ANNUAL REVIEWS

Annual Review of Statistics and Its Application

A Practical Guide to Family Studies with Lifetime Data

Thomas H. Scheike[1] and Klaus Kähler Holst[2]

[1]Division of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark, DK-1014; email: ts@biostat.ku.dk

[2]Maersk, Global Data Analytics, Copenhagen, Denmark, DK-1098

Annu. Rev. Stat. Appl. 2022. 9:47–69

First published as a Review in Advance on November 10, 2021

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040120-024253

Keywords

age of onset, casewise concordance, censoring, competing risks, concordance, dependence, polygenic modeling, time to event

Abstract

Familial aggregation refers to the fact that a particular disease may be overrepresented in some families due to genetic or environmental factors. When studying such phenomena, it is clear that one important aspect is the age of onset of the disease in question, and in addition, the data will typically be right-censored. Therefore, one must apply lifetime data methods to quantify such dependence and to separate it into different sources using polygenic modeling. Another important point is that the occurrence of a particular disease can be prevented by death—that is, competing risks—and therefore, the familial aggregation should be studied in a model that allows for both death and the occurrence of the disease. We here demonstrate how polygenic modeling can be done for both survival data and competing risks data dealing with right-censoring. The competing risks modeling that we focus on is closely related to the liability threshold model.

1. INTRODUCTION

In statistical genetics, it is common to compare traits within families (for example, twins) to study the resemblance with respect to a particular trait. When the trait is the time to an event (for example, age of onset of a particular disease or age of death), one typically needs to deal with right-censored outcomes, the situation in which not all individuals experience the event of interest by the end of follow-up. Failure to do so will bias estimates of familial resemblance. Survival analysis methods have been developed to deal with right-censoring, so that under independence between censoring and the event of interest, population quantities are consistently estimated. In this review, we describe how to apply survival analysis methods when the aim is to describe familial resemblance.

The survival model (**Figure 1*a***) describes the situation where subjects will transition from alive to dead at some point in time t. This transition is often modeled and described by the hazard rate $\alpha(t)$, which gives the instantaneous risk of dying for a subject who is at risk at time t. Alternatively, as shown in **Figure 1*b***, subjects may experience one of several events (die from one of several diseases) or experience cancer or die, as described by the competing risks model. In this context, modeling is done using either the cause-specific hazards $\alpha_1(t)$ and $\alpha_d(t)$ of the event of interest or death, or the cumulative probabilities of seeing either of the events, $F_1(t)$ and $F_d(t)$, as a function of time. The cause-specific hazards are defined as the instantaneous risks of experiencing event 1 or dying among those that are still alive and have not experienced event 1, respectively. When interest is on the possible relationship between multiple types of cancer, such as ovarian and breast cancer, one could consider the competing risks model with three competing risks (**Figure 1*c***), since in addition to the two causes of interest, we will also have death as a competing risk. This extended model would also require modeling of either the cause-specific hazards or the cumulative incidence probabilities of the causes described in the model. The two main classes of models used to describe possible dependence in multivariate event history data are random effects models, also called frailty models, and copula models, where one would more generally describe dependence parameters for a multivariate distribution. Both these approaches are described in several excellent books (Hougaard 2000, Wienke 2011, Duchateau & Janssen 2007). In addition, one may also use standard conditional hazard models (Andersen et al. 1993), which can give a simple and useful

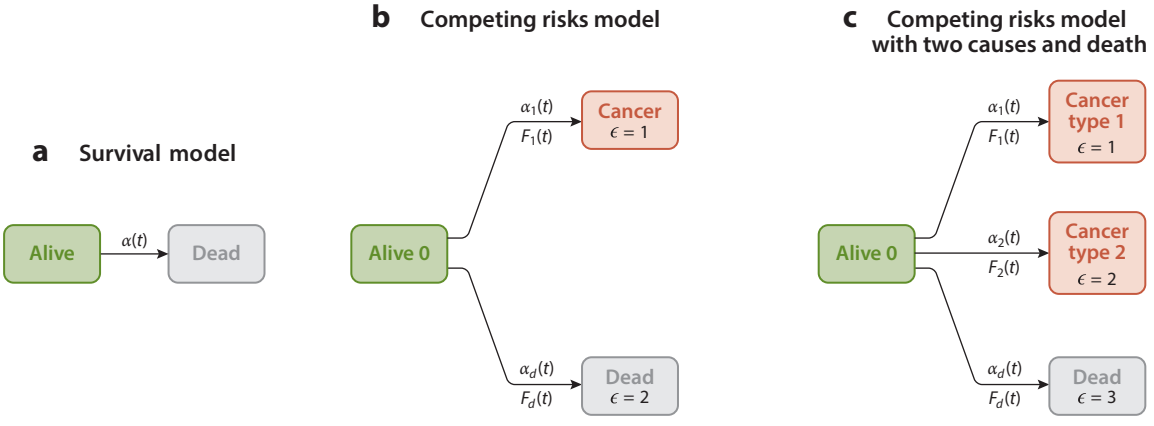

Figure 1

Survival model and competing risks models.

description of how events for one family member change our prediction of the risk for another family member.

Random effects models make it possible to decompose the dependence in a family into, for example, environmental and genetic sources using polygenic modeling—that is, models with several additive random effects that account for these different sources (see Falconer 1967, Neale & Cardon 1992, Falconer & Mackay 1994, Lynch & Walsh 1998, Sham 1998). There is also a long tradition of population genetics in the animal breeding field (Hill 2014), where heritability of longevity and age of onset has been examined in a survival analysis setting (see, for example, Ducrocq & Casella 1996, Yazdi et al. 2002). Below, we illustrate how to use such models in the twin setting.

Sørensen et al. (2013) considered the age of menarche based on data from the Danish Twin Registry consisting of monozygotic (MZ) and dizygotic (DZ) female twin pairs to learn about a possible genetic component in the timing of menarche. The time of menarche was described by a survival model, since death prior to menarche is extremely rare. The authors compared association in the MZ and DZ twins using a polygenic model for the dependence, which leads to a decomposition of the correlation into genetic and environmental effects. The sizes of these effects were partly summarized by reporting a heritability of 61% [95% confidence interval (CI): 0.38–0.84]—that is, the part of the variation that is due to genes. We later consider the age of menopause for Danish twins and do a similar decomposition as well as describing the dependence in MZ and DZ twins.

Another equally important problem is to study heritable factors in the causation of cancer, as illustrated by Lichtenstein et al. (2000), Hjelmborg et al. (2014), Möller et al. (2016), and Mucci et al. (2016), where the interest is in studying the resemblance in the occurrence of a particular type of cancer. This problem is different from the study of dependence in the timing of menarche, in that the occurrence of cancer may be prevented by death and therefore needs to be studied in the competing risks context (see **Figure 1*b***), where causes compete for the first event. When the interest is in describing correlation or cooccurrence between family members for one of the causes, then one is also forced to consider the other cause as well due to a possible cross-dependence—for example, if the death of one family member before cancer is predictive of the risk of cancer for the other family member. In addition, there might also be interest in possible dependence between multiple competing events such as breast cancer and ovarian cancer, and that death is a competing risk that needs to be taken into account (see **Figure 1*c***). We illustrate the methods for competing risks data considering the age of prostate cancer in a twin study that resembles that of Hjelmborg et al. (2014), which found a heritability of 58% (95% CI: 0.52–0.63).

1.1. Data and Notation

We consider $k = 1, \ldots, K$ independent clusters with $i = 1, \ldots, n_k$ subjects within each cluster. For each cluster, we are given a set of independent random effects $V_k^{\mathrm{T}} = (V_{k1}, \ldots, V_{km})$. The dependence structure between the subjects in a cluster is now described by design vectors $Q_{k1}, \ldots, Q_{kn_k}$, such that the ith subject has random effect $Q_{ki}^{\mathrm{T}} V_k$. $Q_{ki} = (Q_{ki1}, \ldots, Q_{kim})^{\mathrm{T}}$ is often specified as a vector of ones and zeroes that makes relevant parts of V_k active and possibly shared for each subject.

In addition to survival times, T_{ki}, and the possible cause indicators for the competing risks data $\epsilon_{ik} \in \{1, 2, 3\}$, which are not needed when we consider survival data, we assume that we have independent right-censoring times, U_{ki}, such that given V_k and the covariates Q_{ki} and X_{ki}, $i = 1, \ldots, n_k$, $(U_{k1}, \ldots, U_{kn_k})$ are conditionally independent of $(T_{k1}, \ldots, T_{kn_k}, \epsilon_{k1}, \ldots, \epsilon_{kn_k})$, and the conditional censoring distribution does not depend on V_k. We thus observe $\tilde{T}_{ki} = T_{ki} \wedge U_{ki}$, the event indicator $\delta_{ki} = I(T_{ki} \leq U_{ki})$, and a possible cause of death $\tilde{\epsilon}_{ki} = \delta_{ki}\epsilon_{ki}$. The censoring

distribution conditional on covariates, X, is denoted by $G_c(\cdot|X)$. We can also express the survival times via counting processes $N_{ki}(t) = I(T_{ki} \leq t, T_{ki} \leq U_{ki})$ with at risk indicators $Y_{ki}(t) = I(T_{ki} \geq t, U_{ki} \geq t)$. For a p-dimensional vector x, let $x^{\otimes 2} = xx^{\mathrm{T}}$.

