Published in final edited form as:
Lifetime Data Anal. 2013 January ; 19(1): . doi:10.1007/s10985-012-9229-1.

Competing risks with missing covariates: effect of haplotypematch on hematopoietic cell transplant patients

Thomas H. Scheike,
Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Martin J. Maiers,
National Marrow Donor Program, Minneapolis, MA, USA

Vanderson Rocha, and
Hematology Bone Marrow Transplant Department, Hospital Saint-Louis, Paris, France

Mei-Jie Zhang
Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA

Thomas H. Scheike: ts@biostat.ku.dk

Abstract

In this paper we consider a problem from hematopoietic cell transplant (HCT) studies where there is interest on assessing the effect of haplotype match for donor and patient on the cumulative incidence function for a right censored competing risks data. For the HCT study, donor's and patient's genotype are fully observed and matched but their haplotypes are missing. In this paper we describe how to deal with missing covariates of each individual for competing risks data. We suggest a procedure for estimating the cumulative incidence functions for a flexible class of regression models when there are missing data, and establish the large sample properties. Small sample properties are investigated using simulations in a setting that mimics the motivating haplotype matching problem. The proposed approach is then applied to the HCT study.

Keywords

Binomial modeling; Bone marrow transplant; Competing risks; Haplotype effects; Haplotype match; Missing covariates; Inverse-censoring probability weighting; Nonparametric effects; Non-proportionality; Regression effects

1 Introduction

For the medical studies involving competing risks, one often wishes to estimate and model the cumulative incidence probability, the marginal probability of failure for a specific cause. The cumulative incidence curves and cause specific hazard functions for all causes contain the same information but are represented in a different ways and thus leads to different interpretations. Both quantities are generally of interest. Let $\lambda_1(t; z)$ be the cause-specific hazard of a cause one event and $\lambda_2(t; z)$ be the cause-specific hazard of an event of other causes than one, where both hazards are conditionally given by a set of covariates z. Assuming that cause one is the primary cause of interest, the cumulative incidence curve for

Correspondence to: Thomas H. Scheike, ts@biostat.ku.dk.

cause one, that is the probability of experiencing cause one before time t, given covariates z which is given by

$$P_1(t;z)=P(T \leq t, \varepsilon=1|z)=\int_0^t \lambda_1(s;z) \exp\left[-\int_0^{s^-} \{\lambda_1(u;z)+\lambda_2(u;z)\}\, du\right] ds,$$

where ε indicates the type cause of failure. Recently, several new methods have been developed to directly model the cumulative incidence probability of a specific cause of failure (Fine and Gray 1999; Scheike et al. 2008; Scheike and Zhang 2008).

The aim of this paper is to consider the situation where there are missing covariates for all individuals. We consider data from a hematopoietic cell transplantation (HCT) study. HCT is a life saving procedure for many cancer patients. With shrinking family sizes, the lack of human leukocyte antigen (HLA) matched sibling is common, and this has increased the use of alternative donor graft sources. One of the alternatives is to use a donor who is not from the patient's family. Unfortunately, in addition to its curative effect, HCT also has potentially lethal complications, especially for unrelated transplants. Severe graft-versus-host disease (GVHD) is one of the major causes of treatment related death. To reduce GVHD and to increase engraftment, a fully HLA-matched graft needs to be selected. For this purpose, the genotypes of the patient and potential donors are determined. Scientists found that many of the more than 400 genes have immune-related functions, and haplotypes that share the same HLA alleles may also share discrete blocks of highly conserved sequences in linkage disequilibrium with those HLA alleles. HLA haplotype analysis has be considered in a HCT study (Petersdorf et al. 2007).

For the HCT data, the donor and patient HLA-genotypes are fully observed and matched, but haplotypes are unknown in current practice, since standard genotyping techniques cannot distinguish the two homologous chromosomes of an individual, thus they cannot determine the haplotype pairs, that is the specific sequence of nucleotides on the chromosomes. In a fully HLA-matched unrelated transplant, the donor and patient have identical genotypes, but may not necessarily have matched haplotype pairs. For example, consider HLA loci A and B. Suppose the donor and patient have identical genotypes of $G = \{A = (1, 3), B = (7, 8)\}$. There are two potential haplo-types that are consistent with this genotype, namely $(h_1, h_2) = ((A1, B7), (A3, B8))$ and $((A1, B8), (A3, B7))$. When the donor and the patient have identical haplotypes, then they are haplotype matched in addition to being genotype matched. Otherwise, they are haplotype mismatched. With more HLA loci considered the number of possible haplotypes corresponding to a single genotype is much larger; Petersdorf et al. (2007) give a detailed example with three loci A, B and DRB1. Laboratory techniques have been used to determine haplotypes, but these methods required for HLA haplotyping are technically complex and not easily adaptable for routine diagnostic use. The method previously described by Petersdorf et al. (2007) requires the manual purification of intact, long strand (2 Mbp-long) DNA, construction of probe-based DNA arrays, and hybridization of long strand DNA to solid phase probes. These are very complex techniques that are not easily adaptable to routine diagnostic testing and would be challenging even for advanced research laboratories, and often these methods are cost-prohibitive (Fallin and Schork 2000). The existence of a haplotype matching effect on transplant survival outcomes among fully HLA-genotype matched unrelated transplants has not been fully investigated.

