Incubation and latency time estimation for SARS-CoV-2

Arntzen, V.H.

This chapter will be published soon as Vera H. Arntzen, Marta Fiocco, Inge M.M. Lakeman, Maartje Nielsen and Mar Rodríguez-Girondo. A new inverse probability of selection weighted Cox model to deal with outcome-dependent sampling in survival analysis. Biometrical Journal.

A new inverse probability of selection weighted Cox model to deal with outcome-dependent sampling in survival analysis

Contents

5.1 Introduction 129
5.2 Weighted Cox regression to deal with outcome-dependent sampling 131
5.3 Simulation study 137
5.4 Real data applications 142
5.5 Discussion 147
5.6 Supplementary material 149

Abstract

Motivated by the study of genetic effect modifiers of cancer, we examined weighting approaches to correct for ascertainment bias in survival analysis. Outcome-dependent sampling is common in genetic epidemiology leading to study samples with too many events in comparison to the population and an overrepresentation of young, affected subjects. A usual approach to correct for ascertainment bias in this setting is to use an inverse probability-weighted Cox model, using weights based on external available population-based age-specific incidence rates of the type of cancer under investigation. However, the current approach is not general enough leading to invalid weights in relevant practical settings if oversampling of cases is not observed in all age groups. Based on the same principle of weighting observations by their inverse probability of selection, we propose a new, more general approach. We show the advantage of our new method using simulations and two real datasets. In both applications the goal is to assess the association between common susceptibility loci identified in Genome Wide Association Studies (GWAS) and cancer (colorectal and breast) using data collected through genetic testing in clinical genetics centers.

Keywords Cox regression □ genetic epidemiology □ outcome-dependent sampling □ survival analysis □ weighting

5.1 Introduction

Outcome-dependent sampling is common in genetic epidemiology. Since harmful variants in cancer associated high risk genes are typically rare, an efficient sampling strategy to find carriers of these variants is to oversample affected individuals with a family history of a specific disease. For example, carriers of pathogenic variants in the Lynch syndrome associated gene *PMS2* and the breast- and ovarian cancer associated genes *BRCA1* and *BRCA2*, are often detected through genetic screening programs in which testing is targeted to families with multiple cases. Due to this testing strategy, the available study cohorts to investigate modifiers of cancer risk are often non-representative samples of the population of interest: carriers with an early diagnosis of cancer are more frequently included in the sampled population compared to those with delayed cancer diagnoses or individuals who remain disease-free.

In the context of survival analysis, family-based outcome-dependent sampling results in an over-representation of events and short lifetimes, which without adjustment, leads to biased estimates of covariate effects when using, for example, a Cox proportional hazards model. This happens because the sampling mechanism affects the joint distribution of the time-to-event and covariate.

To solve this problem, two main approaches have been proposed in the literature: methods based on retrospective likelihood [Carayol and Bonaïti-Pellié, 2004; Chatterjee et al., 2006; Barnes et al., 2013] and the weighted cohort method [Antoniou et al., 2005] based on weighted Cox regression. The general idea of the methods based on retrospective likelihood is to formulate the likelihood of the observed covariate values conditional on the observed outcomes. These methods typically require to know the familial relations within the sample and the distribution of the covariate of interest, leading to analytically complex and computationally intensive methods. When the overall age-specific incidence rates in the population of interest are known, an alternative approach to estimate the association between a set of covariates and time to cancer diagnosis under outcome-dependent sampling is to use a weighted Cox regression model [Antoniou et al., 2005]. The general idea is to propose a weighting scheme with different weights for affected (observed events) and unaffected (right-censored) individuals according to an external source so that the resulting weighted sample mimics the true target population [Antoniou et al., 2005; Barnes

et al., 2012] in terms of the age-specific proportions of affected and unaffected individuals.

Due to its simplicity, this is an attractive approach. However, the proposed weighted scheme has some limitations: it often leads to invalid weights in relevant practical situations since it is only workable under particular sampling schemes, such as those involving substantial oversampling of cases.

The primary objective of this study is to introduce a novel and more versatile inverse probability of selection weighting scheme, utilizing population-based age-specific incidence rates of the event of interest. This leads to the development of a generalized weighted cohort method capable of accommodating arbitrary levels of outcome-dependent sampling, offering an improved alternative to the existing approach. As a secondary goal, we aim to conduct a sensitivity analysis to assess the performance of the weighted approaches in the presence of unobserved heterogeneity, particularly exploring within-family correlations arising from shared, unobserved factors. Despite the frequent inclusion of multiple family members in studies employing the original weighted cohort method (see Supplemental Table 5.5 for details), the influence of unobserved heterogeneity in this context remains unexplored. This aspect merits thorough investigation.

The rest of the paper is organised as follows. In Section 5.2, the commonly used weighted cohort Cox approach is revisited and its assumptions are discussed. A new alternative weighting scheme is proposed in Section 5.2.2. In Section 5.3, both weighting schemes are compared by means of an intensive Simulation study. In Section 5.4, we present two real data illustrations. In both illustrations, the role of genetic variants as modifiers of cancer risk is studied using datasets of affected individuals and family members ascertained through genetic counseling in a clinical genetic center. In the first application, we focus on colorectal cancer in carriers of the pathogenic variant *PMS2*, and in the second application, we analyse the association between a Polygenic Risk Score (PRS) based on common breast cancer-associated variants and breast cancer risk in multiple case families. Main conclusions, recommendations, and a final discussion follow in Section 5.5.

5.2 Weighted Cox regression to deal with outcome-dependent sampling

Let T be the time to event of interest in the target population of interest. The typical target population in our context comprises individuals who are carriers of a specific rare mutation of interest. Denote by C the right censoring time, assumed to be uninformative. Denote by Z the covariate of interest. Since sampling schemes in genetic epidemiology are typically family-based, denote the observed sample information by $(t_{ij}, \delta_{ij}, z_{ij})$, where $i = 1, \ldots, n_j$ index all included individuals in the sample belonging to family j. It is important to note that even though the sampling is family-based, this does not necessarily imply the inclusion of multiple family members in the resulting sample. Assume that N families are observed, with varying observed size $n_1, \ldots, n_N$, so that $n = \sum_{j=1}^{N} n_j$ individuals are included in the sample. The observed time to event for individual i in family j (denoted from now on by ij), is given by $t_{ij} = \min(T_{ij}, C_{ij})$. Define the non-censoring indicator $\delta_{ij} = I(T_{ij} \leq C_{ij})$ where δ_{ij} is 1 if the event is observed or 0 if observation ij is right censored. z_{ij} denotes covariate value for individual ij.

The observed data is collected through a family-based outcome-dependent sampling scheme. The selection process begins by testing the first individual in a family, focusing on those diagnosed with cancer at a young age and with a family history of cancer. If this initial individual tests positive for the mutation, the rest of the family is invited to participate in genetic testing. This approach may identify additional carriers of the mutation within the family, though this is not always the case. Consequently, due to variations in age and family history criteria across studies, influenced by disease severity and prevalence, the level of outcome-dependent sampling varies across studies. Despite the diversity in the final configurations, samples of carriers of rare genetic variants obtained via genetic testing typically result in an over-representation of young cases in the sample.

