Prediction of hematopoietic cell transplantation success from HLA matching

Bence Mélykúti

Dissertation

September 2007

Supervisor:
Gil McVean (Department of Statistics)

University of Oxford
Life Sciences Interface Doctoral Training Centre

Abstract

In bone marrow transplantation finding a donor whose HLA genes are identical or almost identical to those of the transplant recipient is immensely difficult due to the high polymorphism of these genes, but is crucial to the success of the intervention.

A statistical meta-analysis of survival data for 1347 hematopoietic cell transplantations is presented. In a newly developed parametric model of survival time distributions, prospects of recipients were described by a function of mismatch counts between donor and recipient for six HLA genes (HLA-A, -B, -C, -DRB1, -DQB1, -DPB1). Using the maximum likelihood framework to tune parameters to find the best fit of a given model to the data, and also to compare different models, we carried out exhaustive variable selection cycles for different model subtypes.

We have shown that under our modelling assumptions mismatches of HLA-C & HLA-DPB1 alleles between donor and recipient have a negative effect on recipients' survival prospects.

There have already been reports showing the importance of HLA-C matching, but there have been no strong evidence for association between HLA-DPB1 mismatches and transplant-related risk increase. If other studies support the finding that HLA-DPB1 allele mismatches have a negative effect on survival, it should have an impact on clinical protocols for finding appropriate hematopoietic cell donors for patients.

Contents

1 Introduction

1.1 Hematopoietic cell transplantation

Hematopoietic cell transplantation (HCT) is a treatment option for more than twenty different life-threatening diseases, including leukæmias and lymphomas, inherited immune system disorders and inherited metabolic diseases[4,29]. It means the transplantation of bone marrow, peripheral blood, or umbilical cord blood stem cells.

The World Marrow Donor Association reported 7226 blood stem cell transplantations and 1126 patients who were given cord blood units worldwide for the year 2004[31]. Worldwide more than 50 000 patients have received allogeneic (i.e. genetically non-identical, from an unrelated donor) hematopoietic cell transplantation so far, 5000 have received allogeneic umbilical cord blood transplantation[31], and since 1987 alone the US-based National Marrow Donor Program® (NMDP) have helped more than 25 000 patients to receive bone marrow or cord blood transplantation[29]. At any one time 7000 patients throughout the world may be needing hematopoietic cell transplants[30]. The UK's leading bone marrow register searches for donors on behalf of at least 3000 newly diagnosed patients each year[30]. But the number of patients who could benefit from safer transplants is much bigger: different reports and expert opinion both support the claim that even today HCT therapy is broadly underused[4,8].

Our well-established understanding is that a key determinant of the success of HCT is the best possible matching of the human leukocyte antigen (HLA) genes between donor and transplant recipient. Clinically, protein serotypes from at least five of these genes are determined and compared to potential donors' to find one whose genetic profile is similar enough to the patient in need of a transplantation. Although there is substantial linkage disequilibrium among the HLA genes[7], one HLA gene does not determine others on the same chromosome. Therefore a complete match of ten alleles for the five genes (each of which allele is a sample from an imaginary pool of up to hundreds) between patient and any one potential donor is very unlikely. Siblings and close relatives of the patient are most likely to have identical types, but when they do not (which happens in about 60–70% of the cases[31]), an unrelated donor has to be searched for.

This combinatorial variability implies the need for vast numbers of people to volunteer to become donors if needed. Around the globe there is serotype data stored in registries from more than 11.5 million volunteers or cord blood units[28].

Despite this large number of registered potential donors it may happen that no perfectly matching donor can be found. If there is no alternative treatment option and the patient urgently needs a transplant, clinicians have to choose a donor with an incomplete match.

The aim of this study is to add to our knowledge how to choose the one donor. A typical question may be whether there are genes for which the perfect match of both alleles is necessary for the patients' recovery or not, or whether there are any genes for which a mismatch is more tolerable.

It might be the case that the key determinant is not the matches between single alleles but a certain complicated notion of match between the system of genes in the patient and the donor, or even that not the HLA genes themselves are important but one or more regions which are in linkage disequilibrium with them.

1.2 MHC region and the HLA genes

The human Major Histocompatibility Complex (MHC) is a large array of genes on chromosome 6. MHC proteins are almost like military insignia or signalling flags helping our adaptive immune system recognise self against non-self in the never-ending war against pathogens.

In addition to the genomic organisation of the MHC region, the serological and DNA sequence level characterization of most known alleles are known[26], and allele, haplotype and genotype frequencies in different populations are available[6,23,24,25,12,14]. The functions and structure of MHC proteins are increasingly understood.

MHC proteins were first identified as the main antigens recognised in transplantation reactions. Skin grafting experiments conducted in the 1950s with mice proved that graft rejection is an adaptive immune response to the foreign antigens on the surface of grafted cells.[2]

MHC proteins bind foreign protein products in the cytoplasm of cells infected by microbes (class I MHC proteins) or in endocytosed extracellular fluid (class II MHC proteins) and then present them on the cell surface. While class I MHC proteins (such as those coded by HLA-A, -B, -C genes) are expressed in most nucleated cells, class II MHC proteins (coded by HLA-DR, -DQ, -DP and some more other genes) are typically found on antigen-presenting cells such as dendritic cells, macrophages and B cells.[2,21]

T cells are constantly surveying the surface of cells in our body searching for signs of malfunction. A cytotoxic T cell with its receptors can interact with the complex of a class I MHC protein and the protein fragment bound to it and in the meantime its CD8 co-receptor can bind to the non-binding part of the MHC protein to ensure the specificity of T cell recognition. Similarly, helper T cells recognise foreign peptide–class II MHC protein complexes on antigen-presenting cells and use CD4 co-receptors to enhance recognition specificity.[2,21]

Indeed, Zinkernagel & Doherty (1974) showed that a given T cell will only recognise viral peptides when they are bound to a specific MHC protein[33]. This phenomenon is called *MHC restriction*.

Variation in HLA types has been associated with variation in susceptibility and variation in progression of various infectious diseases (e.g. HIV progression, susceptibility to hepatitis B & C, malaria, pulmonary tuberculosis, leprosy)[3,21].

As Klein put it in his 1987 paper[16], the two most profound secrets of the MHC is its true function, and the origin and significance of its polymorphism.

Although, as it has just been outlined, much is known about the MHC function, there is still much left to be learned.

As to the question of polymorphism, the extreme diversity of the HLA genes is well illustrated by the fact that up to 1st July 2007, 580 HLA-A, 921 HLA-B, 312 HLA-C, 501 HLA-DRB1, 86 HLA-DQB1, and 127 HLA-DPB1 alleles have been named including synonymous and non-expressed types[6,23,24,25,30].

Doherty & Zinkernagel (1975) were who first argued that heterozygosity at HLA loci increases individual fitness because it results in successful responding to a wider range of pathogens[9].

There is wide agreement in the scientific community that on the population level selection acts on MHC genes: there is compelling evidence that it is not neutral evolution, but balancing selection which shapes the MHC region[21], that is, selection which enhances

polymorphism.

Pathogen driven mechanisms are strongly believed to be causes of balancing selection, in the form of *overdominant selection* (which is the advantage of heterozygotes), *frequency dependent selection* (being a minority is an advantage, if a certain pathogen adapts to individuals who have a frequent allele) or *fluctuation in selection pressures* (when fitness changes as a function of frequency of pathogens, and not allele frequencies across the population). Many observations suggest that reproductive mechanisms may also play an important role in shaping HLA diversity: HLA types may have an effect on *mate choice* or on the *probability that pregnancy will be completed successfully*[21].

The standard textbook view[2] is that the role of MHC proteins in binding and presenting foreign oligopeptides provides an explanation for the high MHC polymorphism. To illustrate the difficulties in this field, let us mention that Klein (1987) disagreed both with the claim that an arms race between the immune system and pathogens causes the MHC variability, and the hypothesis that there is strong selection pressure on the MHC[16]. More recent findings weaken his position[21].

Let us close this introductory part about the MHC and continue with necessary preparations that will be needed to understand our investigations.

1.3 HLA nomenclature

Each HLA allele has a unique name (e.g. HLA-DRB1*1602). The letters HLA and the specification of the locus are followed by a four-digit number. The first two digits identify the type of the allele, which in most cases corresponds to the serological type. The third and fourth digits refer to subtypes: these distinguish between different amino acid sequences in coded MHC proteins of the same type.

In several cases further digits are used. Fifth and sixth digits are used to indicate different but synonymous DNA sequences. Alleles that only differ by sequence polymorphisms in introns or in the $5'$ or $3'$ untranslated regions are coded with different seventh and eighth digits.[20,27]

Reasons for no one-to-one correspondence between two-digit codes and serotypes may be that there are more than 99 coding variants in one type, as in the case of the HLA-B*15 family: new allele types of this family named after the HLA-B*1599 get HLA-B*95 codes. (Similarly, the recently discovered HLA-A*02 alleles got A*92 codes.) Another reason may be that there are more than 99 types in a family, like in the HLA-DPB1 family: HLA-DPB1*0102 followed the sequence of DPB1*0101, DPB1*0201, ..., DPB1*9901.[20,26,27]

1.4 Patient–donor matching guidelines

In this section we describe the currently implemented criteria for donor searching.

The Anthony Nolan Trust in its HLA Typing and Matching Guidelines[30] recommends patients and unrelated donors should be matched on high resolution HLA-A, -B, -C and -DRB1 types. (High resolution in their definition means resolving polymorphisms within exons 2 and 3 for HLA-A and -B, and the same within exon 2 for HLA-DRB1.) HLA typing of the patient must be DNA typing, at minimum at HLA-A, -B, -C and -DRB1 loci, desirably also at HLA-DRB3, -DRB4, -DRB5, and -DQB1, and optionally at -DPB1.

When needed, partially matched donors can be chosen, but all levels of mismatch would be subject to review. In this case both patient and donor typing should be at high resolution so that the degree of mismatch can be thoroughly assessed.

The NMDP requires matching of 5 of the 6 HLA-A, -B, and -DRB1 alleles as a minimum[13], the argument being that there is abundant data to show that this level of matching can lead to successful transplantation outcomes. It is also added that transplantation success can be improved by stricter matching criteria (e.g. matching for HLA-C). The optimal matching criteria are given as allele-level match for HLA-A, -B, -C, and -DRB1. (The criteria are less strict in umbilical cord blood transplantation.[18,29])

Interesting to note that patients who know their HLA types can search the NMDP registry themselves online to see their prospects for finding a matching donor[29].

In addition to HLA matching, other factors are also considered when selecting the donor: cytomegalovirus (CMV)–negative serology is recommended for patients with negative CMV–serology, larger donor body weight and male sex (on average such donors provide more stem cells), ABO compatibility, matched race, and younger age.[13]

When multiple highly matched potential donors are available, as the NMDP Guidelines say, matching HLA-DQB1, -DPB1, and -DRB3/4/5 loci might be beneficial[13]. In addition to this, it is pointed out that the association between HLA-DQB1 and -DPB1 mismatching and mortality is unproven, and the same has not been studied for HLA-DRB3/4/5 loci.