In HCT studies, aGVHD (acute GVHD) and death without aGVHD are two competing risks, and cancer recurrence (relapse) and treatment-related-mortality (TRM, defined as death without relapse) are another pair of commonly studied competing risks. It has not been fully investigated whether there is a haplotype matching effect on cumulative incidence function of aGVHD, TRM and relapse among those HLA-genotype fully matched unrelated transplants. If a positive beneficial haplotype matching effect can be identified then patient and physician should search for a haplotype matched donor among those available to improve the transplant result. It has been shown that the probability of a HLA-haplotype match is around 80 % among the HLA-genotype matched patients and donors in the HCT setting (Petersdorf et al. 2007).

The Petersdorf et al. (2007) analysis was based on observing the haplotype directly and was carried out as a simple logistic regression analysis or a simple Cox regression analysis. They showed that haplotype mismatched HCT had a higher aGVHD rate and a lower relapse rate, and had no effect on TRM and overall mortality. Their study was based on a relative small sample size with 246 cases. This is the only available study for the haplotype matching effect. Their results have not been confirmed by others. Recently, Scheike et al. (2008) proposed using inverse weighting technique to directly model the cumulative incidence function and proposed a class of flexi-ble regression models. In this paper we study the haplotype matching effect for the competing risks data only based on HLA-genotype data, this makes it possible to directly use the vast amounts of available data for studying the possible effects of haplotype match. We also stress that there are many other related blood and marrow transplant treatments that have similar issues, and where the methods we develop here can be used. We here develop our techniques for the HCT study, and consider a similar cohort of patients with HLA-genotype identical unrelated HCT for leukemia. The data was selected from the center for international blood and marrow transplant research (CIBMTR). More generally our techniques can make use of already available data and lead to new knowledge in the field. Our analysis indicates that haplotype matching has same direction of effects as in the Petersdorf, but are non significant for all events. To further verify our conclusion, we consider a simulation study, which shows that proposed method works well.

We have in earlier work assessed the effect of haplotype match on the overall survival in HCT studies using missing data techniques for hazard estimation (Scheike et al. 2010, 2011). Also Flanders et al. (2005) considered testing of haplotype association on the hazard scale based on genotype data. These techniques can not be directly applied in a competing risks setting because the competing risk may also depend on the unobserved haplotye match thus leading to dependent censoring if the cause specific hazard is estimated. We also stress that we here consider the cumulative incidence regression models that models the absolute risk of experiencing the different causes of death in the competing risks setting and that this quantity depends on the underlying cause specific hazards for all causes.

The paper is structured as follows. In Sect. 2 we develop a model and show how one can estimate the parameters of the model and derive the asymptotic results that can be used for inference. Simulation results to examine the small sample performs are presented in Sect. 3. Section 4 contains a worked example based on HCT data. Finally, Sect. 5 contains some discussion.

2 Model specification

Let $\tilde{T_i}$ and C_i be the event time and right censoring time of the i th individual and let $\tilde{\varepsilon_i}\ \varepsilon\ \{1, 2\}$ denote the failure type. Let $N_i(t) = \mathcal{I}(\tilde{T_i} \le t, \tilde{\varepsilon_i} = 1)$ be the underlying counting process associated with cause 1 and define the indicator $\Delta_i = \mathcal{I}(\tilde{T_i} \le C_i)$ that is one when the observation is uncensored. Note that $N_i(t)$ are not fully observed for censored individuals. However, the observed counting process, $\{\Delta_i N_i(t)\}$ are computable for all t. In addition to the risk covariates (X_i, Z_i), we also observe the common genotype of the patient and donor, that we denote G_i. We assume that we observe n independent identically distributed (*i.i.d.*) replications of $\{(T_i, \varepsilon_i, X_i, Z_i, G_i), i = 1,\ldots, n\}$, where $T_i = \min(\tilde{T}_i, C_i)$, $\varepsilon_i = \tilde{\varepsilon}_i \Delta_i$, $\mathrm{X}_i = (1, X_{i,1},\ldots, X_{i,p})^\top$, and $Z_i = (Z_{i,1},\ldots, Z_{i,q})^\top$. We assume that $(\tilde{T}_i, \tilde{\varepsilon}_i)$ are independent of C_i given covariates of (X_i, Z_i, G_i).

Let $H_d = (H_{d1}, H_{d2})$ and $H_p = (H_{p1}, H_{p2})$ be the underlying unobserved haplotype pairs for the donor and patient, respectively. To assess the effect of haplo-type matching, we consider various regression models for the cumulative incidence function on the form

$$g\{1-P_1(t;H_d, H_p, X, Z)\}=\eta(t)^\top X(H_d, H_p)+g_1\{\gamma, Z(H_d, H_p), t\}, \quad (1)$$

for a known link function g and a known regression function g_1.

