012)
1. Were the two groups similar and recruited from the same population?	Unclear	Yes	Unclear	Yes	Unclear	Yes	Yes
2. Were the exposures measured similarly to assign people to both exposed and unexposed groups?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3. Was the exposure measured in a valid and reliable way?	No	Unclear	No	No	No	Unclear	Unclear
4. Were confounding factors identified?	Yes	No	No	No	No	No	No
5. Were strategies to deal with confounding factors stated?	Yes	No	No	Yes	No	Yes	No
6. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7. Were the outcomes measured in a valid and reliable way?	Yes	Unclear	Unclear	Unclear	Unclear	No	Unclear
8. Was the follow up time reported and sufficient to be long enough for outcomes to occur?	Unclear	Yes	Yes	Yes	No	Unclear	Unclear
9. Was follow up complete, and if not, were the reasons to loss to follow up described and explored?	No	No	No	Unclear	No	Yes	Unclear
10. Were strategies to address incomplete follow up utilized?	Yes	No	No	No	No	No	Unclear
11. Was appropriate statistical analysis used?	Unclear	No	No	Yes	No	Unclear	No

Table 2.2 (continued) Adherence to items listed in the JBI Checklist for Cohort Studies

	(Shitara et al., 2010)	(Sundelöf et al., 2008)	(Thrift, Nagle, Fahey, Russell, et al., 2012)	(Trivers et al., 2005)	(Wu et al., 2007)	(Araujo et al., 2016)	(Dandara et al., 2016)
1. Were the two groups similar and recruited from the same population?	Unclear	Yes	Unclear	Unclear	Unclear	Unclear	Yes
2. Were the exposures measured similarly to assign people to both exposed and unexposed groups?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3. Was the exposure measured in a valid and reliable way?	No	No	No	No	Yes	No	No
4. Were confounding factors identified?	Yes	Yes	Yes	No	Yes	No	No
5. Were strategies to deal with confounding factors stated?	Yes	Yes	Yes	No	Yes	Yes	No
6. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7. Were the outcomes measured in a valid and reliable way?	Unclear	Unclear	No	Unclear	Unclear	Unclear	Unclear
8. Was the follow up time reported and sufficient to be long enough for outcomes to occur?	Yes	Yes	Yes	Yes	Unclear	Yes	Unclear
9. Was follow up complete, and if not, were the reasons to loss to follow up described and explored?	Unclear	Unclear	Unclear	Yes	Unclear	Unclear	Unclear
10. Were strategies to address incomplete follow up utilized?	Yes	No	Unclear	Unclear	Unclear	No	Unclear
11. Was appropriate statistical analysis used?	No	Yes	Unclear	No	Unclear	Unclear	No

Table 2.2 (continued) Adherence to items listed in the JBI Checklist for Cohort Studies

	(He et al., 2020)	(Korpanty et al., 2017)	(Loehrer et al., 2020)	(Macfarlane et al., 2015)	(Okada et al., 2017)	(Y. P. Pan et al., 2018)	(Spence et al., 2018)
1. Were the two groups similar and recruited from the same population?	Unclear	Unclear	Unclear	Yes	Unclear	Yes	Unclear
2. Were the exposures measured similarly to assign people to both exposed and unexposed groups?	Yes	Unclear	Yes	Unclear	Yes	Yes	Yes
3. Was the exposure measured in a valid and reliable way?	Unclear	Unclear	No	No	Unclear	No	Unclear
4. Were confounding factors identified?	No	No	No	No	Yes	No	Yes
5. Were strategies to deal with confounding factors stated?	No	No	Yes	Yes	Yes	No	Yes
6. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	Yes	Yes	Yes	Yes	Unclear	Yes	No
7. Were the outcomes measured in a valid and reliable way?	Unclear	Yes	Unclear	Unclear	Unclear	Unclear	Unclear
8. Was the follow up time reported and sufficient to be long enough for outcomes to occur?	Yes	Yes	No	No	Yes	Yes	Yes
9. Was follow up complete, and if not, were the reasons to loss to follow up described and explored?	No	Unclear	No	No	Unclear	Unclear	Unclear
10. Were strategies to address incomplete follow up utilized?	No	Yes	Unclear	No	No	Unclear	No
11. Was appropriate statistical analysis used?	No	No	Unclear	Yes	Unclear	No	Unclear

Table 2.2 (continued) Adherence to items listed in the JBI Checklist for Cohort Studies

	(A. Spreafico et al., 2017)	(Talukdar et al., 2021)	(X. P. Wang et al., 2016)	(Yao et al., 2014)	(R. M. Zhou et al., 2018)
1. Were the two groups similar and recruited from the same population?	Unclear	Unclear	Unclear	Unclear	Unclear
2. Were the exposures measured similarly to assign people to both exposed and unexposed groups?	Yes	Yes	Unclear	Unclear	Yes
3. Was the exposure measured in a valid and reliable way?	No	No	Unclear	Unclear	No
4. Were confounding factors identified?	Yes	No	No	No	No
5. Were strategies to deal with confounding factors stated?	Yes	No	No	Yes	No
6. Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?	Yes	Unclear	Unclear	Yes	Yes
7. Were the outcomes measured in a valid and reliable way?	Unclear	Unclear	Unclear	Unclear	Unclear
8. Was the follow up time reported and sufficient to be long enough for outcomes to occur?	Yes	Yes	Unclear	Yes	No
9. Was follow up complete, and if not, were the reasons to loss to follow up described and explored?	Unclear	Yes	Unclear	Unclear	Unclear
10. Were strategies to address incomplete follow up utilized?	Unclear	Unclear	Unclear	Unclear	Unclear
11. Was appropriate statistical analysis used?	No	No	No	Unclear	No

In this review JBI item 1 *"Were the two groups similar and recruited from the same population?"* was marked as 'unclear' if there was potential for people with specific behaviours to be under-represented in the sample. For example, in one study (Trivers et al., 2005) baseline interview was completed for about 80% of eligible cases. This leaves some potential that patients who were say smokers, were more ill, more likely to have died and less likely to be included in the study than non-smokers. Further, as the outcome variable is survival time, JBI item 6 *"Were the groups/participants free of the outcome at the start of the study (or at the moment of exposure)?"* may also be sensitive to the delay between diagnosis and measurement. Generally, studies attempted to minimise the delay between diagnosis and enrolment but one notable exception required everyone to survive for 6 months in order to qualify for inclusion (Spence et al., 2018).

It is expected that JBI item 2 *"Were the exposures measured similarly to assign people to both exposed and unexposed groups?"* was always true as the same questionnaires and clinical records were used to obtain behavioural data for all participants. However, this item has been marked 'unclear' when there was no description of how the behavioural measures were defined or collected.

JBI item 3. *"Was the exposure measured in a valid and reliable way?"* was marked as 'no' or 'unclear' throughout. All studies relied on self-report, often with long recall periods. Only one of the 26 studies discussed the validity of their behavioural measures (Wu et al., 2007). That validity test demonstrated clear evidence of a distinction between self-reported smokers and non-smokers, but a less compelling evidence of the distinction between self-reported alcohol consumption and non-consumption (Lin et al., 2011).

JBI item 4 *"Were confounding factors identified?"* was marked as 'yes' if the hazard ratio for the behaviour was adjusted for both age and cancer stage at diagnosis. Age is an

important confounder because age is associated with survival time (Njei, McCarty, & Birk, 2016) and health behaviours change with age. Cancer stage at diagnosis is included as a required confounder as earlier stage is associated with much longer survival (Zhang, 2013) and it is plausible that people with different health behaviours may tend to have different cancer stage at diagnosis. Some studies adjusted for confounders other than cancer stage (Araujo et al., 2016; Loehrer et al., 2020; Macfarlane et al., 2015; S. M. Park et al., 2006; Samadi et al., 2007; Yao et al., 2014). However, most of these other variables (such as comorbidities, or treatments received) were probably on the causal pathway (Cole et al., 2010) and so do not merit confounder adjustment.

JBI item 5 *"Were strategies to deal with confounding factors stated?"* was marked 'yes' if a confounder adjusted hazard ratio was derived from multivariable regression model such as Cox regression or Poisson regression. However, adjustment for cancer stage was often incomplete. For example, in one study only 28% of patients had a cancer stage recorded (Aghcheli et al., 2011) and another had 50% recorded (Thrift, Nagle, Fahey, Smithers, et al., 2012). Where behavioural measures were not the sole focus of the research, a number of studies presented unadjusted hazard ratios for the behavioural predictors, even though they applied confounder adjustment to other analyses (Cescon et al., 2009; He et al., 2020; Korpanty et al., 2017; Y. P. Pan et al., 2018; Salek et al., 2009; Trivers et al., 2005; R. M. Zhou et al., 2018).

JBI item 7 "*Were the outcomes measured in a valid and reliable way?"* requires both the date of diagnosis and the date of death to be obtained. Some studies failed to provide any description of how survival time was defined or measured (Dandara et al., 2016; Y. P. Pan et al., 2018; Salek et al., 2009; Sehgal et al., 2012; A. Spreafico et al., 2017; Talukdar et al., 2021; R. M. Zhou et al., 2018). All other studies identified information sources such as clinical records, clinical visits, telephone contacts, contacts with relatives or data linkage to

various cancer registries or national death indexes, but none discussed how valid or reliable these data collection practices were.

Oesophageal cancer has a short survival time. In the US in 2020 with an estimated 5-year relative survival of just under 20% (National Cancer Institute, 2021a). In this review a median follow-up times of more than 18 months was deemed sufficient to address JBI item 8. *"Was the follow up time reported and sufficient to be long enough for outcomes to occur?"*. One study stated that all patients were follow-up to death (Talukdar et al., 2021). Other studies had median follow-up times of 21 months (Araujo et al., 2016), 29 and 36 months (A. Spreafico et al., 2017), 33 months (Cescon et al., 2009), 39 months (Jing et al., 2012), 59 months (Korpanty et al., 2017), 1.3 years (Spence et al., 2018), 5.3 years (Yao et al., 2014), 5.6 years (Shitara et al., 2010), 6.1 years (Okada et al., 2017), 6.4 years (Thrift, Nagle, Fahey, Smithers, et al., 2012) and mean follow-up of 29 months (Y. P. Pan et al., 2018), 3.03 years (S. M. Park et al., 2006) and in the range of 5 to 7 years (Trivers et al., 2005), 7 to 10 years (Sundelöf et al., 2008). However a median follow-time of 6 months (Salek et al., 2009) was regarded as insufficient. Studies quoting a wide range of follow-up times - 2 months to 8 years (Wu et al., 2007) or up to 5 years (Dandara et al., 2016) were judged to be unclear as to whether follow-up time was sufficient. The remaining studies failed to quantify their follow-up times.

JBI item 9 *"Was follow up complete, and if not, were the reasons to loss to follow up described and explored?"* was surprisingly poorly complied with. One study claimed that all patients were followed till death (Talukdar et al., 2021), one stated only 2/518 patients had been lost to follow-up (Trivers et al., 2005) and one stated 62/420 were lost to follow-up due to inaccessible location or change of address. Many studies relying on clinical records or data linkage seemed to *assume* that there was no loss to follow-up without any supporting

evidence. Consequently, JBI item 10 *"Were strategies to address incomplete follow up utilized?"* was also poorly addressed.

For JBI item 11 *"Was appropriate statistical analysis used?"*. Appropriate statistical analysis was defined as appropriate regression modelling with confounder adjustment and checking of assumptions. This was demonstrated in three studies (Macfarlane et al., 2015; S. M. Park et al., 2006; Sundelöf et al., 2008) Many studies reported results from Cox regression models without first checking the proportional hazards assumption (Aghcheli et al., 2011; Araujo et al., 2016; Loehrer et al., 2020; Okada et al., 2017; Samadi et al., 2007; Shitara et al., 2010; Spence et al., 2018; Sundelöf et al., 2008; Thrift, Nagle, Fahey, Smithers, et al., 2012; Wu et al., 2007; Yao et al., 2014). Two studies employed since discredited stepwise methods for model building (Shitara et al., 2010; A. Spreafico et al., 2017). All other studies failed to correct for any confounding variables when looking at behavioural variables.

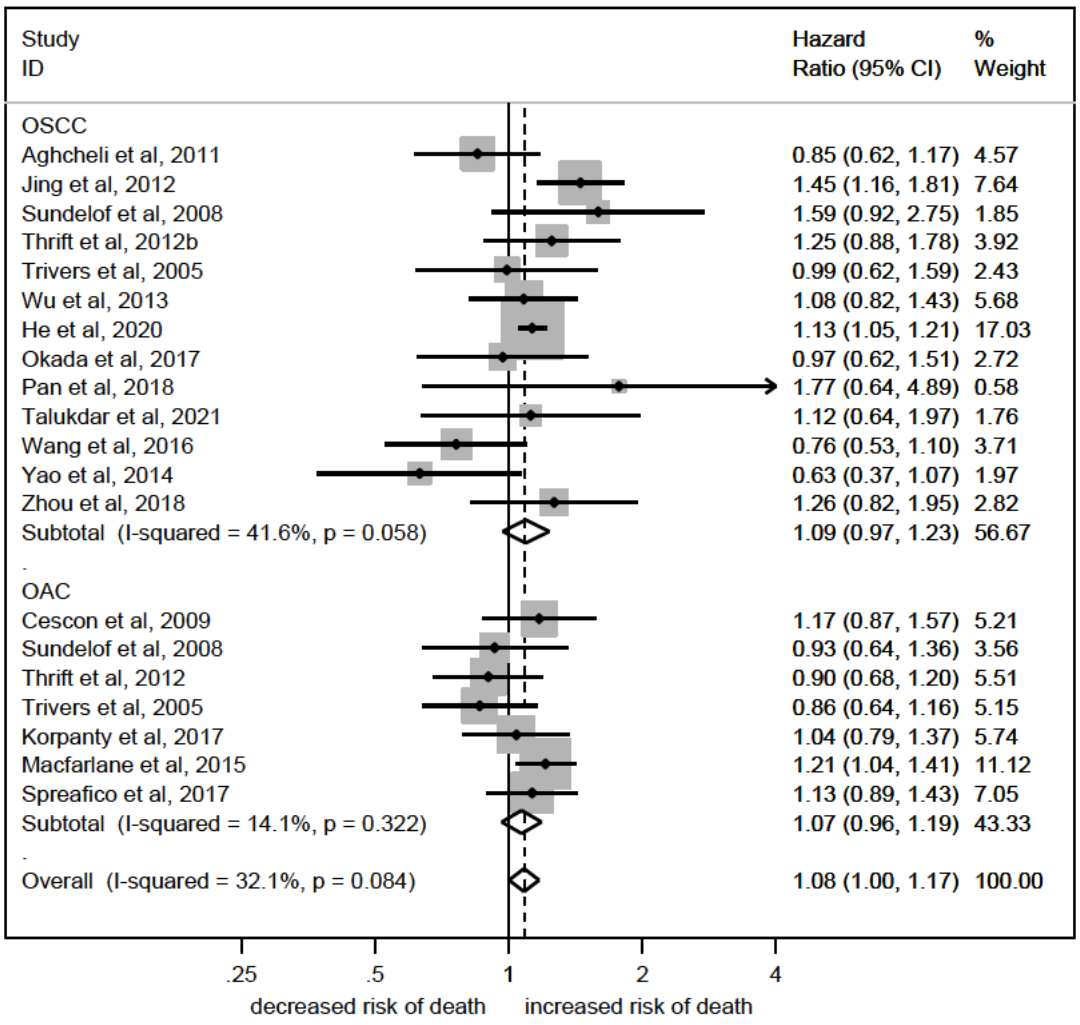
Tobacco smoking

Twenty-three of the 26 studies reported on the relationship between tobacco smoking and risk of death. Smoking was reported according to whether or not participants smoked and/or the number of pack years smoked. The most commonly reported measures were never (<100 cigarettes in lifetime) or ever (Aghcheli et al., 2011; He et al., 2020; Korpanty et al., 2017; Macfarlane et al., 2015; Okada et al., 2017; Y. P. Pan et al., 2018; Samadi et al., 2007; Talukdar et al., 2021; Trivers et al., 2005; S. M. Wang et al., 2016; Wu et al., 2007; R. M. Zhou et al., 2018) and never, former or current (Cescon et al., 2009; S. M. Park et al., 2006; A. Spreafico et al., 2017; Sundelöf et al., 2008; Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012) but eleven different category systems were used. Five of the 17 articles with OSCC results (He et al., 2020; Jing et al., 2012; Sehgal et al.,

2012; Shitara et al., 2010; Sundelöf et al., 2008) and two of seven articles with OAC results (Cescon et al., 2009; Macfarlane et al., 2015) reported at least one smoking category with a statistically significantly increased risk of death compared to the never or least smoking category.

Figure 2.3 presents a forest plot of the HRs of ever versus never smoking status for OSCC and OAC and both subgroups combined. If not directly available, the results for the 'ever' smoking category was created wherever possible by combining results from other smoking categories. For example, results for 'former' and 'current' smoker categories were combined to produce results for the 'ever' smoking category. Often the majority of 'ever' smokers are 'former' smokers; for example 65% (Cescon et al., 2009), 66% of OSCC sub-group (Sundelöf et al., 2008), 59% (Thrift, Nagle, Fahey, Russell, et al., 2012), 73% (Thrift, Nagle, Fahey, Smithers, et al., 2012). There was one mis-specified confidence interval (HR 1.04, 95% CI 0.76,1.04) (Korpanty et al., 2017) which was replaced by the presumed correct bounds of 0.76 to 1.32. One study, lacking standard deviations and confidence intervals (Dandara et al., 2016), was excluded.

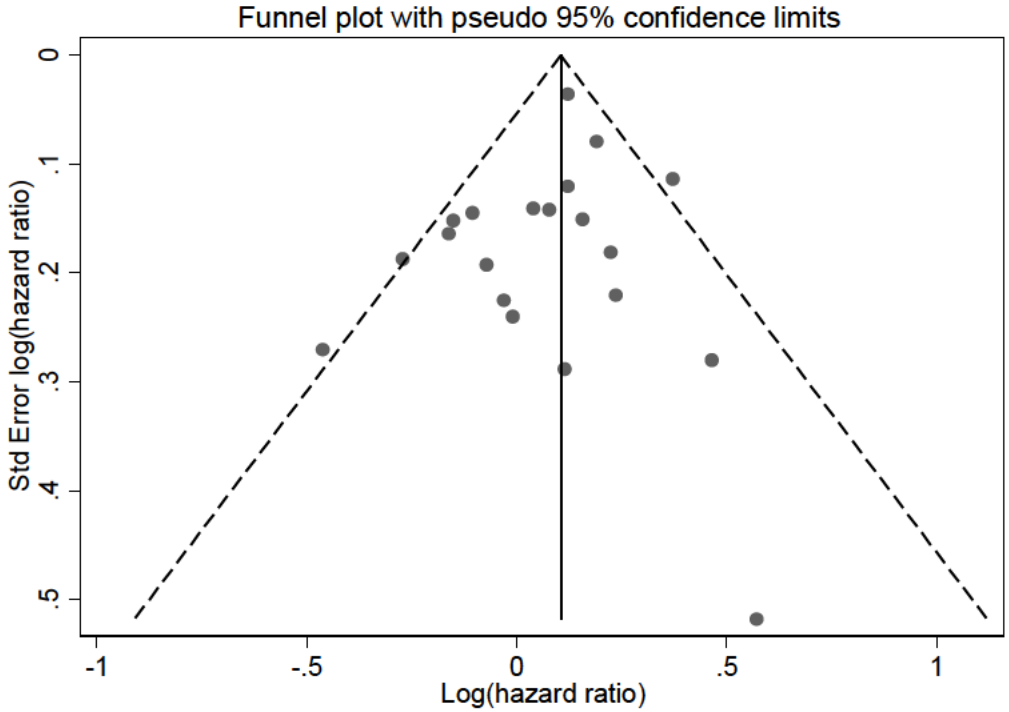
Figure 2.3 Forest plot of risk of death for ever versus never smoked (pre-diagnosis) by cancer type.



The pooled results shown in Figure 2.3 suggested a clinically small effect (HR 1.08, 95% CI 1.00,1.17) with marginal statistical significance (p=0.048), similar in both histological sub-groups (OSCC HR 1.09, OAC HR 1.07) subject to slight to moderate heterogeneity (I^2=32.1%).

The associated funnel plot (Figure 2.4) revealed no evidence of publication bias (Egger's test p=0.270). Sensitivity analyses found that restricting analyses to studies with adjustment for other covariates only did not change the overall results (OC HR=1.03 95% CI 0.90,1.19, I^2=38.8%).

Figure 2.4 Funnel plot of results for ever against never smoked pre-diagnosis.

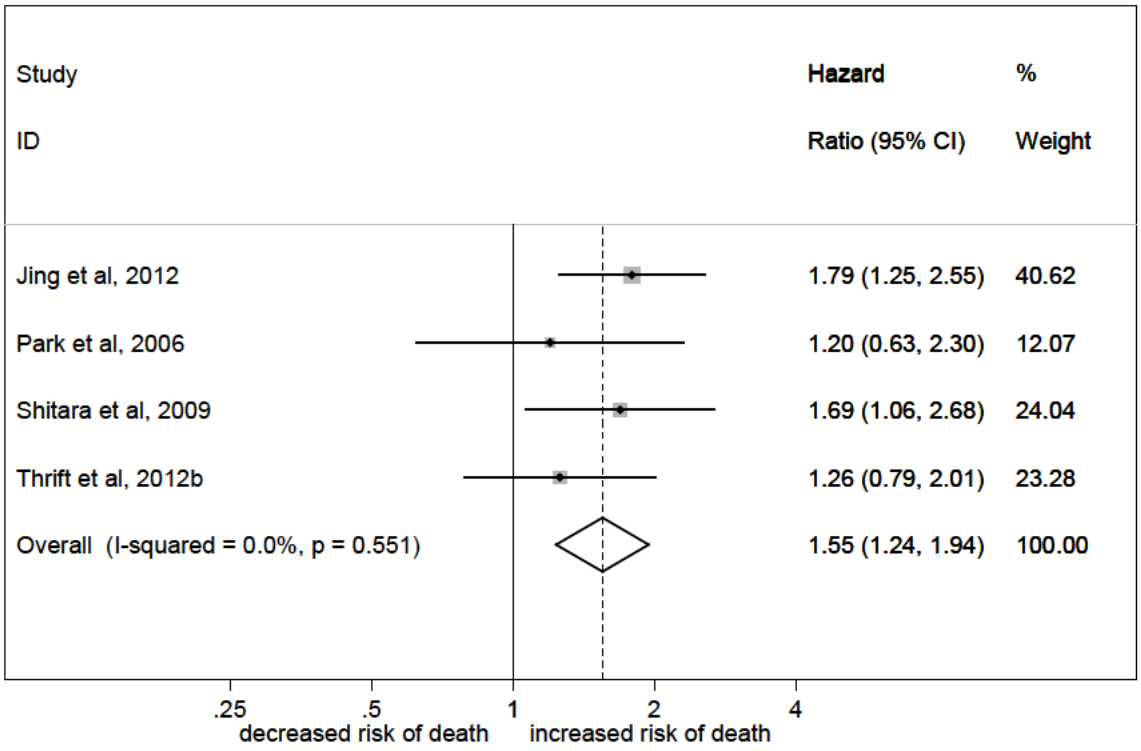


As ever smoking can incorporate relatively small exposures, some studies also reported packyears. The sole study which analysed packyears as a continuous variable (A. Spreafico et al., 2017) found strong evidence of association (HR for each additional 20 pack years 1.22, 95% CI 1.15,1.43, p<0.001) in OAC. The only other result for pack years in OAC found that ≥30 packyears had a HR of just 1.07 (95% CI 0.75,1.52) compared to never smoker.

In the OSCC sub-group, four articles analysed packyears as a categorical variable (Jing et al., 2012; S. M. Park et al., 2006; Shitara et al., 2010; Thrift, Nagle, Fahey, Russell, et al., 2012). Figure 2.5 presents the pooled results for highest versus lowest packyear categories. While highest exposure categories differed between 20 packyears (S. M. Park et al., 2006), 30 packyears (Thrift, Nagle, Fahey, Russell, et al., 2012) and 40 packyears (Jing et al., 2012; Shitara et al., 2010), the risk estimates were homogeneous (I^2=0%). The combined HR for highest pack years category versus lowest was 1.55 (95% CI 1.24,1.94). Sensitivity

analysis including Asian OSCC studies alone increased the HR marginally to 1.65 (95% CI 1.27,2.14) without loss of homogeneity. In contrast including only adjusted HRs decreased the pooled HR to 1.26 (95% CI 1.00,1.58).

Figure 2.5 Forest plot of risk of death for highest against lowest pre-diagnosis smoking pack years, OSCC.



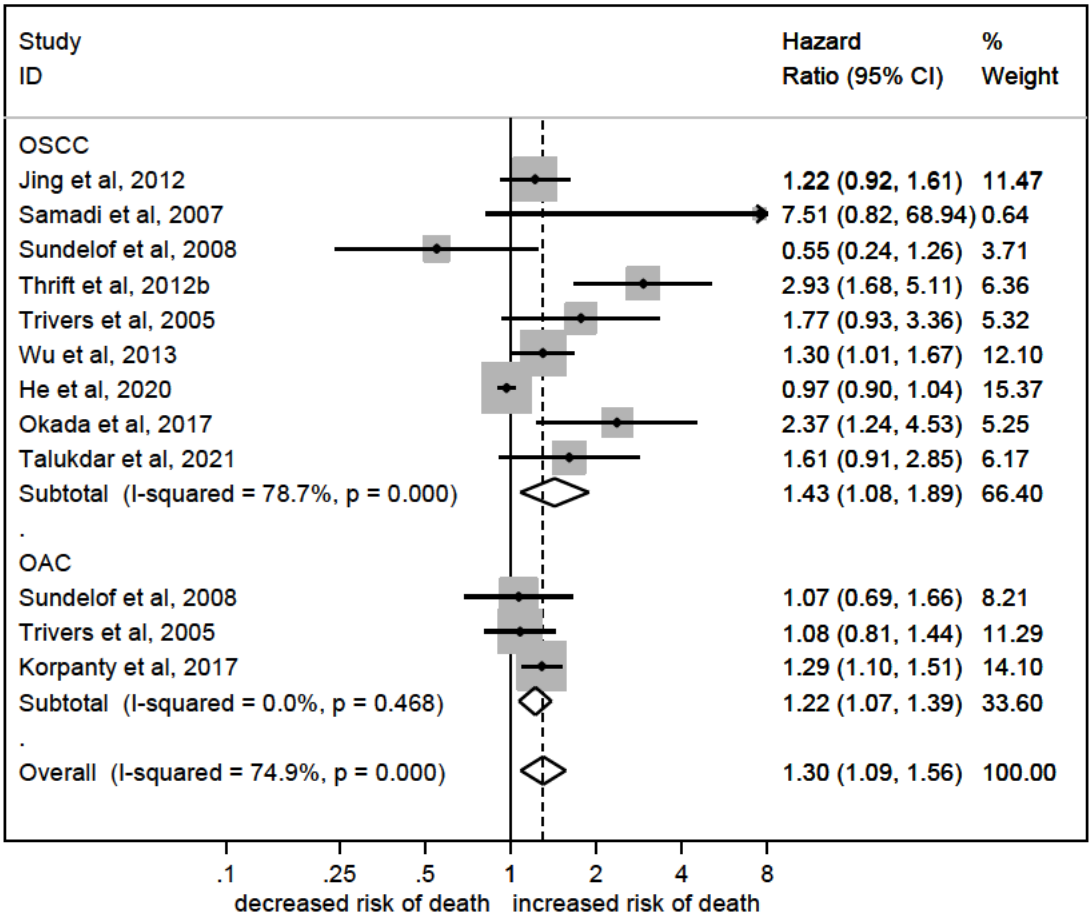
Alcohol consumption

Seventeen of the 27 articles reported alcohol consumption. Thirteen reported never regular drinker as the lowest category, but highest consumption categories varied between ever drinker (Korpanty et al., 2017; Okada et al., 2017; Samadi et al., 2007; Thrift, Nagle, Fahey, Russell, et al., 2012; Trivers et al., 2005; Wu et al., 2007) and from ≥50 grams/week (Dandara et al., 2016) to ≥250 grams/week (Jing et al., 2012). Four of 14 articles reporting results for the OSCC sub-group studies (Okada et al., 2017; Talukdar et al., 2021; Thrift,

Nagle, Fahey, Russell, et al., 2012; Wu et al., 2007) reported at least one alcohol category with a statistically significantly increased risk of death but none of the five articles reporting results for OAC reported significant associations.

Figure 2.6 presents a forest plot of the HRs of regular alcohol use pre-diagnosis against never (after collapsing categories where possible). Alcohol consumption is associated with increased risk of death in OC (HR=1.30, 95% CI 1.09, 1.56) but the high heterogeneity (I^2=74.9%, Egger's test p=0.015) suggesting other unknown factors are impacting on these results. The increased risk of death in OSCC subgroup (HR 1.43) is potentially higher than the increased risk of death in the OAC subgroup (HR 1.22). However, there is considerable heterogeneity in results in OSCC with point estimates of the hazard ratios ranging from 0.55 to 7.51.

Figure 2.6 Forest plot of risk of death for ever versus never alcohol use pre-diagnosis by cancer type.

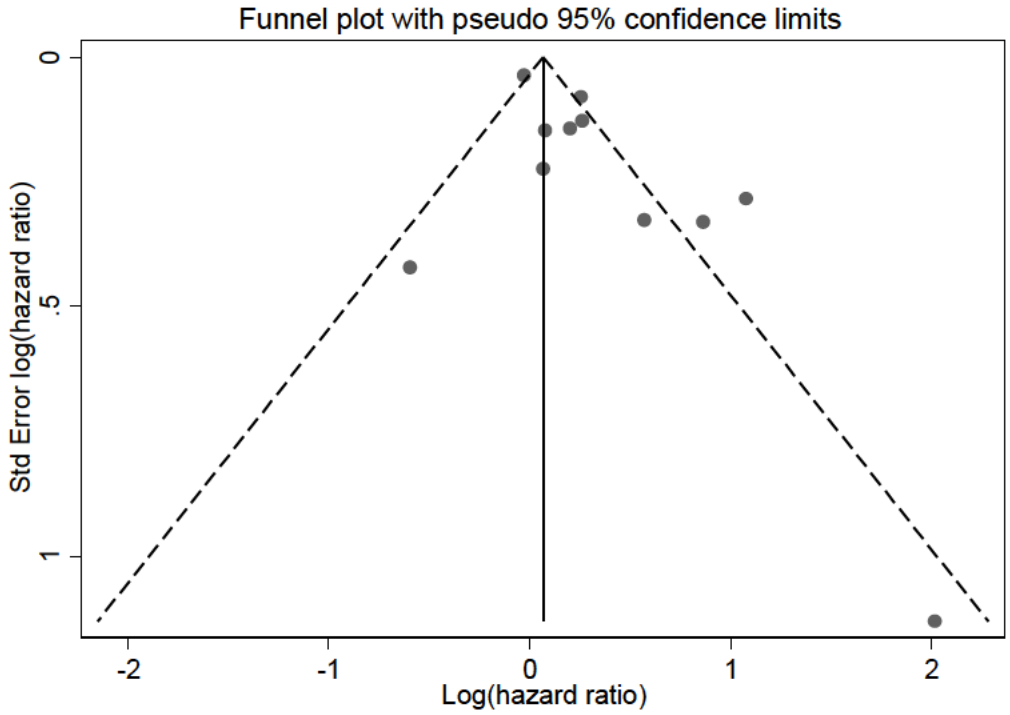


Five of OSCC HRs and one of the OAC HRs were adjusted for other predictors and potential confounders. Sensitivity analysis found that heterogeneity among these five adjusted OSCC studies remained at I^2=75.7% (HR=1.69, 95% CI 0.95,3.02). Similarly, both Asian OSCC studies (6 studies, I^2=73.8%, HR 1.32, 95% CI 1.01,1.71) and non-Asian OSCC studies (3 studies, I^2=81.6%, HR 1.48, 95% CI 0.60,363) recorded high heterogeneity.

There was visual evidence of publication bias (Figure 2.7) confirmed by Egger's test p=0.015. The outlier at the bottom right of the funnel plot came from Iran (Samadi et al., 2007) where alcohol consumption is illegal for the Muslim majority. Just 4 of 222

respondents in this study fell into the alcohol 'ever' group. However, removing the outlier had little impact on overall heterogeneity (I^2=79.7%, Egger's test p=0.030).

Figure 2.7 Funnel plot of results for ever against never alcohol use pre-diagnosis.

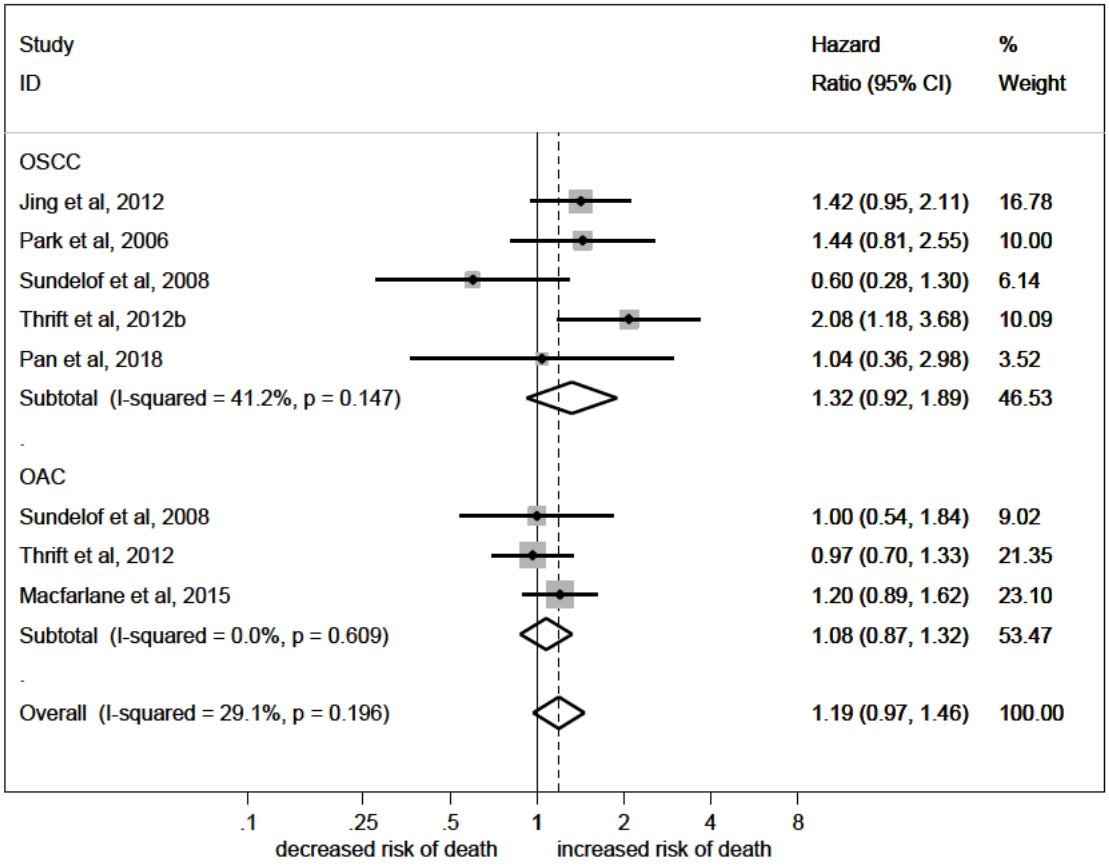


Seven studies recorded alcohol consumption in categories. One (Y. P. Pan et al., 2018) only specified frequency, at ≥4 times per week, without quantity. All others were converted into grams of alcohol per week by equating one standard drink with 10 grams of alcohol. The highest consumption categories varied between ≥50 grams per week (g/w) (Dandara et al., 2016), ≥70 g/w (Sundelöf et al., 2008), ≥124 g/w (S. M. Park et al., 2006), ≥140 g/w for women and ≥210 g/w for men (Macfarlane et al., 2015), ≥230 g/w (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012), ≥230 g/w (Shitara et al., 2010), ≥250 g/w (Jing et al., 2012).

Comparing the highest weekly alcohol consumption with the lowest (Figure 2.8) did not produce statistically significant evidence of effect (HR 1.19, 95% 0.97,1.46). Two studies were excluded from this meta-analysis: one (Dandara et al., 2016) did not report standard

deviations or confidence intervals and the other (Shitara et al., 2010) used a reference category which included up to 230 g/w.

Figure 2.8 Forest plot of risk of death for highest pre-diagnosis alcohol consumption category against lowest by cancer type.

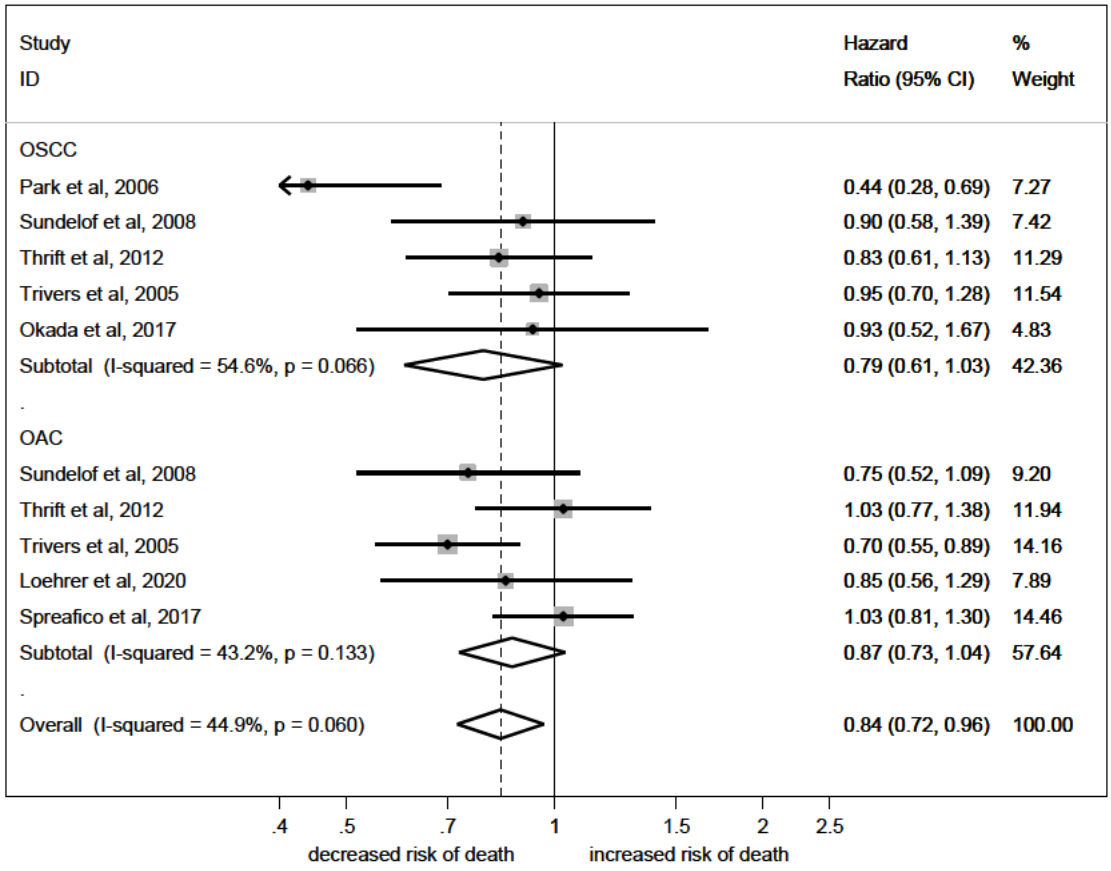


Body Mass Index

Ten articles (from nine studies) reported on Body Mass Index (BMI). For eight articles the BMI category $\geq 25 kg/m^2$ 'overweight and obese' was either provided or was created by combining across categories. One study (He et al., 2020) reported the alternate $\geq 23 kg/m^2$ cutpoint sometimes used for Asian populations (Hubbard, 2000) and another had a highest BMI category of $\geq 20\ kg/m^2$ (Salek et al., 2009). The reference category was variously defined as $<18.5 kg/m^2$ (He et al., 2020),$<20\ kg/m^2$ (Salek et al., 2009), $<23 kg/m^2$

(S. M. Park et al., 2006), <25kg/m^2 (Trivers et al., 2005), 18.5 to <25kg/m^2 (Loehrer et al., 2020; Okada et al., 2017; A. Spreafico et al., 2017; Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012) or 22 to <25kg/m^2 (Sundelöf et al., 2008). The first two studies, comparing against underweight, were excluded from the further analysis. For the other studies, risk of death was compared for those with 'normal' BMI against those who were 'overweight or obese' and the pooled results are presented in Figure 2.9.

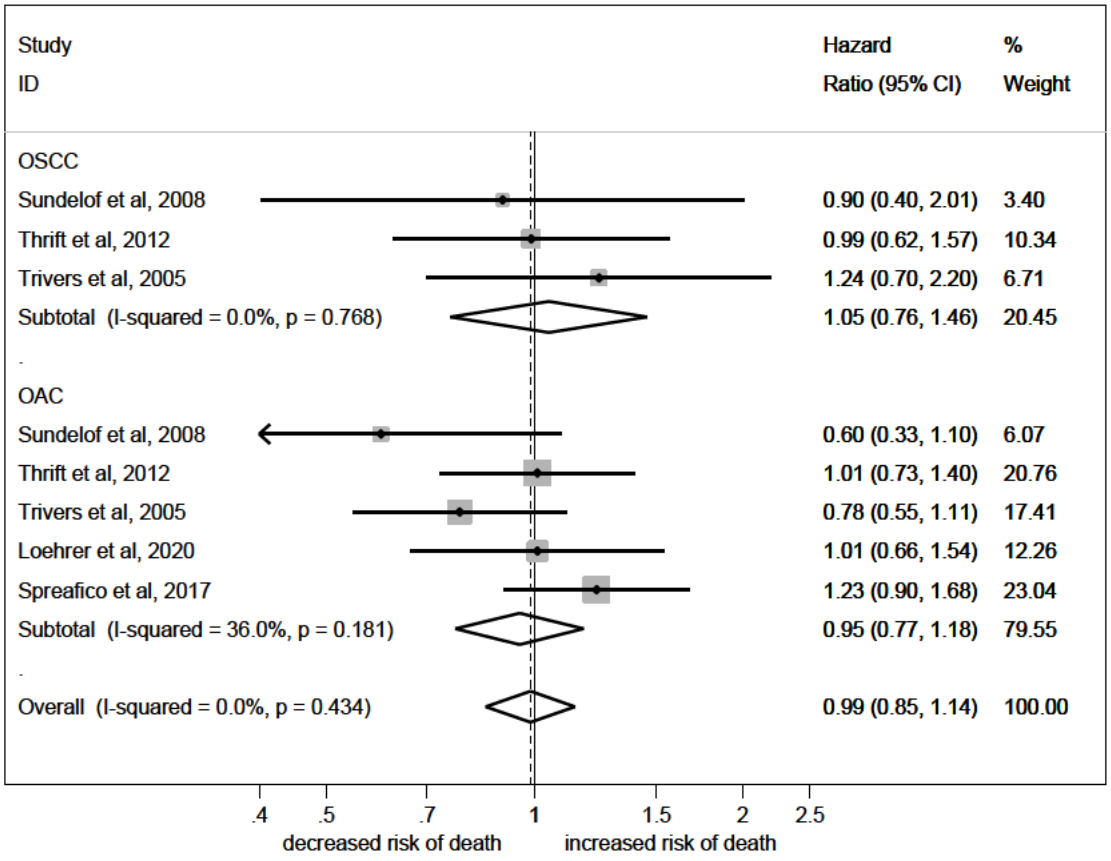
Figure 2.9 Forest plot of risk of death for overweight or obese versus 'normal' pre-diagnosis BMI by cancer type.



Results are suggestive that overweight or obesity decreases risk of death by 16% in OC, with little evidence for difference between OSCC and OAC, but there is substantial heterogeneity in estimates across studies (HR 0.84, 95%CI 0.72,0.96, I2=44.9%). In sensitivity analysis, removing the single study with unadjusted HRs (Trivers et al., 2005) had

little impact on the pooled result (HR 0.81, 95%CI 0.66,0.98, I2=43.0%). Removing the two Asian studies (Okada et al., 2017; S. M. Park et al., 2006) from OSCC reduced heterogeneity but also reduced the strength of the association (HR=0.89, 95% CI 0.73-1.08, I^2=0%). No statistically significant decreased risk was found for obesity alone (BMI≥30, Figure 2.10, OC HR 0.99, 95%CI 0.85,1.14).

Figure 2.10 Forest plot of risk of death for obese against 'normal' pre-diagnosis BMI by cancer type.



Regular physical activity

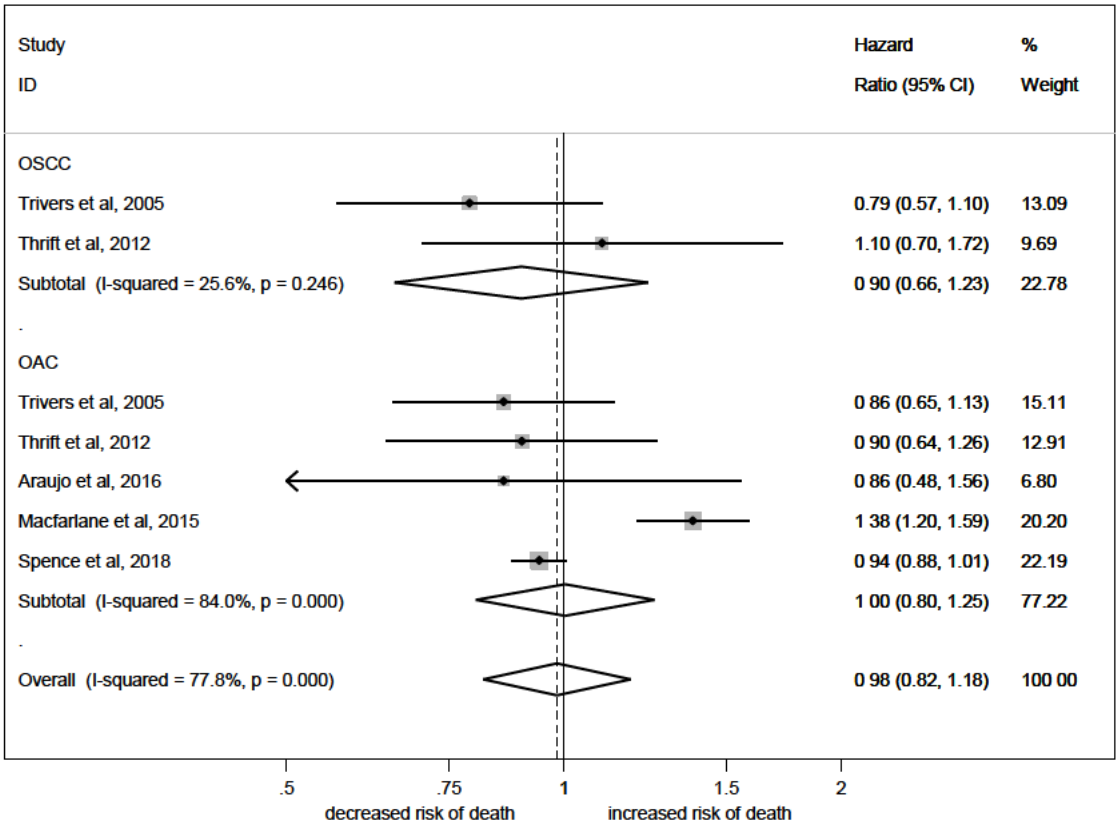
Only three articles reported on the association between physical activity and risk of death from oesophageal cancer. Physical activity categories were reported as '1st low', '2nd', '3rd', and '4th high' (Sundelöf et al., 2008) or 'low', medium', and 'high' (Thrift, Nagle, Fahey, Smithers, et al., 2012) and 'No habit', '1-2 times/week', '≥3 times/week', and 'unknown' (Okada et al., 2017). These categories were sometimes defined based on a series of questions which were not explained in full. Comparing the highest physical activity category to the lowest, the pooled hazard ratio was 1.07 (95% CI 0.87,1.32) with little difference between OSCC (2 studies, HR 0.92, 95% CI 0.67, 1.27) and OAC (2 studies, HR 1.20, 95% CI 0.91,1.58). There was no statistically significant evidence of association.

Regular NSAID consumption

Six articles (from five studies) reported regular NSAID consumption. Three articles looked at all NSAIDs combined (Araujo et al., 2016; Macfarlane et al., 2015; Spence et al., 2018; Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012; Trivers et al., 2005) and three presented results specific to aspirin (Araujo et al., 2016; Macfarlane et al., 2015; Spence et al., 2018; Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012; Trivers et al., 2005). NSAID use was defined as '≥1 tablet per week for ≥6 months' or not (Trivers et al., 2005), 'never', 'less than weekly' or 'at least weekly' in the 5 years prior to diagnosis (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012), evidence of at least one prescription (Macfarlane et al., 2015; Spence et al., 2018,) or incompletely defined as 'no' or 'yes' (Araujo et al., 2016).

Figure 2.11 presents a forest plot of the hazard ratios of the highest NSAID use category compared to the lowest. The pooled hazard ratio was 0.98 (95%CI 0.82,1.18) with high heterogeneity (I^2=77.8%) and little difference between OSCC (HR 0.90, 95%CI 0.66,1.23, I^2=25.6%) and OAC (HR 1.00, 95%CI 1.80,1.25, I^2=84.0%).

Figure 2.11 Forest plot of risk of death for regular NSAID use by cancer type.



2.2.6 **Discussion**

The aim in this paper was to systematically review and quantify the risk of death from oesophageal cancers associated with pre-diagnosis behaviour. The review located 27 eligible articles from 26 studies and found that OSCC cases with the highest lifetime smoking exposure had a 55% increased risk of death relative to non-smokers. OC cases who had ever consumed alcohol had a 30% increased risk of death relative to those who had not, but results were heterogeneous and displayed evidence of publication bias. Being overweight or obese pre-diagnosis was found to decrease risk of death (HR 0.72, 95%CI 0.72,0.96) in OC but not obese alone (HR 0.99, 95%CI 0.85,1.14). Fewer studies investigated the impact of pre-diagnosis physical activity levels and results from regular NSAID consumption displayed considerable heterogeneity. Although the high risk of bias and high heterogeneity between studies weakens the interpretation, the current results suggest that there are some associations between pre-diagnosis health behaviours and post-diagnosis risk of death in OC.

In this study we have found a statistically significant hazard ratios of 1.08 (95% CI 1.00,1.07) for ever smoked compared to never smoked and 1.55 (95% CI 1.24, 1.94) for highest smoking category versus lowest in OSCC. These results are broadly consistent with a 2016 meta-analysis of the relationship between smoking (not restricted to pre-diagnosis) and OC survival time found a pooled HR of 1.41 (95% CI 1.22,1.64) based on just 5 studies of mainly OSCC patients (Kuang et al., 2016) and a 2017 meta-analysis reported HRs of 1.07 (95% CI 0.86-1.32) for ever smoking in OSCC, 1.41 (95% CI 0.96,2.09) for current smoking in OSCC and 0.99 (95% CI 0.73,1.36) for current smoking in OAC based on pooling 8, 2 and 2 papers respectively (McMenamin, McCain, & Kunzmann, 2017).

It is also consistent with findings from other similar patient groups. In stomach cancer for example, a meta analysis (Ferronha, Bastos, & Lunet, 2012) found increased risks of death of 1.08 (95% CI 0.90-1.30) for smokers versus never smokers. In colorectal cancer a

meta-analysis of six studies (Walter, Jansen, Hoffmeister, & Brenner, 2014) found that smoking increases risk of death by 1.26 (95% CI 1.15-1.37) although this was not confined to pre-diagnosis behaviour and included some cohorts selected according to treatment (such as surgical only).

The lower risk for ever smoking is consistent with a scenario in which ever smoking includes a large proportion of former smokers and these former smokers have much lower risks of death than current smokers.

In relation to alcohol consumption, this review found a 30% increase in risk of death associated with alcohol consumption (HR 1.30, 95%CI 1.09,1.56) and a non-statistically significant HRs of 1.19 when comparing the highest to lowest alcohol consumption categories. Point estimates for OSCC (HRs of 1.43 and 1.32) were higher than for OAC (HRs of 1.22 and 1.08) in both situations. These results are consistent with previous studies. A recent meta-analysis comparing highest to lowest alcohol consumption patterns in OC found a statistically significant pooled HRs for overall OC (HR=1.48, 95% CI 1.19,1.84) and OSCC (HR=1.26, 95% CI, 1.01,1.60) but no corresponding effect for OAC (HR=1.01, 95% CI 0.70,1.47) based on 3, 9 and 3 papers respectively (L.-P. Sun et al., 2020). A meta-analysis of stomach cancer (Ferronha et al., 2012) found increased risks of death 1.13 (95% CI 1.00-1.28) for drinkers versus never drinkers.

For overweight and obese, this study found that pre-diagnosis overweight and obese were associated with decreased risk of death in OC (HR=0.84, 95%CI 0.72,0.96) with little evidence of difference between OSCC and OAC. However, obese alone was not protective (HR=0.99, 95% CI 0.85,1.14). Perhaps overweight provides some buffer against the wasting nature of OC, but obese introduces new complications and comorbidities.

In the wider literature, most interest has focussed on the relationship between obesity at surgery and post-surgical survival times. For example, a 2017 meta-analysis reported post-surgical pooled HRs for obesity to be 1.05 (95% CI 0.80, 1.39) for 6 studies of OSCC patients and 0.91 (95% CI 0.81,1.02) for 7 studies of combined OC cases (Gao et al., 2018). The single OAC study had HR=0.95 (95% CI 0.85,1.07). But earlier, a meta-analysis of post-surgical risk for oesophageal cancer reported an aggregated risk of death of 0.79 (95% CI 0.65-0.95) for overweight (Hong et al., 2013). However, these results may be quite different from the research question of this thesis as those receiving surgical treatment are a sub-group of all OC and the criteria for who is offered surgery could influence the results.

In contrast, for colorectal cancer, a systematic review of five studies (E. Parkin, O'Reilly, Sherlock, Manoharan, & Renehan, 2014) concluded that pre-diagnosis obesity conferred higher risks of death, especially in women, although these observations were not quantified through meta-analysis.

Several recent studies of treatment outcomes have found that weight loss prior to diagnosis is a strong predictor of survival (Yu et al., 2018) particularly for early stage cancers (Shen et al., 2017). So, it is possible that individuals with a history of overweight or obesity may not be detected due to pre-diagnosis weight loss.

In this review we found just 3 studies which looked at the relationship between pre-diagnosis physical activity and post-diagnosis survival in OC with no statistical evidence of association. While little has been published about the relationship between pre-diagnosis physical activity and post-diagnosis survival in OC, physical activity has been found to be protective of the incidence of OAC (RR=0.79, 95% CI 0.66,0.94), but not OSCC (McTiernan et al., 2019).

Finally, this review found no evidence of association between regular consumption of NSAIDs (including aspirin) prior to diagnosis and post-diagnosis survival time in OC (HR 0.98, 95% CI 0.82,1.18). Again, while little is known about survival in OC, a recent meta-analysis has confirmed that regular aspirin consumption is protective of OC incidence (RR=0.75, 95% CI 0.65,0.86) (Qiao et al., 2018).

Our review was conducted in compliance with the MOOSE guidelines, using multiple search engines and independent reviewers. The inclusion and exclusion criteria were carefully defined to avoid potential biases which could arise from particular medical conditions and/or the decision whether or not to treat. Appropriate statistical methods and interpretation have been used throughout. The key limitations on the interpretability of the current results are the high risk of bias and the heterogeneity of studies. Fifteen of 26 studies found were from Asia – which could be seen as appropriate given that Asia accounts for about 75% of oesophageal cancers incidence (Ferlay et al., 2015). However, it is known that behaviours and other risk factors differ widely across continents (Hongo et al., 2009) and there are considerable differences in health services between countries. Data were collected at different time points (before, at and after diagnosis) and asked about behaviour at different time points (lifetime behaviours, prior to diagnosis and at diagnosis). Further, there was little standardisation in the measurements used. Smoking was categorized in eleven different ways, alcohol in eight and most studies reporting BMI used slightly different categories. Finally, studies varied markedly on how they adjusted for other factors which affect survival time; with nine articles included no adjustment at all.

There is also considerable potential for measurement and recall bias with all studies relying on self-reported behaviour and the majority requiring recall of behaviour a year or more prior to interview. However, as over- and under-reporting of behaviour is unlikely to be

associated with survival time these biases may change the estimated size of relationships but are unlikely to create any spurious relationships.

The current review includes more studies than other reviews in this general topic area, but have tended to find somewhat smaller associations (HRs closer to 1.0) than some of the other published meta-analyses. One possible explanation is that these reviews include a greater proportion of poorer quality studies (such as results which are unadjusted for possible confounders). Alternatively, other reviews often allowed inclusion of post-diagnosis behaviours, and it may be that pre-diagnosis behaviour has a genuinely smaller association than post-diagnosis behaviour.

In summary, conducting these reviews has provided some confidence that pre-diagnosis smoking and pre-diagnosis alcohol are associated with higher risk of death in OSCC and that pre-diagnosis physical activity is associated with lower risk of death. All other associations seem subject to doubt due to insufficient information, concerns over the quality of the reported results and unexplained inconsistencies in the reported results.

Our review has suggested a need for standardisation of measurement categories and adjustment for other risk predictors. Retrospective studies seeking to record pre-diagnosis behaviour after diagnosis should be discouraged as they are subject to recall and survivor bias. Prospective studies are necessarily large, time consuming and expensive. Record linkage between behaviour cohorts and cancer registry data sets may be required.

This review has found that a number of modifiable behaviours known to be associated with risk of oesophageal cancer also have a carryover effect on risk of death after diagnosis. Successfully addressing these risk behaviours may have the dual benefits of both decreasing incidence and decreasing risk of death after diagnosis.

2.3 Critique of the method

The MOOSE and PRISMA guidelines were followed in the design as well as the reporting of the review. In the original published paper, the main omission in the review processes was the rating of the scientific quality of the study methods of each of the studies. This was justified by questioning the measurement of quality: what constitutes quality and how can it be accurately measured? Are quality scales sufficiently sensitive? A 1999 paper by Juni et al (Jüni et al., 1999) was cited to support this argument.

In retrospect, this was out-of-date. Whereas in 1999-2000 only 22% of systematic reviews of observational studies incorporated a quality assessment component each having to devise their own assessment criteria, by 2003-2004 there were 10 different quality rating scales available and 50% of systematic reviews of observational studies incorporated quality assessment (Mallen, Peat, & Croft, 2006). There was growing agreement on the need for quality assessment. *"'Quality' is an amorphous concept. A convenient interpretation is 'susceptibility to bias', …" (Sanderson et al., 2007)*. The current update assessed risk of bias by applying the JBI Checklist for Cohort Studies.

2.4 Closing comments

This Chapter addressed the first research aim of this thesis by describing current knowledge of the association between pre-diagnosis health behaviours and post-diagnosis survival times for OC. To obtain as complete and accurate answer as possible a structured approach was followed consistent with best practice methods recommended by the Cochrane Collaboration (Higgins et al., 2019). A larger number of published articles were included in the review than other systematic reviews and meta-analyses in this topic area. Despite this,

insufficient information was found in the existing published literature to confidently address the thesis aim.

In the context of the overall thesis, the reviews in this Chapter have provided some hints and reference points as to what associations should be expected in the subsequent analyses (for example, smoking and alcohol consumption should be expected to be associated with decreased survival in OSCC (see Chapters 4,5 and 6). More often, the systematic review has highlighted the gaps and weaknesses in current knowledge, confirming the need for new methods.

Following Chapters address the second aim of this thesis, namely to develop, describe and evaluate new methods to augment cancer registry data with measures of pre-diagnosis behaviour. Chapter 3 presents a review of the data sets, variables and conceptual approach to analyses, before the outcome of analyses are described in Chapter 4, 5 and 6.

Chapter 3 Data sets, variables, and missing variables

3.1 Introduction

The first aim of this thesis is to describe the association between pre-diagnosis health behaviours and post-diagnosis survival times in OC. In Chapter 2 a systematic review of the literature was conducted to address this question. While this review provided evidence and indications of some associations between behaviour and survival time in OC, the final conclusions were that current information is sparse, and sometimes contradictory, and that more research is required to address the study aim.

The second aim of this thesis is to develop, describe and evaluate a new method to augment cancer registry data with external data on health behaviour when one-to-one data linkage is not available. Conceptually, the fact that behaviour differs across demographic groups, implying that demographic characteristics should convey a small quantum of information about health behaviour shall be leveraged. It is anticipated that large data sets and informative demographic variables will be needed to convey sufficient information from the external health behaviour data set to the cancer registry data base.

In the discussions below, cancer registry data for OC patients are sought with a corresponding external source of health behaviour data. Desirable characteristics of both data sets are to be large, have low error rates and to have a range of potentially informative demographic measures in common.

Cancer registries can be characterised by their scope and location. Population-based cancer registries record data from all cancer cases in a defined population (usually a geographic area such as a country or state). They are used to monitor cancer rates, guide

health service planning and needs analysis, and undertake epidemiological and other cancer research. Hospital-based cancer registries record all cancer cases within that institution and help improve patient care and needs assessment at that hospital. Special cancer registries collect data on a particular type of cancer and help research and planning specific to the needs of that particular cancer (National Cancer Institute Surveillance Epidemiology and End Results Program, 2021). To meet the need for large data sets with good data quality, this research focusses on large, population-based cancer registries from selected developed countries.

In Section 3.2 possible cancer registry, health survey pairs are reviewed in Australia, UK and the USA. Section 3.3 critiques the measures of pre-diagnosis behaviour available for analysis. Section 3.4 explains the data sets selected for analysis in this thesis. Section 3.5 commences consideration of how information may be transferred between the selected data sets. The thesis Appendix shows examples of the R code used to extract and format the data sets. Section 3.6 reviews the populations and measures which have been selected for analysis within the context of the wider theses.

3.2 Cancer registry and health survey combinations

The choice of data sets in this research was driven by a) data accessibility and b) optimising the range of demographic and health behaviour data.

The choice of data sets was not necessarily driven by the later representativeness and generalisability of the results. The priority was to see if interpretable results could be derived under the best possible conditions. Hence, the cancer registries and health surveys reviewed below are generally from developed, English speaking countries: with particular focus on Australia; England; and the USA.

3.2.1 **Australia**

In Australia, all cancers cases, except some skin cancers, are reported to the relevant State or Territory population-based cancer registry by hospitals, pathology laboratories, radiotherapy centres and registries of birth, deaths, and marriages. Data from the eight State and Territory registries are pooled by the Australian Institute for Health and Welfare to form the Australian Cancer Database (Australian Institute of Health and Welfare, 2018). The Australian Cancer Database contains data on all reported primary cancers in Australia since January 1, 1982.

While there were only an estimated 1,587 new OC cases in Australia in 2020, aggregating across years can generate large sample sizes. For example, the aggregate number of new OC cases is 27,153 from 2001 to 2020 or 43,073 from 1982 to 2021 (Australian Institute of Health and Welfare, 2020a).

The largest corresponding health survey in Australia is the National Health Survey. The National Health Survey is conducted irregularly with surveys in 1989-90, 1995, 2001, 2004-05, 2007-08, 2011-12, 2014-15, 2017-18 and 2021 (in progress). The survey enrols about 20,000 people at each iteration. For example, in 2017-18 data were collected from 21,315 people of whom 16,049 were 20 or more years of age (Australian Bureau of Statistics, 2021d).

The National Health Survey collects many measures of health behaviour relevant to the current study. For example, the 2017-18 survey the interviewer administered questionnaire included questions on tobacco consumption (focussing on days smoker per week, amount smoked, and age commenced and quit regular smoking), alcohol consumption (focussing on last time consumed alcohol, days in last week drank alcohol, details of alcohol consumed ay most recent episode and usual alcohol consumption), and exercise (focussing on

walking, moderate to vigorous exercise, and strength building activity), as well as interviewer measured height, weight and waist circumference (Australian Bureau of Statistics, 2021a).

The demographic measures in common between the Australian Cancer Database and the National Health Survey are sex, age, country of birth and state/territory of usual residence (Australian Bureau of Statistics; Australian Institute of Health and Welfare, 2018).

However, to access unit record data from Australian Cancer Database, would typically require Ethics Committee approval from each of the relevant data custodians (that is, each of the eight State and Territory cancer registries) and the Australian Institute of Health and Welfare usually charges a data extraction fee (Australian Institute of Health and Welfare, 2018). Similarly, a State specific data set could be obtained by approaching a State or Territory Cancer Registry directly, obtaining (just one) ethics approval and paying a data extraction fee (Cancer Institute NSW, 2021). In contrast, de-identified unit record data from the National Health Survey data is provided free of charge to Australian universities through the Australian Bureau of Statistics' Microdata Download facility (Australian Bureau of Statistics, 2021c) as negotiated through the ABS/Universities Australia access agreement (Australian Bureau of Statistics, 2021b).

3.2.2 **England**

In England in 2017 there were 7,569 new OC cases recorded with a 5-year cumulative 37,300 OC cases from 2013 to 2017 (Office for National Statistics, 2021a).

The National Cancer Registration and Analysis Service (NCRAS), within Public Health England, compiles data from more than 500 local and regional datasets into a comprehensive population-based cancer registry. This registry compiles cancer cases'

demographic details, cancer presentation details and staging, comorbidities, treatment, and survival measures (but not pre-diagnosis behaviour). Requests to use the data set must be submitted to Public Health England via the Office of Data Release. The application process requires submission of a Data Request Form, a project protocol specifying how the data will be used, details of the data items requested and how the data set will be secured. If approved, researchers are required to pay a data retrieval fee (GOV.UK, 2021).

The Health Survey for England has been conducted annually since 1991 enrolling about 8,000 adults and 2,000 children per year (NHS Digital, 2021). Behavioural measures include smoking (ever smoked, current smoking, starting and quitting age, number of cigarettes per day, how soon after waking), alcohol consumption (age of first drink, current drinker, drank in past 12 months, frequency drank in past 12 months, drinking diary for past 7 days, drinking recall for past 12 months, drinking more than 5 years ago), physical activity (last 7 days diary of walking, moderate activity, vigorous activity and sedentary activity) and measured health and weight.

Access to the data set is available through the UK Data Service, a government funded service administering access to over 850 data set (UK Data Service, 2021). The UK Data Service maintains three levels of data access: ‘Open’, ‘Safeguarded’ and ‘Controlled’. The Health Survey for England data files have ‘Safeguarded’ access – which effectively means the UK Data Service will keep a record of who is using the data files and, possibly, for what reason. While access is automatic for researchers in many countries (e.g. European Economic Area, Argentina, New Zealand), Australia is not on the list of countries having “adequate level of data protection”. Australian researchers must seek approval directly from the owners of each data set which they wish to use.

Demographic variables in common between the National Cancer Registry and the Health Survey for England are limited to age (5-year groups), sex, and ethnic background. The cancer registry does not record marital status and the health survey only codes location of residence as rural versus not rural.

3.2.3 **United States of America**

An estimated 19,260 new OC cases will be recorded in the US in 2021, contributing to an estimated 5-year cumulative 89,580 new OC cases from 2017 to 2021 (American Cancer Society, 2021a).

In the US, each State, the District of Columbia and most territories (such as Puerto Rico, US Virgin Islands) run individual population-based cancer registries. Access to data from these individual registries is highly variable, with different approval processes, different patient and physician consent requirements, presence or absence of data processing fees, and different time delays in accessing data. (CDC Cancer Registry Data Access for Research Project, 2018).

However, all US cancer registries contribute agreed data to one or both of two national programs. The Surveillance, Epidemiology, and End Results (SEER) program arose from the National Cancer Act of 1971. It is administered by the National Cancer Institute and has collated cancer registries' data since 1973. Currently it compiles data from 21 geographic areas, covering about 28% of the US population (National Cancer Institute, 2021d). The National Program of Cancer Registries (NPCR) was established by US Congress in 1992 and is administered by Centres for Disease Control and Prevention. The NPCR currently supports cancer registries in 46 States plus District of Columbia and Territories (Centers for Disease Control and Prevention, 2021c). In combination SEER and NPCR cover the entire US.

The NPCR's agreement with contributing States does not allow the release of de-identified unit record data for research without registry permission. Summary statistics (data cubes) are publicly available (U.S. Cancer Statistics Working Group, 2020). In contrast, at the commencement of the current project, the SEER program was providing relatively open access to unit record data for research purposes, with no fee. Upon registering and signing a Data Use Agreement, researchers were able to access the 'case listing' data through SEER's data access software package SEER*STAT (Surveillance Research Program, 2021). In the 10 years from 2006 to 2015 the SEER data set recorded 39,233 new OC cases.

Most recently, the NPCR have provided access to their cancer registry data through the SEER*STAT facility. However, as not all States have provided permission for the release of unit record data, the 'case listing' option of SEER*STAT has now been disabled.

A number of national health surveys run in the US. Firstly, the National Health Interview Survey (NHIS) is administered by the Centers for Disease Control and Prevention's National Center for Health Statistics. This survey, which has run since 1957, uses a complex multi-stage sampling scheme, to select households for face-to-face interviews. The design and content of this survey is updated about every 10 to 15 years. The most recent period of stability was from 1997 to 2018. During this period, the adult version of the questionnaire recorded data on smoking (ever smoked 100 cigarettes, starting and quitting age for regular smoking, whether currently smoke every day or some days, and number of cigarettes per day), alcohol consumption (ever had 12 drinks in a year, 12 drinks in entire life, average number of drinks on days when did drink, binge drinking in the past year, drinking diary and binge drinking in the past 30 days), leisure-time physical activity (how often vigorous activity, how long are the vigorous sessions, how often light or moderate activity, how long are the light to moderate sessions, resistance training) and self-reported height and weight (National Center for Health Statistics, 2021a).

The number of adult respondents to the survey in 2018 was 25,417 and the 5-year cumulative total of adult respondents is 152,556 from 2014 to 2018 (National Center for Health Statistics, 2021a)

In 2019, the National Health Interview Survey underwent a significant redesign. As well as changes in the individual data items, questions on alcohol and physical activity were relegated to a 'rotating core' which means these questions will no longer be asked every year and may be as infrequent as every third year. (National Center for Health Statistics, 2021a)

A second national survey is the National Health and Nutrition Examination Survey (NHANES). This combines in-person interviews with standardised physical examinations and laboratory tests. This survey is also administered by the Centers for Disease Control and Prevention's National Center for Health Statistics and also uses a complex 4-stage sample design, randomly selecting counties from across the US. With a history back to 1959, The NHANES has only become a continuous survey since 1999. It runs with a 2-year cycle with a target sample size of approximately 10,000 per cycle. For example, in the 2017-18 cycle the interviewed sample was 9,254 of whom 8,704 completed the physical examination. Of these 5,569 and 5,265 were 20 or more years of age (National Center for Health Statistics, 2021b).

The National Health and Nutrition Examination Survey collects data on smoking (ever smoke 100 cigarettes, starting and quitting age, whether currently smoke every day or some days, and number of cigarettes per day, how soon smoke after wake up, type of cigarette), physical activity (how often vigorous activity, how long are the vigorous sessions, how often moderate activity, how long are the moderate sessions, walking and bike riding and sport), height and weight (now and 1 year ago). There are no questions on alcohol use (National Center for Health Statistics, 2021b).

Finally, the Behavioural Risk Factor Surveillance System is a telephone survey which has been conducted annually since 1984 and has been a nationwide survey since 1993. BRFSS is run as a separate random sample in each State and Territory. In 2018 sample sizes ranged from 2,758 in Alaska to 35,767 in New York State. A mobile phone survey was first included in 2008. Behavioural measures in the BRFSS include self-reported smoking (ever smoked 100 cigarettes, current smoke every day or some days, time since quit smoking), alcohol (any drink in past 30 days, days per week of drinking in past 30 days, drinks per day when drinking, in last 30 days ever more than 5 (males) or 4 (females) drinks on one occasion, last 30 days largest number of drinks on one occasion), physical activity (any physical activity other than regular job, plus other questions in some years) and height and weight. The 2019 BRFSS data set contains 418,268 data records, with a cumulative 5-year total of 2,233,479 data records for 2015-2019 (Centers for Disease Control and Prevention, 2021a).

For all 3 health surveys, de-identified unit-record data sets are available for download, without registration or fee, from webpages maintained by the Centers for Disease Control and Prevention.

Demographic variables in the SEER cancer registry data set included age, sex, marital status, race and State of residence. Each of these variables were also available in BRFSS data set. The NHANES and NHIS do not release the State of residence: the multi-stage sampling used in these surveys meant that participants would be drawn from relatively few States.

3.2.5 **Other countries**

Many developed countries maintain equivalent cancer registries and national health surveys. However, each country's data systems have unique idiosyncrasies and limitations which would need to be considered when applying the methods in this thesis. Here are further brief examples Canada, South Africa and New Zealand as to what information may and may not be available.

In Canada, the National Cancer Registry is a central compilation of data sets provided from 13 separate province and territories cancer registries (Statistics Canada, 2022a). Each province has mandatory data collection. But there are some weaknesses in the carry over to the National Cancer Registry. For example, new cancers are not available from Quebec after 2010 and the variation in procedures between provinces and delays in reporting results in some undercounting (Statistics Canada, 2022a). More importantly for the methods in this thesis, the only demographic variables passed to the National Cancer Registry are age and sex (Statistics Canada, 2022a).