Liability threshold model: a model where a binary (or ordinal) trait, Y, is related to a latent continuous variable Y^* (the liability) through the relation $Y = I(Y^* > \kappa)$ for some threshold value $\kappa > 0$

Broad-sense heritability: the proportion of phenotypic variation described by genetic factors

1.2. Polygenic Modeling

To briefly present the random effects models that are often used to decompose the dependence into different sources, we consider the analysis in the combined Nordic study of the Danish, Finnish, and Swedish twin registries (Lichtenstein et al. 2000, Hjelmborg et al. 2014, Möller et al. 2016, Mucci et al. 2016) that aimed to model the dependence in twin pairs in the occurrence of breast cancer. There is a large literature on polygenic modeling that aims to decompose the correlation into different sources of variation using, for example, the liability threshold model (see, for example, Falconer 1967, Neale & Cardon 1992, Falconer & Mackay 1994, Lynch & Walsh 1998, Sham 1998).

Here, the event of interest is $Y_{ki}(\tau) = I(T_{ki} \leq \tau, \epsilon_{ki} = 1)$, i.e., the outcome that the subject gets cancer before τ and before dying, obviously. In a general family setting, we assume that given covariates, X_{ki}, and a set of random effects, V_{ki}, for each family member $i = 1, \ldots, n_k$, the cancer occurrences among family members are independent and follow a probit model:

$$\text{probit}(\mathrm{P}(Y_{ki}(\tau) = 1|X_{ki}, V_{ki})) = X_{ki}^{\mathrm{T}}\beta + V_{ki}, \quad i = 1, \ldots, n_k, k = 1, \ldots, K. \tag{1.}$$

Note that we here assume that the survival time is fully observed to discuss how a population model might be structured and formulated, and we return to an inverse probability of censoring weighting (IPCW) adjustment for the right-censoring later.

A properly designed family study may make it possible to decompose the random effect variance into genetic and environmental components, $V_{ki} = V_{\text{gene},\,ki} + V_{\text{env},\,ki}$. Thus, assuming independence between genetic and environmental effects makes it possible to quantify the heritability as the fraction of the total variance due to genetic factors. The probit model, Equation 1, is equivalent to a latent variable model formulation where the binary outcome is defined from a conditionally normally distributed latent variable,

$$Y_{ki}^* = X_{ki}^{\mathrm{T}}\beta + V_{\text{gene},ki} + V_{\text{env},ki} + V_{\mathrm{E},ki},$$

such that $Y_{ki}(\tau) = I(Y_{ki}^* > \kappa)$. For identification, the threshold is fixed at $\kappa = 0$, and $V_{\mathrm{E},k1}, \ldots, V_{\mathrm{E},kn_k}$ are independent standard normally distributed. Under an assumption of no gene-environment interaction, this model then leads to a broad-sense heritability, conditional on covariates, defined by

$$H^2 = \frac{\mathrm{Var}(V_{\text{gene},ki})}{\mathrm{Var}(V_{\text{gene},ki}) + \mathrm{Var}(V_{\text{env},ki}) + \mathrm{Var}(V_{\mathrm{E},ki})}.$$

Under additional assumptions of parents not transmitting their environmental effects to their offspring, random mating (no inbreeding), linkage equilibrium, and no epistasis, the genetic variation can be further decomposed into additive and dominant genetic components (see Lange 2002):

$$\mathrm{Var}(V_{\text{gene},ki}) = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{D}}^2, \tag{2.}$$

where σ_{A}^2 is the variance of additive genetic effects and σ_{D}^2 is the variance of the dominant genetic effects. Furthermore, the genetic correlation can, in this case, be determined from the familial

Table 1 Kinship and fraternity coefficients for different family members

	Δ_7	Φ
Parent-offspring	0	$\frac{1}{4}$
Grandparent-grandchild	0	$\frac{1}{8}$
Great grandparent-great grandchild	0	$\frac{1}{16}$
Half siblings	0	$\frac{1}{8}$
Full siblings, DZ twins	$\frac{1}{4}$	$\frac{1}{4}$
MZ twins	1	$\frac{1}{2}$
Uncle/aunt-nephew/niece	0	$\frac{1}{8}$
First cousins	0	$\frac{1}{16}$
Double first cousins	$\frac{1}{16}$	$\frac{1}{8}$
Second cousins	0	$\frac{1}{64}$

Δ_7 is the fraternity coefficient, which describes the probability that, at a given locus, both alleles for the two relatives are identical by descent. Φ is the kinship coefficient, which is the probability that two randomly selected alleles from the same locus of relatives are identical by descent.

resemblance such that for relatives i and j the genetic covariance structure is defined by

$$\mathrm{Cov}(V_{\mathrm{gene},ki}, V_{\mathrm{gene},kj}) = 2\Phi_{ij}\sigma^2_{\mathrm{A}} + \Delta_{7ij}\sigma^2_{\mathrm{D}}, \qquad 3.$$

where, as in Lynch & Walsh (1998) and Lange (2002), Φ_{ij} is the kinship coefficient, which is the probability that two randomly selected alleles from the same locus of relatives i and j are identical by descent, and the fraternity coefficient Δ_{7ij} describes the probability that at a given locus both alleles for the two relatives are identical by descent (see **Table 1**). It is through this dependence structure that the variance components can be identified by the inclusion of different types of relatives in the study, such as MZ and DZ twins in the classical twin study design. The model with both additive (A) and dominant (D) genetic effects as well as shared (C) and individual (E) environmental effects is typically denoted the ACDE model. For identification reasons, subsets—e.g. ADE, AE, or ACE models—are typically considered.

For the survival data, it can be even harder to find an appropriate scale on which to report a heritability, and we therefore suggest focusing more on the variances of the shared random effects and, in addition, trying to compute other more easily interpretable summary measures to describe the dependence between family members.

2. SURVIVAL DATA

When we observe multivariate survival data from a family or, equivalently, multivariate data from the competing risks model where the competing risks can be assumed independent, the aim is often to describe the strength of dependence between different family members. The standard tools are to use either random effects or copula models where the multivariate survival distribution is given via the marginal survival distributions and a copula (see, for example, Hougaard 2000, Duchateau & Janssen 2007, Wienke 2011). When random effects models are used for modeling dependence in a family, or in a pair, one can further apply the structured polygenic random effects models that makes assumptions about how the sources of dependence can be split up into genetic and environmental components.

Structured polygenic modeling will typically describe the multivariate distribution of the whole family. One could also model and describe only pairwise dependence for specific pairs in a family.

As an alternative to the random effects or copula modeling, one can specify directly conditional models for how the information about one subject alters the prediction of the instantaneous risk for another using conditional intensity models as described by Andersen et al. (1993). Note that random effects models lead to specific conditional models, but it is often useful to specify other, simpler intensity models that are easier to fit. We return to this point in Section 2.3.

2.1. Pairwise Dependence Modeling

A simple and useful first attempt to learn about genetic dependencies represented by the degree of familial aggregation or dependence is to consider pairwise modeling of the dependence, as is often done. One might describe the dependence between MZ and DZ twins or, in the broader context of family studies, look at dependence between siblings, or parents and offspring, or in general for any two subjects in a family. In a survival context, a simple approach is to use, for example, the Clayton-Oakes model (Clayton 1978, Oakes 1982, Glidden 2000), thus assuming that we have marginals for the members of our clusters on, for example, Cox form, such that marginally given covariates X_{ki} and T_{ki} follow a Cox model:

$$\lambda^m_{ki}(t; X_{ki}) = \lambda_0(t) \exp(X_{ki}^{\mathrm{T}} \beta). \tag{4.}$$

We let the related marginal survival function be denoted as $S^m_{ki}(t) = S^m(t|X_{ki})$. Clearly, other marginals may be used and the simple Cox model may be extended in various ways.

Then, for two such members of a family, the Clayton-Oakes copula model specifies that the bivariate survival distribution is given as

$$S_{k,ij}(t,s; X_{ki}, X_{kj}) = \Psi(\nu(i,j), \nu(i,j), \Psi^{-1}(S^m_{ki}(t)) + \Psi^{-1}(S^m_{ki}(s))), \tag{5.}$$

where $\Psi(\nu, \nu, \cdot)$ denotes the Laplace transform of the Gamma distribution $\Gamma(\nu, \nu)$ with mean 1 and variance ν^{-1}, Ψ^{-1} is the inverse Ψ, and by $\nu(i,j)$ we indicate that the dependence parameter depends on the considered type of pair, such as siblings, half-siblings, or parents-offspring.

The dependence parameters are easy to estimate using two-stage fitting, where the marginals are fitted in the first step and then used in pseudolikelihood to estimate $\nu(i,j)$ (as in Glidden 2000, Glidden & Self 1999, Shih & Louis 1995).

One useful simple extension of this model is to allow a regression structure on the dependence parameters, for example, for MZ and DZ twins. We thus suggest modeling $\nu(i,j) = Z(i,j)\theta$, where $Z(i,j)$ is a regression design depending on the considered pair.

Clearly, it can be of interest to use other copula models, and it is useful to investigate the goodness of fit for the chosen pairwise dependence model. For the Clayton-Oakes model, there are several suggestions for extensions of the Gamma-frailty model to allow dependence parameters locally in time (Glidden 1999; see also Nan et al. 2006, Shih & Albert 2010, Hu et al. 2011, Scheike et al. 2015b).

2.2. Hazard Random Effects Modeling

When the interest is in separating the dependencies into different sources using polygenic models, the additional structure from random effects models is needed. In the context of survival data, the natural starting point is therefore to consider polygenic random effects models in this context. We describe such models and present in further detail one class of models that has certain computational advantages.

