Published in final edited form as:

Scand Stat Theory Appl. 2011 September ; 38(3): 409–423. doi:10.1111/j.1467-9469.2010.00720.x.

The additive risk model for estimation of effect of haplotype match in BMT studies

Thomas H. Scheike

Department of Biostatistics, University of Copenhagen

Torben Martinussen

Department of Biostatistics, University of Southern Denmark

Mei-Jie Zhang

Division of Biostatistics, Medical College of Wisconsin

Summary

In this paper we consider a problem from bone marrow transplant (BMT) studies where there is interest on assessing the effect of haplotype match for donor and patient on the overall survival. The BMT study we consider is based on donors and patients that are genotype matched, and this therefore leads to a missing data problem. We show how Aalen's additive risk model can be applied in this setting with the benefit that the time-varying haplo-match effect can be easily studied. This problem has not been considered before, and the standard approach where one would use the EM-algorithm cannot be applied for this model because the likelihood is hard to evaluate without additional assumptions. We suggest an approach based on multivariate estimating equations that are solved using a recursive structure. This approach leads to an estimator where the large sample properties can be developed using product-integration theory. Small sample properties are investigated using simulations in a setting that mimics the motivating haplo-match problem.

Keywords

Aalen model; additive risk model; estimating equation; flexible modeling; haplo-match effects; haplo-type effects; missing data; semiparametric modeling; survival data

1 Introduction

Hematopoietic cell transplantation (HCT) is a life saving procedure for many cancer patients, but it also has side effects, especially for transplants from unrelated donors, for example when the donor is not from the patient's family. Graft-versus-host disease (GVHD) complications is one of the major causes for treatment related death. To reduce GVHD, human leukocyte antigen (HLA) fully matched graft needs to be selected. When a patient does not have an HLA-identical sibling donor, then an HLA fully matched unrelated transplant often will be considered. For an HLA fully matched unrelated transplant, the donor is selected such that the patient and donor have identical HLA genotypes, but they may not necessarily have matching haplotype pairs. For example, suppose the donor and patient have identical HLA-A/B/DRB1 alleles of $\{(A1, A3), (B8, B7), (DR3, DR2)\}$. Now, there are four potential haplotypes that are consistent with this genotype, namely $\{(A1, B8,$

DR3), (A3, B7, DR2)}, {(A3, B8, DR3), (A1, B7, DR2)}, {(A1, B8, DR2), (A3, B7, DR3)}, and {(A3, B8, DR2), (A1, B7, DR3)}. Here, a donor with {(A1, B8, DR3), (A3, B7, DR2)} and a patient with {(A1, B8, DR3), (A3, B7, DR2)} are haplotype matched, but a donor with {(A3, B8, DR3), (A1, B7, DR2)} and a patient with {(A1, B8, DR3), (A3, B7, DR2)} are haplotype miss-matched, see Petersdorf et al. (2007) for details. Among HLA-genotype matched transplants with unrelated donors it has not been fully investigated whether there is an effect of haplotype matching. This is an important question since if a positive beneficial haplotype matching effect can be identified, then patient and physician need to search for a haplotype matched donor among those available HLA-genotype fully matched potential donors. Among a patient and donor with identical HLA genotype, the typical probability of a HLA haplotype match is around 70% (Petersdorf et al., 2007).

Recently, Petersdorf et al. (2007) studied the effect of major histocompatibility complex (MHC) haplotype match of donor and patient after HLA-identical unrelated HCT for leukemia. Their analysis was based on observing the haplotype pairs directly and therefore was carried out as a Cox regression analysis. Identifying the haplotype pairs is a very costly, time-consuming and not entirely unproblematic procedure. In this study, we show how modeling of the missing haplotypes for donor and patient can also be used to address the issue using only the observed genotype and failure time data. For illustration purpose, we consider a similar cohort of patients with HLA-identical unrelated HCT for acute leukemia. The data was selected from the Center for International Blood and Marrow Transplant Research (CIBMTR). We found a strong time-varying effect for the haplotype matching, and therefore the use of Cox's proportional hazards model is not appropriate to assess the effect of haplo-type match for this transplant data. We instead consider Aalen's additive intensity model (Aalen, 1989). This model allows the haplotype matching effect to be varying over time and the time-dependent excess risk can be estimated directly. This excess risk translates directly into a simple relative survival percentage. One particular advantage of this approach is that the absolute size of the effect is directly accessible and that the model does not rely on any particular structure for the effect.

Missing data for survival analysis has been considered in detail in many different settings for Cox's regression model. For example with missing partial covariates for some patients and in frailty models for unobserved cluster effect. Estimating haplotype effects using Cox's regression model has been considered in a large number of papers (Schaid et al., 2002; Zaykin et al., 2002; Schaid, 2004; Zhao et al., 2003; Tregouet & Tiret, 2004; Lin & Zeng, 2006; Epstein & Satten, 2003; Chen et al., 2004; Scheike et al., 2010). One of the attractions for fitting a Cox's regression model is that the EM-algorithm can be applied in this setting. This makes it easy to deal with the missing data. Alternative estimating equations have been considered in Zucker (2005); Gorfine et al. (2005); Scheike (2006); Scheike et al. (2010).

