Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

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

Vera H. Arntzen[1] | Marta Fiocco[1,2] | Inge M.M. Lakeman[3,4] | Maartje Nielsen[3] | Mar Rodríguez-Girondo*[2]

[1]Mathematical Institute, Leiden University, Leiden, The Netherlands
[2]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
[3]Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
[4]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
*Mar Rodríguez-Girondo, Email: m.rodriguez_girondo@lumc.nl

Present Address
Einthovenweg 20, Postbus 9600, 2300 RC Leiden

Abstract

Motivated by the study of genetic effect modifiers of cancer, we examined weighting approaches to correct for ascertainment bias in survival analysis. Family-based 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:
survival analysis, outcome-dependent sampling, weighting, Cox regression, genetic epidemiology

1 | INTRODUCTION

Family-based 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

[0]**Abbreviations:** GWAS, Genome Wide Association Studies; BRCA, breast (and ovarian) cancer associated (genes); reBias, relative Bias; MSE, mean squared error; CRC, colorectal cancer; DAG, directed acyclic graph; SNP, Single Nucleotide Polymorphism (genetic variant)

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: families with members that are diagnosed at a young age, and multiple cases are more likely to be included in the sample.

In the context of survival analysis, family-based outcome-dependent sampling results in an overrepresentation 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 (1–3) and the weighted cohort method (4) 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 (4). 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 (4, 5) 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, for example with strong oversampling of cases only.

This study has two objectives. Firstly, we propose an alternative, more general, inverse probability of selection weighting scheme using population-based age-specific incidence rates of the event of interest. Accordingly, a generalized weighted cohort method that can deal with arbitrary levels of outcome-dependent sampling is proposed and compared with the existing approach. A second goal of this study is to investigate the role of within-family correlation due to unobserved shared factors in the performance of weighted cohort approaches. Despite the presence of multiple members of the same family being relatively common in studies relying on the original weighted cohort method (see Table S1 for details), the impact of unobserved heterogeneity in this context has not been previously evaluated. It is an aspect that deserves attention.

The rest of the paper is organised as follows. In Section 2, the commonly used weighted cohort Cox approach is revisited and its assumptions are discussed. A new alternative weighting scheme is proposed in Section 3. In Section 4, both weighting schemes are compared by means of an intensive Simulation study. In Section 5, 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 6.

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 and 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 $i = 1, \dots, n_j$ index all included individuals in the sample belonging to family j, and assume that M families are observed, with varying observed size $n_1, \dots, n_M$, so that $n = \sum_{j=1}^{M} n_j$ individuals are included in the sample. The observed time to event for individual 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 an outcome-dependent sampling scheme. The specific selection criteria vary over studies, depending on the severity and prevalence of the disease, but are usually of the form $A_j = \left\{ \sum_{i=1}^{n_j} \delta_{ij} \right\} \geq n_A$. That is, if and only if there are at least n_A family members in family j who experienced the event of interest, family j is included in the study. Often, not only δ_{ij}, $i = 1, \dots, n_j$ but also t_{ij}, $i = 1, \dots, n_j$ is used to define the ascertainment event A_j. For example, including families with at least three members with cancer diagnosis below the age of 60 would correspond to $A_j = \left\{ \sum_{i=1}^{n_j} \delta_{ij} \times I(t_{ij} \leq 60) \right\} \geq 3$. In general, if a family j is included in the study, i.e. if $A_j = 1$, all its members are invited to participate, also those who did not experience the event yet (right-censored observations). Despite the diversity of possible family configurations, there is typically an overrepresentation of young cases of the family. Often, only the carriers of a certain genetic pathogenic variant are of interest and the target population is the population of carriers of such a variant. In those situations, the former definitions still apply, but the family size is restricted to the family members who are mutation carriers.

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 overrepresentation of events and short event times in the sample due to outcome-dependent sampling alters 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.

2.1 | The weighted cohort approach revisited

When T represents the age at cancer diagnosis, or another common disease, registry data about the distribution of T in the target population is often available. Let us assume, as it is common in practice, that the available information in the external source is aggregated into K disjoint age intervals, defined as $I_1 = [a_0, a_1), I_2 = [a_1, a_2), \dots I_K = [a_{K-1}, a_K)$. In the case of cancer studies, the typical available external information is the population cancer incidence rate μ_k for each age interval I_k, $k = 1, \dots K$. The seminal paper of *Antoniou et al. ((**?**))* proposed a weighted Cox regression model with sampling weights derived such, that the incidence rates in each interval I_k in the resulting pseudo-population after weighting agrees with the incidence rates μ_k in the target population, available and used as external data to calculate the weights.

