Estimating heritability for cause specific mortality based on twin studies

Thomas H. Scheike, Klaus K. Holst
Department of Biostatistics, University of Copenhagen
Øster Farimagsgade 5, DK-1014 Copenhagen K, Denmark
email: ts@biostat.ku.dk, kkho@biostat.ku.dk

and Jacob B. Hjelmborg[2]
Department of Biostatistics, University of Southern Denmark
J. B. Winsløwsvej 9B, DK-5000 Odense, Denmark
email: jhjelmborg@health.sdu.dk

Summary

There has been considerable interest in studying the magnitude and type of inheritance of specific diseases. This is typically derived from family or twin studies, where the basic idea is to compare the correlation for different pairs that share different amount of genes. We here consider data from the Danish twin registry and discuss how to define heritability for cancer occurrence. The key point is that this should be done taking censoring as well as competing risks due to e.g. death into account. We describe the dependence between twins on the probability scale and show that various models can be used to achieve sensible estimates of the dependence within monozygotic and dizygotic twin pairs that may vary over time. These dependence measures can subsequently be decomposed into a genetic and environmental component using random effects models. We here present several novel models that in essence describe the association in terms of the concordance probability, i.e., the probability that both twins experience the event, in the competing risks setting. We also discuss how to deal with the left truncation present in the Nordic twin registries, due to sampling only of twin pairs where both twins are alive at the initiation of the registries.

Some key words: cause specific hazards, competing risks, delayed entry, left truncation, heritability, survival analysis,

1 Introduction

There has been considerable interest in characterizing the role of genetic and environmental factors and how they affect specific disease such as breast cancer or prostate

cancer. The much cited paper by Lichtenstein et al. (2000) reported for example that the variation due to genes for prostate cancer was 42% with 95 % confidence interval (29%-50%), similarly the heritability of breast cancer was estimated to be 27% (4%-41%). These numbers are deceptively simple to report, but the interpretation is really quite complex. The heritability estimates were based on a combined twin study of the Danish, Finnish, and Swedish twin registries, and random effects modeling of the occurrences of cancer using a Probit link. The methodology applied ignores the timing of the events, but the key problem is that the analyses do not correct for censoring. The resulting variance components, and derived measures of dependence may therefore be severely biased. We later give a simulation study that shows that there are substantial bias in the dependence estimates related to genetic and environmental effects. We show how the censoring can be taken into account, and how the method is closely related to the cumulative incidence function for the competing risks model.

The analyses of Lichtenstein et al. (2000) was based on describing the correlation on the probability scale, that is on the concordance probabilities that are defined as the probability that both twins experience a specific event. It is equally relevant to consider the dependence on the hazard scale as in for example Locatelli et al. (2004, 2007) that aimed to address the same issue using survival analysis techniques, and focused on the heritability of breast cancer. In addition they extended standard modeling to allow that a fraction of the population could be immune to the disease evaluated. This was based on correlated frailty models. This analysis ignores the correlation between different causes, and therefore the estimates of variance components associated with genes and environmental factors can be severely biased. Only when the random effects associated with other causes (death for example) are fully independent of the ones for breast cancer (for example) this will yield a correct analysis. In Ripatti et al. (2003) an extended model that allowed random effects for different causes to be dependent is considered, and recently this has been taken up again by Gorfine & Hsu (2010).

Another approach to characterize the dependence among the cause-specific failure times within a cluster is to consider cross hazard ratios for different causes, for example Bandeen-Roche & Liang (2002); Bandeen-Roche & Ning (2008). This also provides a very useful summary of the dependence structure, but it is quite different from our measures that describe dependence within pairs on the probability scale as will be discussed below. There is no simple link between dependence on the hazard scale and probability scale.

The estimates and separation between genetic and environmental factors have wide ranging consequences and it is therefore of critical importance to estimate such numbers correctly and to have a valid and useful interpretation. Clearly, this needs to be done under modeling assumptions, and understanding these and their consequences for estimation are equally important.

The aim of our work is to discuss a framework for talking about heritability for specific diseases and to suggest a number of models that can be used to describe the correlation within twins. This embeds polygenic models from quantitative genetics and naturally leads to components due to genetic and environmental factors, and with an interpretation on the probability scale in terms of the cumulative concordance probabil-

ities. The approach is based on direct modeling of dependence measures and random effects modeling of the probability of observing cancer in the competing risks model. We here extend Scheike et al. (2010) to a more elaborate modeling of the random effects in a twin setting. Further, the approach extends the direct estimation of dependence as in Scheike & Sun (2011) to a more elaborate and robust estimation of dependence which then subsequently can be separated into genetic and environmental sources. We also show how to test whether or not a genetic or environmental association is present in this framework.

In addition to our discussion of different models that can be used to separate genetic and environmental effects we also address the fact that all twin registries have left-truncated event times. This occurs in varying degree for the Nordic twin registries, but they all have some amount of delayed entry, in the sense that they only include twin pairs where both where alive at a given date. In the Danish twin registry that we use to illustrate this point all twin pairs included were both alive in 1943 when the cancer registration started. The left truncation is more severe for the other Nordic twin registries. Dealing with the multivariate delayed entry is no routine extension as it involves the parameters of interest and additional quantities related to this. Recently, there have been some work on handling delayed entry for regression models for the cumulative incidence function Zhang et al. (2011); Geskus (2010), and we here suggest a related approach for the direct binomial approach (Scheike et al., 2008) that is extended to the multivariate data considered here.

The paper is structured as follows. First we present the competing risks framework and discuss some consequences of this model in the context considered here. In particular, we note that dependence between frailty terms on the hazard scale and between events on the probability scale are quite different. We then specify a dependence measure in this context and show how to do regression modeling that quantifies genetic and environmental sources of dependence. We also suggest new variance components modeling with the aim of estimating the degree of association due to genes and environmental factors through random effects. After having introduced the basic modeling approaches we show how to deal with left truncation for estimation of these dependence measures. A specific section shows the results of ignoring censoring for the liability threshold model. The last part of the paper considers a worked example based on the Danish twin registry.

2 Competing Risks Modeling

When a specific cause of death or disease is of interest and other causes may interfere or compete with this cause, we need to consider the competing risks model. The competing risks model takes its basis in terms of the cause specific intensities. If we for example consider the occurrence of a specific cancer form, then the death of a subject will prohibit the subject from experiencing the specific cancer form, and is thus a competing event. With these two causes in the competing risks models we can define the two cause specific intensities as $\lambda_{\text{cancer}}(t)$ and $\lambda_{\text{non-cancer}}(t)$. With T the event time and ϵ the cause related

to the event time, the cause specific hazard of cancer is defined as

$$\lambda_{\text{cancer}}(t)dt = P(T \in [t, t+dt), \epsilon = \text{cancer} | \text{alive at time t})$$

and similarly for the non-cancer deaths.

We have that the probability of a cancer given an event at time t is

$$\frac{\lambda_{\text{cancer}}(t)}{\lambda_{\text{cancer}}(t) + \lambda_{\text{non-cancer}}(t)}.$$

Further, in the special case where the cause specific hazards do not depend on time, we have that the probability of a subject getting cancer is

$$P(\text{cancer}) = \frac{\lambda_{\text{cancer}}}{\lambda_{\text{cancer}} + \lambda_{\text{non-cancer}}}. \tag{1}$$

It is useful to keep this simple relationship in mind. Further, when additional right censoring is also present and the censoring has constant intensity λ_c then the probability of seeing a cancer occurrence is

$$P(\text{“seeing cancer”}) = \frac{\lambda_{\text{cancer}}}{\lambda_{\text{cancer}} + \lambda_{\text{non-cancer}} + \lambda_c}. \tag{2}$$

We note that the probability of “cancer” is different from the probability of “seeing cancer” with censored data. This creates a problematic bias for models of the lifetime risk of cancer or the cumulative concordance probability of cancer for monozygotic or dizygotic twins.

Now considering a pair of twins, that we denote 1 and 2, and random effects associated with the two causes $Z_{\text{cancer},i}$ and $Z_{\text{non-cancer},i}$ for the two twins $i = 1, 2$, such that the cause specific hazard of cancer is $Z_{\text{cancer},i}\lambda_{\text{cancer}}$ and the cause specific hazard of other events is $Z_{non-cancer,i}\lambda_{\text{non-cancer}}$ for the two twins $i = 1, 2$. Then given the random effects we get

$$P(\text{twin “1” gets cancer}) = \frac{Z_{\text{cancer},1}\lambda_{\text{cancer}}}{Z_{\text{cancer},1}\lambda_{\text{cancer}} + Z_{\text{non-cancer},1}\lambda_{\text{non-cancer}}}$$

$$P(\text{twin “2” gets cancer}) = \frac{Z_{\text{cancer},2}\lambda_{\text{cancer}}}{Z_{\text{cancer},2}\lambda_{\text{cancer}} + Z_{\text{non-cancer},2}\lambda_{\text{non-cancer}}}$$

We note that if $Z_{\text{cancer},i} = Z_{\text{non-cancer},i}$ for $i = 1, 2$ such that there is the strongest possible dependence between the random effects on the hazard scale, then the random effects cancel out on the probability scale and the probabilities are independent. This model is not likely to be realistic but it is likely that the random effects are correlated. It is useful to note that dependence on the hazard scale is different from dependence on probability scale. If the underlying cause specific hazards depend on time, then random effects will not cancel out, and there will be correlation present among occurrence of events. This correlation will vary with time. If in addition censoring independent of the random effects is present and with constant hazard λ_c then the probabilities will also have the λ_c term in the numerator of these expressions. One consequence of this is that if censoring is not corrected for then there will suddenly be correlation in the occurrences of the 'observed cancers' inducing false dependence either stronger or weaker.