A common approach to estimate the effect of covariate Z on T is to use the Cox proportional hazards model with hazard function $h(t|z) = h_0(t)\exp(\beta z)$ where $h_0(t)$ is the baseline hazard. With prospective cohort data, the parameter β can be estimated maximizing the partial likelihood. However, the over-representation of events and short event times in the sample due to outcome-dependent sampling affects the risk set composition along the follow-up time in comparison to the true population, which may

result in biased estimation of the covariate effect. A possible solution to this problem is to consider a weighted Cox model using external information about the distribution of T in the population to construct weights reflecting individuals' selection probabilities.

5.2.1 The weighted cohort approach revisited

When T represents the age at cancer diagnosis, or another common disease, registry data about the marginal distribution of T in the target population is often accessible. In practical scenarios, the available external information is typically aggregated into K distinct age intervals, defined as $I_1 = [a_0, a_1)$, $I_2 = [a_1, a_2)$,..., $I_K = [a_{K-1}, a_K)$. For cancer studies, the commonly available external data comprises the population cancer incidence rate μ_k for each age interval I_k. The seminal work of Antoniou et al. [2005] introduced a weighted Cox regression model with sampling weights derived in such a way that the incidence rates in each interval I_k, in the resulting pseudo-population after weighting, align with the incidence rates μ_k in the target population.

However, before presenting the specific calculation of these weights, it is essential to acknowledge two main assumptions regarding the observed data in this context, given that the externally available data is discrete in time. First, we assume constant hazards within each interval I_k for $k = 1, \ldots, K$. Second, we assume that right-censoring is also discrete and occurs at the specified time points defining the intervals. This implies that if censored observations happen to fall within interval I_k, we assume that the censoring took place at point a_k. When these two prerequisites are met, the marginal distribution of T in the resulting weighted pseudo-population, based on interval-specific incidence rates, will follow the same distribution as in the reference population.

Let r_k denote the number of individuals experiencing the event within the age interval, $I_k = [a_{k-1}, a_k)$, $k = 1, \ldots K$. Similarly, s_k denotes the number of individuals right-censored within the age interval I_k (i.e. follow-up ends between age a_{k-1} and a_k without the event being observed). The term $p_k = \sum_{\{ij: t_{ij} \in I_k, \delta_{ij}=1\}} t_{ij}$ denotes the total follow-up time accumulated by all r_k individuals experiencing the event in age interval I_k; the equivalent total follow-up time accumulated by the s_k right-censored individuals is denoted by $q_k = \sum_{\{ij: t_{ij} \in I_k, \delta_{ij}=0\}} t_{ij}$. Then, all r_k cases in interval I_k are assigned weight w_k and all s_k right-censored individuals in interval I_k are assigned weight v_k such that:

$$\mu_k = \frac{w_k r_k}{w_k p_k + v_k q_k + (a_k - a_{k-1}) \sum_{l>k} (v_l s_l + w_l r_l)}. \tag{5.1}$$

In the right part of Expression (5.1) the weighted total of affected observations is divided by the weighted total of observations at risk. Then, this weighted ratio is imposed to be equal to the population incidence rate μ_k. As a result, after weighting the sample age-specific incidence rates resemble the age-specific incidence rates of the population. However, since equation (5.1) alone does not guarantee unique weights w_k and v_k, the following constraint is incorporated to guarantee unique weights:

$$\frac{w_k r_k + v_k s_k}{r_k + s_k} = 1. \tag{5.2}$$

Combining equations (5.1) and (5.2) provides unique expressions for v_k and w_k:

$$w_k = \frac{\mu_k (q_k (r_k + s_k) + (a_k - a_{k-1}) s_k \sum_{l>k} (r_l + s_l))}{r_k s_k + \mu_k (q_k r_k - p_k s_k)}, \tag{5.3}$$

where $\sum_{l>k}(r_l + s_l)$ are all observations in age groups older than k. The weight equation for censored individuals is given by

$$v_k = \frac{1}{s_k}(r_k + s_k - w_k r_k). \tag{5.4}$$

Once weights v_k, w_k for each age interval I_k, $k = 1, \ldots, K$ are calculated, the regression parameter β can be estimated using the following weighted score equation:

$$U_a(\beta) = \sum_{ij:\delta_{ij}=1} z_{ij} - \sum_{ij:\delta_{ij}=1} \frac{\sum\limits_{l \epsilon R(t_{ij})} W_l z_l \exp[\beta z_l]}{\sum\limits_{l \epsilon R(t_{ij})} W_l \exp[\beta z_l]}, \tag{5.5}$$

where $R(t_{ij})$ is the set of individuals still at risk just before t_{ij}, i.e. $R(t_{ij}) = \{l : t_{ij} \leq t_l\}$, and weight W_{ij} for individual ij ($i = 1, \ldots, n_j$, $j = 1, \ldots, N$) is defined as

$$W_{ij} = \begin{cases} w_k, & \text{if } \delta_{ij} = 1 \text{ and } t_{ij} \in [a_{k-1}, a_k) \\ v_k, & \text{if } \delta_{ij} = 0 \text{ and } t_{ij} \in [a_{k-1}, a_k). \end{cases} \tag{5.6}$$

The derivation and unbiasedness of the estimator resulting from Expression (5.5) are outlined in Supplement 5.6.1. We followed the same reasoning as presented by Mandel *et*

al. [2017], who introduced an inverse probability weighted Cox model to address double truncation. According to the authors, their findings extend beyond the context of double truncation to deal with a broader range of biased sampling scenarios. To elaborate further, Supplement 5.6.1 provides a detailed explanation. Crucially, in our setting, we assume conditional independence between the selection event and covariate values, given the observed event time. This assumption allows us to apply the results of Mandel *et al.* [2017], with the selection event treated as a function of the observed event times.

A number of conditions are required to guarantee finite and positive weights w_k and v_k, namely:

$$r_k > 0 \tag{5.7}$$

$$s_k > 0 \tag{5.8}$$

$$r_k > \mu_k p_k \frac{s_k}{s_k + \mu_k q_k} \tag{5.9}$$

$$w_k < 1 + \frac{s_k}{r_k}. \tag{5.10}$$

Conditions (5.7) and (5.9) are required to get proper w_k weights for the cases, while conditions (5.8) and (5.10) are required to get valid v_k weights for those that are censored. Condition (5.7) implies the observation of events in all the considered intervals while condition (5.8) implies the presence of right-censored observations in all intervals under consideration. Conditions (5.9) and (5.10) are more difficult to interpret and evaluate beforehand, but they are both related to the level of oversampling of events. As discussed by [Antoniou et al., 2005], if oversampling of events occurs in all considered age groups both conditions are typically fulfilled. However, as we will show in our real data application, oversampling of young cancer cases is the norm in genetic epidemiology, but not necessarily the case at older ages, so condition (5.9) and especially (5.10) might not be fulfilled in relevant practical scenarios.

Under oversampling of events at interval I_k, condition (5.9) is verified since $r_k > \mu_k p_k$, i.e., the observed number of events in interval I_k is larger than the expected number of events assuming the population incidence rate (μ_k). Actually, since $\mu_k q_k$ is usually positive, $0 < \frac{s_k}{s_k + \mu_k q_k} < 1$ in general which implies that condition (5.9) is fulfilled even when no oversampling of events is observed in interval I_k. However, condition (5.10) is cumbersome. Since it involves the estimated weight for events w_k together with the ratio of events and

right-censored observations in interval I_k ($\frac{s_k}{r_k}$), this condition is often not satisfied when there is no clear oversampling of cases in interval I_k. In such situations, w_k can still be calculated but it becomes small, which leads to violating condition (5.10).