These flexible models allow some covariates, $X(H_d, H_p)$, to have time-varying effects and other covariates, $Z(H_d, H_p)$, to have constant effects. This distinction becomes very important for the latter application to the HCT study data. Model (1) contains the Fine and Gray (1999) proportional subdistribution hazards model as a special case and Scheike and Zhang (2008) used this model to provide goodness-of-fit procedures for model identification and to check whether a specific covariate has a time-varying effect. The proposed models rely on the unobserved haplotype pairs for the patient and donor. The specific covariate designs we have in mind are to model the effect of matching haplotype nonparametrically, by letting $x(H_d, H_p) = \{1, x, I(H_d = H_p)\}$, or parametrically by letting $z(H_d, H_p) = \{z, I(H_d = H_p)\}$.

Commonly used “cloglog” and “log” link functions can be considered here, which lead to a semiparametric multiplicative model

$$\mathrm{cloglog}[1-P_1\{t;X(H_d, H_p), Z(H_d, H_p)\}]=\eta(t)^\top X(H_d, H_p)+\gamma^\top Z(H_d, H_p) \quad (2)$$

and a semiparametric additive model

$$-\log[1-P_1\{t;X(H_d, H_p), Z(H_d, H_p)\}]=\eta(t)^\top X(H_d, H_p)+\{\gamma^\top Z(H_d, H_p)\}t, \quad (3)$$

respectively.

When the haplotype model that links the observed genotype to the haplotype frequencies can be identified, then we can estimate these parameters consistently and derive the large sample properties.

To model the haplotype matching effect based on observed genotype data, we also need a model that relates the observed genotypes to the underlying haplotypes. With the haplotype pair $H = (h_1, h_2)$, we assume Hardy-Weinberg equilibrium such that

$$P\{H=(h_i, h_j)\}=\pi_i\pi_j,$$

where π_i, $i = 1,.., K$ gives the frequencies of the considered haplotypes. Based on this we can infer the conditional distribution of the underlying haplotypes given the observed genotype, $P(H = h|G = g)$. Specifically, $P\{H = (h_i,h_j)|G = g\} = P\{H = (h_i, h_j: h_i * h_j = g)\}/P(G = g)$, where $h_i * h_j = g$ means that the haplotypes h_i and h_j are consistent with observed genotype g, and $P(G = g) = \Sigma h_i, h_j : h_i * h_{j = g} P\{H = (h_i, h_j)\}$. In the HCT data we have a very large haplotype space (with many haplotypes) and limited observed data. We therefore make further simplifying assumptions about the haplotype frequencies letting several rare haplotypes share common frequencies in our model specifications concerning the haplotype frequencies that follows in the next section.

For the fully observed covariate $V = (X, Z, G)$, the cumulative incidence function has the form:

$$\begin{aligned} P_1(t;X,Z,G) &= P(T \le t, \varepsilon=1|X,Z,G) \\ &= \sum_{h_d,h_p:h_p*h_d \in G} P(H_d=h_d, H_p=h_p|X,Z,G) P_1(t;X,Z,h_d,h_p) \\ &= \sum_{h_d,h_p:h_p*h_d \in G} P(H_p=h_p|G) P(H_d=h_d|G) P_1(t;X,Z,h_d,h_p) \end{aligned}$$

where $h_p * h_d \, \varepsilon \, G$ if $h_{p1} * h_{p2} = h_{d1} * h_{d2} = g \, \varepsilon \, G$. The last equality follows by the assumption that the risk covariates do not affect the haplotype distribution given the genotype, and the patient and donor have independent haplotypes given G. The last assumption is consistent with the fact that the donor and patient are unrelated, but for related donor and patient this part of the expression must be changes appropriately.

To estimate the parameters of the underlying model (1) we consider an inverse probability censoring weighting method as used in Scheike et al. (2008). This is based on the fact that

$$\begin{aligned} E\left\{\frac{\Delta N(t)}{S_C(T|X,Z,G)}|X,Z,G\right\} &= E\left[E\left\{\frac{\Delta N(t)}{S_C(T\left|X,Z,G\right)}|T,\varepsilon,Z,X,H_d,H_p\right\}|X,Z,G\right] \\ &= E[E\{P_1(t;X,Z,H_d,H_p)\}|X,Z,G] \\ &= \sum_{h_d,h_p:h_d*h_p \in G} P(H_p=h_p|G,X,Z) P(H_d=h_d|G,X,Z) P_1(t;X,Z,h_d,h_p) \\ &= P_1(t;X,Z,G), \end{aligned}$$

where $S_C(t|X_i, Z_i, G_i)$ is the survival distribution for the censoring time. We will for simplicity in the presentation assume that the censoring distribution does not depend on any of the observed covariates, denoted as $S_C(t)$.

These calculations show how the basic principle for modelling the missing data relies on computing the conditional mean of the missing covariate, (H_d, H_p), given the observed data (X, Z, G). It is evident how our approach more generally deal with missing covariates in the context of cumulative incidence estimation, but we here focus on the haplotype match problem that has been motivating our developments.

2.1 Estimation

We suggest a simple and robust two-stage procedure for estimation, where we first estimate the haplotype-probabilities and then use these estimated frequencies for estimating the regression parameters given in model (1) for the cumulative incidence functions. This will lead to some loss of efficiency but keeps things simple, and in our experience this loss of efficiency will typically be minor.