The Canadian Community Health Survey has roots back to 1991 but has been run in its current form since 2007. It has run continually since 2007 enrolling around 65,000 respondents per year. Participation is voluntary and data are compiled into data sets every two years (Statistics Canada, 2022b). Data are self-reported with respondents answer questions online (or via telephone if required). Survey modules vary from year to year but questions on smoking (and e-cigarettes more recently), alcohol use, physical activity, NSAID use, sleep, weight and height, and eating habits are often included. The survey also includes a range of demographic variables including sex, age, main daily activity (employment, school, etc), place of birth, population group, language group, food security and income (Statistics Canada, 2022b), most of which are not matched in the National Cancer Registry.

Data files can be freely downloaded from https://www150.statcan.gc.ca/n1/en/catalogue/82M0013X.

In South Africa, the South Africa National Cancer Registry is pathology-based rather than population-based. As data are only collected for pathology laboratories those who were diagnosed or died without pathological confirmation are not recorded and this cancer registry represents an undercount of total cancers in the country (CANSA, 2022). Annual summary tables are provided online (CANSA, 2022), but there is no formal mechanism for external researchers to apply to access unit record data.

South Africa conducts national health surveys from time to time. The most recent is the South Africa Demographic and Health Survey (SADHS) conducted in 2016 with 11,083 households interviewed (National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), & ICF, 2019). Again, there does not appear to be any formal mechanism for external researchers to apply to access unit record data.

The population of New Zealand is less than 2% of the population of the USA and around less than 20% of the population of Australia (Worldometer, 2022) leading to smaller data sets.

The NZ Cancer Registry covers the entire country and reporting is mandatory under the Cancer Registry Act 1993 (New Zealand Legislation, 2021). Primary cancer site, date of diagnosis, date of death, age, sex, ethnicity, and census area of residence are all recorded. Stage of disease is recorded using the SEER summary staging rather than the AJCC staging system (National Cancer Institute, 2021c). The continuous National Health Survey collects data from around to 10,000 adults per year. It includes objective measurement of height and weight as well as self-reported tobacco use, alcohol use, physical activity, dietary habits, age,

gender, ethnicity and possibly deprivation index could be used for matching. (Manatu Hauora Ministry of Health, 2020). Other demographic variables such as education, employment status, income and housing cannot be used for matching as they do not appear in the Cancer Registry data set.

The New Zealand Health Survey has been a continuous survey since 2011 and aims to enrol around 14,000 adults each year, oversampling Maori, Pacific Islander and Asian ethnic groups. The survey includes objective measurement of height, weight, waist circumference and blood pressure as well as self-reported tobacco use, alcohol use, physical activity, dietary habits, body size, age, gender, ethnicity, education, employment status, income, housing and household composition, area deprivation and rurality. as well as self-reported behaviours and has a 77% response rate (Manatu Hauora Ministry of Health, 2021).

In New Zealand, access to unit record files for both the cancer registry data sets and national health survey data sets are restricted to people physically located in New Zealand. These data are provided free of charge to approved research projects (Manatu Hauora Ministry of Health, 2020). NZ has an integrated data infrastructure which means that the government links the data with these linked data similarly available for approved projects.

3.3 Behaviour measures

The core data items compiled in cancer registry data sets are well standardised (D. M. Parkin, 2006) and there is not much ambiguity of which data items are used to measure diagnosis and which are used to describe survival status and survival time. But deciding how to best measure health behaviour is more complex.

In this project health behaviour data is obtained from national health surveys; because national health surveys cumulate across years to produce large data sets which are often readily accessible for research purposes. The primary role of health surveys is to estimate prevalence of health behaviour in the population to support public health intervention and policy (D. M. Parkin, 2006). But, when investigating the impact of pre-diagnosis health behaviour on post-diagnosis survival, measures of cumulative exposure, the stage of life of exposure and/or the intensity of exposure may also be important predictors.

However, by conducting secondary analyses on existing data sets, investigations were restricted by the study design and variable definitions inherent in these surveys. For example, most national health surveys record participants' self-report of health behaviour rather than more objective direct observation. Further, national health surveys will tend towards a wide scope of questions in preference to detailed exploration of individual health behaviours. This could restrict the number of available data items for any given behaviour. Finally, as it may be necessary to cumulate data sets over time to create larger sample sizes, the consistency of the data items definitions and inclusion over repeated surveys needed to be considered.

3.3.1 **Tobacco smoking**

It has been proposed that smoking causes OC via tobacco carcinogens (particularly nitrosamines) coming into contact with the oesophageal mucosa (Zhang, 2013). No specific mechanism has been proposed to explain any potential link between pre-diagnosis smoking and post-diagnosis survival in OC.

Tobacco smokers display quite different smoking behaviours. Some, for example, inhale more deeply and/or more often than others, altering the actual dose per cigarette. Analyses on the sample of OC patients in the geographic catchment of the cancer registries in the study years will only describe the "average" tobacco smoker.

Aspects of smoking which are most likely to be of interest to researchers, policy makers and health promotion workers using the results of this thesis are a) whether the OC case did smoke pre-diagnosis, b) how many cigarettes they smoked per week and c) how many years they smoked.

Self-reported smoking status and smoking frequency likely underestimates true smoking status and true smoking frequency, and that the level of under-reporting differs according to social circumstances (International Agency for Research on Cancer, 2008). For example, patients in treatment may feel uncomfortable disclosing their tobacco smoking to treating clinicians, children responding to surveys may not disclose their tobacco smoking if a parent is present and even adult survey respondents may be embarrassed when disclosing their tobacco smoking status to interviewers.

Some researchers have attempted to develop more objective measures of tobacco smoke exposure to illuminate or avoid under-reporting. A commonly used biomarker of tobacco smoke exposure is cotinine, a by-product of nicotine, which can be detected through saliva, urine, or blood tests (Florescu et al., 2009). But, whether due to expense, logistics

and/or invasiveness, these biomarkers for smoking are not routinely available in national health survey data.

The World Health Organisation and US Centers for Disease Control (2011) collaborated to provide a preferred list of tobacco smoking questions for national health surveys; based on the Global Adult Tobacco survey. The first three questions are self-reports of current smoking status, past smoking status, and number of tobacco products smoked per day. These questions appear to be core questions in most national health surveys.

In the previous systematic review (Section 2.2) pooled results for smoking status and smoking packyears were presented. A pack year is 20 cigarettes smoked every day for one year. Total packyears is calculated by multiplying the number of years a person has smoked by the number of packets of cigarettes smoked per day (National Cancer Institute, 2021b). However, packyears has been criticised for misrepresenting cumulative risk (see for example (Peto, 2012) in relation to lung cancer and (Nance et al., 2017) in relation to cardiovascular disease). It is argued that half a pack per day for 40 years is associated with significantly higher health risks than two packs per year for 10 years, although both equate to cumulative 20 packyears of smoking. It is recommended that smoking intensity (cigarettes per day) and smoking duration (years) be kept as separate predictors (Peto, 2012) or that only smoking intensity is included (Nance et al., 2017).

3.3.2 **Alcohol consumption**

Alcohol consumption is thought to increase the risk of OC in multiple ways. For example (National Cancer Institute, 2022a),

- the body metabolises ethanol to acetaldehyde, a toxic chemical which can damage both DNA and proteins,
- reactive oxygen species are generated, which can damage DNA, proteins, and lipids in the body through oxidation, and
- the body's ability to break down and absorb a variety of nutrients (including vitamin A, folate, vitamin C, vitamin D, vitamin E, and carotenoids) is impaired.

However, clinical mechanisms for associations between pre-diagnosis alcohol consumption and post-diagnosis survival time in OC have not yet been documented.

In most countries the focus of health promotion activities is to address 'excessive' alcohol consumption rather than eliminate all alcohol. In Australia, the current guideline to reduce the risk of alcohol-related harm in adults is:

"To reduce the risk of harm from alcohol-related disease or injury, healthy men and women should drink no more than 10 standard drinks a week and no more than 4 standard drinks on any one day. The less you drink, the lower your risk of harm from alcohol"

(Australian Government Department of Health, 2020)

In the US the National Institute on Alcohol Abuse and Alcoholism and the Substance abuse and Mental Health Services Administration maintain separate but similar definitions of high-risk alcohol consumption. "Binge drinking" is defined as 5 or more drinks for males (4 or more for females) on the same occasion. "Heavy alcohol use" is defined as either more

than 4 drinks (3 drinks) on any one day or more than 14 drinks (7 drinks) per week for males (females) or binge drinking on 5 or more days in the past month (National Institute on Alcohol Abuse and Alcoholism, 2021b).

The previous systematic review (Section 2.2) reported pooled estimates of impact of 'ever consuming alcohol pre-diagnosis' (compared to never) and 'highest alcohol category pre-diagnosis' (compared to lowest) on post-diagnosis survival in OC. These analyses did not address the high-risk alcohol consumption patterns defined above.

Self-report of alcohol consumption significantly underestimates true alcohol consumption. For example, it has been estimated that self-reported alcohol consumption underestimates total alcohol sales in a community by about 40-60% (that is, an undercount of about one bottle of wine per adult per week in the UK) (Nugawela, Langley, Szatkowski, & Lewis, 2016). This is partly due to the under-participation of heavy drinkers in surveys, as well as accidental or deliberate under-reporting by those who do respond. There are many social pressures encouraging the under-reporting of alcohol (see for example "impression management" (Davis, Thake, & Vilhena, 2010)).

One study reviewing national health surveys, found this under-reporting was worse in light drinkers (Stockwell et al., 2016), but the method used (comparing two self-reports from the same person) seems more sensitive to accidental rather than deliberate under-reporting Other studies suggest greater under-reporting by heavy drinkers (Boniface, Kneale, & Shelton, 2014). The truth may be universal under-reporting.

One objective measure of alcohol consumption is direct detection of ethanol in the breath (breathalyser) or blood. But, on average, it only takes about one hour for the body to metabolise one standard drink (DrugRehab.com, 2021). So ethanol in the body is too transient to be used be used to measure long-term patterns of exposure. Biomarkers of more

long-term alcohol abuse have relatively low sensitivity and specificity or are too challenging for routine clinical usage (Torrente, Freeman, & Vrana, 2012). Biomarker profiles, combining across biomarkers, can be more reliable at detecting alcohol abuse (Torrente et al., 2012), but increase the costs and complexity. Biomarkers of alcohol abuse are not routinely available in national health surveys.

Self-reporting of alcohol consumption has been operationalised in terms of a wide range of different measure tools but can be categorised into three main groups a) retrospective summary measures b) retrospective daily drinking measures and c) concurrent daily drinking measures (Patterson, Hogan, & Cox, 2019). The summary measures of alcohol consumption include quantity-frequency measures (average quantity of alcohol consumed on a drinking occasion and average frequency of drinking occasions for a specified period in the past). The daily drinking measures include diaries of number of drinks on each specific day of the period covered. Quantity-frequency approaches have been shown to underestimate alcohol consumption relative to diary-based measures (Patterson et al., 2019), but retrospective diaries can only be applied over relatively short time periods to allow accurate recall. Concurrent recording is known to decrease alcohol consumption during the recording period (Patterson et al., 2019), also leading to underestimation of longer-term exposures.

A recent review of international guidelines for measuring alcohol consumption in population surveys found four sets of guidelines all of which had quantity-frequency questions at their core: specifically alcohol drinking status in the past year or lifetime, volume of alcohol consumption in the past year and frequency and volume of heavy drinking episodes in the past year (Nugawela et al., 2016). This combination of questions would, in the absence of under-reporting, allow identification of both ‘heavy drinking’ and ‘binge drinking’. Alcohol diaries are infrequently incorporated into national health surveys, no doubt due to the complexity of these measures and time constraints of the surveys.

3.3.3 **Physical activity**

People are constantly moving during their normal activities of daily living. But, people require a certain minimum level, intensity, and variety of movements to maintain their physical health:

"Physical activity for health benefit comprises several components (e.g. intensity, frequency, duration and type) that can be carried out in different settings or contexts (e.g. leisure-time, occupational, incidental and transport). Measurement is further complicated because there are several dimensions of physical activity related to health (e.g. energy expenditure, fitness, strength and flexibility)." (Armstrong, Bauman, & Davies, 2000, p. 10)

The physical activity guidelines for Australian adults between 18 and 64 years of age recommend at least either 2.5 to 5 hours of moderate intensity physical activity of 1.25 to 2.5 hours of vigorous intensity activity (or equivalent combination) per week. For older Australians the recommendation is at least 30 minutes of moderate intensity physical activity on most, preferably all, days. Examples of moderate intensity physical activities provided include brisk walk, golf, mowing the lawn and swimming. Examples of vigorous intensity activities include jogging, aerobics, fast cycling, soccer or netball (Australian Government Department of Health, 2021). The US (US Department of Health and Human Services, 2018) and UK (Department of Health & Social Care, 2019) guidelines are very similar. Of course, these guidelines are focussed on general good health and/or cardiovascular health and are not specifically validated for maximising post-diagnosis survival after OC.

Physical activity profiles can also vary according to countries' economic development. Residents of less developed countries have higher levels of physical activity on average and a preponderance of physical activity in work and transport. Individuals in more developed countries tend to have more sedentary work and transport leaving leisure time a

proportionately larger contributor to total physical activity (Bauman et al., 2011; Ng & Popkin, 2012).

There is now a daunting array of physical activity measurement tools available. The 'doubly labelled water' approach using biomarkers is often viewed as the gold standard measurement of total physical activity, but it's too expensive to be used in large scale surveys (National Institute for Health Research, 2021). There is also a range of wearables available for physical activity, such as accelerometers, pedometers, heart rate monitors, and armbands (National Institute for Health Research, 2021; Sylvia, Bernstein, Hubbard, Keating, & Anderson, 2014). Some disadvantages with devices include costs and logistics (getting the devices to participants, getting participants to wear them, retrieving data, and retrieving the devices) and lack of context as to what type of activity is being performed. Finally, there are also many physical activity questionnaires and diaries which have been developed and validated in different populations in different times.

Self-report measures have the advantage of being low cost and relatively easy to administer. As with alcohol consumption, recording of self-report can be achieved using quantity-frequency questionnaires or diaries of activity. There is no overall pattern of agreement between self-report and physical measures of physical activity (Prince et al., 2008), however, with careful planning and management (Ainsworth et al., 2012), self-reported physical activity can sometimes produce good agreement with physical measurements (Besson, Brage, Jakes, Ekelund, & Wareham, 2010), appropriate for population-level investigations (Haskell, 2012).

One attempt to standardise physical activity measurement at the national level is the Global Physical Activity Questionnaire developed, first in 2002. It has 16 questions around three domains - occupational, transport and leisure time - and asks the respondent to recall

physical activity which lasts for at least 10 minutes with moderate and vigorous intensity in a typical week. It does not measure light intensity physical activity (Keating et al., 2019). A recent systematic review located 26 studies of the validity of the Global Physical Activity Questionnaire. Association between GPA and accelerometers was only poor to fair. (Keating et al., 2019).

Overall, accurate measurement of physical activity appears to require physical devices or many item questionnaires, both of which may be difficult to implement in the context of national health surveys. Physical activity measurement is also not always a core item in national health services, sometimes being included intermittently.

3.3.4 **Obesity**

Obesity is not a behaviour as such but is a combined outcome of high energy diet and insufficient physical activity. Perhaps the most known definition of obesity is high Body Mass Index (BMI). A person's BMI is their weight measured in kilograms divided by the square of their height measured in metres. The World Health Organisation has defined obesity as BMI $\geq$30 kg/m^2 (WHO, 1995), sometimes modified to BMI $\geq$27.5 kg/m^2 when applied to Asian populations (WHO Expert Consultation, 2004).

BMI, requiring only weight and height, is relatively easy to collect via either self-report or objective measurement with a stadiometer and weight scales. It has been shown that average self-reported BMI is almost always lower than average BMI from objective measurement, due to both overestimation of height and underestimation of weight but the magnitude of this underreporting varies considerably between studies (Gorber, Tremblay, Moher, & Gorber, 2007). National health surveys which are conducted face-to-face will often

incorporate the objective measurement of height and weight, while telephone-based surveys must rely on self-report.

While the BMI-based definition of obesity is widely used, it has also been criticised as a poor indicator of body fat as a health issue (Lewis et al., 2009; Nuttall, 2015). For example, the accumulation of visceral fat (measured by waist circumference) has been shown to be a better predictor of the health risks for cardiovascular disease (Neeland, Poirier, & Després, 2018). But waist circumference is a problematic alternative as the cutpoint used to distinguish between obese and non-obese differs by ethnic group (Misra, Wasir, & Vikram, 2005).

3.4 Data sets used in this thesis

Throughout this thesis analyses are based on the US SEER cancer registry data and US BRFSS national health survey. The reasons for selecting these data sources were sample size and data availability for a wide range of socio-demographic and health behaviour variables potentially associated with OC. SEER was the only cancer registry data which did not require a formal project review by the data custodians, nor any data extraction fees. BRFSS has the largest total sample size of the surveys reviewed above and was the only health survey conducted across the entire catchment of the SEER cancer registry. (The NHIS and NHANES are conducted on smaller samples and these samples come from randomly selected US Counties which may or may not be within the SEER catchment.) As an additional benefit, the SEER and BRFSS data sets had more demographic variables in common than any other cancer registry and national health survey pair.

The BRFSS health survey suffers from non-response. In 2015 the median response rates for the BRFSS health survey were 68.0% of all persons contacted or 47.2% of all in the

sampling frame adjusted for the expected number ineligible. (Centers for Disease Control, 2015). Non-response biases have been documented (Schneider, Clark, Rakowski, & Lapane, 2012). The BRFSS custodians compute and attach survey weights to each data record to allow users to weight the sample to something more demographically consistent with the US population (Iachan, Pierannunzi, Healey, Greenlund, & Town, 2016). However, throughout this thesis the BRFSS data set will only be used to augment the SEER cancer registry data sets. The BRFSS health survey weights will not be used and the demographic idiosyncrasies of the BRFSS health survey data set will be removed through data matching in the augmentation process.

The primary data set, the SEER cancer registry data, has much lower potential for bias. The SEER cancer registry standard for case ascertainment is reportedly 98% (Mahnken, Keighley, Cumming, Girod, & Mayo, 2008). Strong quality assurance activities are employed (H. S. Park, Lloyd, Decker, Wilson, & Yu, 2012) with evidence that the SEER data records have few missing data points in demographic variables (Kuo & Mobley, 2016). However, the population covered by the SEER cancer registries is only a subset of the US population. The population covered by the SEER cancer registries have been shown to have slightly higher proportions of younger, poorer, lower educated and unemployed people and slightly lower proportion of white people than the general US population (Kuo & Mobley, 2016). For example, SEER currently listed coverage includes only 42.0% of Whites compared to 44.7% of African Americans, 66.3% of Hispanics, 59.9% of American Indians and Alaska Natives, 70.7% of Asians, and 70.3% Hawaiian/Pacific Islanders (National Cancer Institute, 2021d).

Another constraint imposed by using national data sets is the delay between the event of interest and the commencement of analysis and production of results. Both the SEER cancer registry data and the BRFSS health survey are based on continuous data collection, but both are only compiled and released annually. Processes involved in checking and

compiling these data sets can add considerable delays to data availability. For example, the 2017 BRFSS health survey data set was released on October 9, 2018 (Centers for Disease Control, 2018). This means that interviews conducted in early 2017 only became available for analysis 18 months later. https://www.cdc.gov/brfss/annual_data/annual_2017.html As the SEER cancer registry compiles data from other more local registries, the delays are longer. For example, the SEER cancer registry data for 2017 were released in April 2020 (National Cancer Institute, 2022b), three years after diagnosis for those diagnosed early in 2017.

3.5 Variables used in this thesis

This choice of data sources partially defines the scope of analyses in this thesis. Firstly, the SEER cancer registry data definition of an OC case is used. That is, OC cases were those with a newly reported primary malignant cancer coded in the range C15.0 to C15.9 on the International Classification of Diseases for Oncology 3rd edition (World Health Organization, 2013).

Secondly, analyses included the geographic coverage of the SEER cancer registries. This coverage, during the study period, consisted of whole State cancer registries (Connecticut, Georgia Center for Cancer Statistics, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, Utah), registries for smaller geographic regions (Los Angeles, Greater Bay Area Cancer Registry, Greater California, Detroit, Seattle-Puget Sound) and sub-populations (the Alaska Native Tumour Registry).

Thirdly, the years used in the analyses were determined by the availability and consistency of data. 2015 was the latest year of SEER cancer registry data available at the commencement of the analyses. 2001 was the first year with consistent definitions of the

health behaviour variables in the BRFSS health survey. Hence all analyses are restricted to the 2001 to 2015 time period.

Finally, the variables used in the analyses in this thesis were ether provided by, or derived from, either the SEER cancer registry and/or BRFSS health survey data. Table 3.1 describes each of the variables used.

Table 3.1 The variables used in this thesis

SEER variables
Outcome variables
Survival time recode Number of whole months from date of diagnosis to either date of death, date last known to be alive, or follow-up cutoff date for the SEER file (additional days are ignored)
Vital status recode (study cutoff used) 'dead' or 'alive' at follow-up cutoff date.
Variables in common with BRFSS
Age recode with <1 year olds Age at diagnosis coded as '<1 year', '1-4 years', '5-9 years', … , '85+ years' Analysed as: '80-84 years' and '85+ years' combined for compatibility with BRFSS
Sex 'male', 'female'
Marital status at diagnosis 'Married (including common law)', 'Divorced or Separated'; 'Single (never married)', 'Unmarried or domestic partner', 'Widowed', 'Unknown' Analysed as: 4 category marital status 'Unmarried or domestic partner' combined with 'Single/never married' Analysed as: 2 category marital status 'Divorced or Separated', 'Single (never married)', 'Unmarried or domestic partner' and 'Widowed' all combined into a single 'Unmarried' category
Race recode (W,B,AI,API) 'White', 'Black', 'American Indian/Alaska Native', 'Asian or Pacific Islander', 'Unknown'
State-county State and county of residence at diagnosis Analysed as: Only State code used
Year of diagnosis Calendar year
Potential confounding variables
Histologic type ICD-O-3 Histology codes from ICD-O-3 Analysed as: 8050-8089 categorised as squamous cell carcinoma, 8140-8389 categorised as adenocarcinoma

Table 3.1 (continued) The variables used in this thesis

Derived AJCC Stage Group, 6th ed (2004+) Only available for 2004 onwards. Stage category from the AJCC 6th edition. 'I', 'IIA', 'IIB', 'III', 'IV', 'IVB', 'IVNOS', 'Unknown' Analysed as: 'IIA' and 'IICB' combined as stage 'II', 'IV', 'IVB', 'IVNOS' combined as stage 'IV'
Reason no cancer directed surgery 'Not performed, patient died prior to recommended surgery', 'Not recommended', 'Not recommended, contraindicated due to other cond, autopsy only (1973-2002)', 'Recommended but not performed, patient refused', 'Recommended but not performed, unknown reason', 'Recommended, unknown if performed', 'Surgery performed', 'Unknown, death certificate, or autopsy only (2003+)' Analysed as: Treated with curative intent (yes/no) 'Surgery performed' combined with all categories with 'recommended' surgery into new category of 'yes, curative intent', all other categories combined into 'no curative intent'
BRFSS variables
Behavioural variables
Four-level smoker status 'now smokes every day', 'now smokes some days', 'former smoker', 'never smoked' Analysed as: 'now smokes every day' and 'now smokes some days' combined to for 'current smoker' category, 'former smoker' and 'never smoked' combined to form 'not current smoker' category'
Binge drinkers (males ≥5 drinks on one occasion, females ≥4 drinks on one occasion) 'binge drinker', 'not binge drinker'
Heavy drinkers (adult men ≥2 drinks/day, adult women ≥1 drinks/day) 'heavy drinker', 'not heavy drinker'
Adults reporting physical activity in past 30 days other than regular job) 'reported physical activity', 'did not report physical activity'
BMI body mass index Analysed as: 'obese' if BMI≥30, 'not obese' if BMI<30
Drinks per day Calculated total number of alcoholic beverages consumed per day Analysed as: ≥1 drink per day on average combined with current smoker categorised as 'regular smoker and drinker', <1 drink per day on average and/or not current smoker categorised as 'not regular smoker and drinker'

Table 3.1 (continued) The variables used in this thesis

Variables in common with SEER
Imputed age value collapsed above 80 Age at interview categorised as '18-24 years', '25-29 years', '30-34 years', … , '80+ years'
Sex 'male', 'female'
Marital status 'Married', 'Divorced', 'Widowed', 'Separated', 'Never Married', 'A member of an unmarried couple'. 'Refused" Analysed as: 4 category marital status 'A member of an unmarried couple' combined with 'Married' and 'Separated' combined with 'Divorced' Analysed as: 2 category marital status 'Divorced', 'Widowed', 'Separated' and 'Never Married' combined as 'unmarried' and 'Married', 'A member of an unmarried couple' combined as 'married'
Preferred race category 'White', 'Black or African American', 'American Indian or Alaskan Native', 'Asian', 'Native Hawaiian or other Pacific Islander', 'Other race', 'No preferred race', 'Multiracial but preferred race not answered', 'Don't know/not sure', 'Refused' Analysed as: 'Asian', 'Native Hawaiian or other Pacific Islander' combined to be compatible with SEER data, 'Other race', 'No preferred race', 'Multiracial but preferred race not answered', 'Don't know/not sure', 'Refused' were all excluded from analyses
State FIPS code State of residence
Interview year Calendar year

The variables extracted from the SEER and BRFSS data placed several constraints on the proposed analysis. Firstly, some of the variables are lacking in detail. For example, survival time is measured in months rather than days, and age is measured in 5-year age categories rather than single years. Further, the oldest age category in the BRFSS health survey is '80+ years' sometimes forcing SEER cancer registry data to be aggregated across '80-84 years' and '85+ years' age categories. The somewhat less detailed information going into the analyses will tend to lead to less precision in the results (i.e. widening confidence intervals).

There is one missing variable problem. Derived AJCC cancer stage at diagnosis is unavailable before 2004. It also has a high rate of missing data (18%) in the years when it is

available. Cancer stage has been retained as a potential confounding variable. It is known to have a close association with cancer survival time and could potentially be associated with health behaviour (such as if smokers were more likely to be diagnosed at a later cancer stage than non-smokers). It could be contended that cancer stage at diagnosis is not a confounder, but rather and intermediary on the disease pathway between behaviour and survival time, and that adjustment for an intermediary in conditional regression models induces collider stratification bias (Cole et al., 2010). However, it is so fundamental to clinicians and patients that most readers and reviewers expect it to be considered.

Given the missing years and other missing data in the measurement of cancer stage, surgical treatment variable was considered for use as a surrogate. Removal of the primary tumour becomes moot in metastatic cancer. Therefore, stage IV cancer is strongly associated with no surgery. There will also be a small proportion of OC cases where surgery is not recommended due to comorbidities, but generally the surgery versus no surgery divide is a reasonable proxy for the cancer stages I, II, III, versus cancer stage IV.

Another issue arising from the selection of variables is that there are some mismatches in the variables that are in common to both data sets. For example, while State of residence occurs in both data sets, the catchment areas are slightly different. For example, SEER's catchment of area of Detroit can only be equated with BRFSS catchment area of Michigan State and SEER's catchment area of Seattle-Puget Sound can only be equated to the BRFSS catchment area of Washington State. Also, as race is coded into just 4 categories the SEER cancer registry data set, BRFSS health survey respondents coded as 'multi-racial, but preferred race not answered' have no equivalent in the SEER data set.

The health behaviour variables were obtained from the BRFSS health survey data set and are all self-reported. Self-report is known to underestimate the prevalence of higher risk

behaviours. For example, one study concluded that self-reported obesity is underestimated by 16% in the BRFSS data set in comparison to the objectively measured BMI in the National Health and Nutrition Examination Survey (NHANES) (28.67% in BRFSS compared to 34.01% in NHANES) (Ward et al., 2016)

Dichotomous measures of health behaviour (displays the behaviour, does not display the behaviour) were used throughout. Dichotomous variables contain the least amount of information for analysis but represent individuals most fundamental decisions: Shall I stop smoking? (yes/no)?; Shall I stop drinking in excess (yes/no)?; Shall I take up a regular exercise (yes/no)? Given the potential measurement biases in self-reported behaviour, it was determined to focus on the most fundamental measures of behaviour.

More specifically, the tobacco smoking questions in the BRFSS health survey asked if respondents had ever smoked 100 cigarettes, followed by current smoking frequency categorised as “some days”, “every day” or ‘not at all’. As there was little information on quantity smoked, and as less than 10% of respondents fell into the “some days” category, the simple dichotomy between current smokers and not current smokers seemed to be the most informative measure.

The alcohol consumption questions related to self-reported behaviour in the past 30 days. These were frequency and quantity questions including the number of drink days, the average number of drinks per drink day and the largest number of drinks on any one day. The BRFSS had already identified individuals with behaviours consistent with the US Substance Abuse and Mental Health Services Administration’s definitions of binge drinking and heavy alcohol use (National Institute on Alcohol Abuse and Alcoholism, 2021b). These two dichotomous measures were included for the analyses.

The core physical activity questions in the BRFSS health survey were specific to discretionary or leisure time activities in the past 30 days. That is, the measures specifically excluded activities in respondents' regular occupational setting. The questions related to whether they participated in any activity, what was their most common activity, and the number of times per week and average duration of sessions of this one most common activity. As the frequency quantity questions were not inclusive of all leisure time activities, the more basic question of any leisure time activity at all (yes/no) was used.

Obesity was derived from self-reported height and weight using the standard BMI ≥30 kg/m^2 cut point (WHO, 1995).

An additional behavioural variable was created describing the joint behaviour of smoking and regular alcohol consumption. It has previously been suggested that smokers who drink an average of 1.5 standard drinks per day have 8.05 (95% CI 3.89,16.60) times the risk of contracting OC than never smokers who drink an average of 0.5 standard drinkers per day (Steevens, Schouten, Goldbohm, & van den Brandt, 2010). To investigate this, a dichotomous behavioural measure was defined comparing respondents who were current smokers and drank an average of at least 1.0 drink per day to all other respondents. The lower threshold of 1.0 drinks per day had been used in previous research (Thrift, Nagle, Fahey, Russell, et al., 2012) to boost sample size in this group. It was anticipated that larger sample sizes would also be required for the methods in this thesis. In hindsight, however, the lower alcohol threshold may have undermined the differentiation of the clinically important risk behaviours.

3.6 Addressing missing variables

The SEER cancer registry data contains all required variables except pre-diagnosis health behaviours. The BRFSS health survey does record the required health behaviour. However, the SEER cancer registry and BRFSS health survey data sets don't have a lot in common: one is a census the other a sample survey and they are designed and administered by different organisations for different research purposes. The main commonality is that both collect information from the same population (US adults), albeit different people from this population.

In 2014 there were 464,664 BRFSS respondents from the 244,737,285 adults in the US. That is, the proportion of US adults who responded to the BRFSS was 0.001787. The SEER cancer registry data set recorded 3,816 OC cases in 2015. Crudely, 0.001787 of 3,816 or about 6 of the 2015 OC cases could be expected to have responded to the BRFSS in the previous year. Thus, one-to-one data linkage of the BRFSS health survey to the SEER cancer registry data would appear futile – the expected sample size is tiny, and the matching errors would overwhelm the few true matches which were identified.

Given the two data sets cover the same catchment, it would be possible to undertake 'ecological inference' (Freedman, 1999). That is, change the unit of analysis from individual to geographic area. The first step would be to summarise available data by a geographic measurement in common between the two data sets. The smallest geographic areas defined in both SEER and BRFSS are the US States (the SEER cancer registries do provide information down to US counties, but the smallest geographic units in the BRFSS health survey is a rural versus urban variable). One example of ecological inference is to use the proportion of people in the State with the behaviour as a predictor of the median survival time for OC within the State. In doing so, the sample size is reduced to just 10 (the number of States in the SEER data set) and change the focus of the research question to describing

differences in OC incidence as a reflection of health systems, social conditions, or data collection between the States. Results of ecological inference tell nothing about the relationship between pre-diagnosis health behaviour and post-diagnosis survival time at the individual level. Any attempt to extrapolate from the State-level results back to individuals would be an example of the 'ecological fallacy' (Freedman, 1999). Ecological inference does not address the research questions of this thesis and is not pursued.

The second aim of the current research is to develop a new approach to combining information from the cancer registry and the national health survey. The starting position of this thesis is to consider the six behavioural variables as present in the SEER data set, with 100% missing data. There are many methods for correcting for missing data during analysis (Little & Rubin, 2019), some of which may be applicable to, or extendable to, variables with 100% missing data.

Note that with 100% missing data, there are no biases: every observation has an identical probability (1.00) of being missing. Therefore, the data are missing completely at random (MCAR), simplifying the analyses.

Little and Rubin (2019) classify missing value analyses into four (possibly overlapping) types. The first two of these, exclusion and weighting, are simply impossible with 100% missing data. That is, excluding all data records with missing data results in the exclusion of the entire data set and there are no actual observations to give higher or lower weight to. Little and Rubin's third and fourth approaches to missing value analysis, imputation and modelling, are the focus of this current research.

Suppose imputation is conceptualised as making the best possible guesses of the values of the missing data points and model-based methods are making the best possible predictions for the values of the missing data points. In either case, finding the best possible

solution involves using all available information which is informative of the missing data points.

This thesis focusses on six dichotomous pre-diagnosis health behaviours which are 100% missing on the SEER cancer registry data set. The only available information which is even partly informative of the missing data points are the health behaviours of those other members of the same underlying population who responded to the BRFSS health survey. That is, individuals with similar demographic characteristics will have a slightly increased likelihood of displaying the same behaviours compared to individuals with different demographic backgrounds (see section 1.8).

That is, the only information available to inform the imputation or prediction of the missing health behaviours are the relatively small variations in the distributions of health behaviour between demographic groups.

3.7 Closing comments

In this Chapter it has been confirmed that it is common for developed countries to support both population cancer registries and national health surveys. Cancer registry data do not record pre-diagnosis health behaviour, but these health behaviour data are collated in national health surveys that have overlapping geographic catchments in some instances.

Unit record data was readily accessible from both the cancer registries and the national health surveys. Accessing unit record data from the cancer registries usually involves obtaining approval for the study protocol and paying a data access fee, but accessing unit record data from national health surveys rarely requires a study protocol and often does not involve a data access fee; particularly for researchers from within the country.

This means that any methods developed to augment cancer registry data with pre-diagnosis health behaviour from a national health survey will be applicable across many developed countries. The methods of this thesis will not be restricted to the specific data sets used in this thesis. A pragmatic choice was made to only use the US SEER cancer registries data sets and the US BRFSS health survey data sets during the analyses in this thesis, as these were routinely available data and provided the largest sample size on an array for health behaviours for analysis.

While the cancer registries represent censuses of the cancer populations in their catchments, the national health surveys are usually conducted on relatively small samples of the general population of the country. As relatively few of the cancer cases could be expected to have contributed to a previous health survey, record linkage would return unhelpfully small sample sizes for any but the most common cancer types. Record linkage has the additional complication of increasing both the risk and the consequences of a data confidentiality breach, as more data items are compiled for each individual.

Even if containing no individuals in common, human data sets still tend to have variables in common, particularly demographic variables. Demographic variables are known to have associations with health behaviours. So knowing someone's demographics provides a some quantum of information about their likelihood of displaying particular health behaviours. The second aim in this thesis is to investigate methods to exploit this relationship between demographics and health behaviour to augment cancer registry data with information on pre-diagnosis behaviour.

By using existing data sources, the time and expense of data collection, data cleaning and data management was avoided but at the cost of having no input into the definition of the variables. This thesis focusses on a group of six self-reported dichotomous measures of health

behaviour. The choice of behavioural measures was based on the availability of variables that could be most informative of OC post-diagnosis survival times. Dichotomous variables contain the least possible amount of information of any measurement type. However, it is possible that these relatively simple variables may be less affected with bias and may be computationally more straightforward to analyse than more complex measures.

As a first step towards augmenting the cancer registry data set, six pre-diagnosis health behaviour variables are conceptualised as being present in the cancer registry data with 100% missing data. Methods for addressing missing data are well established. This thesis now seeks to apply or extend these methods to address the 100% missing data in the pre-diagnosis variables.

For the remainder of this thesis both model-based and imputation-based methods are explored for addressing the missing health behaviour data for OC cases in the SEER cancer registry data, using the BRFSS health survey data as an external reference.

Chapters 4 to 6 document analyses that aim to solve this problem. Chapter 4 describes the trial of a model-based method, and Chapters 5 and 6 document an imputation-based method.

Chapter 4 A model-based approach to addressing missing pre-diagnosis health behaviour

4.1 Background

The second aim of this thesis was to develop, describe and evaluate a novel algorithm for investigating the relationship between pre-diagnosis health behaviour and post-diagnosis survival time in OC. Chapter 3 described the plan to explore both model-based and imputation-based methods for addressing the 100% missing pre-diagnosis behaviour data in the SEER cancer registry data set using the BRFSS health survey data as an external reference. This Chapter describes the development of a model-based solution to the 100% missing value problem.

The BRFSS health survey provided data on the health behaviours of participants and their demographic characteristics. The SEER data set contained the same demographic variables. A model of health behaviour in the BRFSS health survey using only demographic data can then be applied to the SEER OC cases to predict their missing health behaviour. The model provides the mechanism for information transfer between the BRFSS health survey data set and the SEER cancer registry data set.

The 3-step analysis plan comprises: 1) using the BRFSS data set to develop logistic models describing the association between each health behaviour and the available demographic variables; 2) using these models to estimate the probability of pre-diagnosis health behaviour of OC cases based on their demographic variables; and 3) analysing the relationship between predicted pre-diagnosis health behaviour and post-diagnosis survival time in the OC using Cox regression.

This approach has been published as a peer-reviewed research paper which is presented in Section 4.2. Key examples of the R code used in the analysis are described in the thesis Appendix. Section 4.3 provides a critique the methods that were used and Section 4.4, reviews this work in relation to the overall goals of the thesis.

4.2 Using estimated probability of pre-diagnosis behaviour as a predictor of cancer survival time: an example in oesophageal cancer.

The material in this section has been published as a peer reviewed journal article. The citation is: Fahey, P. P., Page, A., Stone, G., & Astell-Burt, T. (2020). Using estimated probability of pre-diagnosis behavior as a predictor of cancer survival time: an example in esophageal cancer. *BMC Medical Research Methodology*, 20, 1-9.

In 2020, the journal *BMC Medical Research Methodology* has a Q2 rating in Epidemiology in the Scimago rating scheme and a 2-year impact factor of 4.402.

Following the CRediT Taxonomy (National Information Standards Organization, 2021), author contributions were:

- Paul Fahey contributed to Conceptualisation, Data curation, Formal analysis, Methodology and Writing the original draft.
- Andrew Page, Glenn Stone and Thomas Astell-Burt contributed to Supervision
- All authors contributed to Review and Editing of the draft paper.

This material has been reformatted to suit the thesis with supplementary materials re-integrated into the paper, tables and figures re-numbered, and text converted from US to Australian English.

4.2.1 Abstract

Background: Information on the associations between pre-diagnosis health behaviour and post-diagnosis survival time in oesophageal cancer could assist in planning health services but can be difficult to obtain using established study designs. It is postulated that, with a large data set, using estimated probability for a behaviour as a predictor of survival times could provide useful insight as to the impact of actual behaviour.

Methods: Data from a national health survey and logistic regression were used to calculate the probability of selected health behaviours from participant's demographic characteristics for each oesophageal cancer case within a large cancer registry data base. The associations between survival time and the probability of the health behaviours were investigated using Cox regression.

Results: Observed associations include: a 0.1 increase in the probability of smoking one year prior to diagnosis was detrimental to survival (Hazard Ratio (HR) 1.21, 95% CI 1.19,1.23); a 0.1 increase in the probability of hazardous alcohol consumption 10 years prior to diagnosis was associated with decreased survival in squamous cell cancer (HR 1.29, 95% CI 1.07, 1.56) but not adenocarcinoma (HR 1.08, 95% CI 0.94,1.25); a 0.1 increase in the probability of physical activity outside the workplace is protective (HR 0.83, 95% CI 0.81,0.84).

Conclusions: It is concluded that probability for health behaviour estimated from demographic characteristics can provide an initial assessment of the association between pre-diagnosis health behaviour and post-diagnosis health outcomes, allowing some sharing of information across otherwise unrelated data collections.

4.2.2 Background

With an incidence of 9.3/100,000 males and 3.5/100,000 females per year, oesophageal cancer led to more than half a million deaths worldwide in 2018 (Bray et al., 2018). The majority of these deaths arise from modifiable lifestyle factors. In the US in 2014 it was estimated that 71% of male and 59% of female oesophageal cancer deaths arose from modifiable lifestyle factors and that cigarette smoking, alcohol consumption and excess body weight could account for up to 50%, 17% and 27% of deaths respectively (Islami et al., 2018).

While there is considerable documentation of associations between health behaviour and onset of oesophageal cancer (Clara Castro et al., 2018), the impact of health behaviour on survival times is less well understood (Fahey, Mallitt, Astell-Burt, Stone, & Whiteman, 2015). A more thorough understanding of predictors of survival time is needed to assist in anticipating health service needs and for health services planning.

Health behaviour prior to a cancer diagnosis is often different from health behaviour post-diagnosis. Behaviour prior to diagnosis can be influenced by public health activity but post-diagnosis behaviour is strongly influenced by the diagnosis itself (Toohey, Pumpa, Cooke, & Semple, 2016) and by treatment (Demark-Wahnefried et al., 2005; Rock et al., 2012). As oesophageal cancer has relatively short survival times (in the US, just 19% of cases survive 5-years (Siegel, Miller, & Jemal, 2019)), pre-diagnosis behaviour could have a strong carry over effect on survival time.

Unfortunately, investigating the effect of pre-diagnosis behaviour on post-diagnosis survival can be difficult and expensive. As the disease is relatively rare, a prospective cohort study would be inefficient (on the figures above, surveillance of 100,000 men for 10 years would be expected to yield just 93 new oesophageal cancer cases). Retrospective studies

which enrol newly diagnosed cancer patients and ask them to recall their prior health behaviour still involve considerable expense and are fraught with recall and survivor biases. In one example, an Australian study enrolling newly diagnosed oesophageal cancer patients reported that patients with late-stage disease were difficult to enrol and under-represented (Smithers et al., 2010).

Secondary analyses of already existing data can provide alternate, cost-effective opportunities. It is now common for governments to sponsor both regular health behaviour surveys and mandatory cancer registries. For those cancer cases who contributed to a survey prior to diagnosis, their health behaviour and cancer outcomes can be linked to produce a retrospective cohort. Data linkage avoids recall and survivor biases and is cost efficient (as the required data are already collected, compiled and cleaned).

But data linkage may not be feasible either. Confidentiality is one issue. But more fundamentally, as oesophageal cancer is relatively rare, the number of cancer cases who happened to have previously participated in the health survey is likely to be very small. If data linkage cannot be applied, is there any other way in which these rich (and expensive) data sets can be used to help provide insights into the association between pre-diagnosis behaviour and post-diagnosis survival times?

Often the only measures in common between cancer registries and national health surveys are the demographic characteristics of participants. It is known that demographically similar people are more likely to display similar health behaviour than people from different demographic groups (Morris, D'Este, Sargent-Cox, & Anstey, 2016). That is, different demographic groups have a different likelihood for particular behaviours. Probability of behaviour calculated from demographic variables, may be a weak indicator of actual behaviour, but with large data sets even weak signals are detectable.

This study investigated whether or not useful information on the association between pre-diagnosis health behaviours and post-diagnosis survival times could be obtained by analysing cancer cases estimated probability of engaging in these behaviours. The analyses used US data and focused mainly on the three modifiable lifestyle factors identified above: cigarette smoking, alcohol consumption and excess body weight.

4.2.3 **Methods**

The data sets

Unit record data on oesophageal cancer cases and their outcomes was extracted from the Surveillance, Epidemiology, and End Results Program (SEER) cancer registry (Surveillance Epidemiology and End Results (SEER) Program). The SEER system is administered by the National Cancer Institute. SEER currently compiles data from cancer registries covering about 28% of the US population across 13 States. Most cancers, including oesophageal cancers, are recorded. De-identified unit record data made available for research include demographic measures, medical details of the cancer, treatment and outcomes (including survival time). 95.1% of oesophageal cases had positive histology with just 0.4% clinical diagnosis only; the remainder having unknown (2.4%) or other confirmation methods.

Data on health behaviour was extracted from the Behavioral Risk Factor Surveillance System (BRFSS) health survey (Centers for Disease Control and Prevention (CDA)). The BRFSS is an annual national survey of health. It commenced in 1984 and now collects data from more than 400,000 telephone interviews each year covering adult residents of all US States and three Territories. The de-identified unit record information made available for

research included demographic and health behaviour measures, and State population sampling weights.

Both collections provided access to cleaned, de-identified unit record data at no cost to the researcher. Although both data collections are large, with less than 0.2% of American adults participating in BRFSS and around 4,000 oesophageal cancer cases being recorded in the SEER data set each year, only about eight new oesophageal cancer cases each year could be expected to have participated in the previous BRFSS survey.

Inclusions and exclusions

This analysis focusses on the 15-year period from 2001 to 2015. Data prior to 2001 are excluded due to changes in the definitions of some health behaviours variables and because earlier data may be less relevant to current behaviour and outcomes. 2015 was the most recent year of SEER cancer registry data.

As oesophageal cancer is rare in young ages, all cancer cases who were less than 35 years of age are excluded as being atypical. 201 of 57025 (0.3%) cases are excluded. For the BRFSS health survey, all data records from respondents 25 or more years of age who lived in one of the 13 US States represented in the SEER cancer registries are included. Including the younger respondents allows information on health behaviour up to 10 years prior to cancer diagnosis to be retained.

Outcome variable

The outcome of interest is post-diagnosis survival time in months as recorded in the SEER cancer registry data set. That is, all cases with survival less than 30.4 days after diagnosis (including cancers detected post-mortem) have a survival time of 0 months, those who died between 30.4 and 60.8 days have a survival time of 1 month, etc. The maximum

possible survival time is 179 months. For those who are still alive and those who are lost to follow-up, survival time is censored at the date of last follow-up.

Health behaviour variables

The research focused mainly on measures relating to cigarette smoking, alcohol consumption and excess body weight. The choice of variables was restricted to measures available through the BRFSS health survey. The following variables, all recording self-reported behaviour, were included:

- Current smoker (yes/no) which includes those who smoke daily or less than daily;
- Alcohol - heavy drinking (yes or no), which is defined as more than two standard drinks per day for men and more than one standard drink per day for women in the month prior to survey;
- Alcohol - binge drinking (yes or no), which is defined as males reporting having five or more standard drinks or females reporting 4 or more standard drinks on one occasion in the month prior to survey;
- Current smoking and alcohol consumption (yes/no), which is defined as both current smoker and an average consumption of ≥1 standard drink of alcohol per day in the past month.
- Obese (yes/no) which is BMI ≥ 30 kg/m2
- Undertook physical activity or exercise in the past 30 days other than regular job (yes or no)

Demographic variables

As the cancer registry data did not include information on pre-diagnosis health behaviour, the probability of each pre-diagnosis health behaviour for each cancer case was estimated using the available demographic variables.

Of the variables in common between the SEER cancer registry and the BRFSS health surveys it was hypothesized that year, age, sex, race, marital status and State of residence could be helpful for predicting health behaviour. For example, race is known to be associated with smoking (Jamal, 2016) and alcohol dependence (Gilman et al., 2008) in the US. Also, living as married ameliorates social isolation and social isolation is associated with adverse health behaviours such as smoking, higher BMI, and lower desire for exercise (Lauder, Mummery, Jones, & Caperchione, 2006).

As age was recorded in 5-year age groups in the SEER cancer registry data, the same categories were applied to the BRFSS health survey data. Race was categorized as White; Black; Asian or Pacific Islander; and American Indian or Alaskan native. Participants in the BRFSS health survey who self-reported as mixed race (n=44,670, 3.1% of total) were omitted as there was no corresponding code in the SEER cancer registry data set. Marital status was categorized as married or living as married; divorced or separated; widowed; and single.

Other factors considered

Post-diagnosis survival time is sensitive to a range of factors, some of which could potentially confound associations with pre-diagnosis health behaviour and survival time. For example, the association between health behaviours and incidence of oesophageal cancer is known to differ by histological type (Clara Castro et al., 2018; Steevens et al., 2010) and

these differences appear to carry over into survival time (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012). Therefore, sub-group analyses were conducted for squamous cell carcinoma (OSCC) and adenocarcinoma (OAC). Also age is associated with survival time (Njei et al., 2016) and health behaviour can change with age. Age, recorded in 5-year age groups but treated as a continuous variable, is included in the final models as a potential confounder.

Somewhat more difficult was how to address cancer stage. Cancer stage at diagnosis is an important predictor of survival time (Njei et al., 2016) and could perhaps be associated with health behaviour, although this association may be an intermediary step between health behaviour and survival time rather than a true confounder. For completeness cancer stage was adjusted in the models. Disease stage at diagnosis (clinical assessment) was coded by SEER according to the according to the AJCC Cancer Staging Manual 6th Edition (Greene et al., 2003).

Recording of cancer stage at diagnosis was incomplete in the SEER cancer registry data; being unavailable from 2001 to 2003 and having 18% missing data across the other years. Cancer stage was excluded prior to 2004 and categorized it into 5 categories (stage I, stage II, stage III, stage IV, not specified) from 2004 onwards.

Other potential confounders of the association between behaviour and survival were considered to be of lesser impact or potentially on the disease pathway. For example, while the relationship between smoking history and post-diagnosis survival may differ by gender, the effect may be small. In contrast, the choice between curative or palliative treatment is a strong predictor of survival time but may partially lie on the association pathway. (Smoking, for example, may lead to a higher probability of significant co-morbidities and these in turn influence the decision of curative treatment and, hence, survival time.) Adjustment for

variables on the association pathway may remove some of the true association between health behaviour and survival time.

Eligible data records

56,824 SEER oesophageal cancer cases and 1,450,775 BRFSS health survey respondents met the eligibility criteria. Table 4.1 summarizes the characteristics of the two samples. Among the cancer cases, median time till death was 7 months with median follow-up time of censored observations (18.6%) was 30 months. 52.9% of cases were OAC and 33.7% OSCC. 16.1% of the BRFSS respondents were current smokers and 4.8% were judged to be heavy drinkers of alcohol. The BRFSS respondents included higher proportions of younger people and females than the SEER cases.

Table 4.1 Disease characteristics and outcomes of eligible SEER cancer registry oesophageal cancer cases and BRFSS health survey respondents 2001-2015.

	Eligible oesophageal cancer cases 2001-2015 (n=56,824)		Eligible BRFSS respondents 2001-2015 (n=1,450,775)	
	n	%	n	%
Number who died				
died	46,242	81.40%		
censored	10,582	18.60%		
Cancer type				
Squamous cell carcinoma (OSCC)	19,130	33.70%		
Adenocarcinoma (OAC)	30,067	52.90%		
other	7,627	13.40%		
Cancer stage				
I	7,215	15.60%		
II	7,756	16.70%		
III	8,390	8.10%		
IV	14,852	32.00%		
Unknown	8,164	17.60%		
Not collected	10,447			

Table 4.1 (continued) Disease characteristics and outcomes of eligible SEER cancer registry oesophageal cancer cases and BRFSS health survey respondents 2001-2015.

	Eligible oesophageal cancer cases 2001-2015 (n=56,824)		Eligible BRFSS respondents 2001-2015 (n=1,450,775)	
	n	%	n	%
Current smoker				
yes			233,533	16.10%
no			1,195,828	82.40%
missing			21,414	1.50%
Alcohol - binge drinking				
yes			152,087	10.50%
no			1,246,293	85.90%
missing			52,395	3.60%
Alcohol - heavy drinking				
yes			68,968	4.80%
no			1,325,544	91.40%
missing			56,263	3.90%
Current smoking and alcohol consumption				
yes			39,431	2.70%
no			1,348,373	92.90%
missing			62,971	4.30%
Obese				
yes			370,146	25.50%
no			1,004,366	69.20%
missing			76,263	5.30%
Undertook exercise in past 30 days other than regular job				
yes			1,071,481	73.90%
no			353,711	24.30%
missing			25,583	1.80%
State of residence				
Alaska[a]	93	0.20%	41,515	2.90%
California	20,478	36.00%	135,572	9.30%
Connecticut	3,222	5.70%	99,883	6.90%
Georgia	6,299	11.10%	89,692	6.20%
Hawaii	825	1.50%	82915	5.70%
Iowa	2,893	5.10%	83015	5.70%
Kentucky	3,571	6.30%	121288	8.40%
Louisiana	3,412	6.00%	92248	6.40%
Michigan	3,412	6.00%	114870	7.90%
New Jersey	6,950	12.20%	159302	11.00%
New Mexico	1,251	2.20%	90128	6.20%
Utah	1,004	1.80%	109605	7.60%
Washington	3,414	6.00%	230742	15.90%

Table 4.1 (continued) Disease characteristics and outcomes of eligible SEER cancer registry oesophageal cancer cases and BRFSS health survey respondents 2001-2015.

	Eligible oesophageal cancer cases 2001-2015 (n=56,824)		Eligible BRFSS respondents 2001-2015 (n=1,450,775)	
	n	%	n	%
Year of diagnosis or survey				
2001	3,519	6.20%	54078	3.70%
2002	3,410	6.00%	58489	4.00%
2003	3,518	6.20%	76551	5.30%
2004	3,791	6.70%	77720	5.40%
2005	3,476	6.10%	92681	6.40%
2006	3,675	6.50%	95708	6.60%
2007	3,737	6.60%	94696	6.50%
2008	3,810	6.70%	101275	7.00%
2009	3,936	6.90%	116729	8.00%
2010	3,879	6.80%	111233	7.70%
2011	3,942	6.90%	127934	8.80%
2012	3,979	7.00%	120058	8.30%
2013	3,951	7.00%	112713)	7.80%
2014	4,063	7.20%	106771	7.40%
2015	4,138	7.30%	104139	7.20%
Age				
25-29 years	excluded	excluded	75,187	5.20%
30-34 years	excluded	excluded	95,776	6.60%
35-39 years	341	0.60%	110,567	7.60%
40-44 years	910	1.60%	123,417	8.50%
45-49 years	2,311	4.10%	137,293	9.50%
50-54 years	4,310	7.60%	156,343	10.80%
55-59 years	6,608	11.60%	158,319	10.90%
60-64 years	8,171	14.40%	151,202	10.40%
65-69 years	8,622	15.20%	132,403	9.10%
70-74 years	7,896	13.90%	107,064	7.40%
75-79 years	7,288	12.80%	84,215	5.80%
80-84 years	5,664	10.00%	105,523*	7.30%
85+ years	4,703	8.30%		
Missing			13,466	0.90%
Sex				
male	43,856	77.20%	568,196	39.20%
female	12,968	22.80%	882,579	60.80%

Table 4.1 (continued) Disease characteristics and outcomes of eligible SEER cancer registry oesophageal cancer cases and BRFSS health survey respondents 2001-2015.

	Eligible oesophageal cancer cases 2001-2015 (n=56,824)		Eligible BRFSS respondents 2001-2015 (n=1,450,775)	
	n	%	n	%
Marital status				
married/defacto	30,553	53.80%	860,291	59.30%
divorced/separated	6,761	11.90%	240,596	16.60%
widowed	7,816	3.80%	185,631	12.80%
single/never married	8,769	15.40%	156,718	10.80%
missing	2925	5.10%	7539	0.50%
Race				
white	47,447	83.50%	1,217,366	83.90%
black	6,422	11.30%	108,396	7.50%
Pacific or Asian	2,475	4.40%	79,878	5.50%
American Indian/Alaskan Native	340	0.60%	26,893	5.50%
Missing	140	0.20%	18242	1.30%
excluded			44,670	

[a] For Alaska, the SEER cancer registry records American Indians and Alaskan natives only

Statistical analysis

The characteristics of eligible cancer registry cases and health survey respondents are summarized using counts and percentages, with the exception of survival time which is summarized using medians, quartiles and maximums.

The main analysis involves three discrete steps. Firstly, the probability of engaging in each health behaviour were estimated from the BRFSS health survey data using logistic models; with a separate model for each behaviour. Each modelled the probability of having the behaviour of interest based on year of survey, age, sex, race, marital status and State of residence. Differences in the probability of health behaviours between sexes and between marital statuses at different ages were allowed for by including age by sex, age by marital status and marital status by sex interaction terms in each logistic model.

For example, let i represent an eligible individual from the BRFSS data set and $p_i(\widehat{smoker})$ represent the estimated probability that person i is a smoker, then the logistic model has the form

$$logit\big(p_i(\widehat{smoker})\big) = \boldsymbol{x_i}\widehat{\boldsymbol{\beta}} \tag{1}$$

where

$$\begin{aligned}\boldsymbol{x_i}\widehat{\boldsymbol{\beta}} = \widehat{\beta_0} + \widehat{\beta_1}(year_i) + \widehat{\beta_2}(age_i) + \widehat{\beta_3}(sex_i) + \widehat{\beta_{4-6}}(race_i) + \widehat{\beta_7}(marital\ status_i) \\ + \widehat{\beta_{8-19}}(State\ of\ residence_i) + \widehat{\beta_{20}}(age_i)(sex_i) \\ + \widehat{\beta_{21}}(age_i)(marital\ status_i) + \widehat{\beta_{22}}(sex_i)(marital\ status_i)\end{aligned}$$

and the $\hat{\beta}$'s quantify the relationships between the demographic characteristics of the respondents and their likelihood of smoking.

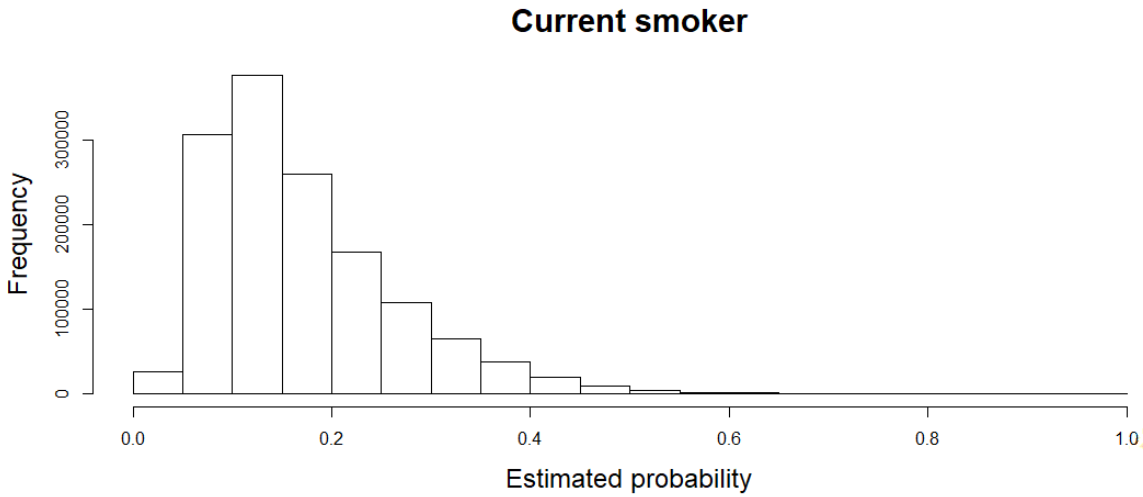
To correct for the complexities in the BRFSS health survey sampling and non-response the logistic models were weighted by the sampling weights provided. In 2011, the BRFSS introduced a new method of calculating sampling weights which improved the weighting of some variables including race and marital status. However, as both systems weight to the State totals, the analyses do not differentiate between the different type of weights. Data records with extreme sampling weights were excluded: those which fell in either the top or bottom 0.5% of the distribution. Firth's bias reduced penalized-likelihood was used to assist models to converge during fitting; using the `logistf()` package (version 1.23) in R software (version 3.5.2). The fitted models are summarized in Table 4.2 and Figure 4.1.

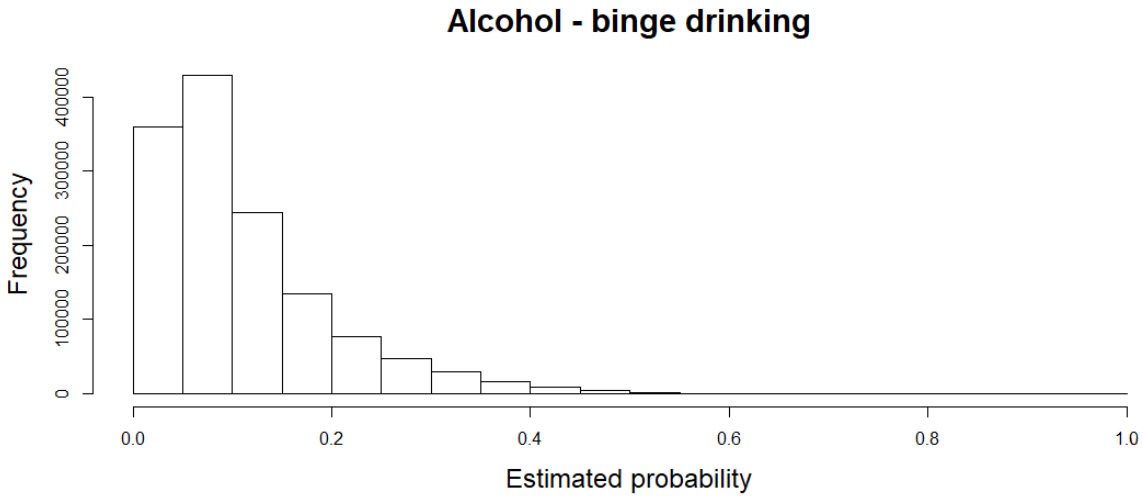
Table 4.2 Logistic regression models predicting health behaviours from demographic variables.

	Year	Age (5 year categories)	Sex	Married	Race- black	Race – American native	Race – Asian pacific	Age by sex	Age by married	Sex by married
Current smoking	-0.0224	-0.1218	-0.2229	-1.1577	-0.0867	0.4801	-0.7785	-0.0076	0.0237	0.1009
Undertook exercise in past 30 days other than regular job	0.0007	-0.0800	-0.1807	0.2287	-0.3992	-0.3856	-0.1156	-0.0015	0.0054	0.0153
Obese	0.0341	-0.0428	0.0473	0.4698	0.5654	0.4089	-1.0469	0.0124	0.0148	-0.4166
Alcohol - binge drinking	0.0298	-0.2490	-0.9603	-0.7847	-0.5513	-0.0700	-0.8542	-0.0055	0.0494	-0.0261
Alcohol - heavy drinking	0.0231	-0.1333	-0.6150	-1.7313	-0.5986	-0.0407	-0.9449	0.0171	0.0831	0.3203
Current smoking and alcohol consumption	-0.016	-0.1671	-1.2999	-1.2934	-0.2675	0.1399	-0.9132	0.0138	0.0364	0.1058

Year and age are treated as numeric. All other variables are binary (dummy) variables.

Figure 4.1 Histograms displaying the estimated probabilities of each health behaviour modelled on age group, sex, marital status, race, State of residence and year.





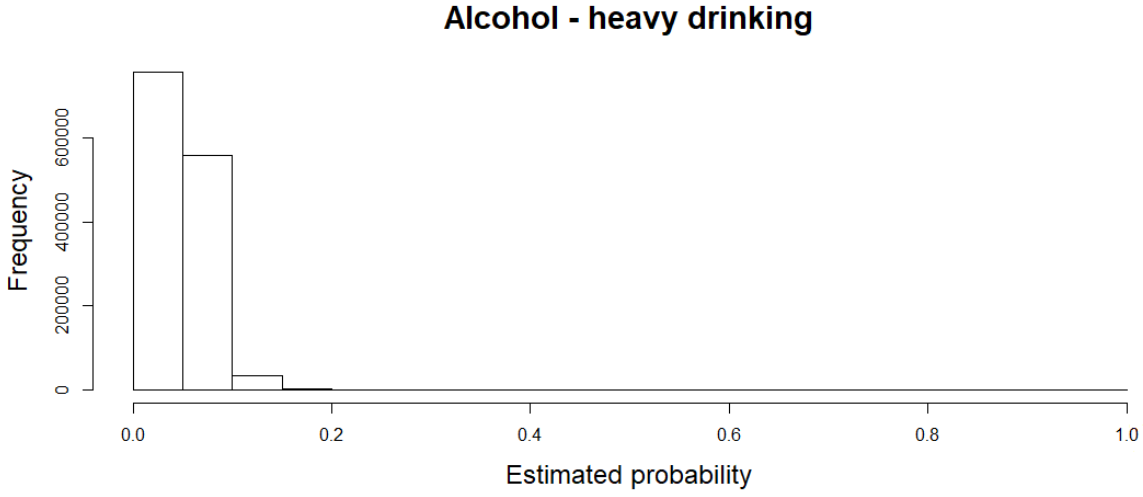
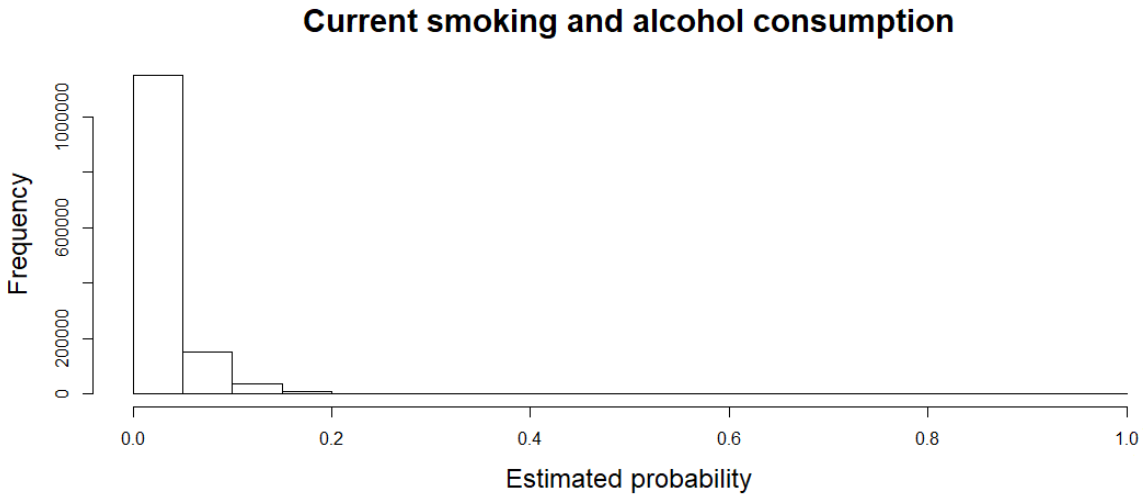
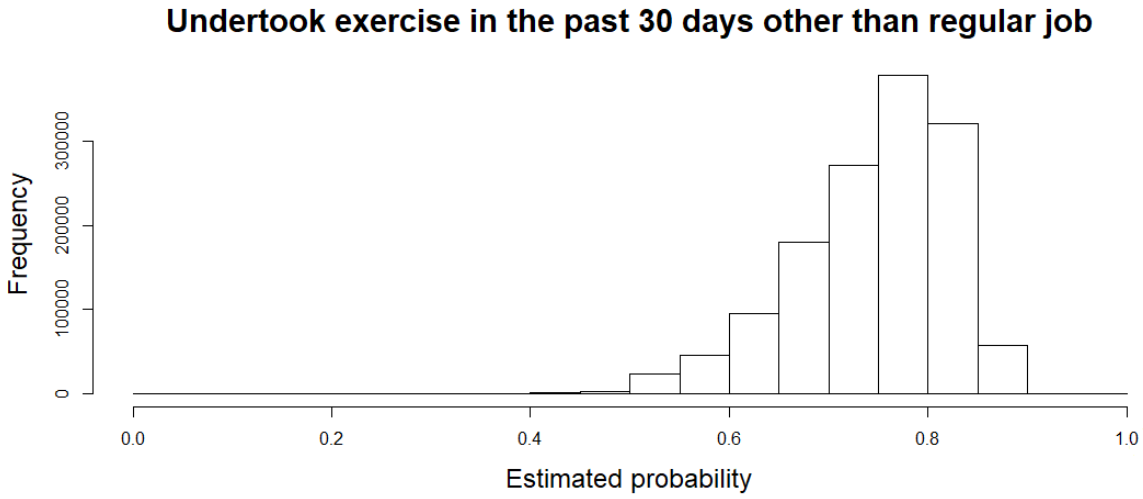
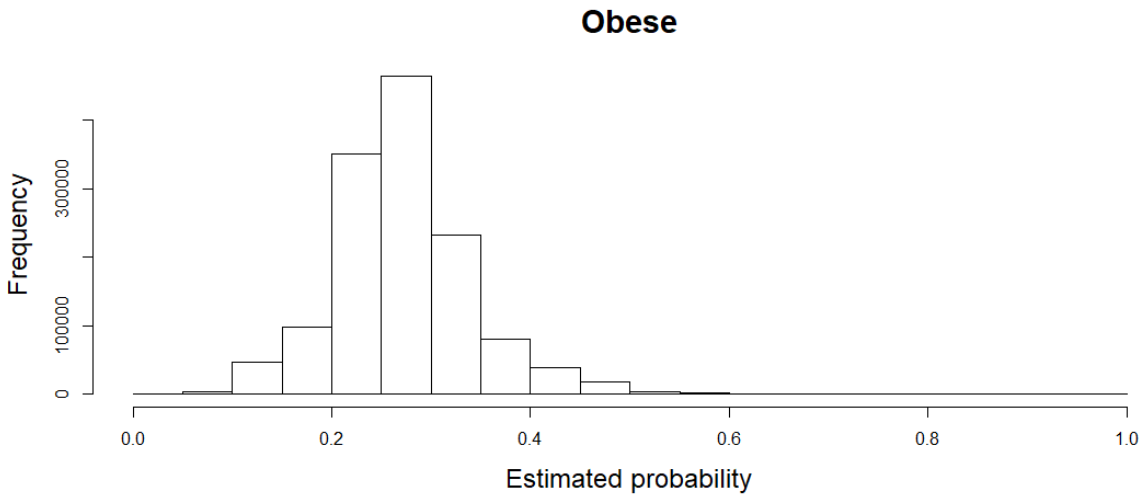


Figure 4.1 (continued) Histograms displaying the estimated probabilities of each health behaviour modelled on age group, sex, marital status, race, State of residence and year.







Year and age category were fitted as numeric variables while sex, race, marital status and State of residence are categorical. Preliminary investigations (not reported) confirmed that a linear model was reasonable for both year and age category. Year is coded as 0 for 2001 through to 14 for 2015 for analysis.

The chosen risk profiling variables were confirmed to be predictors of each health behaviour by visual inspection of odds ratios from logistic regression models. To help gauge the predictive ability of each demographic variable areas under the curve (AUC) of the receiver operating characteristic (ROC) curve are presented for each predictor alone and for the full logistic model using the pROC package (version 1.13.0) in R software. The higher above 0.5 the AUC, the greater the ability of the model to predict the health behaviour.

In the second step of the analysis, for each oesophageal cancer case in the SEER cancer registry, their probability of participating in each health behaviour was estimated by substituting their demographic characteristics into the logistic predictive model for that behaviour (Table 4.2).

For example, let *j* represent an eligible cancer case from the SEER data set and $\boldsymbol{x_j}$ the set of observed values of the demographic variables for individual *j* and $\widehat{\boldsymbol{\beta}}$ represent the regression coefficients for the model predicting smoking (equation 1 above), then the estimated probability of cancer case *j* being a smoker is

$$p_J(\widehat{smoker}) = \frac{e^{x_j\widehat{\beta}}}{1+e^{x_j\widehat{\beta}}} \qquad (2)$$

Given the specific interest in health behaviour prior to diagnosis three pre-diagnosis time points, 1, 5 and 10 years prior to diagnosis, were trialled. This entailed substituting diagnosis year minus 1, 5 or 10 as the year variable of the logistic model and 5-age group minus 0, 1 or 2. To avoid extrapolating earlier than the observed data, the 5-year lag analysis

was restricted to oesophageal cancer cases from 2006 to 2015 and the 10-year lag model was restricted to cases from 2011 to 2015.

In the third step of the analysis, the relationship between the estimated probability of each behaviour and survival was investigated using Cox regression models using the survival package (version 2.43-3) in R software. Separate models were fitted for each behaviour. Results are presented as hazard ratios (HRs) with associated 95% confidence intervals (CIs) and p-values. Models were fitted with and without correction for age and cancer stage at diagnosis.

For example, the Cox model of survival time of cancer case j relative to their estimated probability of smoking, adjusting for age and disease stage, could be written

$$S(t, x, \beta) = [S_0(t)]^{exp\left(\beta_1^*(p_J(\widehat{smoker}))+\beta_2^*(age_j)+\beta_3^*(cancer\ stage_j)\right)} \quad (3)$$

where $p_J(\widehat{smoker})$, a number between 0 and 1, is the estimated probability that the SEER cancer case is a smoker from Equation (2). The * superscript denotes that these β's are different to the β's listed in Equation (1). Under this model $e^{\beta_1^*}$ is the hazard ratio for the estimated probability of smoking, adjusted for age and disease stage.

Subgroup analyses were performed for OSCC and OAC histological types. Missing values were excluded from analysis.

4.2.5 **Results**

Each of the risk profile variables were related to each of the health behaviours (Table 4.3). For example, the prevalence of smoking decreased over the study period (odds ratio (OR)=0.98, 95% confidence interval (CI) 0.98-0.98 for each later year); the prevalence of obesity increased over time (OR=1.03, 95% CI 1.03-1.03 for each additional year); each 5-year increase in age is associated with decreasing prevalence of smoking (OR=0.90, 95% CI 0.90-0.90) and decreased risk of binge drinking (OR=0.82, 95% CI 0.82-0.82); females have lower prevalence of smoking (OR=0.74, 95% CI 0.74,0.74); when compared to those who are married, people who are single have higher prevalence of daily smoking (OR=2.14, 95% CI 2.14-2.14), risk of binge drinking (OR=1.90, 95% CI 0.90-0.90) and risk of concurrently smoking and regular drinking (OR=2.50, 95% CI 2.50-2.50); people classifying as American Indian or Alaskan Native have higher prevalence of daily smoking (OR=1.69, 95% CI 1.68-1.69) and people classified as black have higher risk of obesity (OR=1.75, 95% CI 1.75-1.75) than those who are classified as white; residents of Kentucky are more likely to smoke (OR=2.50, 95% CI 2.49-2.50) residents of Utah are less likely to be heavy drinkers (OR=0.52, 95% CI 0.52-0.52) than Californians.