There has been no related statistical work on how to fit an additive intensity model in this setting where no simple likelihood is available. In this paper, we suggest a new approach that fits the additive risk model without smoothing or additional assumptions. The proposed method is simple to implement and use. The basic idea is to solve the estimating equations recursively to deal with the structure of the model appropriately. It turns out that the large sample properties of the proposed estimators can be derived based on product integration theory (Gill & Johansen, 1990).

The model and estimating procedures are presented in section 2. Section 3 gives large sample properties and also a test for investigating time-invariance. Small sample performs are assessed by simulation studies that are presented in section 4. In section 5 we consider a worked example that assesses the effect of haplotype matching for HCT patients. Section 6 concludes with some final remarks. Proofs are collected in the Appendix.

2 The additive risks model for effect of haplo-match

Let $\tilde{t}$ be the failure time of interest and let C be a potential censoring time such that we observe $T = \tilde{t} \wedge C$ and $\delta = I(\tilde{t} \leq C)$. Let $N(t) = I(T \leq t, \delta = 1)$. We assume that, given covariates X and observed haplotype pairs for the donor $H_d = (H_{d1}, H_{d2})$ and the patient $H_p = (H_{p1}, H_{p2})$, the counting process $N(t)$ has an additive intensity

$$\lambda(t) = R(t) \lambda(t; H_d, H_p, X) = R(t) \{X(H_d, H_p)^T \alpha(t)\} \quad (1)$$

where $R(t) = I(T \geq t)$ indicates individual is at risk at time t , and $X(H_d, H_p)$ is a function of the observed covariates and haplotypes of donors and patients. In this paper, we consider the design where $X(H_d, H_p)^T = \{X^T, I(H_d = H_p)\}$. This simple design gives the effect of haplotype matching between donor and patient.

If H_d and H_p are known, a simple standard survival analysis based on the additive model can be used. However, the problem here is that we only observe the underlying HLA genotypes related to H_d and H_p . For HLA-identical unrelated transplant data, the patient and donor have identical genotypes, but may not necessarily have matched haplotype pairs (see section 1).

To estimate the effect of haplo-match we need to model how the haplotype pairs relate to the observed genotypes. We assume that the probability of a specific haplotype pair for both donor and patient $H = (h_i, h_j)$ can be computed under Hardy-Weinberg equilibrium as

$$P\{H = (h_i, h_j)\} = \pi_i \pi_j \quad (2)$$

where $\pi_i, i = 1, \dots, K$ gives the distribution of the considered haplotypes. One may consider any relevant parametric model for the frequencies of the haplotype pairs. Explicitly, we parameterize the haplotype frequencies as

$$\pi_j = \frac{\exp(\eta_j)}{\exp(\eta_1) + \dots + \exp(\eta_K)}$$

with $\eta_j \in]-\infty, \infty[$ and with the identifiability restriction $\eta_K = 0$ ($\sum \pi_i = 1$). We denote the haplotype parameters as $\eta = (\eta_1, \dots, \eta_{K-1})$. For the HCT data, these parameters are unknown and must be estimated from the data, but one may also search known data-bases to get estimates of these parameters. We here treat the haplotype parameters as unknown. In the HCT data, we have a very rich haplotype space. Thus, a very large number of parameters need to be estimated and since we only have limited observed data, we will make further simplifying assumptions about the haplotype frequencies (see section 5).

Based on this model we can infer the conditional distribution of the underlying haplotype pairs given the observed genotype, $P(H = h/G = g)$. The observed intensity given genotype can be written as

$$\begin{aligned}
\lambda(t; G, X) &= \frac{\sum_{h_d, h_p \in G} P(H_d = h_d, H_p = h_p | G) \exp\left\{-\int_0^t \lambda(s; h_d, h_p, X) ds\right\} \lambda(t; h_d, h_p, X)}{\sum_{h_d, h_p \in G} P(H_d = h_d, H_p = h_p | G) \exp\left\{-\int_0^t \lambda(s; h_d, h_p, X) ds\right\}} R(t) \\
&= \frac{\sum_{h_d, h_p \in G} P(H_d = h_d, H_p = h_p | G) \exp\left\{-\int_0^t \lambda(s; h_d, h_p, X) ds\right\} R(t) X(h_d, h_p)^T}{\sum_{h_d, h_p \in G} P(H_d = h_d, H_p = h_p | G) \exp\left\{-\int_0^t \lambda(s; h_d, h_p, X) ds\right\}} \alpha(t) \\
&= R(t) \frac{T(t, v, G, X)^T}{N(t, v, G, X)} \alpha(t) = R(t) f\{v, A(t^-)\}^T \alpha(t),
\end{aligned}$$

where $A(t) = \int_0^t \alpha(s)$ and $\sum_{h_d, h_p \in G}$ is the sum over all haplotype pairs that are consistent with the observed genotype. For unrelated donors and patients, we can assume that $P(H_d = h_d, H_p = h_p | G) = P(H_d = h_d | G) \times P(H_p = h_p | G)$. For related donors and patients this independence assumption does not hold. Here, the observed hazard function has an additive structure and it also depends on the cumulated regression function evaluated at time t^- , i.e. just before time t .

The likelihood for a subject is

$$P(G) \times \lambda(T; G, X)^\delta \exp\left\{-\int_0^T \lambda(s; G, X) ds\right\}$$

and the score equation for v can be partitioned into two parts

$$\tilde{U}(v) = U^G(v) + U^N(v)$$

where U^G is the part of the score that relates to $P_v(G)$ and U^N relates to the information contained in the survival data.