To estimate the haplotype parameters we consider the log-likelihood of the genotype data $\Pi_i P(G_i = g_i)$. We consider the logistic regression model for the haplotype frequencies,

$$\pi_j=\frac{\exp(\alpha_j)}{\exp(\alpha_1)+\cdots+\exp(\alpha_{K-1})+1}$$

where $(\alpha_1,\ldots,\alpha K_{-1})^\top = X\theta$, and X is a design matrix of size $(K-1) \times m$, with $m \le K-1$ and $\theta = (\theta_1,\ldots,\theta_m)^\top$

Let $\tilde{U}\theta(\theta) = \Sigma_i\ \tilde{U}_{\theta,i}(\theta)$, where $\tilde{U}_{\theta i}(\theta)$is the score of the log-likelihood for the i th subject. $\hat{\theta}$ is estimated by solving $\tilde{U}_\theta(\hat{\theta}) = 0$. By standard asymptotic theory of the MLE, $\sqrt{n}(\hat{\theta}-\theta)$ is asymptotically normal and asymptotically equivalent to

$$\sqrt{n}\sum_i U_{\theta,i}(\hat{\theta}),$$

where $U_{\theta,i}(\theta) = \mathcal{I}_\theta(\theta)\}^{-1}\ \tilde{U}_{\theta,i}(\theta)$, and $\mathcal{I}_\theta(\theta) = -\partial\tilde{U}_\theta(\theta)/\partial\theta.$

Now, given $\hat{\theta}$we can solve score equations for $\eta(t)$ and γ simultaneously using inverse-censoring weighted technique. Let $P_1^{(i)}(t,\eta(t),\gamma,\theta)$ be the $n \times 1$ vector of $P_1(t; x_i, z_i, g_i)$ for $i = 1,\ldots, n$, let $R(t)$ be the $n \times 1$ vector of adjusted responses $\Delta_i N_i(t)/\hat{S}_C(T_i)$, and let $D_\eta(t, \eta(t),\gamma,\theta)$ and $D_\gamma(t,\eta(t),\gamma, \theta)$ be matrices with with the i th rows equal to $D_{\eta,i}(t,\eta(t),\gamma, \theta) = \partial P_1(t; x_i, z_i, g_i)/\partial\eta(t)$ and $D_{\gamma,i}(t, \eta(t),\gamma) = \partial P_1(t; x_i, z_i, g_i)/\partial\gamma$, respectively. Define similarly $D_\theta(t, \eta(t),\gamma, \theta)$ as the matrix with i th row equal to $\partial P_1(t; x_i, z_i, g_i)/\partial\,\theta.$

The regression functions $\eta(t)$ and regression parameters γ can be estimated based on the following estimation equations for fixed θ:

$$U_\eta(t,\eta(t),\gamma,\theta)=D_\eta^\top(t,\eta(t),\gamma,\theta)\,\{R(t)-P_1(t,\eta(t),\gamma,\theta)\}=0, \quad (4)$$

$$U_\gamma(\tau,\eta(\cdot),\gamma,\theta)=\int_0^\tau D_\gamma^\top(t,\eta(t),\gamma,\theta)\,\{R(t)-P_1(t,\eta(t),\gamma,\theta)\}\,dt=0, \quad (5)$$

where τ is the last event time point. Note that the estimates of $\eta(t)$ will be piecewise constant functions that change their value only after events of type one, so we only need to consider the score equations for $\eta(t)$ in the jump times.

The large sample properties in the case of known haplotype parameters, θ, follows the developments of Scheike et al. (2008) and Scheike and Zhang (2008). We show that how the asymptotics are changed due to the additional uncertainty that comes from the estimates of the haplotype frequencies.

For known θ, we have shown that the distributions of $\sqrt{n}\{\hat{\eta}(t)-\eta(t)\}$ and $\sqrt{\eta}(\hat{\gamma}-\gamma)$ are asymptotically equivalent to the following i.i.d. decomposition of

$$\sqrt{n}\{\mathscr{I}_\eta(t)\}^{-1}\sum_i W_{i1}(t)\quad\text{and}\quad\sqrt{n}\{\mathscr{I}_\gamma\}^{-1}\sum_i W_{i2},$$

respectively, where

$$\begin{aligned}\mathscr{I}_\eta(t)&=D_\eta(t)^\top D_\eta(t)\\ \mathscr{I}_\gamma&=\int_0^\tau D_\gamma(t)^\top H(t)D_\gamma(t)dt\\ H(t)&=I-D_\eta(t)\{\mathscr{I}_\eta(t)\}^{-1}D_\eta(t)^\top,\end{aligned}$$

and detailed formula of (W_{1i} (t), W_{2i}) and its consistent estimates ($\hat{W}_{1i}$ (t), $\hat{W}_{2i}$) are given in Scheike et al. (2008) and Scheike and Zhang (2008).

