2010

Establishing benchmarks for predictive performance in liver transplant survival models

Elizabeth Renouf
Western University

Establishing benchmarks for predictive performance in liver transplant survival models

(Spine title: Predictive performance in liver transplant survival models)

(Thesis Format: Monograph)

by

Elizabeth Renouf

Department of Statistical and Actuarial Sciences

Collaborative Program in Biostatistics

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

The School of Graduate and Postdoctoral Studies

The University of Western Ontario

London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO

School of Graduate and Postdoctoral Studies

CERTIFICATE OF EXAMINATION

Supervisor:

Dr. David Bellhouse

Examiners:

Dr. Wenqing He

Dr. Tina Mele

Dr. Duncan Murdoch

The thesis by

Elizabeth Renouf

entitled:

Establishing Benchmarks for Predictive Performance in Liver Transplant Survival Models

is accepted in partial fulfillment of the

requirements for the degree of

Master of Science

.....
Date

.....
Chair of the Thesis Examination Board

I Abstract

This study evaluated the performance of prognostic models for survival after liver transplant. Assessment of the adequacy of such models is difficult and quantitative benchmarks are needed for measuring performance. We examined the commonly used Cox proportional hazards (PH) model for survival analysis and compared to simpler models using survival trees. Models were evaluated using the integrated Brier score on an independent test set, allowing comparison of models based on prediction error. We also evaluated Harrell's concordance statistic in the Cox PH model. We found that two important predictors of survival violated the PH assumption, suggesting that both the PH model and the concordance statistic are inappropriate for transplant data. We found that the scientific significance of the predictive accuracy gained through the use of the models tested here was limited. Benchmarks for performance evaluation are an important tool for accurate decision making in medicine.

II Keywords

survival analysis, integrated Brier score, concordance, liver transplantation, prediction error

III Acknowledgements

This thesis was carried out with data kindly provided by the Arbor Research Collaborative for Health as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. This study satisfied the criteria for IRB exemption according to guideline 1-G-002 of the University of Western Ontario Research Ethics Board. Financial support was provided by NSERC in the form of a Canada Graduate Scholarship for one year. I extend much gratitude to my advisors Dr. David Bellhouse and Dr. Vivian McAlister for their consideration and insight.

Contents

List of Figures

List of Tables

1 Introduction

The body of literature on the subject of survival analysis methodology has grown at a rapid pace. In 1995, Wyatt and Altman published a paper examining why so many prognostic models are published and then promptly forgotten, citing a lack of evidence of both accuracy and generalizability. Among other things, they suggest authors provide a validation of the model on an independent data set, preferably prospectively, in addition to proof of low prediction error. In 2000, Altman and Royston followed up with a tutorial on model validation: "The idea of validating a prognostic model is generally taken to mean establishing that it works satisfactorily for patients other than those from whose data it was derived." More recently, Altman and colleagues published a valuable series of papers in the *British Medical Journal* describing best practices for developing and validating prognostic models (Altman, 2009; Moons, 2009).

New prognostic models for liver transplant survival appear frequently, most of which utilize the popular Cox proportional hazards (PH) regression model (Cox, 1972). With this paper we aim to produce a quantitative evaluation of some statistical methods that are popular in survival analysis. We chose survival after liver transplant as our example because there is much current research being done in this area. The last twenty years of transplant research have produced extensive insight into possible factors affecting survival after liver transplant. Several prognostic models have been proposed. Almost all use the Cox PH model for survival analysis and Harrell's concordance statistic as a measure of model adequacy. Very few authors include a discussion of whether the assumptions of the methods used were met. Difficulties in assessment and comparison of models are further complicated by the fact that quite often, papers analyzing the same data set return contradictory results. This is due to the employment of differing time frames, the use of different inclusion or exclusion criteria, and using cutpoints or dichotomization of continuous variables. Authors will condense categorical variables into fewer categories, and often what is included in each category is impossible to determine. Categorizations of continuous

variables may not be specified in detail, making study results difficult to compare.

Although validation of prognostic models has been firmly established as an absolute necessity, many papers do not test their model performance at all, or they employ a measure of performance based on the same data set with which the model was built, resulting in an overly optimistic assessment. By convention, most transplant researchers use Harrell's concordance statistic (Harrell et al., 1982) as a measure of model performance, although other new methods for measuring the probability of concordance have been proposed, some of which employ measures to better handle censored data (Gonen and Heller, 2005). Missing from current research are large sample analyses with quantitative measures which evaluate prognostic models on independent data. Wyatt and Altman (1995) write of the necessity of separate testing - in time and place - of the model on a new test set: "In view of the established difficulty of transferring prognostic models, it is surprising that such follow up testing is seldom performed."