The NMDP website gives information about each U. S. transplant centre in the NMDP Network[29], including the matching criteria used by them. The requirements vary from centre to centre.

For bone marrow or peripheral blood stem cell (PBSC) transplants, in a typical example 10 of 10 matches are required at HLA-A, -B, -C, -DR and -DQ antigens with an unrelated donor, but a mismatch might be allowed on a case-by-case basis.

Some centres only allow mismatches at the C or DQ loci (e.g. Hawaii Medical Center). On the other hand, some allow mismatches only at the A, B or DR loci (e.g. Loma Linda University Medical Center)!

Some centres set 8 of 8 matches at A, B, C and DR antigens (others: alleles) as standard, potentially with one mismatch at any locus, or allow mismatches only at A, B or C loci.

The standard match level in the UCSF Medical Center is 8 of 8 matches at A, B, C and DR alleles for bone marrow, 10 of 10 matches at A, B, C, DR and DQ alleles for PBSC transplants.

Some centres only disclose that the required level of matching varies depending on treatment protocol.

To conclude this section, we may say even if we assume that these standards are constantly revised at each centre, and that they were valid only at the moment of communicating them to the NMDP, the variance clearly shows that our knowledge about genetic determinants of HSC transplant success is limited.

1.5 Research underpinning patient–donor matching standards

Many statistical analyses of transplant outcomes have focused on dependence from disease, disease stage, origin of transplant (whether autologous or from HLA-identical sibling or allogeneic), or transplant patients were compared with patients receiving different

treatment. (For an overview of these topics see the NMDP website[29].)

There have also been many studies to evaluate the role of HLA matching in transplant outcome.

The aforementioned NMDP commentary (2003)[13] focused on recent large studies by three groups. It expressed that the effect of specific HLA mismatches on specific outcomes, such as graft failure, acute and chronic graft-versus-host disease (GVHD), should be secondary to expected survival times in donor selection, instead they should be used to choose a specific risk-adapted treatment strategy for the patient.

It is now widely accepted that 4-digit matches (that is, allele, instead of only serological) should be sought for.

Flomenberg et al. (2001) reported 8–12% reduction in survival at 5 years after transplant for patients with a single allele mismatch at HLA-A, -B, -C or -DRB1 compared to no mismatch, in a pool of recipient–donor pairs who were matched at the 2-digit level for HLA-A, -B, and -DRB1[10].

Greinix et al. (2004) showed significant ($P = 0.03$) decrease in three-year survival for patients who had at least two allele mismatches in the HLA class I region compared to those with zero or one mismatch, and concluded that selection of unrelated donors should be based on high-resolution HLA class I typing[11].

NMDP data suggests that a single allele-level mismatch is preferable to an antigen-level mismatch, but for instance no such rule has been inferred to decide between a single antigen mismatch and two allele-level mismatches[13]. There are studies suggesting that multiple mismatches may pose cumulative or even synergistic risk on recipients. But beyond these guidelines, the NMDP did not rank the relative importance of matching at particular HLA loci (HLA-A, -B, -C or -DRB1) or could not predict permissible mismatches. These questions remain open.

In real life the physician should consider the patient's clinical status very carefully before deciding on the length of time it is feasible to search for a donor, if only partially matched donors are available. For instance, newly diagnosed chronic myelogenous leukæmia can be relatively stable, allowing for a search time of 4 months, while for an acute leukæmia patient transplantation may be feasible for only a brief period. Waiting for a better matching donor may expose patients to further toxic chemotherapy, an increased risk of infection or relapse. Besides differences in life expectancy, the quality of life achievable with HCT from the best available donor should be compared to what the patient can expect from other therapies.[13]

Further investigations of effects of HLA mismatches on HCT outcome are still needed, because this is the only way to reduce risks to patients if transplantation is the best treatment option for them, and to increase the availability of HCT treatment to a wider range of patients.

2 Data source

For the current meta-analysis we have been using the dbMHC database[12], which was generated by the Hematopoietic Cell Transplantation component of the International Histocompatibility Working Group (IHWG) through data contribution from fifteen countries. The database was last updated on 1st November 2004.

For each 1347 recipient–donor pair the data matrix has the following entries:

- Day: time after transplant in days.
- Died: a binary variable.
- Diagnosis.
- Recipient and donor ages at transplantation/donation in years.
- Recipient and donor genders.
- Recipient and donor HLA alleles at HLA-A, -B, -C, -DRB1, -DQB1, and -DPB1 loci (4-digit codes). (The genetic information contributes $2 \times 2 \times 6 = 24$ entries.)

Patient age was unknown for 134 pairs, donor age for 569 pairs, patient gender for 170 pairs, and donor gender for 271 pairs.

There were five pairs for whom one of four HLA-C alleles was known to 2-digit resolution only, and one pair in which both the recipient and the donor had one allele typed for 2 digits only.

HLA-DRB1 types were missing completely for one pair.

HLA-DPB1 types were missing completely for 293 pairs. There were further three pairs with one donor, and one pair with one patient and one donor HLA-DPB1 type missing.

The observation that there were two transplants both with exactly 92.97-year-old donors (who had different HLA types) raises doubts about the reliability of the dataset.

3 Basic statistical tools

3.1 Survival analysis

The data we have been working with is so-called *survival data*. Such time-to-failure data often arises in medical (patient follow-up) and engineering (product testing) studies[17,19].

Event times measured from the day of transplantation, and an indicator variable are given for each patient, the latter telling whether the event was *death* or *census*. In our dataset census always means *right censoring*: it is known that the patient was alive on that day, or equivalently, the day is a lower bound of actual survival time.

Different causes might lead to censoring, for example the patient might lose contact with the transplant centre, the centre might finish the follow-up, or census might simply mean that when the centre communicates its follow-up data, the patient is alive.

3.2 Survival and hazard functions

Suppose that survival time Z is a positive random variable with a continuous cumulative distribution function (cdf) F. In our parametric models we will always assume that it is absolutely continuous, and its probability density function (pdf) will be denoted by f.

We need to introduce some mathematical notions. Detailed explanations of these can be found in standard survival data analysis textbooks, such as the work by Klein & Moeschberger (2003)[17].

The probability of surviving to time t as a function of t is often called the *survival function*: $S(t) = 1 - F(t)$.

The instantaneous rate of death is the *hazard function* (or *hazard rate*) h. More precisely, it is defined by

$$h(t) = \lim_{\Delta t \searrow 0} \frac{P\Big(t \le Z < t + \Delta t \,\Big|\, t \le Z\Big)}{\Delta t},$$

or, assuming absolute continuity,

$$h(t) = \frac{f(t)}{S(t)} = -\Big(\log S(t)\Big)'.$$

A related quantity, the *cumulative hazard function* H is given by

$$H(t) = \int_0^t h(x)\,\mathrm{d}x = -\log S(t).$$

Consequently, the distribution (given by cdf or pdf or survival function) uniquely determines the hazard function, and vice versa: by the formula

$$S(t) = e^{-H(t)} = \exp\left(-\int_0^t h(x)\,\mathrm{d}x\right)$$

the hazard uniquely determines the distribution. Moreover, any $g : [0, \infty[\to [0, \infty[$ function is a valid hazard function if

$$\int_0^\infty g(x)\,\mathrm{d}x = \infty,$$

in the sense that it uniquely defines a distribution.

For the reader who knows the exponential distribution, the hazard function *'is the thing which is constant in the exponential distribution'*: if the exponential distribution has mean $1/\lambda$, then the hazard is $h(t) = \lambda$.

3.3 Non-parametric estimation of basic quantities for right-censored data: Kaplan–Meier estimator, Nelson–Aalen estimator

The standard estimator of the survival function from survival-type data, the *Kaplan–Meier estimator (KME)* or *Product-Limit estimator*, was first proposed by Kaplan & Meier (1958)[15].

Let $t_1 < t_2 < \ldots < t_K$ be the distinct event times (census or death), allowing multiple events happening at the same time. The idea is that the survival function can be written as a product of a chain of conditional probabilities: if $t_{k-1} < t \leq t_k$ for some $2 \leq k \leq K$, then

$$S(t) = P\Big(Z > t \,\Big|\, Z > t_{k-1}\Big) P\Big(Z > t_{k-1} \,\Big|\, Z > t_{k-2}\Big) \ldots P\Big(Z > t_1 \,\Big|\, Z > 0\Big) P\Big(Z > 0\Big).$$

Under our assumptions $P(Z > 0) = 1$.

Let Y_i denote the number of people at risk just before time t_i, that is, the number of those whose event time is not less than t_i. Further assume that d_i failures happen at time t_i. A natural estimator of one such conditional probability is

$$P\Big(Z > t_i \,\Big|\, Z > t_{i-1}\Big) \approx \frac{Y_i - d_i}{Y_i}.$$

This motivates the definition of the Kaplan–Meier estimator:

$$\widehat{S}(t) = \begin{cases} 1, & \text{if } t < t_1, \\ \prod_{t_i \leq t} \left(\frac{Y_i - d_i}{Y_i}\right), & \text{if } t_1 \leq t \leq t_K \end{cases}.$$

The resulting function is a step function with jumps at the uncensored event times. In Kaplan–Meier plots census events are often indicated by tick marks.

It is not well defined what the estimate should be after t_K. As one will see, in our data the very last observation is a death, which makes the Kaplan–Meier estimator jump down to zero. This single observation might be treated as an outlier[19].

Although the formula $H(t) = -\log S(t)$ motivates an estimator for the cumulative hazard function: $\widehat{H}(t) = -\log \widehat{S}(t)$, there is an alternative which performs better on small sample sizes.

The *Nelson–Aalen estimator* of the cumulative hazard was first suggested by Nelson in 1972[22], then it was reinvented by Aalen in 1978[1]. Its definition is as follows:

$$\widetilde{H}(t) = \begin{cases} 0, & \text{if } t < t_1, \\ \sum_{t_i \leq t} \frac{d_i}{Y_i}, & \text{if } t_1 \leq t \leq t_K \end{cases}.$$

Based on the Nelson–Aalen estimator one can derive an alternative approximation of the survival function: $\widetilde{S}(t) = \exp(-\widetilde{H}(t))$.

The hazard function h can be estimated by the jumps of $\widetilde{H}$, but one always needs to smooth this crude estimate by some weighted averaging over neighbourhoods. (See e.g. Klein & Moeschberger[17], Section 6.2 to learn about smoothing with parametric kernels.)

3.4 Model choice

We summarise some methods used in model selection which we will later rely on.

Our investigations will be based on comparisons of maximum (log-)likelihoods (ML) achievable with different models. Bigger likelihood generally means better fit of model to data, but one must never forget that a better fit might be a worse explanation if it is driven by unnecessary variables that are only good at fitting to the random variation in data instead of to the underlying relations.

In the case of nested models, when the simpler model is a special case of the more complex one (that is, constraining some parameters to certain values yields the simpler model), the addition of a new parameter almost always increases the maximal likelihood. Therefore the likelihood in itself cannot tell whether the introduction of the new variable improved the model. Let us introduce the three criteria for assessing the goodness of a model which we will later use in model selection.