When any of the conditions (5.7)-(5.10) are not satisfied, the weighted cohort method can still be applied by merging intervals, however then the method becomes less precise and dependent on sample specific characteristics which may hamper comparability among studies using this method.

We therefore propose an alternative, more general weighting scheme which allows to overcome the aforementioned limitations.

5.2.2 The new generalised weighted cohort approach

We propose a new weighting scheme to correct outcome-dependent sampling using external information. In contrast to the previous weighted cohort method, the new approach is more general, as it can be applied with arbitrary levels of over or under-representation of events.

Similar to the original method, the newly proposed weights represent sampling probabilities given the observed time to event of each individual so that the resulting pseudo-population matches the target population of reference in terms of the marginal distribution of T. The same score function (5.5) and justification of its validity applies. However, here we take a different approach to derive the weights. Instead of directly using the incidence rates, we focus on the risk sets at the beginning at each interval I_k and weight the individuals so that the resulting weighted risk set presents the same ratio of events and non-events as one would expect if the sample would have been randomly drawn from the target population.

Let N_k denote the number of individuals at risk (those who did not experience the event yet) at the beginning of the interval I_k in our sample, denoted by $\mathcal{S}_O$, potentially drawn under an outcome-dependent sampling mechanism. Now denote by $\mathcal{S}_P$ a hypothetical random sample of the target population with the same N_k number of individuals at risk at the beginning of the interval I_k. In both cases, N_k can be split into two disjoint parts: (i) the number of individuals that experience the event within the interval I_k and (ii) those experiencing the event in later intervals. However, if $\mathcal{S}_O$ is obtained using outcome-dependent sampling, the expected number of individuals belonging to each of these two parts in $\mathcal{S}_O$ and $\mathcal{S}_P$ will, in general, differ.

For the hypothetical random sample $\mathcal{S}_P$, N_k can be decomposed as follows:

$$N_k = N_k S_k + N_k(1 - S_k) \tag{5.11}$$

where $S_k = P(T > a_k | T > a_{k-1})$ represents the conditional probability of experiencing the event in a later time interval than interval I_k given that the event has not been experienced before interval I_k in the reference population. S_k can be directly calculated from the typically available population cancer incidence rates μ_k for each age interval I_k, since $S_k = e^{-\mu_k(a_k - a_{k-1})}$, $k = 1, \dots K$. Accordingly, $1 - S_k$ is the probability of experiencing the event in the interval I_k given that it has not been experienced before, in the reference population. From Expression (5.11) follows that the ratio between events and non-events in interval I_k in the reference population is given by $\frac{1-S_k}{S_k}$. Under the assumption of constant hazards within each pre-specified interval I_k, the ratio of events to non-events completely determines the incidence rate in interval I_k, thereby entirely characterizing the marginal distribution of T.

The same decomposition of the risk set at the beginning of interval I_k can be made for the observed sample $\mathcal{S}_O$, potentially subject to outcome-dependent sampling:

$$N_k = N_k S_k^o + N_k(1 - S_k^o) \tag{5.12}$$

where S_k^o is the observed proportion of individuals at risk at time a_{k-1} experiencing the event beyond I_k, calculated with the sample data.

In our new approach, we keep those subjects non-experiencing the event at interval I_k unweighted ($v_k = 1$) while we assign specific weights (w_k) to those subjects experiencing the event of interest in interval I_k making use of the decompositions given by Expressions (5.11) and (5.12). Specifically, weights w_k correct the oversampling (or undersampling) of cases, such that the ratio between events and non-events on the interval I_k in the resulting pseudo-population after weighting is the same as in the reference population:

$$w_k = \frac{(1 - S_k)}{S_k} \frac{S_k^o}{(1 - S_k^o)}. \tag{5.13}$$

Equation (5.13) illustrates that the population ratio between events and non-events within interval I_k, denoted as $\frac{1-S_k}{S_k}$, is multiplied by the inverse quantity based on the observed data, $\frac{1-S_k^o}{S_k^o}$. After weighting, the composition of the risk set within interval I_k, in terms of the ratio of events to non-events, resembles the composition of the risk set within

interval I_k in the reference population. As a result, under oversampling of cases, weights for affected individuals in interval I_k are $w_k < 1$, representing the inverse of the probability of being selected. Alternatively, under undersampling of cases, $w_k > 1$. Interestingly, in absence of outcome-dependent sampling, i.e. under random sampling, $w_k = 1$ and the new method coincides with the regular unweighted Cox model.

With our new proposal, two conditions need to be fulfilled in order to get valid weights: $(1 - S_k^o) > 0$, $k = 1, \ldots, K$ and $S_K > 0$. The first condition $(1 - S_k^o) > 0$ is satisfied if events are observed in each interval I_k, so as the original weighted cohort method, observation of events in all group ages is a requirement of our new method. However, the new method does not require the presence of right-censoring which makes it a more general and natural approach. The condition $S_K > 0$ only involves the last interval and implies that the method is suitable for studying events not experienced by a part of the population during the relevant follow-up time. This is a mild condition that is always satisfied when studying defective distributions ($S(\infty) > 0$) such as time to cancer or other diseases since not all population members will develop the event of interest. Even if our interest would be to study time to death or the target population would be a highly susceptible population to a specific cancer with lifetime risk of 1, the new weights could still be applied with an appropriate choice of the upper limit of the last interval K.

Once the weights are calculated, the regression parameter β can be estimated using the weighted score equation (5.5) and robust estimates of the standard errors can be obtained using a sandwich estimator, as proposed by Antoniou *et al.* [2005] in the original weighted cohort approach.

In summary, both the existing weighted cohort and the new generalised weighted cohort approaches generate pseudo-populations by means of inverse probability of selection weighting, but these pseudo-populations are different. The method developed in this study, is more general since it does not require oversampling or undersampling of events in all or at specific intervals and does not make assumptions about the right-censoring distribution.

5.3 Simulation study

A simulation study was conducted to assess the new generalized weighted cohort method's performance and compare it with the existing approach in several scenarios intended to

mimic relevant situations in practice. We consider two main simulation settings. First, we generate data so that the (weighted) Cox approaches are well specified. Second, we study the sensitivity of the weighted methods to model misspecification due to the presence and failure to adjust for unobserved heterogeneity.

5.3.1 Simulation setup I

Simulated data was generated using the following model:

$$\lambda_{ij}(t) = \lambda_0 \exp(\beta z_{ij}), \tag{5.14}$$

where t is the observed event time, $\lambda_0 = \frac{1}{60}$ represents the constant baseline hazard, Z is a continuous covariate assumed to be normally distributed ($Z \sim N(0,1)$) and β represents the associated log-hazard ratio. If the resulting event times were larger than 100, these were set to 100. Censoring times were sampled from an exponential distribution ($C \sim Exp(60)$) and the family size in the population is set to n_j (family size of size $n_j = 2$ and 5 members were considered). In each Monte Carlo trial, we generated N families (N = 250, 500, 750). Family-based outcome-dependent sampling was implemented by including families in the sample if for at least n_A family members the event was observed before the end of follow-up ($n_A = 1, 3$). The different combinations of n_A and n_j lead to three different scenarios with increasing level of outcome-dependent sampling: scenario 1 (A1) with $n_j = 5$ and $n_A = 1$ represents the mildest level of selection, scenario 2 (A2) with $n_j = 5$ and $n_A = 3$ represents a medium level of outcome dependent selection, and scenario 3 (A3) with $n_j = 2$ and $n_A = 1$ represents the strongest level of outcome dependent sampling in the simulation study. Moreover, all included families had at least one 'young affected' defined as having observed event time smaller than the first quartile of simulated T distribution. This mimics the common practice in clinical genetics centers: families are invited to participate in genetic studies when a young family member is diagnosed with the event at a young age. In terms of covariate effect, the null case ($\beta = 0$) and two alternative scenarios ($\beta = 0.3, 1$) were considered.