Table 4.3 Associations between the selected demographic variables and health behaviours in the BRFSS health survey data set

	Current smoking			Alcohol - binge drinking			Alcohol - heavy drinking		
	OR	95% CI	AUC	OR	95% CI	AUC	OR	95% CI	AUC
State of residence									
Alaska [a]	1.93	1.93,1.93		1.36	1.36,1.36		1.14	1.14,1.14	
California	1.00			1.00			1.00		
Connecticut	1.25	1.25,1.25		1.06	1.06,1.06		0.93	0.93,0.93	
Georgia	1.63	1.63,1.63		0.85	0.85,0.85		0.74	0.74,0.74	
Hawaii	1.23	1.23,1.24		1.18	1.18,1.18		1.15	1.15,1.15	
Iowa	1.56	1.56,1.56		1.35	1.35,1.35		0.89	0.89,0.90	
Kentucky	2.5	2.49,2.50		0.71	0.70,0.71		0.6	0.60,0.61	
Louisiana	2.04	2.04,2.04		1.01	1.01,1.01		0.85	0.85,0.85	
Michigan	1.83	1.83,1.83		1.22	1.22,1.22		0.94	0.94,0.94	
New Jersey	1.32	1.32,1.32		0.96	0.96,0.97		0.74	0.74,0.74	
New Mexico	1.6	1.59,1.60		0.84	0.84,0.84		0.8	0.80,0.80	
Utah	0.79	0.79,0.79		0.65	0.65,0.65		0.52	0.52,0.52	
Washington	1.35	1.35,1.35	0.574	1.05	1.05,1.05	0.558	0.97	0.97.0.97	0.567
Year of survey									
per additional year	0.98	0.98,0.98	0.544	1.03	1.03,1.03	0.516	1.02	1.02,1.02	0.523
Age									
per additional 5-year age category	0.9	0.90,0.90	0.608	0.82	0.82,0.82	0.68	0.95	0.95,0.95	0.546
Sex									
male	1.00			1.00			1.00		
female	0.74	0.74,0.74	0.519	0.37	0.37,0.37	0.62	0.81	0.81,0.81	0.529
Marital status									
married/defacto	1.00			1.00			1.00		
separated/divorced	2.44	2.44,2.44		1.13	1.13,1.13		1.38	1.38,1.38	
widowed	0.95	0.94,0.95		0.31	0.31,0.31		0.77	0.77,0.77	
single	2.14	2.14,2.14	0.596	1.9	1.90,1.90	0.584	1.54	1.54,1.54	0.551
Race									
white	1.00			1.00			1.00		
black	1.29	1.29,1.29		0.75	0.75,0.75		0.62	0.62,0.62	
Pacific or Asian	0.59	0.59,0.59		0.66	0.66,0.66		0.49	0.48,0.49	
American Indian or Alaskan Native	1.69	1.68,1.69	0.522	1.25	1.25,1.25	0.512	1.04	1.04,1.04	
All predictors combined			0.682			0.738			0.624

Table 4.3 (continued) Associations between the selected demographic variables and health behaviours in the BRFSS health survey data set

	Current smoking and alcohol consumption			Undertook exercise in past 30 days other than regular job			Obese		
	OR	95% CI	AUC	OR	95% CI	AUC	OR	95% CI	AUC
State of residence									
Alaska [a]	1.73	1.72,1.74		0.96	0.96,0.96		1.21	1.21,1.21	
California	1.00			1.00			1.00		
Connecticut	1.17	1.16,1.17		0.96	0.95,0.96		0.92	0.92,0.92	
Georgia	1.2	1.20,1.20		0.76	0.76,0.76		1.27	1.27,1.27	
Hawaii	1.38	1.37,1.38		1.06	1.07,1.06		0.85	0.85,0.85	
Iowa	1.39	1.39,1.39		0.81	0.80,0.81		1.27	1.27,1.27	
Kentucky	1.14	1.14,1.14		0.58	0.58,0.58		1.38	1.38,1.38	
Louisiana	1.49	1.49,1.49		0.56	0.56,0.56		1.5	1.49,1.50	
Michigan	1.47	1.47,1.47		0.83	0.83,0.83		1.37	1.36,1.37	
New Jersey	1.04	1.04,1.04		0.77	0.77,0.77		0.99	0.99,0.99	
New Mexico	1.26	1.26,1.26		0.92	0.92,0.92		1.03	1.03,1.03	
Utah	0.8	0.79,0.80		1.17	1.17,1.17		1.02	1.02,1.03	
Washington	1.28	1.27,1.28	0.542	1.20	1.20,1.20	0.573	1.09	1.09,1.09	0.547
Year of survey									
per additional year	0.99	0.99,0.99	0.525	1.00	1.00,1.00	0.499	1.03	1.03,1.03	0.529
Age									
per additional 5-year age category	0.87	0.87,0.87	0.616	0.93	0.93,0.93	0.573	0.99	0.99,0.99	0.516
Sex									
male	1.00			1.00			1.00		
female	0.33	0.33,0.33	0.627	0.8	0.80,0.81	0.528	0.93	0.93,0.93	0.507
Marital status									
married/defacto	1.00			1.00			1.00		
separated/divorced	2.03	2.03,2.03		0.67	0.67,0.67		1.19	1.19,1.19	
widowed	0.73	0.73,0.74		0.49	0.49,0.49		0.94	0.94,0.94	
single	2.5	2.50,2.51	0.606	0.86	0.86,0.86	0.566	1.17	1.17,1.17	0.525
Race									
white	1.00			1.00			1.00		
black	0.98	0.98,0.99		0.64	0.64,0.64		1.75	1.75,1.76	
Pacific or Asian	0.55	0.55,0.56		1.01	1.01,1.01		0.44	0.44,0.44	
American Indian or Alaskan Native	1.48	1.48,1.48	0.510	0.72	0.72,0.72	0.522	1.60	1.60,1.60	0.54
All predictors combined			0.713			0.623			0.589

Of the fitted logistic models, the model predicting binge drinking (AUC 0.74) displayed the greatest ability to discriminate and the model for predicting obesity (AUC 0.59) showed the least ability to discriminate behaviour.

Table 4.4 shows the associations between post-diagnosis survival time and probability of each pre-diagnosis health behaviour. Each line presents results from separate Cox regression models, for each health behaviour. The columns present results from three separate models: the unadjusted model with the probability of behaviour one year prior to diagnosis as the only predictor; the one-year lag model adjusted for age and cancer stage at diagnosis; and the adjusted model with a 10-year lag. The hazard ratios reported show the impact of a 0.1 increase in the probability of participating in that behaviour. Tables 4.5 and 4.6 provide the same results for the OSCC and OAC histological types separately. Both adjusted variables (age and cancer stage at diagnosis) are significant predictors of survival (Table 4.7). Result for the 5-year lag model (Table 4.8) are similar to the corresponding one-year lag models.

Table 4.4 Association between survival time and probability of pre-diagnosis health behaviour; all oesophageal cancers.

	1 year lag, unadjusted			1 year lag, adjusted [b]			10 year lag, adjusted [b]		
Health behavior	HR[a]	95% CI	P value	HR[a]	95% CI	P value	HR[a]	95% CI	P value
Current smoker	0.99	0.99,1.01	0.252	1.20	1.18,1.22	<0.001	1.18	1.15,1.21	<0.001
Alcohol - Heavy drinking	0.53	0.50,0.56	<0.001	0.82	0.76,0.88	<0.001	1.16	1.04,1.30	0.011
Alcohol - Binge drinking	0.76	0.74,0.77	<0.001	0.96	0.93,0.99	0.012	1.04	0.99,1.08	0.093
Current smoking and alcohol consumption	0.87	0.83,0.92	<0.001	1.93	1.79,2.07	<0.001	1.69	1.56,1.84	<0.001
Undertook exercise in past 30 days other than regular job	0.78	0.77,0.79	<0.001	0.82	0.81,0.84	<0.001	0.80	0.78,0.83	<0.001
Obese	0.95	0.94,0.97	<0.001	1.04	1.03,1.06	<0.001	1.10	1.07,1.14	<0.001

Abbreviations: CI, confidence interval; HR, hazard ratio

[a] The hazard ratio describes the impact of a 0.1 increase in the probability of having the specified health behaviour.

[b] Adjusted for age and cancer stage at diagnosis

Table 4.5 Association between survival time and probability of pre-diagnosis health behaviour; oesophageal squamous cell carcinomas.

Health behavior	1 year lag, unadjusted			1 year lag, adjusted [b]			10 year lag, adjusted [b]		
	HR[a]	95% CI	P value	HR[a]	95% CI	P value	HR[a]	95% CI	P value
Current smoker	0.99	0.97,1.01	0.215	1.20	1.17,1.23	<0.001	1.19	1.16,1.22	<0.001
Alcohol - Heavy drinking	0.50	0.45,0.56	<0.001	0.78	0.69,0.88	<0.001	1.30	1.08,1.57	0.007
Alcohol - Binge drinking	0.75	0.73,0.78	<0.001	0.95	0.90,1.00	0.035	1.09	1.02,1.17	0.013
Current smoker and ≥1 alcoholic drink /day	0.85	0.79,0.93	<0.001	1.93	1.72,2.16	<0.001	1.68	1.51,1.88	<0.001
Did exercise in past 30 days other than regular job	0.78	0.76,0.79	<0.001	0.82	0.80,0.85	<0.001	0.79	0.77,0.82	<0.001
Obese	0.97	0.94,0.99	0.004	1.07	1.04,1.10	<0.001	1.08	1.04,1.11	<0.001

Abbreviations: CI, confidence interval; HR, hazard ratio

[a] The hazard ratio describes the impact of a 0.1 increase in the probability of having the specified health behavior.

[b] Adjusted for age and cancer stage at diagnosis

Table 4.6 Association between survival time and probability of pre-diagnosis health behaviour; oesophageal adenocarcinomas.

	1 year lag, unadjusted			1 year lag, adjusted [b]			10 year lag, adjusted [b]		
Health behavior	HR[a]	95% CI	P value	HR[a]	95% CI	P value	HR[a]	95% CI	P value
Current smoker	1.00	0.98,1.01	0.625	1.20	1.18,1.23	<0.001	1.18	1.16,1.21	<0.001
Alcohol - Heavy drinking	0.55	0.51,0.59	<0.001	0.85	0.77,0.93	<0.001	1.10	0.95,1.26	0.216
Alcohol - Binge drinking	0.76	0.74,0.78	<0.001	0.97	0.93,1.01	0.121	1.01	0.96,1.07	0.722
Current smoker and ≥1 alcoholic drink /day	0.89	0.83,0.95	<0.001	1.93	1.76,2.11	<0.001	1.79	1.65,1.94	<0.001
Did exercise in past 30 days other than regular job	0.78	0.77,0.80	<0.001	0.82	0.81,0.84	<0.001	0.83	0.81,0.85	<0.001
Obese	0.94	0.92,0.96	<0.001	1.03	1.00,1.05	0.028	1.07	1.04,1.10	<0.001

Abbreviations: CI, confidence interval; HR, hazard ratio

[a] The hazard ratio describes the impact of a 0.1 increase in the probability of having the specified health behavior.

[b] Adjusted for age and cancer stage at diagnosis

Table 4.7 Relationship between adjusted variables and survival time

	Relationship with survival time		
	HR	95% CI	Pseudo R^2
Age			
per additional 5 year age category	1.09	1.09,1.09	0.026
Cancer stage			
- Stage I	1.00		
- Stage II	1.17	1.13,1.22	
- Stage III	1.58	1.52,1.64	
- Stage IV	3.34	3.23,3.46	
- Unknown	2.5	2.41,2.60	0.151

Abbreviations: CI, confidence interval; HR, hazard ratio

Table 4.8 Association between survival time and 5-year pre-diagnosis health behaviour.

Health behaviour	All oesophageal cancer 5 year lag, adjusted [b]			OSCC 5 year lag, adjusted [b]			OAC 5 year lag, adjusted [b]		
	HR[a]	95% CI	P value	HR[a]	95% CI	P value	HR[a]	95% CI	p-value
Current smoker	1.18	1.17,1.20	<0.001	1.19	1.16,1.22	<0.001	1.18	1.16,1.21	<0.001
Alcohol - Heavy drinking	0.98	0.90,1.06	0.568	0.85	0.75,0.98	0.026	1.05	0.95,1.16	0.343
Alcohol - Binge drinking	1.00	0.97,1.03	0.818	0.97	0.92,1.02	0.21	1.01	0.97,1.05	0.515
Current smoking and alcohol consumption	1.75	1.64,1.87	<0.001	1.68	1.51,1.88	<0.001	1.79	1.65,1.94	<0.001
Undertook exercise in past 30 days other than regular job	0.81	0.80,0.83	<0.001	0.79	0.77,0.82	<0.001	0.83	0.81,0.85	<0.001
Obese	1.07	1.05,1.09	<0.001	1.08	1.04,1.11	<0.001	1.07	1.04,1.10	<0.001

Abbreviations: CI, confidence interval; OAC, oesophageal adenocarcinoma; OSCC, oesophageal squamous cell carcinoma; HR, hazard ratio

[a] The hazard ratio describes the impact of a 0.1 increase in the probability of having the specified health behaviour.

[b] Adjusted for age and cancer stage at diagnosis

Smoking one year prior to diagnosis appears to be unrelated to survival until adjustment for age and disease stage at diagnosis. In the adjusted model, each 0.1 increase in the probability of pre-diagnosis smoking is associated with a 20% (HR 1.20, 95% CI 1.18-1.23) relative increase in post-diagnosis hazard with no discernible difference in results for OSCC and OAC subgroups.

Results for alcohol consumption are mixed. When using behaviour one year prior to diagnosis as the predictor, a 0.1 increase in the probability of heavy drinking appears to be protective of survival even after adjustment for age and cancer stage at diagnosis (HR 0.82, 95% CI 0.76-0.88). However, when looking at behaviour 10 years prior to diagnosis, the adjusted model finds heavy drinking to be detrimental to post-diagnosis survival in OSCC (HR 1.30, 95% CI 1.08-1.57) and with no discernible association in OAC (HR 1.10, 95% CI 0.95-1.26). The pattern of results for binge drinking is quite similar.

A 0.1 increase in the probability of concurrently smoking and drinking ≥1 standard drink per day in the year prior to diagnosis is associated with double the risk of death (HR=1.93, 95% CI 1.79- 2.07), after adjustment for age and cancer stage with no difference between OSCC and OAC.

After adjustment, a 0.1 increase in probability of obese one year prior to diagnosis is associated with an apparently trivial increase in post-diagnosis hazard (HR 1.04, 95% CI 1.03-1.06). A slightly larger hazard (HR 1.10, 95% CI 1.07-1.14) was recorded for a 0.1 increase in the probability of obese 10 years prior to diagnosis. A 0.1 increase in the probability of exercise outside employment one year prior to diagnosis is associated with improved survival (HR 0.82, 95% CI 0.81-0.84) with little difference between OSCC and OAC.

4.2.6 **Discussion**

The results above appear to support of the proposition that demographic-derived estimates of the probability of health behaviours can assist in identifying association between pre-diagnosis health behaviour and post-diagnosis survival time in oesophageal cancer. The hazard ratios quoted in this paper show the increased hazard of death associated with each additional 0.1 probability of the health behaviour of interest. That is, the association between the estimated likelihood of engaging in a particular behaviour and survival time is reported. This is quite different from the association between the actual health behaviour and survival time and more difficult to interpret. Never-the-less, there is consistency between the results of the present study and previously published results: especially in the presence and direction of associations.

A 0.1 increase in the probability of smoking one year prior to diagnosis, adjusted for age and cancer stage at diagnosis, had an estimated HR of 1.20 (95% CI 1.18-1.22) in oesophageal cancer survival. This association is consistent with findings from previous meta analyses such as HR 1.41 (95% CI 1.22,1.64) (Kuang et al., 2016) for smoking status at time of diagnosis in mainly OSCC patients and HR 1.19 (95% CI 1.04,1.36) for ever smoking (Fahey et al., 2015) in OSCC (although no evidence of association in OAC). Some more recently published studies found similar statistically significant HRs including HR=1.28 (Ma et al., 2016) and HR=1.34 (P. Sun et al., 2016) both from China, and HR=1.22 from a study across two sites in US and Canada (A. Spreafico et al., 2017). In contrast, recent results from Japan HR=0.97 (Okada et al., 2017) failed to find evidence of association between pre-diagnosis smoking and post-diagnosis survival time. A study from South Africa reported an unadjusted HR=0.92 (Dandara, Robertson, Dzobo, Moodley, & Parker, 2015) but the present

study has shown the importance of adjustment for confounders such as age and cancer stage at diagnosis.

The current analyses found that increased probability for ‘at risk’ alcohol consumption in the year prior to diagnosis were generally protective of survival but that a 0.1 increase in ‘at risk’ alcohol behaviour 10 years prior to diagnosis was detrimental to survival in OSCC (heavy drinking HR 1.30 95% CI 1.08-1.57, binge drinking HR 1.09, 95% CI 1.02-1.17). The 10 year results are consistent with a previous meta-analysis (4) which found that ever drinking alcohol produced a significant increase in hazard (HR 1.36, 95% CI 1.15, 1.61) in OSCC but non-significant HR of 1.08 (95% CI 0.85, 1.37) in OAC although ever drinking and ‘at risk’ drinking are widely separate. More recent results from China HR=1.58 (Ma et al., 2016), HR=1.45 (P. Sun et al., 2016) and Japan HR=2.37 (95% CI 1.24,4.53) (Okada et al., 2017) also support the detrimental impact of pre-diagnosis alcohol consumption on post-diagnosis survival.

The unexpectedly protective result for alcohol consumption one-year prior to diagnosis could indicate insufficient adjustment for confounding (such as comorbidities or health symptoms) or weaknesses in the measurement tool (such as biases in the self-reporting of alcohol consumption in standard drinks).

Previous authors have found that pre-diagnosis smoking and alcohol consumption combined produce a disproportionately high risk to post-diagnosis survival (for example, HR 3.84, 95% CI 2.02,7.32 for OSCC (Thrift, Nagle, Fahey, Russell, et al., 2012)). It has also been found that a 0.1 increase in the probability of concurrent daily smoking and consuming one or more alcoholic drinks per day one year prior to diagnosis, adjusted for age and cancer stage at diagnosis, had a relatively high estimated HR of 1.93 (95% CI 1.79,2.07).

A 0.1 increase in the probability of obese one year prior to diagnosis was associated with slightly higher risk of death adjusted HR=1.04 (95% CI 1.03,1.06) mainly associated with OSCC (HR 1.07 95% CI 1.04,1.10). The association seems small and the literature on obesity is sparse with mixed findings. One review found pre-diagnosis obesity could be associated with higher risks of death in cancer (specifically breast, prostate and colorectal cancers) (Parekh, Chandran, & Bandera, 2012) but a later study reported that pre-diagnostic obesity increased hazard for all cancers except cancers of the upper digestive tract (obese compared to normal weight HR 0.87, 95% CI 0.62,1.22) (Reichle, Peter, Concin, & Nagel, 2015). More recently a North American study (A. Spreafico et al., 2017) found recalled obesity in early adulthood was associated with lower survival times than normal weight (HR 1.77, 95% CI 1.25, 2.51). The measure of obesity available in this study may not be optimal.

Finally, it was found that a 0.1 increase in probability of pre-diagnosis physical activity outside of the workplace was associated with improved survival (adjusted HR=0.82, 95% CI 0.81,0.84). This is consistent with a recent review (Lynch & Leitzmann, 2017) which found the relative risk of death between the highest versus lowest category of physical activity to be 0.71 (95% CI 0.57,0.89) for oesophageal cancer.

Strengths and weaknesses

Our analyses using estimated probability for health behaviours has produced results which have some face validity. A strength of this example is that the data sets used are large, public domain and well understood. Any interested researcher can reproduce, refine and/or extend these analyses using the same data sets.

Both the data sets and the analysis technique used have some limitations and weaknesses. In relation to the data sets, there are response biases within the BRFSS (Schneider et al., 2012) which the sampling weights may not have fully addressed. Further,

the measures of behaviour available are limited and are dictated by the existing data base which was designed for other purposes and is not optimized for the research question of this paper.

For the model, estimating the probability of a behaviour is less accurate than a direct measure of behaviour and conveys less information about that behaviour: so, will have less power for detecting associations. There may be residual confounding from unmeasured variables (such as education, socio-economic status or comorbidities). Finally, omitting interactions with year may have contributed to the apparent lack of difference in outcomes between behaviour one, five and ten years prior to diagnosis.

4.2.7 **Conclusion**

The rarer the disease, the less feasible it is to conduct either prospective cohort studies or record linkage (retrospective cohort) studies. Retrospective data collection (including case-control studies) are fraught with recall and survivor biases. Exploiting existing data provides cost-effective opportunities for investigations but may require different methodologies.

Analyses of the associations between estimated probability for pre-diagnosis health behaviour (based on demographic characteristics) and survival time in oesophageal cancer produced results with some face validity. Expressing associations in units of changes in the probability of the health behaviour was cumbersome. However, the required data are already available, allowing relatively quick and inexpensive investigations of possible associations between pre-diagnosis behaviour and post-diagnosis outcomes for relatively rare diseases. And of course, most diseases are relatively rare.

4.3 Critique of the method

The above published peer-reviewed paper describes a model-based method for addressing 100% missing data in the SEER Cancer registry data set. The missing data was replaced with estimates of the probability of each behaviour rather than actual behaviour. The fitted Cox models describe the association between the estimated probability of each pre-diagnosis behaviour and post-diagnosis survival time in OC. All results are expressed in units of estimated probability of the behaviour.

This is a very unusual measurement scale and is likely to have poor clinical interpretability. The hazard ratios quoted, describing relative effects of probabilities, are cognitively difficult for humans (Reyna & Brainerd, 2008) and there does not appear to be any precedents of using this type of measurement.

The actual numeric values of the hazard ratios can be, and have been, manipulated. Hazard ratios are presented for a 0.1 increase in the estimated probability of displaying the behaviour mainly to produce hazard ratios of about the same magnitude as hazard ratios associated with measured behaviour. Choosing a 0.01 or 0.2 increase in estimated probabilities would have been just as valid but would have changed the magnitude of the hazard ratios. No justification was provided for visually comparing the hazard ratios from a 0.1 increase in the estimated probability of displaying the behaviour against the previously published hazard ratios associated with actual behaviours. A comparison of hazard ratios obtained from actual behaviour with those obtained from estimated probability of behaviour is needed but would require data on the actual behaviour of the cancer registry cases. The lack of such data is the reason for this research.

To avoid using the estimated probability of a behaviour the probability could be converted into an imputed behaviour. For example, a cutpoint could be applied such that

probabilities above the cutpoint are assigned to ‘behaviour present’ and those below are assigned to ‘behaviour absent’. This would mean that every OC case from the same demographic group would be assigned the same behaviour. Obviously, this is not realistic; not every individual within the same demographic group follows the same behaviour. Alternatively, OC cases could be randomly assignment to the behaviour present category relative to their estimated probability (the higher the estimated probability, the more likely that OC case would be randomly assigned to behaviour present). But random assignment results in misclassification. For example, true smokers will be assigned a non-smoking behaviour and vice versa. These issues relate to the imputation-based methods for addressing 100% missing variables which shall be considered in more detail in the following Chapters. Here the possibility that model-based methods could be used in conjunction with imputation-based methods is simply flagged.

Given the difficulties with interpreting the hazard ratios for estimated probabilities, the focus could be reduced to whether or not there is evidence of a relationship and if so, in which direction. Although less information, these observations could still have important roles to play. For example, as the analysis is relatively quick and simple, the modelling of behaviour as a 100% missing variable could provide an intermediate check on the feasibility of a larger, more complex study such as cohort design which could give more interpretable results.

In the above analyses the estimated probability for pre-diagnosis health behaviour was derived using a logistic model in which behavioural status was regressed on observed demographic characteristics. Compare this with *“the propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics. The estimated propensity score is the predicted probability of treatment derived from the fitted regression model”* (Austin, 2011, p. 403). Health behaviour is an

exposure rather than a treatment but all else is similar. The ‘estimated probability of behaviour’ could be considered a parallel to a ‘propensity score’ and among other things, propensity scores are useful in matching and stratification (Austin, 2011). Propensity score matching forms pairs of treated and untreated subjects based on their propensity allowing their outcomes to be compared in a matched analysis. An equivalent process in the current application could be to match SEER cancer registry cases with BRFSS health survey respondents based on their estimated probability for pre-diagnosis health behaviour. Comparing outcomes between these pairs would not be relevant as the BRFSS respondents do not have cancer outcomes. However, matching and stratifying on estimated probability for pre-diagnosis health behaviour could be helpful in improving both the cold deck imputation and confounder adjustment methods which will be described in Chapter 5 and 6.

“All models are wrong, but some are useful.” (Box & Draper, 1987)

Standard modelling methods were used throughout. Most quantitative researchers will be familiar with logistic regression and Cox proportional models.

It was noted in Table 4.1 that some behaviours are quite uncommon: 4.8% were at risk of heavy drinking and just 2.7% were smokers and regular drinkers. With small numbers of events in logistic regression the maximum likelihood estimation is biased towards underestimating the probabilities of these events (King & Zeng, 2001). Firth’s correction on the maximum likelihood procedure addresses this bias. However, the bias occurs when there are small numbers of events rather than small rate of events. Even at a rate of 2.7% there were still 39,431 eligible BRFSS health survey respondents with smoking and regular drinking. With such large numbers of events, standard maximum likelihood approach is effectively unbiased and would have given identical results to the models fitted using Firth’s method.

A major issue which has not yet been addressed in the analyses above is the cumulative error across the models. The logistic model-derived estimates of the probability of behaviour contain errors (for example, the AUC for the logistic models varied between 0.738 for at risk of binge drinking to just 0.589 for obesity). It is simply not possible to accurately predict behaviour based on demographic variables alone. But the problem is that these prediction errors have not been carried through to the Cox regression analysis. That is, the Cox regression assumes the predictor variable (the estimated probability of the behaviour) was measured without error. The estimated probabilities of obesity are treated the same as the estimated probability of binge drinking, even though the estimated probabilities of binge drinking appear to better discriminate behaviour according to the AUC statistics. By ignoring this error source, the confidence intervals from the Cox regression are narrower than they should be.

4.4 Closing comments

This Chapter developed and evaluated a model-based method for addressing 100% missing data for pre-diagnosis health behaviours among OC cases in the SEER cancer registries data set. This partially addresses the second aim of this thesis.

However, the second aim was to develop the method such that it addressed the first aim of describing the relationship between pre-diagnosis health behaviours and post-diagnosis survival in OC. In this, the method developed in this Chapter has been quite unsuccessful. For example, the hazard ratios produced from the Cox model failed to quantify the association between pre-diagnosis health behaviour and survival, concentrating on predicted health behaviour instead of actual health behaviour. Also, the confidence intervals

around these hazard ratios failed to capture all of the error present in the analyses and are incorrectly narrow.

With these somewhat difficult problems arising from this model-based method of analysis the following chapter considers whether an alternate approach could produce more interpretable results. In Chapter 3 it was noted that imputation may be an alternative to model-based methods for addressing 100% missing data and this possibility shall be explored further next.

In Chapters 5 and 6 an alternate method for addressing 100% missing data, this time imputing the missing values, while retaining the same data sets and variables described in Chapter 3.

Chapter 5 An imputation-based approach to addressing missing pre-diagnosis health behaviour

5.1 Background

In the previous chapter a model-based approach was used to estimate the probability of pre-diagnosis behaviours for OC cases and then used these estimated probabilities as predictors of post-diagnosis survival using Cox regression models. It was observed that the resulting hazard ratios were difficult to interpret and that estimating the precision these hazard ratios may be difficult (the calculated confidence intervals were too narrow). In this Chapter an imputation-based approach for replacing the missing behavioural data is introduced.

Imputation provides actual values for the missing data, avoiding the difficulties of interpreting estimated probability of behaviour. But the values provided by the imputation are not the true behaviours of the OC cases. The imputed behaviour is, at best, an informed guess of the true health behaviour of each OC case.

There are many different approaches to imputation, varying according to the information available to inform the guesswork. In this thesis a method is sought to impute OC cases unknown health behaviour using the known behaviour of the different individuals in the BRFSS health survey. It has been noted previously (Section 1.8) that individuals with the same demographic background are more likely to display similar behaviours than individuals from different backgrounds. So, a reasonable first step is to match each OC case with a demographically similar individual from the BRFSS health survey data set. If health

behaviours are clustered by demographic group, the demographically matched BRFSS respondents should contain some information about the health behaviours of the OC cases.

The amount of information which can be conveyed from the BRFSS health survey respondents about OC cases behaviour will be determined by how strongly health behaviour clusters by demographic group. Given such clustering is likely to be modest, the health behaviours of BRFSS respondents are expected to provide little real information about the health behaviour of the demographically similar OC cases. It is certainly possible to detect these weak relationships, but sample sizes may need to be quite large.

Cold deck imputation is the process of using 'matching' data records from an external data set as donors of missing data. In this thesis, BRFSS health survey respondents are matched to OC cases according to age group, gender, marital status, race, State of residence and year. For each OC case, a matching BRFSS health survey data record is selected at random to 'donate' their health behaviour data. The donated health behaviour becomes the OC case's imputed behaviour.

In effect this is saying, "because Ms A is a smoker and because Ms A is of the same age, sex, marital status, race, State of residence and recording year as Ms B, then we'll call Ms B a smoker too". Obviously, these imputed values will be incorrect often. Ms B may be misclassified as a smoker when in truth she has never smoked. If smoking presents a hazard to survival, then misclassifying smokers as non-smokers will make the non-smoking group look like it has worse survival than it really does. Alternatively, classifying non-smokers as smokers, will make the smoker group look like it has better survival than it really does. The net impact is to attenuate any measure of risk towards the null. Section 5.2 explores cold deck imputation in more detail, quantifying the attenuation and factors which affect it.

If the misclassification rate can be measured, the associated attenuation can be corrected for during the analysis. Section 5.3 presents a published research article in which cold deck imputation is applied to estimate relative risks for one-year survival in OC while estimating and correcting for the misclassification error.

Section 5.4 critiques the strengths and weakness of the algorithm and results. The coding approach used for the analyses in this Chapter are described in the thesis Appendix.

The closing comments in section 5.5 review the material in this Chapter in the context of the research questions of the thesis and foreshadows the need for the additional analyses in next Chapter.

5.2 Attenuation of risk in cold deck imputation

Consider the hypothetical relationship between pre-diagnosis smoking status (yes or no) and 1-year post-diagnosis survival (yes or no). Suppose 30% of cancer cases were smokers and that smoking increases the proportion dying at one year post-diagnosis at one year from 0.4 to 0.6 – a relative risk of 1.5. Suppose that 20% of respondents in the wider population smoke. Suppose each of 10,000 OC cases are matched with randomly selected BRFSS health survey respondents from the same demographic group and impute the OC cases behaviour to be that of their matching BRFSS respondent (i.e. cold deck imputation).

If smoking status does not cluster in demographic groups, the cold deck imputation will contain no information on the smoking status of OC cases, and the scenario summarized in Table 5.1 would be expected. In this case, both smoking OC cases and non-smoking OC cases would be merged to 20% smokers and 80% non-smokers and the relationship between smoking and survival would be completely lost.

Table 5.1 Expected attenuated relative risk when the true relative risk is 1.5, the samples size is 10,000 and the cold deck imputation of behaviour is no better than random chance.

a) Results of matching categorised by true smoking status of cancer cases

true smoking status	expected number (a)	true death rate (b)	imputed smoking status	expected number (c)	expected deaths (d) = (b)x(c)
smoker (30%)	3,000	0.6	smoker (20%) non-smoker (80%)	3,000x0.2=600 3,000x0.8=2,400	600x0.6=360 2,400x0.6=1440
non smoker (70%)	7,000	0.4	smoker (20%) non smoker (80%)	700x0.2=1,400 700x0.8=5,600	1,400x0.4=560 5,600x0.4=2240

b) Results of matching categorised by imputed smoking status

imputed smoking status	expected number from (c)	expected deaths from (d)	expected death rate (e)=(d)/(c)	attenuated relative risk
smoker	600+1,400=2,000	360+560=920	920/2,000=0.46	
non smoker	2,400+5,600=8,000	1,440+2,240=3,680	3,680/8,000=0.46	
				0.46/0.46=1.0

If, however, the cold deck matching increases the proportions of correct matches (smokers matched to smokers and non-smokers matched to non-smokers) by 10% the scenario shown in Table 5.2 would apply. The 10% increase in correct matches would mean 22% of smokers could be expected to be matched with smokers (instead of 20%) and 88% of non-smokers to be matched with non-smokers (instead of 80%).

Table 5.2 Expected attenuated relative risk when the true relative risk is 1.5, the samples size is 10,000 and the cold deck imputation produces 10% more correct coding of behaviour than would random chance.

a) Results of matching categorised by true smoking status of cancer cases

true smoking status	expected number (a)	true death rate (b)	imputed smoking status	expected number (c)	expected deaths (d) = (a)x(c)
smoker (30%)	3,000	0.6	smoker non-smoker	3,000x0.22=660 3,000x0.78=2340	660x0.6=396 2340x0.6=1404
non-smoker (70%)	7,000	0.4	smoker non smoker	7,000x0.12=840 7,000x0.88=6160	840x0.4=336 6,160x0.4=2464

b) Results of matching categorised by imputed smoking status

imputed smoking status	expected number from (c)	expected deaths from (d)	expected death rate	attenuated relative risk
smoker	660+840=1,500	396+336=732	732/1,500=0.488	
non-smoker	2,340+6,160=8,500	1,404+2,464=3,868	3,868/8,500=0.455	
				0.488/0.455=1.07

Table 5.2 demonstrates a scenario where informative cold deck matching flows through to an attenuated relative risk is 1.07 with an associated 95% confidence interval of 1.01 to 1.14. This attenuated relative risk is statistically significantly different from the null (RR=1) with a p-value of 0.0159. Therefore, in this scenario the cold deck imputation has preserved a positive signal of the association between pre-diagnosis smoking and post-diagnosis survival but information on the strength of this relationship − that the relative risk is equal to 1.5 − has been under-estimated.

For completeness, Table 5.3 demonstrates that in the unlikely event that cold cell imputation leads to entirely accurate assignment of behaviours (introduces no new error) the true relative risk will be preserved.

Table 5.3 Expected attenuated relative risk when the true relative risk is 1.5, the samples size is 10,000 and the cold deck imputation produces entirely correct coding of behaviour.

a) Results of matching categorised by true smoking status of cancer cases

true smoking status	expected number (a)	true death rate (b)	imputed smoking status	expected number (c)	expected deaths (d) = (a)x(c)
smoker (30%)	3,000	0.6	smoker non-smoker	3,000 0	3,000x0.6=1,800 0
non-smoker (70%)	7,000	0.4	smoker non smoker	0 7,000	0 7,000x0.4=2,800

b) Results of matching categorised by imputed smoking status

imputed smoking status	expected number from (c)	expected deaths from (d)	expected death rate (e)=(d)/(c)	attenuated relative risk
smoker	3,000	1,800	1,800/3,000=0.6	
non-smoker	7,000	2,800	2,800/7,000=0.4	
				0.6/0.4=1.5

Table 5.2 demonstrates that cold deck imputation could, in some circumstances, preserve information on the existence of association between the pre-diagnosis health behaviour and the one-year survival in OC. Indeed, if all of the parameters in the above hypothetical scenario were able to be quantified, the ‘true’ relative risk could be derived from

the attenuated relative risk. That is, in Table 5.2, the attenuated relative of risk of 1.07 could be equated with a true relative risk of 1.5 via the known parameters: sample size; the increase in correct coding produced by cold deck imputation; the smoking rate in the OC population; and the smoking rate in the general population.

5.2.1 **Parameters affecting the results**

For cold deck imputation to produce statistically significant evidence of a relationship between OC cases health behaviour and, say, one-year survival the following are required:

a) the relationship must exist and be strong enough to be detected;
b) the cold deck matching algorithm must provide sufficient information about the OC cases' health behaviour; and
c) the sample size must be large enough to detect the signal.

Here some quantitative examples are presented of how strong the relationship may need to be, how informative the cold deck imputation may need to be and how large a sample size may need to be to produce statistically significant evidence of the relationship.

Figures 5.2 to 5.4 present expected relative risks with associated 95% confidence intervals for a range of scenarios. Figure 5.1 shows how the formulas for these statistics were derived by formalising the calculations used in Tables 5.1 to 5.3.

Figure 5.1 Calculation of the post-matching confidence intervals

Let n represent the number of cases
p_c represent the proportion of OC cases who smoked
p_p represent the proportion of the general population who smoked
r_s represent the risk of death (within a given time period) for OC cases who smoked
r_n represent the risk of death (within a given time period) for OC cases who did not smoked
m represent improvement in matching derived from the matching variables

Following the layout of Tables 5.1 to 5.3 the example can be generalised as follows:

a) Results of matching categorised by true smoking status of cancer cases

true smoking status	expected number (a)	true death rate (b)	imputed smoking status	expected number (c)	expected deaths (d) = (a)x(c)
Smoker	np_c	r_s	smoker non-smoker	$np_c p_p(1+m)$ $np_c\left[1-p_p(1+m)\right]$	$np_c p_p(1+m)r_s$ $np_c\left[1-p_p(1+m)\right]r_s$
non-smoker	$n(1-p_c)$	r_n	smoker non smoker	$n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]$ $n(1-p_c)\left(1-p_p\right)(1+m)$	$n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]r_n$ $n(1-p_c)\left(1-p_p\right)(1+m)r_n$

b) Results of matching categorised by imputed smoking status

imputed smoking status	expected number from (c)	expected deaths from (d)	expected death rate	attenuated relative risk
smoker	b= $np_c p_p(1+m)+$ $n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]$	a= $np_c p_p(1+m)r_s+$ $n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]r_n$	a/b	
non-smoker	d= $np_c\left[1-p_p(1+m)\right]+$ $n(1-p_c)\left(1-p_p\right)(1+m)$	c= $np_c\left[1-p_p(1+m)\right]r_s+$ $n(1-p_c)\left(1-p_p\right)(1+m)r_n$	c/d	
				$\frac{a/b}{c/d}$

Thus the components are as follows:

a. The expected number of cases classified as smokers who die within the time period
$$np_c p_p(1+m)r_s + n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]r_n$$
b. The expected number of cases classified as smokers
$$np_c p_p(1+m) + n(1-p_c)\left[1-\left(1-p_p\right)(1+m)\right]$$
c. The expected number of cases classified as non-smokers who die within the time period
$$np_c\left[1-p_p(1+m)\right]r_s + n(1-p_c)\left(1-p_p\right)(1+m)r_n$$
d. The expected number of cases classified as non-smokers
$$np_c\left[1-p_p(1+m)\right] + n(1-p_c)\left(1-p_p\right)(1+m)$$

Which can be used to construct the attenuated relative risk (Altman, 1990) as

$$\widehat{RR} = \frac{a/b}{c/d}$$

and the variance of the log of the attenuated relative risk (Altman, 1990) as

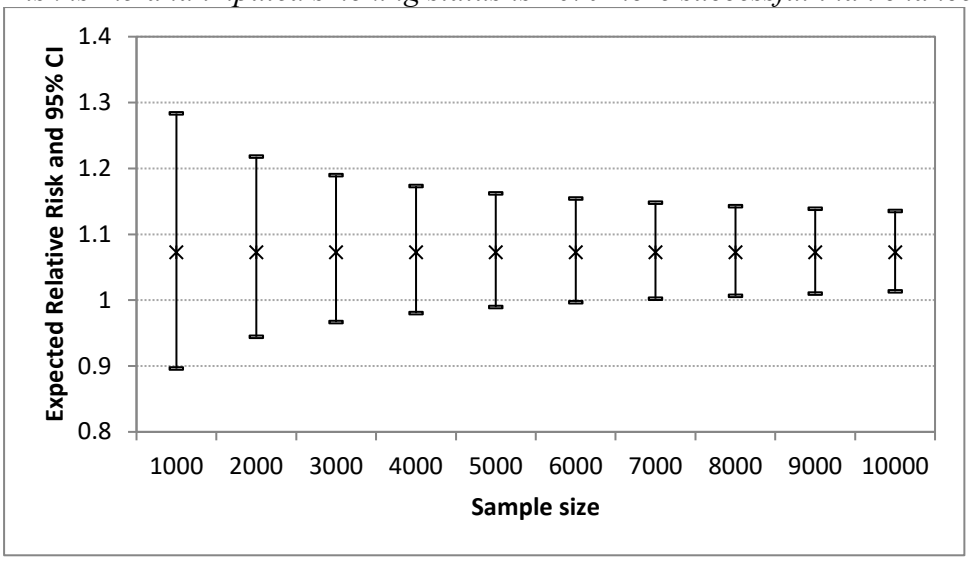
$$\text{var}\left[\log\left(\widehat{RR}\right)\right] = \frac{1}{a} + \frac{1}{c} - \frac{1}{b} - \frac{1}{d}$$

The associated 95% confidence interval is

$$\exp\left(\ \ln\left(\widehat{RR}\right) \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{b} - \frac{1}{d}}\right)$$

Figure 5.2 shows the expected relative risks and associated 95% confidence intervals from the scenario in Table 5.2 but with varying sample sizes. That is, the true relative risk between smoking and one-year survival is 1.5, 30% of OC cases smoked ($p_c = 0.3$), 20% of the wider population smoked $(p_p = 0.2)$, cold deck imputation is 10% better than random chance at imputing OC cases smoking status ($m = 0.1$) and sample sizes vary from $n = 1000$ to $n = 10{,}000$ in steps of 1000.

Figure 5.2 Expected confidence intervals for sample sizes between 1000 and 10,000 when relative risk is 1.5 and imputed smoking status is 10% more successful than chance.

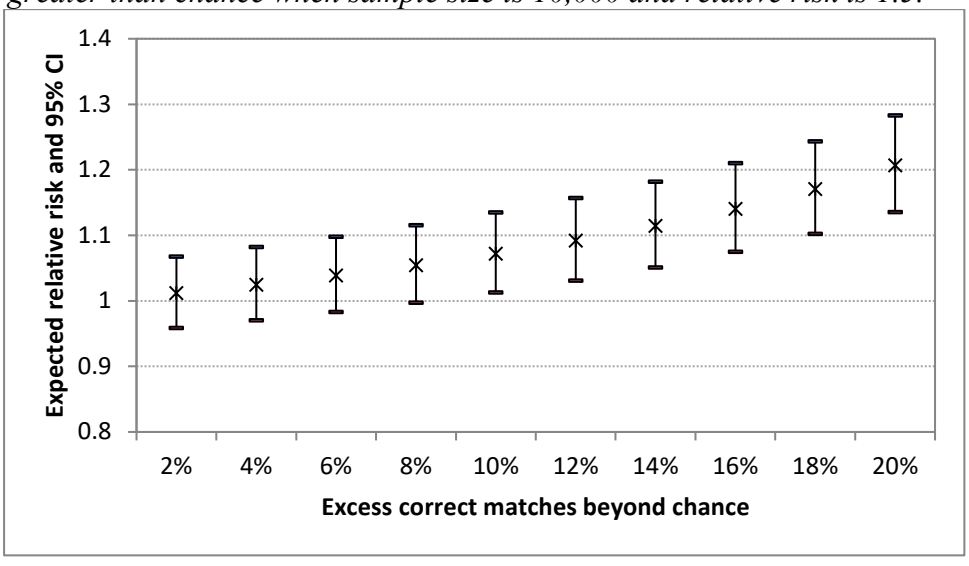


As expected, Figure 5.2 confirms that the attenuated relative risk is not affected by variation in sample size and remains at the 1.07 computed in Table 5.2. Also as expected, the varying sample size changes the width of the confidence interval. The first confidence interval to exclude the null (RR=1) occurs at 7000. That is, a sample size of 7000 or more is needed in order to return statistically significant evidence of relationship when the true relationship has an relative risk of 1.5 and cold deck imputation is 10% better than random chance at assigning the correct behaviour to OC cases.

Equivalent results can be obtained by varying the percentage improvement over chance in the cold deck imputation, where the sample size and relative risk are held constant and by varying the true relative risk while holding percentage improvement over chance and sample size constant. Example results are provided in Figures 5.3 and 5.4.

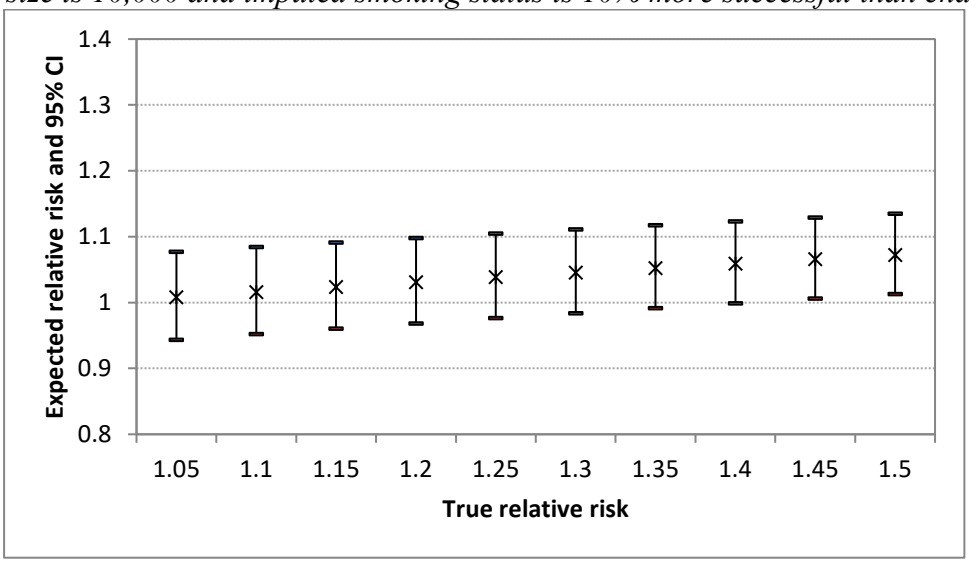
Figure 5.3 shows relative risks and associated 95% confidence intervals obtained in the same scenario, except varying the percentage improvement over chance in the cold deck imputation from 2% to 20% in steps of 2%. That is, the true relative risk is set at 1.5, the sample size at 10,000, the proportion of OC cases who smoke is 30% and the proportion of the wider population who smoke is 20%. As expected, the higher the rate of correct matches in excess of chance, the more likely it to find statistically evidence for the true association. For example, obtaining statistically significant evidence of a relationship when the true relative risk is 1.5 and the sample size is 10,000, would need the cold deck imputation to produce about 10% higher matches than chance.

Figure 5.3 Expected confidence intervals for improvement in behaviour matching from 2% to 20% greater than chance when sample size is 10,000 and relative risk is 1.5.



Finally, in Figure 5.4 the 10,000 sample size, a 10% improvement in correct matches from the imputation relative to chance and the same proportions of smokers are maintained but the true relative risk varies from 1.05 to 1.50 in 0.05 increments. In this scenario, the true relative risk would need to be 1.45 before the cold deck imputation process would provide statistically significant evidence of a relationship between smoking and one-year survival.

Figure 5.4 Expected confidence intervals for relative risks between 1.05 and 1.50 when sample size is 10,000 and imputed smoking status is 10% more successful than chance.



The above results confirm that if the underlying association between health behaviour and one-year survival is large enough and the sample size is large enough and the clustering of behaviour within demographic groups strong enough, the cold deck imputation process can provide statistically significant evidence that the relationship exists.

Unfortunately, providing statistical evidence of the relationship between pre-diagnosis health behaviour and one-year post-diagnosis survival in OC is insufficient. The effect size and its clinical importance for decision making need to be established. (For example, a true relative risk of 1.20 will be clinically less interesting than a true relative risk of 2.10 even if both are found to be statistically significant.) In the scenario presented in Table 5.2 where all parameters were known, the attenuated relative risk of 1.07 could be equated with a true

relative risk of 1.5. But in a real world setting, it would appear difficult to quantify how much the proportion of correctly imputed behaviour from cold deck imputation differs from chance.

At this point the focus turns from the proportion of OC cases receiving correctly imputed behaviour to the proportion receiving incorrectly imputed behaviour. Incorrectly imputed behaviour (such as true smokers imputed to be non-smokers and true non-smokers imputed to be smokers) is a misclassification and misclassification leads to attenuation. If the misclassification rate can be measured the attenuation can be adjusted for during the statistical analysis.

The approach to measuring and correcting for misclassification arising through cold deck imputation is described in the published research paper presented in the next section.

5.3 Augmenting cancer registry data with health survey data with no cases in common: the relationship between pre-diagnosis health behaviour and post-diagnosis survival in oesophageal cancer.

The material in this section has been published as a peer reviewed journal article. The citation is: Fahey, P. P., Page, A., Stone, G., & Astell-Burt, T. (2020). Augmenting cancer registry data with health survey data with no cases in common: the relationship between pre-diagnosis health behaviour and post-diagnosis survival in oesophageal cancer. *BMC Cancer*, 20, 1-11.

In 2020, the journal *BMC Cancer* has a Q2 rating in Cancer Research in the Scimago rating scheme and a 2-year impact factor of 4.069.

Following the CRediT Taxonomy (National Information Standards Organization, 2021), author contributions were:

- Paul Fahey contributed to Conceptualisation, Data curation, Formal analysis, Methodology and Writing the original draft.
- Andrew Page, Glenn Stone and Thomas Astell-Burt contributed to Supervision
- All authors contributed to Review and Editing of the draft paper.

This material has been reformatted to suit the thesis with supplementary materials re-integrated into the paper, tables and figures re-numbered, and text converted from US to Australian English.

5.3.1 **Abstract**

Background: For epidemiological research, cancer registry datasets often need to be augmented with additional data. Data linkage is not feasible when there are no cases in common between data sets. A novel approach to augmenting cancer registry data by imputing pre-diagnosis health behaviour and estimating its relationship with post-diagnosis survival time is presented.

Methods: Six measures of pre-diagnosis health behaviours (focussing on tobacco smoking, ‘at risk’ alcohol consumption, overweight and exercise) were imputed for 28,000 cancer registry data records of US oesophageal cancers using cold deck imputation from an unrelated health behaviour dataset. Each data point was imputed twice. This calibration allowed the estimation of the misclassification rate. Statistical correction for the misclassification is applied to estimate the relative risk of dying within one year of diagnosis

for each of the imputed behaviour variables. Subgroup analyses were conducted for adenocarcinoma and squamous cell carcinoma separately.

Results: Simulated survival data confirmed that accurate estimates of true relative risks could be retrieved for health behaviours with greater than 5% prevalence, although confidence intervals were wide. Applied to real datasets, the estimated relative risks were largely consistent with current knowledge. For example, tobacco smoking status 5 years prior to diagnosis was associated with an increased age-adjusted risk of all cause death within one year of diagnosis for oesophageal squamous cell carcinoma (RR=1.99 95% CI 1.24,3.12) but not oesophageal adenocarcinoma (RR=1.61, 95% CI 0.79,2.57).

Conclusions: A novel imputation-based algorithm for augmenting cancer registry data for epidemiological research has been demonstrated which can be used when there are no cases in common between data sets. The algorithm allows investigation of research questions which could not be addressed through direct data linkage.

5.3.2 **Background**

In 2011 it was estimated that that the cost of maintaining the United States' National Program of Cancer Registries was $US60.77 per case (Tangka et al., 2016). The estimated number of new United States cancer cases in 1999 was 1,291,451 (Centers for Disease Control and Prevention) and 1,762,450 in 2019 (American Cancer Society, 2019) an increase of 36% in 20 years. As in any public investment, there is always a need to maintain, and indeed increase, benefits of cancer registries relative to costs.

The role of cancer registries has changed considerably over time (Roder, Fong, Brown, Zalcberg, & Wainwright, 2014). Since the 1990s, for example, the development of specialised data linkage infrastructure has open wide new research applications (Roder et al.,

2014). However, data linkage may not be feasible in all circumstances. There are still research questions which are waiting for a suitable method of analysis.

Oesophageal cancer is the seventh most common cancer by site (Bray et al., 2018), has low survival (Siegel et al., 2019), and caused an estimated 1 in 20 cancer deaths worldwide in 2018 (Bray et al., 2018). It has been estimated that 71% of male and 59% of female oesophageal cancer deaths in the US arise from modifiable health behaviours: including smoking (50%), alcohol consumption (17%) and excess body weight (27%) (Islami et al., 2018). The impact of pre-diagnosis health behaviour on oesophageal cancer survival is uncertain. As survival times are short, the carry-over effect of pre-diagnosis behaviour may be important, and potentially impact treatment choices (Shitara et al., 2010). Further, as health behaviours in populations change over time (Grant et al., 2017; Méndez, Tam, Giovino, Tsodikov, & Warner, 2016), predicting the impact of behaviour on cancer survival would assist in forecasting future disease burden and health service requirements.

Associations between oesophageal cancer incidence and health behaviour (including tobacco smoking, alcohol consumption, body mass index and physical activity) differ by histological sub-type (Clara Castro et al., 2018; Steevens et al., 2010) with oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC) usually examined separately. Similar differences may exist for survival time (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012).

Nowadays, cancer survival data is generally available through cancer registries (A. H. Siddiqui & Zafar, 2018), but not data on pre-diagnosis health behaviour. Registry data needs to be augmented with additional data collection or linkage to external data sources. Additional data collection can be time consuming, expensive and subject to survivor bias (Smithers et al., 2010) and data linkage needs the same individuals to be present and

identifiable in both data collections and is less feasible for rare disease like oesophageal cancer.

When faced with missing data, researchers sometimes use imputation (De Waal, Pannekoek, & Scholtus, 2011). Imputing data is likely to lead to misclassification of health behaviours (such as smokers classified as non-smokers and vice-versa). However, repeated observations of the same behaviour can be used to quantify, and subsequently correct for misclassification (de Klerk, English, & Armstrong, 1989). In this paper the possibility is investigated that, with large datasets and careful calibration, imputing a completely missing variable could return valid results. An algorithm is described and evaluated for assessing the relationship between pre-diagnosis health behaviours and survival at one-year post-diagnosis for oesophageal cancer where survival is derived from cancer registry data and key health behaviours are fully imputed using unrelated health survey data.

5.3.3 **Methods**

Data sources

Oesophageal cancer cases were extracted from the Surveillance, Epidemiology, and End Results Program (SEER) cancer registries database, which combines data from cancer registries in up to 13 US States covering up to 28% of the US population (Surveillance Epidemiology and End Results (SEER) Program). Available data included patient demographics and outcomes (including survival time).

All records of primary oesophageal cancers diagnosed between 2006 to 2014 were downloaded using the SEER*Stat utility (Surveillance Research Program, 2020). After

excluding 112 cases <35 years of age as atypical, the dataset contained 34,972 oesophageal cancer cases.

Health behaviour data of US residents were extracted from the Behavioural Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention (CDA)). This telephone survey of the adult population of US residents (all States) has been conducted annually since 1984. All 3,018,830 records from 2001 to 2009 were included.

Given that health behaviour can change after diagnosis (Demark-Wahnefried et al., 2005; Toohey et al., 2016) the BRFSS health behaviour best represented the health behaviour of oesophageal cancer cases pre-diagnosis. A 5-year lag was added to minimise the risk of early symptoms influencing behaviour. The initial year was the earliest year in which BRFSS used a consistent definition for health behaviours selected for the present study. The end year was the most recently available SEER cancer registry data which allowed at least 12-months follow-up.

Outcomes, predictors and subgroups

The dichotomous outcome was all-cause mortality within one year of diagnosis.

Six self-reported measures of health behaviour were selected based on previous associations with oesophageal cancer (Clara Castro et al., 2018; Fahey et al., 2015) and availability in the BRFSS dataset:

- Current tobacco smoking (yes or no), defined as daily or less than daily smoking;
- Alcohol consumption – possible binge drinking (yes or no), defined as ≥5 standard drinks for males or ≥4 standard drinks for females on at least one occasion in the month prior to survey;

- Alcohol consumption – possible heavy drinking (yes or no), defined as >2 standard drinks per day for men and >1 standard drink per day for women in the month prior to survey;
- Physical activity (yes or no), defined as any physical activity or exercise in the past 30 days other than for regular job;
- Obese (yes/no), defined as body mass index ≥ 30 kg/m^2; and
- Current tobacco smoking with regular alcohol (yes or no), defined as current tobacco smoking with ≥1 standard drink of alcohol per day on average in the previous month.

Histological subgroups were defined using International Classification of Diseases for Oncology, third edition (ICD-O-3) with 805-808 indicating OSCC (n=10,454) and 814-838 indicating OAC (n=17,950).

Imputation method and covariates

The complete absence of data on health behaviour meant that regression-based imputation and multiple imputation could not be used (Sterne et al., 2009). Random cold deck imputation (De Waal et al., 2011) based on demographic strata was appropriate, as there were demographic variables in common between the two datasets and individuals from the same demographic group have a greater likelihood of engaging in similar health behaviours (Moore et al., 2016).

In random cold deck imputation individuals are allocated into strata according to auxiliary variables and then, within each stratum, one ‘donor’ record is randomly selected for each ‘recipient’ record. The BRFSS health behaviour data were the donor records and the SEER cancer registry data were the recipients. The recipient record is assigned the behaviour of the donor record. The more, and the more informative, the auxiliary variables the greater the chance the imputed behaviour will be correct.

Six auxiliary variables were used:

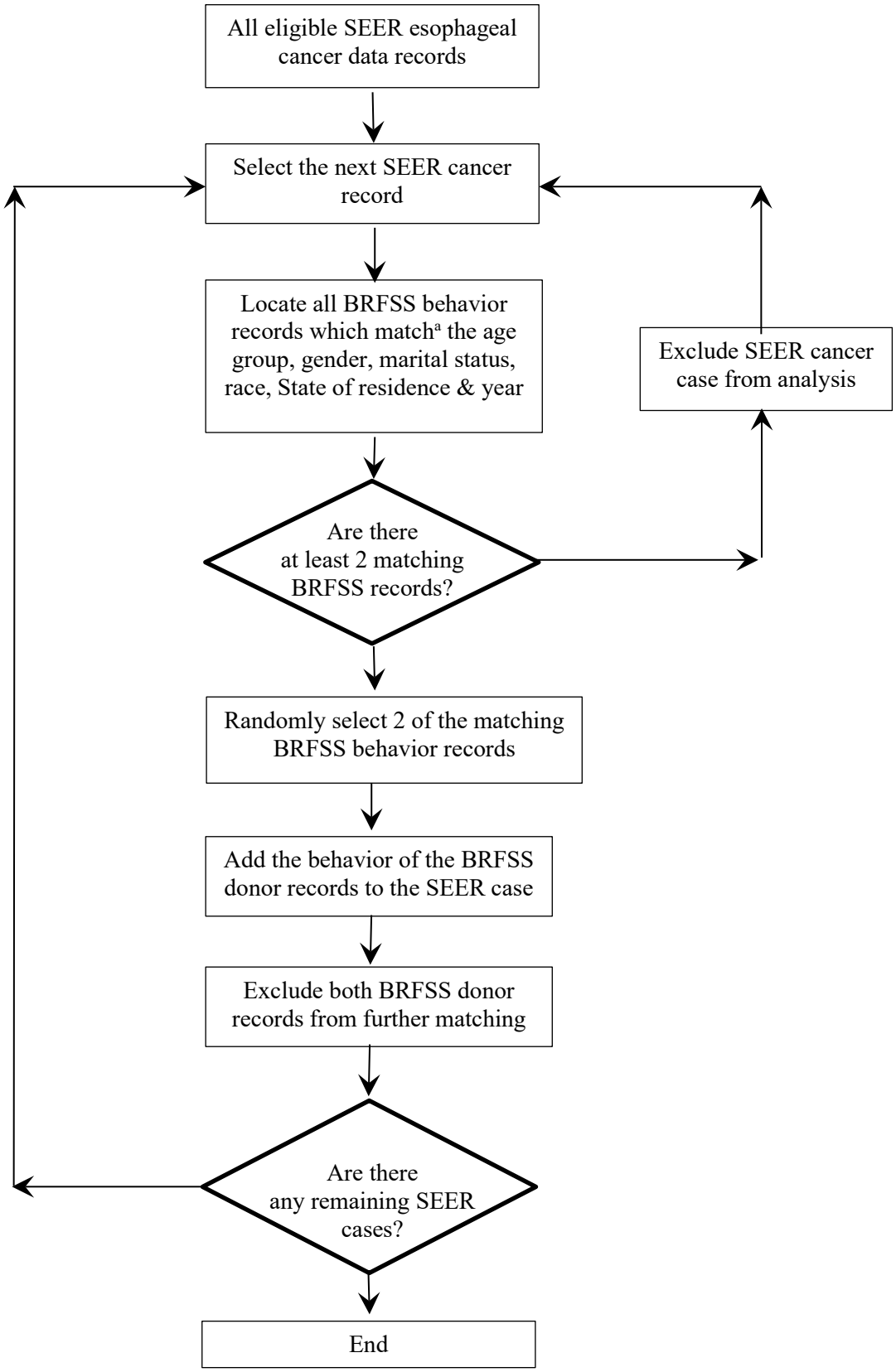
- Age category at diagnosis (5-year groups from 35-39y to 75-79y then ≥80y);
- Gender (male; female);
- Marital status (married, including common law; single or never married; widowed; divorced);
- Race (white; black; Asian or Pacific Islander; American Indian or Alaska Native);
- State of residence (Alaska; California; Connecticut; Georgia; Hawaii; Iowa; Kentucky; Louisiana; Michigan; New Jersey; New Mexico; Utah; Washington);
- Year of diagnosis (2006 to 2014).

To produce the 5-year lag, the donor records were defined to be BRFSS health behaviour records which were five years earlier and one age-group younger than the corresponding SEER cancer case.

There were 37,440 possible combinations of the auxiliary variable categories, 7,397 of which occurred within the SEER oesophageal cancer cases. Of these, 6,986 (94.4%) contained at least one eligible BRFSS donor record.

To allow calibration, two BRFSS donor records were randomly selected for each SEER case (without replacement), such that each cancer case had two imputed values for each lifestyle variable. Where donor records were exhausted before cancer cases, the cancer case was omitted from the analysis. The conceptual steps of the imputation algorithm are summarised in the flow chart in Figure 5.5. In practice, this algorithm was operationalized strata by strata rather than case by case. The imputation algorithm (and all analyses) was written in R software.

Figure 5.5 A conceptual map of the algorithm used to impute the pre-diagnosis behaviour of oesophageal cancer cases.



[a] Matches are BRFSS health behaviour records which are 5 years earlier in time and one age-group younger than the corresponding SEER cancer case, with all other variables equal.

Missing data, exclusions, and the final dataset

Approximately 80% of the 35,084 eligible oesophageal cancer cases were included in the analyses (Figure 5.6). SEER cases were excluded for missing survival time or auxiliary variables (n=2784, 8.0%) or failing to find two donor records (from 4353 to 4453 (12.4% to 12.7%) varying between health behaviours).

Figure 5.6 Flow chart of inclusions and exclusions of SEER oesophageal cancer cases

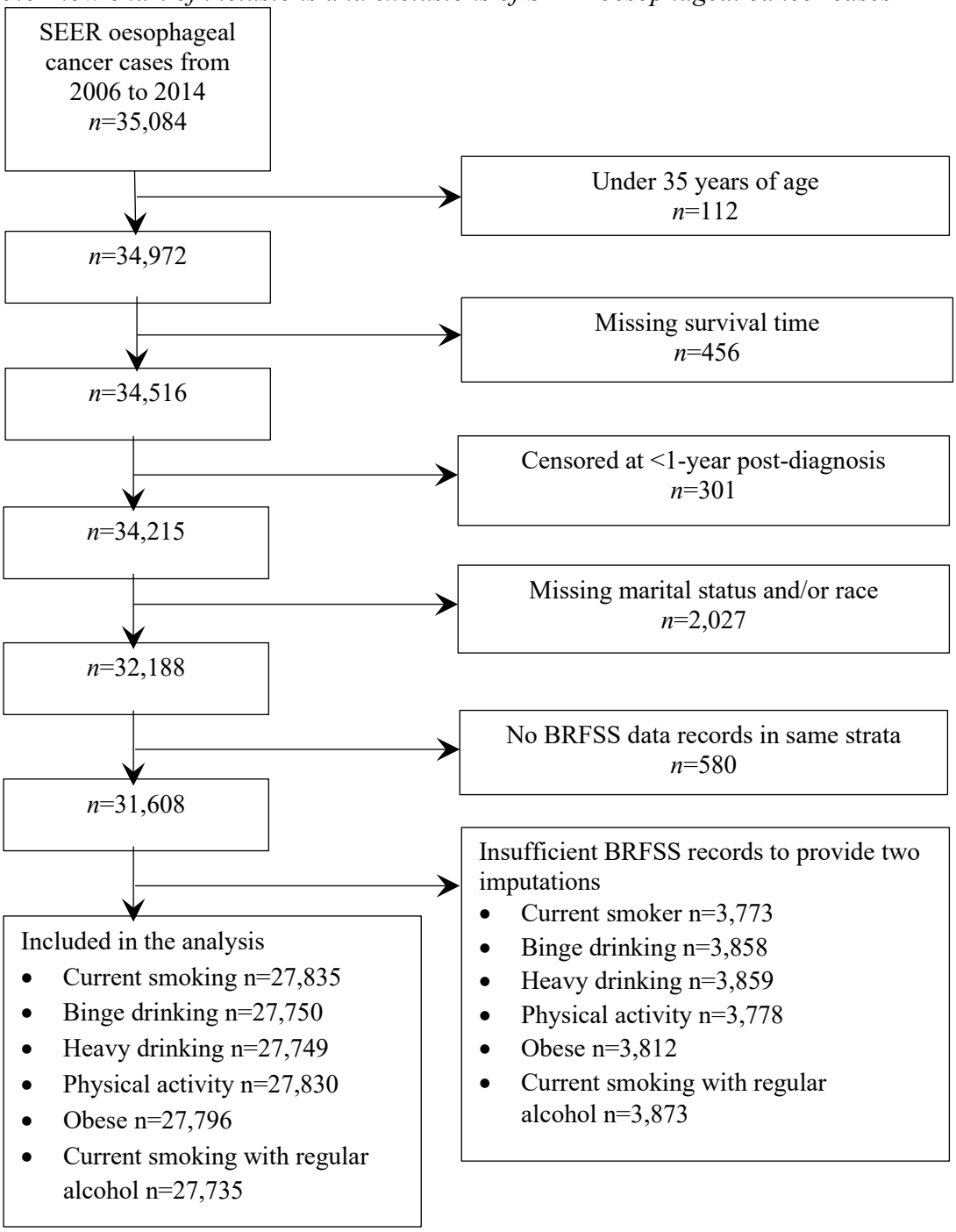


Table 5.4 shows the proportions of eligible SEER cancer cases that were unable to be matched with two donor records with non-missing smoking status. Proportions of unmatched cases are shown for each category of each of the auxiliary variables used in matching. Results for the other behaviours were similar – varying slightly due to the different distribution of missing data in the different variables.

Table 5.4 Number of SEER oesophageal cancer cases seeking donor records for current smoking behaviour and the proportion of these failing to obtain two donor records.

	Seeking Donor Records		Failed to Obtain 2 Donor Records	
	Frequency	% of total	Frequency	% of group
Total	32,188	100.0%	4,353	13.5%
Status at one year				
- Dead	17,490	54.3%	2,563	14.7%
- Alive	14,698	45.7%	1,790	12.2%
Cancer type				
- ESCC	10,454	32.5%	1,540	14.7%
- EAC	17,950	55.8%	2,224	12.4%
Year				
- 2006	3,364	10.7%	691	20.0%
- 2007	3,519	10.9%	659	18.7%
- 2008	3,539	11.0%	620	17.5%
- 2009	3,675	11.4%	727	19.8%
- 2010	3,564	11.1%	502	14.1%
- 2011	3,591	11.2%	423	11.8%
- 2012	3,636	11.3%	496	13.6%
- 2013	3,613	11.2%	148	4.1%
- 2014	3,597	11.2%	87	2.4%
Age group				
- 35-39	179	0.6%	2	1.1%
- 40-44	477	1.5%	5	1.0%
- 45-49	1,258	3.9%	25	2.0%
- 50-54	2,399	7.5%	138	5.8%
- 55-59	3,849	12.0%	368	9.6%
- 60-64	4,919	15.3%	692	14.1%
- 65-69	5,028	15.6%	810	16.1%
- 70-74	4,273	13.3%	725	17.0%
- 75-79	4,010	12.5%	736	18.4%
- 80+	5,796	18.0%	852	14.7%
Sex				
- Male	25,131	78.1%	4,059	16.2%
- Female	7,057	21.9%	294	4.2%
Marital status				
- Married (including common law)	18,363	57.0%	2238	12.2%
- Divorced				
- Widowed	4,046	12.6%	287	7.1%
- Single (Never married)	4,394	13.7%	518	11.8%
	5,384	16.7%	1,309	24.3%
Race				
- White	27,184	84.5%	3,115	11.5%
- Black	3,410	10.6%	681	20.0%
- Asian or Pacific Islander	1,406	4.4%	474	33.7%
- American Indian or Alaska Native	187	0.6%	82	43.9%

Table 5.4 (continued) Number of SEER oesophageal cancer cases seeking donor records for current smoking behaviour and the proportion of these failing to obtain two donor records.

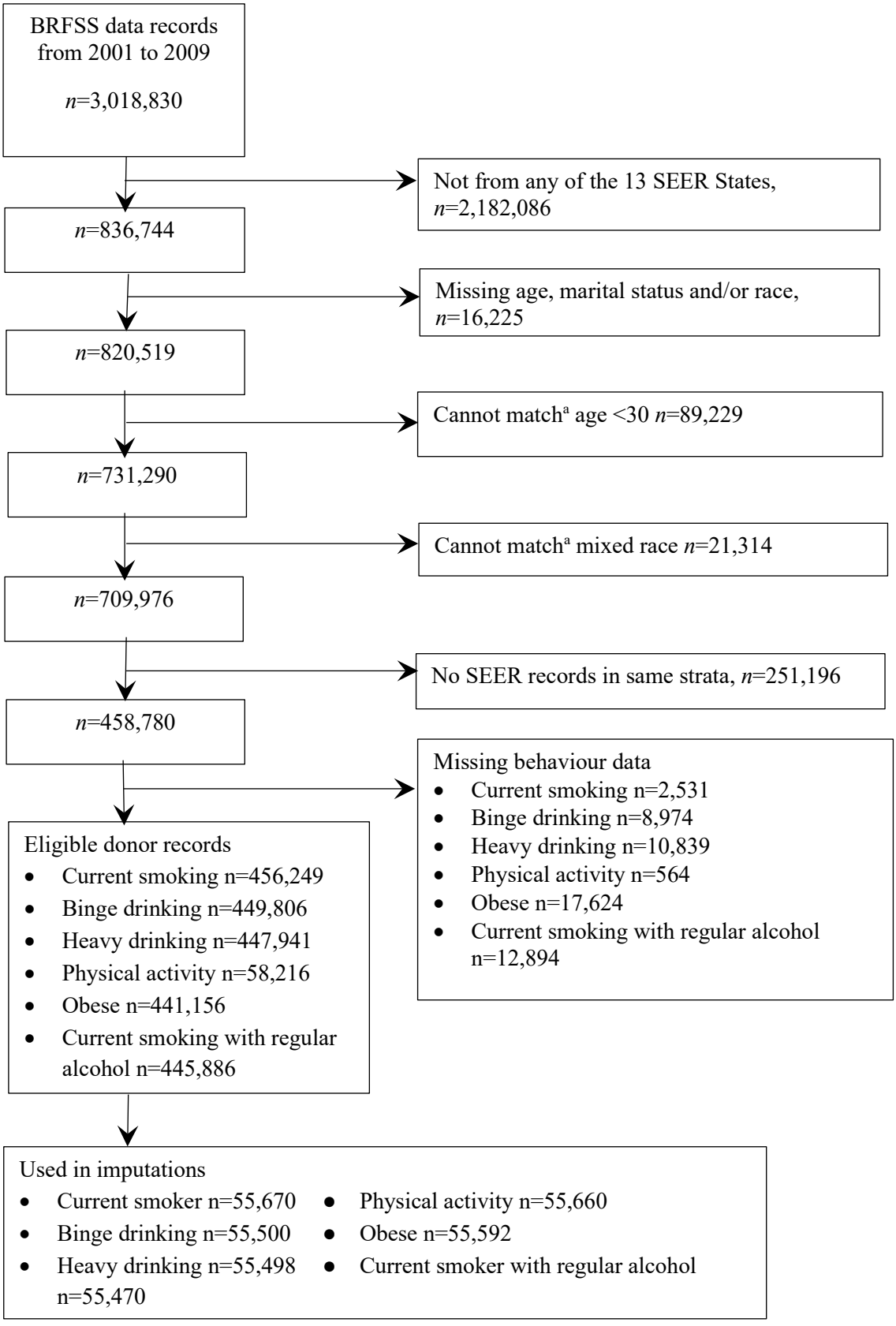
	Seeking Donor Records		Failed to Obtain 2 Donor Records	
	Frequency	% of total	Frequency	% of group
Total	32,188	100.0%	4,353	13.5%
State of residence				
- Alaska	39	0.1%	39	100.0%
- California	11,672	36.3%	3,551	30.4%
- Connecticut	1,799	5.6%	34	1.9%
- Georgia	3,599	11.2%	308	8.6%
- Hawaii	479	1.5%	55	11.5%
- Iowa	1,727	5.4%	35	2.0%
- Kentucky	3,028	9.4%	53	1.8%
- Louisiana	1,972	6.1%	81	4.1%
- Michigan	1,902	5.9%	70	3.7%
- New Jersey	3,629	11.3%	81	2.2%
- New Mexico	648	2.0%	9	1.4%
- Utah	613	1.9%	10	1.6%
- Washington	1,980	6.2%	26	1.3%

Table 5.4 shows that failure to find donor records were more common:

- in earlier years when the number of people interviewed by BRFSS was smaller.
- in the older age categories where there are more cancer cases and relatively fewer BRFSS respondents.
- in California, where the population (and hence number of cancer cases) is quite large. The BRFSS surveys a similar number per State regardless of population size.
- In marital and race groups where the BRFSS is known to be under-representative of the true population (Schneider et al., 2012).