Let $M_1 \subset M_2$ be two nested models with parameter vectors θ_1 and θ_2 (their dimensions are p_1 and p_2), respectively.

The *likelihood ratio test (LRT)* is a statistical hypothesis test. Under M_1, which is the null hypothesis, if the maximum likelihood estimates (MLE) of parameters are $\widehat{\theta}_1$ and $\widehat{\theta}_2$, then

$$\Lambda = 2 \log \frac{P\Big(\text{data} \,\Big|\, \widehat{\theta}_2, M_2\Big)}{P\Big(\text{data} \,\Big|\, \widehat{\theta}_1, M_1\Big)}$$

follows (in most cases) a chi-square distribution: $\Lambda \sim \chi^2_{p_2-p_1}$. Therefore if Λ is bigger than the upper quantile corresponding to a pre-set p-value, then we can reject the null hypothesis M_1 in favour of the more complex M_2. (See — among very many alternatives — the review by Whelan, Liò & Goldman (2001)[32].)

The other two criteria penalise the inclusion of new parameters, and simply choose the bigger penalised log-likelihood value as the better model. For sample size n, models $i = 1, 2$, the *Akaike information criterion (AIC)* uses the expression

$$S_i = \log P\Big(\text{data} \,\Big|\, \widehat{\theta}_i, M_i\Big) - p_i,$$

the *Bayesian information criterion (BIC)* uses

$$S_i = \log P\Big(\text{data} \,\Big|\, \widehat{\theta}_i, M_i\Big) - \frac{1}{2} p_i \log n.$$

4 Patient population

The most fundamental information about the 1347 hematopoietic cell recipients in our study is summarised in Table 1. (An index of abbreviations can be found in Appendix A.)

Mean age of patients (years)	33.0
Range	0.7–65.7
Mean age of donors (years)	37.3
Range	19.9–92.97

Gender (patient)		*Diagnosis*	
Male	682	CML	988
Female	495	ALL	138
Unknown	170	AML	134
Gender (donor)		MDS	44
Male	650	AA	18
Female	426	NHL	8
Unknown	271	MPS	4
Gender (patient/donor)		MDS/MPS	3
Male/male	410	CLL	2
Male/female	215	Autoimmune disease	1
Female/male	238	Myeloma	1
Female/female	211	Other or not specified	6
Either unknown	273		

Table 1: Characteristics of donors and patients.

Most studies on HCT survival data use non-parametric methods (like the Kaplan–Meier analysis) and semiparametric methods (like the Cox proportional hazards model[5,17]) to assess the effects of allele or serotype mismatches[10,11,13,18]. We present some Kaplan–Meier plots for illustration purposes (Figures 1, 2 and 3), but we will not do the formal statistical tests that would be needed to use these for rigorous inference. We will instead develop a fully parametric model.

4.1 Non-parametric descriptions

Figure 1 clearly shows that different diagnoses give differing prospects for survival. One may think this is not necessarily directly related to disease, but it may be a consequence of the slower progression of CML, which allows more time to search for a highly matched donor. This does not seem to be the case: among the 1050 pairs for whom DPB1 types are known, on average the 787 CML patients had 1.42 2-digit and 2.08 4-digit mismatches, while the other 263 patients had a mean of 1.36 and 1.87 2- and 4-digit mismatches, respectively.

This finding will lead us to the stratification of the data by diagnosis in some instances. When large sample volumes are important, we will not differentiate between different diseases.

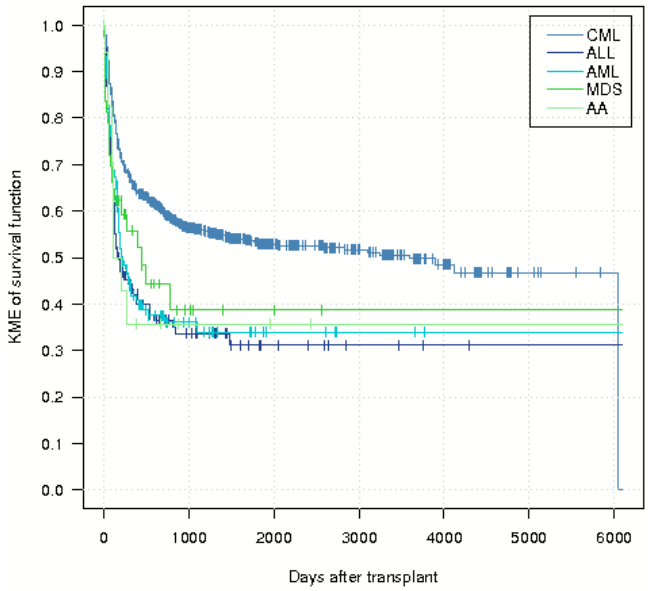

Figure 1: Kaplan–Meier estimates of the survival function for different diseases.

The busy Figure 2 gives some assessment of the detrimental effects of HLA mismatches on patients' prospects. We restricted the plotting to the 787 CML patient–donor pairs who had been HLA-DPB1 typed to obtain a homogeneous cohort of patients. Table 2 gives the ranking among strata with different numbers of matches at certain times after transplant.

A trend is palpable — more mismatches mean higher risk. This holds both for 2-digit and 4-digit mismatches.

It is now a straightforward idea to plot something similar for each six typed HLA gene. In order to assess the effect of mismatches at one locus, one should restrict the investigation to those pairs, who had no mismatches at other HLA genes. Otherwise the random occurrence of mismatches at other loci would have an effect on survival that might well be greater than the one of the locus under consideration. We cannot even assume uniform distribution of mismatches at other loci (see Table 3).

Unfortunately, choosing pairs who were typed for all six genes, picking one gene whose effects we want to study, and further restricting the population to those who had no mismatches anywhere but (potentially) at this gene can only be done for the HLA-DPB1 gene, if one wants a sensible population size (Figure 3). The reason is in Table 3: if any other gene is picked, it is likely, that there will be a mismatch at HLA-DPB1, and additionally, there might be a few elsewhere. Indeed, for HLA-C there are only 12 pairs (5 of which is uncensored) that had at least one 4-digit mismatch at C, but none elsewhere. Similarly, for HLA-DQB1 there are 7 pairs (4 uncensored) with at least one 4-digit mismatch at DQB1 and none elsewhere, and even less for the remaining sites, HLA-A, -B and -DRB1.

If one is not so cautious and uses data from every pair without considering disease

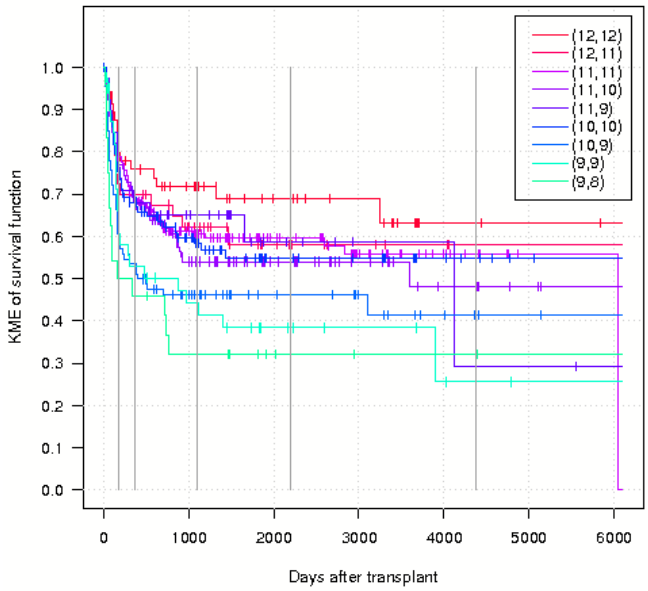

Figure 2: Kaplan–Meier estimates of the survival function for HLA-DPB1 typed CML patient–donor pairs stratified by number of matches (2-digit, 4-digit). A curve is drawn if there were at least 15 different days after transplant on which uncensored event occurred. The vertical lines are at years 1/2, 1, 3, 6, 12.

and the number of mismatches at genes that are different to the one on which the stratification is based, one gets plots in which one 4-digit mismatch gives worse probability of survival than zero mismatch, except for loci HLA-A and -DQ, where differences are marginal. Two allelic mismatches are either too rare to base inference on those few patients (HLA-A, -DRB1, -DQB1), or show roughly the same effect as one mismatch, or only marginally worse (HLA-B, -C, -DPB1). (Data not shown.)

Figure 3 shows the effect of allelic mismatches on survival at HLA-DPB1 among the homogeneous CML diagnosed group (787 patients), and it also compares it to the effect of one mismatch elsewhere. It suggests that no mismatch is certainly better for a patient than any DPB1 mismatch. We get the same for AML patients (108 patients, graph not shown). On the CML plot 2 mismatches seem to give a better chance of survival than 1 mismatch, but on the whole DPB1 typed population (1050 patients), among AML patients, and among the 85 ALL patients the difference vanishes (graphs not shown).

Surprisingly, in the AML group no mismatch seems to be an omen of bad luck compared to one or two HLA-DPB1 allelic mismatches (graph not shown) — this might be due to the small sample size and random fluctuations.

Both among CML patients and more pronouncedly for the whole population (graph not shown) one allelic mismatch at DPB1 might give worse prospects until about 3 years after transplant than a mismatch elsewhere. After three years this difference disappears.

These observations are for information only, without statistical testing they do not form a basis for conclusive claims.

Years	1/2	1	3	6	12
Ranking	(12,12)	(12,12)	(12,12)	(12,12)	(12,12)
	(11,11)	(12,11)	(11,9)	(11,11)	(12,11)
	(11,10)	(11,11)	(12,11)	(11,9)	(11,11)
	(11,9)	(11,10)	(11,11)	(12,11)	(10,10)
	(10,10)	(10,10)	(10,10)	(10,10)	(11,10)
	(12,11)	(11,9)	(11,10)	(11,10)	(10,9)
	(9,9)	(10,9)	(10,9)	(10,9)	(9,8)
	(10,9)	(9,9)	(9,9)	(9,9)	(11,9)
	(9,8)	(9,8)	(9,8)	(9,8)	(9,9)

Table 2: Ranking with respect to estimated probability of survival until 1/2, 1, 3, 6, 12 years after transplant of nine strata of HLA-DPB1 typed CML patients, stratified by the number of matches with donor HLA types (2-digit, 4-digit).

	HLA-A	*HLA-B*	*HLA-C*	*HLA-DRB1*	*HLA-DQB1*	*HLA-DPB1*
0	93.9%/86%	96.9%/83.8%	70.5%/65.4%	99.6%/93.8%	95.4%/89.3%	22.9%/16.2%
1	5.9%/13.2%	3.1%/14.8%	26.3%/29.8%	0.4%/5.9%	4.5%/9.8%	56.6%/53.9%
2	0.1%/0.7%	0%/1.4%	3.2%/4.8%	0%/0.2%	0.1%/0.9%	20.5%/29.9%

Table 3: Empirical distribution of the number of mismatches in a transplantation in our full dataset for each typed HLA gene (2-digit mismatches/4-digit mismatches). At each locus only the transplants for which at least 2-digit codes were known are taken into account. Where third and fourth digits were missing we assumed no match with that gene. For this table where third and fourth digit types were ambiguous (that is, it was AB, meaning either 01 or 02), we assumed match with the other individual's AB or 01 or 02 allele.