For the nonparametric additive structure of the hazard function, the likelihood is hard to work with even when haplotype pairs, H , are fully observed. Therefore, we consider an alternative approach for estimating the nonparametric regression functions. We assume that n i.i.d. observations are available in the time-interval $[0, \tau]$, for some $\tau < \infty$. Let $N(t) = \{N_1, \dots, N_n(t)\}^T$ be the n -dimensional counting process of all subjects with corresponding intensity $\lambda(t) = \{\lambda_1(t), \dots, \lambda_n(t)\}^T$.

We first note that estimation can be done recursively. Assuming for the moment that v is known and $\hat{A}(t) = 0$ for all t smaller than the first jump-time. The intensity of $N_i(t)$ can be written as $\tilde{Y}_i\{t, A(t^-)\}^T \alpha(t)$, where $\tilde{Y}_i\{t, A(t^-)\} = R_i(t) f_i\{v, A(t^-)\}$. This leads to a least square estimator

$$\hat{A}(t) = \int_0^t \tilde{Y}^- \{s, \hat{A}(s^-)\} dN(s),$$

where the i th row of $\tilde{Y}\{t, A(t^-)\}$ is $\tilde{Y}_i\{t, \hat{A}(t^-)\}$ and $\tilde{Y}^-(\cdot)$ is the generalized weighted inverse of $\tilde{Y}(\cdot)$, which is given by

$$(\tilde{Y}^T W \tilde{Y})^{-1} \tilde{Y}^T W.$$

Here the diagonal elements of the weight function, W are $N(t, v, G_i, X_i)/T(t, v, G_i, X_i)$. The efficient weight is equivalent to the efficient weight in the one-dimensional case, which involves inverse of $X(h_d, h_p)^T \alpha(t)$, and it can be implemented using a two-stage approach as for the additive risk model (Huffer & McKeague, 1991).

Next, we look at an estimating equation for v based on the survival data. We first need some definitions. Let the i th row of $\tilde{Z}(t)$ be $R_i(t) \tilde{Z}_i(t, v, G, X)$ where

$$\tilde{Z}_i(t, v, G, X) = \left[\frac{D_v \mathcal{T}(t, v, G_i, X_i)}{\mathcal{T}(t, v, G_i, X_i)} - \frac{D_v N(t, v, G_i, X_i)}{N(t, v, G_i, X_i)} \right] R_i(t)$$

is the derivative of the intensity with respect to v multiplied by the weight described above. Here $D_v h(x)$ denotes the derivative with respect to v evaluated at x .

This leads to an estimating equation for the unknown v based on the survival data

$$\begin{aligned} U^N(v, A) &= \int_0^\tau \tilde{Z}^\top(t) \{dN(t) - \lambda(t; G, X) dt\} \\ &= \int_0^\tau \tilde{Z}^\top(t) \{dN(t) - \tilde{Y}\{t, A(t^-)\} dA(t)\}. \end{aligned}$$

We now replace $A(t)$ with $\hat{A}(t)$ and add the score contribution coming from $P(G)$ to obtain the following estimating equation for v :

$$U(v, \hat{A}) = \int_0^\tau \tilde{Z}^\top(t) \{I - \tilde{Y}\{t, \hat{A}(t^-)\} \tilde{Y}^\top\{t, \hat{A}(t^-)\}\} dN(t) + U^G(v). \quad (3)$$

Then, v is estimated by solving $U(v, \hat{A}) = 0$ and denoted by $\hat{v}$. Finally, we then estimate $A(t)$ by least square estimator with $\hat{v}$.

One may ignore the information about v contained in the survival data and base estimation solely on the genotype data. This we refer to as the two-stage approach.

3 Large sample properties

In this Section we outline the large sample properties and derive estimators for the variance of the suggested estimators. When the estimating function, U , is evaluated in the true parameter values v_0 and A it becomes a martingale, but the asymptotic analysis is complicated by the fact that A is estimated using a recursively constructed estimator.

With $\hat{v}$ the proper root of (3) and with additional regularity conditions it follows that it is consistent and asymptotically normal, we refer to the Appendix for further details. But the key arguments are based on the one dimensional case as in Bagdonavicius & Nikulin (1999); Chen et al. (2002); Dabrowska (2005); Gorfine et al. (2005); Zucker (2005). The key to the extension to the multivariate additive risk model is based on the powerful and elegant product integration theory given by Gill & Johansen (1990). This also is utilized by Dabrowska (2005).

It can be established that

$$\begin{aligned} n^{-1/2} U(t, v_0) &= n^{-1/2} \sum_{i=1}^n U_i(t, v_0) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^n \int_0^t q_i(s, v_0) dM_i(s) + n^{-1/2} \sum_{i=1}^n U_i^G(v_0) + o_p(1) \end{aligned}$$

where $U_i(t, v_0)$ is defined in the second equation, $M_i(t) = N_i(t) - \int_0^t \lambda_i(s) ds$ is the basic martingale and $q_i(s, v)$ is a predictable function given in (9) in the Appendix, and $U_i^G(v_0)$ the score for subject i corresponding to the genotype log-likelihood. Based on this i.i.d. decomposition a simple robust estimator of the variance of $n^{-1/2}U(t, v_0)$ is given as

$$\widehat{\Sigma}(t, \widehat{v}) = n^{-1} \sum_{i=1}^n \widehat{U}_i(t, \widehat{v})^{\otimes 2}, \quad (4)$$

where $\widehat{U}_i(t, \widehat{v}) = \int_0^t \widehat{q}_i(s, \widehat{v}) d\widehat{M}_i(s) + U_i^G(\widehat{v})$ and $\widehat{q}_i$ and $\widehat{M}_i$ are estimators of q_i and M_i obtained by using the estimates and empirical versions of all expressions.