3 Dependence measures for cancer events

The analysis in the combined Nordic study of the Danish, Finnish and Swedish twins registries, Lichtenstein et al. (2000) was based on the assumption that the probability of breast cancer occurrence (denoted as cancer below) for twin j in a twin pair was of the form

$$\text{Probit}(P(\text{twin j gets cancer}|X_j, Z_j)) = X_j^T\beta + Z_j, \quad j = 1, 2 \tag{3}$$

where X_j are possible covariates that we wish to adjust for and Z_1 and Z_2 are correlated normal random effects associated with the twin pair. The classical twin design where both MZ and DZ pairs are included in the study, makes it possible to decompose the random effect variance into genetic and environmental components by noteworthy assuming independence of these random effects, $Z_j = Z_{j,\text{gene}} + Z_{j,\text{env}}$. The heritability of getting cancer in this model is then defined in terms of the random effects as the fraction of the total variance due to genetic factors. We will adopt the standard polygenic model for twin data based on further decomposing the random effects Z_j for each twin in a pair into the genetic components of additive (A) and dominant (D) effects and the environmental effects into shared (C) and non-shared effects (E):

$$Z_j = Z_{A,j} + Z_{D,j} + Z_C + Z_{E,j}, \quad j = 1, 2. \tag{4}$$

We denote the variances of each ot the terms as $\sigma_A^2, \sigma_D^2, \sigma_C^2$ and σ_E^2, respectively. The A, D, and C components are all assumed to be shared for MZ twins, whereas for DZ twins who genetically are like normal siblings, we have

$$\text{cov}(Z_{A,1}, Z_{A,2}) = 0.5\sigma_A^2, \qquad \text{cov}(Z_{D,1}, Z_{D,2}) = 0.25\sigma_D^2.$$

All other pairs of components are assumed to be independent. This is the basis for the standard heritability estimate, defined as the part of the total variance due to genes $((\sigma_A^2 + \sigma_D^2)/(\sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2))$. For quantitative data where the observations Y_j equals Z_j, all terms can be identified under the independence assumptions. In the Probit model the E component is modeled indirectly through the link-function, which is equivalent to a latent variable model formulation where the binary outcome is defined from a conditionally normal distributed latent variable

$$Y_j^* = X_j^T\beta + Z_{A,j} + Z_{D,j} + Z_C + Z_{E,j}$$

such that

$$\text{cancer}_j = \begin{cases} \text{Yes}, & Y_j^* > \delta \\ \text{No}, & \text{otherwise.} \end{cases}$$

where cancer_j is the binary cancer indicator for subject j. This model is termed the liability-threshold model see Falconer (1967); Falconer & Mackay (1994). For identification the threshold is fixed at $\delta = 0$ and $Z_{E,1}$ and $Z_{E,2}$ are independent standard normal

distributed. Further, in practice only two of the variance components Z_A, Z_C and Z_D can be identified unless information on for example adoption status are included in the model. For more on different genetic models see for example Neale & Cardon (1992) or Sham (1998). Note, however that the random effects relate to the Probit-probability scale, and that there therefore are additional variation present in the data due to the probability itself. Using only the random effects to define a heritability estimate as above is thus not comparable to the one from the standard normal model where all the variation is included in the heritability estimate. The Probit random effects analyses can be (and have been) criticized for completely ignoring the time-aspect and the fact that the analyses do not take the censoring into account which makes it impossible to interpret the results, as indicated above. We later give a simulation study that shows that the bias can be large in the standard liability-threshold model in the presence of right censoring.

3.1 Cumulative incidence modeling

We now take an approach that has the same aim but considers the time-aspect and corrects for censorings and thus leads to interpretable estimates.

First we define the transition probability of cancer in the competing risks model as

$$F_1(t;X) = P(T \leq t, \epsilon = \text{cancer}|X)$$

that is the probability of cancer before time t given possible covariates X. This quantity is the cumulative incidence function. We note that $F_1(\infty;X)$ is a consequence of model (3), and is achieved after integrating out the random effect. The key interest here is now to specify dependence measures for these probabilities, and then subsequently to separate the amount of dependence due to heritable factors and shared environmental factors by comparing the dependence for monozygotic and dizygotic twins.

The cumulative incidence model was recently extended to clustered data by random effects models in Katsahian et al. (2006); Scheike et al. (2010). Let the event times and cause indicators for the two twins be (T_1, ϵ_1) and (T_2, ϵ_2) and denote their covariates as X_1 and X_2, respectively. Let the causes of events be either "cancer" or "death". Now the random effects models for the cumulative incidence are based on assuming that conditional on a random effect Z which is shared among the two twins and is independent of the observed covariates $X_j, j = 1, 2$, the cumulative incidence of cancer for the twins are independent and given as

$$F(t;X_j,Z) = P(T_j \leq t, \epsilon_j = \text{cancer}|Z, X_j) = 1 - \exp\left(-Z\Psi_\theta^{-1}\left[\Lambda(t,X_j)\right]\right), \tag{5}$$

where $\Psi_\theta(t) = E_\theta\{\exp(-Zt)\}$ is the Laplace transform of the random effect Z that here is assumed gamma distributed with shape parameter $1/\theta$ and scale parameter θ and $\Lambda(t,X_j) = \exp(-\alpha(t) - (\gamma^T X_j)t)$, where $\alpha(t)$ is a baseline regression function and γ is a q-dimensional vector of parameters. This leads to random effects with mean 1 and variance θ. A useful property of this model is that the marginal cumulative incidence given only X_j is of the simple additive form

$$F_1(t;X_j) = 1 - \exp(-\alpha(t) - t\gamma^T X_j) \tag{6}$$

see Scheike et al. (2008) for more details on this model. In the worked example considered later there are no covariates present in the model, so this is simply the marginal cumulative incidence of breast cancer, and then this is simply a re-parameterization of a non-parametric cumulative incidence function. To estimate such marginal cumulative incidence for correlated data, see Chen et al. (2008); Scheike et al. (2010). Now we can compute the bivariate cumulative incidence on copula form as

$$P(T_1 \leq t, \epsilon_1 = \text{cancer}, T_2 \leq s, \epsilon_2 = \text{cancer} | X_1, X_2) = \\ 1 - (1 - F_1(t; X_1)) - (1 - F_1(s; X_2)) + \Psi_\theta \left(\Psi_\theta^{-1} \left[\Lambda(t, X_1) \right] + \Psi_\theta^{-1} \left[\Lambda(s, X_2) \right] \right).$$

For this model the variance of the random effect, θ, is an indication of the amount of positive correlation. In the next section we shall consider more elaborate random effects modeling in this model.

We define the cross-odds ratio as the ratio between the conditional odds and the marginal odds

$$\text{cOR}(t) = \frac{\text{odds}(P(T_1 \leq t, \epsilon_1 = \text{cancer} \mid T_2 \leq t, \epsilon_2 = \text{cancer}))}{\text{odds}(P(T_1 \leq t, \epsilon_1 = \text{cancer}))} \tag{7}$$

where $\text{odds}(P(A)) = P(A)/(1-P(A))$ and $\text{odds}(P(A|B)) = P(A|B)/(1-P(A|B))$. This measure can be estimated directly (Scheike & Sun, 2011) and can be modeled depending on covariates. It was also shown that the random effects model lead to a (almost) constant cross-odds ratio, $\text{cOR}(t) = 1 + \theta$, for the probability of a twin experiencing cancer given a cancer occurrence for the other twin.

We also consider direct estimation of the relative risk dependence measure

$$\text{RR}(t) = \frac{\mathcal{C}(t)}{P(T_1 \leq t, \epsilon_1 = \text{cancer}) P(T_2 \leq t, \epsilon_2 = \text{cancer})} \tag{8}$$

where $\mathcal{C}(t) = P\left(T_1 \leq t, \epsilon_1 = \text{cancer}, T_2 \leq t, \epsilon_2 = \text{cancer}\right)$. The numerator, $\mathcal{C}(t)$, that we denote as the cumulative concordance probability is modeled relative to the concordance under independence. Similarly, the cOR also leads to estimates of the cumulative concordance probability as shown in Scheike & Sun (2011).