For unknown θ, based on a two-stage approach using the MLE $\hat{\theta}$ and under regularity conditions, it therefore can be shown that $\sqrt{n}\{\hat{\eta}(t)-\eta(t)\}$ and $\sqrt{\eta}(\hat{\gamma}-\gamma)$ are jointly asymptotically Gaussian, both zero mean, and with distributions that are asymptotically equivalent to

$$\sqrt{\eta}\{\mathscr{I}_\eta(t)\}^{-1}\sum_i W_{\eta,i}(t)\quad\text{and}\quad\sqrt{\eta}\{\mathscr{I}_\gamma\}^{-1}\sum_i W_{\gamma,i},$$

respectively, where

$$\begin{aligned}W_{\eta,i}(t)&=W_{i1}(t)-\mathscr{I}_\eta^{-1}(t)D_\eta^T(t)D_\theta(t)U_i(\theta_0),\\ W_{\gamma,i}&=W_{i2}-\left\{\mathscr{I}_\gamma^{-1}\int_0^\tau D_\gamma(t)^\top H(t)D_\theta(t)dt\right\}U_i(\theta_0).\end{aligned}$$

Let $\hat{\mathscr{I}}_\eta$ (t), $\hat{W}_{\eta,i}$ (t), $\hat{\mathscr{I}}_\gamma$ and $\hat{W}_{\gamma i}$ be the estimators of their population counter parts by plugging estimates of all needed quantities. The asymptotic distributions of $\sqrt{n}\,\{\hat{\eta}(t)-\eta(t)\}$ and $\sqrt{n}(\hat{\gamma}-\gamma)$ are asymptotically equivalent to the following conditional multiplier version of the i.i.d. decompositions $\sqrt{\eta}\{\widehat{\mathscr{I}_\eta}(t)\}^{-1}\sum_{i=1}^n E_i\widehat{W}_{\eta,i}(t)$ and $\sqrt{\eta}\{\widehat{\mathscr{I}_\gamma}\}^{-1}\sum_{i=1}^n E_i\widehat{W}_{\gamma,i}$, respectively, where $E_1,.., E_n$ are i.i.d. standard normals. It follows that the asymptotic variance of $\sqrt{\eta}\{\hat{\eta}(t)-\eta(t)\}$ and $\sqrt{n}(\hat{\gamma}-\gamma)$ can be estimated consistently by

$$\begin{aligned}\widehat{\sum}_\eta(t)&=n\left\{\widehat{\mathscr{I}_\eta}(t)\right\}^{-1}\left[\sum_{i=1}^n\left\{\widehat{W}_{\eta,i}(t)\right\}^{\otimes 2}\right]\left\{\widehat{\mathscr{I}_\eta}(t)\right\}^{-1},\\ \widehat{\sum}_\gamma&=n\widehat{\mathscr{I}}_\gamma^{-1}\left[\sum_{i=1}^n\left\{\widehat{W}_{\gamma,i}\right\}^{\otimes 2}\right]\widehat{\mathscr{I}}_\gamma^{-1},\end{aligned}$$

respectively, where $a^{\otimes 2} = aa^\top$.

A (1 - α) × 100 % asymptotic confidence band for $\hat{\eta}$(t) over a fixed time interval can be constructed using resampling technique. These resampling results can be used to construct confidence band for the predicted cumulative incidence function as in Scheike et al. (2008).

3 Simulation study

To investigate the fixed sample properties we did a simulation study that mimics the data example analyzed in Sect. 4. We considered 3,712 patients with genotypes equivalent to the observed ones. Based on the haplotype frequencies described earlier we then sampled haplotype pairs consistent with the observed genotypes for both patient and donor, and then simulated a cumulative incidence function of additive form with approximately 815 events of interest for different levels of the effect of haplotype match. We generate data from a semiparametric multiplicative model (2),

$$\text{cloglog}[1-P_1\{t;X(H_d,H_p),Z(H_d,H_p)\}]=\eta_0(t)+\gamma I(H_d=H_p),$$

with $\gamma = \{-0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5\}$. Observed coverage probabilities are reported in Table 1.

The results of the simulations are given in Table 1. We see that the estimator is almost unbiased for all considered levels of the effect size. The variance is well estimated and the coverage is close the nominal level. All in all we conclude that the finite sample performance is quite good in a sample size similar to the one in the considered data.

4 HCT patients

Petersdorf et al. (2007) studied the effect of MHC haplotype match based on 246 leukemia patients who received a HLA-genotype fully matched unrelated HCT from 1986 to 2003. Their analysis was based on observing the haplotypes directly. This is a very costly and time-consuming procedure (Fallin and Schork 2000). They identi-fied 191(78 %) and 55(22 %) transplants were haplotype matched and miss-matched, respectively. They demonstrated that haplotype matched transplant had a lower incidence rate of grade III–IV aGVHD (odds-ratio=0.22, $p < 0.0001$) and a higher cancer relapse rate (hazard-ratio=2.22, $p = 0.03$), but had no impact on the TRM and overall mortality. In this study, we show how modeling of the missing haplotype data can also be used to address these issues. For the illustration purpose, a similar transplant patient cohort was selected from the statistical center of the center for international blood and marrow transplant research (CIBMTR). The analysis has not been reviewed or approved by the Advisory or Scientific Committee of the CIBMTR. The CIBMTR is comprised of clinical and basic scientists who confidentially share data on their blood and bone marrow transplant patients with CIBMTR Data Collection Center located at the Medical College of Wisconsin. The CIBMTR is a repository of information about results of transplants at more than 450 transplant centers worldwide.