We assessed the performance of predictive models for survival after transplant using the Cox PH model. We tested whether the assumptions of the Cox model and the concordance statistic were met and found that they were not. We evaluated the prediction error of the Cox PH model with a method called the integrated Brier score which is a valid measure of model performance even when the PH assumption does not hold or the model is otherwise mis-specified. We also compared the performance of a simpler model using survival trees with the more complex Cox models that are largely present in transplant literature.

2 Evaluators of Predictive Performance

Critical evaluation of prognostic models is often overlooked but it is essential for obtaining a model that predicts as accurately as possible. In 2001, Christensen published a review of currently used prognostic models for survival in chronic liver disease (without transplant) and expressed concern that measures of prediction error were lacking. The

accuracy of prognostic models for survival after liver transplantation has not been tested except in a very limited way. In addition, there are currently no standards for evaluating the accuracy of prognostic models. Accuracy in the context of this study refers to predictive performance. A more comprehensive view of accuracy is given by Hand (1997). Hand also points to a lack of standard definitions for terminology used in performance assessment, with words such as discriminability, reliability, and imprecision being used interchangeably with inaccuracy.

2.1 Concordance Statistic

The concordance statistic, also called the c-statistic or c-index, is one of the most commonly used methods to assess the performance of survival models. The c-statistic is a measure of predictive discrimination, which to most scientists means how well the model distinguishes between patients who experience an event and those who do not. This measure was introduced by Harrell et al. in 1982 (see also Harrell et al., 1983). He describes the measure as an estimate of the probability that, of two randomly chosen patients, the patient with the higher prognostic score will outlive the patient with the lower prognostic score. To determine the c-statistic, all possible pairs of observed data where one subject failed and the other did not are examined. Harrell's c-statistic is then calculated as the number of pairs where the ordering of the observed survival times agrees with that predicted by the model, divided by the total number of pairs for which the ordering of survival times could be inferred. When the survival times predicted by the model are identical for a pair, then 0.5 is added to the count of concordant pairs, instead of 1, while 1 is added to the count of usable pairs. Thus, the c-statistic excludes patients for which the ordering of survival times cannot be determined. If both subjects are censored, their survival times cannot be ordered and the c-statistic cannot be calculated (Harrell et al., 1983). Similarly, survival times cannot be ordered for two subjects where one has failed and the censoring time of the other is less than the one who failed. A c-statistic greater than 0.7 is considered to be clinically useful,

while a c-statistic of 0.5 would indicate random prediction. A model that predicts perfectly would give a c-statistic of 1. The c-statistic can be expressed as $c = \frac{(C+\frac{P}{2})}{C+D+P}$ where C is the number of concordant pairs, P is the number of ties and D is the number of discordant pairs.

Several biostatisticians have criticized Harrell's c-statistic as a biased measure (Graf, 1999; Schumacher, 2003; Gonen, 2005). One concern is that the c-statistic is usually positively biased, indicating that the predictive accuracy of the model is overly optimistic when based on the c-statistic. Using simulation, Gonen (2005) showed how the value of Harrell's c-statistic increased with the proportion of censoring in the Cox PH model. Graf (1999) criticized the c-statistic as a biased performance measure because it is based on predicted survival times. It has long been recognized that survival time, or the time-to-event, cannot adequately be predicted (Parkes, 1972; Forster 1988; Henderson, 1995). Estimates of duration of survival are often overly optimistic. Most researchers today will use the predicted probability of surviving until a fixed timepoint rather than the predicted survival time to calculate the c-statistic, and this is considered acceptable as long as the two estimates are one-to-one functions of each other. Harrell (1996) notes that this situation holds as long as the PH assumption is satisfied. Because of these limitations, many statisticians prefer methods to assess prediction error that consider "individual vital status as a prediction outcome variable instead of observed survival time" (Schumacher, 2003).

2.2 Brier Score

Several statistical papers have suggested using the Brier score instead of commonly used methods such as p-values, the c-statistic, or receiver operating characteristic curve methodology (Schumacher, 2003; Kronek, 2009; Haibe-Kains, 2008; van Wieringen, 2009; Ikeda, 2001). These latter methods, while well understood by medical practitioners, are limited in the presence of censored data. On the other hand, the Brier score is a useful measure of the predictive performance of prognostic survival models with censored data.

While the Brier score as a method of model assessment is not new, it is virtually unused in transplant-related survival analysis. A search of www.pubmed.com for "Brier score" and "transplant" returned only 1 result, a 2009 paper on stem cell transplantation (Hari et al., 2009).

The method behind the Brier score was developed in 1950 by meteorologist Glenn W. Brier for measuring the accuracy of weather predictions (Brier, 1950). The Brier score, in the context of survival analysis, is a measure of the expected squared difference between individual patient status and the survival probability predicted by the model. A score can be obtained for specific time points, or an integrated score for the entire time period of observation can be used, and a prediction error curve over time can be output for each patient.