The basic idea is now to compare these measures for monozygotic and dizygotic twins. Specifically we shall consider regression models of the form

$$\log(\text{cOR}(t)) = \log(\text{cOR}(t, \theta)) = V^T \theta \quad \log(\text{RR}(t)) = \log(\text{RR}(t, \theta)) = V^T \theta \tag{9}$$

where V is twin-pair specific covariate and θ is a parameter vector. In the worked example below (Section 7) we consider both non-parametric and parametric estimates of, for example, $RR(t)$, that we denote as both $RR(t)$ and $RR(t, \theta)$. We simply use a regression vector V that specifies the zygosity of the twins. This simple principle can be used to estimate the amount of the cOR and RR that is due to heritable factors (genes) and shared environmental factors. We return to this point in Section 7.

Note that the cumulative concordance probability, $\mathcal{C}(t)$, can be estimated under the different dependence models suggested above, and compared with the expected cumulative concordance probability under independence. We provide such estimates in the worked example. Clearly, under models for $RR(t)$ we can compute the concordance as $RR(t)P(T_1 \leq t, \epsilon_1 = \text{cancer})P(T_2 \leq t, \epsilon_2 = \text{cancer})$, and a similar expression can be worked out for the cOR model. In the case without covariates we can of course also estimate the concordance rate non-parametrically by inverse probability of censoring weighting techniques, or by some of the available estimators, see Cheng et al. (2007).

3.2 Biometric modeling

We now consider biometric modeling of twin data in terms of the above risk measures, in particular the heritability in risk of disease and its variation over time. Further, we consider genetic heterogeneity suggesting a mode of action of genetic variants.

For event times T_1 and T_2 in a pair, let $C_{T_1,T_2}(t)$ denote the (cumulative) covariance depending on time, $P(T_1 \leq t, \epsilon_1 = \text{cancer}, T_2 \leq t, \epsilon_2 = \text{cancer}) - P(T_1 \leq t, \epsilon_1 = \text{cancer})P(T_2 \leq t, \epsilon_2 = \text{cancer})$. Then the relative risk in (8) above can be expressed in terms of $C_{T_1,T_2}(t)$ as

$$\text{RR}(t) = 1 + \frac{C_{T_1,T_2}(t)}{P(T_1 \leq t, \epsilon_1 = \text{cancer})P(T_2 \leq t, \epsilon_2 = \text{cancer})} \tag{10}$$

for $t \geq 0$.

We may decompose the genetic effect at a given loci into components of additive genetic covariance, $\sigma^2_A(t)$ and dominant genetic covariance, $\sigma^2_D(t)$ by assuming that $C_{T_1,T_2}(t) = \sigma^2_A(t) + \sigma^2_D(t)$ for MZ pairs and $C_{T_1,T_2}(t) = \frac{1}{2}\sigma^2_A(t) + \frac{1}{4}\sigma^2_D(t)$ for DZ pairs. This decomposition may be extended by the effects of shared environment, $\sigma^2_C(t)$ as seen above. Extending the notion in (Risch, 1990) we define the time dependent *multilocus index* as the ratio

$$\frac{RR_{mz}(t) - 1}{RR_{dz}(t) - 1}$$

and note that this index equals two in case of solely additive contributions to $\sigma^2_A(t)$ and $\sigma^2_D(t)$ from multiple loci whereas the multilocus index exceeds two in case of multiplicative effects of genetic loci, ie. due to interaction of genes. The multilocus index is in case of the natural assumption of equal marginals of MZ and DZ twins in fact the ratio of covariances $C_{T_1,T_2}(t)$ of MZ to that of DZ pairs.

As a straightforward measure of time-varying heritability in risk of disease we adopt the polygenic ACDE model in quantitative genetics (see (4) above) (Sham, 1998) which in case of the estimable ACE submodel is simply the ratio of twice the differences in covariances of MZ and DZ twins to the total variance, hence we define as heritability in

risk of disease,

$$h^2(t) = \frac{2(C^{mz}_{T_1,T_2}(t) - C^{dz}_{T_1,T_2}(t))}{P(T_1^{MZ} \leq t, \epsilon_1^{MZ} = \text{cancer})\,(1 - P(T_1^{MZ} \leq t, \epsilon_1^{MZ} = \text{cancer}))} = \frac{2\,(\mathcal{C}_{\text{MZ}}(t) - \mathcal{C}_{\text{DZ}}(t))}{F_1(t)(1 - F_1(t))} \tag{11}$$

assuming equal marginals of MZ and DZ twins ($F_1(t) = F_1^{MZ}(t) = F_1^{DZ}(t)$) and denoting for example the observations for the MZ twin pair as $(T_1^{MZ}, T_2^{MZ}, \epsilon_1^{MZ}, \epsilon_2^{MZ})$. We note that this measure takes into account the variance on risk scale not accounted for by the heritability in liability to disease from the Probit (or liability threshold) model above. In light of the above time varying heritability, we stress the importance of modeling the concordance functions in a flexible manner that can be faithful to the data. This is pursued in the worked example in Section 7.1.

3.3 Asymptotics

Let T_{ki} and C_k be the event time and right censoring time for the ith individual ($i = 1, 2$) within the kth twin pair ($k = 1, \ldots, K$), respectively, and let $\epsilon_{ki} \in \{j = 1, \ldots, J\}$ denote the failure type. We define the indicator $\Delta_{ki} = \mathcal{I}(T_{ki} \leq C_k)$. Let X_{ki} be an associated covariate. Note that we here assume that both twins are censored at the same time. This is used in deriving the asymptotics for the estimators considered as in Scheike et al. (2010); Scheike & Sun (2011) that used inverse probability censoring weighted estimators under a different censoring assumption. The asymptotics expansions thus differs for those used in these papers that assumed that the censorings were independent within a twin-pair, but follows along the same lines. The censoring assumption is valid for the twin data we consider in this paper because the twins are administratively censored. This makes it a lot simpler to use inverse probabilty of censoring weighting techniques. If a few of the twins should not be censored at the same time due to for example emigration, we can censor these pairs at the first censoring time and then the data we use satisfy the same censoring assumption. We assume that we observe K independent identically distributed replications of this data $(X_{ki}, \tilde{T}_i, \tilde{\epsilon}_i)$, where $\tilde{T}_{ki} = \min(T_{ki}, C_k)$, $\tilde{\epsilon}_{ki} = \epsilon_{ki}\,\Delta_{ki}$, and $X_{ki} = (1, X_{ki1}, \ldots, X_{kip})^{\mathrm{T}}$. We assume that $(T_{ki}, \epsilon_{ki}, i = 1, 2)$ are independent of C_k given covariates.

First we recall that we can estimate the marginal cumulative incidence in a regression setting as in Scheike et al. (2010) by the GEE approach using robust cluster based standard errors. In the special case without covariates we simply use the product limit estimator with GEE type standard errors.

To be concrete we consider marginal regression models of the form

$$F_1(t; X_{ki}) = g(\alpha(t), X_{ki}^T \gamma, t)$$

for some specific link function g, see Scheike et al. (2008) and (6), and assume that we have an estimation procedure at hand that lead to estimators with nice properties. We

denote these estimators $\hat{\alpha}(t)$ and $\hat{\gamma}$, and assume that these quantities have an asymptotic linear influence function such that as $K \to \infty$, $\sqrt{K}(\widehat{\gamma} - \gamma)$ and $\sqrt{K}(\widehat{\alpha}(t) - \alpha(t))$ are asymptotically Gaussian (Gaussian processes) and asymptotically equivalent to the following sums of i.i.d. variables

$$K^{-1/2} \sum_{k=1}^{K} W_{\gamma,k}(\tau) \qquad \text{and} \qquad K^{-1/2} \sum_{k=1}^{K} W_{\alpha,k}(t).$$

Thus $\sqrt{K}(\widehat{\gamma} - \gamma)$ converges in distribution to a mean zero normal random variable, and $\sqrt{K}(\widehat{\alpha}(t) - \alpha(t))$ converges weakly to a mean zero Gaussian process on $t \in [0, \tau]$, for some upper limit $\tau > 0$. One implication of this is that for a given covariate vector X_0 we have that $\widehat{F_1}(t; X_0) = h(\widehat{\alpha}(t), X_0\widehat{\gamma}, t)$ is asymptotically equivalent to

$$K^{-1/2} \sum_{k=1}^{K} g_F(W_{\alpha,k}(t), W_{\gamma,k}(\tau), X_0, t),$$

where g_F is derived from the specific form of the link-function g.

Similarly, for the random effects model (4) and the direct dependence models (7) (cOR and RR) we have an asymptotic expansion such that as $K \to \infty$, $\sqrt{K}(\widehat{\theta} - \theta_m)$ is asymptotically Gaussian and asymptotically equivalent to the following sum of i.i.d. processes as in Scheike et al. (2010); Scheike & Sun (2011)

$$K^{-1/2} \sum_{k=1}^{K} W_{\theta,k}(\tau).$$

Also we can estimate the covariance of these dependence parameters as $\sum_{k=1}^{K} \hat{W}_{\hat{\theta},k}(\tau)^{\otimes 2}$, where for a p-vector $a^{\otimes 2} = aa^T$ i.e. the $p \times p$ matrix.