The example data consists of 3,712 leukemia patients (1,822 for acute myeloid leukemia (AML), 982 for acute lymphoblastic leukemia (ALL) and 908 for chronic myelogenous leukemia (CML)). All patients in the study were HLA-A, B, DRB1 allele matched in high resolution with their donors and transplanted between 1995 and 2007. 1,651; 1,047 and 1,014 patients were transplanted in low, intermediate or high risk of disease statuses, respectively. 647 Males patients received graft from a female donor. 81 % of patients received myeloablative conditioning regimen, 53 % of patients were treated with methotrexale (MTX) + cyclosporin (CsA) ± other or CsA± other for GVHD prophylaxis, and 56 and 44 % of patients received bone-marrow (BM) or peripheral-blood (PB) transplant, respectively.

The genotype data is based on high resolution classification of alleles at three-loci HLA-A, HLA-B and HLA-DRB1. We consider two-stage procedure for this study. First, we need to estimate the haplotype frequencies, π_i or $\theta = (\theta_1,\ldots,\theta_m)^\top$ (see Sect. 2). Two estimation approaches can be considered. One is using MLE method based on current study data. Potentially, we will have a very rich haplotype space. To reduce the total number of possible haplotypes, some additional structure is needed. We here suggest to group the rare frequencies into groups with a common haplotype frequency. Alternatively, we can also use other available data for estimation of the hap-lotype frequencies (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). Clearly using this additional data leads to a more accurate and stable estimation for the haplotype frequencies.

Both methods were used for this HCT study and gave similar results. In this paper, we report the results based on the additional data for the estimation of the haplotype frequencies. The National Marrow Donor Program (NMDP)'s existing potential donor pool as the background population cohort. The NMDP is a nonprofit organization dedicated to creating an opportunity for all patients to receive the bone marrow or umbilical cord blood transplant they need, when they need it. Currently, there are about eight million volunteer adult donors are registered in the NMDP's potential donor pool. It is known that the haplotype frequency is determined by racial category. The donor's and patient's racial categories are fully observed for our study cohort and utilized in our haplotype frequency calculation. Among 3,712 HLA-identical unrelated HCT, the estimated average probability of haplotype matching is 80.7 % which is similar to Petersdorf et al. (2007)'s report.

To excess the haplotype matching effect on aGVHD, TRM, relapse and treatment failure which is defined as TRM or relapse, we fit the multiplicative model

$$\mathrm{cloglog}\{1-P_1(t;H_d,H_p,X,Z)\}=\eta(t)^\top X+\gamma_0 I(H_d=H_p)+\gamma^\top Z. \quad (6)$$

where the first element of X is constant one, and adjusting remaining covariates X of disease type (AML versus ALL versus CML), and covariates Z of patient age ($>$45 versus 19–45 versus ≤18 years old), disease status at pre-transplant (advanced disease versus low or intermediate disease), donor–patient gender match (female to male versus other), conditioning regimen (NST/RIC versus myeloablative), GVHD prophylaxis (MTX+CsA ± Other or CsA±Other versus others), graft type (PB versus BM) and year of transplant (2002–2007 versus 1995–2001). In our study sample, 815(22 %) patients developed grade III–IV aGVHD, 1,488 patients died without aG-VHD and 1,409 patients were censored at end of study. 2,268 Patients were considered as treatment failure in which 1,255 patients died in compete remission (TRM) and 1,013 patients relapsed. Our analysis showed that haplotype match has no effect on aGVHD, relapse, TRM and treatment failure (Table 2).

We also considered the model with non-parametric haplotype match effect where

$$\mathrm{cloglog}\{1-P_1(t;H_d,H_p,X,Z)\}=\eta(t)^\top X+\eta_0(t) I(H_d=H_p)+\gamma^\top Z, \quad (7)$$

where we subsequently performed a resampling test for the constant effect of $H_{0:}$ $\eta_0(t) \equiv \eta_0$ using a supremum Kolmogoroff–Smirnov test. If this test is significant it also suggests that the haplotype match effect is significant with time-varying effect. We stress that the test differs from the test for non-significant effects $H_{0:}$ $\eta_0(t) \equiv 0$, that are not significant for all outcome events. We note that haplotype match has non-significant effect on the probability of aGVHD, relapse, TRM and treatment failure (see Table 2). Figure 1 shows the effect of haplotype match versus haplotype mismatch for TRM, a negative (beneficial) effect for the haplotype matched HCT within first 90 months of transplant and a positive late effect, however, it is not significant since 95 % confidence band contains some straight lines. This is further confirmed by non-significance of the constant effect test ($H_{0:}$ $\eta_0(t) \equiv \eta_0$, for a constant η_0). Constant haplotype matching effects have been observed in all other outcomes. Thus, multiplicative model (6) with constant haplotype matching effect can be considered for all outcomes, which validates the the parametric test from Table 2. This model validation is a critical an important part of such an analysis, and easily done by the developed methodology.