The empirical Brier score, when censoring does not occur, is defined (Schumacher, 2003) as

$$\hat{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(t) - \pi(t|X_i))^2$$

where Y(t) is the individual vital status at time t (zero if an event has occurred before t and one otherwise), X is the vector of covariates, and the index i denotes the i-th patient in $i = 1...n$. The estimated event-free probabilities for an individual with covariate vector X are denoted by $\pi(t|X)$. The extension to right censored data was introduced by Graf et al. in 1999 using observational data on the survival of breast cancer patients. Weighting by means of the distribution of censoring times is used to remove censoring bias. The Brier score for censored data is defined in the following way using the notation of Schumacher (2003):

$$\hat{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(\tilde{T}_i \le t, \delta_i = 1)}{\hat{G}(\tilde{T}_i)}(Y_i(t) - \pi(t|X_i))^2 + \frac{I(\tilde{T}_i > t)}{\hat{G}(t)}(Y_i(t) - \pi(t|X_i))^2\right]$$

where $\tilde{T}_i$ is the minimum of the survival and the censoring time, I is an indicator function,

δ_i is the censoring indicator which is equal to 1 if uncensored, and $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of the censoring distribution in the whole sample. The weights are $\frac{1}{\hat{G}(\tilde{T}_i)}$ if an event occurs before time t and $\frac{1}{\hat{G}(t)}$ if an event occurs after time t. Censored observations with survival times smaller than t are weighted with zero (Graf, 1999). It should be noted that as the end of the study period approaches, $\hat{G}(\cdot)$ becomes very small as censoring approaches its maximum of 84% for the data analyzed in this study.

The Integrated Brier Score (IBS) provides a cumulative prediction error from time zero to some specified time point, $t*$.

$$IBS = \int_0^{t*} BS(t)dt$$

The Brier score will be a number between zero and one. The better a method is at predicting survival on the test set, the *smaller* its Brier score. A perfect prediction model would have a Brier score of 0. Henderson et al. (2001) demonstrated that predictions drawn independently from a uniform [0,1] distribution have an expected Brier score of 0.33. A constant prediction of 0.5 would yield a Brier score of 0.25 (Schumacher, 2003).

Schumacher et al. (2003) have demonstrated that Graf's extension of the Brier score to survival data can provide a meaningful interpretation even when the model is mis-specified, since the score does not depend on the assumed survival model. This means that the Brier score, unlike the c-statistic, is still a valid measure of performance when the PH assumption of the Cox model fails. The main assumption of the weighting scheme proposed by Graf is that the censoring distribution is independent of the covariates.

While medical practitioners may not be as familiar with the Brier score, some statisticians prefer it as an assessment method because it is based on predicted survival curves and is valid for censored data. The Brier score of the Kaplan-Meier estimate where covariate information is ignored can be used as a benchmark for predictive performance. The Kaplan-Meier estimate produces a common survival curve, the same for all patients. The

Brier score from the Kaplan-Meier estimate provides a benchmark value that is similar to that obtained from the null model in linear regression (Haibe-Kains, 2008). Schumacher (2003) writes that the relative gain in predictive power of a prognostic model with respect to the Kaplan-Meier benchmark can be interpreted as an R^2-like measure.

The Brier score is limited in that it can only be applied to right censored data. It is also a measure of overall performance of a model. It is possible that a model that performs well overall may not predict accurately for an individual patient (Gerds, 2006).

There are two packages in R which contain functions for computing the Brier score, `survcomp` for Cox models (Haibe-Kains et al, 2009) and `ipred` for survival trees (Peters and Hothorn, 2009). The functions `sbrier` and `sbrier.score2proba` compute the Brier scores and the corresponding integrated Brier score from a risk score, for every event time. All statistical calculations were carried out using R version 2.10.1 (R Development Core Team, 2009).

2.3 Comparing Brier Scores

Haibe-Kains (2008) used a paired Wilcoxon rank sum test for dependent samples to determine whether one integrated Brier score was significantly better than another. The Wilcoxon rank sum test is a non-parametric test that can be used to determine if the mean or median of one population is shifted to the right or left of another. Thus the Wilcoxon rank sum test can be used to determine whether Brier scores over time in one model are smaller than the Brier scores of another model. For more detailed information on the Wilcoxon rank sum test see Conover (1999).