Fitting polygenic models for survival data can be done by using random effects models, and there are several useful ones. One class of such models are the additive Gamma random effects

models (Petersen et al. 1996, Korsgaard & Andersen 1998, Petersen 1998, Li 1999), where, conditional on covariates and the random effects, the hazard for subject i in cluster k is on the form

$$\left(\sum_{l=1}^{m} Q_{kil} V_{kl}\right) \lambda_0(t) \exp(X_{ki}^{\mathrm{T}} \beta), \qquad 6.$$

with regression effects β, independent Gamma-distributed random effects $V_k^{\mathrm{T}} = (V_{k1}, \ldots, V_{km})$, 1/0 random effects design $Q_{ki}^{\mathrm{T}} = (Q_{ki1}, \ldots, Q_{kim})$, and baseline hazard $\lambda_0(t)$. The estimation for these models is not simple, even though it is possible to use the expectation–maximization algorithm (Klein 1992, Nielsen et al. 1992), and there exists no available software that can fit these models for large registry data, to the best of our knowledge.

An alternative formulation has been considered by Ripatti & Palmgren (2000):

$$\exp\left(Q_{ki}^{\mathrm{T}} V_k\right) \lambda_0(t) \exp(X_{ki}^{\mathrm{T}} \beta), \qquad 7.$$

where it is assumed that V_k follows a multivariate normal distribution. This makes the random effects very flexible, and this can sometimes be a big advantage, in particular when there is negative correlation between random effects. The models in Expression 7, and in Expression 6 with only one random effect, have been implemented in various useful R packages [for example, `coxme`, `phmm`, and `frailtyEM` (Therneau 2020, Donohue & Xu 2019, Vaida & Xu 2000, Balan & Putter 2019)], but again is limited to at most medium-sized data. In particular, Expression 7 is slow to fit because one needs to do numerical integration to approximate the likelihood.

2.2.1. Additive Gamma two-stage hazard modeling. We now present a version of the two-stage model, the copula model (Glidden 2000, Glidden & Self 1999, Shih & Louis 1995), which is easier to use and here is extended to the structured random effects setting. This model has marginals that follow a Cox model and then uses a copula to specify the joint distribution. The model is easier to fit because the semiparametric marginals are fitted first using marginal modeling (Spiekerman & Lin 1998), and then subsequently we estimate the dependence parameters from the random effects using these marginals.

To facilitate the two-stage model, we set up the random effects in a particular way by making sure the total variance of all random effects acting for each subject is the same. We let $(V_{k1}, \ldots, V_{km})$ be independent Gamma-distributed random variables, denoted as $V_{kl} \sim \Gamma(\eta_l, \nu), l = 1, \ldots, m$, such that the random effects have mean $\mathrm{E}(V_{kl}) = \eta_l/\nu$ and variance $\mathrm{Var}(V_{kl}) = \eta_l/\nu^2$. Furthermore, let $\Psi(\eta_l, \nu, \cdot)$ denote the Laplace transform of the Gamma distribution $\Gamma(\eta_l, \nu)$, and let its inverse be $\Psi^{-1}(\eta_l, \nu, \cdot)$. The $\eta = (\eta_1, \ldots, \eta_m)^{\mathrm{T}}$ parameters are given such that $\eta = D\theta$, where D is a $m \times p$ matrix and the parameters $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ are of dimension p. This makes it possible to specify restrictions on the parameters, for example, when considering standard polygenic models that we use (Falconer 1967, Neale & Cardon 1992, Falconer & Mackay 1994, Sham 1998).

The key assumption to make the two-stage construction possible is to assume that the total variance of the random effects for each subject is the same, ν, such that $\nu = Q_{ki}^{\mathrm{T}} \eta$ for all $i = 1, \ldots, n_k$ and all $k = 1, \ldots, K$. Therefore, $Q_{ki}^{\mathrm{T}} V$ is Gamma distributed: $\Gamma(\nu, \nu)$. We get back to specific models where this is the case, but this assumption is often reasonable and needed in the context of polygenic models that aim to characterize genetic effects (Korsgaard & Andersen 1998, Petersen 1998).

Marginally we assume that T_{ki} given covariates X_{ki} follows a Cox model,

$$\lambda_{ki}^m(t; X_{ki}) = \lambda_0(t) \exp(X_{ki}^{\mathrm{T}} \beta),$$

and let the related marginal survival function be denoted as $S_{ki}^m(t) = S^m(t|X_{ki})$. The estimators of the marginal models are obtained by fitting the models as if the data were independent and lead to consistent and asymptotically normal estimators, as pointed out by Spiekerman & Lin (1998).

Kendall's τ: measures dependence as the probability of concordance minus the probability of discordance

We now assume that, given the random effects of the cluster V_k and the covariates X_{ki}, Q_{ki} for $i = 1, \ldots, n_k$, subjects within the cluster are independent with survival distributions

$$S_{ki}(t|X_{ki}, Q_{ki}, V_k) = \exp(-\left(Q_{ki}^{\mathrm{T}} V_k\right) \Psi^{-1}(\nu, \nu, S_{ki}^m(t))).$$

Then, by construction, the marginal survival distribution, when integrating out the random effects, is given by $S_{ki}(t)$. Furthermore, the hazard is given as

$$\lambda_{ki}(t; X_{ki}, V_{ki}, Q_{ki}) = \left(Q_{ki}^{\mathrm{T}} V_k\right) \left[-D_3 \Psi^{-1}(\nu, \nu, S_{ki}^m(t)) S_{ki}^m(t)\right] \lambda_0(t) \exp(X_{ki}^{\mathrm{T}} \beta),$$

where D_3 denotes the partial derivatives with respect to the third argument of Ψ, and is then evaluated at $(\nu, \nu, S_{ki}(t))$. This is a more complicated looking hazard, but the random effects still act in a multiplicative manner.

We can express the multivariate survival distribution as

$$\begin{aligned} S(t_1, \ldots, t_{n_k}) &= \mathrm{E}(\exp(-\sum_{i=1}^{n_k} (Q_{ki}^{\mathrm{T}} V_k) \Psi^{-1}(\nu, \nu, S_{ki}^m(t_i)))) \\ &= \prod_{l=1}^{m} \Psi(\eta_l, \nu, \sum_{i=1}^{n_k} Q_{kil} \Psi^{-1}(\nu, \nu, S_{ki}^m(t_i))). \end{aligned} \quad 8.$$

In the case of bivariate data for subjects i and j within a cluster, we write this function as $S(t_1, t_2) = C(S_{ki}^m(t_1), S_{kj}^m(t_2))$.

This model is easy to fit using our software in the R package `mets` and can be fitted for large registry data by taking advantage of two-stage fitting; the standard errors of the dependence parameters can be derived by extending the original two-stage arguments to this setting, along the lines of Glidden (2000) and Shih & Louis (1995).

2.2.2. Kendall's τ. One consequence of the random effects acting multiplicatively in all the above conditional hazard models is that we can compute Kendall's τ for all models. The probability of concordance minus the probability of discordance in two clusters, 1 and 2, with subjects (i, j), for example, mother and daughter, is

$$\tau = \mathrm{E}\left(\mathrm{sgn}\left[(T_{1i} - T_{2i})(T_{1j} - T_{2j})\right]\right),$$

where sgn$(\cdot)$ gives the sign of its argument. Concordance is thus the probability that those from one cluster either come before or after the cluster to which they are compared—that is, the probability that either $(T_{1i} > T_{2i})$ and $(T_{1j} > T_{2j})$, or $(T_{1i} < T_{2i})$ and $(T_{1j} < T_{2j})$. Discordance is the opposite.

For the extended two-stage model, we can compute this dependence measure specifically as

$$\mathrm{E}\left(\frac{(Q_{1i}^{\mathrm{T}} V_1 - Q_{2i}^{\mathrm{T}} V_2)(Q_{1j}^{\mathrm{T}} V_1 - Q_{2j}^{\mathrm{T}} V_2)}{(Q_{1i}^{\mathrm{T}} V_1 + Q_{2i}^{\mathrm{T}} V_2)(Q_{1j}^{\mathrm{T}} V_1 + Q_{2j}^{\mathrm{T}} V_2)}\right)$$

under the assumption that we compare pairs with equivalent marginals $[S_{X_{1i}}(t) = S_{X_{2i}}(t)$ and $S_{X_{1j}}(t) = S_{X_{2j}}(t)]$ and that $S_{X_{1i}}(\infty) = S_{X_{1j}}(\infty) = 0$. Here, we use that ν is the same across clusters. Kendall's τ would be the same for Equation 6 due to the same additive structure for the frailty terms, and the random effects thus have the same interpretation in terms of Kendall's τ.

When we do not have full follow-up, we can still define Kendall's τ, but now also given two event times, and still compute concordance minus discordance using the formula.

We also note that the model with normally distributed random effects,

$$\exp(Q_{ki}^{\mathrm{T}}V_k)\lambda_0(t)\exp(X_{ki}^{\mathrm{T}}\beta),$$

when the marginal covariates are equivalent, leads to a Kendall's τ for two subjects (i,j) across two clusters, 1 and 2, on the form

$$\mathrm{E}\left(\frac{(\exp(Q_{1i}^{\mathrm{T}}V_1)-\exp(Q_{2i}^{\mathrm{T}}V_2))(\exp(Q_{2j}^{\mathrm{T}}V_1)-\exp(Q_{2j}^{\mathrm{T}}V_2))}{(\exp(Q_{1i}^{\mathrm{T}}V_1)+\exp(Q_{2i}^{\mathrm{T}}V_2))(\exp(Q_{1j}^{\mathrm{T}}V_1)+\exp(Q_{2j}^{\mathrm{T}}V_2))}\right).$$

This quantity can be approximated by numerical integration or with simulations. We note that Kendall's τ is independent of the marginals.