For each considered value of β, the underlying population was generated by simulating a large data set (N = 200 000, $n_j = 1$) without ascertainment and it was used to approximate the population hazards needed to calculate the weights. In all simulation scenarios, we considered five intervals, each with a width of 20 years. Relative bias, mean square

error and coverage proportions of the 95% confidence intervals are reported in Table 5.1. Moreover, the proportion of invalid weights in the M Monte Carlo trials is reported for the two weighted methods.

5.3.2 Simulation setup II

In the previous simulation setting, we have assumed, as the proposed models in Section 5.2, that differences among individuals in terms of hazards can be fully accounted for by including covariates in the Cox proportional hazards model. However, when samples contain multiple members of the same family (often the case when applying the traditional weighted cohort approach as shown in Supplemental Table 5.5), unmeasured hetereogeneity may arise since members of the same family often share common unmeasured characteristics such as genetic, social, dietary or other factors. In this second simulation setting, in order to introduce such unmeasured heterogeneity in the simulated data, we consider an extension of the data generation model specified in Expression (5.14) by adding a latent (frailty) term, U, shared by all members of the same family:

$$\lambda_{ij}(t) = u_j \lambda_0 \exp(\beta z_{ij}), \tag{5.15}$$

where $u_j \sim \Gamma(1, \theta)$ is a latent term (frailty), shared by the n_j members of a given family j. The larger the value of the variance θ, the more family members are alike and the larger the difference between families, yielding larger unobserved family effects. We consider two different values of within-family correlation: 'low' ($\theta = 0.1$) and 'large' ($\theta = 1$). Note that the latent frailty U and the covariate under investigation Z are independent. We expect that, as in the traditional unweighted Cox regression context [Henderson and Oman, 1999], ignoring the presence of U introduces bias in the hazard ratio estimation due to non-collapsability, even when U is independent of the covariate of interest Z. However, we also expect that such bias diminishes in the presence of outcome-dependent sampling and with the use of inverse probability of selection weighted Cox models.

The rest of the simulation settings were as in the previous Simulation setting I, except N, the number of families was fixed to 500 in this setting. For each scenario, weighted Cox models were estimated using the traditional and the generalized weighting scheme. Results obtained with the standard choices of using an unweighted Cox model or a shared gamma frailty model (unweighted) are also reported. Note that the application of the weighted

Table 5.1: Simulation I. Relative bias (reBias), mean square error (MSE) and coverage probability (Coverage) for $\widehat{\beta}$ along 1000 trials. A1: mild level of ascertainment. A2: medium level of ascertainment A3: strong level of ascertainment. N: number of families. For the weighted approaches, the proportion of invalid (negative) weights along 1000 trials is also reported.

			Unweighted			Weighted cohort				Generalized weighted cohort			
β	Scenario	N	reBias	MSE	Coverage	reBias	MSE	Coverage	Invalid weights	reBias	MSE	Coverage	Invalid weights
$\beta = 0$	A1	250	< 0.001	0.003	0.947	< 0.001	0.003	0.939	0.002	< 0.001	0.003	0.944	0.000
		500	0.002	0.002	0.953	< 0.001	0.011	0.953	0.000	< 0.001	0.002	0.949	0.000
		750	< 0.001	0.001	0.942	-0.002	< 0.001	0.939	0.000	-0.002	0.001	0.947	0.000
	A2	250	-0.003	0.004	0.942	< 0.001	0.006	0.938	0.005	< 0.001	0.005	0.940	0.000
		500	0.001	0.002	0.950	< 0.001	0.003	0.943	0.000	< 0.001	0.003	0.939	0.000
		750	-0.002	0.001	0.953	-0.003	0.002	0.941	0.000	-0.002	0.002	0.942	0.000
	A3	250	0.003	0.012	0.948	0.010	0.029	0.925	0.570	0.001	0.034	0.914	0.000
		500	< 0.001	0.006	0.939	< 0.001	0.012	0.937	0.191	< 0.001	0.014	0.935	0.000
		750	-0.003	0.004	0.957	0.002	0.007	0.947	0.058	0.002	0.008	0.951	0.000
$\beta = 0.3$	A1	250	-0.044	0.003	0.948	-0.091	0.004	0.912	0.000	-0.041	0.004	0.944	0.000
		500	-0.046	0.002	0.932	-0.092	0.002	0.887	0.000	-0.042	0.002	0.931	0.000
		750	-0.036	0.001	0.943	-0.100	0.002	0.841	0.000	-0.049	0.001	0.919	0.000
	A2	250	-0.200	0.008	0.818	-0.107	0.007	0.916	0.006	-0.129	0.007	0.893	0.000
		500	-0.195	0.005	0.733	-0.107	0.004	0.907	0.000	-0.128	0.004	0.877	0.000
		750	-0.210	0.005	0.587	-0.115	0.003	0.871	0.000	-0.138	0.003	0.837	0.000
	A3	250	-0.221	0.017	0.885	-0.012	0.024	0.941	0.562	0.071	0.033	0.933	0.000
		500	-0.229	0.010	0.851	-0.034	0.013	0.933	0.208	0.006	0.015	0.937	0.000
		750	-0.220	0.008	0.795	-0.037	0.008	0.947	0.077	-0.005	0.010	0.938	0.000
$\beta = 1$	A1	250	-0.063	0.007	0.805	-0.036	0.006	0.919	0.000	-0.001	0.005	0.946	0.000
		500	-0.064	0.006	0.655	-0.037	0.003	0.893	0.000	-0.003	0.002	0.956	0.000
		750	-0.067	0.034	0.463	-0.038	0.003	0.822	0.000	-0.004	0.002	0.943	0.000
	A2	250	-0.173	0.035	0.294	-0.042	0.008	0.926	0.000	-0.059	0.010	0.891	0.000
		500	-0.173	0.032	0.053	-0.045	0.005	0.911	0.000	-0.060	0.007	0.813	0.000
		750	-0.175	0.032	0.009	-0.047	0.004	0.837	0.000	-0.062	0.006	0.738	0.000
	A3	250	-0.270	0.083	0.228	0.024	0.030	0.907	0.025	0.070	0.043	0.886	0.000
		500	-0.269	0.078	0.054	< 0.001	0.015	0.929	0.014	0.043	0.020	0.925	0.000
		750	-0.269	0.076	0.011	-0.002	0.010	0.938	0.000	0.041	0.014	0.917	0.000

approaches is not possible in the context of frailty models since the correct estimation of the weights would require knowing the true value of the frailty variance, which cannot be correctly estimated under outcome-dependent sampling.