Only 458,780 of the BRFSS health behaviour records matched the SEER cases on the auxiliary variables (Figure 5.7). The number with missing health behaviour ranged from 564 (0.1%) for physical activity to 17,624 (3.9%) for obesity. To avoid imputing a missing value into a missing value, these records were excluded. To avoid cumulative effects, six separate donor datasets were created (each containing complete cases for one of the six health behaviours) and each health behaviour was imputed independently.

Figure 5.7 Flow chart of inclusions and exclusions of BRFSS health behaviour data records



[a] All auxiliary variables could be coded identically in both SEER cancer registry data and BRFSS health behaviour data except the BRFSS data included an additional category for race. Unlike the SEER cancer registry data, the BRFSS data collection allowed respondents to describe their race as "mixed". About 3.5% of BRFSS respondents selected this option. As these records did not match any SEER cancer registry records they did not contribute to the analysis. Similarly, with a minimum age of 35 years for SEER cancer cases and a 5-year lag, BRFSS respondents under 30 years of age could not match any SEER cancer registry records.

Calibrating the effectiveness of imputation

For each health behaviour the agreement between the two donor records was used to estimate of the amount of information on health behaviour retained by the cold deck imputation.

Let p_i represent the proportion of imputed values where the behaviour is present. If the imputation process retained no information on behaviour, the expected proportion of behaviour present to behaviour present matches is p_i^2 - the agreement arising through chance alone. If the imputation process is informative, the proportion of behaviour present to behaviour present matches is greater than chance. Following Lunn & Davies (Lunn & Davies, 1998) this improvement in matching was modelled as $p_i(1-p_i)\,\rho$ where ρ is a measure of correlation.

Table 5.5 summarizes how the agreement between the two imputed values was modelled. The observed number of behaviour present to behaviour present matches is designated *E* and *E* is modelled as the number of behaviour present to behaviour present matches expected by chance alone, np_i^2, plus the excess matches arising from the information retained by the imputation algorithm, $np_i(1-p_i)\rho$.

Table 5.5 Observed and expected agreement between the two sets of imputations.

Imputed behaviour #1	Imputed behaviour #2		
	Behaviour present	Behaviour absent	total
Behaviour present			
-observed	E	F	$E+F$
-expected	$np_i^2 + np_i(1-p_i)\rho$	$np_i(1-p_i) - np_i(1-p_i)\rho$	np_i
Behaviour absent			
-observed	G	H	$G+H$
-expected	$np_i(1-p_i) - np_i(1-p_i)\rho$	$n(1-p_i)^2 + np_i(1-p_i)\rho$	$n(1-p_i)$
Total			
-observed	$E+G$	$F+H$	n
-expected	np_i	$n(1-p_i)$	

The values n, E, F, G and H are obtained by cross tabulation of the two imputed values. Two potential estimates of p_i are the observed proportion in the first set of imputed values and the observed proportion in the second set of imputed values. In the current study p_i is estimated using the average of these two observed proportions:

$$\widehat{p_i} = \frac{1}{2}\left(\frac{E+F}{n} + \frac{E+G}{n}\right)$$

The value of ρ is estimated by the phi coefficient, φ, (the correlation coefficient for dichotomous variables) between the pairs of imputed values.

Statistical analysis

For each behaviour, the first set of imputed values were cross-tabulated against one year survival status and calculated the relative risk of death within one year, RR_i. The subscript i signifies that the imputed data were used in the calculations.

Other potential predictors of survival times were investigated using log-binary regression with associated log likelihood ratio statistics and area under the receiver operator curves (Appendix 1). Age was identified as a confounder as both post-diagnosis survival and proportion recording each health behaviours were lower among older age groups (Appendix 1). To adjust for this, age-adjusted relative risks, $adjRR_i$, were estimated using the Cochran-

Mantel-Haenszel method (Agresti, 2013). Other potential demographic predictors of survival were found to be of lesser impact or confounded with age (Appendix 1).

Beyond the demographic variables, cancer stage at diagnosis (coded by SEER according to the AJCC Cancer Staging Manual 6th Edition (Greene et al., 2003)) was confirmed as a stronger predictor of survival (Appendix 1) but, occurring after health behaviour exposure, may partially lie on the disease pathway. That is, smokers may have more advanced disease at diagnosis due to their smoking and so correcting for cancer stage at diagnosis may falsely attenuate the association between pre-diagnosis smoking and survival post diagnosis (Cole & Hernán, 2002). However, for completeness, subgroup analyses were conducted for cancer stage at diagnosis.

Non-differential misclassification errors will, barring random error and confounding, attenuate the estimated relative risk toward the null (Jurek, Greenland, Maldonado, & Church, 2005). The mathematical relationship between the relative risk using the imputed data, RR_i, and the true relative risk for the cancer cases, RR_T, is derived in Appendix 2. In brief, if the prevalence of behaviour is the same between the donor records and cancer cases in each stratum, the true relative risk can be estimated using

$$RR_T = 1 - \frac{(RR_i - 1)}{(RR_i - 1)p_i(1 - \rho) - \rho}$$

Extreme values of p_i and/or ρ can be problematic. For example, when $\rho = 0$, RR_T is negative: an impossible value for a relative risk.

Random cold deck imputation was repeated 100 times, separately for each of the six health behaviours. As donor records were selected at random within strata, each statistic varied between repetitions. Results were reported as the median value from the 100 repetitions with the associated 2.5 and 97.5 percentiles as empirical 95% confidence

intervals. Subgroup analyses are reported for ESCC and EAC. Where more than 5% of the estimates of the true relative risk RR_T were impossible, the imputation process was labelled as 'failed'.

Checking the algorithm with simulated data

In the absence of a cohort showing the true relationship between pre-diagnosis health behaviour and post-diagnosis survival time, simulated data were used to test the algorithm.

The first set of imputed behaviour was designated to be the 'true' health behaviour of each cancer case. For each health behaviour simulated seven survival status variables were separately simulated (repeated 100 times): to produce relative risks of 0.50, 0.66, 0.80, 1.00, 1.25, 1.50 and 2.00 while maintaining the overall rate of the health behaviour p_i and one year death rate (Appendix 3).

The imputed relative risks were obtained using the second set of imputed health behaviours. As the second set of imputed values were selected independently and without replacement, they had a similar relationship with the first set of simulated data as with the actual cancer cases. The main difference is that the simulated survival data, being based only on the behaviour of interest, have no relationship with (confounding from) any other variables. The true data were likely to display more complex relationships.

5.3.4 **Results**

Calibrating the imputation

The estimated proportion of cancer cases with a given health behaviour, p_i, ranged from a median of 0.737 for physical activity to 0.034 for current smoking with regular drinking (Table 5.6). The phi coefficients, φ, show that there is usually a positive correlation between the two imputed values, albeit weak (medians between 0.008 and 0.077). This confirms that some information about health behaviour is being conveyed through the random cold deck imputation. The value $np_i(1-p_i)\rho$, the number of correct matches greater than would be expected through chance, quantifies the information conveyed through the imputation. 'Heavy drinking', and 'current smoking with regular drinking', had the lowest prevalence (median of 0.05 or less), the lowest correlations between imputed observations (median less than 0.025) and hence lowest information (medians below 20 matches beyond chance).

Table 5.6 The estimated proportions with each health behaviour, the phi coefficient between imputed values and the estimated excess matches for each analysis.

Behaviour 5 years before diagnosis	N	Estimated proportion with behaviour, $\hat{p}_i$		Estimated phi coefficient, $\hat{\rho} = \varphi$		Estimated excess matches, $n\hat{p}_i(1-\hat{p}_i)\hat{\rho}$	
		Median	95% CI	Median	95% CI	Median	95% CI
Current smoking							
- overall	27,835	0.159	0.157,0.162	0.071	0.059,0.084	262.2	220.1,312.2
- ESCC	8,914	0.166	0.162,0.170	0.077	0.061,0.097	94.8	74.5,120.7
- EAC	15,726	0.157	0.153,0.159	0.066	0.052,0.081	137.0	107.4,169.5
Binge drinking							
- Overall	27,750	0.100	0.098, 0.102	0.060	0.049,0.077	150.5	121.5,192.1
- ESCC	8,891	0.086	0.082,0.089	0.060	0.042,0.086	42.2	29.8,61.1
- EAC	15,673	0.109	0.106,0.111	0.058	0.042,0.079	88.6	63.6,120.3
Heavy drinking							
- Overall	27,749	0.048	0.047,0.050	0.011	0.002,0.025	14.3	2.7,32.0
- ESCC	8,888	0.046	0.043,0.049	0.015	-0.002,0.036	5.7	-0.7,14.2
- EAC	15,676	0.050	0.048,0.052	0.008	-0.004,0.028	6.0	-3.0,20.8
Physical activity							
- Overall	27,830	0.737	0.734,0.740	0.034	0.026,0.046	185.1	139.4,247.4
- ESCC	8,912	0.716	0.709,0.721	0.036	0.016,0.056	64.7	29.6,100.2
- EAC	15,724	0.750	0.746,0.754	0.031	0.013,0.047	91.4	40.0,138.4
Obese							
- Overall	27,796	0.257	0.254,0.261	0.030	0.020,0.042	160.2	108.4,226.8
- ESCC	8,898	0.262	0.255,0.268	0.045	0.024,0.061	77.0	41.4,104.6
- EAC	15,709	0.256	0.251,0.261	0.023	0.012,0.041	67.8	35.0,122.4
Current smoker with regular drink							
- Overall	27,735	0.034	0.033,0.035	0.022	0.009,0.038	19.8	8.0,34.2
- ESCC	8,883	0.031	0.029,0.033	0.024	-0.000,0.049	6.2	-0.0,13.5
- EAC	15,670	0.035	0.034,0.037	0.021	0.004,0.042	11.5	2.1,22.4

$\widehat{p_\iota}$ proportion of imputed values where the health behaviour is present

$\hat{\rho} = \varphi$ the correlation between the pairs of imputed values (calculated as the phi coefficient)

$n\widehat{p_\iota}(1 - \widehat{p_\iota})\hat{\rho}$= the excess number of correct matches greater than would be expected through chance alone

Median= median of 100 repetitions of the imputation algorithm,

95% CI= empirical 95% confidence interval created from the 2.5 and 97.5 percentiles obtained from 100 repetitions of the imputation algorithm,

N=number of SEER oesophageal cancer cases receiving data from two donor records from the BRFSS health behaviour datasets.

ESCC=oesophageal squamous cell carcinoma,

EAC=oesophageal adenocarcinoma

Analyses using simulated survival status

The simulated relative risks of survival were accurate to two-decimal places and precise (with a maximum margin of error of 0.07) (Table 5.7). The relative risks obtained by using the (second) imputed behaviour (RR_i) were substantially attenuated toward the null differing from 1.0 only in the second decimal place.

Table 5.7 Result of simulation-based testing of whether or not the imputation can be used to predict relative risk

Target RR	Simulated data RR		Imputed RR (RR_i)		Impossible Result ($RR_T < 0$)	Estimated true RR (RR_T)	
	Median	95% CI	Median	95% CI	Frequency	Median	95% CI[b]
Current smoking							
RR=0.5	0.501	0.475,0.521	0.964	0.934,0.993[a]	0	0.519	0.163,0.904[a]
RR=0.66	0.660	0.635,0.683	0.973	0.944,0.999[a]	0	0.638	0.300,0.985[a]
RR=0.80	0.799	0.771,0.823	0.983	0.952,1.017	0	0.753	0.375,1.226
RR=1.00	1.001	0.976,1.026	0.997	0.967,1.027	0	0.957	0.577,1.444
RR=1.25	1.249	1.220,1.287	1.017	0.989,1.048	0	1.254	0.856,1.793
RR=1.50	1.499	1.465,1.528	1.032	1.005,1.059[a]	0	1.486	1.069,1.947[a]
RR=2.00	2.000	1.974,2.034	1.064	1.034,1.092[a]	0	2.047	1.542,2.532[a]
Binge drinking							
RR=0.5	0.501	0.474,0.526	0.967	0.940,0.996[a]	0	0.478	0.087,0.927[a]
RR=0.66	0.659	0.624,0.692	0.976	0.945,1.015	1	0.629	0.173,1.316
RR=0.80	0.798	0.758,0.830	0.988	0.959,1.025	0	0.805	0.341,1.448
RR=1.00	0.997	0.963,1.033	0.999	0.971,1.032	0	0.981	0.518,1.492
RR=1.25	1.245	1.213,1.278	1.016	0.984,1.054	0	1.271	0.739,2.029
RR=1.50	1.499	1.463,1.534	1.030	0.990,1.068	0	1.517	0.831,2.246
RR=2.00	1.999	1.978,2.028	1.058	1.021,1.093[a]	0	2.014	1.352,2.717[a]
Heavy Drinking							
RR=0.5	0.500	0.450,0.548	0.995	0.945,1.046	40	failed	failed
RR=0.66	0.661	0.606,0.697	0.995	0.946,1.046	34	failed	failed
RR=0.80	0.799	0.746,0.847	0.997	0.944,1.053	43	failed	failed
RR=1.00	0.997	0.949,1.045	0.998	0.940,1.041	32	failed	failed
RR=1.25	1.251	1.210,1.300	1.003	0.959,1.053	22	failed	failed
RR=1.50	1.497	1.459,1.535	1.012	0.956,1.059	24	failed	failed
RR=2.00	Not possible	Not possible					

Table 5.6 (continued) Result of simulation-based testing of whether or not the imputation can be used to predict relative risk

Target RR	Simulated data RR		Imputed RR (RR_i)		Impossible Result ($RR_T < 0$)	Estimated true RR (RR_T)	
	Median	95% CI	Median	95% CI	Frequency	Median	95% CI[b]
Physical activity							
RR=0.5	0.500	0.491,0.509	0.974	0.951,0.997[a]	0	0.504	0.319,0.901[a]
RR=0.66	0.659	0.645,0.671	0.983	0.959,1.006	0	0.632	0.367,1.231
RR=0.80	0.800	0.782,0.818	0.993	0.971,1.017	0	0.833	0.449,1.907
RR=1.00	1.002	0.976,1.022	1.001	0.978,1.021	0	1.025	0.488,2.092
RR=1.25	1.250	1.219,1.276	1.006	0.977,1.030	0	1.206	0.541,2.961
RR=1.50	1.499	1.455,1.549	1.013	0.987,1.037	2	1.514	0.722,4.078
RR=2.00	2.003	1.939,2.083	1.021	1.002,1.047*	3	2.127	1.055,10.987[a]
Obese							
RR=0.5	0.499	0.485,0.517	0.983	0.960,1.008	1	0.550	0.028,1.322
RR=0.66	0.660	0.634,0.680	0.989	0.962,1.016	2	0.665	0.114,1.772
RR=0.80	0.802	0.777,0.823	0.995	0.967,1.015	1	0.846	0.316,1.676
RR=1.00	1.002	0.981,1.024	0.999	0.980,1.024	0	0.962	0.461,2.067
RR=1.25	1.250	1.222,1.274	1.009	0.989,1.030	0	1.335	0.601,2.300
RR=1.50	1.500	1.468,1.534	1.014	0.987,1.039	0	1.440	0.606,2.796
RR=2.00	2.002	1.961,2.041	1.025	0.997,1.044	0	1.995	0.886,3.234
Current smoker with regular drink							
RR=0.5	0.504	0.441,0.550	0.988	0.931,1.034	37	failed	failed
RR=0.66	0.660	0.600,0.713	0.997	0.932,1.066	31	failed	failed
RR=0.80	0.797	0.744,0.863	0.991	0.928,1.052	34	failed	failed
RR=1.00	0.996	0.943,1.049	1.001	0.940,1.059	25	failed	failed
RR=1.25	1.250	1.183,1.298	1.009	0.954,1.059	16	failed	failed
RR=1.50	1.497	1.454,1.545	1.000	0.958,1.065	19	failed	failed
RR=2.00	Not possible	Not possible					

Target RR – the relative risk the simulated data attempted to achieve

Simulated data RR – the relative risk which was actually achieved between the first imputed value and the simulated one-year survival status

Imputed RR (RR_i) – the relative risk calculated using the second imputed data point as the imputed behaviour

Impossible result – instances where the estimated true relative risk was impossible (a negative value)

Estimated True RR (RR_T) – the estimated true relative risk derived from the imputed relative risk and calibration parameters $\widehat{p_i}$ and $\hat{\rho}$

Median= median of 100 repetitions of the imputation algorithm,

95% CI= empirical 95% confidence interval created from the 2.5 and 97.5 percentiles obtained from 100 repetitions of the imputation algorithm,

[a] 95% confidence intervals exclude no association (i.e. exclude relative risk equals 1)

[b] excludes impossible result

Estimation of the true relative risk from the imputed relative risk failed for the two least common health behaviours: 'heavy drinking' and 'current smoking with regular drinking'. For the other four behaviours, the median of the estimated true relative risk was accurate to one, and often two, decimal places. However, the confidence intervals were wide and few excluded no association.

Analyses using true survival status

When imputing the health behaviours onto SEER cancer cases, the median imputed relative risks (RR_i) are attenuated to close to 1.0 (Table 5.8). Less expectedly, most of the median risks are less than 1.0; suggesting that most behaviours were associated with a lower rate of death within one year of diagnosis. Many of the age-adjusted imputed relative risks had the opposite direction of association confirming the potential for confounding by age. Current tobacco smoking 5 years prior to diagnosis was detrimental to one-year survival after diagnosis following adjustment for age, particularly in ESCC where the estimated relative risk was 1.99 (95%CI 1.24, 3.12). For ESCC, the median relative risk for binge drinking 5 years prior to diagnosis was 1.52 although the range of possible relative risks was wide (95% CI 0.44,2.75). Similar results were seen for obesity (ESCC estimated RR 1.73, 95%CI 0.83,4.17). Physical activity 5-years prior to diagnosis was protective for survival with median estimated relative risks of approximately 0.50 (95%CI 0.31, 1.03) for oesophageal cancer overall.

Table 5.8 Estimated relative risks of 1-year survival derived from imputed pre-diagnosis behaviours for SEER oesophageal cancer cases, 2006-2014; unadjusted and age adjusted.

	Imputed RR (RR_i)		Impossible Result ($RR_i < 0$)	Estimated True RR (RR_T)		Age-adjusted Imputed RR ($adjRR_i$)		Impossible Result ($adjRR_i < 0$)	Age-adjusted Estimated True RR ($adjRR_T$)	
	Median	95% CI	Frequency	Median	95% CI	Median	95% CI	Frequency	Median	95% CI
Current smoking										
All	0.986	0.954,1.009	0	0.806	0.380,1.130	1.051	1.014,1.078	0	1.794	1.215,2.357[a]
ESCC	1.025	0.981,1.067	0	1.349	0.733,2.142	1.064	1.016,1.111	0	1.990	1.240,3.117[a]
EAC	0.959	0.914,1.000[a]	5	0.478	0.039,1.003[b]	1.038	0.985,1.085	0	1.613	0.785,2.571
Binge drinking										
All	0.933	0.900,0.964	49	failed	failed	0.997	0.961,1.032	1	0.951	0.445,1.539[b]
ESCC	0.998	0.936,1.059	4	0.991	0.167,1.995[b]	1.033	0.968,1.101	0	1.515	0.440,2.754
EAC	0.914	0.863,0.961[a]	72	failed	failed	0.989	0.935,1.046	3	0.818	0.181,1.890[b]
Heavy drinking										
All	0.981	0.932,1.028	61	failed	failed	1.010	0.963,1.060	23	failed	failed
ESCC	0.995	0.912,1.066	48	failed	failed	1.012	0.929,1.088	36	failed	failed
EAC	0.974	0.907,1.039	66	failed	failed	1.011	0.938,1.077	35	failed	failed
Physical activity										
All	0.954	0.934,0.978[a]	0	0.319	0.165,0.564[a]	0.974	0.956,1.001	0	0.507	0.307,1.030
ESCC	0.959	0.925,0.991	2	0.345	0.073,0.811[a,b]	0.971	0.933,1.003	1	0.452	0.102,1.071[b]
EAC	0.957	0.929,0.986[a]	1	0.311	0.109,0.675[a,b]	0.984	0.954,1.013	0	0.627	0.285,2.180
Obese										
All	0.969	0.946,0.993[a]	24	failed	failed	1.008	0.983,1.036	0	1.262	0.559,2.931
ESCC	1.000	0.968,1.039	0	1.004	0.134,2.378	1.027	0.992,1.068	0	1.733	0.834,4.167
EAC	0.949	0.917,0.987[a]	76	failed	failed	0.996	0.960,1.035	8	failed	failed
Current smoking with regular drinking										
All	0.987	0.930,1.058	40	failed	failed	1.044	0.986,1.120	2	3.254	0.771,11.843[b]
ESCC	1.044	0.946,1.146	12	failed	failed	1.076	0.973,1.180	11	failed	failed
EAC	0.963	0.861,1.052	60	failed	failed	1.032	0.919,1.123	13	failed	failed

Imputed RR (RR_i) – the relative risk calculated using the imputed behaviour

Impossible result – instances where the estimated true relative risk was impossible (a negative value)

Estimated True RR (RR_T) – the estimated true relative risk derived from the imputed relative risk and calibration parameters $\hat{p}_i$ and $\hat{\rho}$

Median= median of 100 repetitions of the imputation algorithm,

95% CI= empirical 95% confidence interval created from the 2.5 and 97.5 percentiles obtained from 100 repetitions of the imputation algorithm,

[a] 95% confidence intervals exclude no association (i.e. exclude relative risk equals 1)

[b] excludes impossible result

Estimates of the relative risks could not be retrieved for the less common behaviours ‘heavy drinking’ and ‘current smoking with regular drinking’. The one relative risk which was retrieved - a median RR of 3.35 for current smoking with regular drinking in all oesophageal cancer - was accompanied by wide uncertainty (95% CI 0.77,11.84).

Subgroup analyses on cancer stage at diagnosis (Table 5.9), suggests that pre-diagnosis health behaviours have stronger relationships with one-year survival in those who are not metastatic at diagnosis. In Table 5.9 the final column shows age-adjusted estimates of the relative risk for each health behaviour for all cases, individuals with stage IV (metastatic) cancer at diagnosis and for stages I-III cancers at diagnosis combined. Results are available for current smoking, non-work related physical activity and obesity each 5 years prior to diagnosis. Review of the relative risks suggests that these pre-diagnosis health behaviours appear to have larger effects on survival for those with stage I-III cancer at diagnosis than for those with metastatic cancers.

Table 5.9 Estimated relative risks of 1-year survival derived from imputed pre-diagnosis behaviours for SEER oesophageal cancer cases, 2006-2014; unadjusted and age adjusted. Showing subgroup analysis for cancer stage at diagnosis.

	Imputed RR (RR_i)		Impossible Result ($RR_i < 0$)	Estimated True RR (RR_T)		Age-adjusted Imputed RR ($adjRR_i$)		Impossible Result ($adjRR_i < 0$)	Age-adjusted Estimated True RR ($adjRR_T$)	
	Median	95% CI	Frequency	Median	95% CI	Median	95% CI	Frequency	Median	95% CI
Current smoking										
All	0.986	0.954,1.009	0	0.806	0.380,1.130	1.051	1.014,1.078 [a]	0	1.794	1.215,2.357[a]
Stage IV	0.996	0.968,1.024	0	0.949	0.531,1.455	1.033	1.000,1.063 [a]	0	1.486	1.006,2.360 [a]
Stage I, II or III	0.969	0.917,1.120	3	0.570	0.125,1.186 [b]	1.085	1.025,1.138 [a]	0	2.499	1.428,3.900 [a]
Binge drinking										
All	0.933	0.900,0.964 [a]	49	failed	failed	0.997	0.961,1.032	1	0.951	0.445,1.539 [b]
Stage IV	0.965	0.933,0.997 [a]	12	failed	failed	1.003	0.967,1.137	0	1.050	0.409,1.833
Stage I, II or III	0.866	0.803,0.928 [a]	97	failed	failed	0.977	0.906,1.052	18	failed	failed
Heavy drinking										
All	0.981	0.932,1.028	61	failed	failed	1.010	0.963,1.060	23	failed	failed
Stage IV	0.993	0.936,1.041	49	failed	failed	1.006	0.948,1.058	31	failed	failed
Stage I, II or III	0.955	0.871,1.052	72	failed	failed	1.006	0.912,1.101	39	failed	failed
Physical activity										
All	0.954	0.934,0.978 [a]	0	0.319	0.165,0.564 [a]	0.974	0.956,1.001	0	0.507	0.307,1.030
Stage IV	0.976	0.946,0.999 [a]	1	0.541	0.207,0.994 [a,b]	0.987	0.958,1.008	0	0.711	0.306,1.474[b]
Stage I, II or III	0.915	0.868,0.947 [a]	17	failed	failed	0.947	0.897,0.978 [a]	3	0.258	0.039,0.588 [a]
Obese										
All	0.969	0.946,0.993 [a]	24	failed	failed	1.008	0.983,1.036	0	1.262	0.559,2.931
Stage IV	0.988	0.963,1.015	8	failed	failed	1.008	0.981,1.034	3	1.267	0.506,4.319 [b]
Stage I, II or III	0.954	0.912,1.002	54	failed	failed	1.018	0.974,1.067	2	1.794	0.376,8.612 [b]
Current smoker with regular drink										
All	0.987	0.930,1.058	40	failed	failed	1.044	0.986,1.120	2	3.254	0.771,11.843
Stage IV	0.983	0.936,1.040	43	failed	failed	1.016	0.966,1.071	22	failed	failed
Stage I, II or III	0.967	0.860,1.097	59	failed	failed	1.073	0.956,1.228	9	failed	failed

Imputed RR (RR_i) – the relative risk calculated using the imputed behaviour

Impossible result – instances where the estimated true relative risk was impossible (a negative value)

Estimated True RR (RR_T) – the estimated true relative risk derived from the imputed relative risk and calibration parameters $\hat{p}_i$ and $\hat{\rho}$

Median= median of 100 repetitions of the imputation algorithm,

95% CI= empirical 95% confidence interval created from the 2.5 and 97.5 percentiles obtained from 100 repetitions of the imputation algorithm,

[a] 95% confidence intervals exclude no association (i.e. exclude relative risk equals 1)

[b] excludes impossible result

5.3.5 **Discussion**

This study shows that an entirely missing variable can be imputed and return accurate estimates of relative risks. Nearly all correlation coefficients were positive, indicating that the imputation conveyed some information about health behaviour, although confidence intervals were wide. Age confounding was evident, and so this discussion focuses on the age-adjusted results. However, for the less common behaviours (heavy drinking and current smoking with regular drinking), no interpretable information could be retrieved.

The choice of health behaviour variables was restricted to measures available through the BRFSS health survey. However, the results are consistent with the literature. Tobacco smoking five years prior to diagnosis was found to be associated with increased risk of death one year after diagnosis in ESCC (RR=1.99, 95% CI 1.24,3.12) and, with less certainty, EAC (RR=1.61, 95% CI 0.79,2.57). Recent meta analyses estimated hazard ratios (HRs) of 1.41 (95% CI 1.22,1.64) and 1.41 (95% CI 0.96,2.09) for current smoking relative to never smoked in mainly ESCC populations (Kuang et al., 2016; McMenamin et al., 2017) and 1.19 (95% CI 1.04,1.36) for ever smoking compared to never smoked in ESCC (Fahey et al., 2015) with no evidence of association between smoking and survival in EAC (Fahey et al., 2015; McMenamin et al., 2017). The unadjusted protective effects of smoking has also been reported (Dandara et al., 2015; Mirinezhad et al., 2012) as has the change in the direction of the association following age adjustment (Mirinezhad et al., 2012).

A previous meta-analysis found that ever drinking alcohol had a detrimental association with survival in ESCC (HR 1.36, 95% CI 1.15, 1.61) but not in EAC (HR=1.08 95% CI 0.85, 1.37) (Fahey et al., 2015). More recent results from China (HR=1.58, 95% CI 1.21,2.07 (Ma et al., 2016; P. Sun et al., 2016), HR=1.45 95% CI 1.13,1.87 (P. Sun et al., 2016)) and Japan (HR=2.37 95% CI 1.24,4.53 (Okada et al., 2017)) also support the

detrimental impact of pre-diagnosis alcohol consumption on survival in ESCC. The association between heavy drinking and survival could not be estimated. However, for binge drinking five years prior to diagnosis, the median relative risk was 1.52 in ESCC, although the confidence interval (95% CI 0.44,2.75) allows no association.

Previous studies have reported that pre-diagnosis smoking with regular alcohol consumption produced a disproportionately high risk to post-diagnosis survival in ESCC (HR 3.84, 95% CI 2.02,7.32 (Thrift, Nagle, Fahey, Russell, et al., 2012)). In the current analyses, a similar association was observed with wider confidence intervals (RR = 3.25, 95% CI 0.77,11.84).

In relation to obesity, a recent North American study (A. Spreafico et al., 2017) found self-reported obesity was associated with lower survival times in EAC compared to normal weight (HR 1.77, 95% CI 1.25, 2.51) and a 27 year follow-up of 29,446 participants in China (S. M. Wang et al., 2016) found higher body mass index protective of death from ESCC (HR=0.97 per unit increase, 95% CI 0.95,0.99) . The results of the current analyses, in contrast, suggested that obesity five years pre-diagnosis may be detrimental to one-year post diagnosis survival for ESCC (median RR=1.73) although confidence intervals were wide (95% CI 0.83,4.17).

One benefit of the algorithm is that it does not add any additional information about individuals to the cancer registry data and so, unlike direct data linkage, does not exacerbate the issues of confidentiality and data security. (The imputed behaviours are only slightly more likely to be correct than an uninformed guess.) The algorithm also provides protection against biases. Data were obtained from the SEER cancer registries which are censuses with good population coverage. Many sampling and non-response biases in the BRFSS health behaviour data (Iachan et al., 2016) are eliminated when using a census as the reference.

However, rigid matching criteria was used and failed to match 20% of cases. Further investigation of the trade-off between exact matching and biases arising from failure to match is required.

As with direct data linkage, the investigations were limited to available health behaviour measures, rather than all clinically important risk factors. Potentially important health behaviours such as diet (Abnet, Arnold, & Wei, 2018; Clara Castro et al., 2018) and hot beverages (Abnet et al., 2018) were unavailable. The number and variety of auxiliary variables available for matching donor to recipient records was also limited. the only investigation of clustering in health behaviours (Meader et al., 2016) in this paper was for the combination of current smoking and regular alcohol consumption.

The results display considerable uncertainty with few instances where the empirical confidence intervals excluded the null. The width of the confidence intervals is sensitive to n, p_i and ρ. Larger n can be achieved by looking at more common cancers, and/or combining data from more cancer registries and/or more years. The proportion with the health behaviour, p_i, can be adjusted through inclusion and exclusion criteria (but will impact on n). Larger ρ requires more informative auxiliary variables for the imputation.

A ‘gold standard’ could not be accessed for validity testing. A gold standard would be an oesophageal cancer dataset where behaviour was measured five years prior to diagnosis.

5.3.6 **Conclusion**

This paper has demonstrated a novel imputation-based algorithm for augmenting cancer registry data for epidemiological research and established its face-validity. The algorithm adds information obtained from an external data set with (presumed) no cases in

common, to the cancer registry data via demographic variables in common. The algorithm is subject to much higher random error than direct data linkage (depending on how informative the demographic variables are), and requires larger sample sizes to compensate. However, it does avoid confidentiality issues (and associated data security costs) arising from direct data linkage.

This algorithm is likely to allow, at least preliminary, investigations of a range of research questions which cannot be addressed through direct data linkage; due to insufficient individuals in common, insufficient matching variables and/or costs associated with data confidentiality and security. By increasing the range of research question which can be addressed with cancer registry data, the algorithm further augments the benefits of cancer registries.

5.3.7 **Appendix 1 The association between health behaviours and cancer stage at diagnosis and survival time**

The relationship between death within 1 year of diagnosis (yes/no) and each predictor was modelled separately using log-binary regression (sometimes called relative risk regression). The explanatory power of each predictor was summarised using the likelihood ratio test statistic and the area under the receiver operator curve. The likelihood ratio test statistics are only comparable if the degrees of freedom is equal. The area under the receiver operator curve is 0.5 for an uninformative predictor variable to 1.0 for a perfectly informative predictor. Results are presented in Table 5.10.

Table 5.10 Relationships between demographic variables and cancer stage at diagnosis and death within the first year of diagnosis.

	Estimated relative risk (RR)	Improvement over the null model (likelihood ratio test statistic)	Area under the receiver operator curve
Year of diagnosis (n=34215)			
- per year	0.992 (0.988,0.995)	18.8, df=1	0.514 (0.507,0.520)
Sex (n=34215)			
- male	reference		
- female	1.031 (1.007,1.054)	6.8, df=1	0.506 (0.501,0.510)
Age group (n=34215)			
- per 5 year age group	1.060 (1.055,1.065)	653.2, df=1	0.575 (0.569,0.581)
Marital status (n=32236)			
- Married	reference		
- Divorced/separated	1.171 (1.136,1.207)		
- Widowed	1.341 (1.306,1.375)		
- Single	1.207 (1.176,1.240)	514.7, df=3	0.565 (0.559,0.570)
Race (n=34140)			
- White	reference		
- Black	1.186 (1.154, 1.219)		
- Asian/Pacific	1.062 (1.013,1.110)		
- American Native	1.122 (0.995,1.246)	133.8, df=3	0.522 (0.518,0.526)
State of residence (n=34215)			
- Alaska	0.832 (0.590, 1.081)		
- California	reference		
- Connecticut	0.857 (0.815,0.901)		
- Georgia	1.034 (1.001,1.067)		
- Hawaii	1.120 (1.041,1.197)		
- Iowa	0.900 (0.855,0.944)		
- Kentucky	1.031 (0.991,1.072)		
- Louisiana	1.044 (1.002,1.085)		
- Michigan	0.939 (0.897,0.981)		
- New Jersey	0.915 (0.884,0.947)		
- New Mexico	1.020 (0.953,1.087)		
- Utah	0.975 (0.903,1.046)		
- Washington	0.956 (0.915,0.997)	128.2, df=12	0.533 (0.527,0.539)
Cancer stage (n=34215)			
- Stage I	0.462 (0.445,0.480)		
- Stage II	0.464 (0.447,0.482)		
- Stage III	0.597 (0.579,0.615)		
- Stage IV	reference		
- Unknown	0.886 (0.866,0.905)	4216.5, df=4	0.689 (0.684,0.695)

df is an abbreviation of degrees of freedom

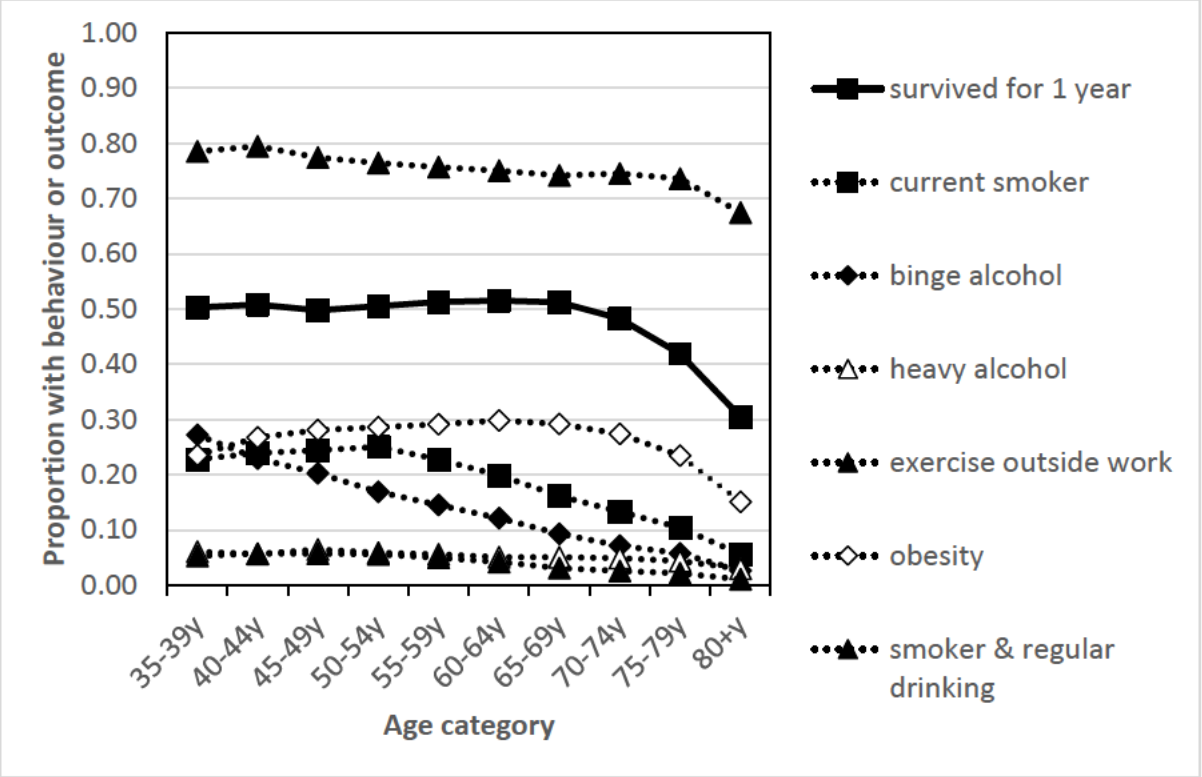
Among the demographic variables, 5-year age group was the strongest predictor of dead within 1 year of diagnosis. Many of the behaviour variables also differ by age group. For example, the relationship between age group and died within 1 year of diagnosis is present in Figure 5.8. To avoid confounding, the main analyses were age-adjusted.

Marital status, the next highest predictor, is strongly related to age. For example, 1.6% of those 35-39 year of age are widowed compared to 46.0% of those 80 or more years of age and 38.0% of those 35-39 years of age are single compared to 7.6% of those 80 or more years of age. Other demographic variables were found to be less predictive of death within one year of diagnosis.

As would be expected, cancer stage at diagnosis is a very strong predictor of death with one year of diagnosis with those diagnosed with stage I and II cancers having less than half the risk of dying as those diagnosed with stage IV disease.

In Figure 5.8 the association between age and one-year survival in SEER oesophageal cancer cases is depicted by the solid line and the relationships between age and health behaviours in the imputed data are shown. Notice that the proportion surviving one-year post diagnosis decreases in the older age groups as does the proportion with each health behaviour. As age is associated with both the predictor variable (health behaviour) and the outcome variable (death within one year of survival) it is likely to confound understanding of the relationship between these variables.

Figure 5.8 Proportion surviving and proportion with health behaviour by age group for SEER oesophageal cancer cases



5.3.8 Appendix 2 The relationship between the true relative risk and the imputed relative risk

In this section the mathematical relationship between the imputed relative risk and the true relative risk is derived. This shows how misclassification errors within the imputed health behaviours are corrected.

Let p_T represent the proportion of cancer cases who have the behaviour and p_i represent the proportion who are imputed to have that behaviour. If imputation were completely uninformative, the expected number of the cancer cases with the behaviour who are correctly imputed to have the behaviour would be np_Tp_i, just by chance alone. If imputation were informative, more of the imputed values would be expected to be correct

than by chance alone. This excess of correct matches can be modelled so long as estimates of p_T, p_i and the correlation between them are available (Oman & Zucker, 2001).

As the health behaviour of cancer patients is not recorded in the SEER cancer registry data, p_T is unknown. One temptation may be to estimate p_T based on the relative risk of incidence. For example, if current smoking is a risk factor for being diagnosed with oesophageal cancer with a relative risk of 2, then twice as many people with current smoking would be expected in the cancer population than in the general population and hence $p_T = 2p_i$. The flaw here is that p_i is not the smoking rate in the general population, it is the smoking rate in a sample with an identical demographic profile to the cancer cases.

For the current analyses it has been assumed that the proportion of imputed values with the behaviour provides a good approximation to the true proportion with the behaviour: $p_T = p_i$. This may or may not be true. (Suppose for example, even after correction for age, sex, race, marital status, State and year, that education status remained an independent predictor of the prevalence of current smoking. Then any difference in education status between the SEER cancer registry cases and BRFSS health survey respondents would produce differences between p_T and p_i.) It is assumed that the auxiliary variables which have been used are sufficient to encompass most of the variation in other factors which differ between the two data sets. That is, it is assumed that differences between the two data sets in, for example, education status are largely explained and corrected for by the existing auxiliary variables age, sex, race, marital status, State and year.

Using p_i as an estimate of p_T and the correlation between the two sets of imputed values as an estimate of the correlation between the true behaviour and the imputed behaviour the relationship between the true and imputed behaviour can be modelled as shown in Table 5.11.

Table 5.11 A model of the association between the true and imputed health behaviours

True (unknown) value of health behaviour	Imputed value of health behaviour: Behaviour present	Behaviour absent	total
Behaviour present			
- observed	A	B	$A+B$
- expected	$np_i^2 + np_i(1-p_i)\rho$	$np_i(1-p_i) - np_i(1-p_i)\rho$	np_T
Behaviour absent			
- observed	C	D	$C+D$
- expected	$np_i(1-p_i) - np_i(1-p_i)\rho$	$n(1-p_i)^2 + np_i(1-p_i)\rho$	$n(1-p_T)$
Total			
- observed	$A+C$	$B+D$	n
- expected	np_i	$n(1-p_i)$	n

Suppose the risk of death within one year of diagnosis is r_p for cancer cases where the behaviour is present and r_a for cancer cases where the behaviour is absent. The true relative risk of death within one year of diagnosis is $RR_T = r_p/r_a$. Notice that cells A and B in Table 5.11 have r_p risk of death and cells C and D have r_a risk. Cross tabulating imputed health behaviour against one-year survival status would produce the results summarized in Table 5.12.

Table 5.12 Cross-tabulation of true one-year survival status against imputed health behaviour

Imputed value of the health behaviour	1-year survival status: died	lived	total
Behaviour present			
- observed	I	J	$I+J = A+C$
- expected	$Ar_p + Cr_a$	$A(1-r_p) + C(1-r_a)$	np_i
Behaviour absent			
- observed	K	L	$K+L = B+D$
- expected	$Br_p + Dr_a$	$B(1-r_p) + D(1-r_a)$	$n(1-p_i)$
Total			
- observed	$I+K$	$J+L$	n
- expected			n

The relative risk calculated using the imputed data is

$$RR_i = \frac{I/(I+J)}{K/(K+L)}$$

$$= \frac{I/np_i}{K/n(1-p_i)}$$

$$= \frac{(Ar_p + Cr_a)/np_i}{(Br_p + Dr_a)/n(1-p_i)}$$

$$= \frac{(Ar_p + Cr_a)}{np_i} \times \frac{n(1-p_i)}{(Br_p + Dr_a)}$$

$$= \frac{(Ar_p + Cr_a)}{(Br_p + Dr_a)} \times \frac{(1-p_i)}{p_i}$$

This equation can be re-arranged to produce the formula for the true relative risk as follows:

$$p_i Br_p RR_i + p_i Dr_a RR_i = Ar_p + Cr_a - Ar_p p_i - Cr_a p_i$$

$$r_p(p_i BRR_i - A + Ap_i) = r_a(C - Cp_i - p_i DRR_i)$$

$$RR_T = \frac{r_p}{r_a} = \frac{C - Cp_i - p_i DRR_i}{p_i BRR_i - A + Ap_i}$$

Now substituting the models for A, B, C and D shown in Table 5.11:

$$RR_T = \frac{n(p_i(1-p_i) - p_i(1-p_i)\rho) - n\big((p_i(1-p_i) - p_i(1-p_i)\rho)\big)p_i - p_i\Big(n\big((1-p_i)(1-p_i) + p_i(1-p_i)\rho\big)\Big)RR_i}{p_i\big(n(p_i(1-p_i) - p_i(1-p_i)\rho)\big)RR_i - n(p_ip_i + p_i(1-p_i)\rho) + \big(n(p_ip_i + p_i(1-p_i)\rho)\big)p_i}$$

Note n and p_i cancel out to give:

$$RR_T = \frac{(1-p_i) - (1-p_i)\rho - (p_i(1-p_i) - p_i(1-p_i)\rho) - \big((1-p_i)(1-p_i) + p_i(1-p_i)\rho\big)RR_i}{(p_i(1-p_i) - p_i(1-p_i)\rho)RR_i - (p_i + (1-p_i)\rho) + p_ip_i + p_i(1-p_i)\rho}$$

$$= \frac{1 - p_i - (\rho - p_i\rho) - \Big(p_i - p_i^2 - (p_i\rho - p_i^2\rho)\Big) - (1 - p_i - p_i + p_i^2 + (p_i - p_i^2)\rho)RR_i}{(p_i - p_i^2 - (p_i - p_i^2)\rho)RR_i - (p_i + \rho - p_i\rho) + p_ip_i + (p_i - p_i^2)\rho}$$

$$= \frac{1 - p_i - \rho + p_i\rho - (p_i - p_i^2 - p_i\rho + p_i^2\rho) - (1 - p_i - p_i + p_i^2 + p_i\rho - p_i^2\rho)RR_i}{(p_i - p_i^2 - p_i\rho + p_i^2\rho)RR_i - p_i - \rho + p_i\rho + p_ip_i + p_i\rho - p_i^2\rho}$$

$$= \frac{1 - p_i - \rho + p_i\rho - p_i + p_i^2 + p_i\rho - p_i^2\rho - (RR_i - p_iRR_i - p_iRR_i + p_i^2RR_i + p_i\rho RR_i - p_i^2\rho RR_i)}{p_iRR_i - p_i^2RR_i - p_i\rho RR_i + p_i^2\rho RR_i - p_i - \rho + p_i\rho + p_ip_i + p_i\rho - p_i^2\rho}$$

$$= \frac{1 - p_i - \rho + p_i\rho - p_i + p_i^2 + p_i\rho - p_i^2\rho - RR_i + p_iRR_i + p_iRR_i - p_i^2RR_i - p_i\rho RR_i + p_i^2\rho RR_i}{p_iRR_i - p_i^2RR_i - p_i\rho RR_i + p_i^2\rho RR_i - p_i - \rho + p_i\rho + p_i^2 + p_i\rho - p_i^2\rho}$$

$$= \frac{-RR_i + 1 + p_i RR_i - p_i - \rho + p_i\rho + p_i RR_i - p_i - p_i^2 RR_i + p_i^2 - p_i\rho RR_i + p_i\rho + p_i^2\rho RR_i - p_i^2\rho}{p_i RR_i - p_i - p_i^2 RR_i + p_i^2 - p_i\rho RR_i + p_i\rho + p_i^2\rho RR - p_i^2\rho - \rho + p_i\rho}$$

$$= \frac{-(RR_i - 1) + p_i(RR_i - 1) - \rho(1 - p_i) + p_i(RR_i - 1) - p_i^2(RR_i - 1) - p_i\rho(RR_i - 1) + p_i^2\rho(RR_i - 1)}{p_i(RR_i - 1) - p_i^2(RR_i - 1) - p_i\rho(RR_i - 1) + p_i^2\rho(RR - 1) - \rho(1 - p_i)}$$

$$= \frac{-(RR_i - 1) + (RR_i - 1)(p_i - p_i^2 - p_i\rho + p_i^2\rho) - \rho(1 - p_i) + p_i(RR_i - 1)}{(RR_i - 1)(p_i - p_i^2 - p_i\rho + p_i^2\rho) - \rho(1 - p_i)}$$

$$RR_T = \frac{(RR_i - 1)(p_i - p_i^2 - p_i\rho + p_i^2\rho) - \rho(1 - p_i) - (1 - p_i)(RR_i - 1)}{(RR_i - 1)(p_i - p_i^2 - p_i\rho + p_i^2\rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)(p_i - p_i^2 - p_i\rho + p_i^2\rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)p_i(1 - p_i - \rho + p_i\rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)p_i(1 - p_i)(1 - \rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(RR_i - 1)}{(RR_i - 1)p_i(1 - \rho) - \rho}$$

Notice that when $RR_i = 1$ then $RR_T = 1$ as would be expected. However, extreme values of p_i and/or ρ may be problematic. When $\rho = 0$ then $RR_T = 1 - \frac{1}{p_i}$ which is always negative and hence an impossible value for a relative risk. Also when either p_i or $(1 - p_i)$ approach 0, RR_T approaches $1 - \frac{(RR_i - 1)}{0 - \rho}$, or $(RR_T - 1)$ approaches $\frac{1}{\rho}(RR_i - 1)$. That is when p_i or $(1 - p_i)$ are close to zero, RR_T becomes particularly sensitive to small ρ.

5.3.9 **Appendix 3 Simulating data with prescribed relative risks**

The aim is to assign health behaviour and survival status to cancer cases in such a way as to produce the target relative risk RR_t while maintaining the proportion of people with the behaviour and proportion of people dying within one year at the true levels. It was assumed that the first set of imputed values were actually the true measurements of behaviour. This

ensured the correct proportion of people with the behaviour. Survival status was then simulated for these behaviours to deliver the target relative risk.

In Table 5.13 below:

$A + B$ is the observed number of cancer cases who have the behaviour;

$C + D$ is the observed number of cancer cases who do not have the behaviour;

$A + C$ is the observed number of cancer cases who died within 1 year;

$B + D$ is the observed number of cancer cases who lived for 1 year or more; and

a' and c' be the number of deaths required to produce a target relative risk, RR_t

Table 5.13 Values required for the simulation

		1 year survival status		
		died	lived	total
Imputed value of the health behaviour	Behaviour present	a'		$A + B$
	Behaviour absent	c'		$C + D$
	Total	$A + C$	$B + D$	n

To achieve any given target relative risk RR_t, a' and c' need to be selected such that

$$\frac{a'/(A + B)}{c'/(C + D)} = RR_t$$

First, note that

$$a' + c' = A + C$$

So a' can be replaced by

$$a' = (A + C) - c'$$

To get:

$$\frac{\big((A+C)-c'\big)/(A+B)}{c'/(C+D)} = RR_t$$

Solving for c' :

$$\frac{(A+C)-c'}{A+B} = \frac{RR_t c'}{C+D}$$

$$(A+C)(C+D) - c'(C+D) = RR_t c'(A+B)$$

$$(A+C)(C+D) = RR_t c'(A+B) + c'(C+D)$$

$$(A+C)(C+D) = c'\big(RR_t(A+B) + (C+D)\big)$$

$$c' = \frac{(A+C)(C+D)}{RR_t(A+B) + (C+D)}$$

Which, in turn, allows the value to be calculated for:

$$a' = (A+C) - c'$$

For data records in the behaviour present group 1-year survival status was randomly assigned with the probability of dying within 1 year equal to $a'/(A+B)$. For those imputed to be in the behaviour absent group 1-year survival status was randomly assigned with the probability of dying within 1 year equal to $c'/(C+D)$.

These random selections were implemented using the `sample()` command in R software.

5.4 Further exploration of matching

Table 5.6 shows that 15.9% of cancer cases were imputed to be current smokers 5 years prior to diagnosis. In the corresponding years, 2001 to 2009, the overall US smoking rate was estimated to have considerably higher, ranging from 22.8% to 20.6% (Méndez et al., 2016). But cancer patients are not expected to be a random sample from the wider population and not expected to have the same health behaviours as the general population. For example, the incidence of OC in the US is known to be higher in older ages and males (American Cancer Society, 2022). Perhaps the older age profile led to the lower smoking rates in the US OC population than in the US general population.

However, other comparable (Western) studies examining pre-diagnosis health behaviours in OC, listed in Table 2.1, also reported higher proportions of pre-diagnosis current smoking in OC. For example, in the US in the slightly earlier period 1999-2004 one study reported that 28% of OC cases had smoked within 1 year prior to diagnosis (Cescon et al., 2009) while another, which included cases from Canada as well as the US, reported that 19.1% of their OC cases were smokers (Anna Spreafico et al., 2017). Further afield pre-diagnosis smoking rates were reported as 22.5% in a 2006-2013 study from Canada (Korpanty et al., 2017), 25.1% in a 2001-2005 study from Australia (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012) and 22.5% in a 1994-1997 study from Sweden (Sundelöf et al., 2008).

Table 5.6 also shows that when using the imputed values, the estimated prevalence of binge drinking (males ≥5 drinks on one occasion, females ≥4 drinks on one occasion) 5 years prior to diagnosis was 10% and the estimated prevalence of heavy drinking (men ≥2 drinks/day, women ≥1 drinks/day) was 4.8%. In contrast the estimated 2018 US prevalence of binge drinking was 26.5% and for heavy drinking was 6.6% (National Institute on Alcohol Abuse and Alcoholism, 2021a). But again, there is no reason to expect cancer cases to be a

representative sample from the wider US population or to exhibit the same profile of alcohol consumption.

No comparable (Western) studies of pre-diagnosis alcohol consumption in OC cases used comparable definitions of 'at risk' alcohol behaviour. One study from Scotland in 1996-2010 applied higher alcohol cutpoints of 2-3 units/day for women and 3-4 units per day for men and classified 8.1% of OC cases as heavy drinkers pre-diagnosis. Studies from Australia 2001-2005 (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012) and Sweden 1994-1997 (Sundelöf et al., 2008) provided lower alcohol cutpoints of ≥70 grams/week per week and >70 grams/week (20 years before interview) respectively and reported prevalences of 63.5% and 33.9% among OC patients respectively. No comparative results were available for 'binge' drinking or 'current smoker with regular drinking'. Although evidence is limited, it seems possible that the prevalence of pre-diagnosis 'at risk' alcohol consumption in OC may be underestimated when using cold deck imputation.

Results for obesity are more varied. From Table 5.6, the estimated rate of pre-diagnosis obesity in OC arising from the imputed data is 25.7%. This is around the centre of the 32.3% reported in a 2004-2016 US study (Loehrer et al., 2020), 19.1% reported in a 1994-2004 US/Canada study (A. Spreafico et al., 2017) and 26.8% reported in a 2001-2005 Australian study (Thrift, Nagle, Fahey, Russell, et al., 2012; Thrift, Nagle, Fahey, Smithers, et al., 2012).

The estimated prevalence of pre-diagnosis physical activity outside of employment within the imputed data was 73.7%, but no comparative results were available.

One possible explanation for these findings is that people who engage in risky behaviours, such as heavy alcohol consumption or smoking, are less likely to participate in

the BRFSS health survey. Alternatively, when these people do participate in the BRFSS health survey they under-report their risk-taking behaviours.

The question arising is whether the potential underestimate of the prevalence of pre-diagnosis health behaviours is of relevance to the research aim of this thesis: to describe the association between pre-diagnosis health behaviour and post-diagnosis survival. If for example, pre-diagnosis smokers are twice as likely to die within 12 months of diagnosis, this is true no matter how many smokers and non-smokers are included in the sample. Relative risk is independent of prevalence.

However, in this thesis behaviour status is not measured but is imputed using a method expected to produce high misclassification rates. Section '5.3.8 Appendix 2 The relationship between the true relative risk and the imputed relative risk' develops the formula for misclassification correction under the assumption that the imputed prevalence of behaviour, p_i, is equal to the true prevalence of behaviour, p_T. Replacing the $p_T = p_i$ assumption with $p_T > p_i$ produces the following alternate calculations and result.

Table 5.14 A model of the association between the true and imputed health behaviours assuming imputation underestimates true prevalence of the behaviour

True (unknown) value of health behaviour	Imputed value of health behaviour		
	Behaviour present	Behaviour absent	total
Behaviour present			
- observed	A	B	$A+B$
- expected	$np_ip_T + np_i(1-p_T)\rho$	$n(1-p_i)p_T - np_i(1-p_T)\rho$	np_T
Behaviour absent			
- observed	C	D	$C+D$
- expected	$np_i(1-p_T) - np_i(1-p_T)\rho$	$n(1-p_i)(1-p_T) + np_i(1-p_T)\rho$	$n(1-p_T)$
Total			
- observed	$A+C$	$B+D$	n
- expected	np_i	$n(1-p_i)$	n

$$RR_T = \frac{n(p_i(1-p_T) - p_i(1-p_T)\rho) - n\big((p_i(1-p_T) - p_i(1-p_T)\rho)\big)p_i - p_i\Big(n\big((1-p_i)(1-p_T) + p_i(1-p_T)\rho\big)\Big)RR_i}{p_i\big(n(p_T(1-p_i) - p_i(1-p_T)\rho)\big)RR_i - n(p_ip_T + p_i(1-p_T)\rho) + \big(n(p_ip_T + p_i(1-p_T)\rho)\big)p_i}$$

Note n and p_i cancel out to give:

$$RR_T = \frac{(1-p_T) - (1-p_T)\rho - (p_i(1-p_T) - p_i(1-p_T)\rho) - \big((1-p_i)(1-p_T) + p_i(1-p_T)\rho\big)RR_i}{(p_T(1-p_i) - p_i(1-p_T)\rho)RR_i - (p_T + (1-p_T)\rho) + p_ip_T + p_i(1-p_T)\rho}$$

$$= \frac{1-p_T - (\rho - p_T\rho) - \big(p_i - p_ip_T - (p_i\rho - p_ip_T\rho)\big) - (1-p_i-p_T+p_ip_T + (p_i - p_ip_T)\rho)RR_i}{(p_T - p_ip_T - (p_i - p_ip_T)\rho)RR_i - (p_T + \rho - p_T\rho) + p_ip_T + (p_i - p_ip_T)\rho}$$

$$= \frac{1 - p_T - \rho + p_T\rho - (p_i - p_ip_T - p_i\rho + p_ip_T\rho) - (1 - p_i - p_T + p_ip_T + p_i\rho - p_ip_T\rho)RR_i}{(p_T - p_ip_T - p_i\rho + p_ip_T\rho)RR_i - p_T - \rho + p_T\rho + p_ip_T + p_i\rho - p_ip_T\rho}$$

$$= \frac{1 - p_T - \rho + p_T\rho - p_i + p_ip_T + p_i\rho - p_ip_T\rho - (RR_i - p_iRR_i - p_TRR_i + p_ip_TRR_i + p_i\rho RR_i - p_ip_T\rho RR_i)}{p_TRR_i - p_ip_TRR_i - p_i\rho RR_i + p_ip_T\rho RR_i - p_T - \rho + p_T\rho + p_ip_T + p_i\rho - p_ip_T\rho}$$

$$= \frac{1 - p_T - \rho + p_T\rho - p_i + p_ip_T + p_i\rho - p_ip_T\rho - RR_i + p_iRR_i + p_TRR_i - p_ip_TRR_i - p_i\rho RR_i + p_ip_T\rho RR_i}{p_TRR_i - p_ip_TRR_i - p_i\rho RR_i + p_ip_T\rho RR_i - p_T - \rho + p_T\rho + p_ip_T + p_i\rho - p_ip_T\rho}$$

$$= \frac{-RR_i + 1 + p_iRR_i - p_i - \rho + p_T\rho + p_TRR_i - p_T - p_ip_TRR_i + p_ip_T - p_i\rho RR_i + p_i\rho + p_ip_T\rho RR_i - p_ip_T\rho}{p_TRR_i - p_T - p_ip_TRR_i + p_ip_T - p_i\rho RR_i + p_i\rho + p_ip_T\rho RR - p_ip_T\rho - \rho + p_i\rho}$$

$$= \frac{-(RR_i - 1) + p_i(RR_i - 1) - \rho(1 - p_T) + p_T(RR_i - 1) - p_ip_T(RR_i - 1) - p_i\rho(RR_i - 1) + p_ip_T\rho(RR_i - 1)}{p_T(RR_i - 1) - p_ip_T(RR_i - 1) - p_i\rho(RR_i - 1) + p_ip_T\rho(RR - 1) - \rho(1 - p_i)}$$

$$= \frac{-(RR_i - 1) + (RR_i - 1)(p_T - p_ip_T - p_i\rho + p_ip_T\rho) - \rho(1 - p_i) + p_i(RR_i - 1)}{(RR_i - 1)(p_T - p_ip_T - p_i\rho + p_ip_T\rho) - \rho(1 - p_i)}$$

$$RR_T = \frac{(RR_i - 1)(p_T - p_i p_T - p_i\rho + p_i p_T \rho) - \rho(1 - p_i) - (1 - p_i)(RR_i - 1)}{(RR_i - 1)(p_T - p_i p_T - p_i\rho + p_i p_T \rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)(p_i - p_i p_T - p_i\rho + p_i p_T \rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)p_i(1 - p_T - \rho + p_T \rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(1 - p_i)(RR_i - 1)}{(RR_i - 1)p_i(1 - p_T)(1 - \rho) - \rho(1 - p_i)}$$

$$= 1 - \frac{(RR_i - 1)(1 - p_i)}{(RR_i - 1)p_i(1 - \rho)(1 - p_T) - \rho(1 - p_i)}$$

$$= 1 - \frac{(RR_i - 1)}{(RR_i - 1)p_i(1 - \rho)\frac{(1 - p_T)}{(1 - p_i)} - \rho}$$

That is, replacing the assumption $p_T = p_i$ with $p_T > p_i$, introduces the new term $\frac{(1-p_T)}{(1-p_i)}$ into the equation. As $p_T > p_i$, $(1 - p_T) < (1 - p_i)$ then $\frac{(1-p_T)}{(1-p_i)} < 1$. If for example, the imputed prevalence of pre-diagnosis smoking p_i=0.159 is only half of the true prevalence then p_T=0.318 then $\frac{(1-p_T)}{(1-p_i)} = \frac{(1-0.318)}{(1-0.159)} = 0.811$. However, as the $p_T > p_i$ assumption may also some impact on estimates of RR_i and ρ there are too many unknowns to predict how the misclassification corrected relative risk will be affected.

Instead simulated data are used to help illustrate and quantify the various relationships. This simulation is described as a series of steps supported by a numeric example.

Step 1: Set the population parameters

The analysis of the association between pre-diagnosis smoking status and post-diagnosis 1 year survival, reported in the first line of Table 5.6, is simulated. The simulation contains the same number of data records (n=27,825), the same proportion of deaths one-year post-diagnosis (14,927/27,825) and the same prevalence of smoking ($p_i = 0.159$) as the true SEER OC data set. The simulation will allow $p_T > p_i$. Specifically, the 'true' prevalence of pre-diagnosis smoking, p_T, will be set twice as high as the prevalence of pre-diagnosis smoking in the imputed behaviour, so $p_T = 2 \times p_i = 0.318$. This is chosen as a plausible upper limit for the true prevalence given the observed prevalences (0.191, 0.225, 0.225, 0.251 and 0.280) from the comparable studies presented above.

Step 2: Randomly allocate pre-diagnosis smoking status (smoker or non-smoker) and survival status (dead 1 year after diagnosis or alive 1 year after diagnosis), with a known relative risk of say 2.0.

The relative risk will be achieved using the method developed in section '5.3.9 Appendix 3 Simulating data with prescribed relative risks'. A cross-tabulation from one simulated data set is shown in Table 5.15.

Table 5.15 Simulated true one-year survival status cross-tabulated against smoking status

Pre-diagnosis smoking status	Post-diagnosis 1-year survival status		
	died	lived	total
Smoker	7,244	1,723	8,967
Non-smoker	7,636	11,232	18,868
Total	14,880	12,955	27,835

This simulated data set has the following prescribed characteristics:

$$n = 27835, \quad n_{dead} = 14880, \quad p_t = \frac{8967}{27835} = 0.322$$

$$RR_t = \frac{7244/8967}{7636/18868} = \frac{0.8079}{0.4047} = 1.996$$

Step 3: Randomly allocate imputed smoking status with prevalence of 0.159.

A new 'imputed' smoking variable is added such that

$$n = 27835, \quad n_{smoker} = 4491$$

$$p_i = \frac{4491}{27835} = 0.161$$

Step 4: Create a relationship between the 'imputed' smoking status and 'real' smoking status.

From the first line of Table 5.6, the median number of matches in excess of chance between the 'real' and 'simulated' pre-diagnosis smoking status is approximately 262. This improvement over chance was simulated by randomly selecting 262 'true' smokers who were 'imputed' to be non-smokers and 262 'true' non-smokers who were 'imputed' to be smokers and changing their imputed smoking status to match the truth. Table 5.16 shows the cross-tabulation of 'true' and 'imputed' pre-diagnosis smoking status in the example data set.

Table 5.16 Simulated data set, cross-tabulation of 'true' and 'imputed' pre-diagnosis smoking status

'True' smoking status	'Imputed' smoking status		
	Smoker	Non-smoker	total
Smoker	1,699 (0.061)	7,268 (0.261)	8,967
Non-smoker	2,792 (0.100)	16,076 (0.578)	18,868
Total	4,491	23,344	27,835

This table meets the required characteristics as

$$p_t = \frac{8967}{27835} = 0.322 \text{ and } p_i = \frac{4491}{27835} = 0.161$$

The results in Table 5.16 allow us to estimate the correlation $\hat{\rho}$ between 'true' and 'imputed' smoking status. When $p_T = p_i$ is assumed, the excess matches are modelled as $np_i(1-p_i)\rho$ (see Table 5.11) which implies

$$\hat{\rho} = \frac{262}{np_i(1-p_i)} = \frac{262}{27825 \times 0.159 \times (1-0.159)} = 0.070$$

but when $p_T > p_i$, the relationship is modelled as $np_i(1-p_T)\rho$ (see Table 5.14) which implies

$$\hat{\rho} = \frac{262}{np_i(1-p_t)} = \frac{262}{27825 \times 0.159 \times (1-0.318)} = 0.087$$

Step 5: Calculate the relative risk using the 'imputed' smoking status

Table 5.17 summarises the relationship between the imputed pre-diagnosis smoking status and 'real' 1 year post-diagnosis survival status for the example simulated data set.

Table 5.17 Simulated true one-year survival status cross-tabulated against smoking status

'Simulated'	1-year survival status		
smoking status	died	lived	total
Smoker	2,509	1,982	4,491
Non-smoker	12,371	10,973	23,344
Total	14,880	12,955	27,835

Notice that this data set complies with the requirements as

$$n = 27835, \quad n_{dead} = 14880$$

$$p_t = \frac{4491}{27835} = 0.161$$

Analysing the imputed smoking status, the relative risk is:

$$RR_i = \frac{2509/4491}{12371/23344} = \frac{0.5586}{0.5299} = 1.054$$

This estimate of the relative risk is attenuated towards the null by the high levels of misclassification error arising from the imputation. Applying misclassification correction under the incorrect assumption $p_t = p_i$ produces a misclassification corrected estimated relative risk of

$$RR_t = 1 - \frac{(RR_i - 1)}{(RR_i - 1)p_i(1 - \rho)\frac{(1 - p_T)}{(1 - p_i)} - \rho}$$

$$RR_t = 1 - \frac{(1.054 - 1)}{(1.054 - 1)0.161(1 - 0.069) - 0.069} = 1.883$$

Knowing the true prevalence of pre-diagnosis smoking was $p_t = 0.312$ and applying the formula allowing $p_t > p_i$ the misclassification corrected estimate of the relative risk is:

$$RR_t = 1 - \frac{(RR_i - 1)}{(RR_i - 1)p_i(1 - \rho)\frac{(1 - p_T)}{(1 - p_i)} - \rho}$$

$$RR_t = 1 - \frac{(1.054 - 1)}{(1.054 - 1)0.161(1 - 0.087)\frac{(1 - 0.322)}{(1 - 0.161)} - 0.086} = 1.670$$

Step 6: Repeat the simulation

This simulation was repeated 100 times and estimated relative risks recorded for each repetition. In the simulation, the 'true' proportion of OC cases who were current smokers pre-diagnosis was set at 0.318, twice as high as the 'imputed' proportion. The relative risk of death within 1-year post-diagnosis was set at 2.0.

It was demonstrated that underestimating the true prevalence of smoking during misclassification correction process caused the misclassification corrected relative risk to be underestimated. If the 'true' proportion of pre-diagnosis smokers, 0.318, were known and adjusted for in the misclassification correction the relative risk of 2.0 was estimated to be just 1.65 with an associated 95% empirical confidences interval of 1.28 to 2.13. This inaccuracy appears to arise through the definition and measurement of the level of excess agreement beyond chance (effectively the information component of the matching process). The higher assumed smoking rate, the higher the number of smoker-to-smoker matches which are expected to arise by chance alone. Some of the excess deaths become falsely attributed to chance and misclassification correction is underestimated.

The problem appears to be partially alleviated by the algorithm which assumes the 'true' proportion of smokers is accurately reflected in 'imputed' smoking status. In this case the median misclassification corrected relative risk was estimated to be 1.83 with an associated 95% empirical confidence interval from 1.35 to 2.50.

The problems associated with underestimating the true prevalence of behaviour appear to be inherent to the calibration step, the measurement of 'excess' information beyond chance produced by imputation. An entirely new approach may be required.

In the meantime, the misclassification corrected relative risks which are being reported may be underestimates of the true relative risk. Fortunately, the effects are modest with the median relative risk of 1.83 being a relatively small underestimate of the true relative risk of 2.0 even when the proportion of smokers was underestimated by a factor of two.

5.5 Critique of the method

The above published peer-reviewed paper describes a method to impute 100% missing health behaviour variables in the SEER cancer registry data set, using the BRFSS health survey as an external reference. The simulation studies (Table 5.7) demonstrated that, so long as the incidence of the behaviour was greater than about 5%, the algorithm returned accurate relative risks estimates (to the first decimal place), but with much reduced precision (wide confidence intervals reflecting the additional random error arising from misclassification).

There were three main steps in the process: imputing the behaviour (twice); measuring misclassification by comparing the two imputed values; and incorporating misclassification correction into the final analysis.

For this trial cold deck imputation was used; allowing BRFSS health survey respondents to donate their behaviour to demographically similar OC cases. The simulation studies demonstrated that the cold deck imputation allowed information to be transferred between these independent data sets. However, the rigid rule that only people from a corresponding age by sex by marital status by race by State of residence by year group could donate to the OC case, left 4,353 (13.5%) of the 32,188 eligible OC cases without the required two donors and hence excluded from the analysis. These exclusions created potential biases with younger, female, and white race OC cases overrepresented in the analysis.

Lack of donor records could be addressed by some ‘nearest neighbour’ matching rules allowing the unmatched OC cases to be matched with BRFSS health survey respondents from different but similar demographic groups (such as one age group younger or older, neighbouring States, adjacent years, etc). This would increase the sample size, increase statistical power and reduce the potential non-response bias for, presumably, very little loss of precision.

Alternatively, the imputation could be based on the estimated probability of engaging in the behaviour obtained from the logistic regression models described in Chapter 4. That is, each OC case could have each behaviour randomly assigned relative to their estimated probability. Using the logistic models would allow behaviour to be imputed for all OC cases and, if required, could be extended to imputing more than two values for each behaviour for each OC case. A potential weakness of this approach is that, as the models do not contain higher order interactions terms, they may fail to identify more isolated pockets of behaviour. Machine learning techniques (Westreich, Lessler, & Funk, 2010) may provide a more natural model for complex systems although some investigators have suggested there is little difference in predictive ability (Christodoulou et al., 2019; Faisal et al., 2020).

To estimate the misclassification rate in the imputed data, the misclassification between two imputed values for the behaviour was used. Consider an example demographic group which contains at least one OC case: 60-64 year old, female, white, widowed, California. There will be hundreds of US adults falling into this demographic group. A small number of these will be diagnosed with OC in a particular year and a small number may respond to the BRFSS health survey in the corresponding year. The estimate of the misclassification rate assumes that, in each particular demographic group, the difference in behaviour between the BRFSS survey respondents is about the same magnitude as the difference in behaviour between the OC cases and BRFSS survey respondents. This

assumption would be untrue if there was another factor which differentiated OC cases and BRFSS survey respondents in the same demographic cluster. Consider if, within the cluster, the OC cases had lower education hence were more likely to engage in smoking and BRFSS survey respondents had higher education and hence less likely to engage in smoking. In this case, the agreement in smoking status between BRFSS survey respondents would be greater than the agreement between OC cases and BRFSS survey respondents. Any such error in the misclassification rate could have a significant impact on the final, misclassification corrected, estimate of relative risk.

The issue could be addressed by adding more, and more informative, demographic variables in the imputation process. In the above example, it would be addressed by adding an education variable. But it is highly unlikely that compatible education variables will be added to both the cancer registry and health survey data collections in the foreseeable future. Instead it is necessary to validate whether the variables which are currently available are sufficient to provide an accurate estimate the misclassification rate between the imputed and true behaviour.

There is also some evidence to suggest that the imputation process underestimated the proportion of people with ‘at risk’ pre-diagnosis behaviour in the OC cases. It has been shown that this could lead to the underestimation of misclassification corrected relative risk. The calibration step warrants further development or complete replacement with a model which allows underestimation of the prevalence of behaviour to be specifically estimated and corrected.

This validation if all the information used in the analysis were available, plus the true pre-diagnosis health behaviour for a group of OC cases and a data set large enough to support the method (i.e. many thousands). But if such a data set existed, the first research aim could

have been addressed using standard analyses. The rationale for this thesis is that the required data set does not exist. It may be possible to conduct the validation step using a more common cancer, such as colorectal cancer, where one-to-one linkage between the cancer registry and health survey data sets may provide a large data set. Of course, colorectal cancer has longer survival times, perhaps weakening the carry-over effects of pre-diagnosis behaviour on post-diagnosis survival.