Based on this it can be derived that $\sqrt{n}(\widehat{v} - v_0)$ converges to a zero-mean normal distribution with covariance matrix that can be consistently estimated by

$$nI^{-1}(\tau, \widehat{v}) \widehat{\Sigma}(\tau, \widehat{v}) nI^{-1}(\tau, \widehat{v}),$$

where I denotes minus of the derivative of U with respect to v .

It also follows that $\sqrt{n}\{\widehat{A}(t) - A(t)\}$ converges to a Gaussian process. The key is again to establish an i.i.d. decomposition. It turns out, based on a Taylor series expansion, that (up to an $o_p(1)$ term)

$$\sqrt{n}\{\widehat{A}(t) - A(t)\} = n^{-1/2} \sum_{i=1}^n L_i(t, v_0)$$

where the L_i 's are defined in (10) in the Appendix and they are i.i.d. processes.

Let $\widehat{L}_i$ denote the estimator of L_i by plugging in the estimates of the unknown quantities, see Appendix for details. The variance of $n^{1/2}\{\widehat{A}(t) - A(t)\}$ may therefore be estimated by

$$n^{-1} \sum_{i=1}^n \widehat{L}_i^{\otimes 2}(t, \widehat{v}). \quad (5)$$

The asymptotic distribution can be approximated by a resampling scheme

$$\Delta(t) = n^{-1/2} \sum_{i=1}^n G_i \widehat{L}_i(t, \widehat{v}), \quad (6)$$

where $G_1, \dots, G_n$ are independent standard normals and are independent of the counting processes and their covariates.

By a resampling technique we can construct $(1 - \alpha)\%$ confidence bands for each cumulative regression coefficient function of $A_j(t)$, for $t \in [0, \tau]$, and derive a supremum test for testing the hypothesis of $H_0: \alpha_j(t) = 0$, for $t \in [0, \tau]$. We omit the details here.

3.1 Testing of constant effects and significant effects

To derive a formal test to see if a given effect varies over time or if it leads to constant excess risk, we consider a supremum test that we evaluate based on the resampling technique. Considering for example the j th component of the model, $\alpha_j(t)$, we now wish to test $H_0 : \alpha_j(t) \equiv \gamma_j$ for $t \in [0, \tau]$. As in Martinussen & Scheike (2006) a test can be based on a supremum test statistic

$$\widehat{W}_j = \sup_{t \in [0, \tau]} \left| \widehat{A}_j(t) - \frac{\widehat{A}_j(\tau)}{\tau} t \right|.$$

Let $\tilde{\Delta}_j^k(t) = n^{-1/2} \Delta_j^k(t)$ where $\Delta_j^k(t)$ is the k th realization of the resampling process of $\sqrt{n} \{ \widehat{A}_j(t) - A_j(t) \}$. Based on derived large sample property and under H_0 , $\widehat{W}_j$ can be asymptotically approximated by

$$\widehat{W}_j^k = \sup_{t \in [0, \tau]} \left| \tilde{\Delta}_j^k(t) - \frac{\tilde{\Delta}_j^k(\tau)}{\tau} t \right|$$

for $\kappa = 1, \dots, K$. A formal test can be derived by comparing observed $\widehat{W}_j$ and simulated $\{\widehat{W}_j^1, \dots, \widehat{W}_j^K\}$

4 Simulations

We simulated data similar to the BMT example data considered in section 5. The underlying additive hazard model had a baseline at 0.3 with the effect of haplotype match being -0.2 in the first 1.5 time units and 0 afterwards. The survival times were right censored at 3 time units. In the 2054 patients we consider below we found that there were 20, 31 and 13 different allele types for the loci A, B, and DRB1, and that this led to a total of 2,325 possible haplotypes. These haplotypes were assumed to have frequencies similar to those we found in a structured MLE based on the genotype data where we found that 1901 of the haplotypes had frequency 9.95×10^{-5} , 214 had frequency 0.00072, 192 had frequency 0.0019, 14 had frequency 0.0087, 2 had frequency 0.027, 1 had frequency 0.042, and 1 had frequency 0.083, respectively. We tried various other haplotype distributions that were not based directly on the underlying structured model, and we will study this issue further in future work. We then considered sample sizes of 2400, 1200, 600 and 300 that lead to approximately 1100, 650, 320 and 160 number of deaths on average, respectively. 2000 random samples were generated for each simulation study, and the resampling based tests were based on 1000 resamples.

We present detailed results for sample size of 600. Figure 1(a) and 1(b) show the mean of the estimated cumulative intensity (broken line) and the true cumulative coefficient (solid line) for the baseline and the effect of the haplotype match, respectively. Figure 1(c) and 1(d) give the mean of the estimated standard errors (broken line) and the sample standard deviation of the realizations (solid line) for baseline and effect of the haplotype match, respectively. Figure 1 shows that the estimators are essentially unbiased and the standard errors are well estimated by the derived estimator. There is some slight underestimation of the variance, that diminishes when the sample sizes increase (figures are not shown here).