4.2 Steps towards a parametric model

711 patients were censored and 636 were registered as deceased. Figure 4 shows the distributions of these event times. While censored observations are broadly scattered with a heavy tail (mean=1490 days, median=1200 d, standard deviation=1205 d, range=1–5841 d), uncensored observations are mainly concentrated to the first year (mean=278 d, median=128 d, st dev=503 d, range=2–6051 d).

4.3 Some thoughts on mismatches

Little is known about how to look at the structure of mismatches to find a good hematopoietic cell donor. Throughout our investigations we simply count the matches or mismatches at each locus and use these as explanatory variables.

Having a certain allele at a certain locus might mean that missing an allele at another locus is not harmful since the functional protein product of the former is so similar to the latter. Consequently, there might be (almost surely there is) some match-mismatch notion across HLA loci, even for the whole MHC of patient and donor. Discovering this would be the ultimate target, but today it seems beyond our reach.

We do not know much about *symmetricity* in tissue compatibility in the sense that if, for instance, Tweedledum has a genotype such that the Mad Hatter could poten-

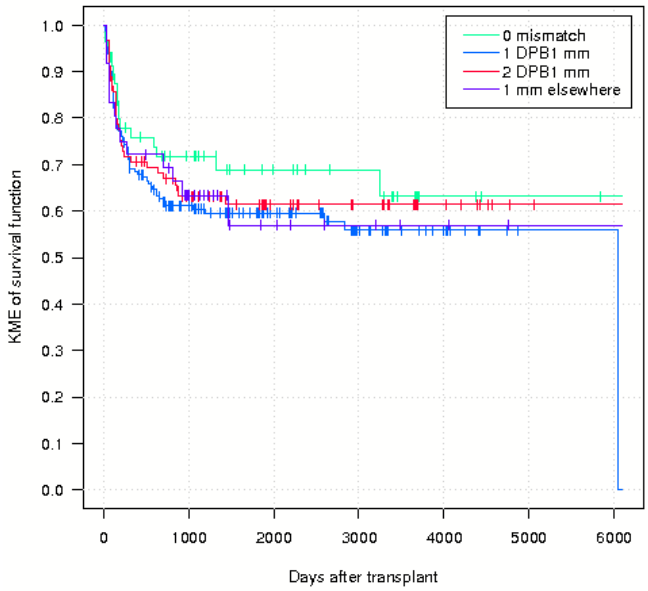

Figure 3: Kaplan–Meier estimates of the survival function for HLA-DPB1 typed CML patient–donor pairs stratified by number of 4-digit mismatches at the HLA-DPB1 locus: the first three strata had no allelic mismatch at other loci, the last stratum consists of all who had one allelic mismatch anywhere but at HLA-DPB1.

tially recover from a HCT with Tweedledum's cells, can Tweedledum live with the Mad Hatter's hematopoietic cells, if this reverse direction transplant is carried out?

In blood donation we know the answer for ABO groups: it is asymmetric. We know how this follows from which phenotype makes which antibodies, or rather, which antibodies are not made in certain types.

In HCT every allele type is expressed, so one does not expect similar asymmetries.

Let us introduce the notion of *non-detrimental (non-deleterious) mismatches*. If one looks at the recipient–donor pairs with a survival time more than, say, ten years, one may say that the mismatches these pairs had are tolerable. One could collate a list of these non-detrimental mismatches and count them as no mismatches.

Again, this approach is based on the assumption that effects of mismatches at different HLA loci are independent, and alleles at one site do not have an effect on how tolerable mismatches are at other sites. The idea could be further developed by assigning weights to each pair or quadruplet of alleles for any gene for a patient–donor pair: instead of having 0, 1 or 2 scores as mismatch counts, one could assign continuous values.

Note that the notion of non-deleterious mismatch is not time-symmetric. Mismatches from patients who died very early are not necessarily harmful, because those patients might have died from other transplant-related complications, or from their original disease.

Unfortunately, in the few instances when we tried using the idea of non-detrimental mismatches in model fitting, it did not improve our results, but made them worse.

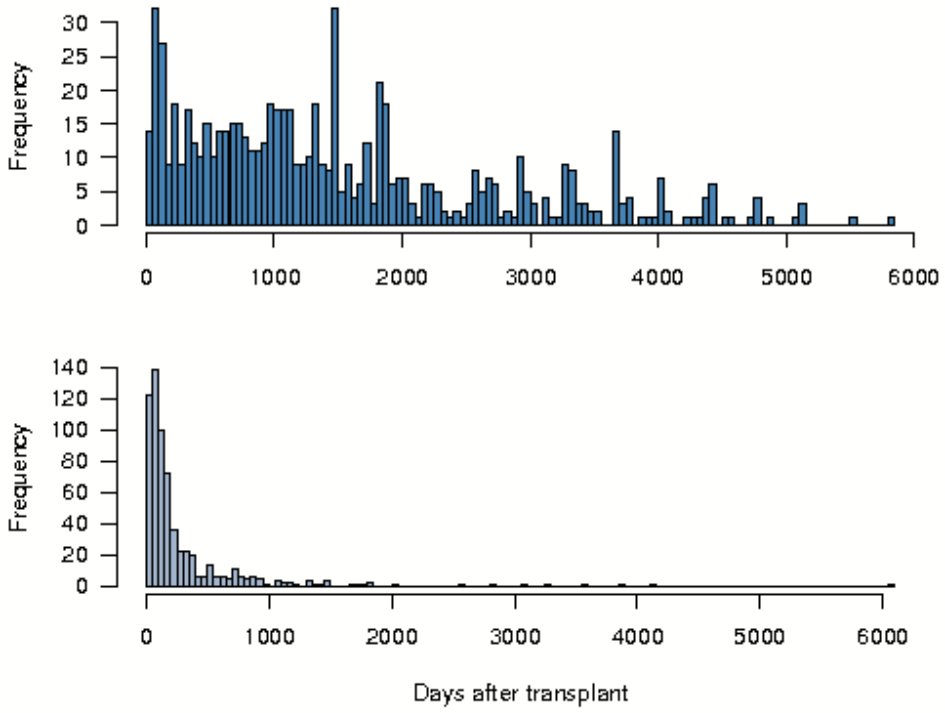

Figure 4: Distribution of time until event. The observed event is either census (top), or death (bottom). Note the different scaling on y axes.

If we use the non-detrimental mismatch approach to investigate recipient–donor symmetry, we can find pairs of recipient–donor pairs where the same mismatch occurred 'in both directions': once recipient had alleles (a_1, a_2) and the donor had (b_1, b_2), on the other occasion the donor had (a_1, a_2) and the recipient (b_1, b_2). This never happens in blood donation. There are examples for symmetric mismatches at HLA-A, -C and -DPB1 genes where both patients were censored after more than 5 years post-transplant: A*0201, A*6801 and A*0201, A*6901 (1 match); Cw*0304, Cw*0401 and Cw*0401, Cw*0702 (1 match); DPB1*0201, DPB1*0301 and DPB1*0401, DPB1*0401 (0 match).

4.4 Partitioning to training and testing data

In order to be able to assess the predictive power of any statistical model we will have proposed, we made preparations for cross-validation. The data was partitioned into a training and a testing set. Later some homogeneous cohorts were partitioned as well, such as patients with same disease, patients with DPB1 types. The ratio of the sizes of the two sets is always two to one.

Two partitioning strategies were used which we call $\mathcal{P}_1$ and $\mathcal{P}_2$. In both cases the censored and uncensored patients were treated separately. In $\mathcal{P}_1$, after sorting each group by the time to event, starting from the smallest, two patient–donor pairs from every successive three were randomly chosen and put into the training set, while the third was assigned to the testing set. Under $\mathcal{P}_2$ simply two thirds of both the censored

and uncensored pairs were randomly picked and put into the training set.

These single random choices were then fixed to form a basis for future model comparisons.

5 Statistical modelling of the data generating process

We assume that survival times of patients are independent random variables which are all realisations of a parameterised, but otherwise single distribution family. For each individual, parameters of this distribution are planned to be estimated by modifying a general description of the whole sample with a function of individual covariates (diagnosis, age, gender, HLA matching between donor and patient).

Note that some further important factors are unknown and cannot be used among the explanatory variables. Probably some stratification by disease progression (pre-transplant risk status) would have been useful, similarly stratification by pre- and post-transplant regimens, or by year (or decade) when transplantations were carried out (the last two are obviously connected).

5.1 Competing risks model

Understanding and modelling the censoring process are crucial to make good use of the survival-type data. According to Klein & Moeschberger[17] (Section 3.2), the most common right censoring strategies are

- censoring after a fixed time has passed since the subject had entered the study,
- censoring at a fixed terminal point, in which case subjects may enter the study at different times,
- censoring at the time when the failure number reaches a predetermined proportion of subjects.

As we did not know anything about the censoring used in generating our dataset, we assumed random censoring. Under this assumption census is a *competing risk* to death.

Generally, competing risks models are used to describe that a subject may encounter different failures (e.g. an individual may contract heart disease or cancer). Here we assume every recipient had an unknown random time to census, and an independent, unknown time to death. Whichever is first is observed and recorded. Hence the observation is the minimum of two independent random variables and the indicator which one was actually observed.

Note the symmetry here. Under this competing risks model the death can be seen as a censoring event too: it hides the actual time of the census. We will exploit this duality in the next section.

5.2 Reference distributions of the time until either of the events

At this stage we try to find two families of distributions that mimic the distribution of the observed census and death times reasonably well. We write F_C for the cdf of our approximation to the time until census, and F_D for the cdf of the time until death. We search for these distributions among the absolutely continuous distributions, and we denote their pdfs by f_C and f_D, respectively. At present we suppose every patient has the same universal F_C and F_D.

In order to find these distributions, the respective hazard functions were estimated from the observed data by the non-parametric Nelson–Aalen estimator (which was introduced in Section 3.3). Figure 5 shows these estimates after smoothing. With the same

technique, estimates of the two pdfs can be drawn as well (graph not shown). These visualisations guided us to find appropriate parametric distributions for our statistical model.