One consequence of these asymptotic results are that when considering the concordance probability function $\mathcal{C}(t) = H(\theta, \alpha, \gamma, X_0)$ where H depends on what specific dependence model we are using, we can derive the standard errors of its estimator $H(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma})$, using the i.i.d. representations. Specifically it follows that the concordance estimator is asymptotically Gaussian with the i.i.d decomposition

$$K^{-1/2} \sum_{k=1}^{K} m(W_{\alpha,k}(t), W_{\gamma,k}, W_{\theta,k}, X_0, t),$$

for some function m that is derived from H. Consequently the estimator of the proposed heritability measure (11) is also asymptotically Gaussian with a similar i.i.d. representation that can be used for estimating its standard error and making confidence bands. Here the computation are done for each specific covariate X_0. In the worked example in Section 7 there are no covariates.

4 Random effects modeling for the cumulative incidence function

In this section we discuss how to formulate random effects models directly that aims at decomposing the amount of dependence that is due to genetic and environmental factors. We do this by extending the simple random effects model to allow a more elaborate additive random effects structure. This is a direct extension of standard genetic random effects models to the cumulative incidence setting, and is of course strongly related to the Probit-random effects models described in the previous section.

For a set of twins with random effects Z_1 and Z_2 with parameters, θ_1 and θ_2, respectively we now assume that their marginal cumulative incidences are independent and given as

$$\begin{aligned} F_1(t; Z_j, X_j) &= P(T_j \leq t, \epsilon_j = \text{cancer} | Z_j, X_j) \\ &= 1 - \exp\left(-Z_j \Psi_{\theta_j}^{-1}\left[\exp\left\{-\alpha(t) - (\gamma^T X_j)t\right\}\right]\right), \end{aligned} \tag{12}$$

for twin j=1,2 with covariates X_j, and where $\Psi_{\theta_j}(\cdot)$ is the Laplace transform of Z_j. We make the additional formal requirement that $P(T_1 \leq t, \epsilon_1 = \text{cancer} | Z_1, X_1, Z_2, X_2) = P(T_1 \leq t, \epsilon_1 = \text{cancer} | Z_1, X_1)$ and similarly for the other twin. We now specify a more detailed polygenic structure for the random effects for the twin design with one random effect due to additive genetic effects (A) and one random effect due to common environmental factors (C), leading to an ACE model (see above and (Neale & Cardon, 1992)) on the probability scale. These random effects are the sources for the dependence. Specifically we assume that a set of random effects (Z_1, Z_2) for a twin pair is given as

$$Z_1 = Z_g + Z_c, \qquad Z_2 = Z_g + Z_c$$

for the monozygotic twins such that all genes and environmental factors are shared and independent, and with $\text{var}(Z_g) = \sigma_g^2$ and $\text{var}(Z_c) = \sigma_c^2$. Whereas for dizygotic twins

$$Z_1 = Z_{gs} + Z_{g,ns,1} + Z_c, \qquad Z_2 = Z_{gs} + Z_{g,ns,2} + Z_c$$

where the random effects $Z_{gs}, Z_{g,ns,a}, Z_{g,ns,b}$ and Z_c are independent with $\text{var}(Z_{gs}) = 0.5\sigma_g^2$, $\text{var}(Z_{g,ns,j}) = 0.5\sigma_g^2$ for $j \in \{a, b\}$, and $\text{var}(Z_c) = \sigma_c^2$, such that only half the genes are shared. To formally fit this model we make the specific assumptions that Z_c is gamma distributed with shape parameter λ_c and scale parameter λ_{tot}^{-1}, Z_g is gamma distributed with shape parameter λ_g and scale parameter λ_{tot}^{-1}, and with Z_{gs}, $Z_{g,ns,j}$ gamma distributed with shape parameter $0.5\lambda_g$ and scale parameter λ_{tot}^{-1}. Here the total variation due to genes and environmental factors is $\lambda_{tot} = \lambda_g + \lambda_c$. This parameterization implies that $E(Z_a) = E(Z_b) = 1$ and with $\text{var}(Z_a) = \text{var}(Z_b) = \lambda_{tot}^{-1}$. Therefore for example also $E(Z_g) = \lambda_g \lambda_{tot}^{-1}$, and $\text{var}(Z_g) = \lambda_g \lambda_{tot}^{-2}$. We now define 5 random effects that are i.i.d. across twins pairs, and that makes notation easier below. We let $\tilde{Z}_1 = Z_g$, $\tilde{Z}_2 = Z_{gs}$, $\tilde{Z}_3 = Z_{gs,ns,1}$, $\tilde{Z}_4 = Z_{gs,ns,2}$, and $\tilde{Z}_5 = Z_c$. We let the parameters of the 5 associated random effects by denoted as λ_j for $j = 1, \ldots, 5$. Then we can write the

random effects for a twin as $\tilde{Z}_1 \cdot I(\text{zygosity} = MZ) + \tilde{Z}_2 \cdot I(\text{zygosity} = DZ) + \tilde{Z}_3 \cdot I(\text{zygosity} = DZ, \text{twin no} = 1) + \tilde{Z}_4 \cdot I(\text{zygosity} = DZ, \text{twin no} = 2) + \tilde{Z}_5$, where "MZ" is a monozygotic twin and "DZ" is a dizygotic twin, and "twin no" gives a numbering of the twins within a twin pair.

Using this parameterization we get that for a pair of twins with survival times (T_1, T_2) and with death causes (ϵ_1, ϵ_2) we can compute the concordance rate as a function of time t as

$$\begin{aligned}\mathcal{C}(t) = 1 - (1 - F_1(t; X_1)) - (1 - F_1(t; X_2)) + \\ \prod_{j=1}^{5} \Psi(\lambda_j, \lambda_{tot}, \Psi^{-1}(\lambda_{tot}, \lambda_{tot}, \exp(-\alpha(t) - X_1^T \gamma t)) + \\ \Psi^{-1}(\lambda_{tot}, \lambda_{tot}, \exp(-\alpha(t) - X_2^T \gamma t))),\end{aligned}$$

where $\Psi(\lambda_j, \lambda_{tot}, \cdot)$ is the Laplace transform of a Gamma distribution with shape parameter λ_j and shape parameter λ_{tot}. We can estimate the parameters of this model as in Scheike et al. (2010), and the asymptotic properties can be derived along the same lines, so we omit these technical details. We outline how this is done in the next section that considers an extension to deal with delayed entry for the suggested models. The key is that we first estimate the marginal parameters, and then subsequently use these for estimation of the dependence parameters.

Similarly, one can also consider alternative genetic models such as the AE submodel, that we consider later in Section 7.

5 Delayed entry

In the Nordic twin registries there is delayed entry. We give more details on the data in the next section. Here we describe how to estimate the dependence parameters for the random effects model under delayed entry. The cOR and RR dependence measures for the cumulative incidence functions can also be estimated along the same lines, but then one needs additional modeling assumptions.

Denoting the event times for the two twins as T_1 and T_2 and the cause indicators as ϵ_1 and ϵ_2 we have the two basic marginal cumulative incidence functions, $F_j(t) = P(T_1 \leq t, \epsilon_1 = j)$ for $j = 1, 2$, and the bivariate cumulative incidence functions are $F_{i,j}(t, s) = P(T_1 \leq t, \epsilon_1 = i, T_2 \leq s, \epsilon_2 = j)$. We denote the truncation time as V, and assume it is the same for both twins as is indeed the case for the Nordic twin registries. We omit the covariates from these expressions for ease of notation. The marginal survival probability is $G(t) = 1 - F_1(t) - F_2(t)$, and the joint survival function $P(T_1 > t, T_2 > t) = 1 - 2F_1(t) - 2F_2(t) + F_{1,1}(t, t) + F_{1,2}(t, t) + F_{2,1}(t, t) + F_{2,2}(t, t)$. These quantities depend on the dependence parameters for different causes, and there are some complicated restrictions that must be satisfied for all the models to be consistent. We therefore avoid modeling all pairwise cumulative incidence functions. This leads to some robustness in terms of the modeling, but there will be some lost efficiency to a full MLE that considers all causes and their combinations.