The standard errors of the cumulative coefficients are approximately halved if the haplotype match had been observed. We also considered the effect of estimating the haplotype parameters and this only had minor impact on the standard errors, estimates and coverage probabilities.

Figure 2 shows the pointwise coverage probabilities for the baseline intensity and for the effect of haplotype match, respectively. The slight bias in the variances result in coverage probabilities that are slightly lower for most of the time-axis. The initial problems with the coverage probabilities are well known and due restrictions close to 0, these problems may be remedied by choosing an appropriate transformation.

We also considered the test for constant effects developed earlier as well as the coverage of the confidence bands. The observed coverage is considered in the time-period from 0.3 to 3 to disregard the initial problematic time period. This resulted in the observed levels reported in Table 1.

The test for constant effect achieves the nominal 5% power for the constant effects for the baseline and the power for the time-varying effect of the haplotype match increases with increasing sample size. The coverage probabilities of confidence bands are in an acceptable level. Slightly to low for the baseline, but close to the true 95% for the haplo-match effect.

The finite sample properties are well described by the estimators. Both the pointwise and uniform description of the uncertainty works well. The slight problems are similar to those found for the additive intensity model with known covariates.

5 Haplotype Match for BMT patients

Recently, Petersdorf et al. (2007) studied the effect of MHC haplotype match based on 246 leukemia patients. All these 246 patients received a HLA-identical unrelated HCT between 1986 and 2003. Their analysis was based on observing the haplotypes directly. This is a very costly and time-consuming procedure. They identified 191 (78%) transplants that were haplotype-matched. They demonstrated that the MHC haplotype match of donor and patient had significant impact on the transplantation outcome of developing acute graft-versus-host disease (GVHD) and recurrent malignancy, but had no impact on the treatment related mortality (TRM) which is defined as death in complete remission and overall mortality. For the illustration purpose, we consider a similar cohort of patients with HLA-identical unrelated HCT for leukemia and transplanted between 1995 and 2007, and donor-patient matched on four loci HLA-A/B/C/DRB1 based on low-resolution (2 digits) classification. The data presented here were obtained from the Statistical Center of the Center for International Blood and Marrow Transplant Research (CIBMTR) located at the Medical College of Wisconsin. The analysis has not been reviewed or approved by the CIBMTR. The CIBMTR is comprised of clinical and basic scientists who share data on their transplant patients with CIBMTR Statistical Center.

The selected data consists of 5,145 leukemia patients (2,054 for acute myeloid leukemia (AML), 1,171 for acute lymphoblastic leukemia (ALL), 1,037 for chronic myelogenous leukemia (CML) and 883 for myelodysplastic syndrome (MDS)). The analysis showed that a haplotype match only had a significant effect on AML patients. For illustration purposes, we here focus on the overall survival of the 2,054 AML patients.

The genotype data used in estimation is based on the three-loci HLA-A, HLA-B and HLA-DRB1. This gives a very rich haplotype space. To reduce the total number of possible haplotypes, low-resolution (2 digits) classification of alleles at loci {A1, A2}, {B1, B2} and {DRB11, DRB12} were used to estimated haplotype frequencies. There were 20, 31 and 13

different allele types for the loci A, B, and DRB1, respectively. This leads to a total of 2,325 possible haplotypes. Some additional structure is needed to be able to estimate the parameters in this setting, and we grouped rare haplotypes into groups with a common haplotype frequency. We ended up with a model with 10 parameters where the 5 largest frequencies have their own parameters and other frequencies were allowed to be in 5 different groups. We tried modifying the models in many different ways but the results were quite robust to these choices. In the regression analysis below we adjust for disease stage at pre-transplant (high risk patient (36%) versus early or intermediate risk patient (64%)), age of patient (> 45 (50%) versus $19 - 45$ (38%) versus ≤ 18 (12%) years), GVHD prophylaxis (Cyclosporine-A (CsA) $\pm$ other (44%) versus others (56%)), graft source (bone marrow (BM) (45%) vs peripheral blood (PB) (55%)) and year of transplant (1995–2001: early, (25%) versus 2002–2007 (75%)). We considered the survival for the first 90 months ($\tau = 90$). There were 1,206 patients died within this study period.

We fitted an additive model to evaluate the effect of haplo-match by $X^T(H_P, H_D) = \{X^T, I(H_P = H_D)\}$, where X consists of a baseline, disease stage, age of patient, GVHD prophylaxis, graft source and transplant year. The estimated cumulative coefficients and 95% pointwise confidence intervals (solid lines) and with 95% confidence bands (broken lines) for covariates and haplo-match are given in Figure 3. Note that it appears that all covariates had time-varying effects and that the haplo-match effect was also timevarying.

The formal tests for a constant excess risk lead to p-values for disease stage, age group $19 - 45$, age group > 45 , GVHD prophylaxis, graft source and year of transplant that are < 0.001 , 0.051 , < 0.001 , 0.002 , 0.014 and 0.002 , respectively. The haplo-match effect lead to a p-value of $p = 0.033$. Figure 3 and these tests results indicate that a similar Cox regression analyses would not be appropriate here.