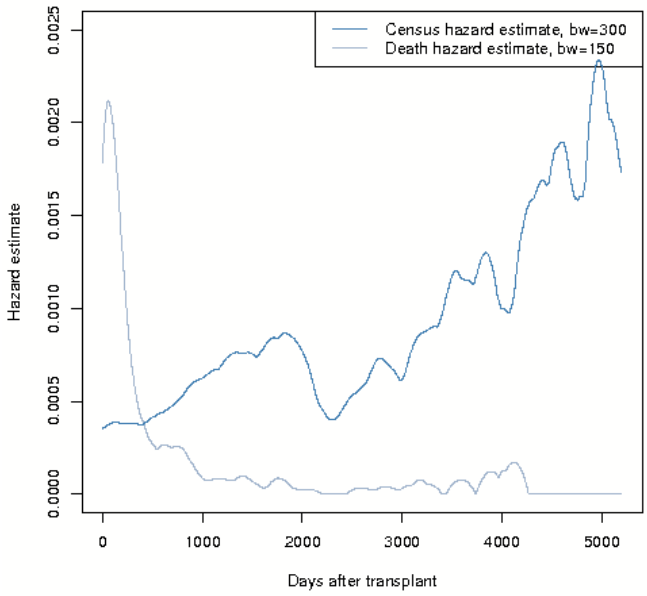

Figure 5: Nelson–Aalen estimates of hazard functions smoothed with an Epanechnikov kernel. Bandwidths of the smoothing (in days) is given in the legend.

The maximum likelihood approach is used to find well fitting distributions. C_i ($i \in \{1, \ldots, I\}$) denoting the days when census happened, D_j ($j \in \{1, \ldots, J\}$) the days of uncensored events, the likelihood* of the data is

$$L \propto \prod_{i=1}^{I} f_C(C_i)\Big(1 - F_D(C_i)\Big) \prod_{j=1}^{J} f_D(D_j)\Big(1 - F_C(D_j)\Big).$$

By reorganising,

$$L \propto \left(\prod_{i=1}^{I} f_C(C_i) \prod_{j=1}^{J} \Big(1 - F_C(D_j)\Big)\right) \left(\prod_{i=1}^{I} \Big(1 - F_D(C_i)\Big) \prod_{j=1}^{J} f_D(D_j)\right). \qquad (1)$$

Because the first factor only depends on F_C, while the second one only on F_D, the likelihood can be maximised by independently maximising the first and the second factors.

In order to ascertain that forthcoming calculations would be kept simple, we were looking for distributions that have at most two parameters.

*We write proportionality instead of equality because we omit a constant factor which comes from the time ordering of our observations and which does not depend on C_is, D_js or any unknown parameters. An explanation of this with further references can be found in Maller & Zhou (1996)[19], pp 98–99.

Maximal log-likelihood values for a few different distributions are presented in Table 4. We found that from our proposed distributions the log-normal distribution gave the best fit for the time until death, and the one with linear hazard function for the day of census.

Distribution of C	*Value for 1st factor*	*Value for 1st factor, $\mathcal{P}_1$ tr. set*
$h_C(t) = at + b$	-5979	-3987
Gompertz(α,β)	-5980	-3987
Weibull(k,λ)	-5982	-3989
$h_C(t) = at^2 + b$	-5984	-3990
$h_C(t) = at^3 + b$	-5988	-3993
Gamma(α,β)	-5988	-3993
Exponential(λ)	-6016	-4011
Log-logistic(α,μ)	-6044	-4031
Log-normal(μ,σ)	-6084	-4061

Distribution of D	*Value for 2nd factor*	*Value for 2nd factor, $\mathcal{P}_1$ tr. set*
Log-normal(μ,σ)	-5075	-3383
Log-logistic(α,μ)	-5107	-3405
Weibull(k,λ)	-5156	-3437
Gamma(α,β)	-5192	-3462
$h_D(t) = at + b$	-5357	-3572
Exponential(λ)	-5452	-3635

Table 4: Maximum log-likelihood values of the two factors of Equation 1 with different distributions assumed. The first table gives $\max \log(\prod_i f_C(C_i) \prod_j (1 - F_C(D_j)))$ values, while the second one gives $\max \log(\prod_i (1 - F_D(C_i)) \prod_j f_D(D_j))$ values. In rows with h_C or h_D the respective distributions are derived from the given hazard functions which uniquely determine them, as it was stated in Section 3.2. The middle column gives these values for the whole dataset, while the last column only uses the data of the $\mathcal{P}_1$ training set (as explained in Section 4.4).

Figures 6 & 7 show Kaplan–Meier estimates of the survival function against the survival functions of the two best and some other approximations when census or death is the uncensored event type, respectively. Parameters for the distributions are the maximum likelihood estimates.

The model fit to time-to-census data is very good, but this turns out to be unimportant: as we believe that genetic information only affects time-to-death, and not time-to-census, we will no longer work with the first factor of Equation 1.

The fit of our models to survival time is poor yet, therefore we have to carry on searching for a better model.

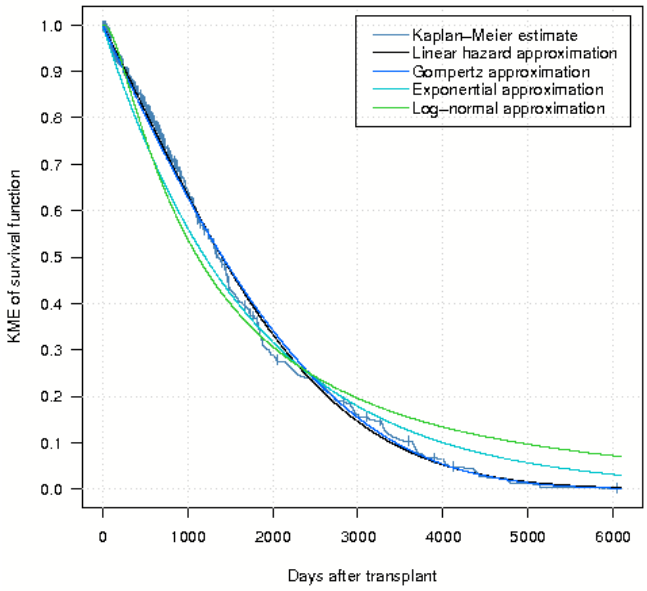

Figure 6: Kaplan–Meier estimate of the survival function for the whole patient population when the census plays the role of death and actual death is treated as census. Survival functions of some parametric models of the random time until census.

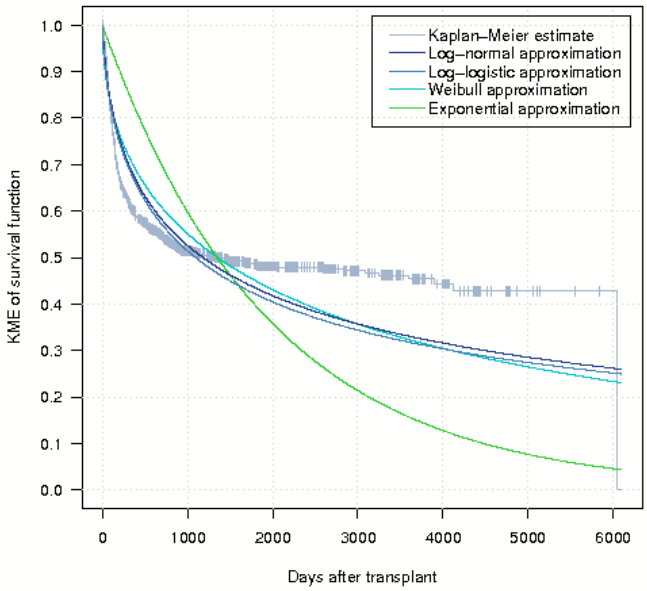

Figure 7: Kaplan–Meier estimate of the survival function for the whole patient population and some of its parametric approximations.

5.3 Mixture models for survival time: mixture model with immunes

The Kaplan–Meier curves that describe the surviving proportion of the patient population seem to level off after some years, modulo we exclude the last patient from our data. This observation suggests there are (as Maller & Zhou call them) *immunes* in the population[19], that is, patients who are cured. This assumption leads to a *mixture model with immunes*.

Mathematically, this means that only a p proportion of the population is mortal ($0 < p < 1$) and, *under this model*, $1 - p$ proportion of the patients are immortal: they will never die, or equivalently, their survival time is ∞. As a consequence, the new cdf F_D^* will specify a subdistribution, which is almost the same as a distribution, with the exception that

$$\lim_{t\to\infty} F_D^*(t) < 1.$$

Given any cdf F_D we have been working with so far and a p ($0 < p < 1$), one can easily derive the new F_D^* by setting $F_D^* = pF_D$. It follows that there exists a pdf, and it is $f_D^* = pf_D$. The likelihood to maximise is

$$L \propto \prod_{i=1}^{I} \Big(1 - F_D^*(C_i)\Big) \prod_{j=1}^{J} f_D^*(D_j) = \prod_{i=1}^{I} \Big(1 - pF_D(C_i)\Big) \prod_{j=1}^{J} \Big(pf_D(D_j)\Big).$$

Now there is the extra p variable in the maximisation in addition to the parameters of F_D.

If one does the log-likelihood maximisation for all the distributions which have been tested so far, one will find that the one with maximum log-likelihood is the log-normal distribution again (log-likelihood of the whole dataset: -4968), followed by the log-logistic distribution (-4970), the improper Gompertz mixture model with immunes (-5003), the Weibull distribution (-5035), the improper Gompertz distribution without immunes (-5038), and others. (The improper Gompertz is already a subdistribution even without introducing immunes to it.[19]) p is estimated 0.53–0.54 unanimously.

The fit of these models to the data is shown in Figure 8. This result is still not satisfactory, particularly because of the existence of immortals, which is an assumption one is reluctant to make.

5.4 Mixture model with two distributions

The best description of the random time until death we found is a mixture model, which models the distribution of the time-to-event as a mixture of *early failure* and *late failure* distributions. Let F_E and F_L denote their cdfs, both having pdfs: f_E and f_L. Here we assume that a q ($0 \le q \le 1$) proportion of patients die from early, and $1 - q$ die from late failure.

At this low level of complexity it is probably not essential to define what we model with early and late failures. However, we can still say some words about it: early should mean lethal complications from the transplantation (e.g. sinusoidal obstruction syndrome, transplantation-related lung injury, transplantation-related infections, immunodeficiency caused by acute GVHD and its treatment with corticosteroids might lead to life-threatening fungal infection, graft rejection, relapse), while late failure might mean both transplant-related causes (immunodeficiency caused by prolonged corticosteroid

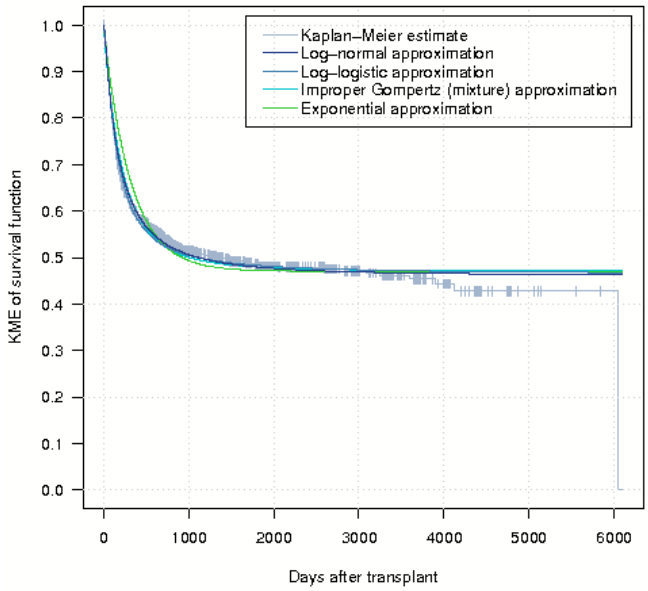

Figure 8: Kaplan–Meier estimate of the survival function for the whole patient population and some of its mixture model approximations with immunes.

treatment of potential chronic GVHD, secondary cancers)[4,8], and other causes of death that occur independently of the transplant.