There are likely to be other methods to measure misclassification rates which have not yet been considered. Strong agreement between two independent measures of the misclassification rate could also help establish the validity of both.

To correct for misclassification, a formula was derived based on the phi statistic as a measure of correlation ρ. It was observed that the method sometimes failed, returning impossible negative relative risks (Table 5.7). This occurred when the imputation process delivered very few correct matches in excess of chance agreement. That is, there wasn't enough information for the algorithm to use despite the very large sample sizes. There are other approaches to misclassification adjustment which have not yet trialled (Barron, 1977; Copeland, Checkoway, McMichael, & Holbrook, 1977). Further research is warranted to clarify whether there are any practical differences between the different approaches to misclassification correction and, if so, which misclassification adjustment is best suited to the very high levels of misclassification that are evident in this thesis.

The imputation-based approach to addressing the missing health behaviour of OC cases has provided results (relative risks) which are more readily interpretable than the results in the previous Chapter (hazard ratios for 10% change in probability of having the behaviour). However, dichotomising survival time into survival status at 1-year failed to use all of the information available in the data set. That is, the 301 (about 1%) OC cases who

were censored at less than 1-year (Figure 5.6) were excluded from the analysis and the continuous measure of survival time in months was dichotomised for all other OC cases.

The thesis research question was not limited to any specific time point and so the analysis above does not fully address the research question.

5.6 Closing comments

The analyses in this Chapter have demonstrated the plausibility of an imputation-based method for quantifying the relationship between pre-diagnosis health behaviour and post-diagnosis 1-year survival in OC cases. As such some progress has been made towards addressing the second aim of this thesis. However, the method is not universal. It appears to be limited to behaviours which are sufficiently common, data sets which share informative demographic measures, and large sample sizes. Further exploration is warranted.

The results in this Chapter (such as Table 5.8) have also progressed the first aim of the thesis which was to describe the relationship between pre-diagnosis health behaviour and post-diagnosis survival times in OC cases. Point estimates of relative risks were obtained for current smoking, binge drinking, and physical activity for ESCC and EAC in agreement with previous research and expectations. But one concern is whether relative risk of survival for some fixed period (1-year in the current analyses) is sufficient to address this first research question. Survival analysis, which uses all of the survival time data and described differences survival trajectories between OC cases with and without the behaviour over time may provide a more complete answer.

Therefore, while progress has been made in addressing both aims of the research thesis, further investigations are warranted to see if a more general solution can be found.

Chapter 6 discusses generalising the ‘impute, impute, calibrate the misclassification, correct for this misclassification’ algorithm for use with survival analysis.

Chapter 6 Extending the imputation-based approach to Cox regression

6.1 Background

In the previous chapter a novel algorithm for imputing the missing pre-diagnosis behaviours of OC cancer registry cases was developed and evaluated. The algorithm was successfully employed to estimate the relative risk of death within 1-year post diagnosis, at least for pre-diagnosis smoking status and exercise behaviours outside of employment. The face validity of these relative risk estimates was confirmed through simulations and comparison with the existing literature. Overall, this algorithm progressed both the second and the first aims of this thesis.

However, the misclassification corrected relative risk in the previous Chapter cannot be easily generalised to include multiple predictors of survival. Many factors impact on a person's risk of death. A model-based approach, such as logistic regression or log-binomial regression, would allow more predictor variables, including multiple health behaviours, to be included within the same model and would allow interactions to be explored. The models that could be used to predict 1-year survival include logistic regression and log-binomial regression. There are established methods for misclassification correction in logistic regression (Spiegelman, Rosner, & Logan, 2000) and applying these would allow a considerably broader range of analyses.

However, even logistic and log-binomial regression models fail to use all of the available information in the cancer registry data set. With survival time in months available for each OC case, it is unnecessarily restrictive to dichotomise these data at 1-year survival

status (or any other single time point). Survival analysis allows inclusion of multiple predictor variables and estimating interactions and uses the full information on survival time.

To extend the initial imputation-based algorithm to be of more practical use, in this Chapter misclassification corrected Cox regression models are developed as the analysis step of the algorithm.

In Section 6.2 the misclassification correction method when fitting Cox proportional hazards models is introduced. Section 6.3 contains a journal article currently under peer-review, which demonstrates the misclassification corrected Cox model within the imputation-based algorithm for replacing 100% missing pre-diagnosis health behaviours.

In the previous Chapter age-confounding was addressed through stratification (Cochran-Mantel-Haenszel method) but in the analyses that follow it is shown that the sample size is too small to support age-stratified misclassification corrected Cox models. In Section 6.4 investigation of sample size and age-stratification is extended with some informal analyses. Examples of the important coding elements in simulating survival times, running the misclassification corrected Cox models and visualisation of the results are presented int eh thesis Appendix. Section 6.5 critiques the methods introduced in this Chapter and Section 6.6 reviews progress towards the thesis goals and introduces the final discussion Chapter of the thesis.

6.2 Misclassification correction in Cox regression

The primary focus of this Chapter is to use misclassification corrected Cox models in the analysis phase of the algorithm for replacing 100% missing health behaviour data. Bang et al (2013) reviewed five distinct approaches to misclassification correction in Cox

regression. However, given the characteristics of the current research (dichotomous misclassified predictor variables and external estimates of misclassification error), only the corrected score method (Zucker & Spiegelman, 2008) was applicable to the analyses in this thesis. The corrected score method is applied as part of the model fitting process for Cox regression. Full details are available in the source article (Zucker & Spiegelman, 2008). This section contains a brief overview how the method is applied.

A simple Cox regression model, without time-varying predictors or censored observations, can be written

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})$$

where

- $\boldsymbol{X}$ represents the matrix of observed values for each individual (rows) on each predictor variable (columns)
- $\lambda(t|\boldsymbol{X})$ represents the hazard (predicted probability of dying at time t given value of the predictor variables in $\boldsymbol{X}$)
- $\lambda_0(t)$ represents the baseline hazard (predicted probability of dying at time t when the value of all the predictor variables in $\boldsymbol{X}$ are zero)
- $\boldsymbol{\beta}$ is the vector of coefficients which describe the relationship between the predictor variables $\boldsymbol{X}$, and the probability of dying at time t, $\lambda(t|\boldsymbol{X})$

Suppose $(Y_i, \boldsymbol{X}_i)$ for individual i was observed, where

- Y_i is the survival time for individual i
- $\boldsymbol{X}_i$ represents the values of the predictor variable for individual i

The aim of fitting the Cox model is to provide the most accurate possible prediction of the survival time Y_i using only the predictor variables $\boldsymbol{X}_i$ for all individuals $i = 1, \dots, n$. In practice, this is achieved by identifying the set of regression coefficients $\boldsymbol{\beta}$ which maximise the probability (likelihood) of the sample displaying the observed survival times $\boldsymbol{Y}$ for the observed predictor variables $\boldsymbol{X}$**.** So, to fit the model the likelihood first needs to be considered.

Suppose each survival time is unique (no ties) such that there are n distinct survival times. Let the risk set be $\boldsymbol{\mathcal{R}}(t) = \{i : Y_i \geq t\}$ the set of individuals who are "at risk" for failure at time t (i.e. have not died before t). At each failure time Y_j , the contribution to the likelihood is:

$$L_j(\boldsymbol{\beta}) = P(\text{individual } j \text{ fails} \mid \text{one failure from } \boldsymbol{\mathcal{R}}(Y_j))$$

$$= \frac{P(\text{individual } j \text{ fails} \mid \text{at risk at } Y_j)}{\sum_{l \in \mathcal{R}(X_j)} P(\text{individual } l \text{ fails} \mid \text{at risk at } Y_j)}$$

$$= \frac{\lambda(Y_j \mid \boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \lambda(Y_j \mid \boldsymbol{X}_l)}$$

$$= \frac{\lambda(Y_j \mid \boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \lambda(Y_j \mid \boldsymbol{X}_l)}$$

$$= \frac{\lambda_0(Y_j)\exp(\boldsymbol{\beta}'\boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \lambda_0(Y_j)\exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}$$

$$= \frac{\exp(\boldsymbol{\beta}'\boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}$$

So the likelihood of individual j dying at time t is determined by their observed risk (determined by their particular combination of predictor variable $\boldsymbol{\beta}'\boldsymbol{X}_j$) relative to the total

risk of everyone who is a risk at time t. Notice that on the fifth line, the baseline hazard $\lambda_0(Y_j)$ was cancelled out. This means that the final equation is a partial likelihood for the death of individual j as the constant baseline hazard component of the hazard has been removed.

The is to fit the model which works best for all individuals in the data set and so the $\boldsymbol{\beta}$ which simultaneously maximises the likelihood of all n individuals in the data set needs to be found. The partial likelihood function for all deaths is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}$$

The task is to identify the values of $\boldsymbol{\beta}$ which maximise the likelihood of obtaining the observed survival times $\boldsymbol{Y}$ given the values of the predictor variables $\boldsymbol{X}$**.** That is, identify the values of $\boldsymbol{\beta}$ which maximise the total likelihood $L(\boldsymbol{\beta})$.

Selecting $\boldsymbol{\beta}$ so as to maximise $L(\boldsymbol{\beta})$ is difficult but some standard apporaches have been developed. Firstly, it has been recognised that the $\boldsymbol{\beta}$ values which maximise the likelihood function $L(\boldsymbol{\beta})$ will also maximise the logarithm of the likelihood function $l(\boldsymbol{\beta})$ and vice-versa.

The fitted Cox model is the model where the observed survival times have the highest probability of occurring. That is, the $\boldsymbol{\beta}$ vector which maximises the likelihood or maximises the logarithm (log) of the likelihood function needs to be identified as this will occur at the same location ($\boldsymbol{\beta}$ coefficients**)**. The advantage of log-likelihood is that it adds the partial likelihoods across individuals rather than multiplies them.

$$l(\boldsymbol{\beta}) = \log\left[\prod_{j=1}^{n} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{X}_j)}{\sum_{l \in \mathcal{R}(Y_j)} \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}\right]$$

$$= \sum_{j=1}^{n}\left[\log(\boldsymbol{\beta}'\boldsymbol{X}_j) - \log\left(\sum_{l \in \mathcal{R}(Y_j)} \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)\right)\right]$$

$$= \sum_{j=1}^{n} l_j(\boldsymbol{\beta})$$

To identify the maximum value of $l(\boldsymbol{\beta})$ calculus is used. At the maximum value, the slope of the line $l(\boldsymbol{\beta})$ in $\boldsymbol{\beta}$ changes direction and so the first derivative in $\boldsymbol{\beta}$ (which is the slope of the line) momentarily achieves a value of zero. So differentiating the log-partial likelihood and setting the results to equal zero, will allow the identification of the values of the $\boldsymbol{\beta}$ vector of coefficients which maximise the partial likelihood function and hence the fit of the Cox regression model.

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta} l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(X_i - \frac{\sum_{l \in \mathcal{R}(Y_j)} X_l \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}{\sum_{l \in \mathcal{R}(Y_j)} \exp(\boldsymbol{\beta}'\boldsymbol{X}_l)}\right)$$

$U(\boldsymbol{\beta})$, the first derivative of the log-likelihood function in $\boldsymbol{\beta}$, is called the score function and this is the level at which the 'corrected score method' for misclassification correction is applied.

The corrected score approach first separates the predictor variables in $\boldsymbol{X}$ into two subgroups: those affected by misclassification error $\boldsymbol{W}$ and those which are not affected $\boldsymbol{Z}$. In the current case, $\boldsymbol{W}$ would contain the single dichotomous health behaviour variable and $\boldsymbol{Z}$ would contain any other predictors in the model (such as age and/or cancer stage). For variables affected by misclassification $\boldsymbol{W}$, the estimated proportion correctly classified and the estimated proportion misclassified in each category (such as proportion of smokers

classified and smokers, non-smokers classified as smokers, non-smokers classified as smokers and non-smokers classified as non-smokers) is recorded in a matrix $\boldsymbol{A}$. The higher the misclassification rate, the greater the attenuation of results towards the null and so the greater the misclassification correction which needs to be applied. Therefore, the magnitude of misclassification correction is determined by $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

As shown in equations 6 to 11 in the source article (Zucker & Spiegelman, 2008), the misclassification corrected score equation splits the X_i and $\boldsymbol{X}_i$ in the $U(\boldsymbol{\beta})$ equation above into misclassification affected $\boldsymbol{W}$ predictors weighted by $\boldsymbol{B}$ and the unaffected $\boldsymbol{Z}$ predictors.

The covariance matrix of the $\boldsymbol{B}$s could be obtained from the second derivative of the log-likelihood equation, but in this thesis empirical confidence intervals are used and the calculation standard errors is ignored.

6.3 Imputing pre-diagnosis health behaviour in cancer registry data and investigating its relationship with oesophageal cancer survival time

The material in this section is currently under review in the peer reviewed journal *PlosOne*.

Following the CRediT Taxonomy (National Information Standards Organization, 2021), author contributions were:

- Paul Fahey contributed to Conceptualisation, Data curation, Formal analysis, and Writing the original draft.
- Paul Fahey, Glenn Stone and Andrew Page contributed to Methodology.
- Andrew Page, Glenn Stone and Thomas Astell-Burt contributed to Supervision

- All authors contributed to Review and Editing of the draft paper.

The analyses in this Chapter uses the same method of imputation and same method of estimation of misclassification rates as in Chapter 5. Indeed Figures 6.1 and 6.2 are copies of Figures 5.6 and 5.7 updated to include an additional year of data and Table 6.1 is an updated version of Table 5.4. All the remaining material is new.

6.3.1 **Abstract**

Background: As oesophageal cancer has short survival, it is likely pre-diagnosis health behaviours will have carry-over effects on post-diagnosis survival times. Cancer registry data sets do not usually contain pre-diagnosis health behaviours and so need to be augmented with data from external health surveys. This paper introduces and tests a new algorithm to augment cancer registries with external data when one-to-one data linkage is not available.

Methods: The algorithm is to use external health survey data to impute pre-diagnosis health behaviour for cancer patients, estimate misclassification errors in these imputed values and then fit misclassification corrected Cox regression to quantify the association between pre-diagnosis health behaviour and post-diagnosis survival.

Results: The algorithm worked effectively on simulated smoking data when there is no age confounding. But age confounding does exist (risk of death increases with age and most health behaviours change with age) and interferes with the performance of the algorithm. The estimate of the hazard ratio of pre-diagnosis smoking was 1.32 (95% CI 0.82,2.68) with 1.93 (95% CI 1.08,7.07) in the squamous cell sub-group and pre-diagnosis

physical activity was protective of survival with a hazard ratio of 0.25 (95% CI 0.03, 0.81). But the method failed for less common behaviours (such as heavy drinking).

Conclusions: Further improvements in the I2C2 algorithm will permit enrichment of cancer registry data through imputation of new variables with negligible risk to patient confidentiality, opening new research opportunities in cancer epidemiology.

6.3.2 **Introduction**

Oesophageal cancer is an important cancer. Worldwide it is estimated that it accounts for 3.1% of all cancers and 5.5% of all cancer deaths (Sung et al., 2021). But it is still a relatively rare disease. In the US for example, the estimated risk of being diagnosed with oesophageal cancer before age 75 is just 1.15% in males and 0.44% in females (Sung et al., 2021). There were an estimated 184,400 new cases (4.3 per 100,000 population) in the US in 2020 with an estimated 5-year relative survival of just under 20% (National Cancer Institute, 2021a) .

Given the relatively short survival time, it is likely that pre-diagnosis health behaviour is an important contributor to post-diagnosis survival time. This relationship should be documented to help understand how current changes in health behaviour (such as decreasing smoking rates (Agaku, Odani, Okuyemi, & Armour, 2020), changes in alcohol consumption patterns (Azagba, Shan, Latham, & Manzione, 2020; Grucza et al., 2018), increasing leisure time physical activity (Morseth & Hopstock, 2020) and increasing obesity rate (Ward et al., 2019)) will impact on the future number of oesophageal cancer survivors and their associated health service needs. Also, at the patient level, clearer understanding of the effect of pre-

diagnosis health behaviour may assist in addressing the current weaknesses in prognostic indexes for oesophageal cancer (Gupta et al., 2018).

Cancer registries provide high quality, census data for cancers and often record patient outcomes such as survival. Unfortunately, cancer registries rarely contain pre-diagnosis behavioural risk factors. Retrospective data collection (contacting cases from the cancer registry and interviewing them about their pre-diagnosis behaviours) are subject to survival and recall biases and data collection costs. An alternative approach is to augment the cancer registry data through record linkage with an external data source. But locating routinely collected pre-diagnosis records of oesophageal cancer patients' health behaviour is challenging. While regular population-based surveys of health behaviour are conducted across a range of settings, the rarity of oesophageal cancer (4.3 people per 100,000 per year in the US (National Cancer Institute, 2021a)), means that very few survey respondents would have gone on to experience oesophageal cancer. Individually linked data sets will be small and, with a low proportion of correct links, and linkage errors may dominate.

An algorithm called the I2C2 (Impute, Impute, Calibrate, Correct) approach was recently described, to investigate the relationship between health behaviour and relative risk of surviving 12 months after diagnosis (Fahey, Page, Stone, & Astell-Burt, 2020a). This does not require cancer registry cases to be present in the health survey data set. It only requires survey participants to be demographically similar to those with data in the cancer registry. This not only avoids the small sample sizes for true matches, but also avoids issues of confidentiality and costs associated with data linkage.

The results of the previous application of this approach (Fahey et al., 2020a) displayed sufficient face-validity to extend the algorithm to the estimation of hazard ratios through Cox regression. Accordingly, the aim of this paper is to describe the relationship

between pre-diagnosis health behaviour and post-diagnosis cancer survival by augmenting a cancer registry data set with information from an external, no-cases-in-common, health behaviour data set.

6.3.3 **Methods**

The project was approved by the Western Sydney University Human Research Ethics Committee (H12305).

Data sources

Cancer registry data were obtained from the US Surveillance, Epidemiology, and End Results Program (SEER) cancer registries data base (https://seer.cancer.gov/data/). The SEER data base is compiled annually and collates data from US cancer registries with a combined coverage of approximately 28% of the US population (Doll, Rademaker, & Sosa, 2018). Data from all 39,233 cases of primary malignant oesophageal cancer for the 10 most recent years (2006-2015 at time of data access) were extracted. The 123 cancer cases aged less than 35 years were excluded as atypical.

Reference data on health behaviours, used to help impute health behaviour for the SEER cancer cases, were obtained from the Behavioural Risk Factor Surveillance System (BRFSS) (https://www.cdc.gov/brfss/data_documentation/index.htm). This is an annual telephone survey conducted in each US State and Territory, compiling behavioural data for more than 400,000 adult residents per year (Centers for Disease Control and Prevention, 2013b). To meet the criteria of 'pre-diagnosis', BRFSS data records were used from 5-years prior to cancer diagnosis. That is, the 3,469,905 BRFSS health survey data records from 2001 to 2010 were included. But the 2,515,009 data records from residents of US States outside

the SEER population catchments were excluded, leaving 954,896 records. (With an estimated incidence of 4.3 per 100,000 (National Cancer Institute, 2021a) just 41 of these 954,896 health survey respondents could be expected to be diagnoses with oesophageal cancer in any given year.)

Variables

The outcome variable was post-diagnosis all cause survival time measured in months. Individuals who survived beyond 2015 and those who were lost to follow-up had their censored survival time recorded at their last known survival date. The maximum possible follow-up time was 119 months. SEER cancer registry records with missing survival time (n=512) were excluded from the analysis.

Self-reported health behaviour variables were selected from the BRFSS health survey, and included:

- Current tobacco smoking (yes or no), defined as daily or less than daily smoking;
- Alcohol consumption – possible binge drinking (yes or no), defined as ≥5 standard drinks for males or ≥4 standard drinks for females on at least one occasion in the month prior to survey;
- Alcohol consumption – possible heavy drinking (yes or no), defined as >2 standard drinks per day for men and >1 standard drink per day for women in the month prior to survey;
- Physical activity (yes or no), defined as any physical activity or exercise in the past 30 days other than for regular job;
- Obese (yes/no), defined as body mass index ≥ 30 kg/m^2; and
- Current tobacco smoking with regular alcohol (yes or no), defined as current tobacco smoking with ≥1 standard drink of alcohol per day on average in the previous month.

The SEER cancer registry variables used to inform the imputation process were:

- Age category at diagnosis (5-year groups from 35-39y to 75-79y then >80y);
- Gender (male; female);
- Marital status (married, including common law; single or never married; widowed; divorced);
- Race (white; black; Asian or Pacific Islander; American Indian or Alaska Native);
- State of residence (Alaska; California; Connecticut; Georgia; Hawaii; Iowa; Kentucky; Louisiana; Michigan; New Jersey; New Mexico; Utah; Washington);
- Year of diagnosis (2006 to 2015).

The same variables were extracted from the BRFSS health survey data sets from 2001 to 2010. The BRFSS health survey records were from 5 years earlier than the SEER cancer cases so as to correspond to 'pre-diagnosis' behaviour. SEER cancer registry data records (n=2,344, 6.0%) and BRFSS health survey data records (n=18,770, 2.0%) with missing data on any of these variables were excluded from further analyses.

Sub-group analyses were conducted on adenocarcinoma and squamous cell carcinoma as different health behaviours may have different impacts on the two types of oesophageal cancer (Fahey et al., 2015).

The I2C2 algorithm

Conceptually, start with the view that the pre-diagnosis health behaviours are variables in the SEER cancer registry with 100% missing data. Then seek to address this missing data by imputation. As health behaviour is missing for all SEER cancer registry records, the imputation requires external data on health behaviour: in this case the BRFSS health survey data set.

In previous publications two different approaches have been described for imputing SEER cancer cases' behaviour from BRFSS health survey data via common demographic variables. One approach was to develop a logistic regression model predicting behaviour from demographic variables in the BRFSS health survey data and then apply that model to each SEER cancer case to estimate their probability of having the behaviour (Fahey, Page, Stone, & Astell-Burt, 2020b). The second approach was to stratify both data sets into age by sex by race by marital status by State of residence by year subgroups, and then within each strata randomly assign one BRFSS health survey data record to donate their behaviour to each SEER cancer registry data record (Fahey et al., 2020a). This latter approach is referred to as cold deck imputation (Nordholt, 1998), and is the method employed in the current study.

Some of the BRFSS health survey data records did not have complete data for all six of the health behaviour measures. To avoid imputing a missing value to replace the missing behaviour, six copies for the BRFSS data set were created, each containing complete data for one of the six behaviours. For each behaviour two BRFSS 'donor' records were randomly assigned, without replacement, to each SEER cancer registry case. Donor records had to be of the same sex, race, marital status and State of residence and had to be recorded 5 years earlier in time and be one 5-year age category younger than the SEER cancer registry case they were assigned to. As the required number of replications of the imputation process increases as the percentage of non-response gets larger (Nordholt, 1998), this cold deck imputation was repeated 100 times for each of the six health behaviours. SEER cancer registry cases with less than two eligible donor records were excluded from subsequent analyses on that behaviour. Figure 6.1 is a flow chart showing SEER cancer cases who were included and excluded from the analysis. Figure 6.2 is an equivalent flow chart for the BRFSS health survey data records.

Figure 6.1 Flow chart of inclusions and exclusions of SEER oesophageal cancer cases

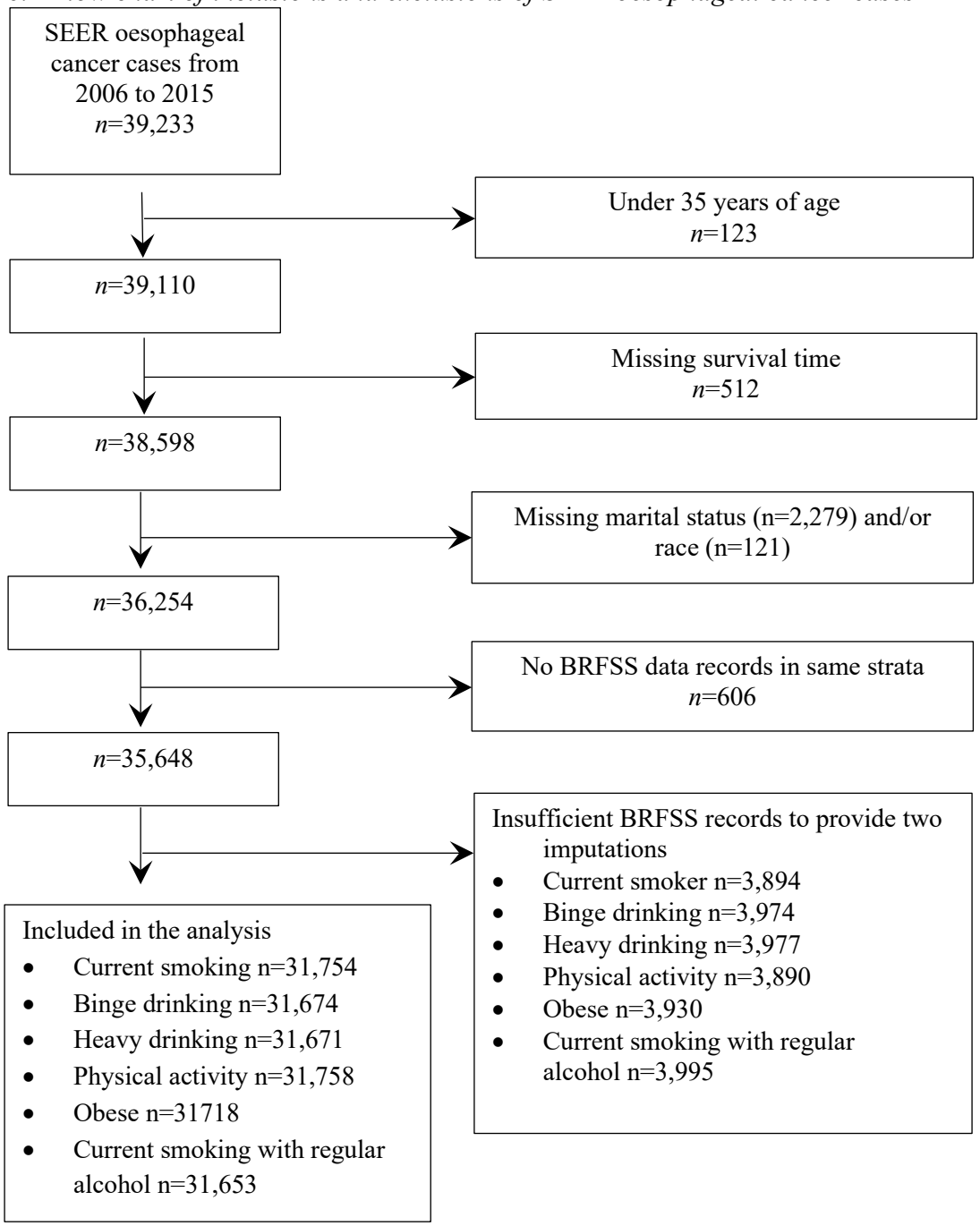
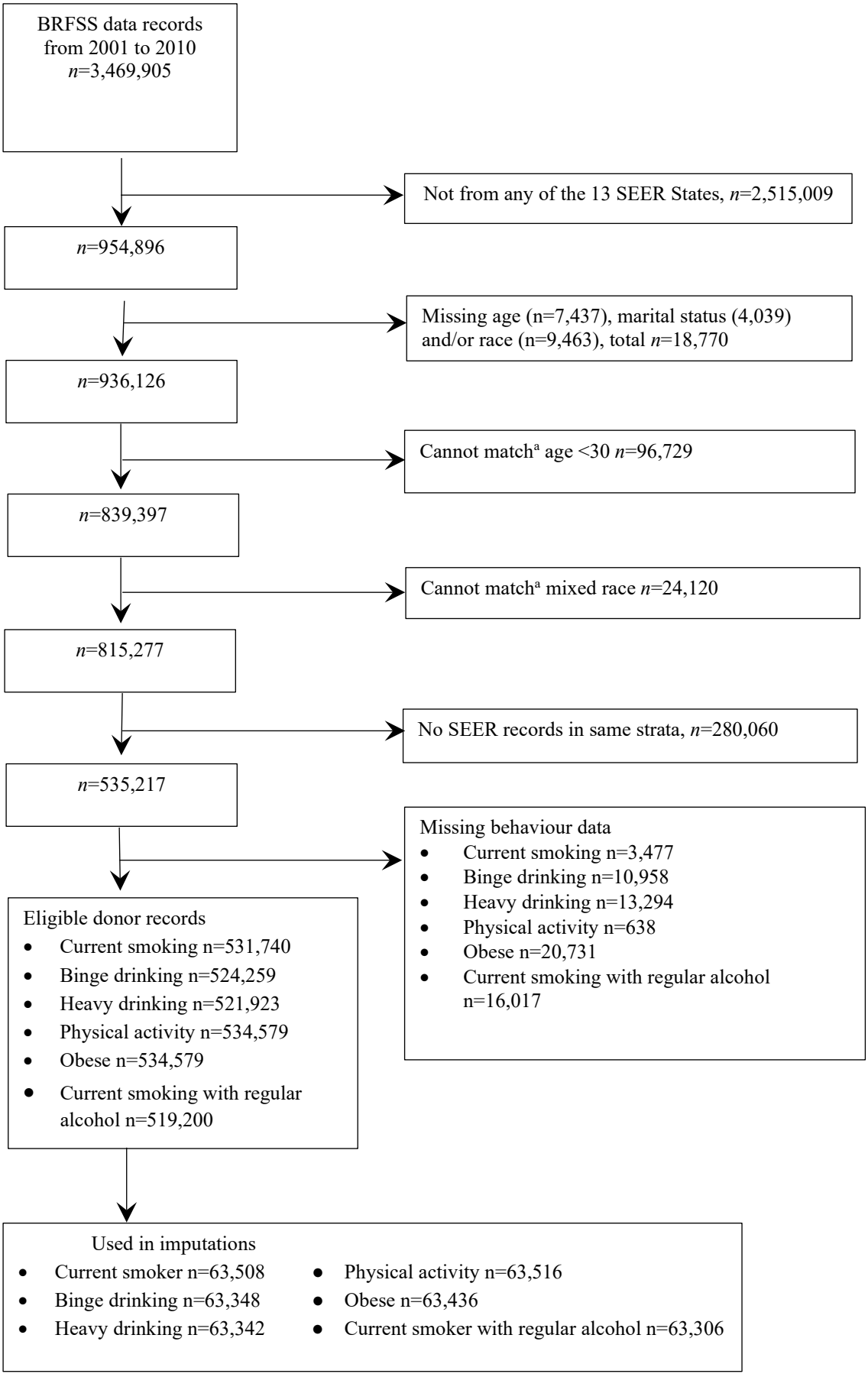


Figure 6.2 Flow chart of inclusions and exclusions of BRFSS health behaviour data records



[a] All auxiliary variables could be coded identically in both SEER cancer registry data and BRFSS health behaviour data except the BRFSS data included an additional category for race. Unlike the SEER cancer registry data, the BRFSS data collection allowed respondents to describe their race as "mixed". About 3.5% of BRFSS respondents selected this option. As these records did not match any SEER cancer registry records they did not contribute to the analysis. Similarly, with a minimum age of 35 years for SEER cancer cases and a 5-year lag, BRFSS respondents under 30 years of age could not match any SEER cancer registry records.

Table 6.1 describes the eligible SEER cancer registry cases and the proportion of these who were excluded due to unavailability of matching BRFSS health survey donor records for 'Current smoking'. Exclusions were higher in earlier years, older age groups, males, non-whites and Californians. The relatively large difference in matching success across follow-up status is a reflection of the poor matching success in the earlier years. For example, as those who are still alive are censored, 46.5% of censored observations occur in the final two study years when failure to match rates were 2.8% and 2.7%. Very similar results would be expected for the other five behaviours.

Table 6.1 Number of SEER oesophageal cancer cases seeking donor records for current smoking behaviour and the proportion of these failing to obtain two donor records.

	Seeking Donor Records		Failed to Obtain 2 Donor Records	
	Frequency	% of total	Frequency	% of group
Total	36,254	100.0%	4,500	12.4%
Cancer type				
ESCC	11,694	32.3%	1,616	13.8%
EAC	20,354	56.1%	2,282	11.2%
Year				
2006	3,462	9.5%	693	20.0%
2007	3,525	9.7%	661	18.8%
2008	3,553	9.8%	622	17.5%
2009	3,693	10.2%	731	19.8%
2010	3,579	9.9%	508	14.2%
2011	3,608	10.0%	426	11.8%
2012	3,653	10.1%	501	13.7%
2013	3,649	10.1%	150	4.1%
2014	3,716	10.2%	100	2.7%
2015	3,816	10.5%	108	2.8%

Table 6.1 (continued) Number of SEER oesophageal cancer cases seeking donor records for current smoking behaviour and the proportion of these failing to obtain two donor records.

	Seeking Donor Records		Failed to Obtain 2 Donor Records	
	Frequency	% of total	Frequency	% of group
Total	36,254	100.0%	4,500	12.4%
Age group				
35-39 years	206	0.6%	2	1.0%
40-44 years	535	1.5%	7	1.3%
45-49 years	1407	3.9%	28	2.0%
50-54 years	2684	7.4%	148	5.5%
55-59 years	4337	12.0%	385	8.9%
60-64 years	5553	15.3%	725	13.1%
65-69 years	5726	15.8%	828	14.5%
70-74 years	4888	13.5%	751	15.4%
75-79 years	4467	12.3%	745	16.7%
80+ years	6451	17.8%	881	13.7%
Sex				
Male	28274	78.0%	4190	14.8%
Female	7980	22.0%	310	3.9%
Race				
White	30619	84.5%	3164	10.3%
Black	3783	10.4%	727	19.2%
Asian or Pacific Islander	1640	4.5%	515	31.4%
American Indian or Alaska Native	212	0.6%	94	44.3%
Marital status				
Married (incl common law)	20621	56.9%	2275	11.0%
Divorced	4584	12.6%	301	6.6%
Widowed	4860	13.4%	532	10.9%
Single (Never married)	6189	17.1%	1392	22.5%
State of residence				
Alaska[a]	48	0.1%	48	100.0%
California[b]	13209	36.4%	3625	27.4%
Connecticut	2024	5.6%	38	1.9%
Georgia	4046	11.2%	355	8.8%
Hawaii	533	1.5%	55	10.3%
Iowa	1946	5.4%	38	2.0%
Kentucky	2384	6.6%	52	2.2%
Louisiana	2209	6.1%	86	3.9%
Michigan	2129	5.9%	70	3.3%
New Jersey	4087	11.3%	87	2.1%
New Mexico	738	2.0%	11	1.5%
Utah	676	1.9%	10	1.5%
Washington	2225	6.1%	27	1.2%
Follow-up status				
Censored	8705	24%	608	7.0%
Died	27549	76%	3892	14.1%

[a] This is a small cancer registry specific to Alaska Natives.

[b] Given the large population of California the number oesophageal cancer cases was higher, but the number of BRFSS health surveys is constant for each State.

Any imputed data will contain errors, and imputation informed by demographic characteristics alone will contain many errors. However, it is known that people from similar demographic groups have a higher likelihood of having similar behaviour than people from different demographic groups (Boersma, Villarroel, & Vahratian, 2020; Centers for Disease Control and Prevention, 2021b; Leventhal, Bello, Galstyan, Higgins, & Barrington-Trimis, 2019; National Center for Health Statistics, 2021c). If the behaviour of the SEER cancer case is donated by a demographically similar individual, then it should have a slightly higher likelihood of being correct than if it were obtained from a completely random donor. Given the selection of donors is random, the resulting misclassification errors will also be random. In statistics, random error is generally controlled through sample size. The more random error, the larger the sample size required to confidently detect the remaining signal amongst the random noise.

The requirement for a large data set may be the limiting factor when the disease is rare. However, the more variables in common and more informative those variables (stronger their relationship with the behaviour) the stronger the information signal (Schneeweiss et al., 2009) and the more likely the analyses will detect it.

The effect of misclassification is to attenuate the results of the analysis towards the null (no effect) (Jurek et al., 2005), but if the misclassification can be measures, it is possible to statistically correct this attenuation.

In the current study misclassification was estimated by imputing the behaviour twice and quantifying the disagreement between these two imputed values. The misclassification between the two imputed values is an estimate of the misclassification between the ‘true’ behaviour and the ‘imputed’ behaviour. That is, the agreement between behaviour between

two individuals from the same demographic strata in BRFSS health survey was used as an estimate of the agreement in behaviour between an individual in the SEER cancer registry and an individual in the BRFSS health survey from the same demographic strata.

Of the 5 methods for misclassification correction for Cox regression models reviewed by Bang et al (Bang et al., 2013), the method used in the current study is corrected score estimation (as it is suitable for a dichotomous predictor, and external measure of the misclassification error) (Zucker & Spiegelman, 2008). For each of the 100 repetitions of the imputations in the data set, the misclassification rate was separately calculated and the misclassification corrected Cox regression model was fitted with corrected score estimation.

Simulation method

As the true hazard ratios (HRs) are unknown, it was not possible to evaluate the effectiveness of the I2C2 algorithm using the real data alone. To test the algorithm, 100 copies of the SEER cancer registry data were simulated, each containing both a 'true' and an 'imputed' smoking status.

Starting with the 100 repetitions of the SEER cancer registry data set with imputed smoking status, censored data were excluded, leaving 23,657 data records in each data set. The proportion of cancer cases in each age group was recorded as well as the average proportion of imputed smokers and non-smokers in each age category, and the average agreement between the two smoking categories in each age category across the 100 data sets. Finally, an appropriate Weibull model was identified to describe the distribution of survival times with the aid of the `fitdistrplus()` package in R software.

Next 100 simulated data sets were created, each of which contained

- A code number of the data set
- Age group
- A smoking status representing the ‘true’ smoking status
- A smoking status representing the ‘imputed’ smoking status
- 14 simulated survival times

In each simulated data set, there were 23,657 data records. These were randomly assigned to age groups and ‘true’ and ‘imputed’ smoking status groups such that the proportion of cases in each age group, the proportion of smokers in each age category and the misclassification between the ‘true’ and ‘imputed’ smoking status were all approximately the same as in the real dataset. The simulated survival times were produced by the method of Bender et al (Bender, Augustin, & Blettner, 2005) for simulating Weibull survival times with specified HRs, and then rounded to the nearest integer (months). The first seven survival times for each data record correspond to HRs of 0.50, 0.67, 0.80, 1.00, 1.25, 1.50 and 2.00 given ‘true’ smoking as the only predictor. The second set of seven HRs are calculated with both smoking and age category as predictors of survival status, with the HR for age set at the value in the SEER cancer registry data and the HRs for ‘true’ smoking at HRs of 0.50, 0.67, 0.80, 1.00, 1.25, 1.50 and 2.00 as above.

The first set of seven survival times are not confounded with age (as age is omitted in the development of the HRs) but the second set of survival times have the same level of confounding by age as the real data (as the proportion of smokers differed by age and both smoking status and age are predictors of survival).

Statistical analysis

Each of the six health behaviours were recorded as dichotomous variables. Each of the 100 repetitions of the data set for each health behaviour contained two records of that health behaviour: two imputed values for the SEER cancer registry data and a designated 'true' and a designated 'imputed' measure in the simulations.

Each statistic was estimated in each of the 100 repetitions of the data set for each health behaviour. Results are presented as the median of these 100 observations and the corresponding 95% empirical confidence interval (CI); the 2.5 and 97.5 percentiles.

To summarise how much information had been retained through the imputation process, the level of agreement beyond chance between the two imputed values (or between the true and imputed value in the simulations) is reported. Both Cohen's Kappa and the direct calculation of the difference in the observed and expected number of people recorded as having the behaviour on both imputations from the cross-tabulation of the two imputed values (or true against imputed values in the simulated data sets) are reported.

Analyses of survival time were conducted using Cox models using the Breslow method for addressing ties. For the simulated data the proportionality assumption was tested using the z-test on Schoenfeld residuals against transformed time. (Given the ideal distributions of the simulation, the median p-value was about 0.5.) Correction for misclassification was applied using the corrected scores estimation method. Results were presented as the median of the estimated HRs with associated 95% empirical CIs.

All analyses were conducted in R v4.0.2. The corrected score estimation software was provided by its creator, Prof David Zucker, as a Fortran 77 program which was called from within R using the foreign function interface (dyn.load() command). It was not possible to

test the proportionality assumption when using corrected scores estimation and analyses were restricted to the Breslow method for addressing ties.

Information passed to the corrected score calculator included the data set and two misclassification rates were passed:

- the proportion positive on the first imputation (or true behaviour) who were coded as negative on the second imputation; and
- the proportion negative on the first imputation (or true behaviour) who were coded as positive on the second imputation; and

Often the imputation process provided insufficient information for the corrected scores estimation to be calculated. While there are a number of ways to check for insufficient information, in this study a pragmatic criterion was employed such that whenever the estimated HRs tended towards positive or negative infinity the model had failed (e.g. Figure 6.3). The number of failures in the 100 data sets are reported in bar charts. Where more than 5 out of 100 data sets return estimated HRs of <0.01 or greater than 100, that analysis was labelled as failed and did not report the results. Where one to five data sets return extreme HRs, these extreme results were omitted from medians and 95% CIs.

Figure 6.3 An example of Cox regression coefficient tending towards negative infinity.

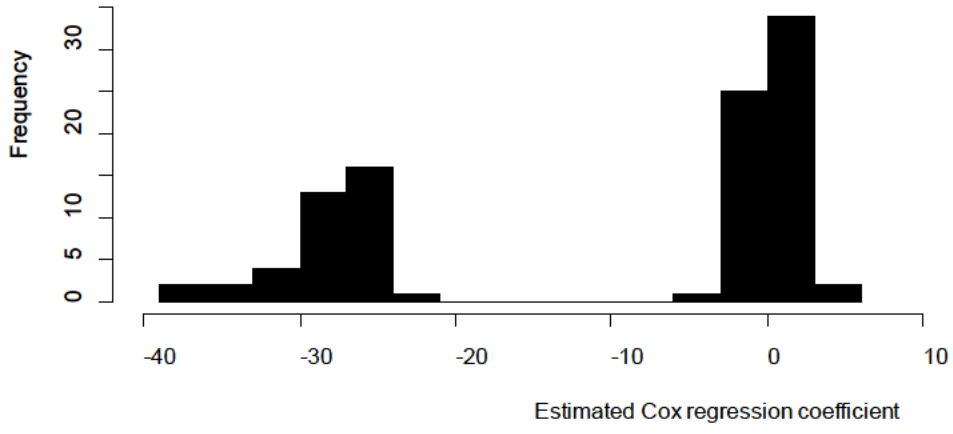


Figure 6.3 shows the estimated Cox regression coefficients for the behaviour ‘Current tobacco smoking with regular alcohol’ after age-standardisation. As more than 5 of the 100 data sets have returned extreme estimates of the coefficients, these results are rejected.

6.3.4 **Results**

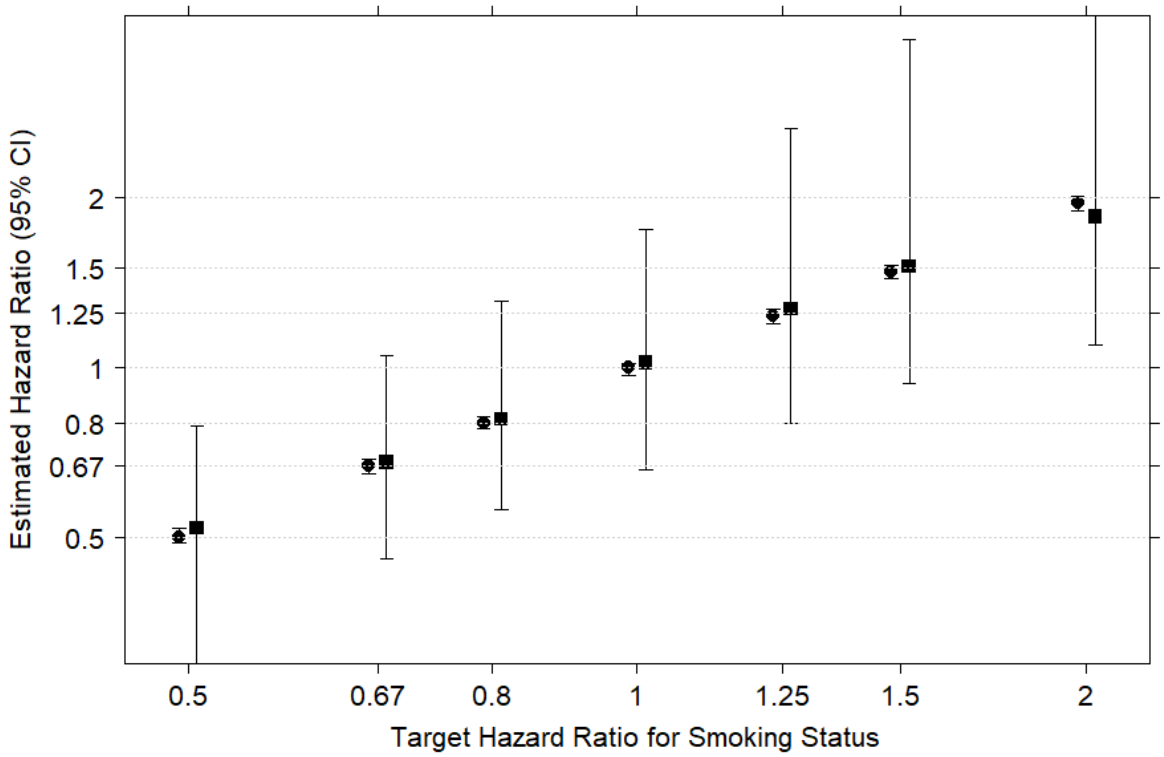
Agreement beyond chance was highest for imputed ‘smoking status’ with a median Kappa of 0.07 and a median of 299 more smoker-to-smoker matches (and 299 more non-smoker to non-smoker matches) observed than expected through chance agreement alone (Table 6.2). This agreement beyond chance scaled to the sample size is estimated to be 0.009. The next highest rates of information transfer were observed for ‘Physical activity’ followed by ‘Obesity’ and ‘Binge drinking’. The two least common behaviours – ‘Heavy drinking’ and ‘Smoking with regular alcohol’ with median prevalence of 4.8% and 3.3% respectively – had virtually no information retained through the imputation process.

Table 6.2 Selected statistics describing the agreement between the two donor records across the 100 repetitions of the SEER cancer registry data for each of the six behaviours.

	Sample size	Proportion with behaviour Median (95% CI)	Kappa (95% CI) Median (95% CI)	Agreement beyond chance (number) Median (95% CI)	Agreement beyond chance (rate) Median (95% CI)
Smoking status	31,754	15.7 (15.7,15.9)	0.07 (0.06,0.08)	299 (246,353)	0.009 (0.008,0.011)
Binge drinking	31,674	10.0 (9.9,10.2)	0.06 (0.04,0.07)	167 (124,203)	0.005 (0.004,0.006)
Heavy drinking	31,671	4.8 (4.7,5.0)	0.01 (<0.01,0.02)	16 (2,33)	<0.001 (<0.001,0.001)
Physical activity	31,758	73.8 (73.6,74.1)	0.03 (0.02,0.04)	214 (145,273)	0.007 (0.005,0.009)
Obesity	31,718	26.1 (25.8,26.3)	0.03 (0.02,0.04)	194 (130,261)	0.006 (0.004,0.008)
Smoking with regular alcohol	31,653	3.3 (3.2,3.4)	0.02 (0.01,0.04)	22 (10,39)	<0.001 (<0.001,0.001)
Simulated smoking status	23,657	15.8 (15.5,16.2)	0.07 (0.06,0.09)	236 (194,279)	0.010 (0.008,0.012)

Figure 6.4 shows the effectiveness of the I2C2 algorithm when applied to simulated data with smoking status as sole predictor of survival time in Cox regression. The median HR obtained from the I2C2 estimation conforms quite well with the median of the true HRs but, as expected, the CIs arising from the I2C2 algorithm are much wider. The wider CIs reflect the additional random error arising from the misclassification errors in the imputation.

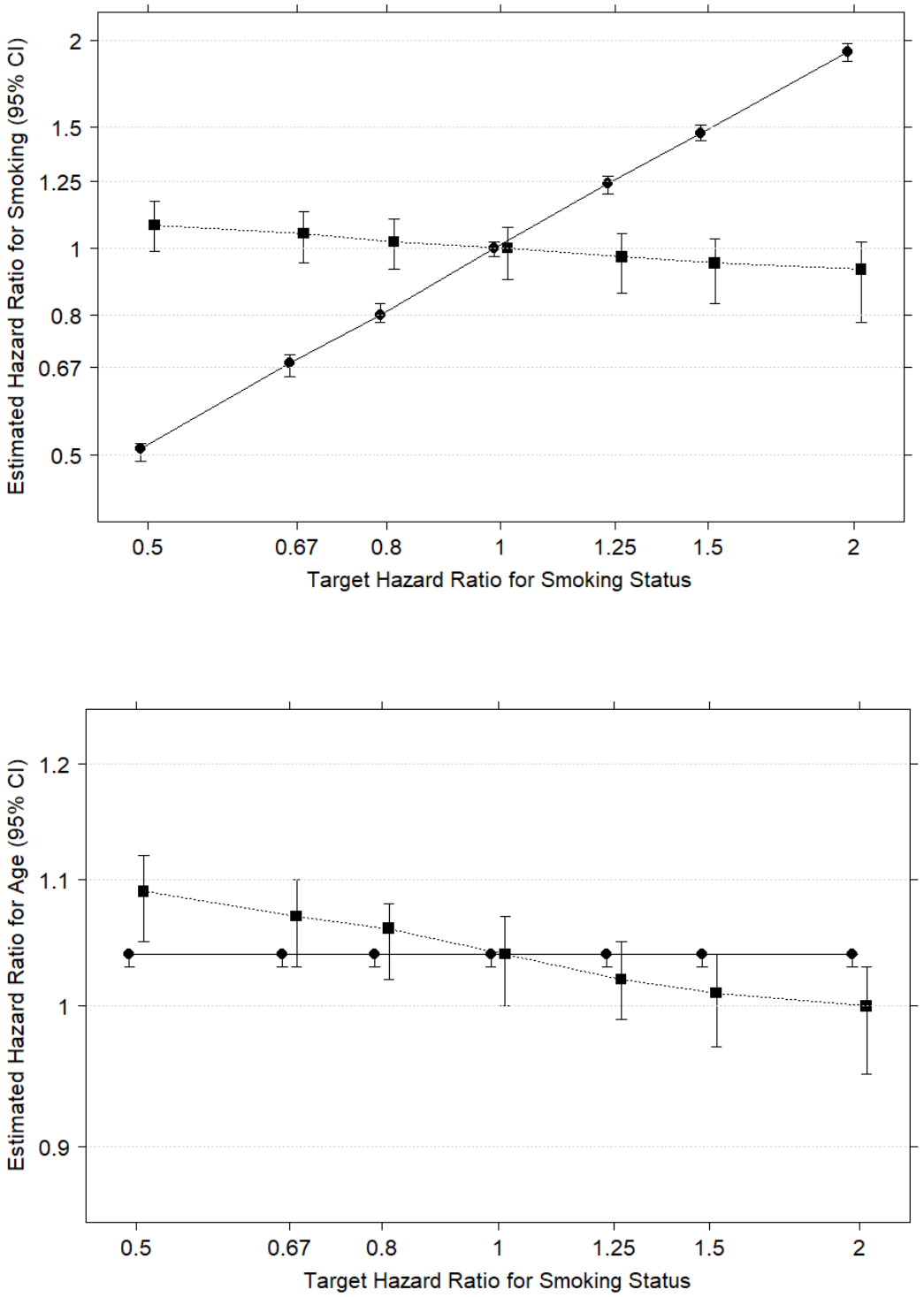
Figure 6.4 Results from 100 simulated data sets each with the same sample size and proportion of smokers as the SEER cancer registry data.



Each simulated data set contains seven different survival times associated with the target HRs on the horizontal axis. The vertical axis shows median and 95% empirical CIs of estimated HRs arising from Cox models. The first CI in each pair is obtained from fitting designated 'true' smoking status as a predictor of survival using the standard Cox model. The second is obtained from fitting designated 'imputed' smoking status using Cox regression with corrected score estimation (the I2C2 algorithm).

Figure 6.5 shows that the I2C2 algorithm fails when the relationship between smoking status and survival time is confounded by age. The ‘effect’ of smoking status (which is subject to high levels of misclassification error) shown in the first graph is being incorrectly attributed to age (which is measured without misclassification error) shown in the second graph.

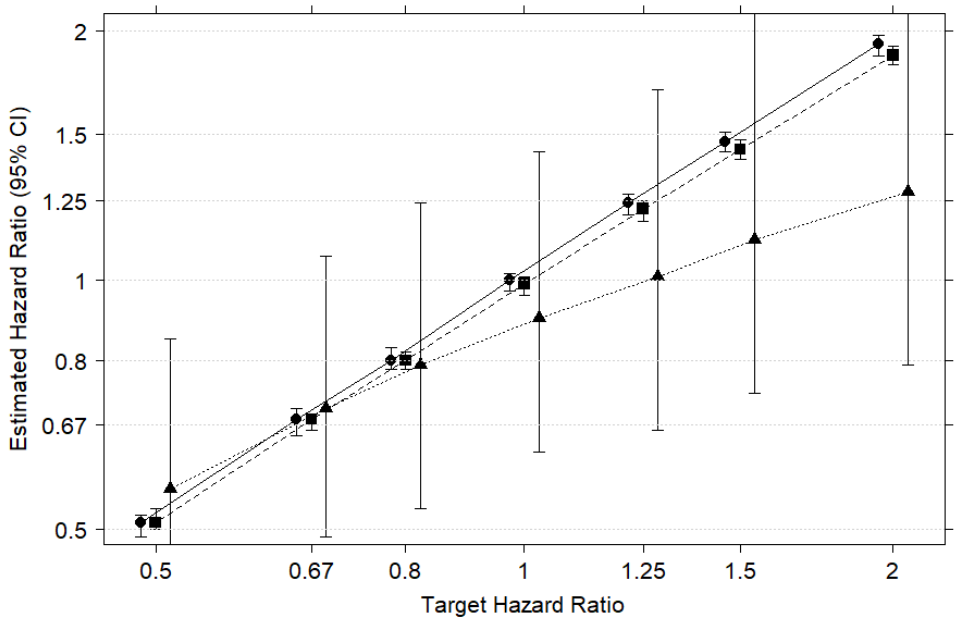
Figure 6.5 Results from 100 simulated data sets containing age confounding.



The first chart shows estimated HRs for smoking status and the second shows estimated HRs for 5-year age group. The CIs with circles joined by solid lines are obtained from fitting the designated 'true' smoking status and 5-year age group as predictors of survival using standard Cox regression. The CIs with squares joined by dotted lines are obtained from fitting the designated 'imputed' smoking status using, adjusted for 5-year age category, using the I2C2 algorithm.

Age confounding was addressed using both standardisation and stratification. All survival times were first standardised to the average age group. That is, the Cox regression was first fitted predicting survival time using age group alone. The observed survival times were then scaled by the predicted age effect (using the `predict()` option in R). Figure 6.6 shows that weight-based confounding adjustment works quite well for the designated 'true' smoking with little residual attenuation to the null. Unfortunately, this small attenuation becomes much larger when magnified by the high levels of misclassification error arising through the I2C2 algorithm. Figure 6.6 shows considerable residual attenuation of the median HR towards null effect (and, as expected, the much wider CIs for the estimated HRs arising from the misclassification errors) associated with 'imputed' smoking status after age standardisation.

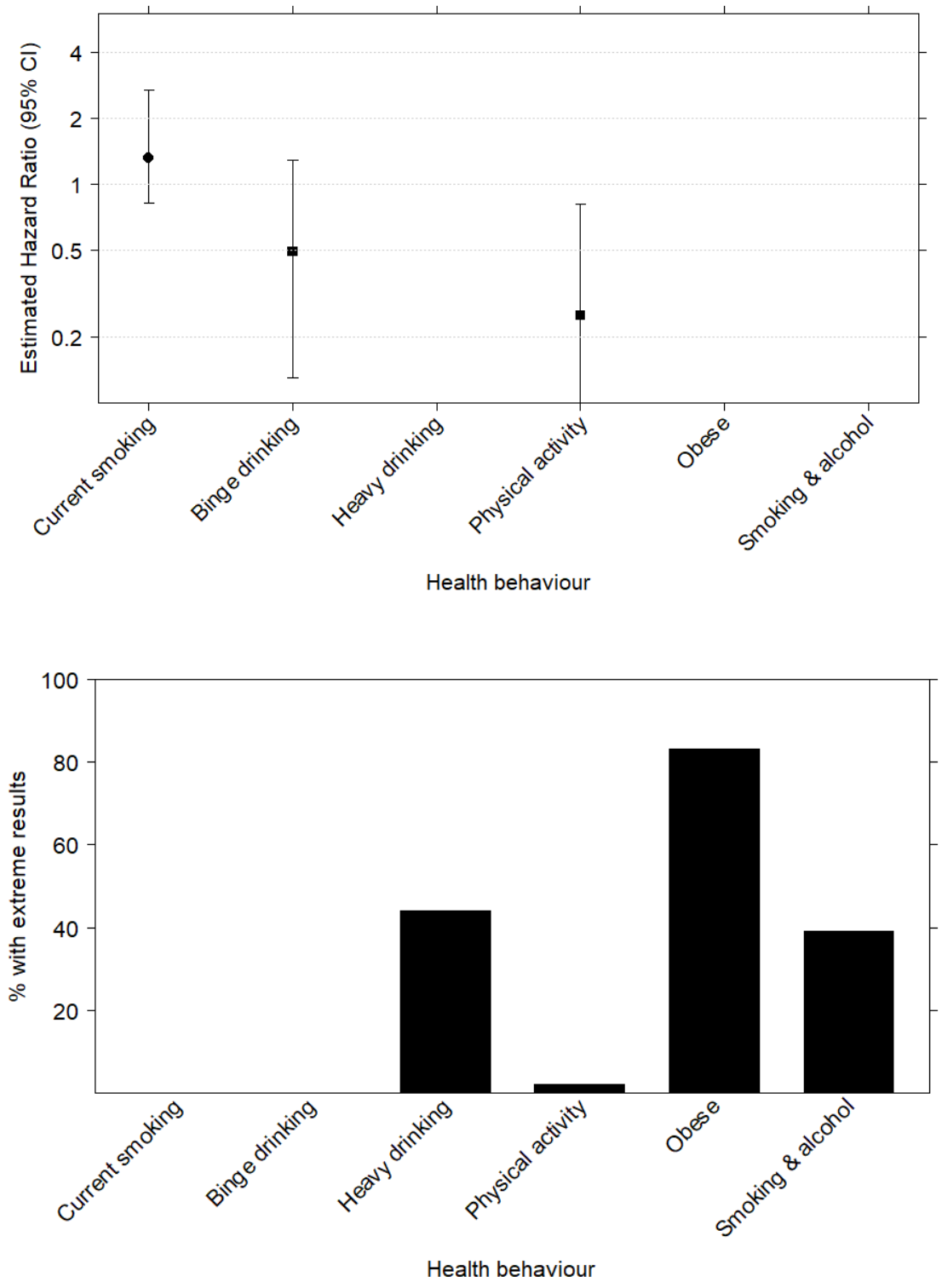
Figure 6.6 Estimates of the HR of smoking 5 years prior to diagnosis on post-diagnosis survival from simulated data sets.



Three CIs are presented at each target HR. The first (denoted by circles joined by a solid line) are the estimated HRs obtained from Cox regression where 'true' smoking status and age are predictors of survival time. The second CIs (squares connected by a dashed line) shows the estimated HRs for 'true' smoking status after age-standardisation of survival times. The third set of CIs (triangles connected by a dotted line) shows the estimated HR for imputed smoking status after age standardisation (I2C2).

Results from applying the I2C2 algorithm to the SEER cancer registry data with imputed behaviour are shown in in Figure 6.7. The bar chart shows that the algorithm has, as expected, often failed to converge on a HR estimate for ‘Heavy drinking’ and ‘Smoking with alcohol’ and has also failed for ‘Obesity’. Therefore, results for these behaviours have been suppressed. The second chart suggests that smoking 5 years prior to diagnosis may be a hazard to survival (HR 1.32, 95% CI 0.82,2.68) but binge drinking could be protective (HR=0.49, 95% CI 0.13,1.29). But both CIs include the null effect. Physical activity outside work is a statistically significantly protective of survival (HR=0.25, 95% CI 0.03,0.81).

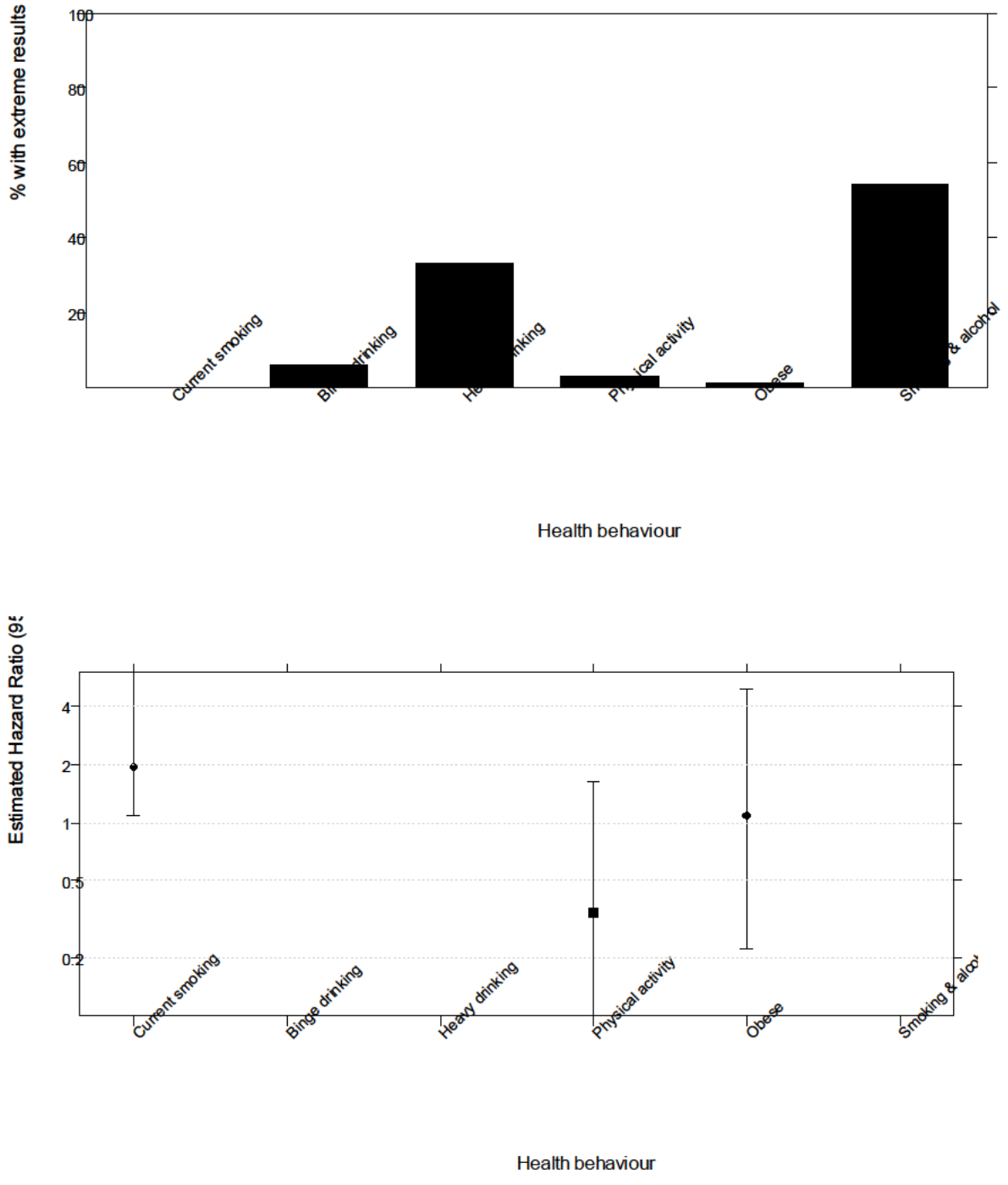
Figure 6.7 A summary of the results obtained after age-standardisation of survival times.



The bar chart shows the number of data sets where the algorithm has returned extreme values (<0.01 or >100) for estimated HRs. The second chart shows the median estimated HR and associated 95% empirical CIs for those behaviours which have recorded 5 or less extreme HRs.

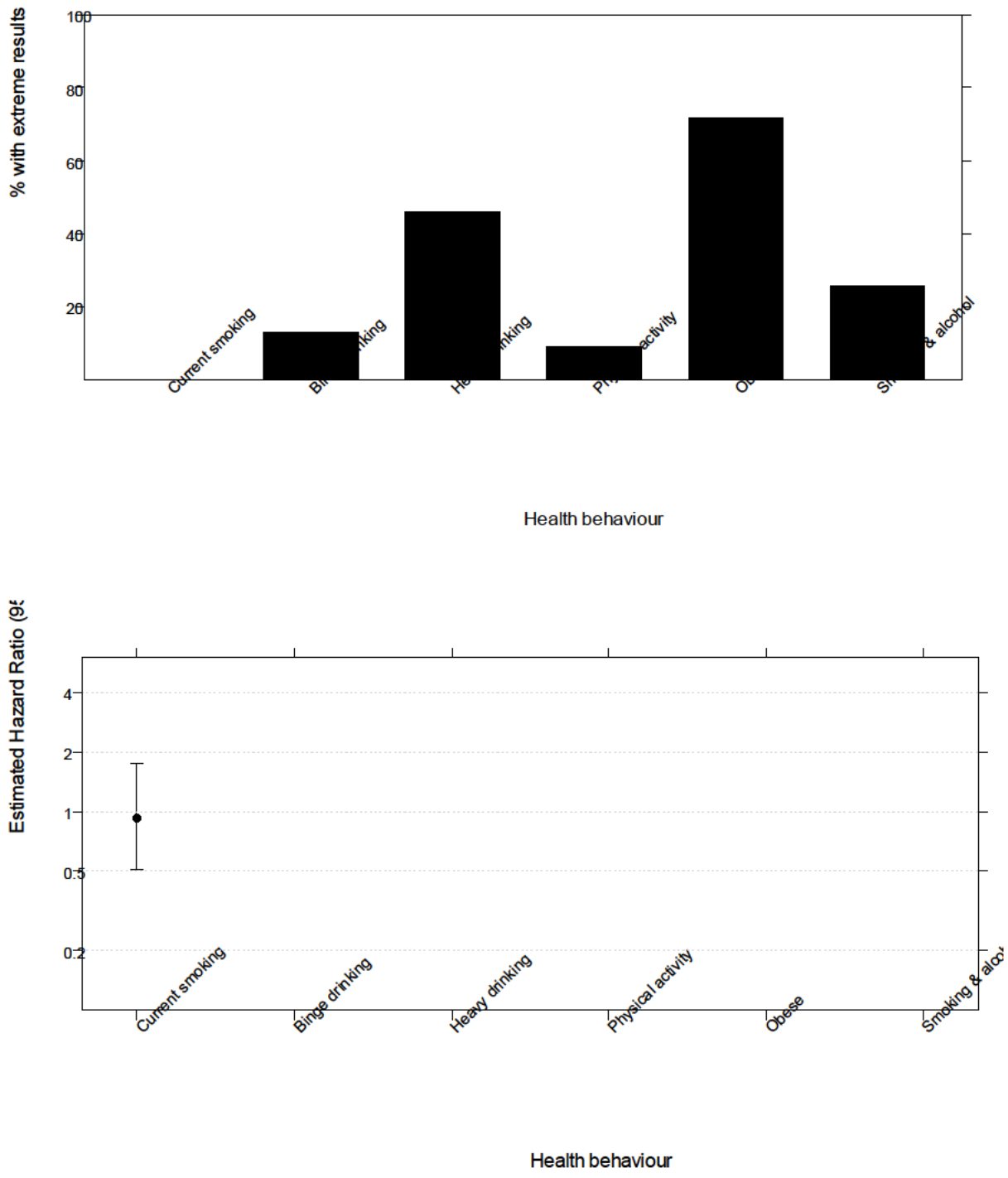
Equivalent analyses for the squamous cell carcinoma sub-group and the adenocarcinoma sub-group are presented in Figure 6.8 and Figure 6.9 respectively. For most of these analyses there was insufficient information passed through the imputation process to allow misclassification correction to converge. 'Smoking status' was associated with a statistically significantly increase in hazard in the squamous cell carcinoma subgroup (HR 1.93, 95%CI 1.08-7.07), but appeared to have no association with the adenocarcinomas (HR 0.92, 95%CI 0.51-1.74), 'Physical activity' may be protective in squamous cell carcinoma (HR 0.34, 95%CI 0.07-1.64), but the results were not statistically significant, and 'Obesity' appears to have no relationship with survival time in squamous cell carcinoma (HR 1.08, 95%CI 0.22-4.96). The simulations suggest these results are likely underestimates.

Figure 6.8 Age-standardised hazard ratios for simulated smoking in squamous cell subgroup.



The bar chart shows the number of data sets where the algorithm has returned extreme values (<0.01 or >100) for estimated hazard ratios. The second chart shows the median estimated hazard ratio and associated 95% empirical confidence intervals for those behaviours which have recorded 5 or less extreme hazard ratios out of 100.

Figure 6.9 Age-standardised hazard ratios for simulated smoking in adenocarcinoma subgroup.



The bar chart shows the number of data sets where the algorithm has returned extreme values (<0.01 or >100) for estimated hazard ratios. The second chart shows the median estimated hazard ratio and associated 95% empirical confidence intervals for those behaviours which have recorded 5 or less extreme hazard ratios out of 100.

Stratification was also trialled for adjusting for the confounding effect of age, but the strata sample sizes proved to be too small for the I2C2 algorithm to produce interpretable results. For the simulated data sets for smoking status, Figure 6.10 shows that the corrected scores Cox regression often failed to converge and Figure 6.11 shows that even where results are available, the median HR from the I2C2 algorithm is a poor point estimate of the 'true' HR.

Figure 6.10 The proportion of data sets failing to converge (HR<0.01 or HR>100) when using I2C2 with age-stratified simulated smoking.