5.3.3 Simulation results

Simulation I

We first present the results obtained when data is generated under the assumption of fully observed heterogeneity. Both the level of outcome-dependent sampling and the covariate effect size determine the observed differences among the studied methods. If the covariate effect is strong ($\beta = 1$), the naive unweighted method is outperformed by the weighted approaches, even when the level of outcome-dependent selection is low (scenario A1). When the covariate effect is weaker ($\beta = 0.3$), and the level of ascertainment is medium

(scenario A2) or high (scenario A3), both weighted cohort methods perform similarly and clearly outperform the naive, unweighted approach. Under a weak level of ascertainment (scenario A1), the new generalized weighted cohort method performs as well as the naive unweighted approach and they slightly outperform the traditional weighted cohort approach. Importantly, the traditional weighted approach is often not applicable when the assumed covariate effect is weak. Negative weights are often obtained in this setting due to violation of condition (Eq. 5.10). The same problem is observed in the null case scenario when assuming $\beta = 0$. The new generalized weighted cohort method does not suffer from this problem, yielding valid and satisfactory results in all studied scenarios.

Simulation II

Table 5.2: Simulation II. Relative bias (reBias), mean square error (MSE) and coverage probability (Coverage) for $\widehat{\beta}$ along 1000 trials. A1: mild level of ascertainment. A2: medium level of ascertainment A3: strong level of ascertainment. N: number of families. Data is generated according to a shared frailty model with frailty variance θ. For the weighted approaches, the proportion of invalid (negative) weights along 1000 trials is also reported.

				Unweighted			**Weighted cohort**				**Generalized weighted cohort**				**Shared frailty**		
θ	β	Scenario	N	reBias	MSE	Coverage	reBias	MSE	Coverage	Invalid weights	reBias	MSE	Coverage	Invalid weights	reBias	MSE	Coverage
$\theta = 0.1$	$\beta = 0$	A1	500	0.002	0.002	0.937	0.003	0.002	0.949	0.000	0.003	0.002	0.953	0.000	< 0.001	0.002	0.939
		A2	500	< 0.001	0.002	0.953	0.003	0.005	0.952	0.000	0.003	0.004	0.952	0.000	< 0.001	0.002	0.954
		A3	500	< 0.001	0.006	0.945	-0.002	0.012	0.941	0.036	< 0.001	0.015	0.936	0.000	0.002	0.006	0.942
	$\beta = 0.3$	A1	500	-0.065	0.002	0.918	-0.092	0.002	0.895	0.000	-0.060	0.002	0.934	0.000	-0.069	0.002	0.914
		A2	500	-0.208	0.006	0.705	-0.087	0.005	0.926	0.000	-0.119	0.004	0.896	0.000	-0.217	0.006	0.688
		A3	500	-0.239	0.011	0.824	-0.037	0.013	0.951	0.206	-0.017	0.016	0.941	0.000	-0.228	0.011	0.848
	$\beta = 1$	A1	500	-0.094	0.010	0.363	-0.059	0.006	0.766	0.000	-0.032	0.003	0.895	0.000	-0.092	0.010	0.394
		A2	500	-0.195	0.040	0.027	-0.057	0.007	0.865	0.000	-0.080	0.010	0.734	0.000	-0.194	0.049	0.030
		A3	500	-0.278	0.083	0.045	-0.012	0.017	0.921	0.000	0.017	0.022	0.912	0.000	-0.278	0.083	0.063
$\theta = 1$	$\beta = 0$	A1	500	0.001	0.002	0.945	< 0.001	0.003	0.940	0.000	0.001	0.002	0.939	0.000	< 0.001	0.006	0.954
		A2	500	< 0.001	0.002	0.944	-0.004	0.008	0.933	0.000	-0.002	0.006	0.926	0.000	-0.002	0.002	0.957
		A3	500	0.001	0.007	0.947	-0.003	0.025	0.940	0.000	-0.002	0.023	0.912	0.000	-0.001	0.007	0.950
	$\beta = 0.3$	A1	500	-0.238	0.007	0.613	-0.130	0.005	0.885	0.000	-0.164	0.006	0.854	0.000	-0.131	0.004	0.830
		A2	500	-0.279	0.009	0.567	0.070	0.008	0.939	0.070	-0.043	0.006	0.940	0.000	-0.231	0.007	0.667
		A3	500	-0.310	0.015	0.793	0.076	0.044	0.879	0.480	-0.051	0.039	0.899	0.000	-0.323	0.016	0.756
	$\beta = 1$	A1	500	-0.263	0.072	0.001	-0.151	0.028	0.391	0.000	-0.177	0.037	0.262	0.000	-0.110	0.015	0.391
		A2	500	-0.293	0.090	0.001	0.001	0.010	0.908	0.000	-0.081	0.014	0.794	0.000	-0.188	0.040	0.123
		A3	500	-0.355	0.135	0.024	-0.045	0.040	0.858	0.492	-0.100	0.051	0.841	0.000	-0.355	0.134	0.028

Table 5.2 shows the results assuming the presence of unobserved family-shared heterogeneity. When unmeasured within-family correlation is mild ($\theta = 0.1$), we found similar results as in the previous simulation study: weighted methods perform similarly and provide better results than the unweighted model. Also, weighted methods, which deal with outcome-dependent sampling but ignore the presence of unobserved heterogeneity, perform better than a gamma shared frailty model which deals with shared unobserved heterogeneity but ignores outcome-dependent sampling.

For strong within-family correlation ($\theta = 1$) the performance of both weighted methods is, in general, good if the level of ascertainment is moderate or high (scenarios A2 and A3). If the level of ascertainment is mild (scenario A1), weighted methods would still outperform the traditional unweighted Cox approach but a shared frailty model seems a better choice in this setting. Bias is still noticeable with the shared frailty model, but of a smaller magnitude. Finally, the original weighted cohort also provided negative weights in this setting with unobserved family-shared heterogeneity, while the newly proposed generalized weighted cohort method proved to be more robust.

Overall, the simulation results show that the new generalized weighted cohort method is preferred over the original weighted cohort approach proposed by Antoniou *et al.* 2005. The original weighted cohort method performs well, in general, in the presence of a combination of a strong covariate effect and strong outcome-dependent sampling, as expected. However, its applicability is restricted to certain scenarios, and it is not general enough. Our sensitivity analysis, based on assuming the existence of unobserved heterogeneity, shows that inverse probability of selection weighted Cox models can still perform properly in the presence of mild unobserved family-shared heterogeneity, but they lead to biased results when the size of the frailty variance is large. Still, weighted methods seem to be preferred over the alternative approach of ignoring outcome-dependent sampling and fitting a shared frailty model if the level of outcome-dependent sampling is strong. If the level of ascertainment is mild, the results indicate a preference for the shared frailty model.

5.3.4 Software implementation

The generalized weighted cohort method developed in this work was implemented in the user-friendly `R` package `wcox`, which can be downloaded from CRAN and `https://github.com/vharntzen/wcox`.

5.4 Real data applications

We present two applications to illustrate the performance of the new generalized weighted cohort method compared to the traditional approaches on real data. In both applications, the goal is to assess the association between common susceptibility *loci* (gene locations on the chromosome) identified in Genome Wide Association Studies (GWAS) and cancer,

using data collected through genetic testing in clinical genetics units. Specifically, the first application is devoted to study the association between a Single Nucleotide Polymorphism (SNP) and colorectal cancer (CRC) in carriers of a pathogenic variant in the *PMS2* gene while the second one focuses on the association of a 161 SNP-based polygenic risk score with breast cancer. The selection of both datasets was based on family history of cancer with oversampling of cancer cases with the aim of finding carriers of certain genetic variants. As a result, the sample used in the first application is composed of *PMS2* mutation carriers. In the second application, the sample is composed of women with a family history of breast cancer and without BRCA1 or BRCA2 mutations.