Under this model the cdf of the random survival time is $F_D = qF_E + (1-q)F_L$, the pdf is $f_D = qf_E + (1-q)f_L$. We now typically have five parameters: two for each of F_E and F_L, and q. The likelihood of the data under this model is

$$L \propto \prod_{i=1}^{I} \Big(1 - qF_E(C_i) - (1-q)F_L(C_i)\Big) \prod_{j=1}^{J} \Big(qf_E(D_j) + (1-q)f_L(D_j)\Big).$$

We assume that both the early and the late failure distributions are one of the following distributions: log-normal, log-logistic, Weibull, gamma, normal, exponential.

Looping through the 36 possible early/late distribution combinations tells us that under the mixture model the maximum achievable log-likelihood of the whole dataset is -4953, and about seven combinations can reach this. We picked the log-normal/log-normal (which was the best with a log-likelihood value of -4952.8, $q = 0.40$, $\text{mean}_E =$ 188 days, mean_L=9700 years) and the log-normal/exponential models (which was seventh with -4953.8, $q = 0.47$, mean_E=238 d, mean_L=62 y) based on their previous performance and their simplicity.

Figure 9 shows their fit to the Kaplan–Meier estimate of the global survival function.

A comparison of our four best models (the log-normal, the log-normal mixture model with immunes, and the two latest mixture models) based on LRT (Section 3.4) shows that the introduction of new parameters with the mixture models gave a significant ($P < 0.001$) improvement in model fit over the non-mixture model, and AIC and BIC values

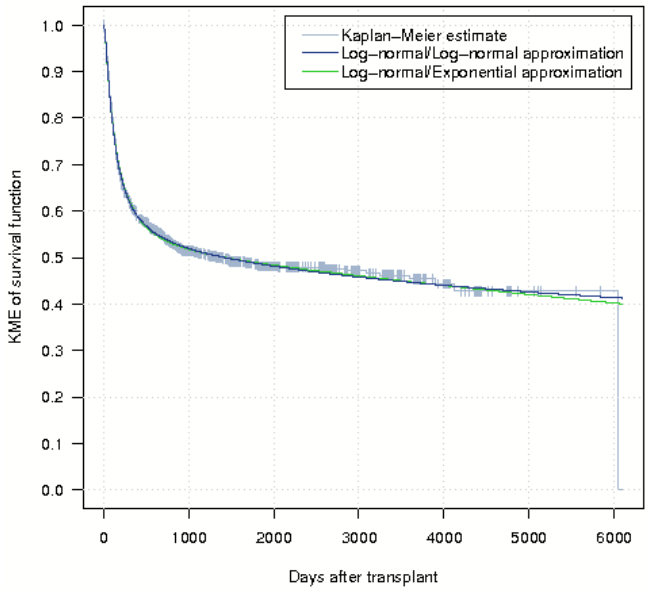

Figure 9: Kaplan–Meier estimate of the survival function for the whole patient population and two mixture model approximations. (The legend gives early failure/late failure distributions, in this order.)

also strongly support the mixture models. With the AIC the log-normal/log-normal, with the BIC the log-normal/exponential is the optimal choice. (Data not shown.)

Assessing predictive performances by estimating the parameters on the training set with MLE, and using these parameters to calculate the log-likelihood of the test set, shows that the mixture models are always better than the basic model, both on the whole dataset and on the three biggest patient cohorts (CML, ALL, AML), when using either $\mathcal{P}_1$ or $\mathcal{P}_2$ partition. One cannot set an unequivocal ranking among the mixture models based on their predictive performances. (Data not shown.)

These tests and the advantage of no need for the modelling assumption of immunes made us use the two latest mixture models.

6 Modelling genetic effects

We reduced the modelling of individual genetic background to counting the number of 4-digit mismatches for each patient–donor pair at each HLA locus and deriving a single value from these.

For pair k the number of mismatches at the six loci are $X_{k,1}, X_{k,2}, \ldots, X_{k,6} \in \{0, 1, 2\}^{\dagger}$. When all types are missing at a locus, it is treated as zero mismatch. Apart from early experiments we always assumed the linear model

$$\beta_0 + \sum_{r=1}^{6} \beta_r X_{k,r}$$

to describe the effect of mismatches for patient k, because this formula resulted in the greatest ML values.

In our mixture models a successful transplant is the one after which the patient dies from late failure (the transplant gives years to their lives). Hence we made our model such that mismatches push patients towards early failure by increasing their individual q through the well-known logistic model:

$$q_k = \frac{1}{1 + \exp\Big(- (\beta_0 + \sum_{r=1}^{6} \beta_r X_{k,r})\Big)}.$$

We did not include other covariates such as gender or age in our models.

Now our model is wholly defined. This is what was used in a variable selection process to find the genes whose mismatches contribute the most to early deaths of HCT recipients.

Putting a variable into a model means we let its β_r coefficient vary. Leaving it out means we fix $\beta_r = 0$. As there are 6 loci, there are $2^6 - 1 = 63$ possibilities to include at least one variable. We tried fitting each alternative model by letting the chosen β_rs and the parameters of F_E and F_L vary. We then compared the resulting log-likelihood values by the LRT, AIC and BIC methods.

Before sharing these results, we describe a test which showed that it is not only noise we are fitting our models to.

Before developing our mixture models we were using the most basic log-normal model. Genetic effects were introduced to the model by multiplying the hazard function of the log-normal distribution by the individual-dependent factor

$$1 + \sum_{r=1}^{6} \beta_r X_{k,r}.$$

Using all six β_rs one can determine the ML value of the survival time for the whole patient population and compare this to the ML value with no β_rs, based on AIC or BIC comparison. The AIC values from the MLEs on both the whole set and the training set ($\mathcal{P}_2$ in this case) is in favour of introducing the six new variables, such as the AIC on the whole dataset. (The BIC comparison, which puts more penalty on new variables, prefers no β_rs on the training set.)

[†]More precisely, between the ambiguous type AB (which is either 01 or 02) and AB or 01 or 02 we counted 0.5 mismatch.

Using partition $\mathcal{P}_2$ to assess the predictive performance on the test set[‡] we get that the prediction with β_rs that were optimised for the training set only is much better than a prediction without β_rs.

The trick was that we carried out the same investigations trying to predict the time until *census* from HLA information. As we expected, it gave poor results: both the AIC (modestly) and BIC (very pronouncedly) showed that the variables did not improve the fit significantly. Prediction on the test set based on training on the training set improved with the introduction of the β_rs, but only marginally, and we think this has happened by chance. (Data not shown.)

6.1 Computational methods

We ran the variable selection loop through the 63 possibilities for the training set of the whole dataset. Later we did this for the training sets of strata of the whole data (patients stratified by diagnosis).

In each such loop we fitted the model to the training set by each partitioning strategy, $\mathcal{P}_1$ and $\mathcal{P}_2$.

As we did not decide between them, we fitted both the log-normal/log-normal and the log-normal/exponential mixture models.

To increase our chance of finding the real maximum of the log-likelihood, we initialised the iterative numerical optimisation with two slightly different parameter settings (β_0, β_rs and parameters of the early and late failure distributions) for two separate runs.

Altogether, for each studied patient cohort and each 63 variable combinations we ran $2 \times 2 \times 2 = 8$ optimisation algorithms. For the whole population this took about 30 hours on a desktop computer.

For each 63 combination of variables, for each of the two mixture models and two partitions the smaller ML value from the two runs was discarded. Within the results of each four set of variable selection loops, the 63 variable combinations were listed in order of their AIC or BIC values on the respective training sets separately.

For each of the two mixture models, and for each of the two ranking aspects (AIC or BIC) separately, a consensus ranking was compiled from the two partitions.[§]

Then compared these results between the two different mixture models, chose variable combinations which were among the best for both models, and collated a list of these best variable combinations for both AIC and BIC separately.

Typically, there was not much difference between rankings with log-normal/log-normal versus log-normal/exponential mixture models, but there was considerable variation between the rankings with $\mathcal{P}_1$ or $\mathcal{P}_2$ partitions.

Running the numerical optimisation with different initialisations proved very important, because the differences for the same computation with different initial values differed with a quantity comparable, or just slightly less than the breadth of the interval

[‡]This is the test for which we defined $\mathcal{P}_2$; $\mathcal{P}_1$ would have probably been too homogeneous for these tests to show strong results.

[§]We chose the best combinations by starting from the best ML values for either $\mathcal{P}_1$ or $\mathcal{P}_2$, and the variable combinations which were among the best with respect to their ranking in both partitions were selected. With this method we collected roughly the best five variable combinations, or we searched through the best 10 of 63 for both $\mathcal{P}_1$ and $\mathcal{P}_2$, and selected all which appeared in both.

spanned by the 63 ML values in some instances (but always for a minority of the 63 combinations only).

The probable effect of this numerical issue is that in a few cases a few good combinations of explanatory variables ended up with worse ML values than what they really had, and at worse positions in the ranking than what they had deserved. These combinations might have evaded our attention, but this almost certainly does not mean that what we believe the best combinations are, are poor in reality.

Further, we will compare the predictive performance on the test set of the best models with that of the model without genetic effects. This assessment does not rely on maximisation, so it is free of its pitfalls.

6.2 Results

In presenting our results about which combinations of covariates describe the variation in survival times best, we will first look for trends instead of an ultimate solution.

The variable combinations selection method presented in the previous section gave these as best combinations for the whole dataset, roughly in this order:

AIC (B,C,DP), (B,C,DR,DP), (B,C,DQ,DP), (B,C),

BIC (C), (B,C), (C,DP), (C,DR), (B).

(For brevity we omit the B1 from DRB1, DQB1 and DPB1.)

The most interesting thing is the frequent appearance of the DPB1 variable, since it is not known to have an effect on survival[13]. It is important to note that in this case pairs without their HLA-DPB1 types were assumed they had complete match at this locus.

A glimpse at Table 3 suggests that assuming one mismatch for them is more sensible. The results of the combination selection process with this assumption are the following:

AIC (B,C,DP), (B,C,DR,DP), (B,C,DQ,DP),

BIC (C), (B,C), (C,DP), (C,DR),

that is, the rankings of the best are the same, but in models with variable DPB1 log-likelihood values improve in almost all instances for $\mathcal{P}_2$, although only marginally. For $\mathcal{P}_1$ it is not so clear: more improve than how many decrease, and the mean change is positive.

We can make three general observations.

First, the best models under BIC have less variables than the best under AIC. This is simply explained by the fact that BIC puts more penalty on the inclusion of new covariates than AIC.