2.3. Conditional Hazard Models

When a family, or a pair of subjects, is observed over time, a possible relationship between the considered subjects will be reflected in the conditional intensity—that is, the instantaneous risk of death for a subject that is under risk given what we observed up to time t for the family, denoted by the history $\mathcal{H}_k(t)$:

$$\lambda_{ki}^{c}(t|\mathcal{H}_k(t),\tilde{T}_{ki}\geq t)=Y_{k1}(t)\lim_{h\to 0}\frac{1}{h}\,\mathrm{P}(T_{ki}\in[t,t+h]|\mathcal{H}_k(t),\tilde{T}_{ki}\geq t).$$

Andersen et al. (1993) provide a general treatment of these models. If this conditional hazard depends on the information we have accumulated about other subjects, it indicates that subjects are related and there will be familial aggregation. The random effects models that we considered above all lead to specific conditional models, and the conditional hazard model, therefore, provides a general approach for learning about dependence. Working with and estimating the parameters of the conditional hazard models can be based on fully specified models or composite likelihood methods (Varin et al. 2011). Considering a pair of twins ($i=1,2$), or a mother and child of a family, we may try to learn about a possible dependence by fitting an intensity model of the form

$$\lambda_{k1}^{c}(t|\mathcal{H}_k(t),\tilde{T}_{k1}\geq t)=Y_{k1}(t)\lambda_0(t)\exp(X_{ki}^{\mathrm{T}}\beta+\gamma N_{k2}(t-)), \qquad 9.$$

where the conditional intensity has an increase in the hazard if the cotwin has died. Note that twin 2 may be censored prior to t if the twins are censored at different points in time. In this model, that can be estimated by standard software and provided with robust standard errors; we consider γ as an indication of the strength of the association between the twins. When the model is fitted for MZ and DZ twins, we can evaluate if the twins are related and how strongly, and if MZ and DZ are differently related.

In the simple case where we do not adjust for covariates, it is useful to remember that if an underlying frailty model is assumed, with conditional hazard $Z\lambda(t)$ for each twin and Z Gamma distributed with mean 1 and variance θ, then the conditional hazard will be on the form

$$\lambda_{k1}^{c}(t|\mathcal{H}_k(t),\tilde{T}_{k1}\geq t)=Y_{k1}(t)\,(\alpha(t)+\theta\alpha(t)N_{k2}(t-))=Y_{k1}(t)\alpha(t)\exp(\log(\theta+1)N_{k2}(t-)),$$

Cause-specific hazard: the instantaneous risk of the event of interest among those still under risk

with $\alpha(t) = \lambda(t)/(1 + \theta(\Lambda(t) + \Lambda(\tilde{T}_{k2} \wedge t)))$, and $\Lambda(t) = \int_0^t \lambda_0(s)ds$. Note that $\alpha(t)$ is in fact also a function of $\tilde{T}_{k2}$ that we have hidden in our notation. This also reveals that the simple conditional hazard model (Equation 9), although useful, does provide parameters that are hard to interpret due to the lack of adjustment for the risk-time of the other twin that needs to be done in this model. Similar complications arise when the conditional intensity modeling is used for the cause-specific hazards in the competing risks models (Eriksson & Scheike 2015).

2.4. Worked Example: Time to Menopause in Twins

We consider 503 pairs of female twins born between 1931 and 1952 that were identified through the Danish Twin Registry. The twins are established as MZ or DZ. There are 269 MZ twin pairs and 234 DZ twin pairs. We here look at the time to menopause, with a broad definition that consisted of either natural menopause; surgical menopause as a result of removal of uterus, cervix, or ovaries; or hormone treatment for menopausal-related issues. Of the 1,006 twins, 845 experienced this broad definition of menopause.

An important first step is to look first at the marginals of the MZ and DZ twins, which often are believed to be the same and, in addition, to be similar to those of the general population. If these have the same marginals and thus the same total variation, it is possible to consider heritability estimates that describe how much of the variation is due to variation in the genes. Here, the marginals appeared similar.

We first estimated the variance of the MZ and DZ twins separately, using the two-stage model with the same marginal for MZ and DZ twins. The variances of the Gamma-distributed random effects (with 95% CI) were 1.08 (0.73; 1.42) and 0.11 (−0.12; 0.36) for MZ and DZ twins, respectively. Comparing the estimates as a test for genetic effects, we found $p < 0.001$, thus clearly rejecting that the dependence is the same for MZ and DZ twins.

Then, fitting the ACE model, we found the environmental effect at the boundary, thus suggesting that there is no shared environmental effect. We then considered the AE model, which leads to an estimate of a genetic variance at 0.98 (0.75; 1.22)—that can be transformed into a Kendall's τ at 0.33 and 0.14 for MZ and DZ sisters, respectively. Note also that the AE model without the C component shows a lack of fit when compared to the variance for MZ and DZ twins, when estimated without restrictions on the parameters. In contrast, the DE model leads to an estimate of a genetic variance at 1.08 (0.82; 1.33), and this model fits the data well. As we also observe below in the worked example of competing risks data, the choice of the preferred polygenic model, however, is highly dependent on the considered scale and modeling approach.

Using the fact that the two-stage hazard model is also a transformation model, with a transformation that depends on the covariates, however, we can still say that for given covariates X, the part of the variation due to genes is 50% using the AE model, since after a linear transformation, the survival times consist of $\log(Z) + \epsilon$, with ϵ being extreme value and $\log(Z)$ a log-transform of the Gamma-distributed random effect that accounts for genetic variation (for more on this, see Korsgaard et al. 1999, Yazdi et al. 2002). The random effects models may further be used to compute relevant probabilities, such as, for example, the probability that both twins have experienced menopause by a given age.

3. COMPETING RISKS MODELING

Generally speaking, there are two modeling approaches for competing risks data, one based on hazard modeling and one based on the cumulative incidence modeling (see **Figure 1*b***). The hazard modeling is based on the cause-specific hazards that give the instantaneous risk of experiencing a

specific cause for those that are still at risk, formally defined as

$$\lambda_k(t,X) = \lim_{h\to 0} \frac{1}{h} \mathrm{P}(T \in [t, t+h], \epsilon = k | T \geq t, X).$$

Cumulative incidence function: the cumulative probability of the event of interest

The cumulative incidence of cancer is given as

$$F_1(t,X) = \mathrm{P}(T \leq t, \text{cancer}|X) = \mathrm{P}(T \leq t, \epsilon = 1|X),$$

that is, the probability of cancer before time t given covariates X.

Similarly, in the multivariate setting with the aim of dependence modeling, one needs to model either the multivariate hazards or multivariate cumulative incidence functions. General discussions and reviews of methods for dependence modeling and familial aggregation for competing risks data are provided by Diao & Zeng (2013) and Bandeen-Roche (2013), respectively.

Having modeled either of these quantities, the model is fully specified. In terms of random effects models, we would need to specify either all cause-specific hazards given random effects, or similarly, all cumulative incidence functions given random effects.

When we have a particular interest in a particular cause, such as in the studies of heritability of cancer, then it is worth noticing that by using cumulative incidence modeling, it is sufficient to model the cumulative incidence of cancer and the strength of the cooccurrence in family members. In contrast, when hazard models are used, a fully specified random effects models for all causes is needed. In the remainder of the article, we focus particularly on cumulative incidence modeling and just briefly mention that there exists much interesting work on how to describe dependence in competing risks data via multivariate cross-ratios of the cause-specific hazards (see, for example, Bandeen-Roche & Liang 2002, Bandeen-Roche & Ning 2008, Shih & Albert 2010, Ning & Bandeen-Roche 2014). The methods described in this work are not directly applicable for polygenic modeling that requires more specific modeling to separate the sources of variation.

In the work of Lichtenstein et al. (2000) and Mucci et al. (2016) studying the heritability of cancer, the dependence modeling was based on cumulative incidence modeling of only the cause of interest, thus considering and modeling, in essence, the joint probability that a pair of family members, or twins, both have experienced cancer. This joint probability is given as the concordance probability

$$\mathcal{C}_{1,1}(t) = \mathrm{P}\left(T_1 \leq t, \epsilon_1 = \text{cancer}, T_2 \leq t, \epsilon_2 = \text{cancer}\right),$$

and this joint probability is often computed and compared with the expected concordance under independence, the recurrence risk, and can be computed and estimated nonparametrically (Scheike et al. 2015a). Fully nonparametric estimators of the full joint distribution of $\mathrm{P}\left(T_1 \leq s, \epsilon_1 = \text{cancer}, T_2 \leq t, \epsilon_2 = \text{cancer}\right)$ were described by Cheng et al. (2007). Below, we describe random effects models for the joint cumulative incidence in a family.

3.1. Competing Risks Hazards Random Effects Modeling

One modeling approach is to describe dependence, or cooccurrence in cancer (as in **Figure 1*b***), modeling the two hazards given covariates and random effects that may influence both hazards. Such random effects will then generate dependence for each of the causes and possible cross-dependence between cancer and death. These models have been used by Wienke (2011) and Eriksson & Scheike (2015) and are generally rather difficult to fit, particularly for large registry data. Given a full description of the cause-specific hazards given random effects, we have a fully specified model, and all consequences and probabilities can be computed. One important point is that we need all full models for all hazards and how they are related.

Concordance: probability of both relatives being cases

Casewise concordance: the risk given the relative was a case

Relative recurrence risk: excess risk to a relative of a case compared with the population risk

One useful observation is that if the competing risks' causes are independent, such that likelihood factors into the likelihood for the different event types (possibly given random effects that also factor), then the two-stage approach described for survival data can be applied to this setting using the cause-specific hazards. Thus, we would first fit the marginal cause-specific hazards and then apply the techniques for multivariate survival data described above. In addition, there is still a Kendall's τ interpretation in the sense that, given the timing of two event times for the cause of interest, the probability of concordance minus discordance is still computed using the formula for Kendall's τ (Equation 9).