Specifically, let r_k denote the number of individuals experiencing the event within the age interval, $I_k = [a_{k-1}, a_k)$, $k = 1, \dots K$. 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:

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

In the right part of expression (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 (1) alone does not guarantee unique weights w_k and v_k, the following constraint is incorporated to guarantee unique weights:

$$\frac{w_k r_k + v_k s_k}{r_k + s_k} = 1, \tag{2}$$

Combining equations (1) and (2) provides unique expressions for v_k and w_k:

$$w_k = \frac{\mu_k (q_k (r_k + s_k) + (a_k - a_{k-1}) s_k \sum_{l>k} (r_l + s_l))}{r_k s_k + \mu_k (q_k r_k - p_k s_k)}, \tag{3}$$

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

$$v_k = \frac{1}{s_k} (r_k + s_k - w_k r_k). \tag{4}$$

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

$$U_a(\beta) = \sum_{ij:\delta_{ij}=1} a_{ij} z_{ij} - \sum_{ij:\delta_{ij}=1} \frac{\sum_{l \epsilon R(t_{ij})} a_l z_l \exp[\beta z_l]}{\sum_{l \epsilon R(t_{ij})} a_l \exp[\beta z_l]}, \tag{5}$$

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

$$a_{ij} = \begin{cases} w_k, & \text{if } \delta_{ij} = 1 \text{ and } t_{ij} \in [a_{k-1}, a_k) \\ v_k, & \text{if } \delta_{ij} = 0 \text{ and } t_{ij} \in [a_{k-1}, a_k). \end{cases} \tag{6}$$

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

$$r_k > 0 \tag{7}$$

$$s_k > 0 \tag{8}$$

$$r_k > \mu_k p_k \frac{s_k}{s_k + \mu_k q_k} \tag{9}$$

$$w_k < 1 + \frac{s_k}{r_k} \tag{10}$$

Conditions (7) and (9) are required to get valid w_k weights for the cases, while conditions (8) and (10) are required to get valid v_k weights for those that are censored. Condition (7) implies the observation of events in all the considered intervals and condition (8) implies the presence of right-censored observations in all considered intervals. Conditions (9) and (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. ((**?**))*, 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 (9) and especially (10) might not be fulfilled in relevant practical scenarios.

Under oversampling of events at interval I_k, condition (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 (9) is fulfilled even when no oversampling of events is observed in interval I_k. However, condition (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 (10).

When any of the conditions (7)-(10) are not satisfied, the weighted cohort method can still be applied by merging intervals, however the method becomes then less precise and dependent on sample specific characteristics which may hamper comparability among studies using this method.

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

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 underrepresentation 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 distribution of T. 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: a) the number of individuals that experience the event within the interval I_k and b) 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:

$$N_k = N_k S_k + N_k(1 - S_k) \tag{11}$$

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 (11) follows that the ratio between events and non-events in interval I_k in the reference population is given by $\frac{1-S_k}{S_k}$.

The same decomposition of the risk set at the beginning of interval I_k can be made for the observed sample S_O, potentially subject to outcome-dependent sampling:

$$N_k = N_k S_k^o + N_k(1 - S_k^o) \tag{12}$$

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 (11) and (12). Specifically, weights w_k correct the oversampling (or undersampling) of cases, such that the ratio between events and non-events in interval I_k in the resulting pseudo-population after weighting is the same as in the reference population:

$$w_k = \frac{(1 - S_k)}{S_k} \frac{S_k^o}{(1 - S_k^o)}, \tag{13}$$

Equation (13) indicates that the population ratio between events and non-events in the interval I_k, given by $\frac{1-S_k}{S_k}$ is multiplied by the inverse quantity based on the observed data, $\frac{1-S_k^o}{S_k^o}$. After weighing, the composition of the risk set at interval I_k resembles the composition of the risk set at 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).

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 subpopopulations are different. The new method 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.

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 under the assumption of fully observed heterogeneity, i.e., differences among individuals in terms of hazards can be fully accounted for by the observed covariates. In a second simulation setting, we consider the presence of unmeasured factors shared within families.

3.1 | Simulation setup I

Simulated data was generated using the following model:

$$\lambda_{ij}(t) = \lambda_0 \exp(\beta z_{ij}), \tag{14}$$

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 its 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 $M = 1000$ families ($M = 125, 250, 500$). 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 simulating a large data set ($N = 200\,000$) without ascertainment and it was used to approximate the population hazards needed to calculate the weights. Bias, mean square error and coverage proportions of the 95% confidence intervals are reported in Table 1. As proposed by *Antoniou et al. (2005)*, robust estimates of the standard errors were obtained using a sandwich estimator in both considered weighted methods. Moreover, the proportion of invalid weights in the M Monte Carlo trials is reported for the two weighted methods.

3.2 | Simulation setup II

In the previous simulation setting, we have assumed, as the proposed models in Section 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 Tabl S1), 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 (14) by adding a latent (frailty) term, U, shared by all members of the same family:

$$\lambda_{ij}(t) = u_j \lambda_0 \exp(\beta z_{ij}), \tag{15}$$

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' (θ=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 (6), the presence of U introduces bias in the hazard ratio estimation, even when U is independent of the covariate of interest Z. However, we also expect that such a bias diminishes under the presence of outcome-dependent sampling.

The rest of the simulation settings were as in the previous Simulation setting I, except M, 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 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.

Table 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. M: number of families.