Figure 3 indicates the overall effect of matching haplotypes is significant even though the significance test only leads to a p-value of 0.055 . Note, also that the test for constant effect also provides an alternative test for significance.

Also, the negative cumulative hazards for the haplo-matching indicates that patient with a haplo-matched HCT has a higher survival probability. Figure 4 gives the predicted survival probabilities of haplo-mismatch (panel a) and haplo-matched (panel b) for a young patient (≤ 18) with early or intermediate disease at pre-transplant, received other than CsA $\pm$ others for GVHD prophylaxis and received a PB transplant between 2002–2007. The survival predictions are provided with 95% pointwise confidence intervals (dotted lines) and with 95% confidence bands (broken lines). The predicted 5 year survival probabilities for haplo-mismatched and haplo-matched HCT are 55% (95% CI: 38%–72%) and 83% (64%–100%), respectively.

This BMT example showed a positive beneficial effect in overall survival for haplotype matched transplant. As we discussed in section 1, this positive effect could lead to a new decision in choosing a best matched donor.

6 Discussion

The popularity of the additive hazards model risk model has grown as it is well suited to study time-dynamics, that is whether or not effects of covariates change with time. In this paper we have demonstrated how the additive risk model can be applied also when there is missing data, and in particular how one can estimate effects of haplo-type match in BMT studies.

The large sample properties of the suggested estimators were derived using product integration theory, and these seem to give an accurate description of the variability even for very small samples. Furthermore, the formal test of whether or not covariate effects vary with time worked quite well in the simulations.

In the context of the the BMT example, we studied the effect of haplotype match, and here specific modeling of the haplotype frequencies were needed. It would be interesting to further consider how to make effects robust to misspecification of the haplotype modeling, and one may pursue ideas like Allen et al. (2005). Further, one may model the haplotypes with additional complexity to allow for lacking Hardy-Weinberg Equilibrium and recombination that is know to present on the HLA gene for the three loci we consider.

The approach is implemented in the R-package HaploSurvival, that is being developed by the first author.

We appreciate helpful and constructive suggestions by the Editors and a referee.

This research was supported by NIH-grant 2 R01 CA54706-10, and a grant from the Danish Research Council on “point process modeling and statistical inference”.

7 Appendix

We start by making the observation that (up to $o_p(n^{-1})$)

$$\widehat{A}(v_0, t) - A_0(t) = \int_0^t \tilde{Y} - (v_0, \widehat{A}) dM - \int_0^t \tilde{Y} - (v_0, \widehat{A}) \{ \tilde{Y}(v_0, \widehat{A}) - \tilde{Y}(v_0, A_0) \} dA_0(s),$$

where $M(s)$ is the (vector) counting process martingale, and where the time arguments were omitted for notational simplicity. Let $S_0(t, v_0) = \tilde{Y}^T(t, v_0, \widehat{A}) \tilde{Y}(t, v_0, \widehat{A})$ and let the limit of this quantity be denoted as $s_0(t)$. Note that the first term, V_n , by the martingale CLT, can be shown to converge to a Gaussian martingale process $V(t)$ with variance $\sigma^2(t) = \int_0^t s_0^{-1} dA_0$. Note also that $V_n(t)$ can be written as

$$V_n(t) = \sum_{i=1}^n \int_0^t S_0^{-1}(s, v_0) \tilde{Y}_i W_i dM_i.$$

Also, a Taylor expansion using matrix derivatives yields (up to $o_p(n^{-1/2})$)

$$\tilde{Y}(t, v_0, \widehat{A}) - \tilde{Y}(t, v_0, A_0) = \sum_j D_{A,j} \tilde{Y}(t, v_0, A_0) \left\{ \widehat{A}_j(t^-) - A_0(t^-) \right\},$$

where the i th row of $D_{A_j}(\tilde{Y}(v_0, A))$ is $D_{A_j} f_i(v, A(t^-))$, and therefore with

$$\begin{aligned} E_j(t) &= \tilde{Y}^-(t, v_0, A_0) D_{A_j} \tilde{Y}(t, v_0, A_0) \\ \tilde{Y} - (v_0, \widehat{A}) \{ \tilde{Y}(t, v_0, \widehat{A}) - \tilde{Y}(t, v_0, A_0) \} &= \sum_j E_j(t) \{ \widehat{A}_j(t^-) - A_0(t^-) \} + o_p(n^{-1/2}). \end{aligned}$$

This implies that $\sqrt{n} \{\widehat{A}(v_0, t) - A_0(t)\}$ converges in distribution to a p -dimensional process W that satisfies the integral equation

$$W(t) = \sum_{j=1}^p \int_0^t -W_j(s^-) e_j(s, v_0) dA_0(s) + V(t).$$

Denote the elements of $e_j(t)$ as $e_{j,km}(t)$. With the $p \times p$ matrices $F(t)$ with elements $F_{kj}(t) = \sum_{m=1}^p e_{j,km}(t) \alpha_{0,m}(t)$ the solution can be written as

$$dW(t) = -W(t^-) F dt + dV(t).$$

Using the product integration formulae (Gill & Johansen, 1990) leads to the solution

$$\begin{aligned} W(t) &= \int_0^t dV(s) \mathcal{F}(s, t), \\ \mathcal{F}(s, t) &= \prod_{u \in]s, t]} (I - F du), \end{aligned} \quad (7)$$

and where $F(s, t)$ is the product integral of F . In the one-dimensional (commutative) case this equals $\exp\left(-\int_s^t e(u, v_0) dA_0(u)\right)$. Note that the matrix $F(s, t)$ is estimated consistently by the product of (the atoms)

$$\widehat{\mathcal{F}}(s, t) = \prod_{]s, t]} (I - d\widehat{F}), \quad (8)$$

where $d\widehat{F}_{jk}(t) = \sum_{m=1}^p E_{j,km}(t) d\widehat{A}_m(t)$. Using the product integral rather than the exponential in the one-dimensional case yields an estimator with the same asymptotic properties.