5.4.1 Application to colorectal cancer

In this application, we consider a sample of male carriers of the germline PMS2 mutation. Motivated by the previous promising findings reported by Ten Broeke *et al.* [2018], we studied the association between the SNP rs1321311 and colorectal cancer in men. The sample consisted of 191 males belonging to 102 different families collected in eight Dutch clinical genetics centers between 2007 and 2016. Details on the selection criteria can be found in Ten Broeke *et al.* [2015]. The distribution of the number of individuals belonging to the same family was very skewed, the mean number of individuals per family was 1.83 and most of the families (55 %) contributed with one single member (Figure 2, left panel). The last age of follow-up ranged between 25 and 88 years, but given that no events were observed after 75 years old, we censored observations at 75 years. The range of observed ages at CRC diagnosis varied between 25 and 75, and 58 events were observed. From the 191 studied individuals, 116 were homozygotes of the non-risk allele, 65 were heterozygotes and 10 were homozygotes of the risk allele. Because of the limited size of the last category, we evaluated the effect of the indicator of being a carrier of the rs1321311 allele.

We considered four different models: unweighted Cox regression, the state-of-the-art weighted cohort method, our new method based on the new and more general weighting scheme, and a shared gamma frailty model as a sensitivity analysis to measure the potential impact of unobserved family-specific heterogeneity. The two studied weighted methods require the knowledge of incidence rates for CRC in carriers of pathogenic variants in *PMS2*. These were obtained by multiplying the population-based incidence rates of

CRC in the Netherlands in 2011 [Netherlands Cancer Registry, 2021] by the previously published [Ten Broeke et al., 2015] age-dependent hazard ratios of CRC for *PMS2* carriers. The choice of the year 2011 as the reference is justified because it is the middle point of the data collection period (2007-2016). The specific age-specific intervals and incidence rates used in this application can be found in Supplemental Table 5.6.

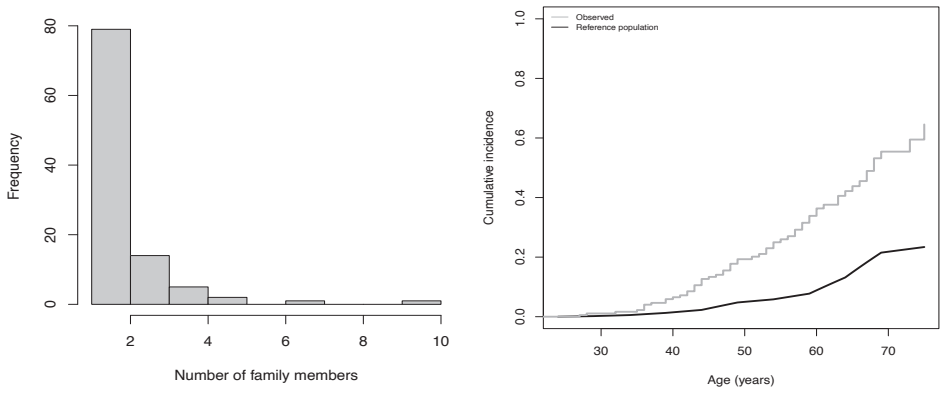

Figure 5.1: Application 1: Study of the association between SNP rs1321311 and CRC cancer in male carriers of a pathogenic variant in the gene *PMS2*. Left panel: Size of the families included in the sample. Right panel: Cumulative incidence of colorectal cancer at different ages. The grey line shows the observed risk in the sample. The black line reflects the expected cumulative colorectal cancer risk for the population of *PMS2* mutation carriers based on previous literature [Ten Broeke et al., 2015]. Specifically, age-specific CRC incidence rates of *PMS2* mutation carriers are obtained multiplying the point estimates of the age-dependent hazard ratios as reported in Table 2 in Ten Broeke *et al.* [2015] by the underlying population-based incidence rates of CRC for males in the Netherlands in 2011 according to the Netherlands Cancer Registry (NCR).

From the results reported in the bottom line of Table 5.3, it is observed that the new generalized weighted cohort method provides slightly larger estimated effects than the well-known (unweighted) Cox regression. In agreement to the result obtained with the unweighted method, the estimated association between the risk allele rs1321311 and CRC was statistically significant at the usual 5% level when using the new method. Importantly, the traditional weighted cohort approach could not be used because negative weights were obtained. Specifically, the oversampling of cases was not strong enough in the age group 65-70 years old and restriction (5.10) discussed in Section 5.2.1 was not met leading to negative weights for unaffected individuals in this age group. The shared frailty model provides the lower estimated covariate effect among the evaluated methods. This is

probably due to the limited sizes of the family clusters and a small underlying unobserved heterogeneity. The estimated frailty variance was 0.15 with a broad confidence interval (0-1), indicating difficulties of the model to give reliable estimates of the level of unobserved heterogeneity. A likely major driving cause for this difficulty is the limited cluster size of this application since most of the families contribute a single individual to the analysis. As a consequence, the shared frailty approach is not recommended in this application and one would rather choose the new generalised weighted cohort approach.

Table 5.3: Application to CRC in male carriers of PMS2. Estimated regression coefficients ($\widehat{\beta}$) and corresponding 95% Confidence Intervals for the effect of the SNP rs1321311 for different Cox models. Case weights are calculated based on incidence rates of CRC for *PMS2* mutation carriers defined as the point estimates of the age-dependent hazard ratios reported in Ten Broeke *et al.* [2018] multiplied by the population-based rates of CRC in Netherlands in 2011.

Model	$\widehat{\beta}$ **(95% CI)**
Unweighted	0.723 (0.182; 1.265)
Frailty	0.671 (0.149; 1.192)
Weighted cohort	- (*negative weights*)
Generalized weighted cohort	0.771 (0.234; 1.308)

5.4.2 Application to breast cancer

In this application, the association between a PRS score and breast cancer was analyzed using a sample of 579 clinically ascertained women belonging to 101 families. On average, six women were included per family (mean family size = 5.73 and standard deviation = 4.66, Figure 2 right panel). The inclusion criterion was two-fold. Per family, one of the women should be tested negative for BRCA1 or BRCA2 pathogenic variants. This was a special feature of this sample and means that family aggregation and early-onset of cancer are not explained by pathogenic variants in these high-risk genes. Furthermore, breast cancer had to occur in at least three female family members or in two females if at least one had bilateral breast cancer before the age of 60. The families were selected between 1990 and 2012 by Clinical Genetic Services in four Dutch cities (Groningen, Leiden, Nijmegen and Rotterdam) and one Hungarian city (Budapest). Given the scarcity of events after 80 years of age (only one observed event at 90), we censored observations at age 80. The PRS was based on 161 SNPs weighted by previously published log-odds ratios (mostly based on population-based case-control studies). Detailed description of the calculation of PRS can