			Unweighted			**Weighted cohort**				**Generalized weighted cohort**			
β	Scenario	M	reBias	MSE	Coverage	reBias	MSE	Coverage	Invalid weights	reBias	MSE	Coverage	Invalid weights
		250	< 0.001	0.003	0.947	< 0.001	0.003	0.939	0.002	< 0.001	0.003	0.944	0.000
	A1	500	0.002	0.002	0.953	< 0.001	0.011	0.953	0.000	< 0.001	0.002	0.949	0.000
$\beta = 0$		750	< 0.001	0.001	0.942	-0.002	< 0.001	0.939	0.000	-0.002	0.001	0.947	0.000
		250	-0.003	0.004	0.942	< 0.001	0.006	0.938	0.005	< 0.001	0.005	0.940	0.000
	A2	500	0.001	0.002	0.095	< 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
		250	0.003	0.012	0.948	0.010	0.029	0.925	0.570	0.001	0.034	0.914	0.000
	A3	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
		250	-0.044	0.003	0.948	-0.091	0.004	0.912	0.000	-0.041	0.004	0.944	0.000
	A1	500	-0.046	0.002	0.932	-0.092	0.002	0.887	0.000	-0.042	0.002	0.931	0.000
$\beta = 0.3$		750	-0.036	0.001	0.943	-0.100	0.002	0.841	0.000	-0.049	0.001	0.919	0.000
		250	-0.200	0.008	0.818	-0.107	0.007	0.916	0.006	-0.129	0.007	0.893	0.000
	A2	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
		250	-0.221	0.017	0.885	-0.012	0.024	0.941	0.562	0.071	0.033	0.933	0.000
	A3	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
	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
$\beta = 1$		750	-0.067	0.034	0.463	-0.038	0.003	0.822	0.000	-0.004	0.002	0.943	0.000
		250	-0.173	0.035	0.294	-0.042	0.008	0.926	0.000	-0.059	0.010	0.891	0.000
	A2	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
		250	-0.270	0.083	0.228	0.024	0.030	0.907	0.025	0.070	0.043	0.886	0.000
	A3	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

3.3 | Simulation results

3.3.1 | 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 (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.

3.3.2 | Simulation II

Table 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. M: number of families. Data is generated according to a shared frailty model with frailty variance θ.

				Unweighted			**Weighted cohort**				**Generalized weighted cohort**				**Shared frailty**		
θ	β	Scenario	M	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
$\theta = 2$	$\beta = 0$	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 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, it is remarkable that weighted methods outperform the fit of a gamma shared frailty model which deals with shared unobserved heterogeneity but ignores outcome-dependent sampling.

For strong within-family correlation ($\theta = 2$) 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 indicate 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, in general, well in presence

of the combination of a strong covariate effect and strong outcome-dependent sampling, as expected. However, its applicability is restricted to certain scenarios and is not general enough. Inverse probability of selection weighted Cox models seem to still perform properly under 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 to 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.

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 https://github.com/vharntzen/wcox.

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 high-risk families without BRCA1 or 2 mutations.

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 (7), 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 (8). The distribution of the family size was very skewed, the mean family size 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. 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 (9) by the previously published ((8)) 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).

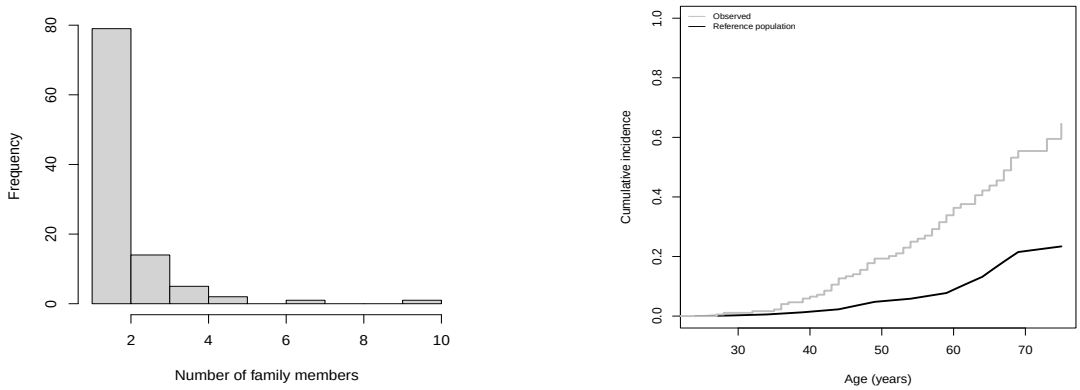

Figure 1 Application 1: Study of the association between SNP rs1321311 and CRC cancer in male carriers of a pathogenic variant in the gene *PMS2*. Left panel: Size of the families included in the sample. Right panel: Cumulative incidence of colorectal cancer at different ages. The gray line shows the 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 (8). 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 (8) 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 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 (11) discussed in Section 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 for the new generalised weighted cohort approach.

Table 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 (7) 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)

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 this PRS can be found elsewhere (10). 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 (9) (mid point of the sample selection period) were used as external input to construct the weights.

From the results reported in Table 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