To construct score equations for the dependence parameter of the concordance probability, note that for each set of causes (i, j)

$$R_1(i,t)R_2(j,t) - P(T_1 \leq t, \epsilon_1 = i, T_2 \leq t, \epsilon_2 = j | T_1 \geq V, T_2 \geq V)$$

with $R_i(j,t) = I(T_i \leq t, \epsilon_i = j)$ for $i = 1, 2$ has mean 0. Further,

$$P(T_1 \leq t, \epsilon_1 = 1, T_2 \leq t, \epsilon_1 = 1 | T_1 \geq V, T_2 \geq V) = \\ (F_{1,1}(t,t) - F_{1,1}(t,V) - F_{1,1}(V,t) + F_{1,1}(V,V))/P(T_1 \geq V, T_2 \geq V)$$

Now, the joint (bivariate) cumulative incidence is denoted as $F_{1,1}(s,t)$ and can be computed on copula form as indicated earlier, and depends on the dependence parameters θ. Suppose that we know the truncation probability, $P(T_1 \geq V, T_2 \geq V)$, then we note that

$$\tilde{U}(\theta, t, V) = R_1(1,t)R_2(1,t) - \frac{F_{1,1}(t,t) - F_{1,1}(t,V) - F_{1,1}(V,t) + F_{1,1}(V,V)}{P(T_1 \geq V, T_2 \geq V)}$$

also has mean 0. We stress that $F_{1,1}(t,s)$ depend on the dependence parameter θ.

To deal with right censoring we consider an IPCW weighted version of the above equation that still has mean 0 if the joint censoring of the twins is correctly modeled given covariates and then

$$U(\theta, t, V) = U(\theta, t, V, T_1, T_2, \epsilon_1, \epsilon_2) \\ = \frac{\Delta_1 \Delta_2 R_1(1,t)R_2(1,t)}{G_c(T_1, T_2, V)} - \frac{F_{1,1}(t,t) + F_{1,1}(V,V) - F_{1,1}(t,V) - F_{1,1}(V,t)}{P(T_1 \geq V, T_2 \geq V)}$$

has mean 0, where $\Delta_j = I(T_j \leq C_j)$ for $j = 1, 2$ and $G_c(T_1, T_2, V)$ is the bivariate survival function for the censoring times among those where both are uncensored at time V.

A special observation here is that the right censoring times for a twin pair are exactly the same, i.e., $C_1 = C_2$ such that $G_c(T_1, T_2, V) = \min(G_c(T_1), G_c(T_2))/G_c(V)$. To use this in practice we replace G_c by an estimator, based on left truncated survival times. Note that the censoring distribution is truncated as well. Further all marginal cumulative incidence functions are also replaced by estimates that take the truncation into account. In addition we also estimate the joint survival function by the two-stage estimator of Glidden (2000) adapted to deal with left-truncation.

Let K be the number of independent twin pairs that are observed, and let the score for each twin pair be defined as above and denoted as $U_k(\theta, t, V_k)$. Then a possible estimating equation for θ that is based only on modeling of the concordance probability function is given by

$$U_\theta(\theta, \hat{G}, \hat{P}) = \int_0^\tau \sum_{k=1}^{K} I(t > V_k) \frac{\partial U_k(\theta, t, V_k)}{\partial \theta} U_k(\theta, t, V_k)\, dt = 0. \tag{13}$$

The estimator $\widehat{\theta}$ of θ is obtained by solving (13), standard errors and asymptotic normality can be established along the lines of Section 3.3. Note, that the truncation weight must be estimated from some other bivariate survival model, and the asymptotics should reflect the uncertainty from this estimate.

6 Bias due to censorings using the Liability threshold model

We simulated data similar to the Danish breast cancer data using the cross-odds ratio model of Scheike & Sun (2011) with a cOR at 3 for the MZ twins and 2 (or 1.5) for the DZ twins similarly to what we saw in the data. These dependencies also generate two interesting models in terms of genetic and environmental effects on the dependence. The marginal cumulative incidence was chosen to mimic the cancer data and resulted in a marginal cumulative incidence at 12% close to what we saw in the data. Based on simulations from this model we estimated the heritability using an ACE-probit model similarly to Lichtenstein et al. (2000). We censored the data by different levels of censoring ranging from 0% to 82% for MZ and DZ twins separately, and then applied uniform censoring on the time-interval. The censoring was assumed to be the same for both twins within a twin-pair to mimic the administrative censoring present in the data. In reality the censoring will typically depend on the zygosity due to the changes in MZ and DZ prevalence over time, see Pinborg (2005) and Skytthe et al. (2003). Thus leading to differences in the censoring pattern for MZ and DZ twins. To see if different censorings degrees for the MZ and DZ twins gave additional complications we further allowed the censoring to depend on zygosity. We simulated 400000 MZ and 400000 DZ twin pairs because the simulations are meant to consider bias issues only. Variance estimation is another topic and beyond the scope of this paper. Results were stable over multiple simulations.

First, we consider the situation with the cOR at 3 and 2 for the MZ and DZ twins, respectively. we note that when ignoring the censoring the marginal cumulative incidence is severely biased (see Table 1), and similarly the concordance lifetime probabilities are completely skewed due to the censorings. Unbiased estimates of the true values are given for the case with censoring being 0% for both MZ and DZ twins. Without censorings the lifetime prevalence is close to the true 12 %, and similarly the model produces a lifetime concordance probabilty close to the true MZ concordance 0.035. Indeed the liability threshold model should produce unbiased estimates in the case without right censoring.

`Table 1 around here`

The marginal estimates are biased in a symmetric fashion for MZ and DZ censoring. the highest degree of censoring (82% for both MZ and DZ) the marginal estimate is reduced to 0.6%. This is related to our discussion in Section 2, and is a direct effect of nuisance due to the censorings. The concordance estimates for the MZ twins based on the liabilty threshold ACE model also shows a large bias dependent on the censoring pattern, the observed quantities range from 6.9% to 0.2% . Generally, the more MZ censoring the higher the concordance probability, this is consistent with the same censoring generating positive dependence in the binary outcomes. The reported concordance is based on an average of the dependence of the MZ and DZ twins using the ACE model for these dependencies.

Subsequently, we look at the dependence estimates in terms of the ACE decomposition of the dependencies based on fitting the standard ACE liability threshold model

ignoring the censorings. Table 2 shows the estimate of the genetic and common environmental components for different censoring patterns for the MZ twins having cOR 3 and the DZ twins having cOR 1.5, similarly to what we found in the breast cancer data that we consider in the next Section. When the data is uncensored we find that the genetic variance (A) of the Liability threshold model is 37% and the common environmental component (C) is 0%. This represents the best ACE approximation of our data generation. The effect of censoring can be seen be comparing to these numbers. On the diagonal where the censoring is the same for MZ and DZ twins we see that the C component increases, due to the dependence generated by the common censoring times that makes the DZ twins increase in correlation to a degree where a C component becomes a possibility. This is reflected also in the decreasing A component but to a lesser extend. The A component is highly sensitive to the censoring patterns and ranges from 99% to 0%. The C component ranges from 64% to 0% over the range of different censoring patterns.

`Table 2 around here`

Subsequently, we considered data where the cOR was 3 for MZ and 2 for the DZ twins to increase the environmental component. This leads to the estimates reported in Table 3. Here the C component is 0 even though the estimate also increases in a fashion similar to the situation in Table 2. The A component is highly sensitive to the censoring patterns and ranges from 99% to 0%. The C component ranges from 65% to 0% over the range of different censoring patterns. Note that the A component decreases on the diagonal with equal censoring in this case. The direction of the bias depends on the censoring pattern.

`Table 3 around here`

We conclude that the bias is substantial and conclusions based on the simple liability threshold model cannot be used for important conclusions about genetic and environmental effects.

7 Worked example: heritability of breast cancer

We apply the method to the population based cohort of Danish female twins born from 1870 with follow-up on death-status till January 1st 2009. Breast cancer occurrences were identified from the National cancer registry which began registration January 1st 1943. Of 31,212 female twins eligible for study we identified 1,100 breast cancer cases, 7,708 were dead at time of follow-up and 22,404 were censored. The number of twin pairs by status are listed in Table 4. We note that 72% of the twins were censored, and the censoring was 74% and 70% among MZ and DZ twins, respectively. Both twins in a pair had to be alive January 1st 1943 at which time cancer-registration starts. This lead to the left-truncation that we discussed how to deal with in a previous section.

`Table 4 around here`

Population based twin registries are very well suited for estimating magnitude and type of heritability, however, as stated above delayed entry and censoring is present to a very large and non-ignorable extent. Lichtenstein et al. (2000) found the estimated

heritability of breast cancer to be 27% (4%-41%) (95 % confidence interval) and similarly the shared environmental factor was 6% (0%-22%). This result is stated to lead to interpretation in a liability threshold model for the dichotomous outcome breast cancer status. It is important to note that the liability threshold model yields a very useful and simple interpretation, that can not be applied here, and we must thus interpret the model directly on the probability scale with the direct consequence that the dependence is reflected in the concordance probability. As pointed out these estimates are incorrect, because they are not corrected for censoring. The dependence measures for the MZ and DZ twins will be severely biased because of the censoring. The censoring is the same for both twins and thus perfectly correlated, thus considering the event cancer versus "death or censored" will thus be positive correlated for twins solely due to the censorings.

We first estimate the marginal cumulative incidence of breast cancer in women, taken the truncation into account. We did this assuming that monozygotic and dizygotic twins had the same cumulative incidence. We formally tested this by fitting a Fine-Gray regression model (Fine & Gray, 1999) with the result that the effect of zygosity (DZ) being -0.11 with standard error 0.07, thus clearly showing that there is no significant difference between monozygotic and dizygotic twins. The marginal cumulative incidence is given in Figure 1 panel (a) as the marginal estimate (dotted line). Approximately 12% of all women experience breast cancer over their lifetime.