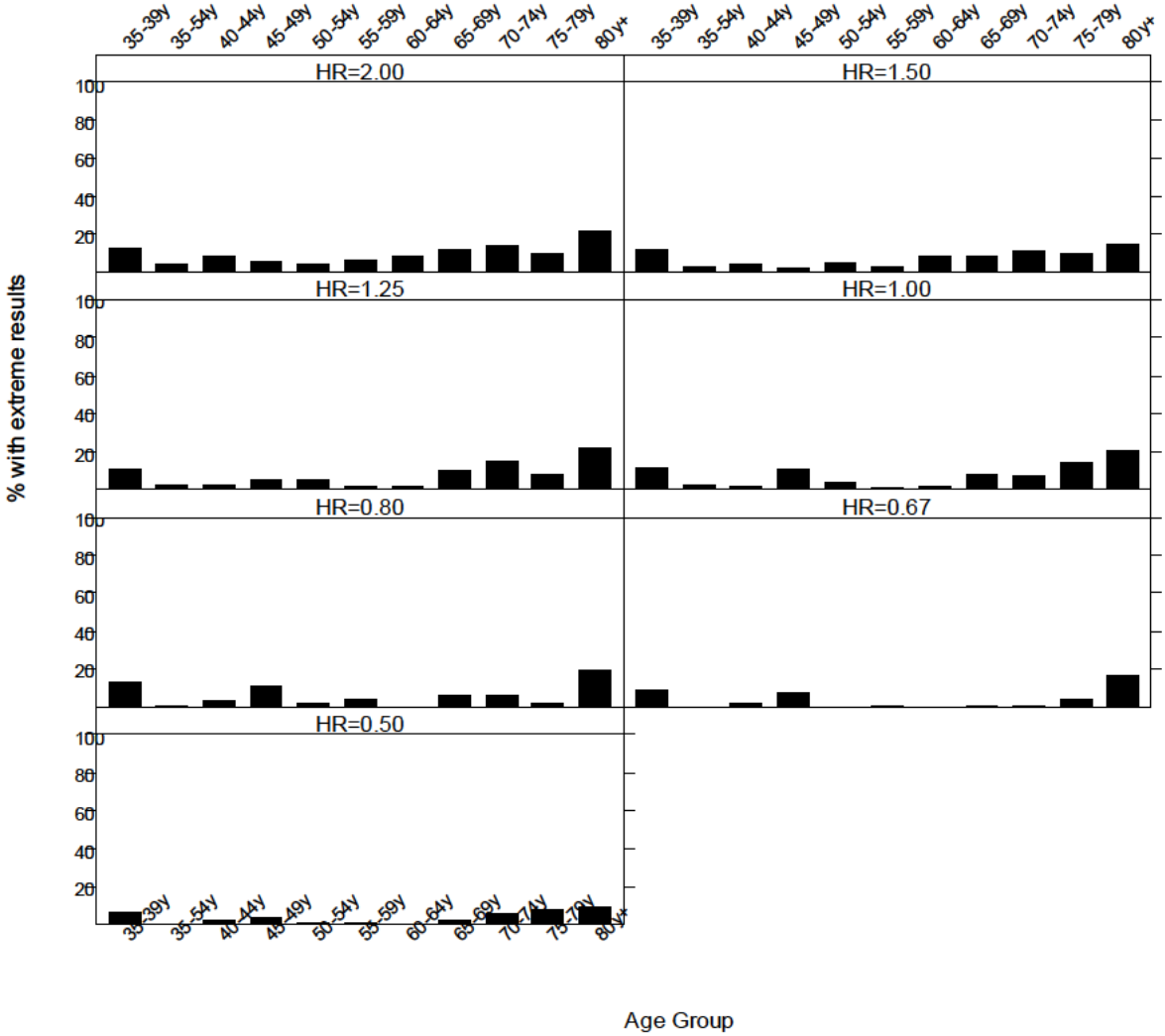
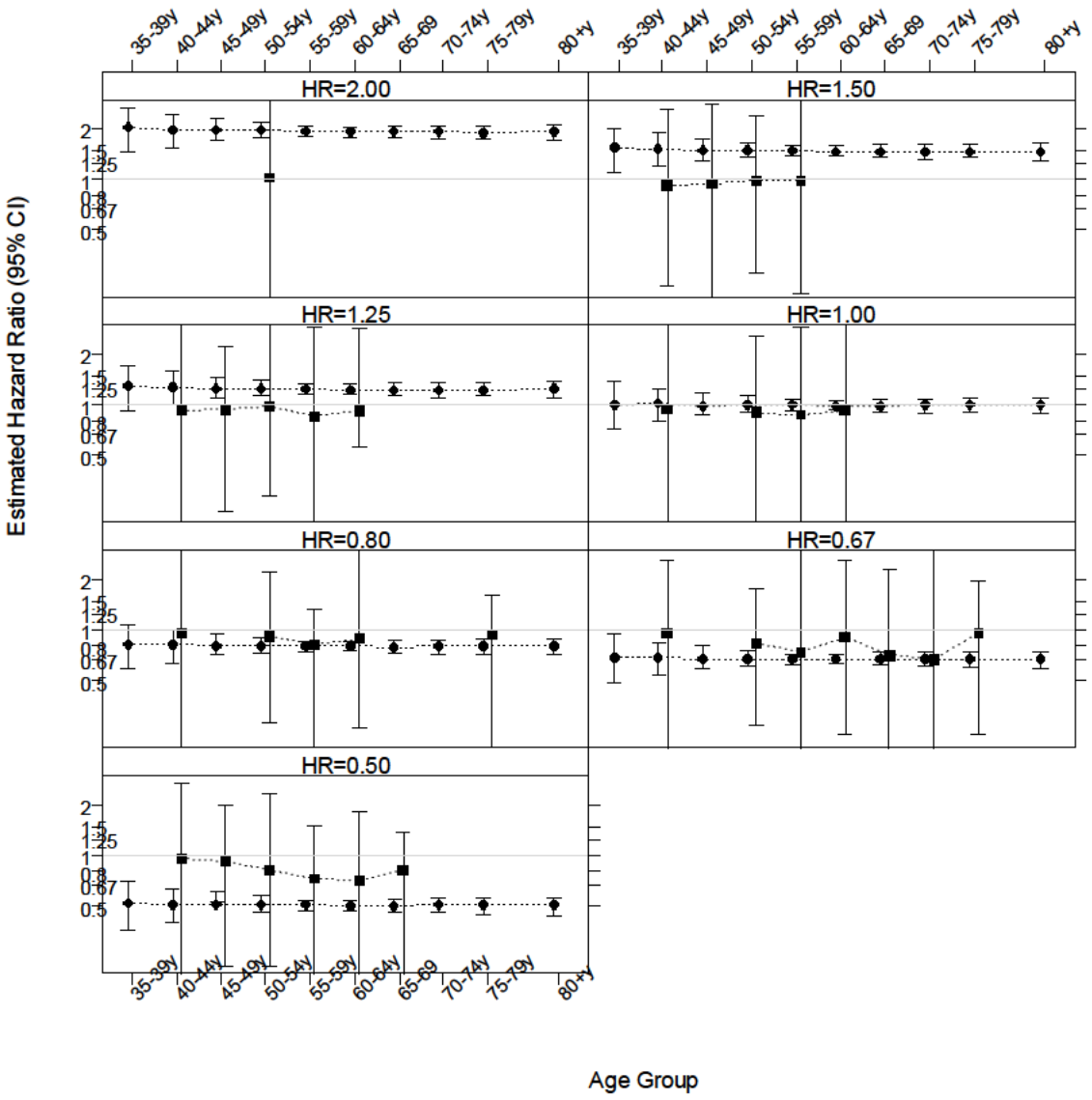


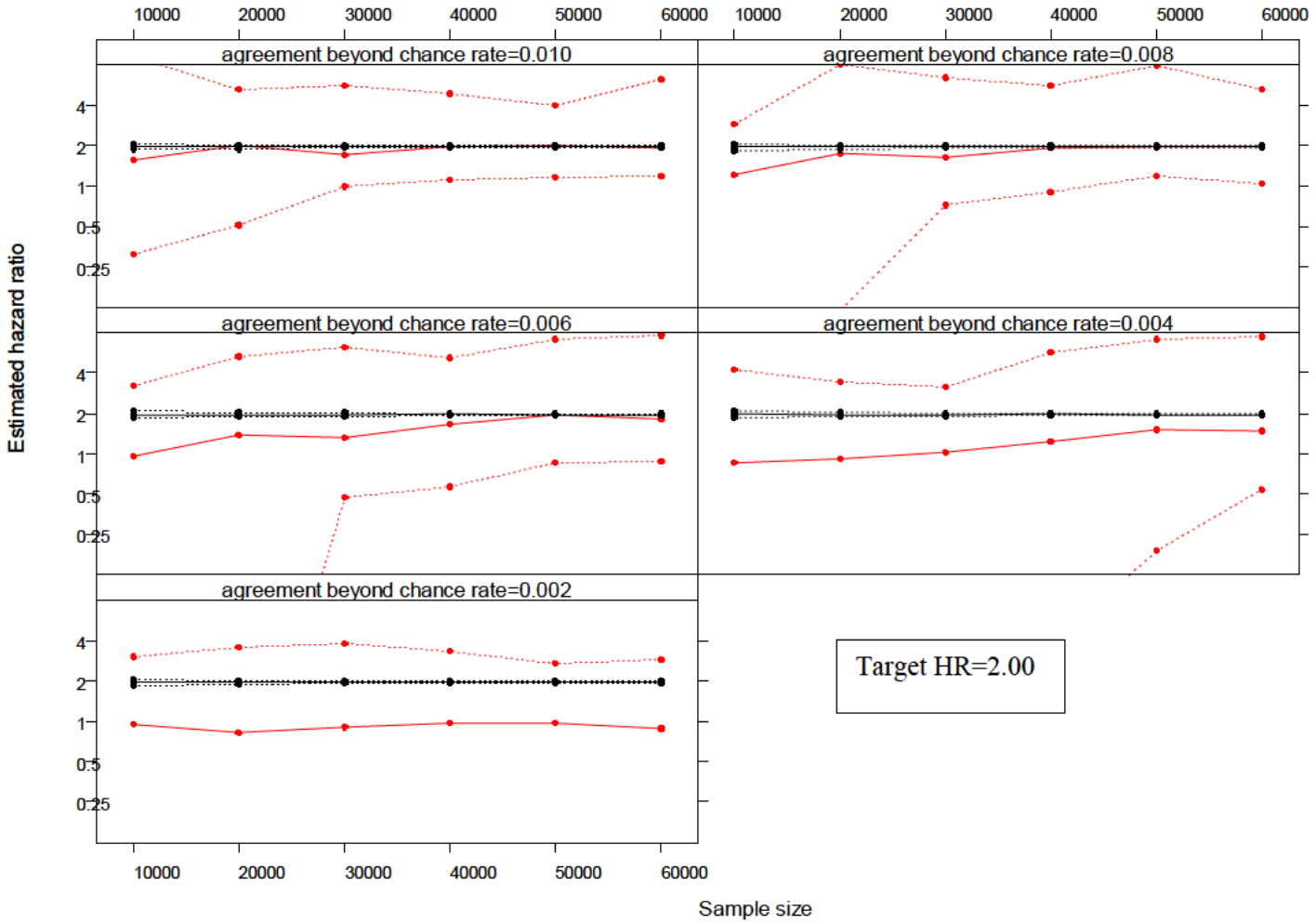
Figure 6.11 Age-stratified hazard ratios for simulated smoking in oesophageal cancer.



The first CI in each pair shows the HRs for 'true' smoking status and the second shows the results obtained using the imputed smoking status with misclassification correction. Where more than 5 in 100 data sets returned a HR<0.01 or HR>100, results are suppressed. The I2C2 algorithm fails to fully correct for the attenuation of results towards the null in any of the age categories.

Figure 6.12 shows the relationship between sample size, agreement beyond chance rate and estimated HRs, using the prevalence of smoking from the impute values across the SEER cancer registry data sets. It shows that with simulated survival times (Weibull, no censoring), the I2C2 algorithm gives accurate point estimate (albeit with wide uncertainty) when agreement beyond chance as a rate is 0.008 or more and sample sizes are 20,000 or more. For agreement beyond chance of 0.006 point estimates start becoming accurate from sample sizes of 40,000 or 50,000 but for lower signal strengths the algorithm seems to generally do poorly. Therefore, the available sample size of around 31,700 (including censored observations), stratified into 5-year age groups seems too small to support age stratified I2C2 analysis in this thesis.

Figure 6.12 Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

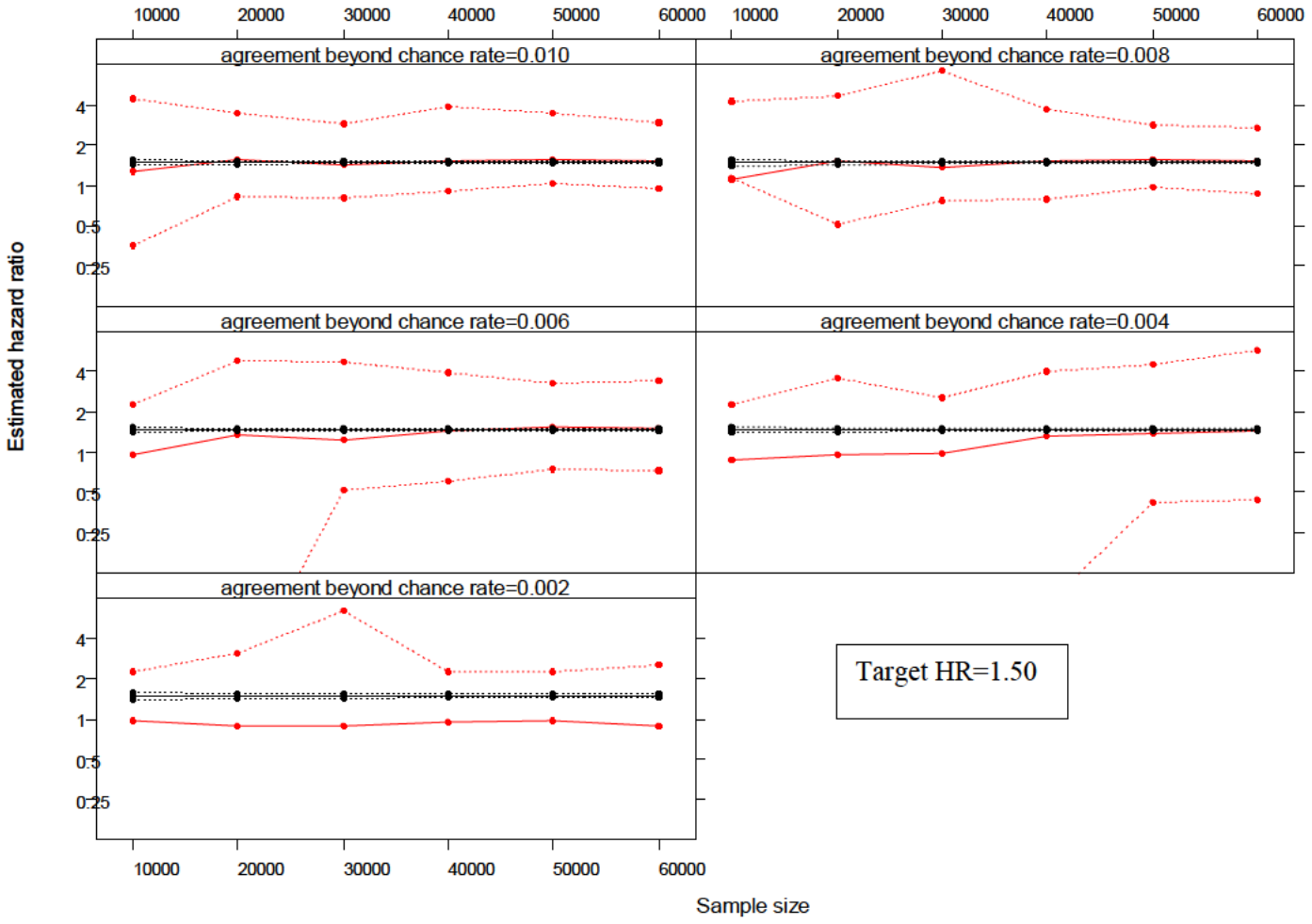


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

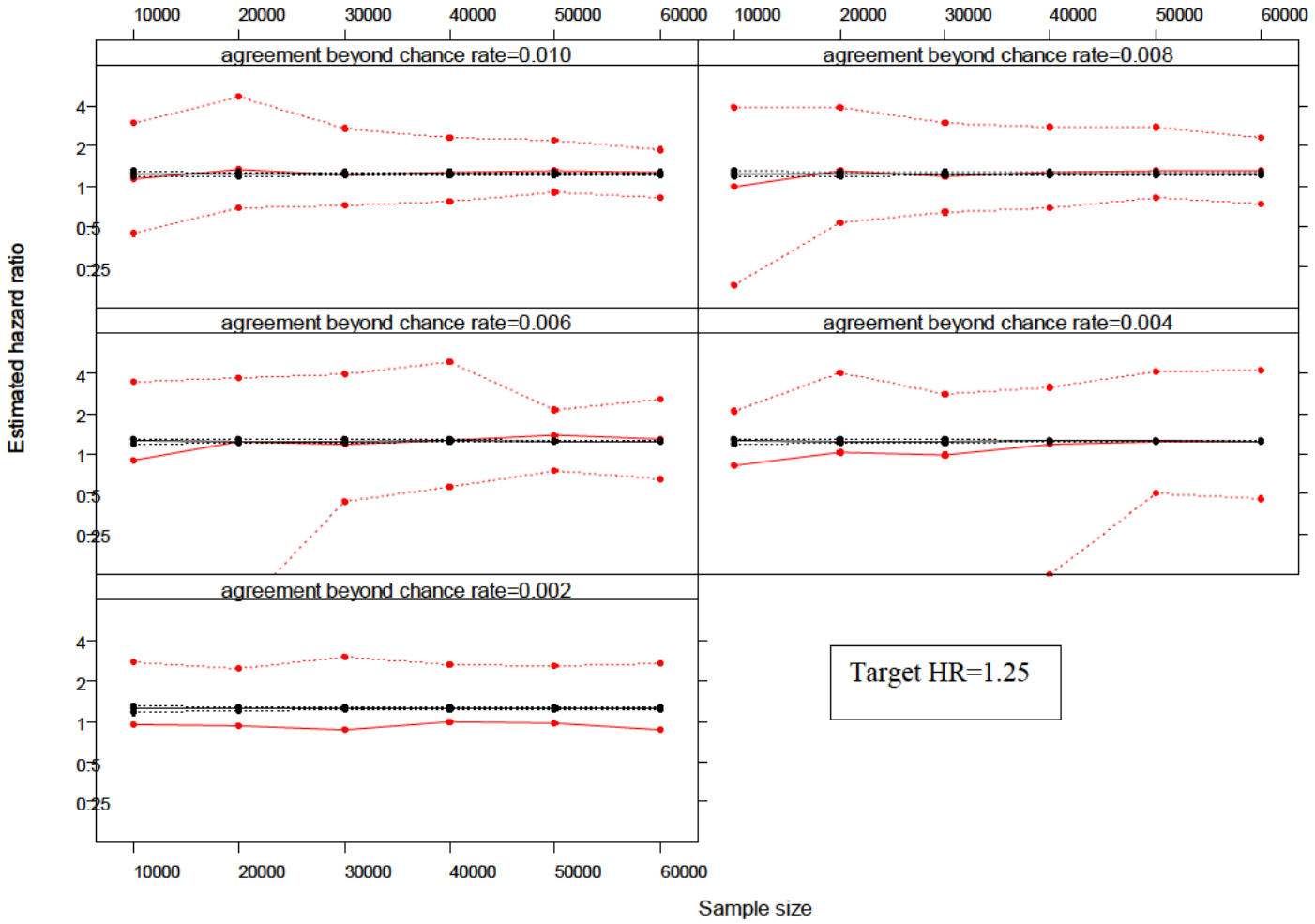


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

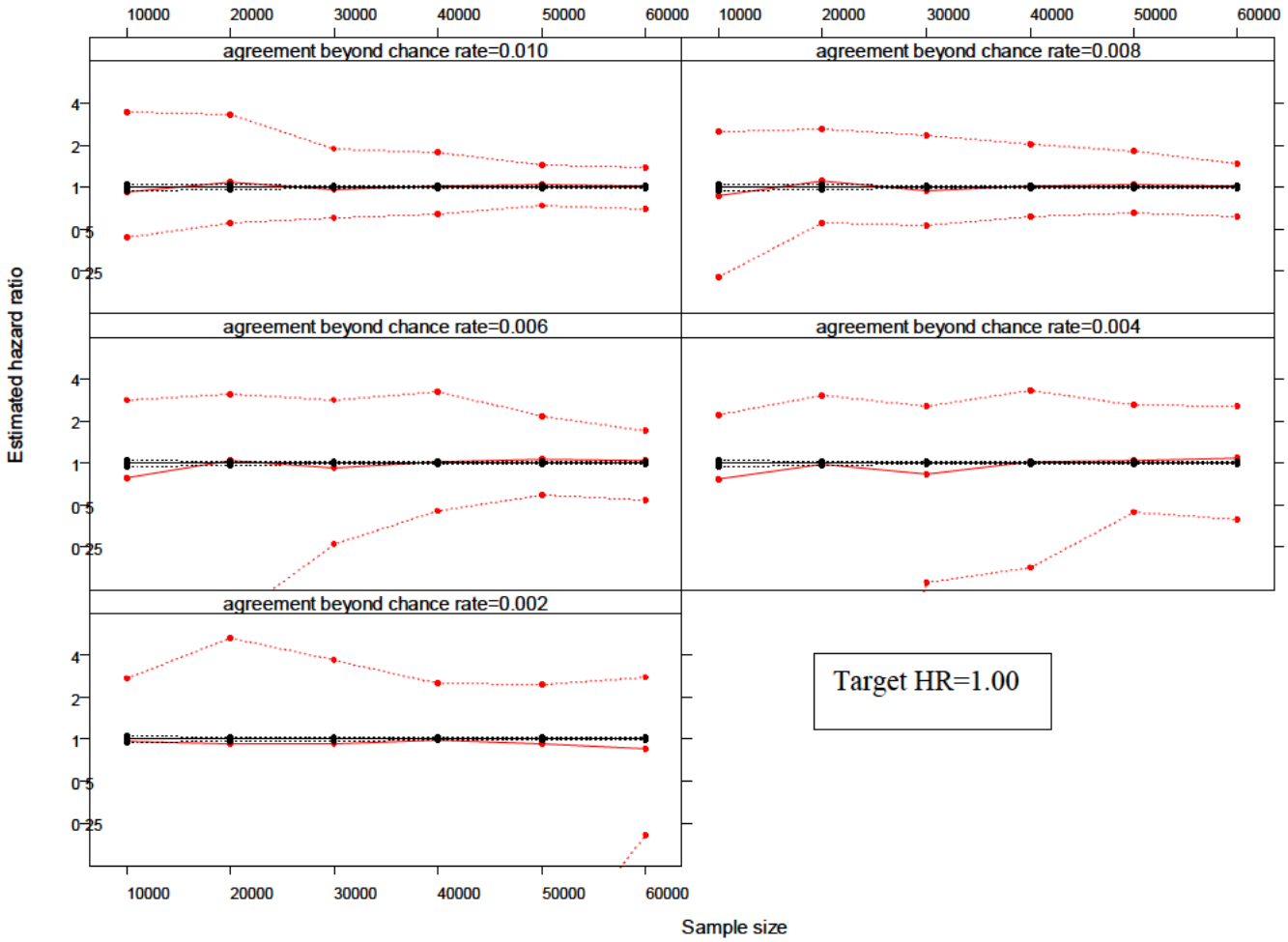


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

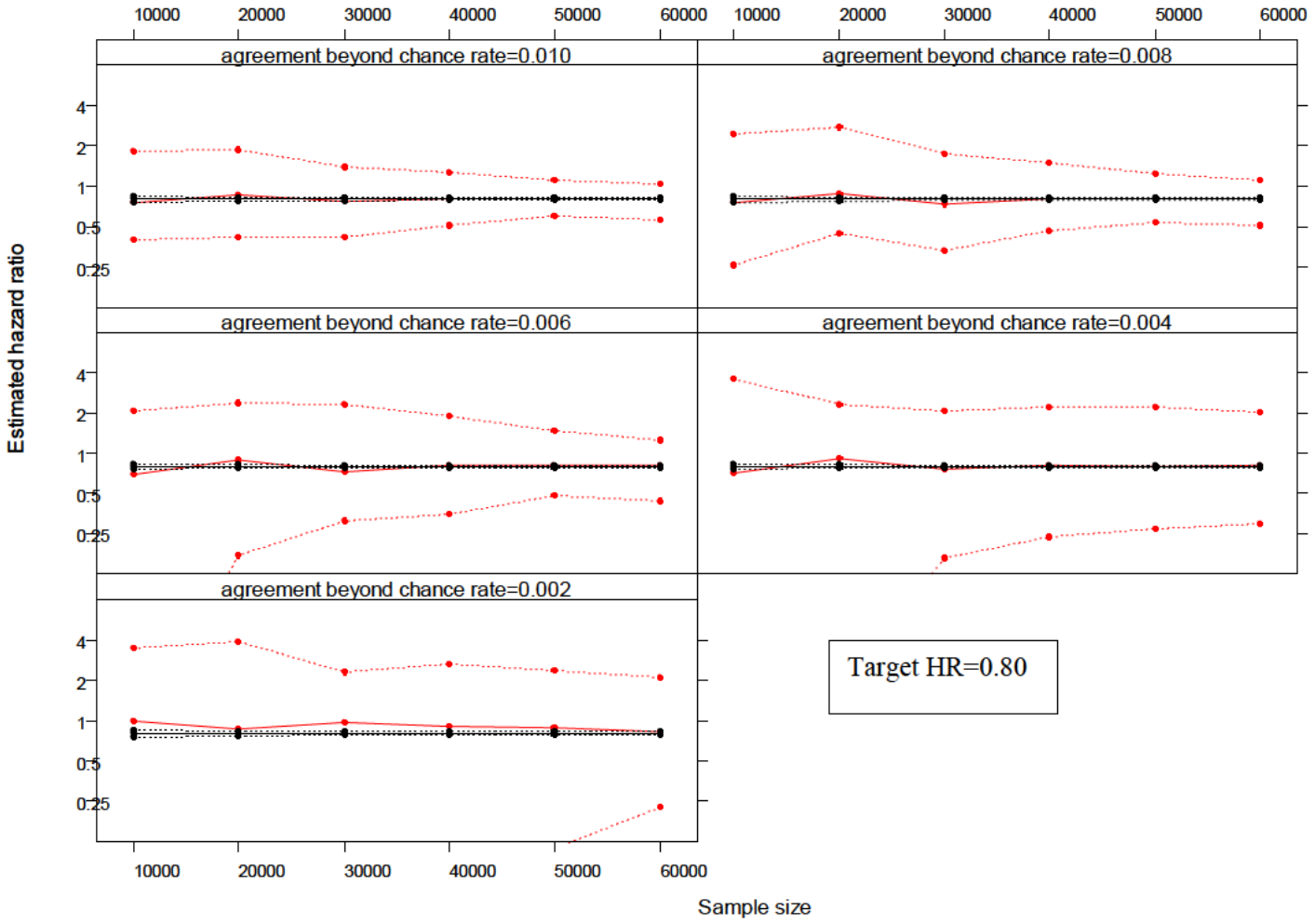


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

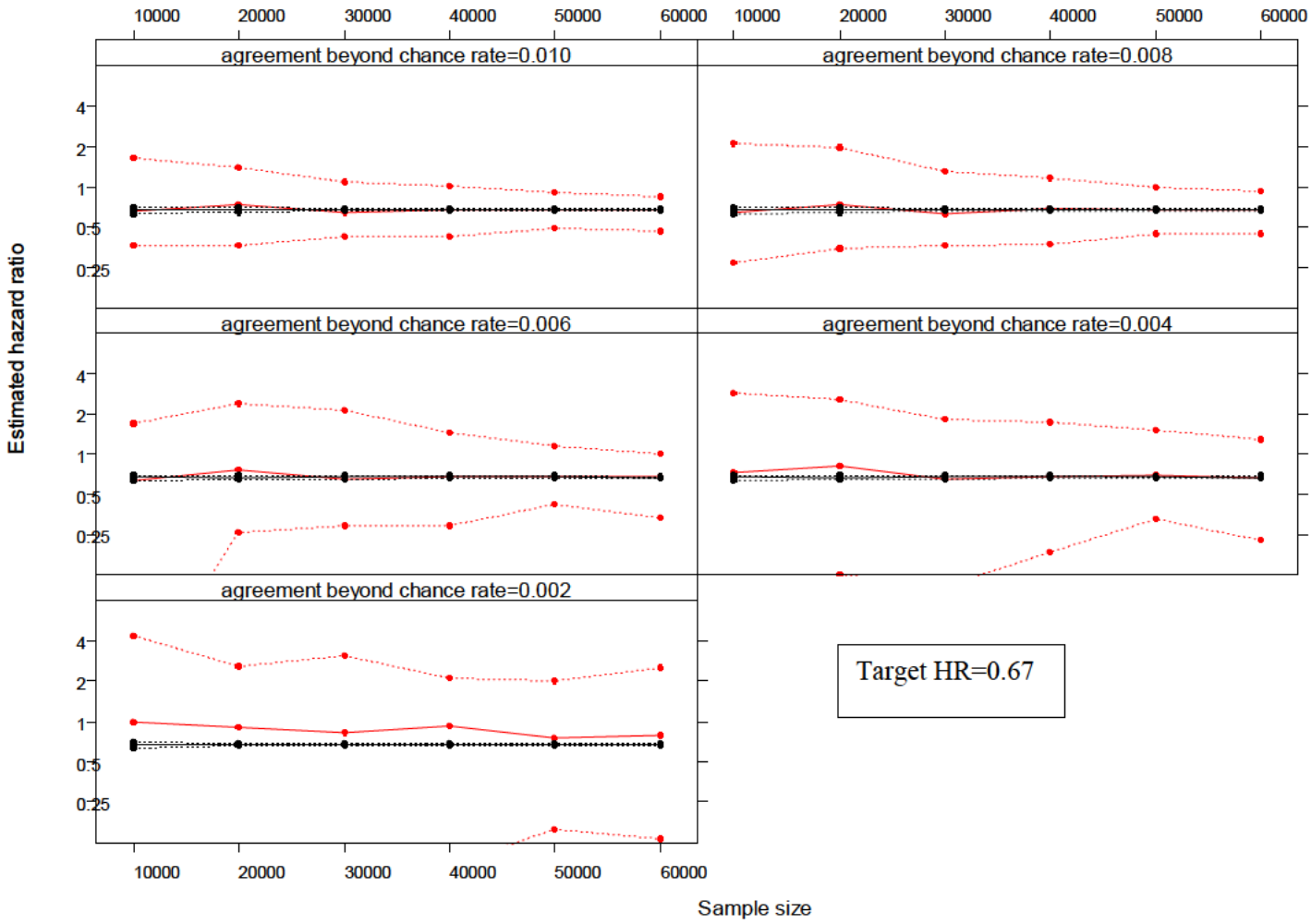


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.

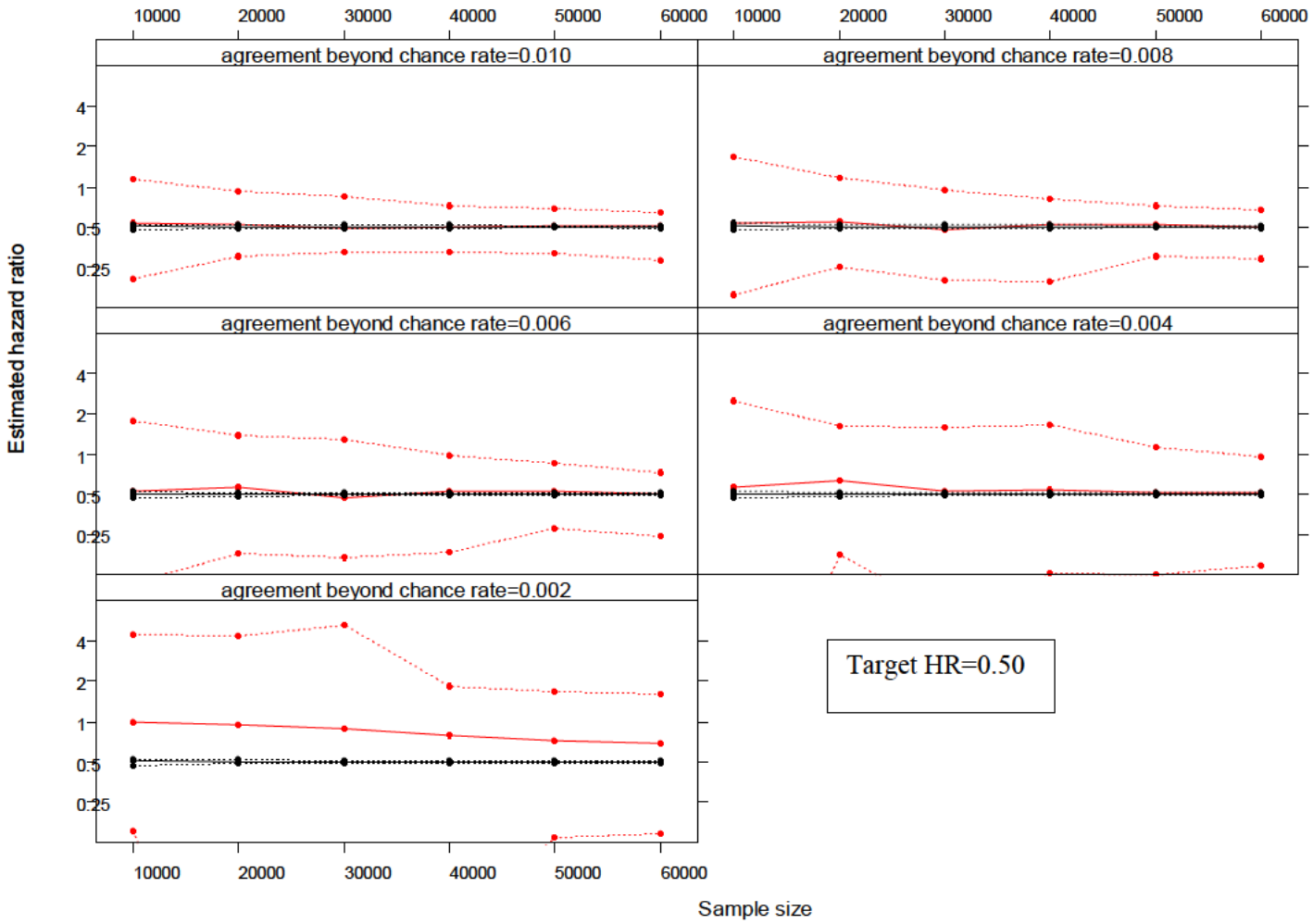


The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

Figure 6.12 (continued) Simulation of the relationship between sample size, agreement beyond chance rate and estimated hazard ratios for smoking status.



The black solid line connects median estimated hazard ratios for 'true' smoking status with the black dotted lines representing the associated 95% confidence intervals.

The red solid line connects median hazard ratios estimated from 'imputed' smoking status using the correct score estimation misclassification correction.

The red dotted lines connect the associated 95% confidence intervals.

6.3.5 **Discussion**

This study investigated the relationship between pre-diagnosis health behaviour and post-diagnosis oesophageal cancer survival by applying an imputation algorithm (the Impute, Impute, Calibrate, Correct, or I2C2, approach) to augment a cancer registry data set with information from an external, no-cases-in-common, health behaviour data set. Findings from the study showed that, using simulated data without age confounding, the I2C2 algorithm provides an accurate estimate of the median HR for the predictor variable, albeit with wider CIs arising from the additional random error arising from misclassification.

In practice, despite encouraging results, the I2C2 algorithm is vulnerable to a number of problems. Sample size and information retention through the imputation process are important issues. In the current example, the I2C2 algorithm failed to produce any estimates for behaviours with low prevalence such as 'Heavy drinking' and 'Current smoking with regular drinking' (both with an estimated prevalence of <5% among the cancer cases) and in age groups with low prevalence such as 'Smoking status' among those 75 or more years of age.

Another issue was age confounding. Survival times decrease as age increases, and prevalence of many health behaviours also differ by age. For example, among the SEER cancer registry cases imputed smoking rates are lower in older age groups, while prevalence of obesity and physical activity outside of employment are both higher in older age groups. Age-standardisation was only partially successful at removing the age effect and sample sizes were insufficient for age-stratification.

The I2C2-based analyses did produce some suggestive results. Of the six behaviour measures studied, imputed smoking status 5-years prior to diagnosis appeared to retain the

most information through the imputation with a median excess smoking-to-smoking matches of 299 (95% CI 246,353), or 0.009 (95% CI 0.008,0.011) of the sample. In the current study the age-standardised estimated median HR for smoking was 1.32 (95% CI 0.82,2.68). But previous studies have suggested the effect of smoking on survival may differ between sub-groups (Fahey et al., 2015).

Recent meta analyses estimated HRs of 1.41 (95% CI 1.22,1.64) and 1.41 (95% CI 0.96,2.09) for current smoking relative to never smoked in mainly squamous cell populations (Kuang et al., 2016; McMenamin et al., 2017) with no evidence of association between smoking and survival in EAC (Fahey et al., 2015; McMenamin et al., 2017). In the current study age-standardised results showed a statistically significantly increased HR in squamous cell carcinoma (HR=1.93, 95%CI 1.08,7.07) with no apparent effect in adenocarcinoma (HR=0.89, 95%CI 0.05,3.53). Based on the simulations, these results are likely to underestimate the true hazards.

Physical activity outside employment 5 years prior to diagnosis, recorded 214 (95% CI 145-273) more agreements between imputed values than predicted by chance. Physical activity appeared to be protective of survival age-standardised estimated HR of 0.25 (95%CI 0.03,0.81). A recent meta-analysis (Friedenreich, Stone, Cheung, & Hayes, 2020) combined results from a US and a Korean study and reported that pre-diagnosis physical activity to be protective of post-diagnosis survival in oesophageal cancer (HR=0.77, 95%CI 0.59-1.00). No evidence was found for a difference between squamous cell and adenocarcinoma in the association between physical activity outside work and survival.

I2C2-derived HR estimates for other health behaviours were sparse and in one case problematic: the tendency for 'binge drinking' 5 years prior to diagnosis to be protective of age-standardised survival (HR 0.49, 95% CI 0.13,1.29) may suggest a failure of age standardisation.

These findings suggest potential in the I2C2 algorithm, and with the benefit of experience, there are many ways to improve its performance. Sample size could be increased in a variety of ways. For example, the collection could be broadened to include more cancer registries or more years of data. However, data from the non-SEER cancer registries in the US are less accessible and the number of years included is limited by changing data definitions and changing clinical practices. But minimising data losses arising from the matching algorithm is readily achievable in the short-term. The variables used in this paper are quite limited: each of the 6 behaviours were dichotomous and obviously dichotomous variables convey the least possible information about behaviour. Some of the demographic variables (such as 5-year age groups, State of residence, etc) could have conveyed more information if divided into smaller categories. The imputation process used was parsimonious: simply excluding cancer registry cases with less than two donor records rather than attempt to find nearest neighbours etc. Model-based imputation (Blanchette, DeKoven, De, & Roberts, 2013) may be more informative than simple donor records. Additional investigation of the measurement of misclassification would also strengthen the approach. For example, internal calibration, confirming the true health behaviour from a sub-set of the cancer cases could give a more direct measure of misclassification. Similarly, a quick and simple method for age-standardisation was used, which has been subject to previous criticism (Nieto & Coresh, 1996). Ideally, it may be possible to refine the corrected scores estimation algorithm to address confounding variables directly (such as ability to specify offsets or fix

the coefficient for age). Finally, the I2C2 needs to be more thoroughly tested. A prospective or retrospective cohort which contained true pre-diagnosis behaviours as well post-diagnosis survival times would allow gold standard comparisons between true the HR and I2C2-estimated HR.

The variety and size of health surveys and prospective cohorts will continue to increase over time. All such large human data sets tend to contain a good variety of demographic measures allowing differentiation of people with similar and dissimilar risk factors. Unlike one-to-one matching, the I2C2 algorithm produces almost no additional confidentiality risk over and above the original cancer registry. (The data which is being added to the cancer registry is largely misclassifications.)

Further improvements in the I2C2 algorithm will permit enrichment of cancer registry data through imputation of new variables with negligible risk to patient confidentiality, opening new research opportunities in cancer epidemiology.

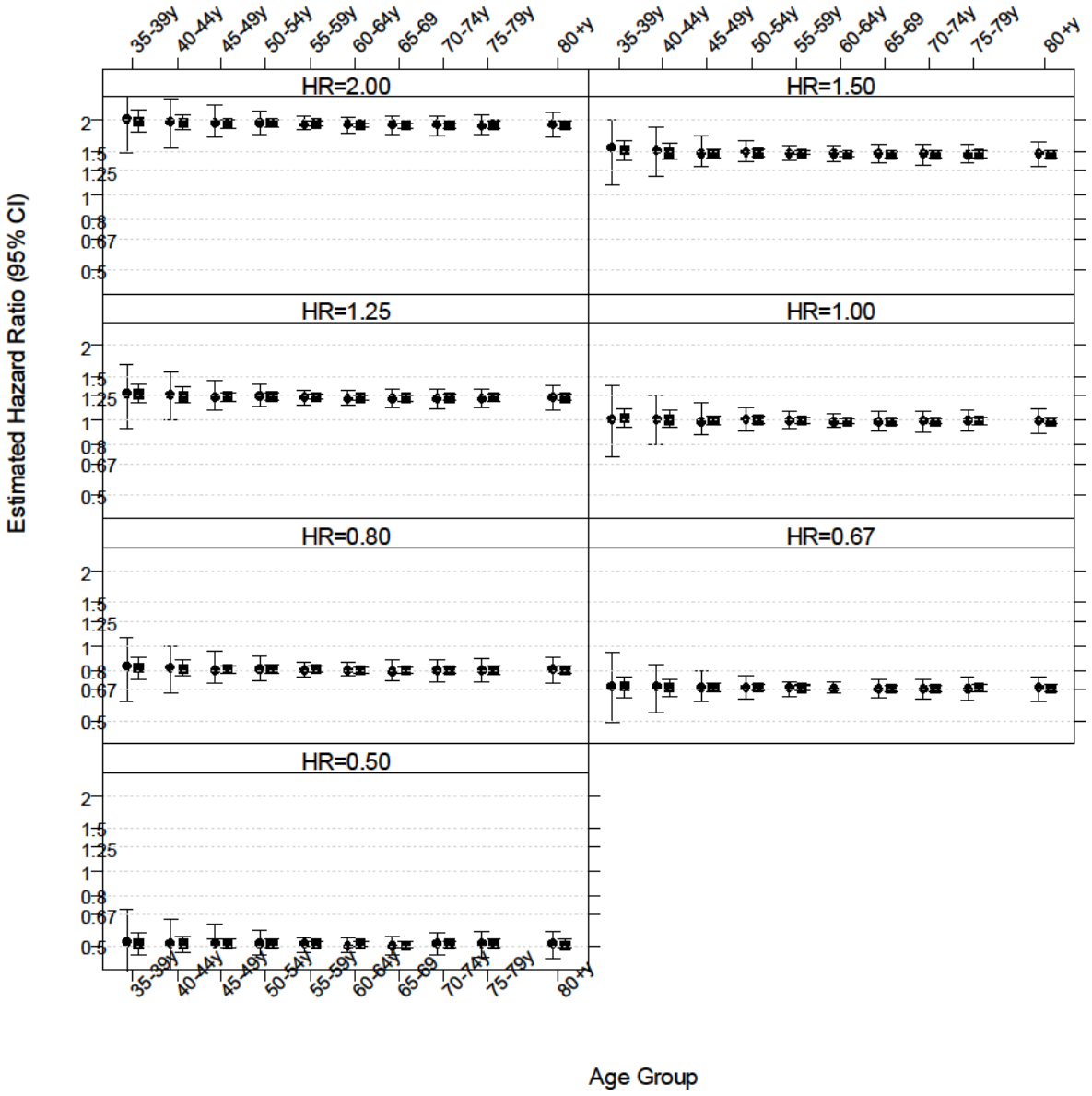
6.4 Further exploration of stratification for age confounding

In the draft manuscript above, stratification for confounding adjustment was dismissed because the sample sizes were too small. For the sake of completeness, this section considers with what may have happened if the sample sizes were increased 10-fold or 20-fold. To generate a larger sample size, 10 of the simulated data sets were randomly selected (with replacement) and combined into a single data set of 10 times the size. This was done 100 times to generate 100 larger data sets for analysis.

Increasing the sample sizes in this way will not affect the median estimated HR (point estimate) but does invalidate the empirical CIs. That is, when combining 10 data sets, those

with randomly lower HRs tend to be combined with those with randomly higher HRs leading to an overall cancelling out in variability. Figure 6.13 provides a visual demonstration of this effect. Here the empirical confidence intervals are calculated for the designated ‘true’ smoking status in the simulated data sets using standard Cox modelling (i.e. no need for misclassification correction because these are the ‘true’ values.). Notice that artificially inflating the sample size, is artificially narrowing the empirical confidence intervals. As the CIs are misleading and they are supressed in all further analyses.

Figure 6.13 Median estimated hazard ratios with associated 95% empirical confidence intervals for each target hazard ratio and each 5 year age category.

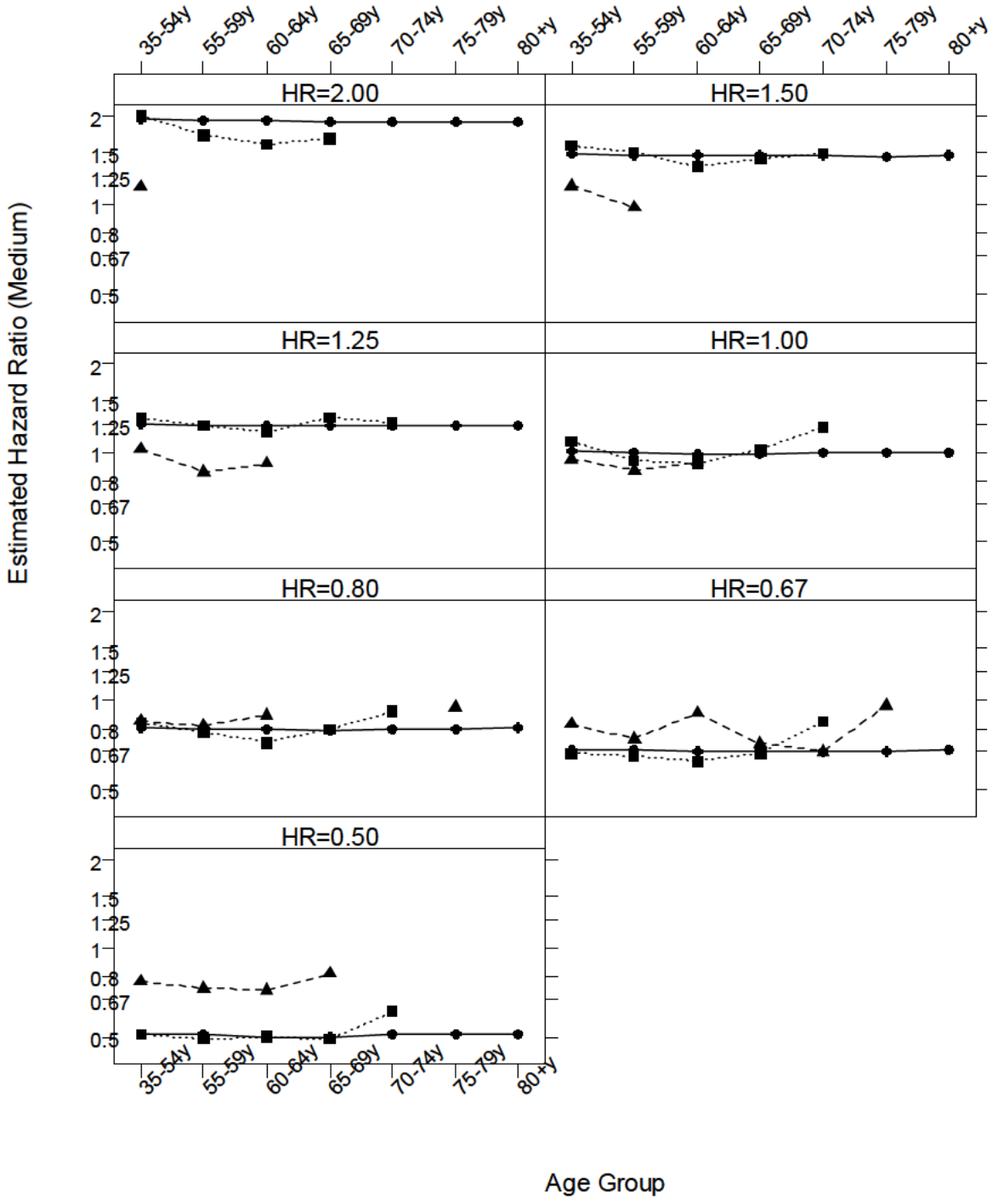


The first confidence interval in each pair shows the hazard ratios for 'true' smoking status and the second shows the hazard ratio for 'true' smoking status after the sample size is artificially inflated 10-fold.

Staying with the simulated data sets for smoking status, Figure 6.14 shows that the median estimated HRs from the I2C2 algorithm from the inflated sample size conform quite closely with the true median HR: particularly in the age groups under 70 years. However, at

the lower end of the age range, the first 4 age groups were concatenated as sample sizes were too small, and at the upper end of the age range sample sizes are large but the number of smokers is so small that there still seems to be insufficient information for the algorithm to work.

Figure 6.14 Trace of the medians of estimated HRs by target HR and age category.



The four youngest age categories have been concatenated. Solid lines with diamonds represent medians from Cox models on true smoking status. Dashed lines with triangles represent medians from the I2C2 algorithm. Dotted lines with squares represent medians from I2C2 algorithm fitted on data sets inflated 10-fold.

The I2C2 algorithm was then applied to the 10-fold inflated SEER cancer registry data augmented by imputed health behaviours. Even with 10-fold inflation of the sample sizes, Figure 6.15 shows that there is insufficient information to support the analysis of most health behaviours in most age groups. Information levels only appear to be sufficient for smoking status among those less than 75 years, binge drinking among those less than 60 years and non-employment physical activity in those 65 years and older. Figure 8 provides the estimated median HRs for these groups.

Figure 6.15 Bar charts of the percentage of data sets returning extreme values (<0.01 or >100) for estimated HRs when using the I2C2 algorithm in the age-stratified cancer registry data sets with 10-fold inflated sample sizes.

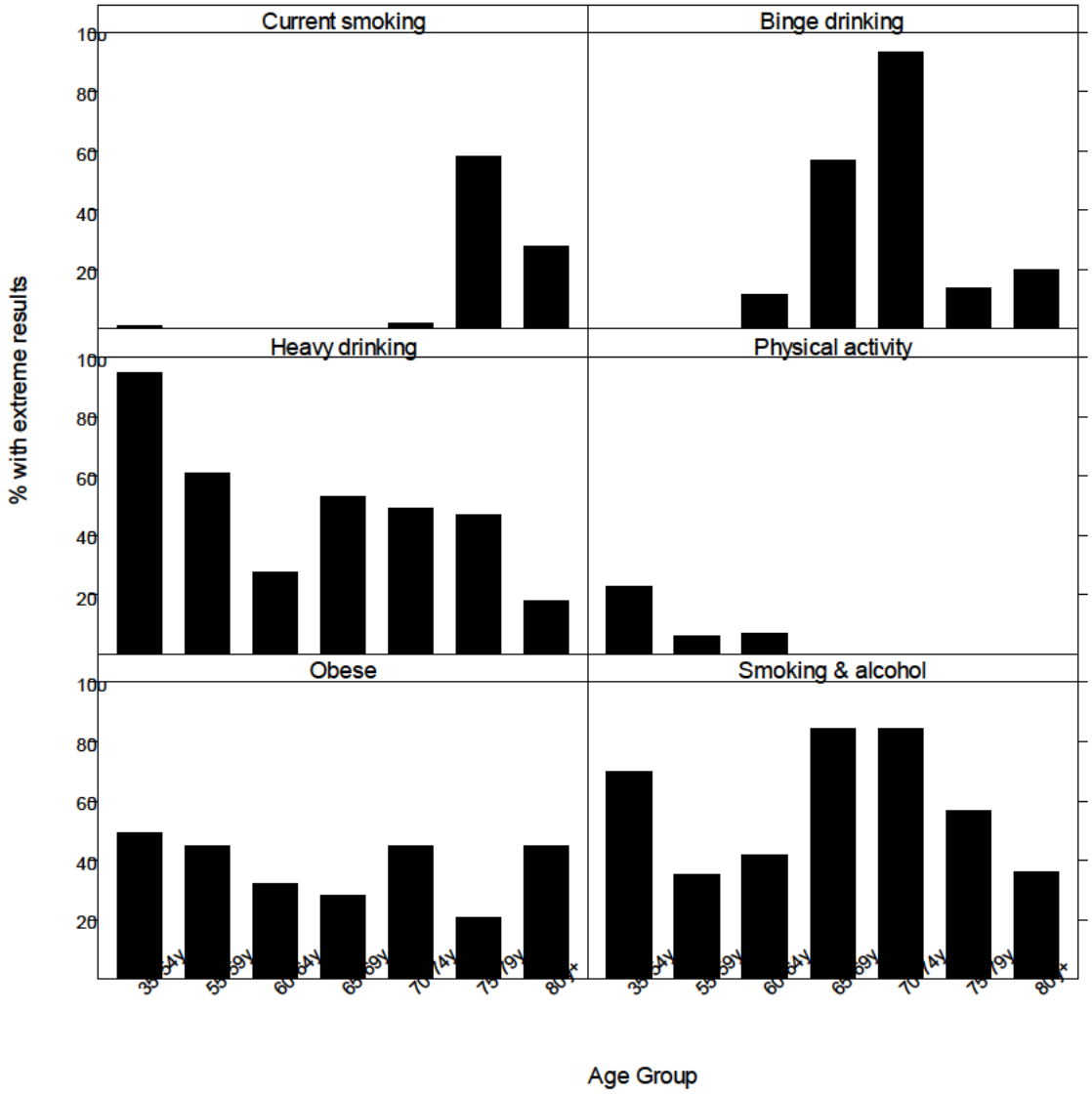
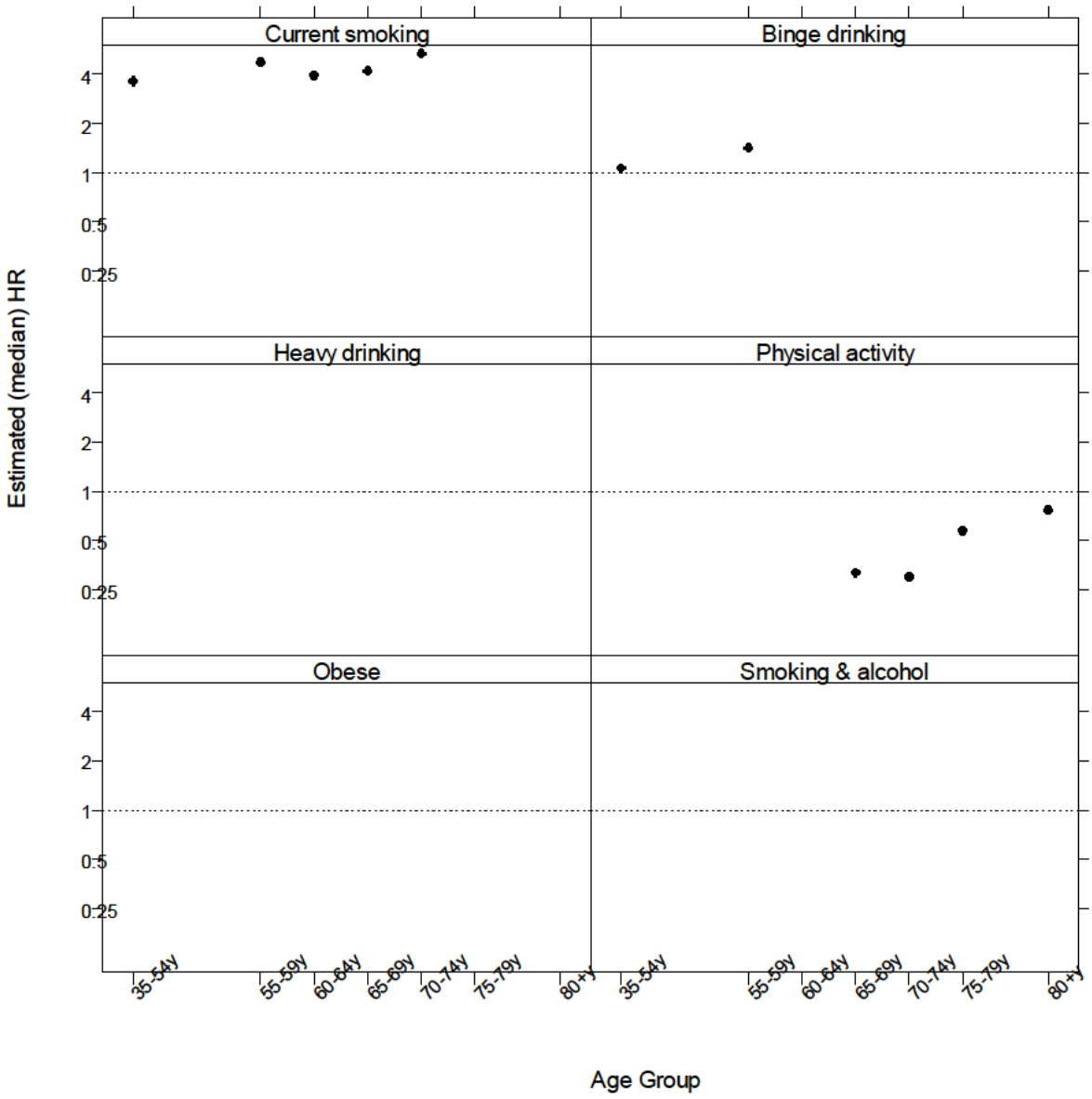


Figure 6.16 shows that the point estimate hazard ratio for binge drinking five years prior to diagnosis is slightly higher than the null (median HR 1.42) in 55-59 year age group. Smoking 5 years prior to diagnosis has consistently high point estimate HRs across the younger age groups (3.65, 4.77, 3.92, 4.15 and 5.4 for 34-54 years to 70-74 year age groups). Physical activity outside employment 5 years prior to diagnosis has point estimate HRs consistent with being protective of survival in the over 65-year age groups (estimated median HRs 0.32 for 65-69 years, 0.30 for 70-74 years and less effective 0.57 and 0.77 in the two older age groups).

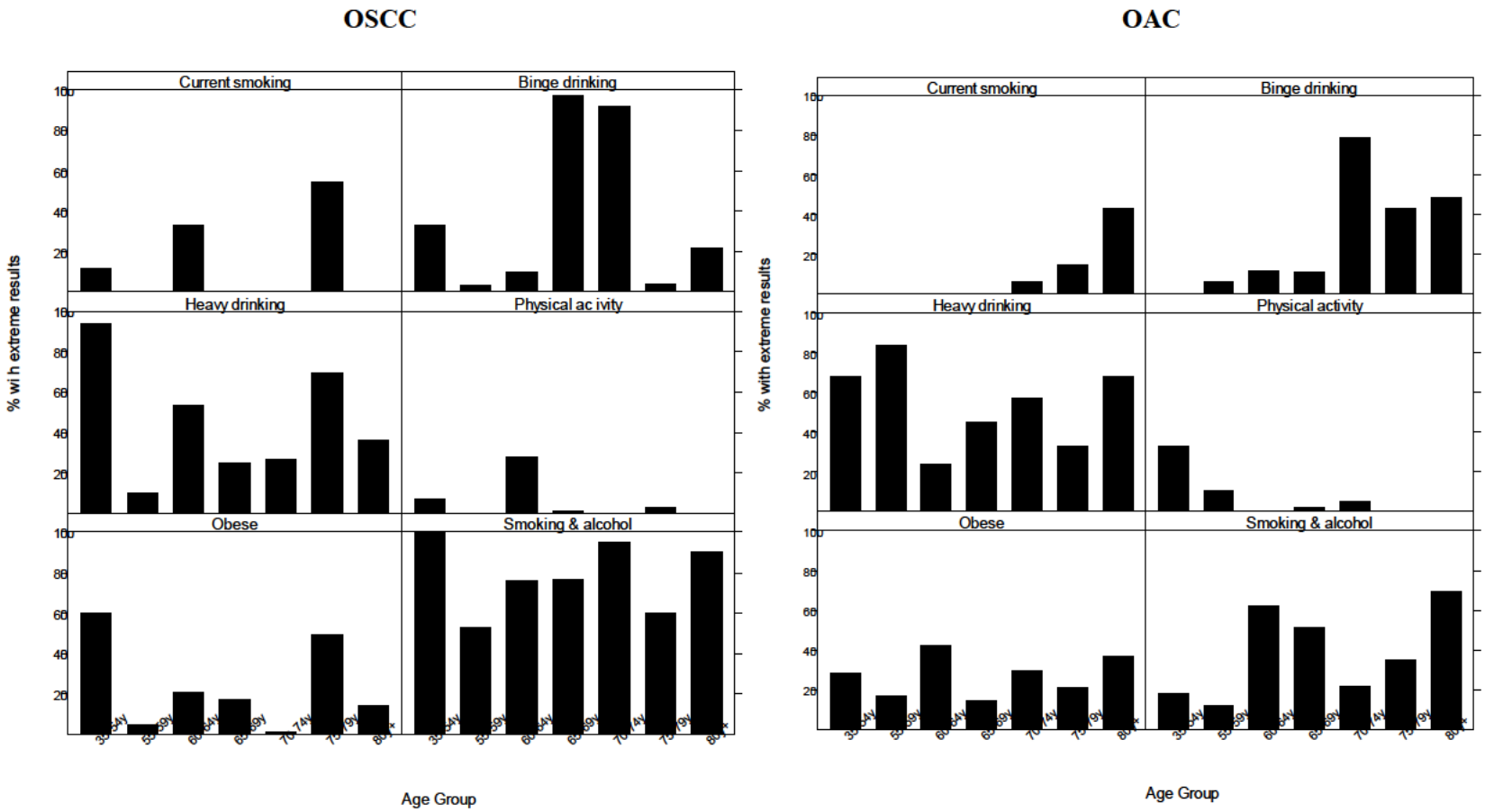
Figure 6.16 Median of the estimated HRs in behaviour age-stratified cancer registry data sets with 10-fold inflated sample sizes.



Age groups where more than 5 of the 100 data sets returned extreme values (<0.01 or >100) for estimated HRs have been omitted.

The same analyses were repeated for the OSCC and OAC sub-groups. With smaller sample sizes in the subgroups, the true sample sizes were inflated 20-fold in an attempt to provide the I2C2 algorithm with sufficient information. Despite this, the results in Figure 6.17 confirm that there is insufficient information for many of the behaviour by age categories.

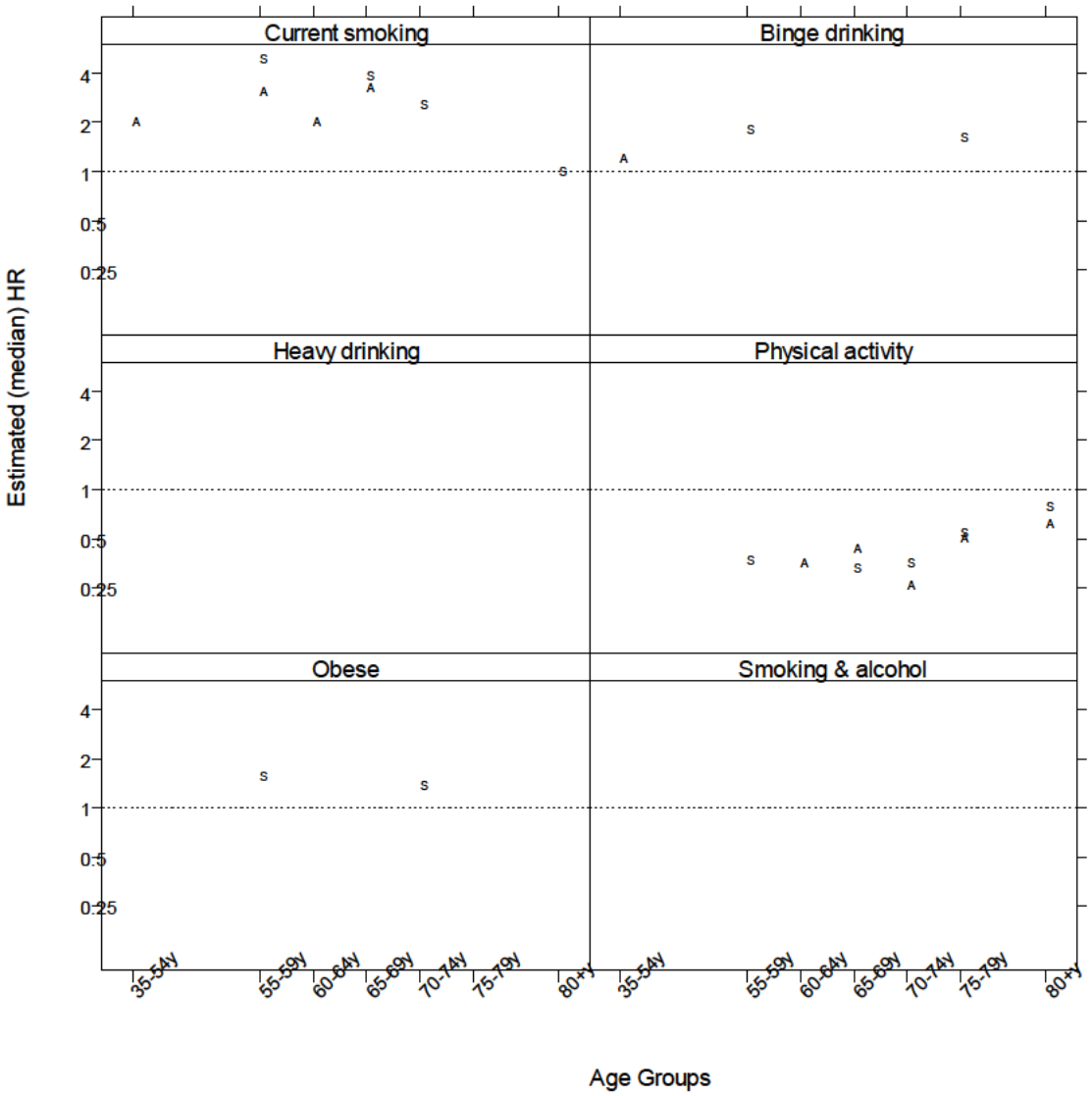
Figure 6.17 Bar charts of the percentage of data sets returning extreme values (<0.01 or >100) for estimated hazard ratios when using the I2C2 algorithm in the age-stratified cancer registry data sets with 20-fold inflated sample sizes



The results which are available are summarised in Figure 6.16. Notice that median estimated HRs tend to cluster above the null for ‘Smoking status’ and below the null for ‘Physical Activity’ with effect sizes reducing for those 75 or more years of age. Results for other behaviours are sparse to non-existent.

For ‘Smoking status’ among those under 75 years, the estimated median HRs are 5.1, 4.0, 2.6 for the squamous cell subgroup, generally higher than the 2.1, 3.2, 2.1 and 3.4 estimated median HRs observed in the adenocarcinoma group, but for ‘Physical activity’ the there are no marked differences between the subgroups (with estimated median HRs of 0.38, 0.34, 0.36 for squamous cell carcinoma and 0.36, 0.45, 0.27 for adenocarcinoma subgroups.).

Figure 6.18 Sub-group median of the estimated HRs in behaviour age-stratified cancer registry data sets with 20-fold inflated sample sizes.



'S' represents squamous cell carcinomas and 'A' represents adenocarcinomas. Age groups where more than 5 of the 100 data sets returned extreme values (<0.01 or >100) for estimated HRs have been omitted.

This exercise has demonstrated that with larger sample sizes, stratification may become viable.

6.5 Critique of the method

In this Chapter the data sets (with one additional year) and variables introduced in Chapter 3 and the method of imputation and estimation of misclassification error introduced in Chapters 5 were used. The issues arising have been discussed previously. The new focus of this Chapter was on identifying and testing misclassification adjusted Cox proportional hazards models for use with the I2C2 algorithm.

The corrected score method for Cox regression was developed by Prof David Zucker and Prof Donna Spiegelman. Prof Zucker kindly provided his program for the corrected score method which he had written in Fortran 77. As the fitting algorithm is complex and written in an unfamiliar programming language, it was difficult to modify this program to suit the specific requirements of this research project.

The maximum number of lines of data which could be processed by the program was less than 100,000. As the estimated HR was of interest but not the associated estimated standard deviations, the error messages from the program were not fully aligned with the needs of this thesis. In this thesis quite crude criteria (HR<0.01 or HR>100) was used to indicate failure of the model. This will need to be reviewed and perhaps improved in the future.

The Fortran 77 standard was released in 1977. Today's computers and computing are very different now. Software such as R or Python are preferred choice for programming statistical algorithms (T. Siddiqui, Alkadri, & Khan, 2017). Coding the corrected score method into one of these more common languages would make the algorithm more accessible to users and aid in its expansion and development. As R is open source, it may possible to copy and edit the macros for fitting Cox regression (such as the `coxph()` command) rather

than developing a whole new program. But this is beyond the scope of the current thesis and recommended as an issue for future research.

The corrected score method of misclassification adjustment worked as expected in the absence of confounding (Figure 6.4). However, when age was added as a confounding variable to the simulated data sets the corrected scores algorithm misassigned much of the effect of smoking (measured with misclassification) to age group (measured without classification error) (Figure 6.5).

Using the corrected scores method the estimated coefficient of age varied instead of the hazard ratio of smoking status. The coefficient of for age group is readily estimated by fitting age as a sole predictor of survival time in the Cox regression. Age group and survival time are both recorded in the cancer registry data set and so are not subject to imputation or misclassification error. Indeed the estimated coefficient, 0.04, was used when simulating the survival times (`beta_age <- 0.040` in the example R code above). Given that the correct coefficient of age was known, this should have been fixed in the model fitting algorithm. A fixed coefficient can be included in a regression analysis using the offset option. An offset variable has a fixed coefficient of 1. So using R

```
coxph(Surv(sm_time2,status)~smoker2+offset(0.04*agegrp),data=...)
```

would produce an estimated coefficient for the second imputed smoking status, where the effect of age group is fixed at the known value of 0.04. Unfortunately, the corrected scores program in Fortran 77 did not provide an option for offset variables and modification of the original program was not possible.

When the coefficient of age is unknown or cannot be offset, the options of correcting age confounding are stratification, standardization, multivariable analysis and propensity

score methods (Kahlert, Gribsholt, Gammelager, Dekkers, & Luta, 2017). Propensity score matching requires the exposed group and unexposed group to be correctly identified, which is not possible with misclassification. The multivariate analysis was not effective for the correct score method as previously described. In Sections 6.3 and 6.4 age stratification was investigated in sufficient detail to demonstrate that very large sample size are required to make this approach potentially viable.

Age-stratification did reveal that the oldest groups have very much reduced prevalence of smoking and the models appear ineffective at these age strata even though the total number of people in the older strata were quite large. Similar effects were found for all behaviours, even though not reported in this thesis. This may warrant further consideration of either the 5-year definition of pre-diagnosis behaviour and/or the inclusion of the older age groups in the analyses. For example, research questions looking at 'lifetime tobacco smoking' or 'ever smoked' may be more interesting than questions about 5-years pre-diagnosis.

The only remaining method for correcting for age confounding appears to be standardisation. In the current analyses survival times were adjusted according to age group. Findings indicated that the hazard ratios produced by this method were slightly attenuated to the null in the absence of misclassification, but that this small residual attenuation became magnified by the misclassification correction process, leading to considerable underestimation of the hazard ratios. One alternate approach which has been suggested is, instead of applying weights to survival times according to age groups, try applying weights to age groups at each observed survival time (Nieto & Coresh, 1996). However, it is not clear if this method could be applied to the current situation where the confounder, age-group, has 10 categories.

The methods described in this thesis have potential but require additional investigation and development. Further research of offset variables and standardisation methods may lead to further improvements.

6.6 Closing comments

This Chapter advanced and attempted to generalise the method to impute 100% missing variables in cancer registry data, formalising the process as the I2C2 algorithm. Additional estimates of the associations between pre-diagnosis health behaviour and post-diagnosis survival in OC were obtained. So both thesis aims were at least partially addressed.

Further research is warranted for each step of the analysis process: how to improve imputation, how to improve the measurement of misclassification and how to improve the misclassification correction algorithm, particularly in relation to confounding. Further research is also required into the accuracy and robustness of the estimates of association between pre-diagnosis health behaviour and post-diagnosis survival in OC, ideally comparing the results of I2C2 against some gold-standard prospective or retrospective (data-linkage) cohort study where results from observed and imputed health behaviours can be compared.

This thesis has made considerable progress in documenting and testing an entirely new research approach that has potential for further development. Chapter 7 reviews the key findings from this research, discusses additional investigations that need to be pursued, and considers applications and implications of the material in this thesis.

Chapter 7 Discussion

7.1 Background

The first aim of this thesis was to document association between pre-diagnosis health behaviour and post-diagnosis survival times in OC. The review of the literature (Chapter 2) confirmed evidence of associations for a variety of health behaviours, but also indicated that more research was needed. As this is an association across time, the main research question would typically be addressed through a cohort study; either prospective or retrospective (record linkage). But suitable data for OC could not be obtained.

When looking for alternative methods to prospective cohorts and record linkage studies, surprisingly little was found. This led to the second aim of this thesis: to develop, describe and evaluate a new method for addressing the first aim.

This chapter reviews what has been achieved, make recommendations for further research and explore the implications of the research to date.

Section 7.2 reviews findings from first aim of this thesis: what has been learned to date, what is not yet resolved and recommendations for future research. Section 7.3 presents a similar review of progress in addressing the second aim of this thesis. In Section 7.4 the potential implications and applications of this research are discussed, and in Section 7.5 conclusions from this thesis are presented.

7.2 Progress in addressing the first aim

OC has poor survival and so it is reasonable to expect that pre-diagnosis health behaviour could have an important carry-over effect on survival times. The clinical mechanisms for any such associations have not yet been explored. This thesis was specifically interested in pre-diagnosis behaviour because this can be influenced by public health programs. Post-diagnosis behaviour is subjected to other factors such as the treatment regimen and the trauma of the cancer diagnosis itself. Pre-diagnosis health behaviour is a factor in survival beyond the influence of the treating medical team, but within the influence of public health professionals.

The systematic review (Chapter 2) investigated the association between pre-diagnosis smoking, alcohol consumption, body mass index, physical exercise and/or regular consumption of non-steroidal anti-inflammatory drugs (NSAIDs) and post-diagnosis survival in OC. The 2014 review located 13 eligible studies: the majority from Asia. The 2021 update identified a further 14 eligible studies. This doubling of eligible studies in the past 7 years is consistent with increasing interest in the research question. The pooled hazard ratios from the systematic review combined study results from around the world with most studies from Asia. These studies were typically relatively small, conducted on a few hundred OC cases.

Analyses of similar pre-diagnosis health behaviours were also conducted using OC cases from the US SEER cancer registry, 2001 to 2015, with around 25,000 to 30,000 OC cases (Chapters 4 to 6). All behaviour measures were self-reported and dichotomised.

7.2.1 **Tobacco smoking**

Within this thesis there was consistent evidence that pre-diagnosis smoking is associated with poorer survival in OC. There is more evidence of association in the OSCC subgroup than in OAC. The results are summarised in Table 7.1 with discussion below.

Table 7.1 Summary of evidence for association between pre-diagnosis smoking and survival

measure	documented	OSCC HR/RR (95%CI)	OAC HR/RR (95%CI)
Pooled hazard ratios - Ever versus never smoked - Highest smoking (pack years) vs lowest	Chapter 2	 1.09 (0.97,1.23) 1.55 (1.24,1.94)	 1.07 (0.96,1.19) n/a
Hazard ratio for 0.1 increase in the probability of being a smoker	Chapter 4	1.20 (1.17,1.23)	1.20 (1.18,1.23)
Age adjusted relative risk of death within 1 year of diagnosis	Chapter 5	1.99 (1.24,3.12)	1.61 (0.79,2.57)
Age adjusted hazard ratio for smoker 5-years pre-diagnosis	Chapter 6	1.93 (1.08,7.07)	failed

All exposure data were based on self-report and it is possible there may be a bias towards under-reporting smoking quantity. For example, an analysis using the US National Survey of Drug Use and Health and US National Health Interview Survey found that self-reported cigarette consumption was only about 65% of cigarette sales (Liber & Warner, 2018). However, most results reported in the current thesis relate to binomial measures of smoking which may have lower bias. For example, 15.7% of OC cases include in the analyses were imputed to be smokers compared to US adult population estimates of 20.6% to 15.1% for 2006 to 2015 (American Lung Association, 2021). Given the OC population is older, with smoking rates decreasing in the oldest age groups, this small disparity may be understandable.