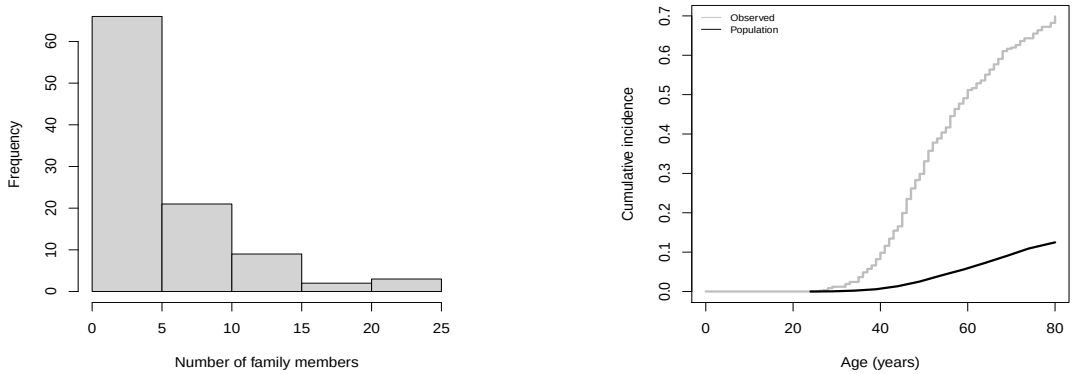

Figure 2 Application 2: Study of the association between a polygenic risk score and female breast cancer. Left panel: Size of the families included in the sample. Right panel: Cumulative incidence of breast cancer at different ages. The gray line shows the observed risk in the sample. The black line shows the population-based (the Netherlands, 2001 (9)) cumulative incidence used as reference in the weighted analyses.

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 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 | 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 overrepresentation 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 number of 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 (4) 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 (4, 5). An extra important assumption, but less often addressed, is fully observed heterogeneity. This means that if residual familiar aggregation is present, caution is required when interpreting the results. We have explicitly investigated this issue. Under the presence of unmeasured shared family factors affecting time to event, the estimation of a covariate effect is problematic, even if this covariate is independent of such unmeasured effects. The mere presence of such unobserved effects will affect the estimation of the hazard ratio of interest and careful interpretation of the results is required. Unobserved familiar heterogeneity induces a selection process based on the unobserved factors, so that the effect of the factor under investigation is attenuated with time, since the frailer individuals experience the event earlier in time. We have seen that also in this case, the use of weighted approaches seems advisable with respect to the naive unweighted approach. Also, if the number of available individuals per family is limited, i.e. cluster size is small, our new method might be the preferred option, outperforming a shared frailty model and the traditional weighted cohort approach. However, we would like to warn about the interpretation of the estimated effect and point out the systematic downwards bias of the regression coefficient in this setting. The extension of weighting approaches to the context of frailty models seems 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.

Financial disclosure

Last author acknowledges financial support by the Grant PID2020-118101GB-I00 (MCIN/ AEI /10.13039/501100011033)

Conflict of interest

The authors declare no potential conflict of interest.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

Figure S1. 500 hPa geopotential anomalies for GC2C calculated against the ERA Interim reanalysis. The period is 1989–2008.

How to cite this article: Arntzen V.H., M. Fiocco, I.M.M. Lakeman, M. Nielsen, and M. Rodríguez-Girondo (2016), A new inverse probability of selection weighted Cox model to deal with outcome-dependent sampling in survival analysis, *Q.J.R. Meteorol. Soc.*, *2017;00:1–6*.

APPENDIX

S1. LITERATURE REVIEW: USE OF WEIGHTED COHORT WITH FAMILY DATA

Among the 81 citations of Antoniou et al.'s 2005 paper (4), 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 **??**. For each study, we calculated the average number of family members included. Note that this may fluctuate: one study (11) described the exact family composition of the sample, see Table **??** 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 (12, 13).

Table S1: 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(4) on 06/02/202, 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.(14)	2020	578 families; 760 carriers	1.6	Breast
Dashti et al.(15)	2018	774 families; 2042 carriers	2.6	Ovarian
Ten Broeke et al.(7)	2018	152 families; 521 samples	3.4	Colorectal
Kamiza et al. (16)	2016	62 families; 260 carriers	4.2	Endometrial
Dashti et al. (17)	2017	761 families; 1925 carriers	2.5	Colorectal
Chae et al. (12)	2016	15049 families; 42489 participants(1)	5.3	Colorectal
Win et al. (18)	2015	330 families; 1098 carriers	3.3	Breast and ovarian
Win et al. (19)	2015	593 families; 854 individuals	1.4	General
Dashti et al. (20)	2015	548 families; 1128 women	2.1	Breast
Ait Ouakrim et al. (13)	2015	748 families; 1858 carriers(2)	2.5	Breast
Pooley et al. (21)	2014	3134 families; 4822 included(3	1.5	Breast
Killick et al. (22)	2014	115 families; 158 included(4)	1.4	Breast
Win et al. (23)	2013	315 families; 927 carriers	2.9	Breast
Pijpe et al. (24)	2012	930 families; 1122 carriers(4)	1.2	Breast
Win et al. (25)	2011	498 families provided 1324 carriers; 287 families provided 1219 non-carriers	2.7 (carriers); 4.2 (non-carriers)	Ovarian
Win et al. (25)	2011	286 families provided 601 carriers; 182 families provided 533 non-carriers	2.1 (carriers); 2.9 (non-carriers)	Breast
Amos et al. (26)	2010	93 families; 489 included (of which 45 married-in)	5.3	Ovarian
Antoniou et al. (27)	2006	392 families; 810 carriers	2.1	Breast
Andrieu et al. (11)	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.

S2. POPULATION INCIDENCE RATES USED IN REAL DATA APPLICATIONS

Table S2: 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 (9). For colorectal cancer, age-specific incidence rates were obtained by multiplying the population-based incidence rates of CRC in the Netherlands in 2011 (9) by the previously published ((8)) 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

References

1. Carayol J, Bonaïti-Pellié C. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset.. *Genetic epidemiology* 2004; 27: 109-117.

2. Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH. Case–control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics* 2006; 62: 36-48.