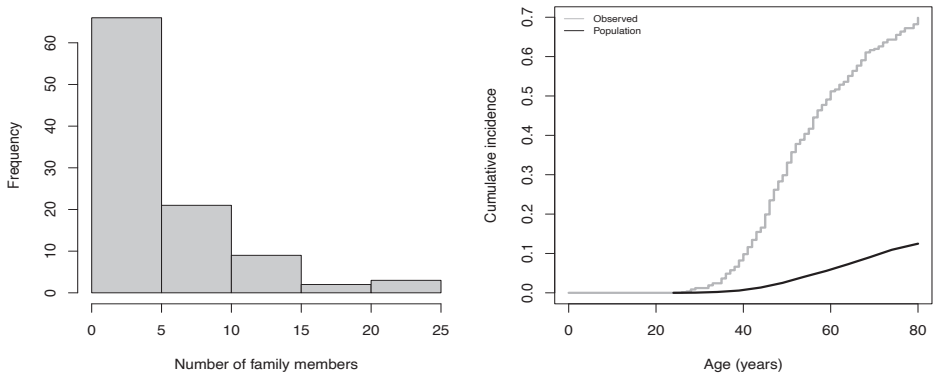

Figure 5.2: Application 2: Study of the association between a polygenic risk score and female breast cancer. Left panel: Size of the families included in the sample. Right panel: Cumulative incidence of breast cancer at different ages. The gray line shows the observed risk in the sample. The black line shows the population-based (the Netherlands, 2001 [Netherlands Cancer Registry, 2021]) cumulative incidence used as reference in the weighted analyses.

be found elsewhere [Lakeman et al., 2019]. As before, to establish the association between the marker of interest, the PRS, and breast cancer, we considered four different models: the traditional unweighted Cox regression, the state of the art weighted cohort method to deal with outcome-dependent sampling, our new weighted method and a shared gamma frailty model. Population-based incidence rates of the Netherlands in 2001 [Netherlands Cancer Registry, 2021] (mid point of the sample selection period) were used as external input to construct the weights. The specific age-specific intervals and incidence rates used in this application can be found in Supplemental Table 5.6.

From the results reported in Table 5.4, we observe that the new method provides a slightly smaller effect than the previously proposed weighted cohort approach and that both provided smaller effects than the unweighted Cox model. None of these three approaches reached statistical significance at the 5% level. In order to estimate the level of heterogeneity due to unmeasured within-family similarity, a shared frailty model was also fitted. The estimated frailty variance was 0.41, indicating that unobserved heterogeneity is not negligible in this application. This, together with the large size of the included families, is probably the reason why the shared frailty model seems to outperform the other methods. The estimated conditional hazard ratio using a shared frailty model is larger

than the ones obtained using unweighted and weighted versions of the Cox model even if statistical significance at the 5% level is also not reached with this approach. According to our simulation results, we infer that the association between PRS and breast cancer is likely obscured by ignoring the strong unobserved heterogeneity and that the frailty approach is preferred in this application.

Table 5.4: Application to female breast cancer in non-BRCA1/2 families. Estimated regression coefficients ($\widehat{\beta}$) and corresponding 95% confidence intervals for the effect of polygenic risk score (PRS) for different Cox models.

Model	$\widehat{\beta}$ **(95% CI)**
Unweighted	0.110 (-0.096; 0.317)
Frailty	0.173 (-0.045; 0.390)
Weighted cohort	0.079 (-0.226; 0.385)
Generalized weighted cohort	0.062 (-0.261; 0.384)

5.5 Discussion

In this paper, we have revisited the analysis of outcome-dependently sampled survival data with weighted Cox regression using external data to construct inverse probability of selection weights. Our research is motivated by the interest in the effect of potential modifying factors on cancer risk using clinically ascertained data. Typically, those data sets are collected through ongoing genetic testing programs, where selection criteria lead to an over-representation of young cases and hence, the resulting samples are not representative of the target population of interest. We proposed a new weighting scheme that restores the expected ratio of events and non-events at each follow-up time using population-based hazard information. Our simulation study has shown that the new method can be applied to a broader set of realistic scenarios. Our real data applications support the same conclusion indicating the broader applicability of the new weighting scheme and it should be the preferred option to analyze data obtained under family-based outcome-dependent sampling when unobserved heterogeneity is negligible or mild.

A strength of the new weighting scheme is that it relies on fewer assumptions to provide valid, non-negative weights. The traditional weighted cohort [Antoniou et al., 2005] approach requires that a number of conditions are fulfilled, which hamper its applicability. Specifically, the original method is problematic if oversampling of cases is not observed in

all age groups. In practice, although overall oversampling of events is expected, it does not necessarily hold for all age groups. Our new method overcomes this restriction and can be applied to a wider set of oversampling schemes, hence it can be regarded as a generalization of the traditional weighted cohort approach. This together with user-friendly implementation makes it an attractive analysis tool for applied researchers in the field.

Likewise the previously proposed weighted cohort method, our approach relies on a number of assumptions. First, a crucial assumption is the existence of a well-established external source of population-based incidence rates. Second, the sampling probabilities of observed individuals depend on the age at onset but they are assumed to be conditionally independent of the risk modifier under investigation. These two assumptions have been previously discussed in the context of the weighted cohort method [Antoniou et al., 2005; Barnes et al., 2012]. Furthermore, the relationship between the hazard and the risk modifier under investigation should approximately follow a proportional hazards specification. We have examined the performance of both the traditional and the new generalised weighted cohort approaches under model misspecification, specifically, under non-collapsability due to the presence of residual familial aggregation. In this case, we have also observed that the use of weighted approaches seems advisable compared to the naive unweighted approach. Additionally, if the number of available individuals per family is limited, which is the most common situation in practice, our new method might be the preferred option, outperforming a shared frailty model and the traditional weighted cohort approach. However, we would like to caution about the interpretation of the estimated effect and point out the systematic downward bias of the regression coefficient in this setting, proposing the systematic inclusion of the results of a shared frailty model as a sensitivity analysis.

The extension of weighting approaches to deal with outcome-dependent sampling to the context of frailty models would be interesting but challenging. Since the estimated incidence rate in the sample depends on the correct estimation of the frailty variance, it would be necessary to know the value of the frailty variance to derive correct weights. However, the frailty variance is latent and hence we anticipate an identifiability problem in such an approach. More sophisticated modeling, using a frailty model with explicit correction for ascertainment is possible but not straightforward and it is left as future research. It is noteworthy that such a complex approach will presumably require large clusters and sample sizes and hence our simpler approach based on borrowing information

from a trustworthy external source will still be preferred in a number of relevant practical situations, such as our application to PMS2 carriers.

In conclusion, for performing regression analysis using survival data obtained under family-based outcome dependently sampling, specialized techniques are required to avoid bias and provide valid inference. We have proposed an accurate and conceptually simple method which generalizes and outperforms existing methods based on weighted Cox regression.

Acknowledgments

The departments of Clinical Genetics and Human Genetics (LUMC, Leiden) are gratefully acknowledged for providing the breast cancer data set.

5.6 Supplementary material

5.6.1 Unbiasedness of the inverse probability of selection weighted Cox approach

Denote by D the selection event, and let $W(t)$ be the probability of being selected given that that the observed event time is t, assume that $W(t) > 0$ on the support of T.