The measures of tobacco smoking also differ between the systematic review and the analyses of the SEER cancer registry OC cases. The smoking measures reported in the systematic review were longitudinal (ever smoked, lifetime exposure in pack years), whereas the analyses of OC cases from the SEER cancer registry was cross sectional (current smoker at some time point prior to diagnosis). As a result, the classification of former smokers differs. In the longitudinal measures, former smokers were included in the exposure groups but in the cross-sectional analyses they were included in the non-exposure (not a current smoker) group. It is known that risk of OC is lower in former smokers than current smokers, particularly for OSCC (Q.-L. Wang, Xie, Li, & Lagergren, 2017) and it may be reasonable to assume that there is a similar effect for survival time. That is, the lower estimated hazard ratios from the systematic review may be consistent with a large group of former smokers whose hazards are much lower than the current smokers with whom they are grouped.

From Table 7.1 the estimated hazard ratios associated with a 0.1 increase in probability of being a smoker presented in Chapter 4 fail to differentiate between OSCC and OAC cases. (The hazard ratios are the same and hence represent the hazard ratio of the combined OSCC and OAC groups.) These results can also be criticised for the arbitrary choice of a 0.1 increase in probability, the difficulty in interpreting the results and the artificially narrow confidence intervals. The results from these analyses should be discounted or even dismissed.

It can be seen in the final row of Table 7.1 that models failed to produce a misclassification corrected, age-adjusted hazard ratio for pre-diagnosis smoking in OAC, even though the OAC sample was more than 1.5 times larger than the OSCC sample. There is a temptation to assume that the algorithm failed because the association between pre-diagnosis smoking and post-diagnosis survival in OAC was very weak. However other

explanations, such as different age structure and/or different survival distributions between OAC to OSCC, are perhaps more likely.

Most confidence intervals for the effect of pre-diagnosis smoking on survival in OSCC excluded the null, but in OAC only the artificially narrow confidence interval for a 0.1 increase in probability of pre-diagnosis smoking excluded the null.

7.2.2 **Alcohol consumption (without and with smoking)**

Different countries have different views on alcohol consumption. Fourteen countries prohibit alcohol entirely (WorldAtlas, 2021), but the majority permit alcohol consumption while designating consumption levels "at risk" to health. Because "at risk" alcohol consumption is based on self-reported consumption it is prone to under-reporting and this under-reporting has been shown to be worse among the "at risk" groups that the analyses in the current thesis were attempting to identify (Boniface et al., 2014). Misclassifying individuals who are truly "at risk" as "not at risk" will weaken observed hazard ratios towards the null.

In Australia in 2019 an estimated 16.8% of people aged 14 years and older consumed 2 standard drinks per day and 25% of people aged 14 and over consumed more than 4 standard drinks in one sitting at least monthly (Australian Institute of Health and Welfare, 2021). While these at-risk behaviours have quite similar definitions to the 'heavy drinking' and 'binge drinking' variables which were obtained from the BRFSS health survey data, the corresponding rates in the imputed behaviours for OC cases from the SEER cancer registries were only 4.8% and 10.0% respectively. These are less than half of the number that might be expected. This suggests considerable under-reporting or under-representation of 'at risk'

behaviour: with ‘at risk’ people selected for inclusion in the BRFSS health survey either under-estimating their alcohol consumption or declining to participate in the survey.

The systematic review produced pooled estimates of hazard ratios for the dichotomous variables ‘ever’ versus ‘never’ and ‘highest’ alcohol consumption versus ‘lowest’. If under-reporting is common across all drinking, the at risk drinkers will still likely cluster in the highest self-report category even though the entire distribution of alcohol consumption has been shifted downward.

The systematic review found a positive hazard of about 1.4 for pre-diagnosis alcohol consumption in OSCC (subject to considerable heterogeneity) and somewhat lower for OAC. The analysis of the cancer registry data found a similar 1.52 hazard ratio associated with binge drinking in OSCC but with wide confidence intervals. The results are summarised in Table 7.2.

Table 7.2 Summary of evidence for association between pre-diagnosis alcohol consumption and survival

measure	documented	OSCC HR/RR (95%CI)	OAC HR/RR (95%CI)
Pooled hazard ratios			
- Ever vs never alcohol consumption	Chapter 2	1.43 (1.08,1.89)	1.22 (1.07,1.39)
- Highest weekly alcohol vs lowest		1.32 (0.92,1.89)	1.08 (0.87,1.32)
0.1 increase in the probability of			
- Binge drinking	Chapter 4	0.95 (0.90,1.00)	0.97 (0.93, 1.01)
- Heavy drinking		0.78 (0.69,0.88)	0.85 (0.77,0.93)
- Smoking and regular drinking		1.93 (1.72,2.16)	1.93 (1.76,2.11)
Age-adjusted relative risk			
- Binge drinking	Chapter 5	1.52 (0.44, 2.75)	0.82 (0.18,1.89)
- Heavy drinking		failed	failed
- Smoking and regular drinking		failed	failed
Age-adjusted Cox regression			
- Binge drinking	Chapter 6	failed	failed
- Heavy drinking		failed	failed
- Smoking and regular drinking		failed	failed

Given the very high levels of misclassification produced by cold deck imputation, when a behaviour is rare, there appears to be insufficient information to support misclassification correction in a number of the analyses. The imputation-based analyses of heavy drinking, with an imputed prevalence of 4.8% among SEER OC cases, failed throughout. For the somewhat more common binge drinking, the age-adjusted relative risk of dying within 1 year returned confidence intervals wide enough to include the null effect for both OSCC and OAC, and the adjusted Cox regression analyses failed.

The results of the investigation of the combined regular drinking with smoking behaviour are also presented in Table 7.2. It had been observed that combining "at risk" alcohol consumption with smoking produces a disproportionate increase in risk on OC (Castellsagué et al., 1999) and there was some evidence of a similar effect on post-diagnosis survival in OSCC (Thrift, Nagle, Fahey, Russell, et al., 2012). No other studies had addressed the association between pre-diagnosis regular alcohol with smoking and post-diagnosis survival and so there were no pooled estimates arising from the systematic review. The definition of 1.0 drinks/day used to the define alcohol consumption component of this variable is lower than the standard definitions of 'at risk' alcohol consumption. Including those who drink less into the 'at risk' group may have diffused the predictive power of this behaviour. The hazard ratios for a 0.1 increase in the prevalence of regular drinking and smoking produced an estimated hazard ratio of 1.93 with an artificially narrow confidence interval. But given the relative rarity of this behaviour combination, a 0.1 increase is extrapolating far beyond the observed data. Indeed, just 3.3% of the SEER OC cases were imputed to have had regular drinking with smoking. The imputation-based methods failed, presumably due to this low prevalence.

7.2.3 **Physical activity**

Physical activity differs with the level of economic development. People in less developed economies have higher levels of work and transport related physical activity while people in more developed economies are more likely to have sedentary work and transport and be more reliant on leisure time physical activity (Bauman et al., 2011). Measuring physical activity can be complex and there are a variety of measurement tools available.

Only three previous studies of the association between pre-diagnosis physical activity and post-diagnosis survival in OC were identified, and these studies used different measures of physical activity.

In the analyses of the SEER OC cancer registry cases the BRFSS health survey dichotomous variable "physical activity other than for regular job" was used, concentrating on leisure time physical activity levels. The results are presented in Table 7.4.

Table 7.3 Summary of evidence for association between pre-diagnosis physical activity and survival

measure	documented	OSCC HR/RR (95%CI)	OAC HR/RR (95%CI)
Pooled hazard ratios - Pre-diagnosis physical activity	Chapter 2	0.92 (0.67,1.27)	1.20 (0.91,1.58)
0.1 increase in the probability of leisure time physical activity	Chapter 4	0.82 (0.80,0.85)	0.83 (0.81,0.85)
Age-adjusted relative risk	Chapter 5	0.45 (0.10, 1.07)	0.63 (0.29,2.18)
Age-adjusted Cox regression	Chapter 6	0.34 (0.07,1.64)	failed

Analyses suggested that pre-diagnosis leisure time physical activity in the US was associated with a lower post-diagnosis risk of death than for those who do not exercise. However, all confidence interval included the null (except the artificially narrow confidence intervals for 0.1 increase in the probability of leisure time physical activity).

7.2.4 **Overweight and obesity**

Overweight and obesity is an indicator of diet and physical activity. Overweight and obesity were defined using the WHO cutpoints on body mass index (Hubbard, 2000). The systematic review found that the combined overweight and obese group had a small survival advantage (HR about 0.80) over normal weight in both OSCC and OAC and the confidence intervals excluded the null, but this disappeared in obese only group. The apparent advantage of overweight may be related to the fact that OC can often interfere with feeding. However, the pooled results come from just 5 papers for OSCC and 5 papers for OAC, with marked heterogeneity between these studies.

All other analyses were conducted on the obese group alone and all confidence intervals included the null (except the artificially narrow confidence intervals for 0.1 increase in the probability of being obese). Overall, any association between pre-diagnosis overweight or obese on survival time appears to be less consequential than the associations observed for smoking and physical activity. The results are summarised in Table 7.2.

Table 7.4 Summary of evidence for association between pre-diagnosis overweight and/or obesity and survival

measure	documented	OSCC HR/RR (95%CI)	OAC HR/RR (95%CI)
Pooled hazard ratios - Overweight or obese versus normal - Obese versus normal	Chapter 2	 0.79 (0.61, 1.03) 1.05 (0.76, 1.46)	 0.87 (0.73,1.04) 0.95 (0.77, 1.18)
0.1 increase in the probability of being obese	Chapter 4	1.07 (1.04,1.10)	1.03 (1.00, 1.05)
Age-adjusted relative risk for obese	Chapter 5	1.53 (0.83,4.17)	failed
Age-adjusted Cox regression for obese	Chapter 6	1.08 (0.22,4.96)	failed

7.2.5 **Regular NSAID consumption**

The systematic review revealed high heterogeneity between studies and the pooled results (HR 0.98, 95%CI 0.82,1.18, I^2=77.8%) provided no evidence that regular consumption of NSAIDs pre-diagnosis was associated with post-diagnosis survival times in OC. NSAID consumption could not be included in the analyses of SEER cancer registry cases, as the BRFSS health survey did not collect this information.

7.2.6 **What has been learnt and what is yet to do**

This thesis has presented consistent evidence that pre-diagnosis smoking is associated with decreased post-diagnosis survival in OSCC, with less consistent evidence for OAC. Smoking about 1 to 5 years prior to OSCC diagnosis was associated with a doubling of the hazard / risk of death. Leisure time physical activity 1 to 5 years prior to diagnosis may be associated with a halving of the post-diagnosis hazard / risk of death in OC, but stronger evidence is required. The association between pre-diagnosis overweight and/or obesity may not be large enough to warrant further investigation. The association between 'at risk' pre-diagnosis alcohol consumption and post-diagnosis survival time is somewhat less certain. A new measure approach may be required to address the apparent under-reporting. In the analyses of the SEER OC cases, the choice of variables was decided by the developers of the BRFSS health survey and so were not optimised for the specific research aims of the thesis.

7.3 Progress in addressing the second aim

A new algorithm was developed to research the association between pre-diagnosis behaviour and post-diagnosis survival time in cancer which started with the SEER cancer registry data and then added pre-diagnosis health behaviour variables. When first added, these variables were empty (that is, contained 100% missing data) and it was hypothesised that a method could be developed for addressing missing data: either through modelling or imputation. As the data were 100% missing, an external data source was required to inform the missing value correction. The BRFSS national health survey was identified as a suitable external source for health behaviour variables. It was assumed unlikely that there would be many individuals in common between the cancer registry and the health survey, but (as often occurs with human data sets) there were a range of demographic measures in common.

To communicate information from the external health survey data to the cancer registry the analyses relied on the fact that two individuals from the same demographic group are more likely to display similar behaviour than two individuals with different demographic backgrounds; in alcohol consumption for example (Boersma et al., 2020) or smoking (Leventhal et al., 2019). That is, it was hypothesised the demographic variables could convey some, albeit small, information about health behaviour. The expected weakness of this signal component suggested the need for a large sample size to detect it.

7.3.1 A model-based solution

Using the health survey data only, a logistic model was used to predict the probability of a behaviour, using only demographic variables as predictors. The model was then applied to each OC cancer registry case, using those demographic variables to estimate their

probability of participating in that behaviour pre-diagnosis. These estimated probabilities of behaviour were then used as the predictor in the Cox regression model for survival time.

There are three main problems identified in using this approach. Firstly, the resulting hazard ratios were for a 0.1 increase in the probability of the behaviour rather than describing the difference between having the behaviour or not. The increase in probability measure is very difficult to interpret clinically. Secondly, the logistic model contained just main effects and a few two-way interactions between the demographic variables. This is unlikely to be sufficient to convey the complexity of the real world or identify behavioural subgroups within the population. Identifying the unique behaviour of young, black, unmarried, males, for example, would require 4-way interactions in the model. Even with a very large data set, such complex models would be affected by areas of sparse data and overfitting. The third issue arising was that the uncertainty of the estimated probabilities was not carried across to the final Cox model. The estimated probability of having the behaviour variable was treated as an accurate measurement rather than an estimate. Thus, the uncertainty in the resulting hazard ratios was understated.

A hybrid solution, using the OC cases' estimated probabilities of having the behaviour to inform the imputation process rather than as a direct predictor, was also considered. This could address the awkward hazard ratio but not the potential importance of complex demographic subgroups and the associated need for modelling higher-level interactions. Overall, these analyses suggested there was little potential for further development of this model-based method.

7.3.2 **An imputation-based solution**

An alternative, an imputation-based algorithm for augmenting cancer registry data was more promising. Using simulated data based on the actual SEER OC cases, this thesis demonstrated that the method produced point estimates of hazard ratios and relative risks which were quite close to true values, albeit with much wider confidence intervals. However, all the difficulties of this approach have not yet been solved, with age-confounding one issue proving to be difficult to address. Indeed, there are considerable opportunities for improvement at each step in this initial algorithm.

The algorithm is referred to as the I2C2 algorithm; standing for impute, impute, calibrate and correct. As the impute step is repeated, the algorithm consists of 3 main steps, described below.

The first step is to impute values for the missing health behaviour data in the SEER OC cancer registry data set. Again, the BRFSS health survey data was chosen as an external reference to help impute the missing health behaviour, relying upon the demographic variables in common to transfer information from the BRFSS health survey to the SEER cancer registry data set.

It was assumed that the two data sets are completely independent, and few imputation methods can be applied to completely independent data sets. OC cancer cases were matched with demographically similar individuals in the BRFSS health survey. The health survey respondents then donated their behaviour to their matched OC cancer cases. This approach is referred to as cold deck imputation (Nordholt, 1998).

The matching process needs to be improved. All OC cases who had less than two matches in the BRFSS health survey data were excluded and this produced considerable

potential for bias (exclusions were higher in older age groups, males, non-whites, etc). Allowing matching to the next most similar individual may have resulted in lower bias, more efficient use of available data allowing the method to be applied smaller data sets.

Further, the more variables and the more informative the variables used for matching; the more information which could be transferred from the BRFSS health survey to the SEER OC data set. There is evidence (section 5.4 Further exploration of matching) that the imputation applied in this thesis produced an underestimation of the prevalence of some pre-diagnosis health behaviours in OC. This may have led to an underestimation of the association between these pre-diagnosis behaviours on post-diagnosis survival. The desire to maximise the number of demographic variables in common was an important contributor to selecting the SEER cancer registry data set and BRFSS health survey. Some other data sets have fewer variables in common, potentially undermining the success of the I2C2 algorithm. The quality of the information is also important. It has been noted that self-reported behaviours are sometime inaccurate (e.g. (Ward et al., 2016)). There may also be errors in self-reported demographic variables such as race. Minimising such errors will increase the power of the I2C2 algorithm. Finally, the information in the available variables is partly determined by the coarseness of the categories. More detailed categories (such as age in years instead of 5-year age groups and smaller geographic regions than States) would equate to more information but would also decrease the likelihood of exact matches. This reinforces the need to develop criteria for matching across similar demographic groups when exact matches are unavailable.

The calibration step is the measurement of misclassification. In all analyses the behaviour twice followed by cross-tabulation of the imputed values. This can only be applied to dichotomous measures of behaviour and alternate approaches for polytomous or

continuous variables (such as pack-years, alcohol consumption, BMI, metabolic equivalent of tasks (METs) have not been considered. Further, the thesis did not consider whether three or more imputed values could provide a more accurate estimate of misclassification than just two.

An additional approach to validating the misclassification measure would be to survey a random sample of SEER OC cases to obtain actual measures of their pre-diagnosis behaviours. This would allow the estimated misclassification rate produced by cross-tabulating the two imputed behaviours to be validated against the true misclassification rate between the true behaviour and the imputed behaviour. It is important to note however, that surveys of OC survivors would be subject to sources of bias such as survivor bias, response bias, recall bias. A more rigorous validation approach would be to find a cohort study (or record linkage) which included pre-diagnosis behaviours, death of OC diagnosis and date of death. Behaviours could be imputed for the cohort and estimated misclassification rates from two imputed values could be compared with true misclassification between true and imputed behaviour. Unfortunately, attempts to access a suitable validation cohort for OC outcomes have so far been unsuccessful, and future investigation of this approach may need to be conducted on alternate health outcomes.

The final step in the I2C2 algorithm is to correct, which means to incorporate misclassification correction into the analysis of the augmented cancer registry data set. Misclassification correction was demonstrated for age-adjusted relative risk of death at 1-year post diagnosis and Cox regression models. Misclassification correction for logistic and/or log-binary models of death at a post-diagnosis timepoint should also be possible (Spiegelman et al., 2000) using methods like the corrected scores algorithm that were trialled in the thesis in the Cox regression models.

Despite high misclassification rates, the misclassification correction algorithms were successful when applied to simulated data, but only when the behaviour was not too rare. A comprehensive rule of thumb to define "not too rare" has not yet been established, although some initial results from simulations were presented which shown that, given the linkage variables used in this thesis, misclassification correction failed for behaviours with a prevalence less than 5%. This issue of feasibility is very important in deciding whether to pursue further development of the I2C2 algorithm and warrants future research priority.

Although the misclassification correction methods worked well with simplified simulated data sets, adding age-confounding into the data sets created some complex problems. Age-confounding was addressed via age-stratification when estimating relative risks for death within 1-year post-diagnosis, and via age-standardisation when estimating hazard ratios from Cox regression models. However, age-stratification did not work for Cox regression, perhaps because the stratified sample sizes were too small and that age-standardisation results seemed to be incompletely adjusted. Alternative methods for confounder adjustment in misclassification correction warrant further research.

Finally, software code for conducting misclassification adjustment is currently under-developed. Code was specifically developed in this thesis for misclassification correction on relative risk, and misclassification correction in Cox regression relied upon Fortran 77 program. This thesis highlights the need to develop misclassification correction options for logistic, log-binary and Cox regression models in modern statistical software.

7.3.3 **What has been learnt and what is yet to do**

This thesis has proposed new methods to augment cancer registry data sets with health behaviour data when data linkage is not possible, and has begun the process of developing such a method.

The conceptual approach of regarding the health behavioural variables as variables with 100% missing data in the cancer registry data set, and applying missing value methods to add whole variables to data sets has not been undertaken previously.

Findings show that predicting missing values using statistical models is likely to be a very complex process as it is difficult for the model to convey the true complexity of the real world. In contrast, imputation-based approaches may be more achievable.

The I2C2 algorithm was developed and tested. The algorithm involved imputing values for missing health behaviour variable using values donated from an external health behaviour data set, repeating the imputation, calibrating the misclassification by cross-tabulating the two imputed values and finally correcting for the attenuation of results arising from the misclassification by employing appropriate misclassification correction during the final analysis.

The I2C2 algorithm can provide accurate estimates of association under the idealised conditions of large, simulated data sets; albeit it with a certain lack of precision manifested as wide confidence intervals. However, there are difficulties in confounder adjustment (in this case age-adjustment) when using the misclassification correction tools and the algorithm has yet to be fully validated using actual data.

There are data sets available which may be appropriate for a validation study of the I2C2 algorithm, and this remains a priority for future research.

7.4 Potential applications of this research

Combinations of public health programs and government regulation can modify health behaviours associated with cancer outcomes. For example, in Australia the proportion people aged 14 and over who were daily smokers has more than halved in less than 30 years, from 24% in 1991 to 11% in 2019 (Australian Institute of Health and Welfare, 2020b). Also, between 2001 and 2019 the proportion of 18-24 year old Australians abstaining from alcohol increased from 10% to 21% but the proportion of those aged 70 or more abstaining from alcohol decreased from 32% to 28% (Australian Institute of Health and Welfare, 2020b). The prevalence of discretionary physical activity among Canadian adults rose from 20.6% to 41.1% between 1981 and 2000 (Craig et al., 2004). Between 2012 and 2021 recreational Marijuana use has been legalised in 19 US States and two Territories (US News, 2021).

Changing health behaviours have impacts on the future health needs of communities. Preparing for future health service needs requires long-term planning as it takes time to train clinicians and other health professionals, and to develop infrastructure. Accurate projections of future health needs are required to inform planning. This includes projections of how many people will be diagnosed with cancer and how long they will survive.

To predict the future number of cancer survivors requires the current number of cancer survivors (current state), the incidence rate (additions) and the mortality rate (subtractions). With changing health behaviours, the future incidence rate and future mortality rate may not be the same as current rates. Many current projection methods take current trends in incidence and mortality rates and extrapolate these into the future (Maddams, Utley, & Møller, 2012; Mariotto, Robin Yabroff, Shao, Feuer, & Brown, 2011), without any consideration of the underlying causes of changing incidence and mortality; such as the changing rates of health behaviour in the community. The work in this thesis to

document the association between health behaviour in the community and survival times can help how changes in the community's health behaviour might alter survival times and thus inform next generation projection models.

At the individual level, the results in this thesis confirm that pre-diagnosis health behaviour is an important factor in the prognosis of OC. Current prognostic indexes for OC have been criticised (Gupta et al., 2018) for not being particularly effective. But none include behaviour. The results in this thesis should help improve prognostic tools for OC, leading to better lifestyle and treatment choices at the individual level.

Finally, the results in this thesis add additional information to the costs and benefits of public health programs. Not only do programs reducing smoking, for example, decrease the risk of cancer, but they may also increase the survival of those who do get cancer. The results generated in this thesis can assist in quantifying the full value of such public health programs and in so improve cost-benefit analyses and program decision making.

The final application of the results presented in this thesis is to encourage further research. This thesis has shown that information on the effect of pre-diagnosis health behaviour on post-diagnosis survival in OC is fragmented, incomplete and perhaps poor quality. More research is required using more sophisticated behavioural measures and truly prospective cohorts.

The I2C2 algorithm is an approach that is still under development. There are opportunities for improvement in each step (imputation, calibration and misclassification correction) and the results of analyses have not been fully validated using data where pre-diagnosis health behaviour is known for cancer. The algorithm requires large data sets and,

even under the ideal conditions offered by simulation, the results have wide uncertainty and wide confidence intervals.

Despite this, it is proposed that the approach has a role as a relatively quick and cheap alternative to prospective cohort studies and record linkage (retrospective cohort) studies. Prospective cohorts are complex and expensive and take many years to produce new knowledge. Record linkage is an alternative, however there are logistical and resource implications associated with data linkage, the need for identifiable data (and attendant confidentiality issues), and issues of statistical power associated with the size of available data sets for linkage..

Confidentiality is protected by legislation and data custodians and Ethics Committees, as well as researchers, can face legal sanction for breaches (Palamuthusingam, Johnson, Hawley, Pascoe, & Fahim, 2019). As data sets continue to get larger with more linkage between them and interrogative methods become more sophisticated, the risk of breaches in confidentiality grow. In response data protection methods grow more complex. It is now common for the potentially identifying demographic data to be separated from the potentially sensitive clinical variables. A label (master key) created from the demographic data but with no information about the demographics can be used to link the clinical data (NSW Health, 2021). It is also common to maintain linked data within secure workspaces where no data can be exported from the system, only vetted results (Sax Institute, 2021). These confidentiality processes add to the complexity and cost of record linkage studies.

A key advantage of the I2C2 algorithm is that the imputed values added to the cancer registry data set adds almost no additional information about any individual in the data set. Given the very high levels of misclassification, the imputed data is probably incorrect most of the time. Rather than make individuals easier to identify, the incorrect imputed values can

obscure individuals and make it more difficult to identify them. Unlike data linkage, the I2C2 algorithm adds virtually no additional threat to the confidentiality of the cancer cases.

The other advantage of I2C2 over record linkage is that it does not require the same individuals to be recorded in both data sets. Locating cancer cases in other data sets requires linkage algorithms which can add costs and time to projects, and are subject to record linkage errors. The rarer the cancer, the fewer the number of individuals who will be located in the external data sets and as the number of correct linkages decrease the relative proportion of linkage errors increases. So the I2C2 is of particular advantage when the disease is rare.

As well as avoiding matching errors for data linkage and participation biases in prospective cohort studies, the I2C2 algorithm can also somewhat avoid the biases within the external reference data within the process. Cancer registries are generally a census of all cancer cases, subject to very little participation bias. If the I2C2 algorithm can be improved to produce imputed values for all cancer cases, then the augmented data set will be similarly unbiased. Even some of the known biases in the BRFSS health survey will be addressed through the cold deck imputation method. That is, groups which are over-represented in the BRFSS such as white, younger age, etc (Schneider et al., 2012) will be sampled less, and groups who are underrepresented in the BRFSS will be sampled more. However other biases, such as under-participation of ‘at risk’ drinkers, may not be resolved by the imputation process.

This thesis has shown that the I2C2 algorithm is feasible for examining the associations between pre-diagnosis behaviour and post-diagnosis survival with appropriate large data sets already existing in most developed countries. The method could easily be applied to other outcome measures such as the association between health behaviour and stage at diagnosis, other tumour characteristics, or choice of curative versus palliative

treatment. The I2C2 method is also readily applicable to predictors of outcome which could be linked through demographics, such as environmental exposure. There may also be instances when different linking variables could be used: say using tumour characteristics as the linking variables to add additional laboratory variables to the cancer registry data set.

With increasing large scale data collection throughout society there are also likely other potential applications of the I2C2 algorithm outside of augmenting cancer registries which have yet to be identified.

7.5 Conclusion

OC survival does appear to be associated with pre-diagnosis behaviour. For example, pre-diagnosis smoking is associated with decreased survival times in OSCC and exercise may be associated with increased survival as would be expected. However, the literature is still incomplete with unresolved questions relating to the measurement tools, and quality and quantity of evidence.

This thesis has described a new method for augmenting cancer registry data bases with 100% imputed pre-diagnosis health behaviour variables, called the I2C2 algorithm. The algorithm still needs further refinement and validation, but it is proposed that it could become an efficient potential alternative to cohort and record linkage studies in preliminary studies of cancer epidemiology.

Appendix 2 Coding elements

Chapter 2

While using R software throughout the rest of the thesis, for the meta-analyses the 'metan' and 'metafunnel' packages in STATA were used. Formatting forest plots is more straightforward in 'metan' than in the 'metafor' package in R software. The code below was run in STATA version 12 but will work in any recent version.

When fitting meta-analyses, ratio measures were log-transformed to a linear scale and the analyses conducted using linear mixed effects models. That is, the logarithm of (hazard A)/(hazard B) is the linear expression log(hazard A)+log(hazard B). Results are back-transformed into the ratio scale using the 'eform' option in metan. Example STATA code is provided below.

```
/*************************************************************/
/* Read in the comma separated data file containing:         */
/*     author, a text field of author and year               */
/*     hr, the hazard ratio for pre-diagnosis smoking        */
/*     lcl and ucl, the associated 95% confidence interval   */
/*     cancercat, indicating which histological subgroup     */
/*                the HR pertains to.                        */
/*************************************************************/

insheet using "oesoph_smoke2.csv", comma

/*************************************************************/
/* Convert the hazard ratios and confidence intervals        */
/* to effect sizes and associated standard errors.           */
/*************************************************************/

gen log_hr=ln(hr)
gen log_se_ucl=(ln(ucl)-log_hr)/(1.959964)
gen log_se_lcl=(log_hr-ln(lcl))/(1.959964)
gen log_se_hr=(log_se_ucl+log_se_lcl)/2

label variable log_hr "Log(hazard ratio)"
label variable log_se_hr "Std Error log(hazard ratio)"
```

```
/*************************************************************/
/* In this case, retain statistics for 'ever smoke' versus  */
/* 'never smoke' for analysis                                */
/*************************************************************/

keep if (reference=="never" & comparison=="ever")

/*************************************************************/
/* Run the random effect meta-analysis and format the forest*/
/* plot using the metan application for STATA using the     */
/* pre-defined format s1mono.                                */
/* The variable cancercat defines the OSCC and OAC subgroups*/
/*************************************************************/

metan log_hr log_se_hr, random /* completes the meta-analysis*/
  /* all of the rest is setting up the forest plot */
  /* present HRs not log(HRs) */
  eform effect("Hazard Ratio")
  /* add author names */
  label(namevar=author)
  /* format the x-axis */
  xlabel (0.25,0.5, 1, 2,4) force nowarning null(1)
  favours("decreased risk of death" # "increased risk of death")
  /* separate into OSCC and OAC subgroups */
  by(cancercat) nooverall
  /* additional formatting on the forest plot */
  textsize(130) diamopt(lcolor(black)) olineopt(lc(black)
  lp(shortdash) lwidth(thin)) scheme(s1mono)

/*************************************************************/
/* Produce a funnel plot, using the pre-defined              */
/* format s1mono                                             */
/*************************************************************/

metafunnel log_hr log_se_hr, scheme(s1mono)

/*************************************************************/
```

Chapter 3

Having selected the data sets and variables for analysis, data were collected for analysis.

The SEER data set was downloaded from https://seer.cancer.gov/ using the SEER*STAT utility. SEER*STAT is used interactively producing no code to show how the data were extracted. The data were downloaded from SEER*STAT in comma separated values (.csv) format. The following examples of code show how these data were read into and R software ready for analysis.

```
#######################################################################
# Read in the comma separated SEER cancer registry data file.         #
#                                                                     #
# The data file was produced using the data listing function          #
# in the SEER*STAT program: all SEER OC cases in 2001 to 2015.        #
#######################################################################

oesoph <- read.csv("../All_oesophagael_2001_15.csv", header=TRUE)

#######################################################################
# recode text variables into numeric variables                        #
#######################################################################

# sex
summary(oesoph$Sex)
oesoph$sex1<- recode(oesoph$Sex,"'Female'=1; 'Male'=0",
       as.numeric=TRUE, as.factor=FALSE)

# Age in 5 year categories
summary(oesoph$Agecat)
oesoph$agegrp<- recode(oesoph$Agecat,
      "'05-09 years'=2; '10-14 years'=3;'15-19 years'=4;
       '20-24 years'=5; '25-29 years'=6;'30-34 years'=7;
       '35-39 years'=8; '40-44 years'=9; '45-49 years'=10;
       '50-54 years'=11;'55-59 years'=12; '60-64 years'=13;
       '65-69 years'=14; '70-74 years'=15;'75-79 years'=16;
       '80-84 years'=17; '85+ years'=17",
       as.numeric=TRUE, as.factor=FALSE)


# AJCC stage, excluding the missing 2001-2003 data
oesoph$stage<- recode(oesoph$AJCCstage,
              "'I'='I'; 'IIA'='II'; 'IIB'='II'; 'III'='III';
              'IV'='IV';'IVA'='IV';'IVB'='IV';'IVNOS'='IV';
               else='UNK Stage'",
               as.numeric=TRUE, as.factor=FALSE)
```

```
for(i in 1:length(oesoph$stage)) {
  if(oesoph$Yeardiag[i]==2001 | oesoph$Yeardiag[i]==2002 |
     oesoph$Yeardiag[i]==2003) {
       oesoph$stage[i]<-NA
        }
      }

...

#######################################################################
# Save the processed data into a new R-formatted data set.            #
#######################################################################

seeroesoph <- oesoph[,c(7,24:34)]
save(seeroesoph,file="all_oesophagael_2001_2015.RData")

#######################################################################
```

The BRFSS health survey data sets are compiled each year separately. The 15 files for 2001 to 2015 were downloaded from https://www.cdc.gov/brfss/ as column formatted text files. Each file was individually read into R and saved in R format. Only variables relevant to the study were retained. The relevant demographic variables were re-coded to match the corresponding SEER cancer registry variables and the chosen behavioural variables were coded to be dichotomous with '1' representing behaviour present and '0' representing behaviour absent. Example R code is provided below.

```
#######################################################################
# Read in the relevant variables from the BRFSS yearly file.          #
# This example is 2001.                                               #
#######################################################################


df1 <- read.fwf("cdbrfss2001asc.ASC", widths=c(2,-96,1,-10,2,-8,1,
       -18,1,-488,2,-56,10,-29,6,-8,1,-5,6,1,1,-3,1),header=FALSE)


#######################################################################
# Name the variables                                                  #
#######################################################################

library(car)
df1 <- rename(df1,replace=c("V1"="state",V2"="smoke100","V3"="age",
```

```
        "V4"="marital","V5"="sex","V6"="prace","V7"="finalwt",
        "V8"="bmi","V9"="smoker3","V10"="drnkday","V11"="rfbinge",
        "V12"="rfdrhvy","V13"="totinda"))
...

#####################################################################
# Recode the demographic variables to be consistent with SEER      #
# coding.                                                           #
# This example is marital status.                                   #
#####################################################################

df1$marital <- recode(df1$marital, "1=1;2=2;3=3;4=2;5=4;6=1;9=NA")
df1$marital <- factor(df1$marital, levels = c(1,2,3,4),
               labels = c("married","divorced/separated","widowed",
                          "single/never married"))
...

#####################################################################
# Recode the behavioural variables as 0=behaviour absent,          #
# 1=behaviour present.                                              #
# This example shows a  number of smoking status variables.        #
#####################################################################

df1$smoke.daily <- recode(df1$smoker3, "1=1; 9=NA; else=0")
df1$smoke.some <- recode(df1$smoker3, "2=1; 9=NA; else=0")
df1$smoke.ex <- recode(df1$smoker3, "3=1; 9=NA; else=0")
df1$smoke.never <- recode(df1$smoker3, "4=1; 9=NA; else=0")
...

#####################################################################
# Save the processed data into a new R-formatted data set.         #
# (finalwt are the sampling weights calculated by the BRFSS        #
# analysts.)                                                        #
#####################################################################

brfss2001 <-data.frame(df1$state,df1$smoke100,df1$age,df1$marital,
               df1$sex,df1$smoke.daily,df1$smoke.some,df1$smoke.ex,
               df1$smoke.never,df1$race,df1$bmicat,df1$drnkany,
               df1$drnkday,df1$rfbinge,df1$rfdrhvy,
               df1$totinda,df1$finalwt)
save(brfss2001,file="BRFSS2001.RData")

#####################################################################
```

The final step was to combine the 15 yearly files into a single R formatted data set for analysis. While doing so the year was added to each file and added some further coding of demographic variables (to be consistent with the SEER cancer registry data set) and added one computed behaviour variable: smoking with regular drinking. The final step was to

exclude all BRFSS health survey data records that did not match the States or race categories within the SEER cancer registry. Example R code is provided below.

```
#######################################################################
# Read in the previously saved BRFSS data files for each year from #
# 2001 to 2015.                                                       #
#######################################################################

load(".../BRFSS2001.RData")
load(".../BRFSS2002.RData")
...
load(".../BRFSS2015.RData")


year <- rep(2001,dim(brfss2001)[1])
brfss2001 <- data.frame(brfss2001,year)
year <- rep(2002,dim(brfss2002)[1])
brfss2002 <- data.frame(brfss2002,year)

brfsstemp <- rbind(brfss2001,brfss2002,brfss2003,brfss2004,
                   brfss2005,brfss2006,brfss2007,brfss2008,
                   brfss2009,brfss2010,brfss2011,brfss2012)

#######################################################################
# Coding age into age groups to match the SEER cancer registry     #
# variable.                                                           #
# There is slightly different code (omitted) for 2013-2015 as age  #
# was provided in 5-year age groups instead of age in years.        #
#######################################################################

brfsstemp$age_cat <- recode(brfsstemp$age, "0=0; 1:4=1; 5:9=2;
      10:14=3; 15:19=4; 20:24=5; 25:29=6; 30:34=7; 35:39=8;
      40:44=9; 45:49=10; 50:54=11; 55:59=12; 60:64=13; 65:69=14;
      70:74=15; 75:79=16; 80:99=17; else=NA")
brfsstemp <- brfsstemp[(brfsstemp$age_cat>=6),]
brfsstemp <- brfsstemp[,-3]


...
brfss <- rbind(brfsstemp,brfss2013,brfss2014,brfss2015)

#######################################################################
# The only new behavioural measure created was smoking with         #
# regular alcohol consumption.                                        #
# Daily smoking and Some smoking were both coded 0,1 and people     #
# could not be coded as 1 on both. Average drinks per day was       #
# coded numerically. (As there is no natural upper bound, I have    #
# used the exaggerated 80 drinks per day.)                            #
#######################################################################

brfss$drnk1 <- recode(brfss$drnkday, "0:0.999=0; 1:80=1")
```

```
brfss$smoker <- brfss$smoke.daily+brfss$smoke.some
temp1 <- brfss$drnk1+brfss$smoker
brfss$smkalc <- recode(temp1, "2=1; 1=0; 0=0")

#####################################################################
# Exclude BRFSS health survey respondents from non-SEER states and #
# those who selected the 'mixed race' option. Mixed race was not   #
# an option in the SEER cancer registry collection.                 #
#####################################################################

save(brfss,file="BRFSS.RData")


brfss_short <- brfss[brfss$state=="Alaska" |
    brfss$state=="California" | brfss$state=="Connecticut" |
    brfss$state=="Georgia" | brfss$state=="Hawaii" |
    brfss$state=="Iowa" | brfss$state=="Kentucky" |
    brfss$state=="Louisiana" | brfss$state=="Michigan" |
    brfss$state=="New Jersey" | brfss$state=="New Mexico" |
    brfss$state=="Utah" | brfss$state=="Washington",]

brfss_short <- brfss_short[brfss_short$race!=7,]

#####################################################################
# Save the processed data into a new R-formatted data set.         #
#####################################################################

save(brfss_short,file="BRFSS_short.RData")

#####################################################################
```

Chapter 4

The first coding tasks was to take the previously prepared BRFSS health survey data set and fit a logistic regression model for predicting each health behaviour using year, age, sex, marital status, race, State of residence, sex by age interaction, marital status by age and sex by marital status interactions. The following R code example illustrates how this was done.

```
#####################################################################
# Read in the previously prepared BRFSS health survey data file.   #
#####################################################################

load(".../BRFSS_short.RData")
```

```
#######################################################################
# Exclude data records falling into the top and bottom 0.5% of      #
# sample weights.                                                   #
# (Data errors are often found at the ends of the distribution).    #
#######################################################################

quantile(brfss_short$finalwt,c(0.005,0.995))
brfss_short <- brfss_short[brfss_short$finalwt>10.00 &
                           brfss_short$finalwt<7478,]

#######################################################################
# Set up categorical variables as factors prior to model fitting.   #
#######################################################################

brfss_short$racef <- factor(brfss_short$race,labels=c("white",
          "black","Asian pacific","Americian native"))
brfss_short$msf <- factor(brfss_short$married,labels=c("married",
          "divor_sep","widow","single"))
...

#######################################################################
# Fit the logistic regression model which predicts smoking status   #
# (in this example) using year, the demographic variables and       #
# the interactions between sex and age and sex and marital status.  #
#######################################################################

smk.r <- logistf(smoker~statef+yearf+age_cat+sexf+msf+racef+
                 sexf*age_cat+msf*age_cat+msf*sexf,
                 data=brfss_short, pl=TRUE, firth=TRUE,
                 weights=finalwt)

#######################################################################
# Document the predictive ability of the model using area under     #
# the receiver operator curve (AUC ROC).                            #
#######################################################################

library(pROC)
roc.smk.r <- roc(smk.r$y~smk.r$predict, na.rm=TRUE, algorithm=2,
                 auc=TRUE, ci=TRUE)
auc(roc.smk.r)


#######################################################################
# Save the (22) coefficients from the fitted model in a vector.     #
#######################################################################

b.smk.r <- coef(smk.r)
b.smk.r.ci <- confint(smk.r)

#######################################################################
```

Having recorded each fitted model as a vector of coefficients, the next task was to apply this model to each of the SEER cancer registry OC cases estimating their probability of undertaking the corresponding health behaviour. The following code example illustrates how this was done.

```
#########################################################################
# Read in the previously prepared SEER cancer registries data file.#
#########################################################################

load("../all_oesophagael_2001_2015.RData")

#########################################################################
# Three options are separately coded (but very similar):           #
#     1 year pre-diagnosis,                                        #
#     5 years pre-diagnosis, 1 age category younger                #
#     10 years pre-diagnosis, 2 age categories younger             #
# The code below is for the 1 year pre-diagnosis option.           #
#########################################################################

#########################################################################
# Exclude OC cases under 35 years.                                 #
# The 80-84 year and 85+ year age categories were combined in the  #
# SEER cancer registry data as the later years of the BRFSS health #
# survey only provide age categories to 80+ years                  #
#########################################################################

seeroesoph <- seeroesoph[seeroesoph$agegrp>7,]
seeroesoph$agegrp <- recode(seeroesoph$agegrp,"18=17")

#########################################################################
# Apply the linear component of the logistic model to the OC cases #
# using the regression coefficients developed on the BRFSS         #
# health survey data set.                                          #
# In this case the model is for smoking status.                    #
#########################################################################

b <- b.smk.r

for(i in 1:dim(seeroesoph)[1]) {
  seeroesoph$risk[i]<-b[1,1] + b[14,1]*(seeroesoph$Yeardiag[i]-1) +
  b[15,1]*seeroesoph$agegrp[i] + b[16,1]*seeroesoph$sex1[i] +
  b[23,1]*seeroesoph$sex1[i]*seeroesoph$agegrp[i]

  if(seeroesoph$state[i]=="AK") {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[13,1] }
  else if(seeroesoph$state[i]=="CT") {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[2,1] }
  ...
  else if(seeroesoph$state[i]=="WA") {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[12,1] }
```

```
  if(seeroesoph$marital2[i]==0) {
    seeroesoph$risk[i] <- NA }
  else if(seeroesoph$marital2[i]==2) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[17,1] +
       b[24,1]*seeroesoph$agegrp[i] + b[27,1]*seeroesoph$sex1[i] }
  else if(seeroesoph$marital2[i]==3) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[18,1] +
       b[25,1]*seeroesoph$agegrp[i] + b[28,1]*seeroesoph$sex1[i] }
  else if(seeroesoph$marital2[i]==4) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[19,1] +
       b[26,1]*seeroesoph$agegrp[i] + b[29,1]*seeroesoph$sex1[i] }
  if(seeroesoph$race1[i]==0) {
    seeroesoph$risk[i] <- NA }
  else if(seeroesoph$race1[i]==2) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[20,1] }
  else if(seeroesoph$race1[i]==3) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[21,1] }
  else if(seeroesoph$race1[i]==4) {
    seeroesoph$risk[i] <- seeroesoph$risk[i] + b[22,1] }
}

#########################################################################
# Convert the linear component of the logistic model into a             #
# probability.                                                          #
#########################################################################

seeroesoph$prob.smk.r <-
(exp(seeroesoph$risk.smk.r)/(1+exp(seeroesoph$risk.smk.r)))

#########################################################################
```

Having estimated the probability of each behaviour for each SEER OC case, this estimated probability was then used as a predictor of survival time through fitting Cox regression models. Unadjusted and adjusted (for age and disease stage) models were fitted. Subgroup analyses were conducted in OSCC and OAC. The following code example illustrates how this was done.

```
#########################################################################
# This probability has been multiplied by 10 prior to being used        #
# in the Cox model to ensure that hazard ratios are associated          #
# with steps of 0.10 probability.                                       #
# The data frame seer_risk below is equivalent to seeroesoph above.     #
#########################################################################

seer_risk$prob10.smk.r<- seer_risk$prob.smk.r*10
```

```
#######################################################################
# Fit the Cox regression model, with and without adjustment for       #
# the potential confounders age and disease stage.                    #
#######################################################################

library(survival)

p.smk.r <- coxph(Surv(survmonth,status1==1~prob10.smk.r,
                 data=seer_risk)
summary(p.smk.r)
p.smk.r.agestg <- coxph(Surv(survmonth,status1==1)~prob10.smk.r+
                        agegrp+stage, data=seer_risk)
summary(p.smk.r.agestg)


#######################################################################
# Conduct subgroup analysis by histological type. In this case        #
# adenocarcinoma.                                                     #
#######################################################################


SCCAC <- seer_risk[complete.cases(seer_risk$type),]
AC <- seer_risk[SCCAC$type==2,]

ac.p.smk.r <- coxph(Surv(survmonth,status1==1~prob10.smk.r,data=AC)
summary(ac.p.smk.r)
ac.p.smk.r.agestg <- coxph(Surv(survmonth,status1==1)~prob10.smk.r+
                                 agegrp+stage,data=AC)
summary(ac.p.smk.r.agestg)

#######################################################################
```

Chapter 5

The following R code segment explains the process used to match SEER OC cases with randomly selected BRFSS health survey respondents. In this example, a 5-year lag is incorporated from BRFSS health survey to the SEER OC case record. This simulates health behaviour 5-years prior to OC diagnosis.

```
#####################################################################
# Read in the SEER OC data set and adjust category codes and        #
# to ensure compatibility with the BRFSS health survey data.        #
#####################################################################

load(".../all_oesophagael_2001_2015.RData")
...
```

```
#####################################################################
# Given that the earliest available BRFSS data set is 2001 and      #
# given a built-in 5-year lag between health behaviour and          #
# cancer diagnosis, the earliest cancer data set used is 2006.      #
#####################################################################

seeroesoph <- seeroesoph[seeroesoph$Yeardiag>=2006 &
                         seeroesoph$Yeardiag<=2015,]

#####################################################################
# The outcome variable is survival status at 1-year                 #
# coded as alive or dead. Records censored at less than 1-year      #
# are excluded, those censored at 1-year or more are alive, etc     #
#####################################################################

n <- dim(seeroesoph)[1]
seeroesoph$dead1y <- rep(-1,n)
for(t in 1:n) {
  if(seeroesoph$survmonth[t]>=12 ) {
         seeroesoph$dead1y[t]<-0
      }
  else {
    if(seeroesoph$status1[t]==1 & seeroesoph$survmonth[t]<12) {
         seeroesoph$dead1y[t]<-1
      }
    }
  }
seeroesoph <- seeroesoph[seeroesoph$dead1y~=-1,]

#####################################################################
# Each OC case belongs to a particular year by age by sex by        #
# martial status by race by State cluster. This adds a cluster      #
# label to each OC case.                                            #
#####################################################################

seeroesoph$Key <- paste(seeroesoph$year,seeroesoph$sex,
                        seeroesoph$agegrp,seeroesoph$marital2,
                        seeroesoph$race,seeroesoph$state, sep="~")

#####################################################################
# Some demographic clusters contain more than one OC case.          #
# Here the clusters which contain at least one OC case are          #
# listed.                                                           #
#####################################################################

ks <- unique(c(seeroesoph$Key))

#####################################################################
# Now read in the BRFSS health survey data, make variables          #
# compatible with SEER exclude data records which will not match    #
# due to the 5-year time lag, missing data, etc.#
#####################################################################

load(".../BRFSS_short.RData")
...
```

```
brfss_short <- brfss_short[brfss_short$year>=2001 &
                           brfss_short$year<=2010,]
brfss_short <- brfss_short[complete.cases(
                           brfss_short[,c(1,2,3,4,9,10)]),]
...

#####################################################################
# Specify the OC cases' year and age group which each BRFSS         #
# survey data record matches with and add the corresponding         #
# cluster label.                                                    #
#####################################################################

brfss_short$year <- recode(brfss_short$year,"2001=2006;
                           2002=2007;2003=2008;2004=2009;2005=2010;
                           2006=2011;2007=2012;2008=2013;2009=2014;
                           2010=2015;2011=2016;2012=2017;2013=2018;
                           2014=2019;2015=2020")
brfss_short$agegrp<- recode(brfss_short$agegrp,
"6=7;7=8;8=9;9=10;10=11;11=12;12=13;13=14;14=15;15=16;16=17;17=17")

brfss_short$Key <- paste(brfss_short$year,brfss_short$sex,
                         brfss_short$agegrp,brfss_short$marital2,
                         brfss_short$race,brfss_short$state,sep="~")


#####################################################################
# Exclude BRFSS health survey data from demographic clusters        #
# where there are no OC cases and vice-versa.                       #
#####################################################################

brfss_short <- brfss_short[brfss_short$Key%in%ks,]
ks_brfss <- unique(c(brfss_short$Key))
seeroesoph <- seeroesoph[seeroesoph$Key%in%ks_brfss,]
ks <- unique(c(seeroesoph$Key))

#####################################################################
# To avoid cumulating missing values across the health behaviour    #
# variables, each health behaviour was analysed separately.         #
#####################################################################

brfss_short$ID <- seq.int(nrow(brfss_short))

brfss_smoker <-brfss_short[complete.cases(brfss_short$smoker),
                           c(1,2,3,4,7,8,9,13,14)]
brfss_binge <- brfss_short[complete.cases(brfss_short$rfbinge),
                           c(1,2,3,4,5,7,8,13,14)]
...
rm(brfss_short)

#####################################################################
# For each OC case, randomly select (without replacement) two       #
# BRFSS data records from the same demographic cluster. Record      #
# the behaviour of both BRFSS health survey records onto the        #
# OC case's data record.                                            #
#####################################################################
```

```
## start a new data frame to record the 100 data sets

loop <- 0; year<-0; status1<-0; survmonth<-0; sex<-0; agegrp<-0;
marital2<-0; race<-0; curative<-0; stage<-0; state<-0; type<-0;
dead1y <- 3; Key <-"2005~0~17~3~1~CA";
smoker1<-3; smoker2<-3; binge1<-3; binge2<-3; ...

samples <- data.frame(year,status1,survmonth,sex,agegrp,marital2,
                      race,curative,stage,state,type,dead1y,Key,
                      smoker1,smoker2,binge1,binge2,...,loop)


## repeat the random selection 100 times, to create 100 data sets

for(loop in 1:100) {

## start a new data frame to record an individual data set

year<-0; status1<-0; survmonth<-0; sex<-0; agegrp<-0; marital2<-0;
race<-0; curative<-0; stage<-0; state<-0; type<-0; dead1y <- 3
Key <-"2005~0~17~3~1~CA";
smoker1<-3; smoker2<-3;binge1<-3; binge2<-3; ...

sample <- data.frame(year,status1,survmonth,sex,agegrp,marital2,
                  race, curative,stage,state,type,dead1y,Key,
                  smoker1,smoker2,binge1,binge2,...)

## address each demographic cluster one at a time

for(k in ks) {

## pull out all of the BRFSS health survey records in that cluster

  miniseer <- seeroesoph[k==seeroesoph$Key,]
  mini_smoker <- brfss_smoker[k==brfss_smoker$Key,]
  mini_binge <- brfss_binge[k==brfss_binge$Key,]
  ...

  n_seer <-dim(miniseer)[1]
  n_smoker <- dim(mini_smoker)[1]
  n_binge <- dim(mini_binge)[1]
  ...

## randomise the order of BRFSS records within the cluster

  mini_smoker$order <- runif(n_smoker,0,1)
  mini_binge$order <- runif(n_binge,0,1)
  ...

  mini_smoker2 <- mini_smoker[order(mini_smoker$order),]
  mini_binge2 <- mini_binge[order(mini_binge$order),]
  ...

## If there is an odd number of number of BRFSS records,
## ignore the last. It can never be part of the two matches.
```

```
  if (n_smoker %% 2 !=0) {n_smoker <- n_smoker-1}

## Write two randomly ordered BRFSS health behaviours side-by-side
## on a single data line.

  if(n_smoker>=2) {
    n2_smoker <- n_smoker/2
    mini_smoker3a <- mini_smoker2[1:n2_smoker,7]
    mini_smoker3b <- mini_smoker2[(n2_smoker+1):n_smoker,7]
    mini_smoker3 <- data.frame(cbind(mini_smoker3a,mini_smoker3b))
    mini_smoker3 <- rename(mini_smoker3,c("mini_smoker3a"="smoker1",
                           "mini_smoker3b"="smoker2"))
  }

  if (n_binge %% 2 !=0) {n_binge <- n_binge-1}
  if(n_binge>=2) {
    n2_binge <- trunc((n_binge+1)/2)
    mini_binge3a <- mini_binge2[1:n2_binge,5]
    mini_binge3b <- mini_binge2[(n2_binge+1):n_binge,5]
    mini_binge3 <- data.frame(cbind(mini_binge3a,mini_binge3b))
    mini_binge3 <- rename(mini_binge3,
c("mini_binge3a"="binge1","mini_binge3b"="binge2"))
  }
...
 }

## This macro adds each pair of health behaviours to next OC
## data record from the same cluster, (checking for instances
## where there are fewer pairs of health behaviours than OC cases).

combine.df <- function(x,y) {
  rows.x <- nrow(x)
  rows.y <- nrow(y)
  if (rows.x > rows.y) {
    diff <- rows.x - rows.y
    df.na <- matrix(NA, diff, ncol(y))
    colnames(df.na) <- colnames(y)
    cbind(x, rbind(y, df.na))
       }
  else {
    diff <- rows.y - rows.x
    df.na <- matrix(NA, diff, ncol(x))
    colnames(df.na) <- colnames(x)
    cbind(rbind(x, df.na), y)
       }
  }

## Run the macro. Notice that it is cumulative: the data frame
## from the previous iteration is fed back into the macro.
## On the first iteration, the two smoking values are attached to
## the OC case record, on the second iteration the two binge
## drinking value are added, etc for all 6 behaviours.

  add_smoker <- combine.df(miniseer,mini_smoker3)
  add_binge <- combine.df(add_smoker,mini_binge3)
```

```
  ...

## Any data records created beyond the available OC cases will
## have missing OC case data, including year. This trims any
## behaviour only data records from the data set.

  sample1 <- add_smkalc[complete.cases(add_smkalc$year),]

## Add the data from this demographic cluster to the data frame

  sample <- rbind(sample,sample1)
  }

## Number this data frame and add it (minus the first dummy line)
## to the 100 samples data frame

sample <- sample[-1,]
sample$loop  <- rep(loop,dim(sample)[1])
samples <- rbind(samples,sample)
}

samples <- samples[-1,]

## Save the 100 repetitions of the data set with imputed behaviours.

save(samples,file=".../sample_5yearlag.RData")

#########################################################################
```

Having created 100 data sets with two random cold deck imputed values for each health behaviour, these two imputed values were cross-tabulated and the results recorded for use in the eventual estimation of the relative risks and associated 95% confidence intervals. Specifically, for each of 6 behaviours in each of the 100 data sets the cell counts from the cross-tabulation (values E, F, G and H values defined in Table 5.5) and the correlation coefficient were recorded. The (first) imputed behaviour was cross-tabulated with 1-year survival and recorded the cell counts (the values I, J, K and L defined in Table 5.12). This level of detail formed a basis from which a variety of intermediatory and diagnostic statistics could be created and tested.

The calculation of age-adjusted relative risk using Cochran-Mantel-Haenszel method has been added into the same loop. This attenuated relative risk, uncorrected for misclassification, is also an input into the final analyses.

```
#######################################################################
# Load the 100 data sets with random cold deck imputed behaviours  #
#######################################################################

load("../sample_5yearlag.RData")


#######################################################################
# These analyses were repeated for various subgroups - this is how #
# squamous cell carcinomas would have been selected.                #
#######################################################################

#samples <- samples[complete.cases(samples$type),]
#samples <- samples[samples$type==1,]

#######################################################################
# For each of the 100 data sets for each of the 6 behaviours       #
# record cell frequencies from the cross-tabulation of the two     #
# imputed values for the behaviour (E, F, G, H) and associated     #
# correlation. Also record the cell frequencies from the cross-    #
# tabulation of imputed behaviour against observed survival status #
# at 1-year (I, J, K, L) for each of the 6 behaviours.             #
#######################################################################

## Create data frame for recording the results

loop<-0; age<--1;
n_sm<--1;E_sm<--1;F_sm<--1;G_sm<--1;H_sm<--1;I_sm<--1;J_sm<--
1;K_sm<--1;L_sm<--1;
n_bi<--1;E_bi<--1;F_bi<--1;G_bi<--1;H_bi<--1;I_bi<--1;J_bi<--
1;K_bi<--1;L_bi<--1;
...
cor_sm<--1;cor_bi<--1; ...
rrcmh_sm<--1;rrcmh_bi<--1;rrcmh_dr<--1; ...
rrcmh2_sm<--1;rrcmh2_bi<--1; ...

results <- data.frame(loop,age,
                      n_sm,E_sm,F_sm,G_sm,H_sm,I_sm,J_sm,K_sm,L_sm,
                      n_bi,E_bi,F_bi,G_bi,H_bi,I_bi,J_bi,K_bi,L_bi,
                      ...,cor_sm,cor_bi,...,rrcmh_sm,rrcmh_bi,...,
                      rrcmh2_sm,rrcmh2_bi,...)
```

```
## This function returns all the cell counts and correlations
## for a single data set.

extract.results <- function(z) {

  # smoker
  rrm_tab_sm<-table(z$smoker1,z$dead1y)
  I_sm <- rrm_tab_sm[2,2]
  J_sm <- rrm_tab_sm[2,1]
  K_sm <- rrm_tab_sm[1,2]
  L_sm <- rrm_tab_sm[1,1]
  n_sm <- sum(rrm_tab_sm)

  rel_tab_sm <- table(z$smoker1,z$smoker2)
  E_sm <- rel_tab_sm[2,2]
  F_sm <- rel_tab_sm[2,1]
  G_sm <- rel_tab_sm[1,2]
  H_sm <- rel_tab_sm[1,1]

  cor_sm <- cor.test(z$smoker1,z$smoker2)$estimate

# binge
  rrm_tab_bi<-table(z$binge1,z$dead1y)
  I_bi <- rrm_tab_bi[2,2]
  J_bi <- rrm_tab_bi[2,1]
  K_bi <- rrm_tab_bi[1,2]
  L_bi <- rrm_tab_bi[1,1]
  n_bi <- sum(rrm_tab_bi)

  rel_tab_bi <- table(z$binge1,z$binge2)
  E_bi <- rel_tab_bi[2,2]
  F_bi <- rel_tab_bi[2,1]
  G_bi <- rel_tab_bi[1,2]
  H_bi <- rel_tab_bi[1,1]

  cor_bi <- cor.test(z$binge1,z$binge2)$estimate

  ...

  loop.result <- data.frame(loop, age,
                      n_sm,E_sm,F_sm,G_sm,H_sm,I_sm,J_sm,K_sm,L_sm,
                      n_bi,E_bi,F_bi,G_bi,H_bi,I_bi,J_bi,K_bi,L_bi,
                      ...,cor_sm,cor_bi,...)
  return(loop.result)
}

## All 100 data sets are stored consecutively in the same data
## frame. 'looplen' is the number of data records per data set.

looplen <- dim(samples)[1]/100

## For each of the 100 data sets ...

for(i in 1:100) {
  start <- (i-1)*looplen+1
  end <- (start-1)+looplen
```

```
  sample <- samples[start:end,]

## extract the required cell counts and correlations.
  age <- 0
  loop.result <- extract.results(sample)


## Then stratify the data set by age ...

sample8<- sample[sample$agegrp==8,]
sample9 <- sample[sample$agegrp==9,]
...

## and for each age, record the numeric elements for the numerator
## and denominator of the Cochran-Mantel-Haenszel relative risk.

sm_rr_tab8 <- table(sample8$smoker1,sample8$dead1y);
n8_sm<-sum(sm_rr_tab8)
top8_sm<-0; bot8_sm <-0
if(dim(sm_rr_tab8)[1]==2) {
  top8_sm <- sm_rr_tab8[2,2]*(sm_rr_tab8[1,2]+sm_rr_tab8[1,1])
  bot8_sm <- sm_rr_tab8[1,2]*(sm_rr_tab8[2,2]+sm_rr_tab8[2,1])
}
sm_rr_tab9 <- table(sample9$smoker1,sample9$dead1y); n9_sm<-
sum(sm_rr_tab9)
top9_sm <- sm_rr_tab9[2,2]*(sm_rr_tab9[1,2]+sm_rr_tab9[1,1])
bot9_sm <- sm_rr_tab9[1,2]*(sm_rr_tab9[2,2]+sm_rr_tab9[2,1])
...

## Calculate the Cochran-Mantel-Haenszel relative risk for this
## data set.

rrcmh2_sm <-top8_sm/n8_sm+top9_sm/n9_sm+...+top17_sm/n17_sm)/
           (bot8_sm/n8_sm+bot9_sm/n9_sm+...+bot17_sm/n17_sm)

## Store all of the results for this data set and repeat.

loop.result1 <- cbind(loop.result,rrcmh_sm,rrcmh_bi,...,
                      rrcmh2_sm,rrcmh2_bi,...)
results <- rbind(results, loop.result1)
}

## Save the data frame containing the results from each of
## the 100 data sets.

results <- results[-1,]
save(results,file=".../samp_result_5yearlag.RData")

#####################################################################
```

The cross-tabulations, correlations and attenuated relative risks can now be processed to estimate p_c the proportion of OC cases who smoked, p_p the proportion of the wider population who smoked (assumed equal to p_c), m the improvement in matching derived from the matching variables and the association between true and imputed health behaviours (summarised as expected cell sizes A, B, C and D) described in Table 5.11. The misclassification adjusted relative risk can be estimated from the observed relative risk from imputed behaviours using

$$RR_T = \frac{C - Cp_i - p_i DRR_i}{p_i BRR_i - A + Ap_i}$$

derived immediately after Table 5.12. Finally, the excess number of OC cases with correctly imputed behaviour codes (such as true smokers imputed to be smokers and true non-smoker imputed to be non-smokers) relative to the number expected by chance alone is calculated as a measure of the amount of information conveyed though the imputation process.

Results are presented as medians and empirical 95% confidence intervals (0.025 and 0.975 percentiles) across the 100 data sets.

```
#####################################################################
# Calculate intermediate statistics, including the relative         #
# risk from the imputed behaviour                                   #
#####################################################################

results$pp_sm <- ((results$E_sm+results$F_sm)/results$n_sm +
                  (results$E_sm+results$G_sm)/results$n_sm)/2
results$pc_sm <- results$pp_sm
results$ms_sm <- sqrt((results$E_sm/results$n_sm-results$pp_sm^2)/
                  (results$pp_sm*(1-results$pp_sm)))
excess_sm <- results$n_sm*((results$pc_sm*(1-results$pp_sm))*
             results$cor_sm)
results$RR1_sm <- (results$I_sm/(results$I_sm+results$J_sm))/
                  (results$K_sm/(results$K_sm+results$L_sm))


#####################################################################
# Estimate the true relative risk using the relative risk           #
# from the imputed behaviours.                                      #
#####################################################################

results$A1_sm <- results$n_sm*(results$pc_sm*results$pp_sm+
```

```
                (results$pc_sm*(1-results$pp_sm))*results$cor_sm)
results$B1_sm <- results$n_sm*(results$pc_sm*(1-results$pp_sm)-
                (results$pc_sm*(1-results$pp_sm))*results$cor_sm)
results$C1_sm <- results$n_sm*(results$pc_sm*(1-results$pp_sm)-
                (results$pc_sm*(1-results$pp_sm))*results$cor_sm)
results$D1_sm <- results$n_sm*((1-results$pc_sm)*(1-results$pp_sm)+
                (results$pc_sm*(1-results$pp_sm))*results$cor_sm)

results$RRt_sm3 <- (results$C1_sm-results$C1_sm*results$pp_sm-
                    results$pp_sm*results$D1_sm*results$RR1_sm)/
                   (results$pp_sm*results$B1_sm*results$RR1_sm-
                    results$A1_sm+results$pp_sm*results$A1_sm)

## Repeat for the age-adjusted relative risk

results$RRt_sm_age_adj3<-(results$C1_sm-results$C1_sm*results$pp_sm-
                     results$pp_sm*results$D1_sm*results$rrcmh2_sm)/
                    (results$pp_sm*results$B1_sm*results$rrcmh2_sm-
                     results$A1_sm+results$pp_sm*results$A1_sm)

#####################################################################
# Present the results.                                              #
#####################################################################

## Count how often estimate of the true relative risk was valid
## (i.e. positive).

RRt_sm <- results$RRt_sm3[results$RRt_sm3>0]
length(RRt_sm)

## For combined age ...

result_all <- results[results$age==0,]

## present the medians and empirical 95% confidence intervals.

quantile(result_all$n_sm,c(0.025,0.5,0.975))
quantile(result_all$pp_sm,c(0.025,0.5,0.975))
quantile(result_all$cor_sm,c(0.025,0.5,0.975))
quantile(excess_sm,c(0.025,0.5,0.975))
quantile(result_all$RR1_sm,c(0.025,0.5,0.975))
quantile(RRt_sm,c(0.025,0.5,0.975))
quantile(result_all$rrcmh2_sm,c(0.025,0.5,0.975))
quantile(RRt_sm_age_adj,c(0.025,0.5,0.975))

#####################################################################
```

The final coding step was to create simulated survival times corresponding to relative risks of 2.0, 1.5, 1.25, 1.0, 0.8, 0.67 and 0.5. Starting with the 100 samples previously generated, the first imputed behaviour was designated as the true behaviour and the second as

the imputation. The methods in Section 5.3.9 above were then used to randomly generate the seven 1-year survival statuses for each of the 6 behaviours.

Please note the coded formula is mathematically identical to the formula in Section 5.3.9.

$$c' = \frac{(B+D)}{1 + RR_t((A+B)/(C+D))}$$

$$= \frac{(B+D)}{\big((C+D) + RR_t(A+B)\big)/(C+D)}$$

$$= \frac{(A+C)(C+D)}{RR_t(A+B) + (C+D)}$$

```
#########################################################################
# Set up a data frame to record the results                             #
#########################################################################

...

#########################################################################
# Determine how many data records to add into each simulated data  #
# set.                                                                  #
#########################################################################

looplen <- dim(samples)[1]/100

## For each of the 100 data sets ...

for(i in 1:100) {

  start <- (i-1)*looplen+1
  end <- (start-1)+looplen
  sample <- samples[start:end,]

#########################################################################
# Designate the first imputed value of each behaviour as the true  #
# value of that behaviour then randomly create 7 survival statuses #
# for each of the 6 behaviours corresponding to true relative      #
# risks of 2.0, 1.5, 1.25, 1.0, 0.8, 0.67 and 0.5 using the method #
# derived in Section 5.3.9.                                             #
#########################################################################

  ...
```

```
## Record the proportion of designated smokers and non-smokers in
## the data set, A+B and C+D.

  wtab_sm <- table(sample$smoker1,sample$dead1y)
  wtab_ab_sm <- wtab_sm[2,1]+wtab_sm[2,2]
  wtab_cd_sm <- wtab_sm[1,2]+wtab_sm[1,1]

## Set the true relative risk to 0.5 and calculate the expected
## proportion of deaths among the smokers and among the non-smokers

  sim_rr <- 0.50
  mult50_sm <- sim_rr*(wtab_ab_sm/wtab_cd_sm)
  wtab50_c_sm <- (wtab_sm[2,2]+wtab_sm[1,2])/(1+mult50_sm)
  wtab50_a_sm <- wtab_sm[2,2]+wtab_sm[1,2]-wtab50_c_sm
  sim_prob_died50_sm <- wtab50_a_sm/wtab_ab_sm
  sim_prob_died50_nonsm <- wtab50_c_sm/wtab_cd_sm

## Repeat for all other relative risks and all other behaviours

  sim_rr <- 0.66
  mult66_sm <- sim_rr*(wtab_ab_sm/wtab_cd_sm)
  wtab66_c_sm <- (wtab_sm[2,2]+wtab_sm[1,2])/(1+mult66_sm)
  wtab66_a_sm <- wtab_sm[2,2]+wtab_sm[1,2]-wtab66_c_sm
  sim_prob_died66_sm <- wtab66_a_sm/wtab_ab_sm
  sim_prob_died66_nonsm <- wtab66_c_sm/wtab_cd_sm
  ...

## Set up vectors to record the imputed survival status under
## each of the 7 relative risks across each of the 6 behaviours.

  sample$sim_died1y50_sm <- rep(-1,looplen)
  sample$sim_died1y66_sm <- rep(-1,looplen)
  ...

## Randomly allocate smokers 1-year survival status and non-smokers
## 1-year survival status to produce the required relative risk

  for(w in 1:looplen) {
    if(!is.na(sample$smoker1[w])) {
      if(sample$smoker1[w]==1 & sim_prob_died50_sm>0 &
                  sim_prob_died50_sm<1)
         sample$sim_died1y50_sm[w]<-
           sample(c(0,1),1,replace=TRUE,p=c(1-sim_prob_died50_sm,
                  sim_prob_died50_sm))
      if(sample$smoker1[w]==0 & sim_prob_died50_nonsm>0 &
                  sim_prob_died50_nonsm<1)
         sample$sim_died1y50_sm[w]<-
           sample(c(0,1),1,replace=TRUE,p=c(1-sim_prob_died50_nonsm,
                 sim_prob_died50_nonsm))
      if(sample$smoker1[w]==1 & sim_prob_died66_sm>0 &
                  sim_prob_died66_sm<1)
         sample$sim_died1y66_sm[w]<-
          sample(c(0,1),1,replace=TRUE,p=c(1-sim_prob_died66_sm,
                  sim_prob_died66_sm))
      if(sample$smoker1[w]==0 & sim_prob_died66_nonsm>0 &
                  sim_prob_died66_nonsm<1)
```

```
          sample$sim_died1y66_sm[w]<-
           sample(c(0,1),1,replace=TRUE,p=c(1-sim_prob_died66_nonsm,
                  sim_prob_died66_nonsm))
      ...

    }
  }

  sample$sim_died1y50_sm <- recode(sample$sim_died1y50_sm,"-1=NA")
  sample$sim_died1y66_sm <- recode(sample$sim_died1y66_sm,"-1=NA")
  ...
...

#####################################################################
```

As with the analyses above, before this loop was closed, the two simulated behaviours were cross-tabulated and also behaviour against 1-year survival status were cross-tabulated and cell frequencies were saved for analysis.

The final analyses, presented in 'Section 5.4 Further exploration of matching' were to explore what might happen if the imputed proportion of OC cases who were pre-diagnosis smokers, were an underestimate of the true proportion. The R code for this simulation is presented below.

```
#####################################################################
# Explore what might happen if the proportion of pre-diagnosis   #
# smokers in OC is an underestimate of the true proportion       #
#####################################################################

set.seed(1111)

## Create a storage device for results from the 100 iterations
i<--1; p_t_sim<--1; RR_t<--1; p_i_sim<--1; phi_t<--1;
phi_i<--1; RR_i<--1; RR_corrected<--1; RR_approx <--1

res <- data.frame(cbind(i,p_t_sim,RR_t,p_i_sim,phi_t,phi_i,RR_i,
                        RR_corrected,RR_approx)

## Simulate 100 data sets
for(i in 1:100) {

#####################################################################
# Step 1 Set the population parameters.                             #
#####################################################################
```

```
## set the sample size
n <- 27835

## set the probability of dying with 1 year post diagnosis
prob_dead <- 14927/27835

## set the 'true' proportion of smokers in OC cases
p_t <- 0.318

## set the 'imputed' proportion of smokers in OC cases
p_i <- 0.159

#####################################################################
# Step 2 Randomly allocate 'true' pre-diagnosis smoking status   #
#        and 'true' survival status, with rr=2.0.                #
#####################################################################

rr <- 2

## Apply the formula in Section 5.3.9 to determine the
## proportion in each cell of the relative risk table,
## for the given RR=2.
c_t <- (prob_dead*(1-p_t))/(rr*p_t+(1-p_t))
a_t <- prob_dead-c_t
b_t <- p_t-a_t
d_t <- (1-p_t)-c_t

# Draw a random sample with the desired proportions
sm1 <- sample(1:4, n, replace=TRUE,prob=c(a_t,b_t,c_t,d_t))

## Code the 'true' smoking status (smoking1) and survival
## status (dead) of individuals in each cell of the relative
## risk table.
smoker1 <- rep(-1,n)
dead <- rep(-1,n)

for(i in 1:n) {
if(sm1[i]==1) {smoker1[i]<-1; dead[i]<-1}
if(sm1[i]==2) {smoker1[i]<-1; dead[i]<-0}
if(sm1[i]==3) {smoker1[i]<-0; dead[i]<-1}
if(sm1[i]==4) {smoker1[i]<-0; dead[i]<-0}
}

## Confirm the proportion of smokers and relative risk
tab_t <- table(smoker1,dead); tab_t
p_t_sim <- (tab_t[2,2]+tab_t[2,1])/sum(tab_t); p_t_sim
RR_t <- (tab_t[2,2]/(tab_t[2,1]+tab_t[2,2]))/(tab_t[1,2]/
        (tab_t[1,1]+tab_t[1,2])); RR_t

#####################################################################
# Step 3: Randomly generate 'imputed' smoking status (smoking2)  #
#         with prevalence 0.159.                                    #
#####################################################################

p_i <- 0.159
```

```
## draw the sample
smoker2 <- sample(c(0,1), n, replace=TRUE,prob=c((1-p_i), p_i))

## Confirm the prevalence
table(smoker2)
prop.table(table(smoker2))

#####################################################################
# Step 4: Create a relationship between the 'imputed' smoking    #
#         status and 'real' smoking status.                         #
#####################################################################

## Combine the variables into a data set.
comb <- data.frame(cbind(dead,smoker1,smoker2))

## set the number of matches beyond chance, c
c<-262

## In first c smoker/non-smoker divergent pairs and the
## first c non-smoker/smoker divergent pairs, recode 'imputed'
## smoking status to match 'true' smoking status
count_sm<-0; count_ns<-0

for(i in 1:n) {
  if(comb$smoker1[i]==1 & comb$smoker2[i]==0 & count_sm<c) {
    comb$smoker2[i]<-1; count_sm<-count_sm+1
    }
  if(comb$smoker1[i]==0 & comb$smoker2[i]==1 & count_ns<c) {
    comb$smoker2[i]<-0; count_ns<-count_ns+1
    }
}

## confirm the matching
a_tab <- table(comb$smoker1,comb$smoker2); a_tab
a_tab_p <- a_tab/sum(a_tab)

## confirm the simulated 'true' smoking status
p_t_sim <- (a_tab[2,1]+a_tab[2,2])/sum(a_tab); p_sim_t

## confirm the simulated 'imputed' smoking status
p_i_sim <- (a_tab[1,2]+a_tab[2,2])/sum(a_tab); p_sim_i

## check the correlation between 'true' and 'imputed'
## smoking status assuming p_t>p_i
phi_t <- c/(n*p_i_sim*(1-p_t_sim)); phi_t

## check the correlation between 'true' and 'imputed'
## smoking status assuming p_t=p_i
phi_i <- c/(n*p_i_sim*(1-p_i_sim)); phi_i


#####################################################################
# Step 5: Estimate relative risk using 'imputed' smoking status  #
#####################################################################
```

```
## create the relative risk table
tab_i <- table(comb$smoker2,comb$dead)
tab_i

## calculate the attenuated estimate of relative risk,
## uncorrected for the misclassification errors
RR_i <- (tab_i[2,2]/(tab_i[2,1]+tab_i[2,2]))/(tab_i[1,2]/
        (tab_i[1,1]+tab_i[1,2]))
RR_i

## estimate of relative risk assuming p_t>p_i
RR_corrected <- 1 - ((RR_i-1)/((RR_i-1)*p_i_sim*((1-p_t_sim)/
                    (1-p_i_sim))*(1-phi_t)-phi_t))
RR_corrected

## estimate of relative risk assuming p_t=p_i
RR_approx<- 1 - ((RR_i-1)/((RR_i-1)*p_i_sim*(1-phi_i)-phi_i))
RR_approx

## store the results of this iteration
loop_res <- data.frame(cbind(i,p_t_sim,RR_t,p_i_sim,phi_t,phi_i,
                             RR_i,RR_corrected,RR_approx))
res <- rbind(res,loop_res)
}

## read the median results and 95% empirical confidence intervals
## from the 100 iterations
res <- res[-1,]

quantile(res$p_i_sim,c(0.025,0.5,0.975))
quantile(res$p_t_sim,c(0.025,0.5,0.975))
quantile(res$phi_i,c(0.025,0.5,0.975))
quantile(res$phi_t,c(0.025,0.5,0.975))
quantile(res$RR_corrected,c(0.025,0.5,0.975))
quantile(res$RR_approx,c(0.025,0.5,0.975))

#####################################################################
```

Chapter 6

The code for data input, imputation and misclassification measurements have been reviewed in previous Chapters. The novel coding elements in this Chapter include simulating the survival time data, running the corrected score software and producing lattice plots.