3. Barnes D, Barrowdale D, Beesley J, et al. Estimating single nucleotide polymorphism associations using pedigree data: applications to breast cancer. *British Journal of Cancer* 2013; 108: 2610-2622.

4. Antoniou AC, Goldgar DE, Andrieu N, et al. A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes.. *Genetic epidemiology* 2005; 29: 1-11.

5. Barnes DR, Lee A, Easton DF, Antoniou AC. Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. *Genetic epidemiology* 2012; 36(3): 274–291.

6. Robin Henderson PO. Effect of frailty on marginal regression estimates in survival analysis.. *Journal of the Royal Statistics Society. Series B* 1999; 61(2): 367-379.

7. Ten Broeke SW, Elsayed FA, Pagan L, et al. SNP association study in PMS2-associated Lynch syndrome.. *Familial cancer* 2018; 17: 507-515.

8. Ten Broeke SW, Brohet RM, Tops CM, Klift v. dHM, Velthuizen MEea. Lynch syndrome caused by germline PMS2 mutations: delineating the cancer risk.. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2015; 33: 319-325.

9. (NCR) NCR. Netherlands Comprehensive Cancer Organisation (IKNL), derived via www.iknl.nl/en/ncr/ncr-data-figures. 2021.

10. Lakeman I, Hilbers F, Rodríguez-Girondo M, et al. Addition of a 161-SNP polygenic risk score to family history-based risk prediction: impact on clinical management in non-BRCA1/2 breast cancer families. *Journal of Medical Genetics* 2019: jmedgenet–2019.

11. Andrieu N, Goldgar DE, Easton DF, et al. Pregnancies, breast-feeding, and breast cancer risk in the International BRCA1/2 Carrier Cohort Study (IBCCS).. *Journal of the National Cancer Institute* 2006; 98: 535-44.

12. Chau R, Dashti SG, Ait Ouakrim D, et al. Multivitamin, calcium and folic acid supplements and the risk of colorectal cancer in Lynch syndrome.. *International journal of epidemiology* 2016; 45: 940-53.

13. Ait Ouakrim D, Dashti SG, Chau R, et al. Aspirin, Ibuprofen, and the Risk of Colorectal Cancer in Lynch Syndrome.. *Journal of the National Cancer Institute* 2015; 107.

14. Borde J. WBNDWLKSGHJQAWLNHJPREANRAGAHJFUDVMAKMWGSWBSCVAGOLAECSMAASRKKHE. Performance of Breast Cancer Polygenic Risk Scores in 760 Female CHEK2 Germline Mutation Carriers.. *Journal of the National Cancer Institute* 2021; 113(7): 893-899.

15. Dashti SG, Win AK, Hardikar SS, et al. Physical activity and the risk of colorectal cancer in Lynch syndrome.. *International journal of cancer* 2018; 143: 2250-2260.

16. Kamiza AB, Hsieh LL, Tang R, et al. TP53 Polymorphisms and Colorectal Cancer Risk in Patients with Lynch Syndrome in Taiwan: A Retrospective Cohort Study.. *PloS one* 2016; 11: e0167354.

17. Dashti SG, Buchanan DD, Jayasekara H, et al. Alcohol Consumption and the Risk of Colorectal Cancer for Mismatch Repair Gene Mutation Carriers.. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2017; 26: 366-375.

18. Win AK, Clendenning M, Crawford W, et al. Genetic variants within the hTERT gene and the risk of colorectal cancer in Lynch syndrome.. *Genes & cancer* 2015; 6: 445-51.

19. Win AK, Reece JC, Buchanan DD, et al. Risk of colorectal cancer for people with a mutation in both a MUTYH and a DNA mismatch repair gene.. *Familial cancer* 2015; 14: 575-83.

20. Dashti SG, Chau R, Ouakrim DA, et al. Female Hormonal Factors and the Risk of Endometrial Cancer in Lynch Syndrome.. *JAMA* 2015; 314: 61-71.

21. Pooley KA, McGuffog L, Barrowdale D, et al. Lymphocyte telomere length is long in BRCA1 and BRCA2 mutation carriers regardless of cancer-affected status.. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2014; 23: 1018-24.

22. Killick E, Tymrakiewicz M, Cieza-Borrella C, et al. Telomere length shows no association with BRCA1 and BRCA2 mutation status.. *PloS one* 2014; 9: e86659.

23. Win AK, Hopper JL, Buchanan DD, et al. Are the common genetic variants associated with colorectal cancer risk for DNA mismatch repair gene mutation carriers?. *European journal of cancer (Oxford, England : 1990)* 2013; 49: 1578-87.

24. Pijpe A, Andrieu N, Easton DF, et al. Exposure to diagnostic radiation and risk of breast cancer among carriers of BRCA1/2 mutations: retrospective cohort study (GENE-RAD-RISK).. *BMJ (Clinical research ed.)* 2012; 345: e5660.

25. Win AK, Dowty JG, English DR, et al. Body mass index in early adulthood and colorectal cancer risk for carriers and non-carriers of germline mutations in DNA mismatch repair genes.. *British journal of cancer* 2011; 105: 162-9.

26. Amos CI, Pinney SM, Li Y, et al. A susceptibility locus on chromosome 6q greatly increases lung cancer risk among light and never smokers.. *Cancer research* 2010; 70: 2359-67.