When the causes are related, in the sense that random effects act on multiple causes, then the models are typically difficult to fit and work with, even though one can still compute the observed conditional hazard; this is particularly true for large registry data where no software exists for this type of modeling.

Without the assumption of independent risks, polygenic random effects modeling (for example, an ACE model) would be quite complicated due to possible shared genetic or environmental components for the different causes. Clearly—and also when considering possible polygenic modeling of multiple related causes (**Figure 1*c***)—some part of the genetic and environmental components must be shared across causes.

3.2. Concordance and Recurrence for Time to Event

We start by making the observation that there exist simple nonparametric estimators of the relative recurrence risk and the casewise concordance that extends those from the binary case that often are computed for binary trait twin/family studies.

Given a binary trait for a twin pair, (Y_1, Y_2), we have the joint distribution of the pair, $p_{ij} = \mathrm{P}(Y_1 = i, Y_2 = j)$ for $i, j = 0, 1$. Focusing on the twin case, we also have symmetry, so that $p_{10} = p_{01}$, thus leading to equivalent marginal rates $p_0 = p_{0\cdot} = p_{\cdot 0} = p_{00} + p_{01}$, and $p_1 = 1 - p_0$. Now, given observations from a cohort of size n, we have the counts $n_{ij} = \sum_n I(Y_1 = i, Y_2 = j)$ for $i, j = 0, 1$ and define $n_d = n_{0,1} + n_{1,0}$ as the number of discordant pairs. Here, $I()$ is the indicator that is one when the condition is fulfilled and zero otherwise.

The casewise concordance is defined as

$$P_c = \mathrm{P}(Y_1 = 1 | Y_2 = 1) = \frac{p_{1,1}}{p_1} = \frac{p_{11}}{p_{01} + p_{11}},$$

and the maximum likelihood estimators (MLEs), given by Witte et al. (1999), are $\hat{\mathrm{P}}_c = 2n_{1,1}/(2n_{1,1} + n_d)$ and $\hat{p}_1 = (2n_{1,1} + n_d)/2n$. Similarly, the relative recurrence risk, $R = P_c/p_1$, can also be computed and estimated by MLEs $\hat{R} = \hat{P}_c/\hat{p}_1$.

A direct analog to the age-of-onset setting is to compute and estimate the basic quantities for a fixed time, such that we define the concordance probability at time t (Scheike et al. 2015a, 2014b) as

$$\mathcal{C}_{i,j}(t) = \mathrm{P}_{i,j}(t,t) = \mathrm{P}(T_1 \leq t, \epsilon_1 = i, T_2 \leq t, \epsilon_2 = j) \quad \text{for} \quad i, j = 1, 2,$$

and given the marginal probability, the cumulative incidence $F_1(t)$, we can still define the casewise concordance, $\mathcal{C}^c_{1,1}(t) = \mathcal{C}_{1,1}(t)/F_1(t)$, and the recurrence risk $\mathcal{R}_{1,1}(t) = \mathcal{C}_{1,1}(t)/F_1^2(t)$. We can also consider the recurrence risk across different causes (see **Figure 1*c***) as the probability $\mathcal{C}_{1+2}(t) = \mathcal{C}_{1,2}(t) + \mathcal{C}_{2,1}(t)$ and the related recurrence risk $\mathcal{R}_{1+2}(t) = \mathcal{C}_{1+2}(t)/(2F_1(t)F_2(t))$.

Critically, one can estimate the concordance probability accounting for right-censoring. A simple approach for estimating concordance probabilities is to act under the same-censoring assumption, that all pairs are censored at the first censoring for each pair. This is typically satisfied automatically in twin studies, and under this assumption we can estimate the concordance

probability by a standard Aalen–Johansen cumulative incidence estimator Andersen et al. (1993) based on observing if the pair moves to the state where both have the cancer of interest or to a competing state, or possibly at some point in time are censored. A useful observation is also to note that regression modeling can be applied to the concordance probabilities, such that summaries can be made of how covariates are important for the concordance probability, $\mathcal{C}_{i,j}(t|X)$, using competing risks regression modeling.

3.3. Cumulative Incidence Random Effects Modeling

The cumulative incidence model has been extended to clustered data by random effects or copula models by Katsahian et al. (2006), Scheike et al. (2010), Dixon et al. (2011), and Cheng & Fine (2012). We again specify a two-stage approach based on marginal cumulative incidence models, where conditional on a random effect V, the cumulative incidence of cancer for a family member is given as

$$F_1(t, X_j, V) = \mathrm{P}(T \leq t, \text{cancer}|V, X_j) = 1 - \exp\left(-V \Psi_\theta^{-1}\left[1 - F_1(t, X_j)\right]\right), \qquad 10.$$

where Ψ_θ is the Laplace transform of the random effect V that here is assumed Gamma distributed with mean 1 and variance θ, and Ψ_θ^{-1} is its inverse, thus leading to the marginal cumulative incidence $F_1(t, X)$ given X. To estimate such marginal cumulative incidence for correlated data, one can apply the marginal modeling approach (see Chen et al. 2008, Scheike et al. 2010). To estimate the dependence parameters, we then note that given the marginals and the dependence parameter, we can compute, for example, the concordance function given covariates $\mathcal{C}_{1,1}(t, X_i, X_j, \theta) = \mathrm{P}(T_1 \leq t, \epsilon_1 = 1, T_2 \leq t, \epsilon_2 = 1|X_i, X_j)$ that can be compared to what is seen in the data. This leads to an IPCW estimating function

$$U(\theta) = \sum_k \sum_{i,j} C_{k,ij}(t) V_{k,ij} \left[I(T_{ki} \leq t, \epsilon_{ki} = 1, T_{kj} \leq t, \epsilon_{kj} = 1) - \mathcal{C}_{1,1}(t, X_{ki}, X_{kj}, V_{k,ij}^{\mathrm{T}} \theta) \right]$$

with $C_{k,i,j}(t) = I(T_{ki} < C_k) I(T_{kj} < C_k) / G_c(\min(T_{ki}, T_{kj}) | X_{ki}, X_{kj}))$, and i, j are two subjects within cluster k. We operate again under the same-censoring assumption to simplify the censoring weight.

The parameters of such a model can be estimated in many ways, but a key feature of this model is, again, that we can fit the marginal cumulative incidence models first using standard competing risks regression methods and then subsequently estimate the dependence parameters. Furthermore, this type of modeling can be extended to structured random effects modeling for polygenic modeling, as by Scheike et al. (2014a).

The simple concordance modeling can be supplemented and summarized differently by the more detailed description of the joint cumulative incidence of prostate cancer, which is modeled to approximate well the concordance probabilities. Furthermore, one could jointly model all cumulative incidence functions to learn about cross-cancer dependence.

3.4. Liability Threshold Model

An alternative modeling approach can be obtained via the probit model (Equation 1) by adjusting for right-censoring through inverse probability weighting (Holst et al. 2016). This is similar to the direct binomial regression models of Scheike et al. (2008) for univariate time to event outcomes. The binary outcome of interest is

$$Y_{ij}(\tau) = I(T_{ij} \leq \tau, \epsilon_{ij} = 1), \quad j = 1, \ldots, n_k; i = 1, \ldots, K,$$

and we let the full data likelihood score function be given by $\mathcal{U}\{\theta; \boldsymbol{Y}_k(\tau), \boldsymbol{X}_k\}$ with $\boldsymbol{Y}_k(\tau) = (Y_{k1}(\tau)\delta_{k1}, \ldots, Y_{kn_k}(\tau)\delta_{kn_k})^T$, $\boldsymbol{X}_k(\tau) = (X_{k1}, \ldots, X_{kn_k})^T$, and $\theta \in \mathbb{R}^p$ is the parameter vector. The

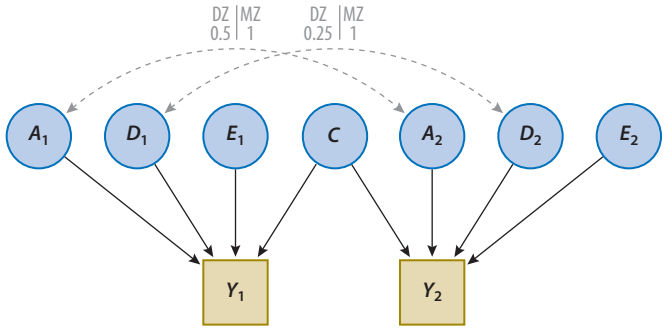

Figure 2

Path diagram for the ACDE polygenic model for twin data. Abbreviations: A, additive genetic effects; C, shared environmental effects; D, dominant genetic effects; DZ, dizygotic; E, individual environmental effects; MZ, monozygotic.

IPCW adjusted estimator is then defined as the root of the estimating function

$$\mathcal{U}_{\text{IPCW}}(\theta; \mathcal{O}_1, \ldots, \mathcal{O}_{n_k}) = \sum_{k=1}^{K} \frac{\prod_{i=1}^{n_k} \delta_{ki}}{G_c(\widetilde{T}_{k1}, \ldots, \widetilde{T}_{kn_k}; \boldsymbol{X}_k)} \mathcal{U}\{\boldsymbol{Y}_k(\tau), \boldsymbol{X}_k; \theta\},$$

where $\mathcal{O}_k = (\boldsymbol{Y}_k(\tau), \boldsymbol{X}_k, \delta_{k1}, \ldots, \delta_{kn_k}, \widetilde{T}_{k1}, \ldots, \widetilde{T}_{kn_k})$ are the observed data. General family dependence structures may be modeled using this approach. However, for larger cluster sizes, n_k, the calculations become computationally intensive due to the numerical integration involved in the calculations of the multidimensional normal probability functions. In this case, pairwise modeling via a composite likelihood approach (Varin et al. 2011) can be applied instead. Practically, the parameter θ can be estimated using a two-stage approach where a plug-in estimator of the censoring distribution is used in the above estimating function. A challenge is that consistency relies on a correctly specified censoring distribution. As previously mentioned, we can, under a same-censoring assumption, use an estimate given by

$$\widehat{G}_c(\widetilde{T}_{k1}, \ldots, \widetilde{T}_{kn_k}; \boldsymbol{X}_k) = \bigwedge_{i=1}^{n_k} \widehat{G}_c(\widetilde{T}_{ki}; X_{ki}),$$

where the marginal censoring distributions $\widehat{G}_c(\widetilde{T}_{ki}; X_{ki})$ may be fitted using a semiparametric model such as a Cox proportional hazards model.