We now consider the estimating function for v based on the counting process data, $\tilde{U}^N(v)$. By the martingale decomposition

$$\tilde{U}^N(v_0) = \int Z^\top \{I - \tilde{Y}(v_0, \widehat{A}) \tilde{Y}^-(v_0, \widehat{A})\} dM + \int Z^\top \{I - \tilde{Y}(v_0, \widehat{A}) \tilde{Y}^-(v_0, \widehat{A})\} \tilde{Y}(v_0, A_0) dA_0.$$

By a Taylor expansion, as before, it may be shown that the difference between the last two terms in the previous equation (up to $o_p(n^{-1/2})$) equals

$$-\sum_{j=1}^p Z^\top \{I - \tilde{Y}(v_0, A_0) \tilde{Y}^-(v_0, A_0)\} D_{A_j} \tilde{Y}(v, A_0) \{\widehat{A}_j(t^-) - A_{0j}(t^-)\} = \sum_{j=1}^p \tilde{s}_j \{\widehat{A}_j(t^-) - A_{0j}(t^-)\}.$$

With the matrices in each of the terms converging to $\tilde{s}_j$.

The last two term therefore equals ($o_p(n^{-1/2})$)

$$\sum_{j=1}^p \int \tilde{s}_j W_j dA_0,$$

that can be written as

$$\int G(t) W(t) dt,$$

with $G(t)$ is an $q \times p$ matrix with elements $G_{kj}(t) = \sum_{m=1}^p \tilde{s}_{j,km} \alpha_{0,m}(t)$. Now changing the order of integration and using the structure of $W(t)$ one gets

$$\int_0^\tau G(t) W(t) dt = \int_0^\tau G(t) \int_0^t \mathcal{F}(s, t) V(s) ds dt = \int_0^\tau g(s) V(s) ds = \sum_{i=1}^n \int_0^\tau g(s) s_0^{-1}(s, v) \tilde{Y}_i dM_i(s),$$

where

$$g(s) = \int_s^\tau G(t) \mathcal{F}(s, t) dt = \int_s^\tau \mathcal{F}(s, t) dG(t).$$

Define

$$\begin{aligned} q_{1i}(t, v) &= \tilde{Z}_i - s_z(t, v) s_0^{-1}(t, v) \tilde{Y}_i W_i, \\ q_{2i}(t, v) &= g(t) s_0^{-1}(t, v) \tilde{Y}_i W_i, \\ q_i(t) &= q_{1i}(t, v) + q_{2i}(t, v). \end{aligned} \quad (9)$$

Then one can write the estimating function as follows (up to an $o_p(1)$ term)

$$n^{-1/2} \tilde{U}^N(v_0) = n^{-1/2} \sum_{i=1}^n \int_0^\tau \{q_{1i}(t, v_0) + q_{2i}(t, v_0)\} dM_i(t).$$

This is a sum of i.i.d. terms (or a martingale) and therefore converges to a normal distribution with variance that is estimated by the robust estimator given in (4). The total score for v is $U^N(v_0) + U^G(v_0)$

Based on a Taylor series expansion (up to an $o_p(1)$ term) we find that

$$\begin{aligned} & \sqrt{n} \{\widehat{A}(t) - A_0(t)\} \\ &= n^{1/2} \int_0^t \frac{\partial}{\partial v} \tilde{Y}^-(t, v_0) dN(\widehat{v} - v_0) + n^{-1/2} \sum_{i=1}^n \int_0^t \mathcal{F}(s, t) s_0^{-1}(s, v_0) \tilde{Y}_i(s) dM_i(s) \\ &= n^{-1/2} \sum_{i=1}^n L_i(t, v_0), \end{aligned}$$

where

$$L_i(t, v) = P(t, v_0) I^{-1}(\tau) \int_0^\tau q_i(t, v_0) dM_i(t) + \int_0^t \mathcal{F}(s, t) s_0^{-1}(s, v_0) \tilde{Y}_i(s) dM_i(s), \quad (10)$$

and with $P(t, v_0)$ the limit of $n^{-1} \widehat{P}(t, \widehat{v})$ defined in (11), with $I(\tau)$ the limit of $n^{-1} I(\tau, v_0)$ (12), and with the product integral $F(s, t)$ defined in (7).