27. Antoniou AC, Shenton A, Maher ER, et al. Parity and breast cancer risk among BRCA1 and BRCA2 mutation carriers.. *Breast cancer research : BCR* 2006; 8: R72.

28. Leffondré K, Wynant W, Cao Z, Abrahamowicz M, Heinze G, Siemiatycki J. A weighted Cox model for modelling time-dependent exposures in the analysis of case-control studies. *Statistics in medicine* 2010; 29(7-8): 839–850.

29. Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *The American Journal of Human Genetics* 2000; 66(3): 1119–1131.

30. Kleinbaum DG, Klein M. *Survival analysis*. 3. Springer . 2010.

31. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian journal of statistics* 2007; 34(1): 103–119.

32. Chen K. Generalized case–cohort sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001; 63(4): 791–809.

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

34. Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *JNCI: Journal of the National Cancer Institute* 2015; 107(5).

35. Metcalfe K, Finch A, Poll A, et al. Breast cancer risks in women with a family history of breast or ovarian cancer who have tested negative for a BRCA1 or BRCA2 mutation. *British journal of cancer* 2009; 100(2): 421.

36. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017; 551(7678): 92.

37. Muranen TA, Mavaddat N, Khan S, et al. Polygenic risk score is associated with increased disease risk in 52 Finnish breast cancer families. *Breast cancer research and treatment* 2016; 158(3): 463–469.

38. Oldenburg R, Kroeze-Jansema K, Houwing-Duistermaat J, et al. Genome-wide linkage scan in Dutch hereditary non-BRCA1/2 breast cancer families identifies 9q21-22 as a putative breast cancer susceptibility locus. *Genes, Chromosomes and Cancer* 2008; 47(11): 947–956.

39. Parkin DM, Pisani P, Ferlay J. Global cancer statistics. *CA: a cancer journal for clinicians* 1999; 49(1): 33–64.

40. Ziegler RG, Hoover RN, Pike MC, et al. Migration patterns and breast cancer risk in Asian-American women. *JNCI: Journal of the National Cancer Institute* 1993; 85(22): 1819–1827.

41. Wang C. Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error. *Statistics & probability letters* 1999; 45(4): 371–378.

42. Antoniou AC, Sinilnikova OM, Simard J, et al. RAD51 135G→ C modifies breast cancer risk among BRCA2 mutation carriers: results from a combined analysis of 19 studies. *The American Journal of Human Genetics* 2007; 81(6): 1186–1200.

43. Wienke A, others . Frailty models. *Max Planck Institute for Demographic Research, MPIDR Working Papers: No. WP-2003-32* 2003.

44. Clayton D, Cuzick J. The semiparametric Pareto model for regression analysis of survival data. *Bulletin de l'Institut International de Statistique* 1985; 51: 175–180.

45. Huster WJ, Brookmeyer R, Self SG. Modelling paired survival data with covariates. *Biometrics* 1989: 145–156.

46. Lin D. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine* 1994; 13(21): 2233–2247.

47. Cai J, Prentice RL. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* 1995; 82(1): 151–164.

48. Balan TA, Putter H. frailtyEM: An R Package for Estimating Semiparametric Shared Frailty Models. 2017.

49. Therneau TM, Crowson CS, Atkinson EJ. Adjusted survival curves. 2015.

50. Klein JP, Van Houwelingen HC, Ibrahim JG, Scheike TH. *Handbook of survival analysis*. CRC Press . 2016.

51. Wang W, Wells MT. Estimation of Kendall's tau under censoring. *Statistica Sinica* 2000: 1199–1215.

52. Bouter L, Van Dongen M, Zielhuis G. Epidemiologische Onderzoek: Opzet en Interpretatie (Vol. 6th). *Houten: Bohn Stafleu van Loghum* 2010.

53. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 1958; 53(282): 457–481.

54. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 1972; 34(2): 187–202.

55. Carr S. Ascertainment Bias in Genetic Pedigree Analysis. 2016.

56. Rodríguez-Girondo M, Deelen J, Slagboom EP, Houwing-Duistermaat JJ. Survival analysis with delayed entry in selected families with application to human longevity. *Statistical methods in medical research* 2018; 27(3): 933–954.

57. Xue X, Agalliu I, Kim M, et al. New methods for estimating follow-up rates in cohort studies. *BMC Medical Research Methodology* 2017; 17. doi: 10.1186/s12874-017-0436-z

58. Lee A, Cunningham A, Kuchenbaecker K, Mavaddat N, Easton D, Antoniou A. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *British journal of cancer* 2014; 110(2): 535.

59. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common Genetic Variation and Breast Cancer Risk—Past, Present, and Future. 2018.

60. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics* 2013; 45(4): 353.

61. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American statistical Association* 1989; 84(408): 1074–1078.

62. RStudio Team . *RStudio: Integrated Development Environment for R, Version 1.0.153*. RStudio, Inc.; Boston, MA: 2017.

63. Dashti SG, Li WY, Buchanan DD, et al. Type 2 diabetes mellitus, blood cholesterol, triglyceride and colorectal cancer risk in Lynch syndrome.. *British journal of cancer* 2019; 121: 869-876.

64. Schrijver LH, Olsson H, Phillips KA, et al. Oral Contraceptive Use and Breast Cancer Risk: Retrospective and Prospective Analyses From a BRCA1 and BRCA2 Mutation Carrier Cohort Study.. *JNCI cancer spectrum* 2018; 2: pky023.