This modeling approach is particularly appealing when the main objective is to separate variation into environmental and genetic components. Here, we adopt the standard polygenic model for twin data based on further decomposing the genetic components, as in Equations 2 and 3, into A and D effects, and the environmental effects into C and E effects. This model is illustrated in **Figure 2**. Here, we consider the case where we allow the parameters to be time dependent as a function of τ,

$$\text{P}(Y_i(\tau) = 1 | X_i, V_i) = \Phi\left\{X_i^{\text{T}}\beta(\tau) + V_{\text{A}(\tau),i} + V_{\text{D}(\tau),i} + V_{\text{C}(\tau)}\right\}, \qquad 11.$$

which then gives us the opportunity to explore how the genetic and environmental variation on the liability scale evolve over the lifetime.

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 that

$$\text{cov}(V_{\text{A}(\tau),1}, V_{\text{A}(\tau),2}) = \tfrac{1}{2}\sigma^2_{\text{A}(\tau)}, \qquad \text{cov}(V_{\text{D}(\tau),1}, V_{\text{D}(\tau),2}) = \tfrac{1}{4}\sigma^2_{\text{D}(\tau)}.$$

In practice, only two of the variance components V_A, V_C, and V_D, which are here assumed to be mutually independent and normal distributed, can be identified unless information on, for example, adoption status is included in the model. More on different genetic models is provided by, for example, Neale & Cardon (1992) and Sham (1998). In a given polygenic model, the heritability, the part of the variation due to genes, can be computed. This type of summary measure has several weaknesses but also some appeal. In the context of the ACE model for the liability threshold model, we would say that the heritability is

$$H^2(\tau) = \frac{\sigma_A^2(\tau)}{\sigma_A^2(\tau) + \sigma_C^2(\tau) + 1}. \qquad 12.$$

Note also that the variances relate to the liability scale and that, therefore, there is additional variation present in the data on the risk scale.

Another reasonable modeling approach is to apply a more flexible bivariate probit model where the marginals are kept identical, but instead of a random effect structure, the bivariate distribution of the twin pairs is captured by a correlation parameter that depends on the zygosity, i.e.,

$$Y_i^*(\tau) = \mu_{\text{zyg}}(\tau) + \epsilon_i(\tau),\ i = 1, 2; \quad (\epsilon_1, \epsilon_2)^T \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{\text{zyg}}(\tau) \\ \rho_{\text{zyg}}(\tau) & 1 \end{bmatrix}\right).$$

Again, by considering the estimates for different time points τ, this gives an alternative and flexible approach to estimating parameters on the probability scale as a function of age, such as concordance probabilities and relative recurrence estimates, as in Section 3.2.

3.5. Worked Example: Time to Prostate Cancer in Twins

We consider a data set in the `mets` R package that resembles the data of Hjelmborg et al. (2014) that were based on the Nordic Twin Study of Cancer, a collaborative research project studying the genetic and environmental components of prostate cancer. The data comprise about 18,000 DZ twins and 11,000 MZ twins. It was a population-based register study based on the Danish, Finnish, Norwegian, and Swedish twin registries.

3.5.1. Hazards random effects modeling. In the case of the simple competing risks model with only death as the competing risk (**Figure 1*b***), it may be a reasonable assumption that the cause-specific hazards can be modeled independently and with different random effects acting on the cause-specific hazards separately. This is in contrast to the case where the interest is on the competing risks model with multiple causes of interest (**Figure 1*c***), where the interest indeed would be in the relationships between the two-cause–specific hazards for the causes of interest. Using a marginal hazard model for the cause-specific hazard of prostate cancer stratified on country, we fitted the two-stage random effects model with different variances for MZ and DZ twins.

This leads to random effects of 1.32 (0.56; 2.08) and 5.42 (3.54; 7.31) for MZ and DZ twins, respectively. These can be translated into Kendall's τ of 0.73 and 0.40 for MZ and DZ twins, respectively, thus showing a very strong positive correlation for both types of twins. An attempt to decompose these variances into genetic and environmental effects led to the conclusion that the ACE and AE did not fit the data particularly well, and we therefore also fitted a DE model that led to a genetic variance at 5.05 (3.44; 6.67) and no suggestion of an environmental effect.

3.5.2. Concordance modeling. First, we estimate the concordance probability—that is, that both twins have prostate cancer at different ages. For simplicity of presentation, we do not initially

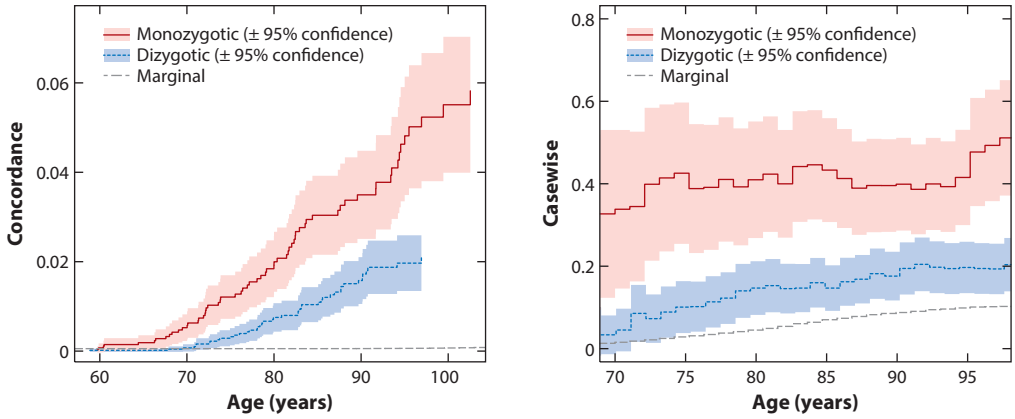

Figure 3

Concordance and casewise concordance for prostate cancer in monozygotic and dizygotic twins.

stratify according to the different countries. The twins are censored at the same time; otherwise, we would enforce this in the data by artificially censoring both twins at the first censoring time. However, given that we have the same-censoring assumption satisfied, we can do the standard Aalen–Johansen product limit estimator of the concordance probabilities for MZ and DZ twins (see **Figure 3**). We see that the concordance is considerably higher for MZ twins compared to DZ twins and that the casewise concordance also suggests clearly that there is positive dependence present for both DZ and MZ twins in terms of the occurrence of prostate cancer. The lifetime risk of a male twin getting prostate cancer (ignoring country differences) was about 10%, and the risk of both twins getting prostate cancer was about 5% and 2% for MZ and DZ twins, respectively, thus suggesting a strong positive cooccurrence in prostate cancer. Similarly, looking at the casewise concordance, if your cotwin has had cancer by the age of 80, your risk is about 40% and 13% for MZ and DZ twins, respectively.

The simple analysis of concordance did not take into account that the cumulative incidence is quite different in the different Nordic countries; in particular, Denmark uses different criteria for the diagnosis of prostate cancer (for more on this, see Hjelmborg et al. 2014). To describe this further, we also considered a competing risks regression model for the concordance. We started by considering an interaction between country and MZ/DZ difference ($p = 0.15$) and then went on to describe the difference between MZ and DZ twins adjusting for country differences (see **Table 2**). Using a logit link, we modeled the concordance probability and found that the concordance was considerably lower in Denmark compared with other Nordic countries, with an odds ratio (OR) of about 3, that MZ twins had a concordance with an OR of about 2.8 compared with DZ twins, and this was similar for all Nordic countries. The logit concordance model makes a statement

Table 2 Concordance regression

	OR Estimate	95% CI
Finland	3.57	1.11–6.03
Norway	2.49	0.54–4.43
Sweden	3.00	1.11–4.90
zygMZ	2.81	1.70–3.93

Abbreviations: CI, confidence interval; OR, odds ratio; zygMZ, effect of being a monozygotic twin pair.

about the concordance probabilities across the time range observed in the study and is thus valid for all ages. Alternatively, one can also do this modeling for a specific point in time, which is more in line with the liability threshold modeling.

3.5.3. Cumulative incidence random effects modeling. The simple concordance modeling can be supplemented and summarized differently by the more detailed description of the joint cumulative incidence of prostate cancer that can be modeled with our random effects models. To do this, we used a model for the marginal cumulative incidence that allowed different marginal cumulative incidences in the different Nordic countries. We then, first, considered the simple unstructured random effects model with different random effect variances for MZ and DZ twins, respectively. This type of modeling can be done by modeling over the entire time range or by considering a fixed point in time. We here considered the entire time range by fitting the model at the ages 50, 60, 70, 80 and 90.

This led to random effects of 3.22 (1.36; 5.08) and 0.789 (0.05; 1.54) for MZ and DZ twins, respectively, thus showing a very strong positive correlation for both types of twins. An attempt to decompose these variances into genetic and environmental effects led to the conclusion that the ACE and AE did not fit the data particularly well, and we therefore also fitted a DE model that led to a genetic variance at 3.34 (1.56; 5.04) and no suggestion of an environmental effect.