Based on fitted models, we can compute the predicted cumulative incidence functions (CIF) of aGVHD, relapse, TRM, and leukemia-free survival probability for a given set values of the covariates with 95 % confidence intervals and confidence bands. For the illustration purpose, we compute the predicted CIF of TRM by haplotype matched versus miss-matched for a patient with AML disease, transplanted in early or intermediate disease stage, donor–patient gender match of male to male (MM) or (MF) or (FF), received myeloablative conditioning regimen, CsA±Other for GVHD prophylaxis, and bone marrow graft source, and transplanted between 2003 and 2007. Resampling method based on 1,000 realizations was used to construct confidence band (See Fig. 2). Figure 2a shows the predicted CIF based on constant multiplicative model (6). Figure 2b shows the predicted CIF based on the alternative non-parametric haplotype match effect model (7), which is a more flexible model allowing haplotype matching effect change over time. Figure 2b shows that haplotype matched transplant has a lower cumulative incidence of TRM initially and a higher incidence rate of TRM later although this change is not significant.

5 Discussion

We have demonstrated how to assess the haplotype matching effect on the competing risks for hematopoietic cell transplant studies based on modeling of the missing data. This opens up for using the huge amounts of available data for studying detailed aspects of the HLA haplotypes on the outcome for blood and marrow transplants studies without cost-prohibitive laboratory typing for the haplotypes for patient and donor.

Another situation where the effect of haplotype matching is of interest is transplantations using umbilical cord blood (CB) which has recently been accepted as an alternative graft source to bone marrow (BM) for HCT (Eapen et al. 2007). Most CB transplants are mismatched at one or two HLA loci. For HLA-genotype mismatched unrelated transplants, the haplotype pairs of the donor and the patient can be either matched on a single haplotype or mismatched on both haplotypes. For example, with patient HLA-genotype $G_p = \{A = (1, 3), B = (7, 8)\}$ and donor genotype $G_d = \{A = (1, 3), B = (7, 13)\}$ which are mismatched at the B locus, it is possible that they share a single haplotype $(A1, B7)$. As far as we know, the question of whether there is an effect of a single haplotype matching versus both haplotypes being mismatched for HLA-mismatched unrelated transplants has not yet been investigated. The methods developed here can be used to study such questions.

We found no significant haplotype matching effect on grade III–IV aGVHD, relapse and TRM when fitting a model with constant effects. A more careful model examination using the non-parametric models validated these conclusions.

We have implemented the methods in the R-package HaploSurvival that is available from R-forge.

An issue for further research is a further study of the robustness to incorrect modeling of the haplotype distribution. Here it could be of interest to develop robust estimating equation along the lines of Allen and Satten (2005) and Allen et al. (2005), but extending these methods to our setting is not obvious.

Acknowledgments

The research was supported by National Cancer Institute Grant 2 R01 CA54706-10. Thomas Scheike was also supported by a grant from the Danish Research Council on “Point process modeling and statistical inference” and the excellence program at the University of Copenhagen.

References

Allen AS, Satten GA. Robust testing of haplotype/disease association. BMC Genet. 2005; 6(Suppl 1):S69. [PubMed: 16451682]

Allen AS, Satten GA, Tsiatis AA. Locally-efficient robust estimation of haplotype-disease association in family-based studies. Biometrika. 2005; 92:559–571.

Eapen M, Rubinstein P, Zhang M-J, Stevens C, Kurtzberg J, Scaradavaou A, Loberiza FRECR, Klein JP, Horowitz MM, Wagner JE. Outcomes of transplantation of unrelated donor umbilical cord blood and bone marrow in children with acute leukaemia: a comparison study. Lancet. 2007; 369:1947–1954. [PubMed: 17560447]

Excoffier L, Slatkin M. Maximum-likelihood estimation of polecular haplotype frequenceis in a deiploid population. Mol Biol Evol. 1995; 12:921–927. [PubMed: 7476138]

Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet. 2000; 67:947–959. [PubMed: 10954684]

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999; 94:496–509.

Flanders WD, Khoury MJ, Yang QH, Austin H. Test of trait—haplotype association when linkage phase is ambiguous, appropriate for matched case–control and cohort studies with competing risks. Stat Med. 2005; 24:2299–2316. [PubMed: 16015677]

Hawley M, Kidd K. Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered. 1995; 86:409–411. [PubMed: 7560877]

Long J, Williams R, Urbanek M. An EM algorithm and testing strategy for multi-locus haplotypes. Am J Hum Genet. 1995; 56:799–810. [PubMed: 7887436]

Petersdorf E, Malkki M, Gooley T, Martin P, Guo Z. MHC haplotype matcing for unrelated hematopoietic cell transplantation. PLOS Med. 2007; 4:59–68.

Scheike T, Martinussen T, Silver J. Estimating haplotype effects for survival data. Biometrics. 2010; 66:705–715. [PubMed: 19764954]

Scheike T, Martinussen T, Zhang M. The additive risk model for estimation of haplotype effects. Scand J Stat. 2011; 38:409–423.

Scheike TH, Zhang M-J. Flexible competing risks regression modelling and goodness-of-fit. Lifetime Data Anal. 2008; 14:464–483. [PubMed: 18752067]

Scheike TH, Zhang M-J, Gerds T. Predicting cumulative incidence probability by direct binomial regression. Biometrika. 2008; 95:205–220.