Second, B and C are the covariates with the most appearances. At least one is always among the variables, but C seems to be more important. It is known that there is strong linkage disequilibrium between HLA-B and -C genes[7], and a Pearson's chi-squared test shows contingency between the existence of mismatches at these two loci in our data ($P < 0.00001$). So we may well expect if mismatches at one locus show a negative effect on patients' survival, mismatches at the other will as well, because one mismatch at the other locus makes a mismatch (and a negative effect) more likely at the first locus.

Third is the surprising observation that HLA-A is not among the important variables. Actually, (A,B,C,DP) would be the next best combination under the AIC, but with $\beta_1 < 0$ — that is, a mismatch at HLA-A would improve patients' prospects. We immediately rejected this model.

A closer inspection tells us that in more than half of the 32 combinations where A appears, it has a negative coefficient, and independent of the sign the coefficient is about one order of magnitude smaller than the other β_rs. This does not imply that HLA-A matching is not an important determinant of tranplant outcome, but it does not seem to have much effect in our data. Table 3 shows that there are few pairs with A disparity, particularly few with two mismatches. This might explain why A is in such a controversial role.

To further investigate the effect of DPB1 mismatches, we restricted our attention to the 1050 pairs who had this gene typed. The best covariate combinations now are

AIC (B,C,DP), (B,C,DR,DP), (B,C,DQ,DP),

BIC (C), (C,DP), (B,C), (B,C,DP), (B),

very similarly to previous results. (A,B,C,DP) would be the next in the AIC line, now with $\beta_1 > 0$ (but small). Anyway, it is never entirely clear where to end the list.

At last in this series, in order to filter the potentially distorting factor of diagnosis, results for the homogenised CML patient cohort are given (787 patients who were DPB1 typed):

AIC (B,C,DP),

BIC (C,DP), (B,C), (C).

The rankings are not shown for the 108 DPB1 typed AML and the 85 ALL patients because it turned out that the optimisation led to model overfitting: predictions on the test set were inferior to predictions without taking account of genetic variation. A closer inspection showed that the β values are unrealistically big.

Values for the β_0 intercept term are usually between (-0.5)–(-0.1), the β_rs ($r = 1, \ldots, 6$) are about 0.2–0.5, and the patient-specific q is typically between 0.3–0.8.

Among the ALL and AML patients β_r values were often hundredfold greater, and q_ks were distributed throughout the whole $[0, 1]$ interval.

The optimisation can be modified to keep β_rs under control by penalising their absolute values with formulæ like in ridge regression or lasso. Given more time, these investigations should be carried out.

6.3 Assessment of predictive power

A critical part of our model selection process is the testing of covariate combinations (and the complete model) in predicting transplant outcomes in the test set. For this, one determines the MLE of all parameters using data solely from the patients in the training set. Once this is done, one applies these parameters and the genetic covariates of patients of the test set to calculate the log-likelihood of the test set data.

If these log-likelihoods are consistently bigger then log-likelihoods derived from predictions made by the model which does not use individual genetic variation data, then one can be reasonably confident that the covariates increase the predictive power.

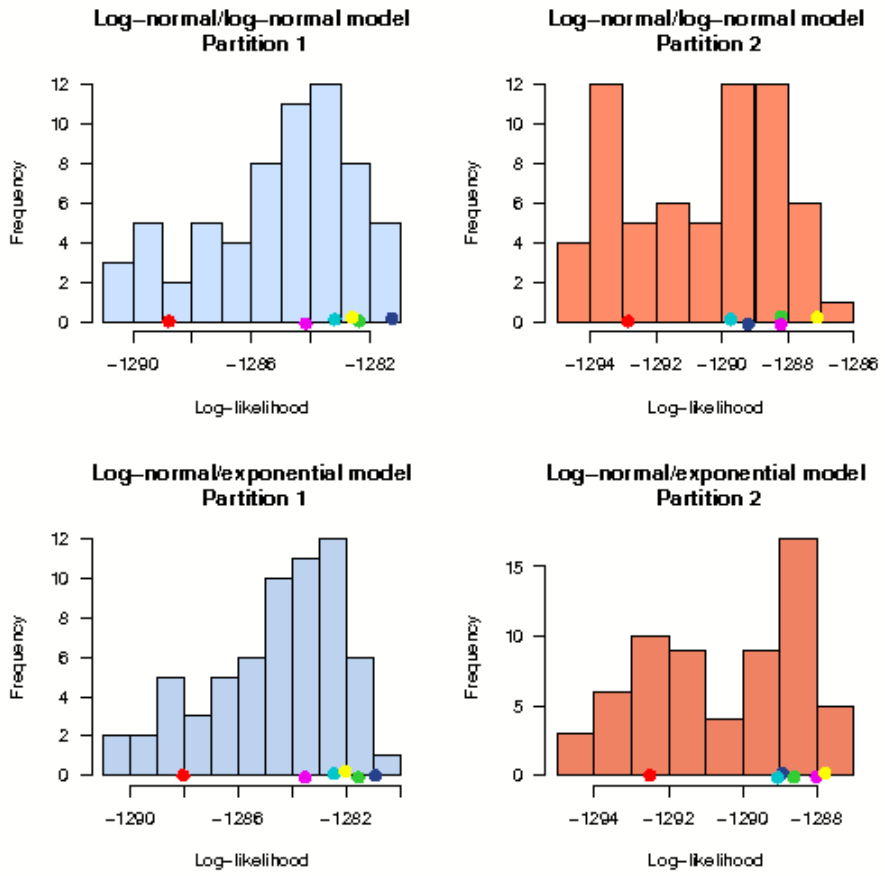

Figure 10: Log-likelihood landscape on the test set for all DPB1 typed recipient–donor pairs, for both mixture models and partitions $\mathcal{P}_1$ and $\mathcal{P}_2$: histograms of the 63 log-likelihood values for the test set using mismatch counts and parameter values from the training. Coloured blobs represent the value for the combinations (B,C,DR,DP)—green, (B,C,DP)—dark blue, (B,C)—yellow, (C,DP)—light blue, (C)—magenta. Log-likelihood of the prediction without mismatch covariates is in red for reference.

Figure 10 and 11 shows log-likelihood values of the test set of predictions from five of our best covariate combinations: (B,C,DR,DP), (B,C,DP), (B,C), (C,DP) and (C). Figure 10 uses the whole DPB1 typed population, and Figure 11 restricts this to the CML patiens.

Each five combination predicts reasonably well. More covariates typically mean more extreme behaviour compared to others: they are sometimes the best among the five, sometimes the worst, even falling below the reference without covariates. This behaviour indicates slight overfitting with too many parameters.

However, starting from the simplest, (C) is always better than the reference, such as (C,DP). These two are the most robust combinations of variables to describe the effect of genetic variation in our data.

If one has to choose only one, then (C,DP) seems to be the best compromise between good fit on training set, small number of covariates, and reasonable predictive power on the test set.

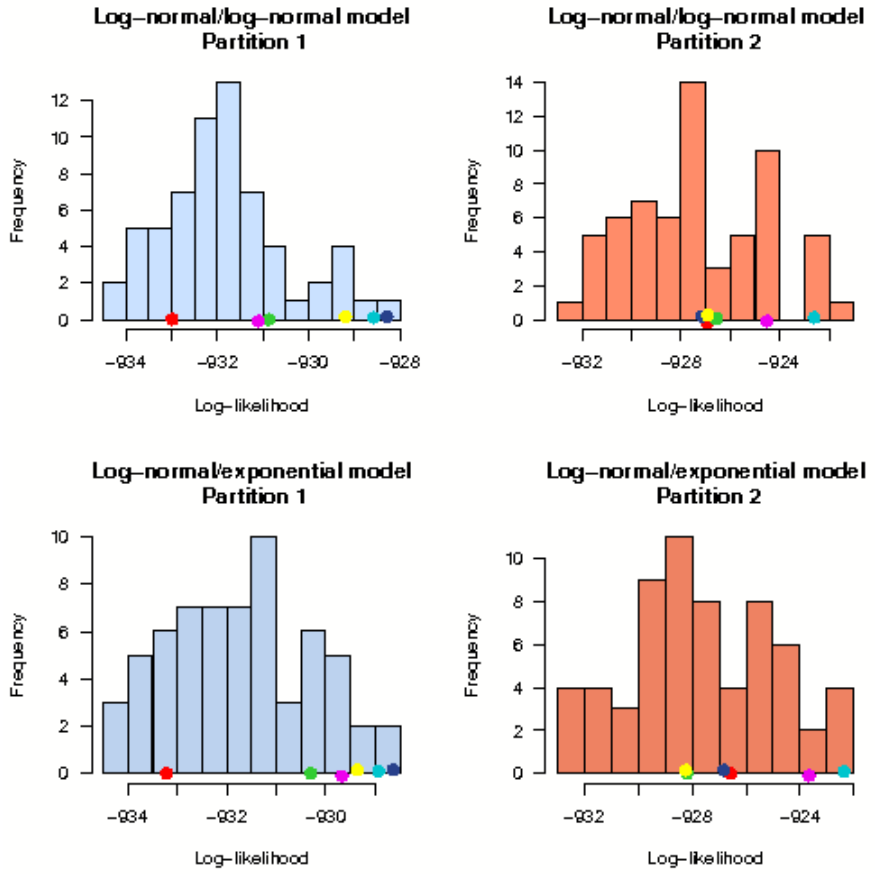

Figure 11: Log-likelihood landscape on the test set for all DPB1 typed CML patients, for both mixture models and partitions $\mathcal{P}_1$ and $\mathcal{P}_2$: histograms of the 63 log-likelihood values for the test set using mismatch counts and parameter values from the training. Coloured blobs represent the value for the combinations (B,C,DR,DP)—green, (B,C,DP)—dark blue, (B,C)—yellow, (C,DP)—light blue, (C)—magenta. Log-likelihood of the prediction without mismatch covariates is in red for reference.

A LRT shows that the inclusion of the C or the (C,DP) covariates brings significant improvement over the basic model. The p-values for the training set of the patients who were DPB1 typed are 0.00013 for C, 0.00007 for (C,DP). For the training set of DPB1 typed CML patients these values are 0.0016 and 0.0008, respectively. (All these numbers are the worst cases among the four different cases: two distributions, two partitions.)

The CML population is a very large subset of the whole population; it is not surprising that we get similar results on both sets. Therefore it would be very interesting to carry out these tests for ALL and AML patients with more attention to good penalty schemes.

7 Discussion and directions for future research

Our final model was a competing risk censoring model where census time and survival time are independent random variables, time-to-census has a distribution with an increasing linear hazard, and survival time is drawn from a mixture of an *early* log-normal and a *late* log-normal or exponential failure distribution. For each patient–donor pair the affine combination of the number of allelic mismatches at each locus gives the probability that the patient will die from an early failure through a logistic model.

Variable selection analyses repeated for different partitions to training and test data, and for the two slightly different probabilistic models, indicate that HLA-C is the most important genetic determinant of survival time after a hematopoietic cell transplant in the studied patient population.