The first section of R code provides an example of how the simulated data for smoking status were produced. The 100 simulated data sets were constructed to have the following similarities with the uncensored cases from the SEER OC data set:

a) the same total number of cases as in the OC cases who died
b) the same overall rate of misclassification (smokers classified as non-smokers and non-smokers classified as smokers) as in the OC cases who died
c) the same number of cases within each age group as in the OC cases who died
d) the same proportion of smokers within each age group as in the OC cases who died
e) the same misclassification rate within each age strata as in the OC cases who died.

Survival times were also simulated to have a distribution as similar as possible as the distribution of survival times in the OC cases who died. Preliminary investigations were conducted to identify the shape of the distribution in the OC cases who died. Two measures of a distribution's shape are skewness and kurtosis; the third and fourth standardized moments of the random variable. The skewness and kurtosis of the observed distribution of survival times were compared against a range of standard distributions (normal, uniform, exponential, logistic, beta, lognormal and gamma) using the skewness-kurtosis plot generated by the `descdist()` command in the `fitdistrplus` library in R. The beta distribution was the best fit but is rarely used in survival analysis. None of the other distributions appeared satisfactory, perhaps because the removal of the censored observations produced artificial changes to the distribution. The Weibull distribution was used as it is more commonly used in survival analysis. The `fitdist()` command was used to identify the shape and scale parameters of the Weibull model which fitted the observed data most closely then modified

these to better align the quartiles of the model with the quartiles in the observed OC survival times and to avoid generating any unreasonably long survival times.

Weibull survival times with specific hazard ratios were simulated using the formula from Table II in Bender et al (2005). For each data set 14 sets of survival times were generated corresponding to hazard ratios of 2.0, 1.5, 1.25, 1.0, 0.8, 0.67 and 0.50 first in the absence of association between age and survival time and then the in the presence of age-confounding.

```
#####################################################################
# Start with the 100 copies of the SEER OC data set which already  #
# contain the imputed health behaviours on each data record.       #
#####################################################################

load(".../sample_5yearlag_inc2015.RData")

#####################################################################
# The simulations are only be performed for smoking status.        #
#####################################################################

samples <- samples[complete.cases(samples$smoker1),]
...

#####################################################################
# The simulations are only performed for uncensored OC cases.      #
#####################################################################

samples <- samples[samples$status1==1,]

#####################################################################
# Record the key characteristics of the OC data set:               #
#   the sample size, the misclassification rate (estimated by      #
#   the agreement between the two imputed value of smoking status),#
#   the proportion within each age group, the proportion of smokers#
#   in each of age group and ....                                  #
#####################################################################

samp_len <- dim(samples)[1]/100
agree_sm <- prop.table(table(samples$smoker1,samples$smoker2),1)
p_age_sm <- prop.table(table(samples$agegrp))
p_sm <- prop.table(table(samples$smoker1,samples$agegrp))
p_sm2 <- as.vector(p_sm)
```

```
#####################################################################
# ... the misclassification rate within each of age strata.         #
#####################################################################

samp_age8 <-samples[samples$agegrp==8,]
agree_sm_age8 <- prop.table(table(samp_age8$smoker1,
                            samp_age8$smoker2),1)
samp_age9 <-samples[samples$agegrp==9,]
agree_sm_age9 <- prop.table(table(samp_age9$smoker1,
                            samp_age9$smoker2),1)
...



#####################################################################
# Create 100 simulated data sets, initially just containing the     #
# data set number, left censoring of survival (all 0), age group,   #
# 'true' smoking status (smoker1) and 'imputed' smoking status      #
# (smoker2).                                                        #
#####################################################################

## Create a data frame to contain the data sets.

set.seed(99)
loop<-0; time0<-0; agegrp<-0; smoker1<-3; smoker2<-3;
sims100_sm <-data.frame(cbind(loop,time0,agegrp,smoker1,smoker2))

## start the loop

for(j in 1:100) {

## create the required number of data records and record the data
## set number and 0 left censoring on each data line.

  sim_sm<-data.frame(rep(1,samp_len))
  sim_sm$loop=j
  sim_sm$time0 <- rep(0,samp_len)

## Within each age group allocate true smoking status(smoker1) in
## the same proportions as in the OC data set and then randomly
## sample imputed smoking status (smoker2) to have the same level
## of misclassification with true smoking status as in the OC data
## set.

  temp <- sample(1:20, samp_len, replace=TRUE, prob=p_sm2)
    for(i in 1:samp_len) {
      if(temp[i]==1) {
        sim_sm$agegrp[i]<-8; sim_sm$smoker1[i]<-0;
        sim_sm$smoker2[i]<-sample(0:1, 1, replace=TRUE,
                            prob=agree_sm_age8[1,])
      }
      else if(temp[i]==2) {
        sim_sm$agegrp[i]<-8; sim_sm$smoker1[i]<-1;
        sim_sm$smoker2[i]<-sample(0:1, 1, replace=TRUE,
                            prob=agree_sm_age8[2,])
      }
```

```
      else if(temp[i]==3) {
       sim_sm$agegrp[i]<-9; sim_sm$smoker1[i]<-0;
       sim_sm$smoker2[i]<-sample(0:1, 1, replace=TRUE,
                          prob=agree_sm_age9[1,])
      }
      else if(temp[i]==4) {
       sim_sm$agegrp[i]<-9; sim_sm$smoker1[i]<-1;
       sim_sm$smoker2[i]<-sample(0:1, 1, replace=TRUE,
                          prob=agree_sm_age9[2,])
      }
      ...
    }

## Remove the initial dummy column and add the newly simulated data
## set to the main data frame.

sim_sm<- sim_sm[,-1]
sims100_sm <- rbind(sims100_sm,sim_sm)
}

## Remove the initial dummy line and save the initial 100 simulation
## data sets

sims100_sm <- sims100_sm[-1,]
save(sims100_sm,file=".../sims100_5yearlag.RData")


#########################################################################
# Explore the shape of the distribution of survival times of OC    #
# cases who died.                                                       #
#########################################################################


library(fitdistrplus)

## Excluded censored data

samp_died <- sample[sample$status1==1,]

## As some distributions do not allow 0 survival time,
## 0 months survived was recoded to 0.5 months.

samp_died$survmonth <- samp_died$survmonth+0.5
descdist(samp_died$survmonth, discrete=FALSE)
fit_died_W<-fitdist(samp_died$survmonth,"weibull", method="mle")

# This analysis suggests the closest Weibull is shape=0.85,
# scale=12.3
...
# But the shape=0.85, scale=10.5 seemed to align better
# with the quartiles of the survival times of the observed OC cases
# who died.
```

```
#####################################################################
# Using the method from Bender et al, simulate survival times      #
# which produce HRs of 2.0,1.5,1.25,1.0, 0.8,0.67,0.5 both for     #
# when age is not associated with survival and when is associated. #
#####################################################################

# Define the constants:
# - the shape (v_d) and scale (l_d) parameters of the Weibull
#   distribution (note Bender et al and fitdisc() command use
#   different representations of the scale parameter)
# - the coefficient of age when fitted as a predictor of survival in
#   the OC cases who died

v_d <- 0.85; l_d <- (10.5)**(-1/v_d)
beta_age <- 0.040
...

## ... and add them to the simulations data frame

sims_sm_len <- dim(sims100_sm)[1]
sims100_sm$v <- rep(v_d,sims_sm_len)
sims100_sm$l <- rep(l_d,sims_sm_len)

## Apply Bender et al's formula to simulate survival times
## when age is associated with survival

sim_work <- sims100_sm

sim_len <- dim(sim_work)[1]
sim_work$top <- log(runif(sim_len,0,1))
sim_work$beta_age <- rep(beta_age,sim_len)


beta_sm_hr2<-log(2); beta_sm_hr15<-log(1.5);
...
sim_work$beta_sm_hr2 <- rep(beta_sm_hr2,sim_len)
sim_work$beta_sm_hr15 <- rep(beta_sm_hr15,sim_len)
...


sim_work$sm_time2 <- (-sim_work$top/(sim_work$l*
                      exp(sim_work$agegrp*sim_work$beta_age+
                      sim_work$smoker1*sim_work$beta_sm_hr2)))**
                      (1/sim_work$v)
sim_work$sm_time15 <- (-sim_work$top/(sim_work$l*
                      exp(sim_work$agegrp*sim_work$beta_age+
                      sim_work$smoker1*sim_work$beta_sm_hr15)))**
                      (1/sim_work$v)
...


## ... and when age is not associated with survival time.

sim_work$sm_time2na <-(-sim_work$top/(sim_work$l*
                      exp(sim_work$smoker1*sim_work$beta_sm_hr2)))**
```

```
                        (1/sim_work$v)
sim_work$sm_time15na<-(-sim_work$top/(sim_work$l*
                      exp(sim_work$smoker1*sim_work$beta_sm_hr15)))**
                      (1/sim_work$v)
...

#####################################################################
# Save the simulated data.                                          #
#####################################################################

sim_work <- sim_work[,-c(6:16)]
sim_work$status <- 1

save(sim_work,file="../sims100_w_time_5yearlag.RData")

#####################################################################
```

Prof David Zucker, Hebrew University, Jerusalem, Israel kindly provided his computer code for running the corrected scores method for misclassification correction in Cox regression. This code was written in Fortran 77. The program was edited to become a subroutine which could be compiled as a shared object (DLL), and then called it from within R software using the `dyn.load()` foreign function interface.

The data file for the corrected scores subroutine needed to be called ‘input.dat’ and needed to contain one line of data per OC case consisting of:

- censoring status (0=censored, 1=uncensored)
- time at entry (all 0 in the current data set, diagnosis is time 0)
- survival time (in months in the current case)
- behaviour (1=present, 2=absent)
- optional covariates (include age group here if age-adjusting)

The instruction file for the corrected scores subroutine needed to be called ‘cntlr.crd’ and needed to contain

- line 1: sample size, number of categories in the behaviour variable, number of predictors with misclassification, number of predictors without misclassification, size of the misclassification matrix, some flags to control output
- lines 2 & 3: values of the misclassified variable (1 and 2 in the current case; representing behaviour present and behaviour absent)
- lines 4 & 5: the misclassification matrix **A**
- lines 6 to 9: the matrix of partial derivatives of the misclassification matrix with respect to its parameters **Ȧ** (1 -1 ; 0 0; 0 0; -1 1 in the current case. See Zucker & Spiegelman (2008) page 1919.)
- line 10 to 11: the covariance matrix **Γ** of the misclassification parameters (a 2x2 diagonal matrix in the current case. See Zucker & Spiegelman (2008) page 1919.)

The results of the corrected score sub-routine are written to a file called 'ercox.out'.

The following R code is an example of a call to the corrected score program.

```
#####################################################################
# Read in the OC cases 100 data sets with imputed behaviours.      #
# In this example the analysis of smoking status is presented.     #
#####################################################################

load(".../working_5yearlag_inc2015.RData")

dat <- samples[complete.cases(samples$smoker1),c(1,8,4,5,6)]

#####################################################################
# Add time at entry - 0 for all cases                               #
#####################################################################

dat$time0 <- rep(0,dim(dat)[1])

#####################################################################
# For each of the 100 data sets ...                                 #
#####################################################################


for(i in 1:100) {
  sampi <- dat[dat$loop==i,c(1,6,2,3,4)]
```

```
#####################################################################
# Calculate the agreement statistics between the two                #
# imputed values of smoking status (the omegas in                   #
# Zucker & Spiegelman)                                              #
#####################################################################

  tab <-table(sampi$smoker1,sampi$smoker2)

  n <- sum(tab)
  om1 <- tab[1,1]/(tab[1,1]+tab[1,2])
  om2 <- tab[2,2]/(tab[2,1]+tab[2,2])


#####################################################################
# Calculate the covariance matrix (capital Gamma in                 #
# Zucker & Spiegelman)                                              #
#####################################################################

  th1 <- (tab[1,1]+tab[1,2])/n
  th2 <- (tab[2,1]+tab[2,2])/n
  biggam1 <- (om1*(1-om1)/th1)/(n/100)
  biggam2 <- (om2*(1-om2)/th2)/(n/100)

#####################################################################
# Write the command file cntlr.crd                                  #
#####################################################################

  n_sampi <- dim(sampi)[1]

  l1_sm <- c(n_sampi,2, 1, 0, 2, 0, 0)
  l2 <- 1
  l3 <- 2
  l4_sm <- c(om1, 1-om1)   # correcting for misclassification
  l5_sm <- c(1-om2, om2)   # correcting for misclassification
  l6 <- c(1,-1)
  l7 <- c(0,0)
  l8 <- c(0,0)
  l9 <- c(-1,1)
  l10_sm <- c(biggam1,0)
  l11_sm <- c(0,biggam2)

  write(l1_sm,file=".../cntlr.crd",ncolumns=7,sep=" ",append=FALSE)
  write(l2,file=".../cntlr.crd",append=TRUE)
  write(l3,file=".../cntlr.crd",append=TRUE)
  write(l4_sm,file=".../cntlr.crd",append=TRUE)
  write(l5_sm,file=".../cntlr.crd",append=TRUE)
  write(l6,file=".../cntlr.crd",append=TRUE)
  write(l7,file=".../cntlr.crd",append=TRUE)
  write(l8,file=".../cntlr.crd",append=TRUE)
  write(l9,file=".../cntlr.crd",append=TRUE)
  write(l10_sm,file=".../cntlr.crd",append=TRUE)
  write(l11_sm,file=".../cntlr.crd",append=TRUE)

#####################################################################
# Write the data file input.dat                                     #
#####################################################################
```

```
  sampi <- sampi[,-5]
  sampi$smoker1 <- recode(sampi$smoker1,"1=2;0=1");
  write.table(sampi,".../input.dat", append=FALSE, sep=" ",
                    dec=".", row.names=FALSE, col.names=FALSE)


#####################################################################
# Run the corrected score subroutine (here called wking.dll)      #
#####################################################################

  dyn.load(".../wking.dll")
  is.loaded("wking")
  .Fortran("wking")
  dyn.unload(".../wking.dll")
  is.loaded("wking")

}


#####################################################################
# Read the fitted coefficients from the corrected score           #
# misclassification corrected Cox regression from the file        #
# ercox.out.                                                      #
#####################################################################

sm_res <- read.table(".../ercox.out")

#####################################################################
# Convert the coefficient of smoking status into a hazard ratio   #
# and present the results as the median HR and empirical 95%      #
# confidence interval.
#####################################################################

sm_res$HR <- exp(temp$V2)
...
quantile(sm_res$HR,c(0.50,0.025,0.975))

#####################################################################
```

The following R code demonstrates the code used to create lattice plots: in this case Figure 6.11.

```
#####################################################################
# Read the results for plotting from the Excel file where they    #
# were stored and ...                                             #
#####################################################################

library(lattice)
library(readxl)
```

```
agestrata <- read_excel(".../sim5y_age_strata.xlsx", sheet="Sheet1")

## add labels and ...

agestrata$HR_f <- factor(agestrata$HR,levels=c(2,1.5,1.25,1,0.8,
                         0.67,0.5),
                         labels=c("HR=2.00", HR=1.50","HR=1.25",
                         "HR=1.00", "HR=0.80","HR=0.67","HR=0.50"))
agestrata$method_f <- factor(agestrata$method,levels=c(1,2,3,4,5),
                             labels=c("Direct","Direct nx10","I2C2",
                             "I2C2 nx10", "I2C2 nx20"))

## select the categories to display.

ages1 <- agestrata[agestrata$method<=2,]
ages1 <- ages1[ages1$agegrp!=45,]

ages2 <- agestrata[agestrata$method==1 | agestrata$method==3,]
ages2 <- ages2[ages2$agegrp!=45,]
ages2 <- droplevels(ages2)


## Where more than 5 data sets returned extreme results exclude
## the results.

n<-dim(ages2)[1]
for(i in 1:n) {
   if(ages2$errors[i]>5) {
      ages2$median[i]<-NA;ages2$lcl[i]<-NA;ages2$ucl[i]<-NA
      }
    }

#######################################################################
# Formatting commands within each pane.                               #
#######################################################################


my.panel<-function(x, y, subscripts, col, pch, group.number, ...) {
## state the locations for the confidence intervals
        low95 <- log2(ages2$lcl)[subscripts]
        up95 <- log2(ages2$ucl)[subscripts]
        myjitter <- c(-0.5,0.5)
## draw 95% confidence interval
        panel.arrows(x+myjitter[group.number], low95,
                     x+myjitter[group.number], up95, angle=90,
                     code=3, length=.05, col=col)
## connect means with line segments
        panel.xyplot(x+myjitter[group.number], y, col='white',
                     pch=16)
        panel.xyplot(x+myjitter[group.number], y, col=col,
                     pch=pch, ...)
## add grid lines
       panel.abline(h=0, col='lightgrey', lty=1)
       }
```

```
#####################################################################
# Define the plot variables and overall formatting.                 #
#####################################################################

xyplot(median~agegrp | HR_f, groups=method_f, data=ages2,
       xlab ='Age Group', ylab='Estimated Hazard Ratio (95% CI)',
       col=c(1,1), pch=c(16,15), ly=ages2$lcl, uy=ages2$ucl,
       layout=c(2,4), as.table=TRUE, ylim=c(0.2,3.0),
       type="o", lty=3,
       scales=list(y=list(log=2,alternating=1,at=c(0.5,0.67,0.8,1.0,
                          1.25,1.5,2)),
                   x=list(alternating=3,at=c(37,42,47,52,57,62,67,
                          72, 77,85), rot=45,
                          labels=c("35-39y","40-44y","45-49y",
                             "50-54y","55-59y","60-64y","65-69",
                             "70-74y","75-79y","80+y"))),
       panel.groups=my.panel, panel="panel.superpose",
       par.settings = list(strip.background=list(col="white")))

#####################################################################
```

References

Abbas, G., & Krasna, M. (2017). Overview of esophageal cancer. *Annals of cardiothoracic surgery, 6*(2), 131.

Abnet, C. C., Arnold, M., & Wei, W.-Q. (2018). Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology, 154*(2), 360-373.

Agaku, I. T., Odani, S., Okuyemi, K. S., & Armour, B. (2020). Disparities in current cigarette smoking among US adults, 2002–2016. *Tobacco control, 29*(3), 269-276.

Aghcheli, K., Marjani, H.-A., Nasrollahzadeh, D., Islami, F., Shakeri, R., Sotoudeh, M., . . . Khalilipour, E. (2011). Prognostic factors for esophageal squamous cell carcinoma—a population-based study in Golestan Province, Iran, a high incidence area. *PloS one, 6*(7).

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Ainsworth, B. E., Caspersen, C. J., Matthews, C. E., Mâsse, L. C., Baranowski, T., & Zhu, W. (2012). Recommendations to improve the accuracy of estimates of physical activity derived from self report. *Journal of Physical Activity and Health, 9*(s1), S76-S84.

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., . . . Bonaventure, A. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). *The Lancet, 385*(9972), 977-1010.

Altman, D. G. (1990). *Practical statistics for medical research*: CRC press.

American Cancer Society. (2011). *Global cancer facts & figures 2nd Edition*. American Cancer Society, Atlanta.

American Cancer Society. (2019). *Cancer Facts & Figures 2019*. Altanta: American Cancer Society.

American Cancer Society. (2021a). Cancer Facts and Figures. Retrieved from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures

American Cancer Society. (2021b). Survival rates for esophageal cancer (2010-2016). Retrieved from https://www.cancer.org/cancer/esophagus-cancer/detection-diagnosis-staging/survival-rates.html

American Cancer Society. (2022). Esophageal cancer risk factors. Retrieved from https://www.cancer.org/cancer/esophagus-cancer/causes-risks-prevention/risk-factors.html

American Joint Committee on Cancer. (2021). Cancer staging system. Retrieved from https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx

American Lung Association. (2021). Overall tobacco trends. Retrieved from https://www.lung.org/research/trends-in-lung-disease/tobacco-trends-brief/overall-tobacco-trends

Antoni, M., & Sakshaug, J. W. (2020). *Data Linkage*: SAGE Publications Limited.

Araujo, J. L., Altorki, N. K., Sonett, J. R., Rodriguez, A., Sungur-Stasik, K., Spinelli, C. F., . . . Abrams, J. A. (2016). Prediagnosis aspirin use and outcomes in a prospective cohort of esophageal cancer patients. *Therapeutic advances in gastroenterology, 9*(6), 806-814.

Armstrong, T., Bauman, A., & Davies, J. (2000). *Physical activity patterns of Australian adults. Results of the 1999 National Physical Activity Survey*. Australian Institute of Health and Welfare, Canberra.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research, 46*(3), 399-424.

Australian Bureau of Statistics. (2021a). 4263.0 - National Health Survey: Users Guide, 2017-18. Retrieved from https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4363.02017-18?OpenDocument

Australian Bureau of Statistics. (2021b). ABS/Universities Australia agreement. Retrieved from https://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+ABS/Universities+Australia+CURF+Agreement?OpenDocument

Australian Bureau of Statistics. (2021c). Microdata Entry Page. Retrieved from https://www.abs.gov.au/websitedbs/d3310114.nsf/home/microdata+entry+page

Australian Bureau of Statistics. (2021d). National Health Survey: First Results methodology. Retrieved from https://www.abs.gov.au/methodologies/national-health-survey-first-results-methodology/2017-18

Australian Government Department of Health. (2020). Australian alcohol guidelines revised. Retrieved from https://www.health.gov.au/news/australian-alcohol-guidelines-revised

Australian Government Department of Health. (2021). Physical activity and exercise guidelines for all Australians. Retrieved from https://www.health.gov.au/health-topics/physical-activity-and-exercise/physical-activity-and-exercise-guidelines-for-all-australians#physical-activity-guidelines-by-age

Australian Institute of Health and Welfare. (2018). About Australian Cancer Database. Retrieved from https://www.aihw.gov.au/about-our-data/our-data-collections/australian-cancer-database/about-australian-cancer-database

Australian Institute of Health and Welfare. (2020a). Cancer data in Australia. Retrieved from https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/data

Australian Institute of Health and Welfare. (2020b). *National Drug Strategy Household Survey 2019*. Drug Statistics series no 32. PHE 270. Canberra.

Australian Institute of Health and Welfare. (2021). Alcohol, tobacco and other drugs in Australia. Retrieved from https://www.aihw.gov.au/reports/alcohol/alcohol-tobacco-other-drugs-australia/contents/drug-types/alcohol

Azagba, S., Shan, L., Latham, K., & Manzione, L. (2020). Trends in binge and heavy drinking among adults in the United States, 2011–2017. *Substance use & misuse, 55*(6), 990-997.

Bang, H., Chiu, Y.-L., Kaufman, J. S., Patel, M. D., Heiss, G., & Rose, K. M. (2013). Bias correction methods for misclassified covariates in the Cox model: comparison of five correction methods by simulation and data analysis. *Journal of statistical theory and practice, 7*(2), 381-400.

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, 414-418.

Bauman, A., Ma, G., Cuevas, F., Omar, Z., Waqanivalu, T., Phongsavan, P., . . . Group, N.-c. D. R. F. P. C. (2011). Cross-national comparisons of socioeconomic differences in the prevalence of leisure-time and occupational physical activity, and active commuting in six Asia-Pacific countries. *Journal of Epidemiology & Community Health, 65*(1), 35-43.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine, 24*(11), 1713-1723.

Besson, H., Brage, S., Jakes, R. W., Ekelund, U., & Wareham, N. J. (2010). Estimating physical activity energy expenditure, sedentary time, and physical activity intensity by self-report in adults. *The American journal of clinical nutrition, 91*(1), 106-114.

Blanchette, C. M., DeKoven, M., De, A. P., & Roberts, M. (2013). Probabilistic data linkage: a case study of comparative effectiveness in COPD. *Drugs in context, 2013*.

Boersma, P., Villarroel, M., & Vahratian, A. (2020). *Heavy drinking among U.S. adults, 2018. NCHS Data Brief, no 374.* Hyattsville, MD.

Boniface, S., Kneale, J., & Shelton, N. (2014). Drinking pattern is more strongly associated with under-reporting of alcohol consumption than socio-demographic factors: evidence from a mixed-methods study. *BMC public health, 14*(1), 1-9.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*: John Wiley & Sons.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces* (Vol. 424): Wiley New York.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin, 68*(6), 394-424.

Cancer Australia. (2021). Oesophageal cancer statistics. Retrieved from https://www.canceraustralia.gov.au/affected-cancer/cancer-types/oesophageal-cancer/statistics

Cancer Institute NSW. (2021). Request unlinked unit record data for research. Retrieved from https://www.cancer.nsw.gov.au/research-and-data/cancer-data-and-statistics/request-unlinked-unit-record-data-for-research

Cancer Research UK. (2021). Oesophageal cancer incidence statistics (2015-2017). Retrieved from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer/incidence

CANSA. (2022). South African Cancer Statistics. Retrieved from https://cansa.org.za/south-african-cancer-statistics/

Castellsagué, X., Muñoz, N., De Stefani, E., Victora, C. G., Castelletto, R., Rolón, P. A., & Quintana, M. J. (1999). Independent and joint effects of tobacco smoking and alcohol drinking on the risk of esophageal cancer in men and women. *Int J Cancer, 82*(5), 657-664.

Castro, C., Bosetti, C., Malvezzi, M., Bertuccio, P., Levi, F., Negri, E., . . . Lunet, N. (2014). Patterns and trends in esophageal cancer mortality and incidence in Europe (1980–2011) and predictions to 2015. *Annals of oncology, 25*(1), 283-290.

Castro, C., Peleteiro, B., & Lunet, N. (2018). Modifiable factors and esophageal cancer: a systematic review of published meta-analyses. *Journal of gastroenterology, 53*(1), 37-51.

CDC Cancer Registry Data Access for Research Project. (2018). *Cancer Registry Research Approval Process: Classification of States by Level of Approval Required*: National Center for Chronic Disease Prevention and Health Promotion.

Centers for Disease Control. (2015). BRFSS Combined Landline and Cell Phone Weighted Response Rates by State, 2015. Retrieved from https://www.cdc.gov/brfss/annual_data/2015/2015_responserates.html

Centers for Disease Control. (2018). 2017 BRFSS survey data and documentation. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2017.html

Centers for Disease Control and Prevention. United States Cancer Statistics: Data Visualizations. Retrieved from https://gis.cdc.gov/Cancer/USCS/DataViz.html

Centers for Disease Control and Prevention. (2013a). Behavioural Risk Factor Surveillance System (BRFSS). Retrieved from http://www.cdc.gov/brfss/

Centers for Disease Control and Prevention. (2013b). *The BRFSS data user guide*. Atlanta: Department of Health and Human Services.

Centers for Disease Control and Prevention. (2021a). Behavioural Risk Factor Surveillance System. Retrieved from https://www.cdc.gov/brfss/data_documentation/index.htm

Centers for Disease Control and Prevention. (2021b). Burden of Cigarette Use in the U.S. Retrieved from https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html

Centers for Disease Control and Prevention. (2021c). National Program for Cancer Registries (NPCR). Retrieved from https://www.cdc.gov/cancer/npcr/

Centers for Disease Control and Prevention (CDA). *Behavioral Risk Factor Surveillance System Survey Data. U.S. Department of Health and Human Servies, Centers for Disease Control and Prevention. Atlanta, Georgia. 2001-2014*.

Cescon, D. W., Bradbury, P. A., Asomaning, K., Hopkins, J., Zhai, R., Zhou, W., . . . Ma, C. (2009). p53 Arg72Pro and MDM2 T309G polymorphisms, histology, and esophageal cancer prognosis. *Clinical Cancer Research, 15*(9), 3103-3109.

Chen, R., Ma, S., Guan, C., Song, G., Ma, Q., Xie, S., . . . Wei, W. (2019). The national cohort of esophageal cancer-prospective cohort study of esophageal cancer and precancerous lesions based on high-risk population in china (NCEC-HRP): study protocol. *BMJ open, 9*(4), e027360.

Chen, Y., Yu, C., & Li, Y. (2014). Physical activity and risks of esophageal and gastric cancers: a meta-analysis. *PloS one, 9*(2), e88082.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology, 110*, 12-22.

Cole, S. R., & Hernán, M. A. (2002). Fallibility in estimating direct effects. *International journal of epidemiology, 31*(1), 163-165.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International journal of epidemiology, 39*(2), 417-420.

Copeland, K. T., Checkoway, H., McMichael, A. J., & Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology, 105*(5), 488-495.

Corley, D. A., Kerlikowske, K., Verma, R., & Buffler, P. (2003). Protective association of aspirin/NSAIDs and esophageal cancer: a systematic review and meta-analysis. *Gastroenterology, 124*(1), 47-56.

Craig, C. L., Russell, S. J., Cameron, C., & Bauman, A. (2004). Twenty-year trends in physical activity among Canadian adults. *Canadian journal of public health, 95*(1), 59-63.

Dandara, C., Robertson, B., Dzobo, K., Moodley, L., & Parker, M. I. (2015). Patient and tumour characteristics as prognostic markers for oesophageal cancer: a retrospective analysis of a cohort of patients at Groote Schuur Hospital. *Eur J Cardiothorac Surg, 49*(2), 629-634.

Dandara, C., Robertson, B., Dzobo, K., Moodley, L., & Parker, M. I. (2016). Patient and tumour characteristics as prognostic markers for oesophageal cancer: a retrospective

analysis of a cohort of patients at Groote Schuur Hospital. *European Journal of Cardio-Thoracic Surgery, 49*(2), 629-634.
Davis, C. G., Thake, J., & Vilhena, N. (2010). Social desirability biases in self-reported alcohol consumption and harms. *Addictive behaviors, 35*(4), 302-311.
de Klerk, N. H., English, D. R., & Armstrong, B. K. (1989). A review of the effects of random measurement error on relative risk estimates in epidemiological studies. *Int J Epidemiol, 18*(3), 705-712.
De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563): John Wiley & Sons.
Delker, E., Brown, Q., & Hasin, D. S. (2016). Alcohol consumption in demographic subpopulations: an epidemiologic overview. *Alcohol research: current reviews, 38*(1), 7.
Demark-Wahnefried, W., Aziz, N. M., Rowland, J. H., & Pinto, B. M. (2005). Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. *J Clin Oncol, 23*(24), 5814.
Department of Health & Social Care. (2019). *UK Chief Medical Officers' physical activity guidelines*. Department of Health & Social Care, Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/832868/uk-chief-medical-officers-physical-activity-guidelines.pdf
Di Pardo, B. J., Bronson, N. W., Diggs, B. S., Thomas, C. R., Hunter, J. G., & Dolan, J. P. (2016). The global burden of esophageal cancer: a disability-adjusted life-year approach. *World journal of surgery, 40*(2), 395-401.
Doll, K. M., Rademaker, A., & Sosa, J. A. (2018). Practical guide to surgical data sets: surveillance, epidemiology, and end results (SEER) database. *JAMA surgery, 153*(6), 588-589.
Dong, J., Gu, X., El-Serag, H. B., & Thrift, A. P. (2019). Underuse of surgery accounts for racial disparities in esophageal cancer survival times: a matched cohort study. *Clinical Gastroenterology and Hepatology, 17*(4), 657-665. e613.
DrugRehab.com. (2021). How long does alcohol stay in your system (blood, urine and saliva)? Retrieved from https://www.drugrehab.com/addiction/alcohol/how-long-does-alcohol-stay-in-your-system/
Dubecz, A., Gall, I., Solymosi, N., Schweigert, M., Peters, J. H., Feith, M., & Stein, H. J. (2012). Temporal trends in long-term survival and cure rates in esophageal cancer: a SEER database analysis. *Journal of Thoracic Oncology, 7*(2), 443-447.
Enzinger, P. C., & Mayer, R. J. (2003). Esophageal cancer. *New England Journal of Medicine, 349*(23), 2241-2252.
Fahey, P. P., Mallitt, K.-A., Astell-Burt, T., Stone, G., & Whiteman, D. C. (2015). Impact of pre-diagnosis behavior on risk of death from esophageal cancer: a systematic review and meta-analysis. *Cancer Causes Control, 26*(10), 1365-1373.
Fahey, P. P., Page, A., Stone, G., & Astell-Burt, T. (2020a). Augmenting cancer registry data with health survey data with no cases in common: the relationship between pre-diagnosis health behaviour and post-diagnosis survival in oesophageal cancer. *BMC cancer, 20*, 1-11.
Fahey, P. P., Page, A., Stone, G., & Astell-Burt, T. (2020b). Using estimated probability of pre-diagnosis behavior as a predictor of cancer survival time: an example in esophageal cancer. *BMC medical research methodology, 20*, 1-9.
Faisal, M., Scally, A., Howes, R., Beatson, K., Richardson, D., & Mohammed, M. A. (2020). A comparison of logistic regression models with alternative machine learning

methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. *Health informatics journal, 26*(1), 34-44.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., . . . Bray, F. (2015). GLOBOCAN 2012 v1. 0, Cancer incidence and mortality worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer, Lyon, France. 2013. *globocan. iarc. fr*.

Ferronha, I., Bastos, A., & Lunet, N. (2012). Prediagnosis lifestyle exposures and survival of patients with gastric cancer: systematic review and meta-analysis. *European Journal of Cancer Prevention, 21*(5), 449-452.

Florescu, A., Ferrence, R., Einarson, T., Selby, P., Soldin, O., & Koren, G. (2009). Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. *Therapeutic drug monitoring, 31*(1), 14.

Forsea, A.-M. (2016). Cancer registries in Europe—going forward is the only option. *Ecancermedicalscience, 10*.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. *International Encyclopedia of the social & Behavioral sciences, 6*(4027-4030), 1-7.

Friedenreich, C. M., Neilson, H. K., & Lynch, B. M. (2010). State of the epidemiological evidence on physical activity and cancer prevention. *European journal of cancer, 46*(14), 2593-2604.

Friedenreich, C. M., Stone, C. R., Cheung, W. Y., & Hayes, S. C. (2020). Physical activity and mortality in cancer survivors: a systematic review and meta-analysis. *JNCI cancer spectrum, 4*(1), pkz080.

Gao, H., Feng, H.-M., Li, B., Lin, J.-P., Yang, J.-B., Zhu, D.-J., & Jing, T. (2018). Impact of high body mass index on surgical outcomes and long-term survival among patients undergoing esophagectomy: a meta-analysis. *Medicine, 97*(28).

Gilman, S. E., Breslau, J., Conron, K. J., Koenen, K. C., Subramanian, S., & Zaslavsky, A. (2008). Education and race-ethnicity differences in the lifetime risk of alcohol dependence. *J Epidemiol Community Health, 62*(3), 224-230.

Gorber, S. C., Tremblay, M., Moher, D., & Gorber, B. (2007). A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review. *Obesity reviews, 8*(4), 307-326.

GOV.UK. (2021). Accessing Public Health England data. Retrieved from https://www.gov.uk/government/publications/accessing-public-health-england-data

Grant, B. F., Chou, S. P., Saha, T. D., Pickering, R. P., Kerridge, B. T., Ruan, W. J., . . . Fan, A. (2017). Prevalence of 12-month alcohol use, high-risk drinking, and DSM-IV alcohol use disorder in the United States, 2001-2002 to 2012-2013: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *JAMA psychiatry, 74*(9), 911-923.

Greene, F. L., Page, D. L., leming, I. D., Fritz, A. G., Balch, C. M., Haller, D. G., & Morrow, M. (2003). *AJCC Cancer Staging Manual,* (6th ed.). Berlin: Springer-Verlag,.

Grucza, R. A., Sher, K. J., Kerr, W. C., Krauss, M. J., Lui, C. K., McDowell, Y. E., . . . Bierut, L. J. (2018). Trends in adult alcohol use and binge drinking in the early 21st-century United States: a meta-analysis of 6 National Survey Series. *Alcoholism: clinical and experimental research, 42*(10), 1939-1950.

Gupta, V., Coburn, N., Kidane, B., Hess, K. R., Compton, C., Ringash, J., . . . Mahar, A. L. (2018). Survival prediction tools for esophageal and gastroesophageal junction cancer: A systematic review. *The Journal of thoracic and cardiovascular surgery, 156*(2), 847-856.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big data & society, 4*(2), 2053951717745678.

Haskell, W. L. (2012). Physical activity by self-report: a brief history and future issues. *Journal of Physical Activity and Health, 9*(s1), S5-S10.

He, Y., Liang, D., Du, L., Guo, T., Liu, Y., Sun, X., . . . Shan, B. (2020). Clinical characteristics and survival of 5283 esophageal cancer patients: A multicenter study from eighteen hospitals across six regions in China. *Cancer Communications, 40*(10), 531-544.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*: John Wiley & Sons.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj, 327*(7414), 557-560.

Hong, L., Zhang, H., Zhao, Q., Han, Y., Yang, J., & Brain, L. (2013). Relation of excess body weight and survival in patients with esophageal adenocarcinoma: a meta-analysis. *Diseases of the Esophagus, 26*(6), 623-627.

Hongo, M., Nagasaki, Y., & Shoji, T. (2009). Epidemiology of esophageal cancer: Orient to Occident. Effects of chronology, geography and ethnicity. *Journal of gastroenterology and hepatology, 24*(5), 729-735.

Hubbard, V. S. (2000). Defining overweight and obesity: what are the issues? In: Oxford University Press.

Iachan, R., Pierannunzi, C., Healey, K., Greenlund, K. J., & Town, M. (2016). National weighting of data from the behavioral risk factor surveillance system (BRFSS). *BMC Med Res Methodol, 16*(1), 155.

International Agency for Research on Cancer. (2008). Methods for evaluating tobacco control policies. IARC handbooks of cancer prevention. Vol. 12. In: Geneva, Switzerland: WHO Press.

Islami, F., Goding Sauer, A., Miller, K. D., Siegel, R. L., Fedewa, S. A., Jacobs, E. J., . . . Soerjomataram, I. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA Cancer J Clin, 68*(1), 31-54.

Jamal, A. (2016). Current cigarette smoking among adults—United States, 2005–2015. *MMWR. Morb Mortal Wkly Rep, 65*.

Jedy-Agba, E. E., Oga, E. A., Odutola, M., Abdullahi, Y. M., Popoola, A., Achara, P., . . . Erinomo, O. (2015). Developing national cancer registration in developing countries–case study of the Nigerian national system of cancer registries. *Frontiers in public health, 3*, 186.

Jing, C., Huang, Z., Duan, Y., Xiao, X., Zhang, R., & Jiang, J. (2012). Folate intake, methylenetetrahydrofolate reductase polymorphisms in association with the prognosis of esophageal squamous cell carcinoma. *Asian Pacific Journal of Cancer Prevention, 13*(2), 647-651.

Jung, H.-K., Tae, C. H., Lee, H.-A., Lee, H., Don Choi, K., Park, J. C., . . . Sung, J. (2020). Treatment pattern and overall survival in esophageal cancer during a 13-year period: A nationwide cohort study of 6,354 Korean patients. *PloS one, 15*(4), e0231456.

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Jama, 282*(11), 1054-1060.

Jurek, A. M., Greenland, S., Maldonado, G., & Church, T. R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol, 34*(3), 680-687.

Kahlert, J., Gribsholt, S. B., Gammelager, H., Dekkers, O. M., & Luta, G. (2017). Control of confounding in the analysis phase–an overview for clinicians. *Clinical epidemiology, 9*, 195.

Keating, X. D., Zhou, K., Liu, X., Hodges, M., Liu, J., Guan, J., . . . Castro-Piñero, J. (2019). Reliability and concurrent validity of global physical activity questionnaire (GPAQ): a systematic review. *International journal of environmental research and public health, 16*(21), 4128.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Korpanty, G. J., Eng, L., Qiu, X., Faluyi, O. O., Renouf, D. J., Cheng, D., . . . Xu, W. (2017). Association of BRM promoter polymorphisms and esophageal adenocarcinoma outcome. *Oncotarget, 8*(17), 28093-28100. doi:10.18632/oncotarget.15890

Kuang, J.-j., Jiang, Z.-m., Chen, Y.-x., Ye, W.-p., Yang, Q., Wang, H.-z., & Xie, D.-r. (2016). Smoking exposure and survival of patients with esophagus cancer: a systematic review and meta-analysis. *Gastroenterol Research Pract, 2016*.

Kuo, T.-M., & Mobley, L. R. (2016). How generalizable are the SEER registries to the cancer populations of the USA? *Cancer Causes Control, 27*(9), 1117-1126.

Lantz, P. M., Janz, N. K., Fagerlin, A., Schwartz, K., Liu, L., Lakhani, I., . . . Katz, S. J. (2005). Satisfaction with surgery outcomes and the decision process in a population-based sample of women with breast cancer. *Health services research, 40*(3), 745-768.

Lauder, W., Mummery, K., Jones, M., & Caperchione, C. (2006). A comparison of health behaviours in lonely and non-lonely populations. *Psychol Health Med, 11*(2), 233-245.

Leventhal, A. M., Bello, M. S., Galstyan, E., Higgins, S. T., & Barrington-Trimis, J. L. (2019). Association of cumulative socioeconomic and health-related disadvantage with disparities in smoking prevalence in the United States, 2008 to 2017. *JAMA internal medicine, 179*(6), 777-785.

Lewis, C. E., McTigue, K. M., Burke, L. E., Poirier, P., Eckel, R. H., Howard, B. V., . . . Pi-Sunyer, F. X. (2009). Mortality, health outcomes, and body mass index in the overweight range: a science advisory from the American Heart Association. *Circulation, 119*(25), 3263-3271.

Liber, A. C., & Warner, K. E. (2018). Has underreporting of cigarette consumption changed over time? Estimates derived from US National Health Surveillance Systems between 1965 and 2015. *American journal of epidemiology, 187*(1), 113-119.

Lin, M.-Y., Chen, M.-C., Wu, I.-C., Wu, D.-C., Cheng, Y.-J., Wu, C.-C., . . . Wu, M.-T. (2011). Areca users in combination with tobacco and alcohol use are associated with younger age of diagnosed esophageal cancer in Taiwanese men. *PloS one, 6*(10), e25347.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793): Wiley.

Ljung, R., Drefahl, S., Andersson, G., & Lagergren, J. (2013). Socio-demographic and geographical factors in esophageal and gastric cancer mortality in Sweden. *PloS one, 8*(4), e62067.

Loehrer, E. A., Giovannucci, E. L., Betensky, R. A., Shafer, A., & Christiani, D. C. (2020). Prediagnostic adult body mass index change and esophageal adenocarcinoma survival. *Cancer medicine, 9*(10), 3613-3622.

Lunn, A. D., & Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika, 85*(2), 487-490.

Lynch, B. M., & Leitzmann, M. F. (2017). An evaluation of the evidence relating to physical inactivity, sedentary behavior, and cancer incidence and mortality. *Curr Epidemiol Rep, 4*(3), 221-231.

Ma, Q., Liu, W., Jia, R., Long, H., Zhang, L., Lin, P., . . . Ma, G. (2016). Alcohol and survival in ESCC: Prediagnosis alcohol consumption and postoperative survival in lymph node-negative esophageal carcinoma patients. *Oncotarget, 7*(25), 38857.

Macfarlane, T. V., Murchie, P., & Watson, M. C. (2015). Aspirin and other non-steroidal anti-inflammatory drug prescriptions and survival after the diagnosis of head and neck and oesophageal cancer. *Cancer Epidemiol, 39*(6), 1015-1022.

Maddams, J., Utley, M., & Møller, H. (2012). Projections of cancer prevalence in the United Kingdom, 2010–2040. *British journal of cancer, 107*(7), 1195-1202.

Mahnken, J. D., Keighley, J. D., Cumming, C. G., Girod, D. A., & Mayo, M. S. (2008). Evaluating the completeness of the SEER-Medicare linked database for oral and pharyngeal cancer. *Journal of registry management, 35*(4), 145.

Mallen, C., Peat, G., & Croft, P. (2006). Quality assessment of observational studies is not commonplace in systematic reviews. *Journal of clinical epidemiology, 59*(8), 765-769.

Manatu Hauora Ministry of Health. (2020). New Zealand Cancer Registry. Retrieved from https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/new-zealand-cancer-registry-nzcr

Manatu Hauora Ministry of Health. (2021). Questionnaires and content guide 2020/21: New Zealand Health Survey. Retrieved from https://www.health.govt.nz/publication/questionnaires-and-content-guide-2020-21-new-zealand-health-survey

Mariotto, A. B., Robin Yabroff, K., Shao, Y., Feuer, E. J., & Brown, M. L. (2011). Projections of the cost of cancer care in the United States: 2010–2020. *Journal of the National Cancer Institute, 103*(2), 117-128.

McMenamin, U. C., McCain, S., & Kunzmann, A. T. (2017). Do smoking and alcohol behaviours influence GI cancer survival? *Best Pract Res Clin Gastroenterol, 31*(5), 569-577.

McTiernan, A., Friedenreich, C. M., Katzmarzyk, P. T., Powell, K. E., Macko, R., Buchner, D., . . . Vaux-Bjerke, A. (2019). Physical activity in cancer prevention and survival: a systematic review. *Medicine and science in sports and exercise, 51*(6), 1252.

Meader, N., King, K., Moe-Byrne, T., Wright, K., Graham, H., Petticrew, M., . . . Sowden, A. J. (2016). A systematic review on the clustering and co-occurrence of multiple risk behaviours. *BMC public health, 16*(1), 657.

Méndez, D., Tam, J., Giovino, G. A., Tsodikov, A., & Warner, K. E. (2016). Has smoking cessation increased? An examination of the US adult smoking cessation rate 1990–2014. *Nicotine Tob Res, 19*(12), 1418-1424.

Mirinezhad, S. K., Somi, M. H., Jangjoo, A. G., Seyednezhad, F., Dastgiri, S., Mohammadzadeh, M., . . . Nasiri, B. (2012). Survival rate and prognostic factors of esophageal cancer in east Azerbaijan province, North-west of Iran. *Asian Pac J Cancer Prev, 13*(7), 3451-3454.

Misra, A., Wasir, J. S., & Vikram, N. K. (2005). Waist circumference criteria for the diagnosis of abdominal obesity are not applicable uniformly to all populations and ethnic groups. *Nutrition, 21*(9), 969-976.

Moola, S., Munn, Z., Tufanaru, C., Aromataris, E., Sears, K., Sfetcu, R., . . . Lisy, K. (2017). Chapter 7: Systematic reviews of etiology and risk. *Joanna briggs institute reviewer's manual. The Joanna Briggs Institute, 5*.

Moore, S. C., Lee, I.-M., Weiderpass, E., Campbell, P. T., Sampson, J. N., Kitahara, C. M., . . . Hartge, P. (2016). Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. *JAMA Intern Med, 176*(6), 816-825.

Morris, L. J., D'Este, C., Sargent-Cox, K., & Anstey, K. J. (2016). Concurrent lifestyle risk factors: Clusters and determinants in an Australian sample. *Prev Med, 84*, 1-5.

Morseth, B., & Hopstock, L. A. (2020). Time trends in physical activity in the Tromsø study: An update. *PloS one, 15*(4), e0231581.

Nance, R., Delaney, J., McEvoy, J. W., Blaha, M. J., Burke, G. L., Navas-Acien, A., . . . McClelland, R. L. (2017). Smoking intensity (pack/day) is a better measure than pack-years or smoking status for modeling cardiovascular disease outcomes. *Journal of clinical epidemiology, 81*, 111-119.

National Cancer Institute. (2021a). Cancer Stat Facts: esophegeal cancer. Retrieved from https://seer.cancer.gov/statfacts/html/esoph.html

National Cancer Institute. (2021b). Dictionary of cancer terms. Retrieved from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pack-year

National Cancer Institute. (2021c). SEER training modules. Staging a cancer case. Retrieved from https://training.seer.cancer.gov/staging/

National Cancer Institute. (2021d). Surveillance, Epidemiology, and End Results Program. Retrieved from https://seer.cancer.gov/

National Cancer Institute. (2022a). Alcohol and Cancer Risk. Retrieved from https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet

National Cancer Institute. (2022b). SEER data change history. Retrieved from https://seer.cancer.gov/data/data-changes.html

National Cancer Institute Surveillance Epidemiology and End Results Program. (2021). What is a cancer registry. Retrieved from https://seer.cancer.gov/registries/cancer_registry/index.html

National Center for Health Statistics. Surveys and data collection systems. Retrieved from https://www.cdc.gov/nchs/surveys.htm#tabs-1-1

National Center for Health Statistics. (2014). *Health, United States, 2013: With special feature on prescription drugs* Hyattsville, MD.

National Center for Health Statistics. (2021a). About the National Health Interview Survey. Retrieved from https://www.cdc.gov/nchs/nhis/about_nhis.htm

National Center for Health Statistics. (2021b). National Health and Nutrition Examination Survey. Retrieved from https://www.cdc.gov/nchs/nhanes/index.htm

National Center for Health Statistics. (2021c). National Health Interview Survey tables of summary health statistics. Retrieved from https://www.cdc.gov/nchs/nhis/shs/tables.htm

National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), & ICF. (2019). *South African Demographic and Health Survey 2016.* Pretoria, South Africa: NDoH, Stats SA, SAMRC, and ICF

National Information Standards Organization. (2021). CRediT. Contributor Roles Taxonomy. Retrieved from http://credit.niso.org/

National Institute for Health Research, Cambridge Biomedical Research Centre,. (2021). Physical activity assessment, objective measures. Retrieved from https://dapa-toolkit.mrc.ac.uk/physical-activity/objective-methods/introduction

National Institute on Alcohol Abuse and Alcoholism. (2021a). Alcohol use in the United States (2019). Retrieved from https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics

National Institute on Alcohol Abuse and Alcoholism. (2021b). Drinking levels defined. Retrieved from https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking

Neeland, I. J., Poirier, P., & Després, J.-P. (2018). Cardiovascular and metabolic heterogeneity of obesity: clinical challenges and implications for management. *Circulation, 137*(13), 1391-1406.

New Zealand Legislation. (2021). Cancer Registry Act 1993. Retrieved from https://www.legislation.govt.nz/act/public/1993/0102/latest/DLM318888.html

Ng, S. W., & Popkin, B. M. (2012). Time use and physical activity: a shift away from movement across the globe. *Obesity reviews, 13*(8), 659-680.

NHS Digital. (2021). Heath Survey for England. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england

Nieto, F. J., & Coresh, J. (1996). Adjusting survival curves for confounders: a review and a new method. *American journal of epidemiology, 143*(10), 1059-1068.

Njei, B., McCarty, T. R., & Birk, J. W. (2016). Trends in esophageal cancer survival in United States adults from 1973 to 2009: a SEER database analysis. *J Gastroenterol Hepatol, 31*(6), 1141-1146.

Nordholt, E. S. (1998). Imputation: methods, simulation experiments and practical examples. *International Statistical Review, 66*(2), 157-180.

NSW Health. (2009). Policy Directive PD2009_012: Cancer registry - Notifying cancer cases to the NSW Central Cancer Registry. Retrieved from https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/PD2009_012.pdf

NSW Health. (2021). CHeReL Centre for Health Record Linkage. Retrieved from https://www.cherel.org.au/

Nugawela, M. D., Langley, T., Szatkowski, L., & Lewis, S. (2016). Measuring alcohol consumption in population surveys: a review of international guidelines and comparison with surveys in England. *Alcohol and alcoholism, 51*(1), 84-92.

Nuttall, F. Q. (2015). Body mass index: obesity, BMI, and health: a critical review. *Nutrition today, 50*(3), 117.

Office for National Statistics. (2021a). Cancer registration statistics, England Statistical bulletins. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/previousReleases

Office for National Statistics. (2021b). Cancer survival in England - adults diagnosed (2013-2017). Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed

Okada, E., Ukawa, S., Nakamura, K., Hirata, M., Nagai, A., Matsuda, K., . . . Tamakoshi, A. (2017). Demographic and lifestyle factors and survival among patients with esophageal and gastric cancer: The Biobank Japan Project. *J Epidemiol, 27*(3s), S29-s35. doi:10.1016/j.je.2016.12.002

Oman, S. D., & Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika, 88*(1), 287-290.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Brennan, S. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *Bmj, 372*.

Palamuthusingam, D., Johnson, D. W., Hawley, C., Pascoe, E., & Fahim, M. (2019). Health data linkage research in Australia remains challenging. *Internal medicine journal, 49*(4), 539-544.

Pan, Y.-P., Kuo, H.-C., Hsu, T.-Y., Lin, J.-Y., Chou, W.-C., Lai, C.-H., . . . Yeh, K.-Y. (2020). Body Mass Index-Adjusted Body Weight Loss Grading Predicts Overall Survival in Esophageal Squamous Cell Carcinoma Patients. *Nutrition and cancer*, 1-8.

Pan, Y. P., Hsu, T. Y., Lin, J. Y., Ho, C. J., Kuan, C. Y., Chou, W. C., . . . Yeh, K. Y. (2018). Prognostic Significance of Low Body Mass Index and Betel-Quid Use in the 5-Year Survival Rates of Esophageal Squamous Cell Carcinoma Patients. *Nutr Cancer, 70*(8), 1315-1321. doi:10.1080/01635581.2019.1588983

Parekh, N., Chandran, U., & Bandera, E. V. (2012). Obesity in cancer survival. *Annu Rev Nutr, 32*, 311-342.

Park, H. S., Lloyd, S., Decker, R. H., Wilson, L. D., & Yu, J. B. (2012). Overview of the Surveillance, Epidemiology, and End Results database: evolution, data variables, and quality assurance. *Current Problems in Cancer, 36*(4), 183-190.

Park, S. M., Lim, M. K., Shin, S. A., & Yun, Y. H. (2006). Impact of prediagnosis smoking, alcohol, obesity, and insulin resistance on survival in male cancer patients: National Health Insurance Corporation Study. *Journal of clinical Oncology, 24*(31), 5017-5024.

Parkin, D. M. (2006). The evolution of the population-based cancer registry. *Nature Reviews Cancer, 6*(8), 603-612.

Parkin, E., O'Reilly, D. A., Sherlock, D. J., Manoharan, P., & Renehan, A. G. (2014). Excess adiposity and survival in patients with colorectal cancer: a systematic review. *Obesity reviews, 15*(5), 434-451.

Patterson, C., Hogan, L., & Cox, M. (2019). A comparison between two retrospective alcohol consumption measures and the daily drinking diary method with university students. *The American journal of drug and alcohol abuse, 45*(3), 248-253.

Pennathur, A., Zhang, J., Chen, H., & Luketich, J. D. (2010). The “best operation” for esophageal cancer? *The Annals of thoracic surgery, 89*(6), S2163-S2167.

Peto, J. (2012). That the effects of smoking should be measured in pack-years: misconceptions 4. *British journal of cancer, 107*(3), 406-407.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature medicine, 25*(1), 37-43.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International journal of behavioral nutrition and physical activity, 5*(1), 1-24.

Qiao, Y., Yang, T., Gan, Y., Li, W., Wang, C., Gong, Y., & Lu, Z. (2018). Associations between aspirin use and the risk of cancers: a meta-analysis of observational studies. *BMC cancer, 18*(1), 1-57.

Reichle, K., Peter, R. S., Concin, H., & Nagel, G. (2015). Associations of pre-diagnostic body mass index with overall and cancer-specific mortality in a large Austrian cohort. *Cancer Causes Control, 26*(11), 1643-1652.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences, 18*(1), 89-107.

Rock, C. L., Doyle, C., Demark-Wahnefried, W., Meyerhardt, J., Courneya, K. S., Schwartz, A. L., . . . McCullough, M. (2012). Nutrition and physical activity guidelines for cancer survivors. *CA Cancer J Clin, 62*(4), 242-274.

Roder, D., Fong, K. M., Brown, M. P., Zalcberg, J., & Wainwright, C. (2014). Realising opportunities for evidence-based cancer service delivery and research: linking cancer registry and administrative data in A ustralia. *European journal of cancer care, 23*(6), 721-727.

Salek, R., Bezenjani, S. E., Saedi, H. S., Ashkiki, M. H. H., Hosainzade, S. M., Mohtashami, S., . . . Ghavamnasiri, M. R. (2009). A geographic area with better outcome of esophageal carcinoma: Is there an effect of ethnicity and etiologic factors? *Oncology, 77*(3-4), 172-177.

Samadi, F., Babaei, M., Yazdanbod, A., Fallah, M., Nouraie, M., Nasrollahzadeh, D., . . . Fuladi, R. (2007). Survival rate of gastric and esophageal cancers in Ardabil province, North-West of Iran. *Arch Iran Med, 10*(1), 32-37.

Sanderson, S., Tatt, I. D., & Higgins, J. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International journal of epidemiology, 36*(3), 666-676.

Sax Institute. (2021). SURE. Retrieved from https://www.saxinstitute.org.au/our-work/sure/

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.), 20*(4), 512.

Schneider, K. L., Clark, M. A., Rakowski, W., & Lapane, K. L. (2012). Evaluating the impact of non-response bias in the Behavioral Risk Factor Surveillance System (BRFSS). *J Epidemiol Community Health, 66*(4), 290-295.

Sehgal, S., Kaul, S., Gupta, B., & Dhar, M. (2012). Risk factors and survival analysis of the esophageal cancer in the population of Jammu, India. *Indian journal of cancer, 49*(2), 245.

Shen, S., Araujo, J., Altorki, N., Sonett, J., Rodriguez, A., Sungur-Stasik, K., . . . Abrams, J. (2017). Variation by stage in the effects of prediagnosis weight loss on mortality in a prospective cohort of esophageal cancer patients. *Diseases of the Esophagus, 30*(9), 1.

Shitara, K., Matsuo, K., Hatooka, S., Ura, T., Takahari, D., Yokota, T., . . . Kodaira, T. (2010). Heavy smoking history interacts with chemoradiotherapy for esophageal cancer prognosis: a retrospective study. *Cancer Sci, 101*(4), 1001-1006.

Short, M. W., Burgers, K., & Fry, V. (2017). Esophageal cancer. *American family physician, 95*(1), 22-28.

Siddiqui, A. H., & Zafar, S. N. (2018). Global availability of cancer registry data. *J Glob Oncol, 4*.

Siddiqui, T., Alkadri, M., & Khan, N. A. (2017). Review of Programming Languages and Tools for Big Data Analytics. *International Journal of Advanced Research in Computer Science, 8*(5).

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J Clin, 69*(1), 7-34.

Singh, S., Devanna, S., Varayil, J. E., Murad, M. H., & Iyer, P. G. (2014). Physical activity is associated with reduced risk of esophageal cancer, particularly esophageal adenocarcinoma: a systematic review and meta-analysis. *BMC gastroenterology, 14*(1), 1-11.

Smithers, B. M., Fahey, P. P., Corish, T., Gotley, D. C., Falk, G. L., Smith, G. S., . . . Whiteman, D. C. (2010). Symptoms, investigations and management of patients with cancer of the oesophagus and gastro-oesophageal junction in Australia. *Med J Aust, 193*(10), 572-577.

Spence, A. D., Busby, J., Johnston, B. T., Baron, J. A., Hughes, C. M., Coleman, H. G., & Cardwell, C. R. (2018). Low-dose aspirin use does not increase survival in 2

independent population-based cohorts of patients with esophageal or gastric cancer. *Gastroenterology, 154*(4), 849-860. e841.

Spiegelman, D., Rosner, B., & Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association, 95*(449), 51-61.

Spreafico, A., Coate, L., Zhai, R., Xu, W., Chen, Z.-F., Chen, Z., . . . Heist, R. S. (2017). Early adulthood body mass index, cumulative smoking, and esophageal adenocarcinoma survival. *Cancer Epidemiol, 47*, 28-34.

Spreafico, A., Coate, L., Zhai, R., Xu, W., Chen, Z. F., Chen, Z., . . . Liu, G. (2017). Early adulthood body mass index, cumulative smoking, and esophageal adenocarcinoma survival. *Cancer Epidemiol, 47*, 28-34. doi:10.1016/j.canep.2016.11.009

Statistics Canada. (2022a). Canadian Cancer registry (CCR). Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3207#a1

Statistics Canada. (2022b). Canadian Community Health Survey - Annual Component (CCHS). Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226

Steevens, J., Schouten, L. J., Goldbohm, R. A., & van den Brandt, P. A. (2010). Alcohol consumption, cigarette smoking and risk of subtypes of oesophageal and gastric cancer: a prospective cohort study. *Gut, 59*(01), 39-48.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br Med J, 338*, b2393.

Stockwell, T., Zhao, J., Greenfield, T., Li, J., Livingston, M., & Meng, Y. (2016). Estimating under-and over-reporting of drinking in national surveys of alcohol consumption: identification of consistent biases across four English-speaking countries. *Addiction, 111*(7), 1203-1213.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., . . . Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Jama, 283*(15), 2008-2012.

Sun, L.-P., Yan, L.-B., Liu, Z.-Z., Zhao, W.-J., Zhang, C.-X., Chen, Y.-M., . . . Liu, X. (2020). Dietary factors and risk of mortality among patients with esophageal cancer: a systematic review. *BMC cancer, 20*, 1-13.

Sun, P., Zhang, F., Chen, C., Ren, C., Bi, X.-W., Yang, H., . . . Jiang, W.-Q. (2016). Prognostic impact of body mass index stratified by smoking status in patients with esophageal squamous cell carcinoma. *Onco Targets Ther, 9*, 6389.

Sundelöf, M., Lagergren, J., & Ye, W. (2008). Patient demographics and lifestyle factors influencing long-term survival of oesophageal cancer and gastric cardia cancer in a nationwide study in Sweden. *European journal of cancer, 44*(11), 1566-1571.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

Surveillance Epidemiology and End Results (SEER) Program. *Research Data (1973-2013). National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission*.

Surveillance Research Program, National Cancer Institute,. (2020). SEER*Stat software (seer.cancer.gov/seerstat) version 8.3.5.

Surveillance Research Program, National Cancer Institute,. (2021). SEER*Stat software. Retrieved from https://seer.cancer.gov/seerstat

Sylvia, L. G., Bernstein, E. E., Hubbard, J. L., Keating, L., & Anderson, E. J. (2014). Practical guide to measuring physical activity. *Journal of the Academy of Nutrition and Dietetics, 114*(2), 199-208.

Talukdar, J., Kataki, K., Ali, E., Choudhury, B. N., Baruah, M. N., Bhattacharyya, M., . . . Medhi, S. (2021). Altered expression of TGF-β1 and TGF-βR2 in tissue samples compared to blood is associated with food habits and survival in esophageal squamous cell carcinoma. *Current Problems in Cancer, 45*(1), 100617.

Tangka, F. K., Subramanian, S., Beebe, M. C., Weir, H. K., Trebino, D., Babcock, F., & Ewing, J. (2016). Cost of operating central cancer registries and factors that affect cost: findings from an economic evaluation of Centers for Disease Control and Prevention National Program of Cancer Registries. *Journal of public health management and practice: JPHMP, 22*(5), 452.

Thrift, A. P., Nagle, C. M., Fahey, P. P., Russell, A., Smithers, B. M., Watson, D. I., . . . Study, A. C. S. C. F. U. (2012). The influence of prediagnostic demographic and lifestyle factors on esophageal squamous cell carcinoma survival. *Int J Cancer, 131*(5), E759-E768.

Thrift, A. P., Nagle, C. M., Fahey, P. P., Smithers, B. M., Watson, D. I., & Whiteman, D. C. (2012). Predictors of survival among patients diagnosed with adenocarcinoma of the esophagus and gastroesophageal junction. *Cancer Causes Control, 23*(4), 555-564.

Thrumurthy, S. G., Chaudry, M. A., Thrumurthy, S. S., & Mughal, M. (2019). Oesophageal cancer: risks, prevention, and diagnosis. *Bmj, 366*.

Toohey, K., Pumpa, K., Cooke, J., & Semple, S. (2016). Do activity patterns and body weight change after a cancer diagnosis? a retrospective cohort study. *Int J Health Sci Res, 6*(10), 110-117.

Torrente, M. P., Freeman, W. M., & Vrana, K. E. (2012). Protein biomarkers of alcohol abuse. *Expert review of proteomics, 9*(4), 425-436.

Tramacere, I., Gallus, S., Zuccaro, P., Colombo, P., Rossi, S., Boffetta, P., & La Vecchia, C. (2009). Socio-demographic variation in smoking habits: Italy, 2008. *Preventive medicine, 48*(3), 213-217.

Tramontano, A. C., Chen, Y., Watson, T. R., Eckel, A., Hur, C., & Kong, C. Y. (2019). Esophageal cancer treatment costs by phase of care and treatment modality, 2000-2013. *Cancer medicine, 8*(11), 5158-5172.

Tramontano, A. C., Nipp, R., Mercaldo, N. D., Kong, C. Y., Schrag, D., & Hur, C. (2018). Survival disparities by race and ethnicity in early esophageal cancer. *Digestive diseases and sciences, 63*(11), 2880-2888.

Trivers, K. F., De Roos, A. J., Gammon, M. D., Vaughan, T. L., Risch, H. A., Olshan, A. F., . . . Stanford, J. L. (2005). Demographic and lifestyle predictors of survival in patients with esophageal or gastric cancers. *Clinical Gastroenterology and Hepatology, 3*(3), 225-230.

U.S. Cancer Statistics Working Group. (2020). U.S. Cancer Statistics Data Visualizations Tool, based on 2019 submission data (1999–2017): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. Retrieved from www.cdc.gov/cancer/dataviz

UK Data Service. (2021). Retrieved from https://www.ukdataservice.ac.uk/

US Department of Health and Human Services. (2018). *Physical activities guidelines for Americans, 2nd edition*. Washington DC: US Department of Health and Human

Services Retrieved from https://health.gov/sites/default/files/2019-09/Physical_Activity_Guidelines_2nd_edition.pdf

US News. (2021). Where is Marijuana legal? A guide to Marijuana legalization. Retrieved from https://www.usnews.com/news/best-states/articles/where-is-marijuana-legal-a-guide-to-marijuana-legalization

Valsecchi, M. G., & Steliarova-Foucher, E. (2008). Cancer registration in developing countries: luxury or necessity? *The lancet oncology, 9*(2), 159-167.

Walter, V., Jansen, L., Hoffmeister, M., & Brenner, H. (2014). Smoking and survival of colorectal cancer patients: systematic review and meta-analysis. *Annals of oncology, 25*(8), 1517-1525.

Wang, L., Li, Y., Wang, N., Huang, X., Cao, S. R., & Zhou, R. M. (2020). Association of cytotoxic T-lymphocyte associated protein 4 gene polymorphisms with the risk and prognosis of oesophageal cancer in a high-incidence region from northern China. *International journal of immunogenetics, 47*(2), 180-187.

Wang, Q.-L., Xie, S.-H., Li, W.-T., & Lagergren, J. (2017). Smoking cessation and risk of esophageal cancer by histological type: systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute, 109*(12), djx115.

Wang, S. M., Fan, J. H., Jia, M. M., Yang, Z., Zhang, Y. Q., Qiao, Y. L., & Taylor, P. R. (2016). Body mass index and long-term risk of death from esophageal squamous cell carcinoma in a Chinese population. *Thorac Cancer, 7*(4), 387-392.

Wang, X. P., Li, X. H., Zhang, L., Lin, J. H., Huang, H., Kang, T., . . . Zheng, X. (2016). High level of serum apolipoprotein A-I is a favorable prognostic factor for overall survival in esophageal squamous cell carcinoma. *BMC cancer, 16*, 516. doi:10.1186/s12885-016-2502-z

Ward, Z. J., Bleich, S. N., Cradock, A. L., Barrett, J. L., Giles, C. M., Flax, C., . . . Gortmaker, S. L. (2019). Projected US state-level prevalence of adult obesity and severe obesity. *New England Journal of Medicine, 381*(25), 2440-2450.

Ward, Z. J., Long, M. W., Resch, S. C., Gortmaker, S. L., Cradock, A. L., Giles, C., . . . Wang, Y. C. (2016). Redrawing the US obesity landscape: bias-corrected estimates of state-specific adult obesity prevalence. *PloS one, 11*(3), e0150735.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology, 63*(8), 826.

WHO. (1995). *Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Consultation. WHO Technical Report Series Number 854* Geneva.

WHO Expert Consultation. (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet (London, England), 363*(9403), 157-163.

World Cancer Research Fund, & American Institute for Cancer Research. (2007). *Food, nutrition, physical activity, and the prevention of cancer: a global perspective* (Vol. 1). Washington DC: Amer Inst for Cancer Research.

World Health Organization. (2013). International classification of diseases for oncology (ICD-O)–3rd edition, 1st revision.

World Health Organization, & Centers for Disease Control. (2011). Tobacco questions for surveys: a subset of key questions from the Global Adult Tobacco Survey (GATS): global tobacco surveillance system.

WorldAtlas. (2021). 14 countries where drinking alcohol is illegal. Retrieved from https://www.worldatlas.com/articles/14-countries-where-drinking-alcohol-is-illegal.html

Worldometer. (2022). Countries in the world by population (2022). Retrieved from https://www.worldometers.info/world-population/population-by-country/

Wu, X., Chen, V. W., Andrews, P. A., Ruiz, B., & Correa, P. (2007). Incidence of esophageal and gastric cancers among Hispanics, non-Hispanic whites and non-Hispanic blacks in the United States: subsite and histology differences. *Cancer Causes Control, 18*(6), 585-593.

Xie, S.-H., Wahlin, K., & Lagergren, J. (2017). Cause of death in patients diagnosed with esophageal cancer in Sweden: a population-based study. *Oncotarget, 8*(31), 51800.

Yao, W., Qin, X., Qi, B., Lu, J., Guo, L., Liu, F., . . . Zhao, B. (2014). Association of p53 expression with prognosis in patients with esophageal squamous cell carcinoma. *International journal of clinical and experimental pathology, 7*(10), 7158.

Yu, X.-L., Yang, J., Chen, T., Liu, Y.-m., Xue, W.-p., Wang, M.-H., & Bai, S.-M. (2018). Excessive pretreatment weight loss is a risk factor for the survival outcome of esophageal carcinoma patients undergoing radical surgery and postoperative adjuvant chemotherapy. *Canadian Journal of Gastroenterology and Hepatology, 2018*.

Zhang, Y. (2013). Epidemiology of esophageal cancer. *World journal of gastroenterology: WJG, 19*(34), 5598.

Zhou, R.-M., Li, Y., Wang, N., Niu, C.-X., Huang, X., Cao, S.-R., & Huo, X.-R. (2020). PARP1 Gene Polymorphisms and the Prognosis of Esophageal Cancer Patients from Cixian High-Incidence Region in Northern China. *Asian Pacific Journal of Cancer Prevention, 21*(10), 2987-2992.

Zhou, R. M., Li, Y., Wang, N., Huang, X., & Cao, S. R. (2018). Phospholipase C ε-1 gene polymorphisms and prognosis of esophageal cancer patients from a high-incidence region in northern China. *Mol Clin Oncol, 8*(1), 170-174. doi:10.3892/mco.2017.1475

Zucker, D. M., & Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in medicine, 27*(11), 1911-1933.