In the absence of covariates, the density of the observed data, obtained under biased (in particular within an outcome-dependent sampling where $D \perp Z|T$) is given by the following weighted density:

$$f_{T|D}(t) = \frac{P(D=1|T=t)f_T(t)}{\int_0^\infty P(D=1|T=s)f_T(s)ds}.$$

Accordingly, the joint density of the sampled information including covariate Z is given by:

$$f_{T,Z|D}(t,z) = \frac{W(t)h(t|z;\beta)exp(-\int_0^t h(y|z;\beta)dy)f_Z(z)}{E(W(t))}.$$

The density of the event times conditional on the covariate is therefore:

$$f_{T|Z,D}(t,z) = \frac{W(t)h(t|z;\beta)exp(-\int_0^t h(y|z;\beta)dy)}{E(W(t)|z)}.$$

Note that $E(W(t)|z) = P(D = 1|t, z)$ depends on β and thus this biased sampling must be accounted for in the estimation of β. However, since the weight $W(t)$ is a function of t alone, based on the key assumption $D \perp Z|T$, it follows that $f_{Z|T,D} = f_{Z|T}$ and the following standard probabilistic result from the Cox model can be used:

$$E(Z|T = t, D) = E(Z|T = t) = \frac{E\left[Ze^{\beta Z}\overline{F}_{T|Z}(t|Z)\right]}{E\left[e^{\beta Z}\overline{F}_{T|Z}(t|Z)\right]},$$

where $\overline{F} = 1 - F$ denotes the survival function.

Let $f_{Z|D} = E(W(t)|z)f_Z(z)/E(W(T))$ denote the marginal weighted density of the covariate, then we can adapt the former general result to our setting and rewrite it as follows:

$$E(Z|T = t, D) = \frac{E\left[Ze^{\beta Z}\overline{F}_{T|Z}(t|Z)/E(W(T)|Z)|D\right]}{E\left[e^{\beta Z}\overline{F}_{T|Z}(t|Z)/E(W(T)|Z)|D\right]}.$$

The former expression still involved functionals of Z and T unconditionally on D. To rewrite the expectation as a function of observed variables we use the expression on the density of the event times conditional on the covariate $f_{T|Z,D}(t, z)$ and the fact that $W(t) > 0$ on the support of T, which implies $\overline{F}_{T|Z}(t|z)/E(W(T)|Z = z) = E\left[W(T)^{-1}I(T \geq t)|Z = z, D)\right]$. As a result:

$$E(Z|T = t, D) = \frac{E\left[Ze^{\beta Z}W(T)^{-1}I(T \geq t)|D\right]}{E\left[e^{\beta Z}W(T)^{-1}I(T \geq t)|D\right]},$$

which implies the unbiasedness of the weighted estimating equation

$$U(\beta) = \sum_{i=1}^{n}\{Z_i - \frac{\sum_{j-1}^{n} Z_j e^{\beta Z_j}(W(T_j)^{-1}I(T_j \geq T_i)}{\sum_{j-1}^{n} e^{\beta Z_j}(W(T_j)^{-1}I(T_j \geq T_i)}\}.$$

5.6.2 Literature review: family size and the use of the weighted cohort approach

Among the 81 citations of Antoniou *et al.*'s 2005 paper [Antoniou et al., 2005], we found 51 papers that used a weighted cohort approach to obtain unbiased Hazard Ratios in a Cox model (list available upon request). The majority (62.7%, n = 32) were studies into breast and ovarian cancer risks. In 48 of the 51 papers applying the weighted cohort approach, the study sample certainly includes multiple members per family, but the exact number of families was only mentioned in 19 papers, shown in Table 5.5. For each study, we calculated the average number of family members included. Note that this may fluctuate: one study [Andrieu et al., 2006] described the exact family composition of the sample, see Table 5.5 footnote 6. The median of the paper-specific, average family cluster sizes was 2.5. Generally, this data was collected at family cancer- or genetics clinics, where relatives of the index case (proband) were invited to be tested. Sometimes this was combined with 'population-based' recruitment [Chau et al., 2016; Ait Ouakrim et al., 2015].

Table 5.5: Family information (when reported) in papers applying weighted cohort approach. This list is a subset of all (81) PubMed citations of the paper of Antoniou *et al.* [2005] on 06/02/2021, with inclusion criteria 1) applying weighted Cox, 2) mentioning the number of families and sample size.

Authors	**Year**	**Sample**	**Average number of relatives per family**	**(Cancer) research area**
Borde *et al.*	2020	578 families; 760 carriers	1.6	Breast
Dashti *et al.*	2018	774 families; 2042 carriers	2.6	Ovarian
Ten Broeke *et al.*	2018	152 families; 521 samples	3.4	Colorectal
Kamiza *et al.*	2016	62 families; 260 carriers	4.2	Endometrial
Dashti *et al.*	2017	761 families; 1925 carriers	2.5	Colorectal
Chau *et al.*	2016	15049 families; 42489 participants[(1)]	5.3	Colorectal
Win *et al.*	2015 (b)	330 families; 1098 carriers	3.3	Colorectal
Win *et al.*	2015 (a)	593 families; 854 individuals	1.4	Colorectal
Dashti *et al.*	2015	548 families; 1128 women	2.1	Breast
Ait Ouakrim et al.	2015	748 families; 1858 carriers[(2)]	2.5	Breast
Pooley et al.	2014	3134 families; 4822 included[(3]	1.5	Breast
Killick et al.	2014	115 families; 158 included[(4)]	1.4	Breast
Win et al.	2013	315 families; 927 carriers	2.9	Breast
Pijpe et al.	2012	930 families; 1122 carriers[(4)]	1.2	Breast
Win et al.	2011 (a)	498 families provided 1324 carriers; 287 families provided 1219 non-carriers	2.7 (carriers); 4.2 (non-carriers)	Colorectal
Win et al.	2011 (b)	286 families provided 601 carriers; 182 families provided 533 non-carriers	2.1 (carriers); 2.9 (non-carriers)	Endometrial
Amos et al.	2010	93 families; 489 included (of which 45 married-in)	5.3	Ovarian
Antoniou et al.	2006	392 families; 810 carriers	2.1	Breast
Andrieu et al.	2006	1074 families; 1601 women	1.5[6]	Breast

(1) Includes some population-based recruitment for which average number of recruited relatives per family was 2.6, vs. 5.3 for clinic-based families.
(2) 25% population-based.
(3) Sampled non-carrying relatives only.
(4) Relatives functioned as controls, i.e. non-carriers.
(5) This concerns only one of the data sets in the paper.
(6) The relative occurrence of relatives per family was 71.1% for size 1, 17.9% for size 2, 6.2% for size 3, 2.8% for size 4 and 2.0% for size 5 up until 11.

5.6.3 Population incidence rates used in real data applications

Table 5.6: Population-based age-specific incidence rates (in cases per 100.000) used for weight construction. For breast cancer, we used the registered data of The Netherlands in 2001 for women [Netherlands Cancer Registry, 2021]. For colorectal cancer, age-specific incidence rates were obtained by multiplying the population-based incidence rates of CRC in the Netherlands in 2011 [Netherlands Cancer Registry, 2021] by the previously published [Ten Broeke et al., 2015] age-dependent hazard ratios of CRC for *PMS2* carriers. The choice of the year 2011 as the reference is justified because it is the middle point of the data collection period (2007-2016).

Age group	Colorectal (men)	Breast (women)
25-30	40.56	9.09
30-34	69.39	33.61
35-39	146.60	72.41
40-44	204.38	156.17
45-49	518.28	241.48
50-54	221.10	323.56
55-59	411.68	310.83
60-64	1214.11	363.63
65-69	2016.21	388.41
70-74	405.12	415.42
75-79	-	293.22