65. Qian F, Rookus MA, Leslie G, et al. Mendelian randomisation study of height and body mass index as modifiers of ovarian cancer risk in 22,588 BRCA1 and BRCA2 mutation carriers.. *British journal of cancer* 2019; 121: 180-192.

66. Kehm RD, Hopper JL, John EM, et al. Regular use of aspirin and other non-steroidal anti-inflammatory drugs and breast cancer risk for women at familial or genetic risk: a cohort study.. *Breast cancer research : BCR* 2019; 21: 52.

67. Terry MB, Liao Y, Kast K, et al. The Influence of Number and Timing of Pregnancies on Breast Cancer Risk for Women With BRCA1 or BRCA2 Mutations.. *JNCI cancer spectrum* 2018; 2: pky078.

68. Qian F, Wang S, Mitchell J, et al. Height and Body Mass Index as Modifiers of Breast Cancer Risk in BRCA1/2 Mutation Carriers: A Mendelian Randomization Study.. *Journal of the National Cancer Institute* 2019; 111: 350-364.

69. Derks-Smeets IAP, Schrijver LH, Die-Smulders dCEM, et al. Ovarian stimulation for IVF and risk of primary breast cancer in BRCA1/2 mutation carriers.. *British journal of cancer* 2018; 119: 357-363.

70. Kamiza AB, Hsieh LL, Tang R, et al. Polymorphisms of DNA repair genes are associated with colorectal cancer in patients with Lynch syndrome.. *Molecular genetics & genomic medicine* 2018; 6: 533-40.

71. Park B, Hopper JL, Win AK, et al. Reproductive factors as risk modifiers of breast cancer in BRCA mutation carriers and high-risk non-carriers.. *Oncotarget* 2017; 8: 102110-102118.

72. Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer.. *Nature genetics* 2017; 49: 1767-1778.

73. Kuchenbaecker KB, McGuffog L, Barrowdale D, et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers.. *Journal of the National Cancer Institute* 2017; 109.

74. Lawrenson K, Kar S, McCue K, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus.. *Nature communications* 2016; 7: 12675.

75. Vigorito E, Kuchenbaecker KB, Beesley J, et al. Fine-Scale Mapping at 9p22.2 Identifies Candidate Causal Variants That Modify Ovarian Cancer Risk in BRCA1 and BRCA2 Mutation Carriers.. *PloS one* 2016; 11: e0158801.

76. Dunning AM, Michailidou K, Kuchenbaecker KB, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170.. *Nature genetics* 2016; 48: 374-86.

77. Hollestelle A, Baan v. dFH, Berchuck A, et al. No clinical utility of KRAS variant rs61764370 for ovarian or breast cancer.. *Gynecologic oncology* 2016; 141: 386-401.

78. Blein S, Bardel C, Danjean V, et al. An original phylogenetic approach identified mitochondrial haplogroup T1a1 as inversely associated with breast cancer risk in BRCA2 mutation carriers.. *Breast cancer research : BCR* 2015; 17: 61.

79. Osorio A, Milne RL, Kuchenbaecker K, et al. DNA glycosylases involved in base excision repair may be associated with cancer risk in BRCA1 and BRCA2 mutation carriers.. *PLoS genetics* 2014; 10: e1004256.

80. Couch FJ, Wang X, McGuffog L, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk.. *PLoS genetics* 2013; 9: e1003212.

81. Gaudet MM, Kuchenbaecker KB, Vijai J, et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk.. *PLoS genetics* 2013; 9: e1003173.

82. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer.. *Nature genetics* 2013; 45: 371-84, 384e1-2.

83. Stevens KN, Wang X, Fredericksen Z, et al. Evaluation of chromosome 6p22 as a breast cancer risk modifier locus in a follow-up study of BRCA2 mutation carriers.. *Breast cancer research and treatment* 2012; 136: 295-302.

84. Kirchhoff T, Gaudet MM, Antoniou AC, et al. Breast cancer risk and 6q22.33: combined results from Breast Cancer Association Consortium and Consortium of Investigators on Modifiers of BRCA1/2.. *PloS one* 2012; 7: e35706.

85. Lecarpentier J, Noguès C, Mouret-Fourme E, et al. Variation in breast cancer risk associated with factors related to pregnancies according to truncating mutation location, in the French National BRCA1 and BRCA2 mutations carrier cohort (GENEPSO).. *Breast cancer research : BCR* 2012; 14: R99.

86. Maxwell CA, Benítez J, Gómez-Baldó L, et al. Interplay between BRCA1 and RHAMM regulates epithelial apicobasal polarization and may influence risk of breast cancer.. *PLoS biology* 2011; 9: e1001199.

87. Mulligan AM, Couch FJ, Barrowdale D, et al. Common breast cancer susceptibility alleles are associated with tumour subtypes in BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2.. *Breast cancer research : BCR* 2011; 13: R110.

88. Cox DG, Simard J, Sinnett D, et al. Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carriers.. *Human molecular genetics* 2011; 20: 4732-47.

89. Rebbeck TR, Mitra N, Domchek SM, et al. Modification of BRCA1-Associated Breast and Ovarian Cancer Risk by BRCA1-Interacting Genes.. *Cancer research* 2011; 71: 5792-805.

90. Antoniou AC, Kartsonaki C, Sinilnikova OM, et al. Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for BRCA1 and BRCA2 mutation carriers.. *Human molecular genetics* 2011; 20: 3304-21.