The derivative of $\tilde{A}(t)$ with respect to v is

$$\tilde{P}(t, v) = -\int_0^t S_0^{-1}(\tilde{A}) D_v S_0(\tilde{A}) S_0^{-1}(\tilde{A}) \tilde{Y}^\tau W dN + \int_0^t S_0^{-1}(\tilde{A}) D_v \tilde{Y}^\tau W dN + \int_0^t S_0^{-1}(\tilde{A}) \tilde{Y}^\tau D_v W dN. \quad (11)$$

The derivative with respect to v of the estimating function, U , is

$$\begin{aligned} I(v) &= I(\tau, v) \\ &= \int D_v(\tilde{Z}^\tau) \{I - \tilde{Y} \tilde{Y}^\tau\} dN \\ &\quad - \int \tilde{Z}^\tau D_v(\tilde{Y}^\tau) S_0^{-1} \tilde{Y}^\tau W dN \\ &\quad + \int \tilde{Z}^\tau \tilde{Y} S_0^{-1} (D_v S_0) S_0^{-1} \tilde{Y}^\tau W dN \\ &\quad - \tilde{Z}^\tau \tilde{Y} S_0^{-1} D_v(\tilde{Y})^\tau W dN \\ &\quad - \int \tilde{Z}^\tau \tilde{Y} S_0^{-1} \tilde{Y}^\tau D_v W dN. \end{aligned} \quad (12)$$

Based on the derivation of the estimating function it straightforward to do a Newton-Raphson iteration to find a solution to the estimating equations. Note that the first term is a martingale when evaluated in the true parameter, and therefore will converge towards 0 in probability.

Note also that an estimator of $F(s, t)$ is given in (8).

References

- Aalen OO. A linear regression model for the analysis of life times. *Statist. Med.* 1989; 8:907–925.
- Allen AS, Satten GA, Tsiatis AA. Locally-efficient robust estimation of haplotype-disease association in family-based studies. *Biometrika.* 2005; 92:559–571.
- Bagdonavicius V, Nikulin M. Generalised proportional hazards model based on modified partial likelihood. *Lifetime Data Anal.* 1999; 5:329–350. [PubMed: 10650741]
- Chen J, Peters U, Foster C, Chatterjee N. A haplotype-based test of association using data from cohort and nested case-control epidemiologic studies. *Human Heredity.* 2004; 58:18–29. [PubMed: 15604561]
- Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika.* 2002; 89:659–668.
- Dabrowska DM. Quantile regression in transformation models. *arXiv:math.ST.* 2005; 05115082v1:1–34.
- Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics.* 2003; 73:1316–1329. [PubMed: 14631556]
- Gill RD, Johansen S. A survey of product-integration with a view towards application in survival analysis. *Ann. Statist.* 1990; 18:1501–1555.
- Gorfine M, Zucker D, Hsu L. Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika.* 2005; 93:735–741.
- Huffer FW, McKeague IW. Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.* 1991; 86:114–129.
- Lin D, Zeng D. Likelihood based inference on haplotype effects in genetic association studies. *J. Amer. Statist. Assoc.* 2006; 101:89–110.
- Martinussen, T.; Scheike, T. *Dynamic Regression Models for Survival Data.* Springer; Verlag New York: 2006.
- Petersdorf E, Malkki M, Gooley T, Martin P, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLOS medicine.* 2007; 4:59–68.
- Schaid DJ. Evaluating associations of haplotypes with traits. *Genetic Epidemiology.* 2004; 27:348–364. [PubMed: 15543638]

- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when the linkage phase is ambiguous. *American Journal of Human Genetics*. 2002; 70:425–434. [PubMed: 11791212]
- Scheike T, Martinussen T, Silver J. Estimating haplotype effects for survival data. *Biometrics*. 2010 DOI: 10.1111/j.1541-0420.2009.01329.x.
- Scheike TH. A flexible semiparametric transformation model for survival data. *Lifetime Data Anal*. 2006; 12:461–80. [PubMed: 17031497]
- Tregouet D, Tivet L. Cox proportional hazards survival regression in haplotype-based association analysis using the stochastic-em algorithm. *European Journal of Human Genetics*. 2004; 12:971–974. [PubMed: 15241485]
- Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*. 2002; 53:79–91. [PubMed: 12037407]
- Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*. 2003; 72:1231–1250. [PubMed: 12704570]
- Zucker DM. A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *J. Amer. Statist. Assoc*. 2005; 100:1264–1277.

Figure 1. Mean of estimated cumulative hazards (broken line) versus true cumulative hazards (solid line) for the baseline and haplomatch effect (Panels a and b), and observed pointwise standard deviation of estimates (solid line) versus mean of estimated standard errors (broken line) (Panels c and d) based on sample size 600

Figure 2. Observed pointwise coverage of 95 % pointwise confidence interval for sample size 500 for the cumulative baseline (solid line) and cumulative excess effect of haplomatch (broken line).

Figure 3.
Cumulative excess hazard for covariates based on additive model with 95% point-wise confidence intervals and 95% confidence bands (broken lines).

Figure 4. Predicted survival probabilities with 95% pointwise confidence intervals (dotted lines) and 95% confidence bands (broken lines) based on additive risk model for a specific patient (see text) with haplotype mismatched HCT (Panel a) and with haplotype matched HCT (Panel b).

Table 1

Observed power for test of constant effects and coverage probabilities for confidence bands for model where baseline is constant and haplotype match effect is non-constant.

	Observed power for test of $H_0 : A(t) = \gamma t$	
n	baseline	haplomatch effect
300	0.06	0.17
600	0.05	0.25
1200	0.05	0.40
2400	0.05	0.63

	Observed coverage of confidence bands	
n	baseline	haplomatch effect
300	0.92	0.96
600	0.93	0.97
1200	0.94	0.97
2400	0.95	0.95