Interestingly, when we based the estimation only on the concordance at 90 years of age, we found, in contrast, random effects 2.67 (1.04; 4.30) and 0.66 (−0.05; 1.36) for MZ and DZ twins, respectively. The DE model led to a genetic variance of 2.74 (1.19; 4.29). There is some suggestion that the Gamma-distributed random effects do not describe the dependence well across the entire age range, by comparing the nonparametric concordance estimator with that obtained from the random effects models. Therefore, in this study, it was a good idea to consider specific time horizons. We return to this point below when considering the liability threshold model with different time horizons.

3.5.4. Liability threshold. We estimated the parameters of a bivariate probit model separately for MZ and DZ twins at $\tau = 95$ years with the right-censoring distribution estimated by the Kaplan–Meier (KM) estimator.

Here, we see a relative recurrence risk of 4.44 (3.50; 5.38) in MZ twins and 2.48 (1.87; 2.81) in DZ twins. This indicates familial aggregation with elevated risk in both groups compared with the background population risk. The marginal risk of getting prostate cancer before age 95 is estimated to be 0.089 (0.081; 0.097). There is significantly stronger dependence between the MZ twins, suggesting that there is a strong genetic component to the familial aggregation. We can also calculate dependence measures on the liability scale in terms of the tetrachoric correlations, with estimates in the MZ group of 0.63 (0.51; 0.72) and the DZ group 0.34 (0.22; 0.45). Next, we estimated a model with an ACE random effects structure. In this case, the fit of the ACE model is indistinguishable from the bivariate probit model, and we obtained a heritability estimate of $H^2 = 0.58$ (0.26; 0.89). A similar conclusion can be drawn when adjusting for country-specific effects in the marginal of the bivariate probit model and censoring distribution (stratified KM), where we see an estimate of $H^2 = 0.59$ (0.25; 0.94).

We also examined how the dependence within MZ and DZ twins evolves with age by choosing different values of τ, with different parameters at each time point, as in Equation 11. The results are shown in **Figure 4**, where we see a minor tendency to stronger dependence at earlier age, though this seems to be happening for both MZ and DZ twins. We also repeated this analysis for the ACE random effects model. The two models are generally well aligned, as also shown by correlation plots from the ACE model in **Figure 5**, which agree nicely with the estimates from the

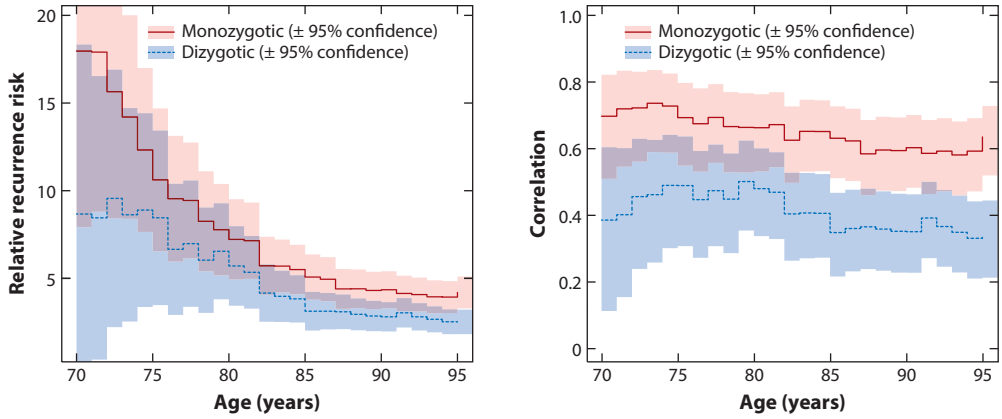

Figure 4

Relative recurrence risk and tetrachoric correlation estimates from IPCW adjusted bivariate probit models with pointwise 95% confidence intervals. Abbreviation: IPCW, inverse probability of censoring weighting.

bivariate probit model. From the heritability plot in **Figure 5**, we see a fairly constant estimate of the broad-sense heritability over time, with an estimate of about 50%. The CIs are broad, especially at earlier ages where there are fewer events.

3.5.5. Summary of worked example. The prostate cancer data were analyzed using event history methods to quantify and compare the dependence in MZ and DZ twins, respectively. Even though modeling of the cause-specific hazards may be carried out under an assumption of independent risks, which may also be validated, we generally believe that it is often more constructive to be on the risk scale, thus modeling the risk of the event of interest. Nevertheless, we found and quantified dependence estimates for the cause-specific hazards for MZ and DZ twins, and these were translated into Kendall's τ for ease of interpretation. The dependence was further quantified and separated into different sources on the hazard scale.

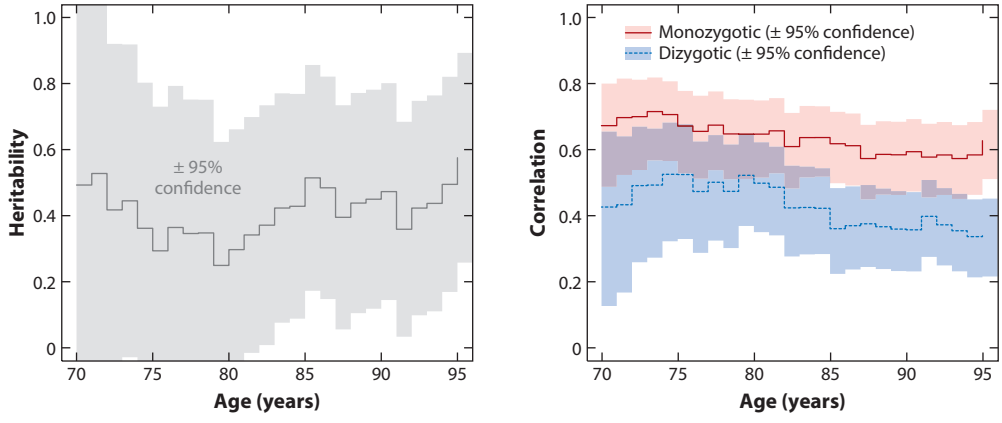

Figure 5

Heritability H^2 and tetrachoric correlation estimates from IPCW adjusted liability threshold model with an ACE random effects structure. Abbreviations: ACE, additive shared and individual environmental effects; IPCW, inverse probability of censoring weighting.

On the risk scale, we estimated nonparametrically the concordance and casewise concordance to quantify the strength of dependence in the twins. Additional modeling was carried out using polygenic random effects models, based on either two-stage Gamma-distributed random effects or the liability threshold model. Even though both models were fitted to approximate the concordance and the marginals in the data, and both fitted the data well, they led to somewhat different conclusions in terms of the polygenic model that was deemed most appropriate. This demonstrates that conclusions based on such models are heavily dependent on the scale on which they take place and therefore should be interpreted cautiously, thus suggesting that conclusions should primarily be based on what is seen on the risk scale.

4. DISCUSSION

Family studies are central tools for understanding the etiology of diseases, i.e., by studying if the risk of disease is greater for individuals with affected relatives, and to investigate to what extent such familial aggregation may be due to genetic factors. A key point is that the timing aspect of the event, in most cases, cannot be ignored. Often the outcome of interest is the lifetime risk of getting the disease, and here special attention should be made to dealing with right-censoring in the data, i.e., the individuals who are still at risk at the end of follow-up. Another main point is that competing risks due to, for example, death should also be taken into account in the analysis to avoid bias in the estimation of dependence measures.

We have discussed several approaches in this review for dealing with these challenges in family studies. In the absence of competing risks, as in the menopause example, or when the competing risks can be assumed to be independent, multivariate survival analysis techniques can be applied, such as frailty models based on additive Gamma or Gaussian distributions. In practice, the implementation of these models can be difficult in large registry studies due to computational challenges related to, for example, numerical integration. Here, two-stage models based on copulas are particularly attractive due to their computational efficiency. Kendall's τ is easily calculated for these models, which gives an interpretable dependence measure on the probability scale.

For competing risks data, several options are available. Random effects modeling is possible either on the hazard scale for each cause or directly on the cumulative incidence scale. In the common case, when considering disease onset and death as competing risks, it is possible to apply the liability threshold model with adjustment for right-censoring via IPCW. This makes it possible to decompose the variation into genetic and environmental components and, thus, to calculate measures of heritability that are commonly reported in quantitative genetics. While the heritability interpretation relies on a number of genetic assumptions and does not capture all the variation on the risk scale, we note that when the polygenic model fits the data well, the heritability measure provides a simple summary measure of the dependence structure. For example, for the ACE model in the classic twin design, the heritability is simply two times the difference in tetrachoric correlation between MZ and DZ twins, $2(\rho_{MZ} - \rho_{DZ})$.

Rather than putting too much emphasis on a single summary of the differences in dependence due to genes according to one genetic model, we recommend that the dependence also be measured using other established measures on the risk scale, such as relative recurrence risks and the casewise concordance. These numbers are more easily interpretable.

We illustrate how to carry out the analyses demonstrated in the review using our R package `mets` in the **Supplemental Appendix**.

Supplemental Material >

An important direction for further development is to consider in further detail the extended competing risks models with multiple causes of interest. Here, particular interest may center on the cross-correlations across causes, but these models require that all multiple causes are modeled

simultaneously and therefore are complicated. The methods and models considered here can be further developed to deal with ascertainment sampling based on probands such as case-control sampling of families (see Chatterjee et al. 2006, Matthews et al. 2008).

In addition, when there is interest on both timing and risk, such as when deciding if early cancers are more heritable, special models need to be developed to deal with such questions (see, for example, Cederkvist et al. 2019). Such models are naturally founded on cumulative incidence models, since the cumulative incidence function gives the timings of the events of interest in contrast to, for example, the cause-specific hazard.