`Figure 1 around here`

We now apply the approaches described in this paper. We estimated the cOR, the RR and fitted various random effects models to describe the dependence for monozygotic and dizygotic twins. The estimates from the different models are shown in Table 5. We note that the dependence is considerably stronger for the MZ twins than for the DZ twins. Under the random effects model the cross-OR should be the random effects variance plus one. Here we get a cOR at 2.95 and 1.50 for the MZ and DZ twins, respectively, thus giving a somewhat higher estimate than from the random effects model for the MZ twins, and an equivalent estimate for the DZ twins.

`Table 5 around here`

The estimates of association can all be translated into estimates of the concordance probabilty function, that is the probability that both twins have experienced cancer at a given time. We note that the cOR, RR and the random effects models lead to very similar estimates of the concordance probabilities. Using the random effects model the concordance probabilty is approximately 2.8% and 1.8% for MZ and DZ twins, respectively. Along these lines we note that the multilocus index (see Section 3.2) is estimated to 2.9 but is not significantly greater than two which would indicate genetic interaction effects for breast cancer risk.

Taking the left-truncation into account has a noticeable impact on the dependence estimates for the Danish twin registry. We expect that the effect of truncation is even more pronounced in the Swedish and Finnish twin registries. To deal with the left truncation we first directly model the time to cancer or death, to estimate the truncation probabilities. We here used a gamma-copula Clayton-Oakes model, Glidden (2000), that lead to dependence estimates (Gamma variances; numbers in parentheses are the

estimated standard errors) for monozygotic and dizygotic twins at 0.40 (0.05) and 0.18 (0.03), respectively. We used the `two.stage` function of the `timereg`-package for R to fit this model. Using this model we estimated the truncation probabilities that are needed in the estimating equations. Clearly, ignoring the truncation will be even more problematic when considering the dependence on the competing event (death prior to breast cancer). Table 5 gives the dependence estimates taking the truncation into account. As expected the dependence is estimated somewhat too high if the truncation is ignored. In Figure 1, we show the concordance probabilty estimates based on the random effects model ignoring the truncation and correcting for it, versus the marginal estimate that would be the strongest possible correlation, and under independence. As expected when correcting for the truncation we get a lower estimate of the concordance probabilities, that are now 2.2% and 1.4% for the MZ and DZ twins, respectively.

The conclusion from these initial analyses and the estimates of dependence on the probability scale is that monozygotic twins are much more strongly dependent than dizygotic twins. A formal test for equivalent variance yield that the dependence is significantly different ($p < 0.001$), see Scheike et al. (2010); Scheike & Sun (2011). This suggest a genetic component in breast cancer. We shall now try to quantify the size and type of the genetic components.

We also considered several structured random effects models with genetic and environmental variance components. All models involving both additive genetic, A, and environmental, C, components did not converge. In particular we found that the ACE model did not fit well. This is particularly so when we corrected for truncation as we should. We use the standard terminology and for example talk about the AE model, even though our E term is included in our link function implicitly due to the fixed marginals. Clearly, this avoids some complications and sensitivity to choice of link-function when estimating the E term that is just a matter of choosing link-function. We here thus only estimate correlation parameters, and later return to how to translate these numbers into heritability estimates.

`Table 6 around here`

We note that all methods lead to comparable conclusions. If we compare the difference random effects models, we see that the simple model with different random effects variances for MZ and DZ twins, is comparable to what is achieved by the AE model, even though the AE model overestimated the dependence for DZ and underestimated the MZ dependence. This is further improved for the DCE and DE model, where D is the dominant genetic component. The ADE model also fits quite well, and almost achieves the estimates from the unstructured model. Clearly, when comparing the derived dependence based on the different models, it is evident that the ADE, DCE and DE model all fits the data very well, thus indicating that the genetic component in part may be driven by dominant genetic effects. We stress that what model to use should also involve biological reasoning in addition to pure statistical fitting.

Correcting for the truncation makes all models with unique environmental factors inconsistent with the data. The DCE model corrected for truncation lead to a C estimate at $-0.02(0.61)$ in a model without positive constraints, thus leading to a C component

that is on the boundary if a positivity constraint is enforced. Similarly, the ACE model lead to a C estimate at $-0.30(0.39)$. Among the remaining models the DE model fits the data quite well and is almost equivalent with the simple unstructured model (UN). We note that the AE model has a rather poor fit, and is thus not consistent with the data.

In the next section we discuss one simple approach for translating the dependence estimates into heritability estimates. The random effects models considered here, clearly suggest that the genetic component is present and must be considerably larger than the shared environmental effects. The shared environmental effects was estimated to 0 for most considered models, and without positive constraints we got negative estimates for the C component of the ACE and DCE models. The different genetic models can be compared in terms of the concordance probabilty function for the MZ and DZ twins, that in essence reflects the size the shared variance.

7.1 Heritability

We argued that the ACE or AE model lead to a heritability

$$h^2(t) = 2\frac{\mathcal{C}_{\mathrm{MZ}}(t) - \mathcal{C}_{\mathrm{DZ}}(t)}{F_1(t)(1 - F_1(t))}$$

where $F_1(t)$ is the marginal cumulative incidence and $\mathcal{C}(t)$ is the cumulative concordance probability (see Section 3). Note, that we here use that the marginals are same for MZ and DZ twins.

First, we aim to estimate a lifetime heritability where we estimate the heritability at age 100 years, denoted as $h^2(100)$, where the concordance probability estimates are based solely on the life-time dependence that is estimated from considering the lifetime occurrences only. This leads to heritability at 0.25 using the RR model. Note that this estimate could be obtained using the Probit model and inverse probabilty of censoring weighting and the derived concordance probabilty estimates from this model. We could also estimate this heritability using the assumption that RR(t) is constant over time and then we get 0.23. The estimate is 0.22 when the truncation is taken into account. Using some of the other models make the estimate change slightly. The lifetime heritability is thus in the neighborhood of 0.22. Similarly, the lifetime multilocus estimate is 3.9 thus suggesting dominant genetic effects, i.e. gene-gene interactions.

Rather than settling for a cumulative lifetime heritability we can also estimate the cumulative heritability for different ages. To do this it is important to model the dependence structure and as a consequence the concordance probabilty functions in a flexible manner that can be faithful to the data. We first considered a fully non-parametric concordance probability estimator based on inverse-probability censoring weighting. This curve is shown in Figure 2 (solid line). A compromise between a fully non-parametric estimator and a structured model is to assume that for example $RR(t) = RR(t, \theta)$ depends on time and model parameters θ and is piecewise constant or linear in time for both MZ and DZ twins. We estimated the heritability assuming that $\log(RR(t))$ was

linear leading to the dashed line in Figure 2. We also estimated the cumulative heritability using two piecewise constant models for $RR(t)$, and these are depicted with the dotted and dash-dotted lines in Figure 2.

`Figure 2 around here`

The non-parametric estimate could be given confidence bands, but it is clear that the variability is very high. A trend test using a log-linear model for $RR(t)$ results in a negative trend for both MZ and DZ twins, and an overall negative trend for the heritability, but the trend is not significant for neither MZ nor DZ twins with $p = 0.15$ and $p = 0.25$, respectively.

We note that the heritability has a vaguely decreasing trend over time, which is consistent with the hypothesis that the early breast cancers in particular are due to heritable factors. All models yield somewhat similar estimates and it is obvious that the assumed structure for the dependence in the $RR(t)$ model has a considerable impact on the heritability estimates. The cOR model leads to a very similar conclusion.

An alternative characterization of the difference between the dependence among MZ and DZ twins is to consider the multilocus index that we introduced earlier. A time-dependent version of it, can be estimated assuming that the $RR(t)$ is piecewise constant as we also did for the heritability modeling above. This leads to the estimates in Table 7.

`Table 7 around here`

We note that the dependence is stronger in the early time-period and seems to decrease over time. This is consistent with the expected stronger dependence among early cancers. The multilocus index suggests that there may be some indication of multiplicative effects, but the index does not differ significantly from 2 in any of the time-intervals.

8 Discussion

We have proposed a framework for estimation of dependence for specific diseases in the presence of competing risks, right censoring and truncation. The dependence measures considered here are based on the cumulative incidence function thus extending the classical Probit random effects model to a more refined framework. In addition we have shown how to estimate these dependence measures in the presence of right censorings. Our simulation set-up revealed that the bias from ignoring the censorings is dramatic and depends on the specific censoring patterns among DZ and MZ twins. The direct modeling approaches in terms of the cross-odds ratios or relative risk measures lead to straightforward interpretations of the dependence. We further separated these measures into heritable factors and shared environmental factors using specific random effects models, such as the ACE model. This leads to a useful separation of the genetic and environmental factors.