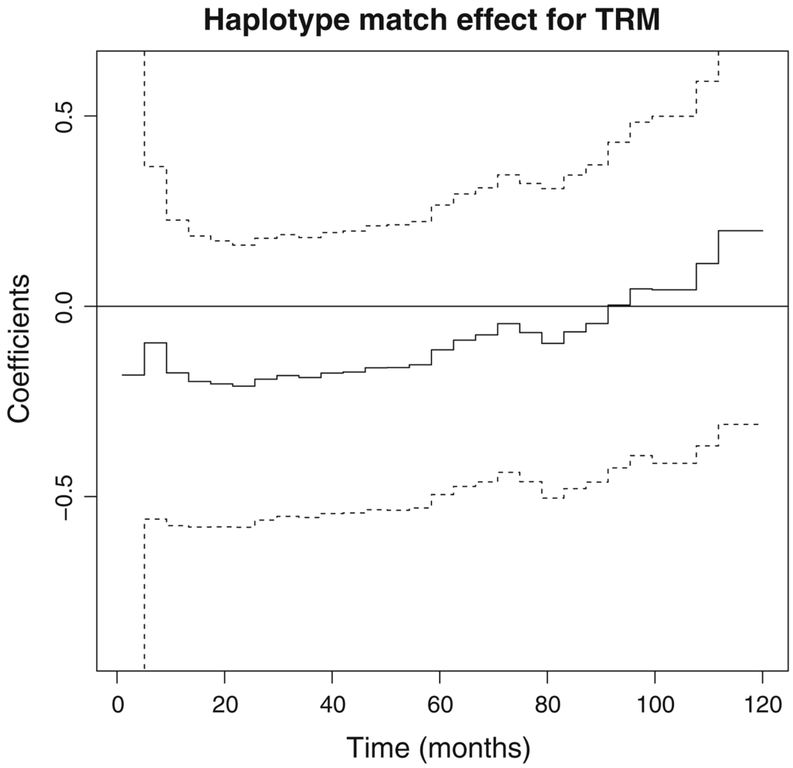

Fig. 1.

Effect of haplo-match for CIF of TRM with corresponding 95 % confidence bands (*dashed lines*)

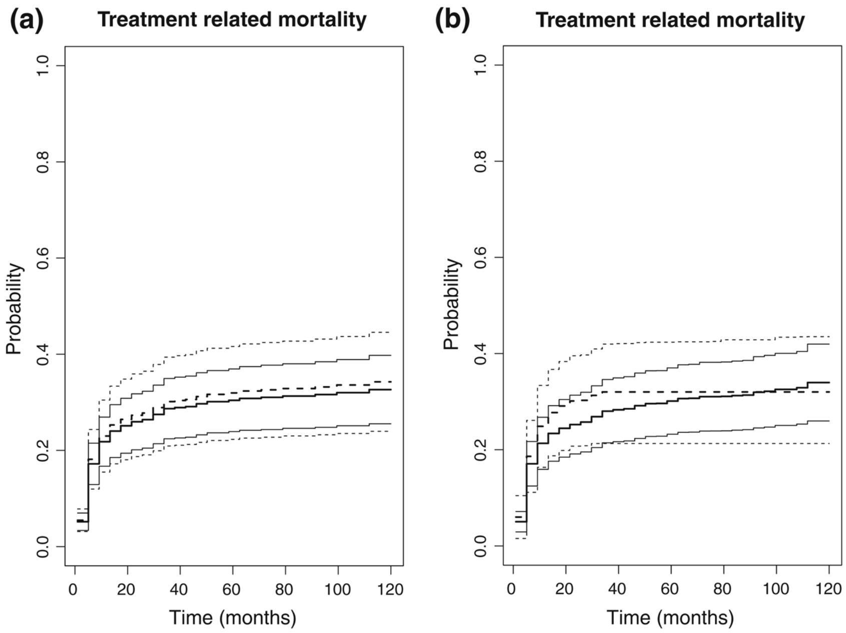

Fig. 2.

Thick solid and *dashed lines* are the CIF of TRM for haplo-matched and haplo-mismatched HCT, respectively. *Light solid* and *dashed lines* represent corresponding 95 % confidence bands. **a** based on parametric model, and **b** non-parametric effect of haplo-match

Table 1

Mean of estimates (mean est.), the standard deviation of estimates (sd. est.), the mean of estimated standard errors (mean SE), and observed coverage of 95 % confidence intervals (coverage %) based on 1,000 realizations for different effect sizes (β)

γ	**Mean est.**	**Sd. est.**	**Mean SE**	**Coverage (%)**
−0.5	−0.470	0.097	0.097	94
−0.3	−0.277	0.063	0.065	93
−0.1	−0.089	0.049	0.051	96
0.0	0.000	0.041	0.041	95
0.1	0.093	0.050	0.050	94
0.3	0.285	0.065	0.063	95
0.5	0.479	0.087	0.088	94

Table 2

Adjusted effect of HLA-A, -B, -CRB1 haplotype matching of risks of grades III–IV aGVHD, relapse, TRM and treatment-failure after HCT from HLA-identical unrelated donors

Outcome	**exp($\hat{\gamma}_0$)**	**95 % CI**	***P*-value**	**Nonparametric test *P*-value**
aGVHD III–IV	0.98	0.69–1.38	0.98	0.46
Relapse	1.03	0.76–1.41	0.83	0.94
TRM	0.93	0.69–1.24	0.62	0.27
Treatment-failure	0.99	0.77–1.25	0.91	0.15