91. Martrat G, Maxwell CM, Tominaga E, et al. Exploring the link between MORF4L1 and risk of breast cancer.. *Breast cancer research : BCR* 2011; 13: R40.

92. Osorio A, Milne RL, Alonso R, et al. Evaluation of the XRCC1 gene as a phenotypic modifier in BRCA1/2 mutation carriers. Results from the consortium of investigators of modifiers of BRCA1/BRCA2.. *British journal of cancer* 2011; 104: 1356-61.

93. Win AK, Dowty JG, Antill YC, et al. Body mass index in early adulthood and endometrial cancer risk for mismatch repair gene mutation carriers.. *Obstetrics and gynecology* 2011; 117: 899-905.

94. Spurdle AB, Marquart L, McGuffog L, et al. Common genetic variation at BARD1 is not associated with breast cancer risk in BRCA1 or BRCA2 mutation carriers.. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011; 20: 1032-8.

95. Ramus SJ, Kartsonaki C, Gayther SA, et al. Genetic variation at 9p22.2 and ovarian cancer risk for BRCA1 and BRCA2 mutation carriers.. *Journal of the National Cancer Institute* 2011; 103: 105-16.

96. Antoniou AC, Beesley J, McGuffog L, et al. Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction.. *Cancer research* 2010; 70: 9742-54.

97. Walker LC, Fredericksen ZS, Wang X, et al. Evidence for SMAD3 as a modifier of breast cancer risk in BRCA2 mutation carriers.. *Breast cancer research : BCR* 2010; 12: R102.

98. Gaudet MM, Kirchhoff T, Green T, et al. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer.. *PLoS genetics* 2010; 6: e1001183.

99. Engel C, Versmold B, Wappenschmidt B, et al. Association of the variants CASP8 D302H and CASP10 V410I with breast and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers.. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2010; 19: 2859-68.

100. Wang X, Pankratz VS, Fredericksen Z, et al. Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers.. *Human molecular genetics* 2010; 19: 2886-97.

101. Pande M, Lynch PM, Hopper JL, et al. Smoking and colorectal cancer in Lynch syndrome: results from the Colon Cancer Family Registry and the University of Texas M.D. Anderson Cancer Center.. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2010; 16: 1331-9.

102. Spurdle AB, Fahey P, Chen X, et al. Pooled analysis indicates that the GSTT1 deletion, GSTM1 deletion, and GSTP1 Ile105Val polymorphisms do not modify breast cancer risk in BRCA1 and BRCA2 mutation carriers.. *Breast cancer research and treatment* 2010; 122: 281-5.

103. Osorio A, Milne RL, Pita G, et al. Evaluation of a candidate breast cancer associated SNP in ERCC4 as a risk modifier in BRCA1 and BRCA2 mutation carriers. Results from the Consortium of Investigators of Modifiers of BRCA1/BRCA2 (CIMBA).. *British journal of cancer* 2009; 101: 2048-54.

104. Jakubowska A, Rozkrut D, Antoniou A, Hamann U, Lubinski J. The Leu33Pro polymorphism in the ITGB3 gene does not modify BRCA1/2-associated breast or ovarian cancer risks: results from a multicenter study among 15,542 BRCA1 and BRCA2 mutation carriers.. *Breast cancer research and treatment* 2010; 121: 639-49.

105. Antoniou AC, Sinilnikova OM, McGuffog L, et al. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers.. *Human molecular genetics* 2009; 18: 4442-56.

106. Rebbeck TR, Mitra N, Domchek SM, et al. Modification of ovarian cancer risk by BRCA1/2-interacting genes in a multicenter cohort of BRCA1/2 mutation carriers.. *Cancer research* 2009; 69: 5801-10.

107. Johnatty SE, Couch FJ, Fredericksen Z, et al. No evidence that GATA3 rs570613 SNP modifies breast cancer risk.. *Breast cancer research and treatment* 2009; 117: 371-9.

108. Osorio A, Pollán M, Pita G, et al. An evaluation of the polymorphisms Ins16bp and Arg72Pro in p53 as breast cancer risk modifiers in BRCA1 and BRCA2 mutation carriers.. *British journal of cancer* 2008; 99: 974-7.

109. Rebbeck TR, Antoniou AC, Llopis TC, et al. No association of TGFB1 L10P genotypes and breast cancer risk in BRCA1 and BRCA2 mutation carriers: a multi-center cohort study.. *Breast cancer research and treatment* 2009; 115: 185-92.

110. Antoniou AC, Spurdle AB, Sinilnikova OM, et al. Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.. *American journal of human genetics* 2008; 82: 937-48.

111. Hughes DJ. Use of association studies to define genetic modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers.. *Familial cancer* 2008; 7: 233-44.

112. Couch FJ, Sinilnikova O, Vierkant RA, et al. AURKA F31I polymorphism and breast cancer risk in BRCA1 and BRCA2 mutation carriers: a consortium of investigators of modifiers of BRCA1/2 study.. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2007; 16: 1416-21.

113. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA).. 2007.

114. Durocher F, Labrie Y, Soucy P, et al. Mutation analysis and characterization of ATR sequence variants in breast cancer cases from high-risk French Canadian breast/ovarian cancer families.. *BMC cancer* 2006; 6: 230.