Our random effects models relied on Gamma distributions for computational convenience, but the choice of the random effects distribution will lead to different types of dependence, see Hougaard (2000). We compared our concordance estimates of different

models and concluded based on this that the Gamma distribution was not seriously violated. Further, research could develop goodness of fit methods and study the different types of dependence induced by the random effects.

A technical aspect was that we showed how to correct for truncation in assessing the dependence. This required additional modeling, and we here used a truncation weight given from some other frailty model.

To compute a heritability measure we tried to separate the variation on the scale of the observed data, thus bringing in the variation due to the probability scale as well. This had the advantage that the otherwise somewhat artificial E component of non-shared random effects was not modeled directly in the random effects models, thus avoiding sensitivity to the choice of the link-function. This separation relied on directly modeling the correlation of the observed binary responses through ACE or DCE type models. We stress that the choice of genetic model should be considered carefully in a biological context. Further as expected the dependence modeling had great impact on the time-dependent heritability estimates obtained from these models. We note that the extension of the classical liability threshold (or probit) modeling to account for censored paired observations by IPCW techniques as devised could be a useful way to correct for bias in many studies of dichotomous traits in twins having time to event available. Note, however, that our heritability measure is different from the heritability on the liability scale. We will explore models that are aimed more directly at estimating heritability in future work. Another topic for further research that we are considering in ongoing work is how to use a full maximum likelihood approach to increase the efficiency.

We have implemented the methods in the R-package **mets** that is available at CRAN.

Acknowledgment

We appreciate constructive and useful comments from the Associate editor and two referees that have improved the presentation of our paper considerably and have raised several interesting issues. We thank our collaborators at the NorTwinCan consortia for stimulating discussions about the twin data.

References

Bandeen-Roche, K. & Liang, K.-Y. (2002). Modelling multivariate failure time associations in the presence of a competing risk. Biometrika **89**, 299–314.

Bandeen-Roche, K. & Ning, J. (2008). Nonparametric estimation of bivaiate failure time associations in the presence of a competing risk. Biometrika **95**, 221–232.

Chen, B. E., Kramer, J. L., Greene, M. H., & Rosenberg, P. S. (2008). Competing Risks Analysis of Correlated Failure Time Data. Biometrics **64**, 172–179.

Cheng, Y., Fine, J. P., & Kosorok, M. R. (2007). Nonparametric Association Analysis of Bivariate Competing-Risks Data. J. Amer. Statist. Assoc. **102**, 1407–1415.

Falconer, D. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Annals of human genetics **31**, 1–20.

Falconer, D. & Mackay, T. (1994). Introduction to quantitative genetics. Addison-Wesley.

Geskus, R. (2010). Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. Biometrics **67**, 39–49.

Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. J. Amer. Statist. Assoc. **94**, 496–509.

Glidden, D. V. (2000). A two-stage estimator of the dependence parameter for the Clayton-Oakes model. Lifetime Data Anal. **6**, 141–156.

Gorfine, M. & Hsu, L. (2011). Frailty-based competing risks model for multivariate survival data. Biometrics **67**, 415–426.

Hougaard, P. (2000). Analysis of multivariate survival data. Statistics for Biology and Health. Springer-Verlag, New York.

Katsahian, S., Resche-Rigon, M., Chevret, S., & Porcher, R. (2006). Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. Statist. Med. **25**, 4267–4278.

Lichtenstein, P., Holm, N., Verkasalo, P., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., & Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. New England Journal of Medicine **343**, 78.

Locatelli, I., Lichtenstein, P., & Yashin, A. (2004). The heritability of breast cancer: A Bayesian correlated frailty model applied to Swedish twins data. Twin Research and Human Genetics **7**, 182–191.

Locatelli, I., Rosina, A., Lichtenstein, P., & Yashin, A. (2007). A correlated frailty model with long-term survivors for estimating the heritability of breast cancer. Statistics in medicine **26**, 3722–3734.

Neale, M. C. & Cardon, L. R. (1992). Methodology for genetic studies of twins and families. Dortrecht: Kluwer Academic Publishers.

Pinborg, A. (2005). IVF/ICSI twin pregnancies: risks and prevention. Human reproduction update **11**, 575–593.

Ripatti, S., Gatz, M., Pedersen, N., & Palmgren, J. (2003). Three-state frailty model for age at onset of dementia and death in Swedish twins. Genetic epidemiology **24**, 139–149.

Risch, N. (1990). Linkage strategies for genetically complex traits. i. multilocus models. American Journal of Human Genetics **46**, 222.

Scheike, T. H. & Sun, Y. (2012). On cross-odds ratio for multivariate competing risks data. Biostatistics **13**, 680–694.

Scheike, T. H., Sun, Y., Zhang, M. J., & Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. Biometrika **97**, 133–145.

Scheike, T. H., Zhang, M.-J., & Gerds, T. (2008). Predicting cumulative incidence probability by direct binomial regression. Biometrika **95**, 205–20.

Skytthe, A., Pedersen, N., Kaprio, J., Stazi, M., Hjelmborg, J., Iachine, I., Vaupel, J., & Christensen, K. (2003). Longevity studies in GenomEUtwin. Twin. Res. **6**, 448–454.

Sham, P. (1998). Statistics in Human Genetics. Arnold Appl. of Statistics.

Zhang, X., Zhang, M., & Fine, J. (2011). A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. Statistics in Medicine **30**, 1933–51.

	Lifetime prevalence				**Lifetime concordance**			
MZ \ DZ	0%	28%	55%	82%	0%	28%	55%	82%
0%	0.120	0.101	0.082	0.064	0.034	0.027	0.021	0.018
25%	0.103	0.083	0.063	0.044	0.032	0.024	0.016	0.013
55%	0.086	0.065	0.044	0.025	0.033	0.023	0.013	0.007
82%	0.078	0.054	0.029	0.006	0.069	0.045	0.021	0.002

Table 1: Lifetime disease prevalence and lifetime concordance for MZ twins estimated using Liability threshold model ignoring censoring for different censorings patterns of MZ and DZ twins. Cross odds-ratio dependence parameter 3 and 2 for MZ and DZ, respectively.

	A				**C**			
MZ \ DZ	0%	27%	55%	82%	0%	28%	55%	82%
0%	0.374	0.221	0.000	0.000	0.000	0.170	0.431	0.528
21%	0.473	0.461	0.046	0.000	0.000	0.026	0.429	0.581
55%	0.627	0.624	0.431	0.000	0.000	0.000	0.167	0.640
82%	0.988	0.979	0.949	0.257	0.000	0.000	0.000	0.494

Table 2: Estimates of variance for genetic (A) and environmental components (C) of threshold liabilty model given different censoring patterns for MZ and DZ twins. Cross odds-ratio dependence parameter 3 and 1.5 for MZ and DZ, respectively.

	A				**C**			
MZ \ DZ	0%	28%	55%	82%	0%	28%	55%	82%
0%	0.316	0.024	0.000	0.000	0.079	0.367	0.454	0.530
21%	0.513	0.255	0.000	0.000	0.000	0.229	0.496	0.585
55%	0.676	0.633	0.253	0.000	0.000	0.019	0.341	0.646
82%	0.990	0.981	0.954	0.179	0.000	0.000	0.000	0.578

Table 3: Estimates of variance for genetic (A) and environmental components (C) of threshold liabilty model given different censoring patterns for MZ and DZ twins. Cross odds-ratio dependence parameter 3 and 2 for MZ and DZ, respectively.

Number of pairs at time of follow-up			
MZ & DZ Status	Breast cancer	No cancer and dead	No cancer and alive
Breast cancer	41 & 44	325	289
No cancer and dead	164	1005 & 1876	932
No cancer and alive	138	446	3967 & 6092

Table 4: Number of pairs by status at time of follow-up with MZ pairs in lower left triangle and DZ pairs in upper. Censorings: 72% of a total of 30,638 twins are alive without (diagnosed) breast cancer at time of follow-up.

	log-cOR	log-RR	log-RR trunc.	random effect	random effect trunc.
MZ	1.01 (0.17)	0.90 (0.16)	0.78 (0.16)	1.48 (0.47)	1.04 (0.40)
DZ	0.40 (0.16)	0.39 (0.16)	0.25 (0.16)	0.48 (0.25)	0.21 (0.21)

Table 5: Within pair association in the occurrence of breast cancer under models for cross-odds ratio, relative recurrence risk and random effects models against zygosity (numbers in parentheses are estimated standard errors). For some models we corrected for truncation.

	Ignoring truncation				
	ADE	DCE	DE	AE	UN
D	0.72 (0.80)	1.20 (0.16)	1.52 (0.47)		
C		0.23 (0.50)			
A	0.72 (0.80)			1.32 (0.39)	
MZ	1.44	1.43	1.52	1.32	1.48 (0.47)
DZ	0.54	0.54	0.38	0.66	0.48 (0.25)
	Correcting for truncation				
	ADE	DCE	DE	AE	UN
D	1.12 (1.11)	1.10 (0.25)	1.06 (0.39)		
C		-0.02 (0.61)			
A	-0.05 (2.91)			0.89 (0.31)	
MZ	1.07	1.08	1.06	0.89	1.04 (0.40)
DZ	0.25	0.26	0.27	0.47	0.21 (0.21)

Table 6: Variances of random effects for different genetic models (numbers in parentheses are estimated standard errors), and derived variance for MZ and DZ twins. UN is simple unstructured random effects model.