SUMMARY POINTS

1. The study of familial aggregation for age of onset of diseases or death needs to be addressed using appropriate methods that take right-censoring into account. When specific diseases are the object of interest, one is forced to study the phenomenon using competing risks modeling due to death as well as other potential competing risks.
2. The two main directions of modeling are based on hazard modeling (or risk modeling) and cumulative incidence modeling; dependence is often described using random effects models.
3. The results of polygenic random effects modeling can be difficult to interpret and depend on the specific models used.
4. Simple and easily interpreted summary measures on the risk scale, such as concordance probabilities, casewise concordance, and relative recurrence risk, are useful to compute.

FUTURE ISSUES

1. The methods introduced here need modification to deal appropriately with ascertained and case-control sampling, which needs to take the sampling into account. In the case of delayed entry, which is often the case for many registries, there is also a need for special attention as software is often not adapted to this situation.
2. Models that deal with the relationship and possible aggregation of different diseases need to be further developed. This type of modeling will typically need to be done jointly with death in a competing risks model.
3. Models that deal with timing as well as risk, two separate components, need to be developed further, and there is often an interest in making statements about a possible relationship between these.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Andersen PK, Borgan Ø, Gill RD, Keiding N. 1993. *Statistical Models Based on Counting Processes*. New York: Springer

Balan TA, Putter H. 2019. `frailtyem`: an R package for estimating semiparametric shared frailty models. *J. Stat. Softw.* 90:1–29

Bandeen-Roche K. 2013. Familial studies. In *Handbook of Survival Analysis*, ed. JP Klein, H van Houwelingen, J Ibrahim, TH Scheike, pp. 549–68. Boca Raton, FL: Chapman and Hall/CRC

Bandeen-Roche K, Liang KY. 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 bivariate failure time associations in the presence of a competing risk. *Biometrika* 95:221–32

Cederkvist L, Holst K, Andersen K, Scheike T. 2019. Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing. *Biostatistics* 20:199–217

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

Chen BE, Kramer JL, Greene MH, Rosenberg PS. 2008. Competing risks analysis of correlated failure time data. *Biometrics* 64:172–79

Cheng Y, Fine JP. 2012. Cumulative incidence association models for bivariate competing risks data. *J. R. Stat. Soc. Ser. B* 74:183–202

Cheng Y, Fine JP, Kosorok MR. 2007. Nonparametric association analysis of bivariate competing-risks data. *J. Am. Stat. Assoc.* 102:1407–15

Clayton DG. 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65:141–51

Diao G, Zeng D. 2013. Clustered competing risks. In *Handbook of Survival Analysis*, ed. JP Klein, H van Houwelingen, J Ibrahim, TH Scheike, pp. 511–22. Boca Raton, FL: Chapman and Hall/CRC

Dixon S, Darlington G, Desmond A. 2011. A competing risks model for correlated data based on the subdistribution hazard. *Lifetime Data Anal.* 17:473–95

Donohue MC, Xu R. 2019. `phmm`: proportional hazards mixed-effects models. *R Package*, version 0.7–11. **https://github.com/mcdonohue/phmm**

Duchateau L, Janssen P. 2007. *The Frailty Model*. New York: Springer

Ducrocq V, Casella G. 1996. A Bayesian analysis of mixed survival models. *Genet. Sel. Evol.* 28:505–29

Eriksson F, Scheike T. 2015. Additive Gamma frailty models with applications to competing risks in related individuals. *Biometrics* 71:677–86

Falconer D. 1967. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* 31:1–20

Falconer D, Mackay T. 1994. *Introduction to Quantitative Genetics*. Boston: Addison-Wesley

Glidden DV. 1999. Checking the adequacy of the Gamma frailty model for multivariate failure times. *Biometrika* 86:381–93

Glidden DV. 2000. A two-stage estimator of the dependence parameter for the Clayton-Oakes model. *Lifetime Data Anal.* 6:141–56

Glidden DV, Self S. 1999. Semiparametric likelihood estimation in the Clayton-Oakes failure time model. *Scand. J. Stat.* 26:363–72

Hill WG. 2014. Applications of population genetics to animal breeding, from Wright, Fisher and Lush to genomic prediction. *Genetics* 196:1–16

Hjelmborg J, Scheike T, Holst K, Skytthe A, Penney K, et al. 2014. The heritability of prostate cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol. Biomarkers Prev.* 23:2303–10

Holst KK, Scheike TH, Hjelmborg JB. 2016. The liability threshold model for censored twin data. *Comput. Stat. Data Anal.* 93:324–35

Hougaard P. 2000. *Analysis of Multivariate Survival Data: Statistics for Biology and Health*. New York: Springer

Hu T, Nan B, Lin X, Robins J. 2011. Time-dependent cross ratio estimation for bivariate failure times. *Biometrika* 98:341–54

Katsahian S, Resche-Rigon M, Chevret S, Porcher R. 2006. Analysing multicenter competing risks data with a mixed proportional hazards model for the subdistribution. *Stat. Med.* 25:4267–78

Klein J. 1992. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 48:795–806

Korsgaard IR, Andersen AH. 1998. The additive genetic Gamma frailty model. *Scand. J. Stat.* 25:225–69

Korsgaard IR, Andersen AH, Jensen J. 1999. Discussion of heritability of survival traits. In *Proceedings of the International Workshop of Genetic Improvement of Functional Traits in Cattle: Longevity, Jouy-en-Josas, France*, pp. 31–35. Uppsala, Swed.: Interbull

Lange K. 2002. *Mathematical and Statistical Methods for Genetic Analysis*. New York: Springer. 2nd ed.

Li H. 1999. The additive genetic Gamma frailty model for linkage analysis of age-of-onset variation. *Ann. Hum. Genet.* 63:455–68

Lichtenstein P, Holm N, Verkasalo P, Iliadou A, Kaprio J, et al. 2000. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* 343:78

Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*, Vol. 1. Sunderland, MA: Sinauer

Matthews AG, Finkelstein DM, Betensky RA. 2008. Analysis of familial aggregation studies with complex ascertainment schemes. *Stat. Med.* 27:5076–92

Möller S, Mucci L, Harris J, Scheike T, Holst K, et al. 2016. The heritability of breast cancer among women in the Nordic Twin Study of Cancer. *Cancer Epidemiol. Biomarkers Prev.* 25:145–50

Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, et al. 2016. Familial risk and heritability of cancer among twins in Nordic countries. *JAMA* 315:68–76

Nan B, Lin X, Lisabeth L, Harlow S. 2006. Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *J. Am. Stat. Assoc.* 101:65–77

Neale MC, Cardon LR. 1992. *Methodology for Genetic Studies of Twins and Families*. Dordrecht, Neth.: Kluwer Academic

Nielsen GG, Gill RD, Andersen PK, Sørensen TIA. 1992. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Stat.* 19:25–43

Ning J, Bandeen-Roche K. 2014. Estimation of time-dependent association for bivariate failure times in the presence of a competing risk. *Biometrics* 70:10–20

Oakes D. 1982. A model for association in bivariate survival data. *J. R. Stat. Soc. Ser. B* 44:414–22

Petersen JH. 1998. An additive frailty model for correlated life times. *Biometrics* 54:646–61

Petersen JH, Andersen PK, Gill RD. 1996. Variance component models for survival data. *Stat. Neerl.* 50:191–211

Ripatti S, Palmgren J. 2000. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 56:1016–22

Scheike T, Hjelmborg J, Holst K. 2015a. Estimating twin pair concordance for age of onset. *Behav. Genet.* 45:573–80

Scheike T, Holst K, Hjelmborg J. 2014a. Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Anal.* 20:210–33

Scheike T, Holst K, Hjelmborg J. 2014b. Estimating twin concordance for bivariate competing risks twin data. *Stat. Med.* 33:1193–204

Scheike T, Holst K, Hjelmborg J. 2015b. Measuring early or late dependence for bivariate lifetimes of twins. *Lifetime Data Anal.* 21:280–99

Scheike TH, Sun Y, Zhang MJ, Jensen TK. 2010. A semiparametric random effects model for multivariate competing risks data. *Biometrika* 97:133–45

Scheike TH, Zhang MJ, Gerds T. 2008. Predicting cumulative incidence probability by direct binomial regression. *Biometrika* 95:205–20

Sham P. 1998. *Statistics in Human Genetics*. New York: Wiley

Shih J, Albert P. 2010. Modeling familial association of ages at onset of disease in the presence of competing risk. *Biometrics* 66:1012–23

Shih JH, Louis TA. 1995. Inference on association parameter in copula models for bivariate survival data. *Biometrics* 51:1384–99

Sørensen K, Juul A, Christensen K, Skytthe A, Scheike T, Jensen TK. 2013. Birth size and age at menarche: a twin perspective. *Hum. Reprod.* 28:2865–71

Spiekerman CF, Lin DY. 1998. Marginal regression models for multivariate failure time data. *J. Am. Stat. Assoc.* 93:1164–75

Therneau T. 2020. `coxme`: mixed effects Cox models. *R Package*, version 2.2-16. **https://CRAN.R-project.org/package=coxme**

Vaida F, Xu R. 2000. Proportional hazards model with random effects. *Stat. Med.* 19:3309–24

Varin C, Reid N, Firth D. 2011. An overview of composite likelihood methods. *Stat. Sin.* 21:5–42

Wienke A. 2011. *Frailty Models in Survival Analysis*. Boca Raton, FL: Chapman and Hall/CRC

Witte J, Carlin J, Hopper J. 1999. Likelihood-based approach to estimating twin concordance for dichotomous traits. *Genet. Epidemiol.* 16:290–304

Yazdi M, Visscher P, Ducrocq V, Thompson R. 2002. Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. *J. Dairy Sci.* 85:1563–77