Surprisingly, the second most important gene seems to be HLA-DPB1 (barring HLA-B, which is correlated with HLA-C both in the human genome and in their effect on HCT success).

Our results did not give a strong case for the importance of HLA-A or -DRB1 matching, but this is probably explained by bias in the data (Table 3). Hence we do not claim these loci are not important, but rather we say that HLA-DPB1 should get more attention from the scientific community. Figure 3 of Section 4.1 also supports our claim. As we have already expressed, given more time, this should be investigated with more rigour.

Similarly, the ALL and AML populations in the variable selection analysis should get attention once again.

At this low level of complexity one should not see any of these models as ends in themselves. Their predictive power is not good enough to rely on them on a case-by-case basis, but one should rather focus on the genes that were picked up by these models.

In correspondence with our finding, Greinix et al. (2004) cite five papers by four research groups which found that HLA-C exerts significant effect on graft failure and survival[11].

The NMDP's recommendations also include the matching of patient and donor HLA-C types, but they do *not* require matching HLA-DPB1 types.[29]

Continuing our investigations is particularly important because the negative effect of HLA-DPB1 disparities on survival after HCT (if there really is one) is overlooked.

7.1 Assessment of bias caused by differing mismatch distributions among HLA loci

The effects of the bias caused by the differences between the distribution of mismatch counts at different loci in our data could be investigated by simulation.

Assuming our mixture model (either the log-normal/log-normal or the log-normal/exponential), setting its parameters and the β_0 coefficient to values borrowed from our model fits, and setting all other β_r coefficients equal to each other (to simulate a hypothesis that all considered HLA loci have the same effect on survival time), we could generate survival (and census) times for patients based on the actual mismatch counts, either by bootstrapping or taking all patient–donor pairs' mismatch counts. Now, using these simulated event times and event types (census or death) we could do the usual

variable selection loop through the 63 possibilities and see which variables are picked most often as the ones which describe the individual variability best.

If HLA-C and -DPB1 are picked up among the most significant contributors much more frequently than they were in our investigations, it would suggest that the bias in mismatch counts in the data had much distortive effect on our findings. Consequently, although C and DPB1 may be important in determining transplant outcome, they would appear less important than A, B or DRB1 in the light of previous studies[11,13].

On the other hand, if C and DPB1 are chosen less frequently or with about the same frequency as they were in our previous variable selection loops, then it would suggest that C and DPB1 are as important factors as A, B or DRB1. According to the literature this latter case seems unlikely.[11,13]

A serious drawback of such a simulation would be the difficulty of interpreting its results, and the quantification of the effects caused by the bias in the mismatch distributions.

7.2 Future research

A much more complex study could be carried out by building our mismatch notion on amino acid sequences instead of allelic types. Much insight might be gained from comparing similarities and differences of proteins, even across genes, in understanding averse effects between graft and host. Our knowledge about the linkage disequilibrium pattern in the MHC region[7] could be used to search for potential further determinants of transplant success which are hidden between the typed HLA genes.

From a methodological point of view, future statistical investigations should be preferentially done in the Bayesian framework.

A most ambitious target would be the detailed mathematical modelling of the complex host–leukæmia–graft system[4,8], probably in a differential equations framework. Our current understanding of the underlying processes is insufficient to carry out such a venture, but modelling always helps to understand what is exactly what we do not know about a complex biological system yet.

Acknowledgements

The author wants to express his gratitude to Gil McVean for guiding his work, letting him do what he felt was important but always helping him when it was needed. The author is indebted to Niall Cardin whose hunch was to use a mixture of two distributions to model the time until death. This work could not have been carried out without financial support from the EPSRC through the Life Sciences Interface Doctoral Training Centre, University of Oxford.

A Abbreviations

A list of common abbreviations of the field of this study is presented for reference.

AA Aplastic anemia

AL Acute biphenotypic leukæmia

ALL Acute lymphoblastic leukæmia (Acute lymphoid leukæmia, Acute lymphoblastic anemia)

AML Acute myelogenous leukæmia

BMT Bone marrow transplantation, often more broadly includes cord blood transplantation

CLL Chronic lymphocytic leukæmia

CML Chronic myelogenous leukæmia

GVHD Graft-versus-host disease

HCT Hematopoietic cell transplantation

HLA Human leukocyte antigen

HSC Hematopoietic stem cell

HSCT Hematopoietic stem cell transplantation

MDS Myelodysplastic syndromes

MHC Major histocompatibility complex

MM Multiple myeloma

MOF Multiorgan failure

MPS-I Mucopolysaccharidosis I, Hurler's Syndrome

NHL Non-Hodgkin's lymphoma

PBSC Peripheral blood stem cells

SAA Severe aplastic anemia

TRM Transplant-related mortality

UCB Umbilical cord blood

UCBT Umbilical cord blood transplantation

URD Unrelated donors

B Methods

All calculations and graph plotting were done using the R statistical software package (version 2.5.1). The built-in `optim` function with the Nelder–Mead method was used to maximise log-likelihoods.

References

[1] O. O. Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726, 1978.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of The Cell*, chapter 24, The Adaptive Immune System, pages 1363–1421. Garland Science, New York and London, fourth edition, 2002.

[3] Graham S. Cooke and Adrian V. S. Hill. Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.*, 2:967–977, 2001.

[4] Edward A. Copelan. Hematopoietic stem-cell transplantation. *N. Engl. J. Med.*, 354(17):1813–1826, 2006.

[5] David R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[6] IMGT/HLA Database. `www.ebi.ac.uk/imgt/hla` .

[7] Paul I. W. de Bakker, Gil McVean, Pardis C. Sabeti, Marcos M. Miretti, Todd Green, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, 38:1166–1172, 2006.

[8] M. Devetten and J. O. Armitage. Hematopoietic cell transplantation: progress and obstacles. *Annals of Oncology*, 2007. doi:10.1093/annonc/mdm064.

[9] P. C. Doherty and R. M. Zinkernagel. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 256:50–52, 1975.

[10] Neal Flomenberg, Lee Ann Baxter-Lowe, Dennis Confer, Marcelo Fernandez-Vina, Alexandra Filipovich, Mary Horowitz, Carolyn Hurley, Craig Kollman, Claudio Anasetti, Harriet Noreen, Ann Begovich, William Hildebrand, Effie Petersdorf, Barbara Schmeckpeper, Michelle Setterholm, Elizabeth Trachtenberg, Thomas Williams, Edmond Yunis, and Daniel Weisdorf. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. *Blood*, 104:1923–1930, 2001.

[11] H. T. Greinix, I. Faé, B. Schneider, A. Rosenmayr, A. Mitterschiffthaler, B. Pelzmann, P. Kalhs, K. Lechner, W. R. Mayr, and G. F. Fischer. Impact of HLA class I high-resolution mismatches on chronic graft-versus-host disease and survival of patients given hematopoietic stem cell grafts from unrelated donors. *Bone Marrow Transplantation*, 35:57–62, 2005.

[12] International Histocompatibility Working Group. dbMHC database. `www.ncbi.nlm.nih.gov/mhc/` .

[13] Carolyn Katovich Hurley, Lee Ann Baxter Lowe, Brent Logan, Chatchada Karanes, Claudio Anasetti, Daniel Weisdorf, and Dennis L. Confer. National Marrow Donor Program HLA-matching guidelines for unrelated marrow transplants. *Biology of Blood and Marrow Transplantation*, 9:610–615, 2003.

[14] Allele Frequencies in Worldwide Populations. `www.allelefrequencies.net` .

[15] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53:457–481, 1958. Online available at `http://www.jstor.org/view/01621459/di985853/98p0114a/0` .

[16] Jan Klein. Origin of major histocompatibility complex polymorphism: The trans-species hypothesis. *Hum. Immunol.*, 19:155–162, 1987.

[17] John P. Klein and Melvin L. Moeschberger. *Survival Analysis, Techniques for Censored and Truncated Data*. Springer, second edition, 2003.

[18] C. Liao, J. Y. Wu, Z. P. Xu, Y. Li, X. Yang, J. S. Chen, X. W. Tang, S. L. Gu, Y. N. Huang, P. H. Tang, and Tsang K. S. Indiscernible benefit of high-resolution HLA typing in improving long-term clinical outcome of unrelated umbilical cord blood transplant. *Bone Marrow Transplantation*, 40:201–208, 2007.

[19] Ross A. Maller and Xian Zhou. *Survival Analysis with Long-Term Survivors*. John Wiley & Sons, 1996.

[20] S. G. E. Marsh, E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, D. E. Geraghty, J. A. Hansen, C. K. Hurley, B. Mach, W. R. Mayr, P. Parham, E. W. Petersdorf, T. Sasazuki, G. M. Th. Schreuder, J. L. Strominger, A. Svejgaard, P. I. Terasaki, and J. Trowsdale. Nomenclature for factors of the HLA system, 2004. *Tissue Antigens*, 65(4):301–369, 2005.
See also previous reports and later updates from the WHO Nomenclature Committee for Factors of the HLA System among the references therein.

[21] Diogo Meyer and Glenys Thomson. How selection shapes variation of the human major histocompatibility complex: a review. *Ann. Hum. Genet.*, 65:1–26, 2001.

[22] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–965, 1972.

[23] J. Robinson, A. Malik, P. Parham, J. G. Bodmer, and S. G. E. Marsh. IMGT/HLA — a sequence database for the human major histocompatibility complex. *Tissue Antigens*, 55:280–287, 2000.

[24] J. Robinson, M. J. Waller, P. Parham, J. G. Bodmer, and S. G. E. Marsh. IMGT/HLA — a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.*, 29:210–213, 2001.

[25] J. Robinson, M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. E. Marsh. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.*, 31:311–314, 2003.

[26] Geziena M. Th. Schreuder, Carolyn K. Hurley, Marie Marsh, Steven G. E.and Lau, Marcelo A. Fernandez-Vina, Harriet J. Noreen, Michelle Setterholm, and Martin Maiers. HLA Dictionary 2004: Summary of HLA-A, -B, -C, -DRB1/3/4/5, -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens*, 66:1–55, 2005.
Also appeared in Human Immunology (2005) 66:170–210, and in International Journal of Immunogenetics (2005) 32:19–69.

[27] Anthony Nolan Research Institute, HLA Informatics Group website. `www.anthonynolan.org.uk/HIG/` .

[28] Bone Marrow Donors Worldwide website. `www.bmdw.org` .

[29] National Marrow Donor Program® website. `www.marrow.org`
Diseases treatable with BMT:
`www.marrow.org/PHYSICIAN/Tx_Indications_Timing_Referral/Diseases_Treatable_by_HCT/index.html`
An overview of outcomes data analyses:
`www.marrow.org/PHYSICIAN/Outcomes_Data/index.html` .

[30] The Anthony Nolan Trust website. `www.anthonynolan.org.uk` .

[31] World Marrow Donor Association website. `www.worldmarrow.org` .

[32] Simon Whelan, Pietro Liò, and Nick Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics*, 17(5):262–272, 2001.

[33] R. M. Zinkernagel and P. C. Doherty. Restriction of *in vitro* T-cell mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature*, 248:701–702, 1974.