Age	DZ: log-RR (SE)	MZ: log-RR (SE)	Multilocus (SE)
0-70	0.944 (0.293)	1.630 (0.260)	2.62 (1.51)
70-80	0.503 (0.219)	1.220 (0.201)	3.63 (2.26)
80-90	0.354 (0.167)	0.838 (0.168)	3.08 (1.95)
90-	0.339 (0.156)	0.675 (0.162)	2.39 (1.53)

Table 7: Piece-wise constant $RR(t)$ for MZ and DZ twins, and resulting multi-locus index (numbers in parentheses are estimated standard errors).

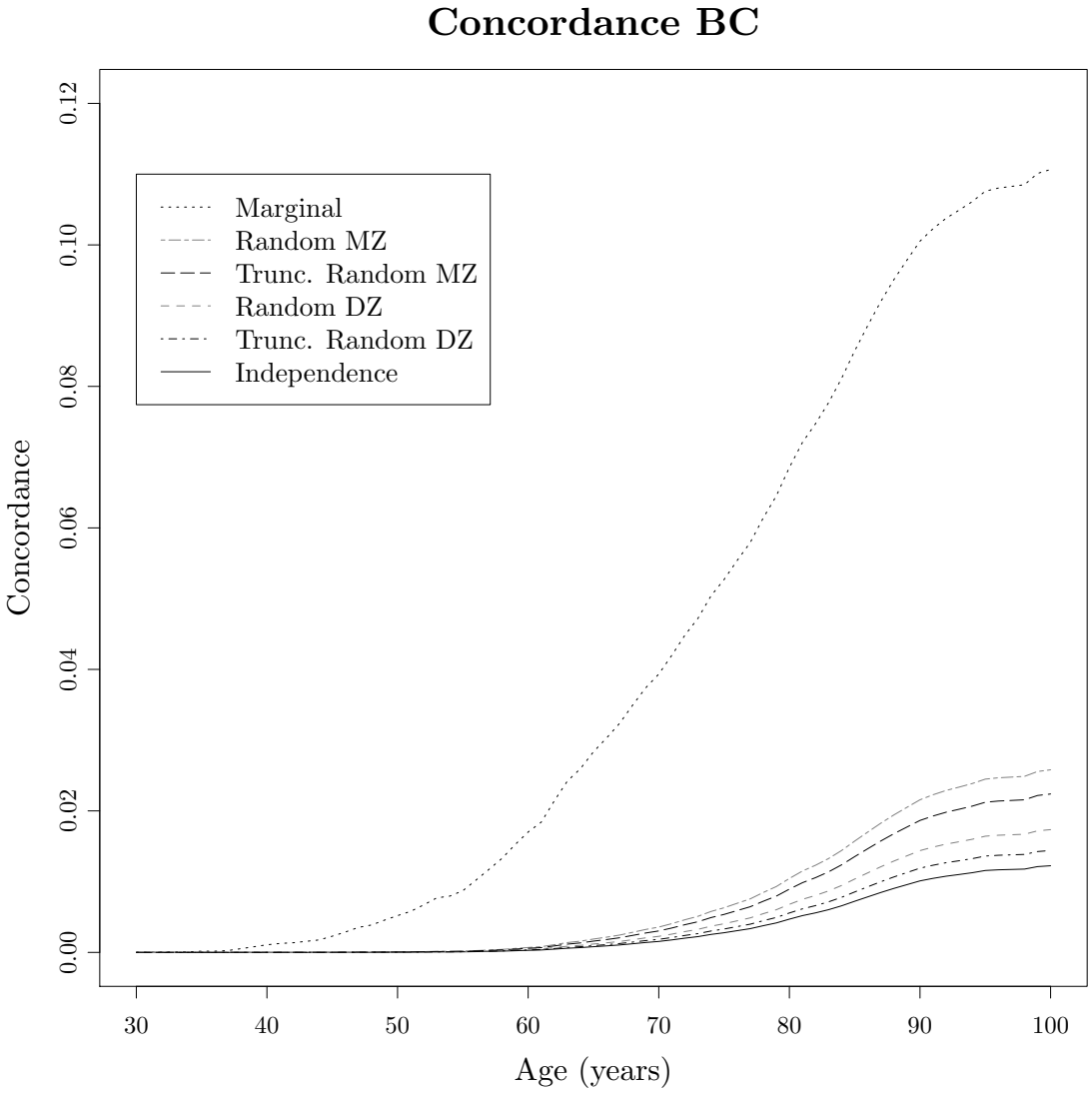

Figure 1: Estimated concordance probability function for breast cancer for random effects model ignoring truncation and correcting for it.

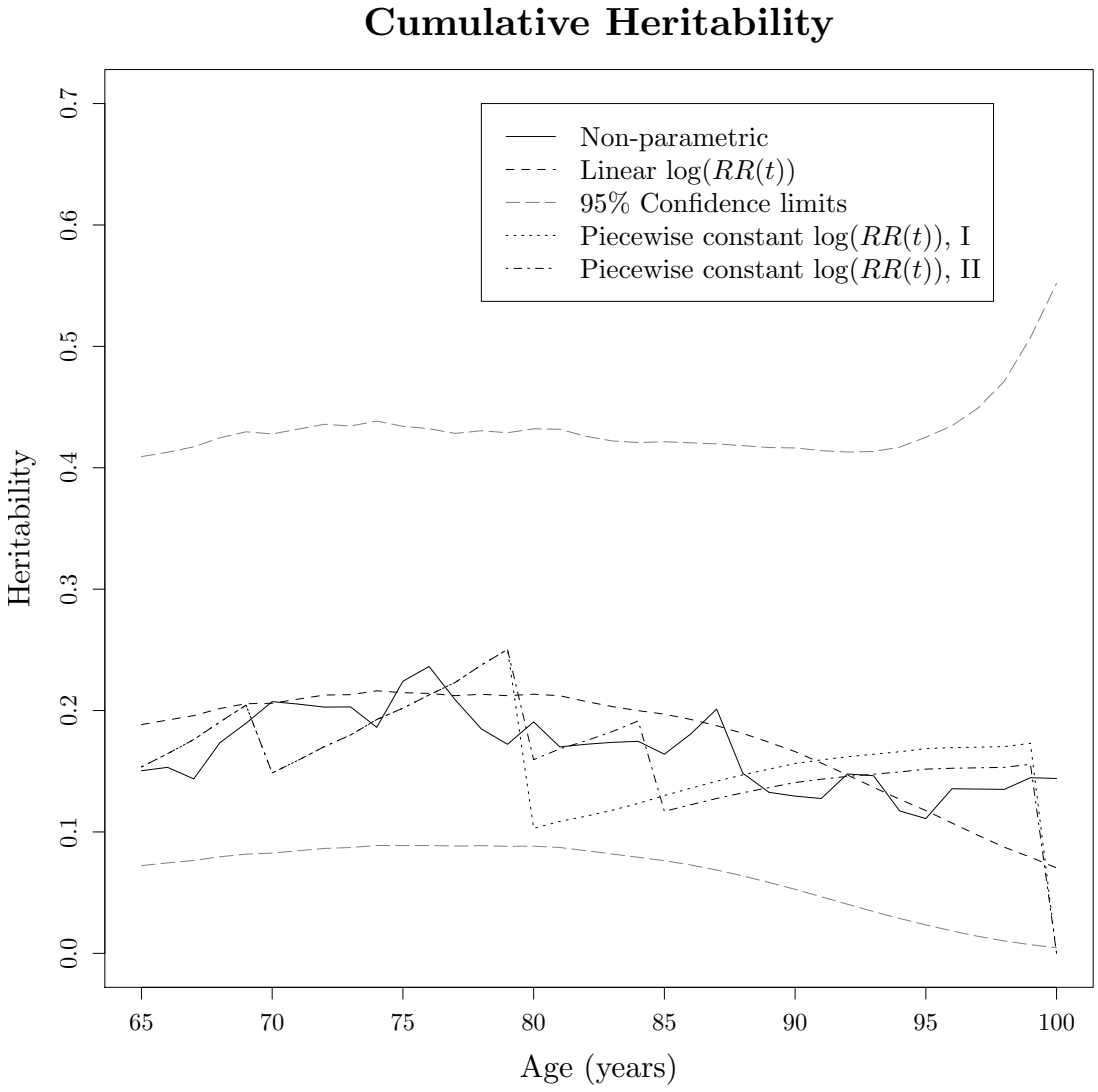

Figure 2: Heritability estimates for breast cancer based with pointwise 95% confidence limits for the linear log(RR(t)) model.