3 Literature Review

In this review we focus on recent papers whose primary goal was prediction of outcome for future patients after liver transplant. An extensive review of the topic of re-

transplantation for patients with recurrent liver disease was undertaken by Rosen et al. (2000; 2003) using Organ Procurement and Transplantation Network (OPTN) data from patients transplanted between 1990 and 1996 (n = 1356). The authors tested Cox PH survival models on an independent data set from outside the United States (n = 281). The predictive performance was evaluated by calculating the risk score from their model and carrying out a Kaplan-Meier analysis stratified by risk score (high, medium, low). The result for each data set was displayed in a graph of survival curves stratified by risk group. While the graphs of the test set looked similar to those of the training set, there was no measure of variability given and so the statistical significance is unknown. No other quantitative evaluation of predictive performance was performed on the independent data set. On the combined data sets the authors evaluated their model with ROC curves and the c-statistic (0.606 to 0.657 depending on the timepoint). No examination of the PH assumption was reported. Ghobrial et al. (2002) developed a prognostic model for survival after liver transplant in Hepatitis C positive patients using Cox PH regression on the OPTN data set. The c-statistic was used to assess model performance on the same data set used to build the model. They found that the c-statistic was 0.69, 0.68, and 0.67, at 3 months, 6 months, and 1 year after transplantation, respectively. There was no report of assessment of the PH assumption.

Merion et al. (2005) introduced a "survival benefit" model for liver transplant as a time-dependent Cox regression model to compare mortality for wait-listed candidates and transplant recipients, at equal duration since wait-listing. The model was based on the Model for End Stage Liver Disease (MELD) score which is currently used to predict survival on the waiting list. A piecewise PH model was used to estimate mortality after transplant at various time windows up to one year. Habib et al. (2006) assessed the value of the pre-transplant MELD score to predict survival after transplant using 12 years of data from the Thomas E. Starzl Transplantation Institute in Pittsburgh. In building a prognostic model using Cox PH regression, the authors considered 2,009 adults receiving deceased

donor transplant but excluded patients with acute failure, hepatocellular carcinoma, and other hepatic malignancies. They evaluated their model by determining the c-statistic on the same data set that was used to build the model (0.63 for patient survival). Habib also used the Grambsch-Therneau test to assess the PH assumption, however the results were not reported. Weismuller et al. (2008) developed a prognostic model using Cox PH regression for survival after liver transplant based on a small data set of 133 patients and validated it on a separate cohort of 87 patients. Performance measures used were ROC curves and a graphical comparison using Kaplan-Meier survival curves. No assessment of the PH assumption was included in the discussion.

In 2008 a new model for predicting three month survival after liver transplantation was proposed by Rana (2008). Logistic regression analysis was employed and a risk score was assigned to each risk factor based on its odds ratio for patient death at three months. Model performance was assessed using ROC curves and the c-statistic (0.70) on the same data set used to build the model. There was no discussion of whether goodness of fit tests for the logistic model were performed.

In a 2009 publication, Schaubel et al. calculated survival benefit as the difference between a candidate's predicted 5-year mean lifetime without a transplant and their predicted 5-year mean lifetime with a transplant. The authors propose an organ allocation model based on predictions of survival benefit for every candidate on the waiting list, assessed each time an organ becomes available. The survival after transplant component of the prediction was developed using a Cox PH model which incorporated 30 covariates and utilized categorization of continuous variables such as recipient age, donor age, recipient pre-transplant serum creatinine and recipient pre-transplant albumin. The model included all adults and children transplanted between September 2001 and December 2007. The fit was evaluated with Harrell's c-statistic (0.63). The model was cross validated by randomly splitting the data set used to build the model repeatedly, fitting the model with one half and calculating the c-statistic on the other. There was no discussion included on assessment

of the PH assumption. In 2009 Ravaioli et al. used European registry data to develop a survival benefit model using Cox regression with liver transplantation as a time-dependent variable. Continuous variables were dichotomized with cutpoints chosen based on prior studies. Variables were assessed using the c-statistic (not reported) and ROC curves on the same data set used to build the model. No evaluation of model performance was carried out and no assessment of the PH assumption was discussed.

A review of prognostic models for survival after liver transplant shows a lack of both independent validation and evidence of model accuracy. The models are likely to be overly optimistic in their assessment of model fit. The ability to compare the prediction error of these prognostic models to a benchmark indicator would vastly improve interpretation of these results.

Very few authors report any examination of the assumptions of the statistical methodology used. As noted by Wyatt and Altman (1995), hundreds of prognostic models are published every year yet very few are actually used in clinical decision making. Too often, the use of the Cox PH model is not followed up by a critical check of model assumptions. Reliable and accurate predictive models for survival after liver transplant would be a valuable tool for physicians determining how to allocate a scarce resource. Examination of performance measures is urgently needed in order to support better decision-making.

4 Application to Liver Transplant Data

4.1 Description of the Data Set

The data set described here is OPTN data provided by the Scientific Registry of Transplant Recipients (SRTR). The data examined comprises patients aged 18 or older who received a first liver transplant from a deceased donor between September 1, 2001 and December 31, 2007. We chose a data set including only adult patients and excluded living donations in the hope that a more homogenous data set would produce a model with greater

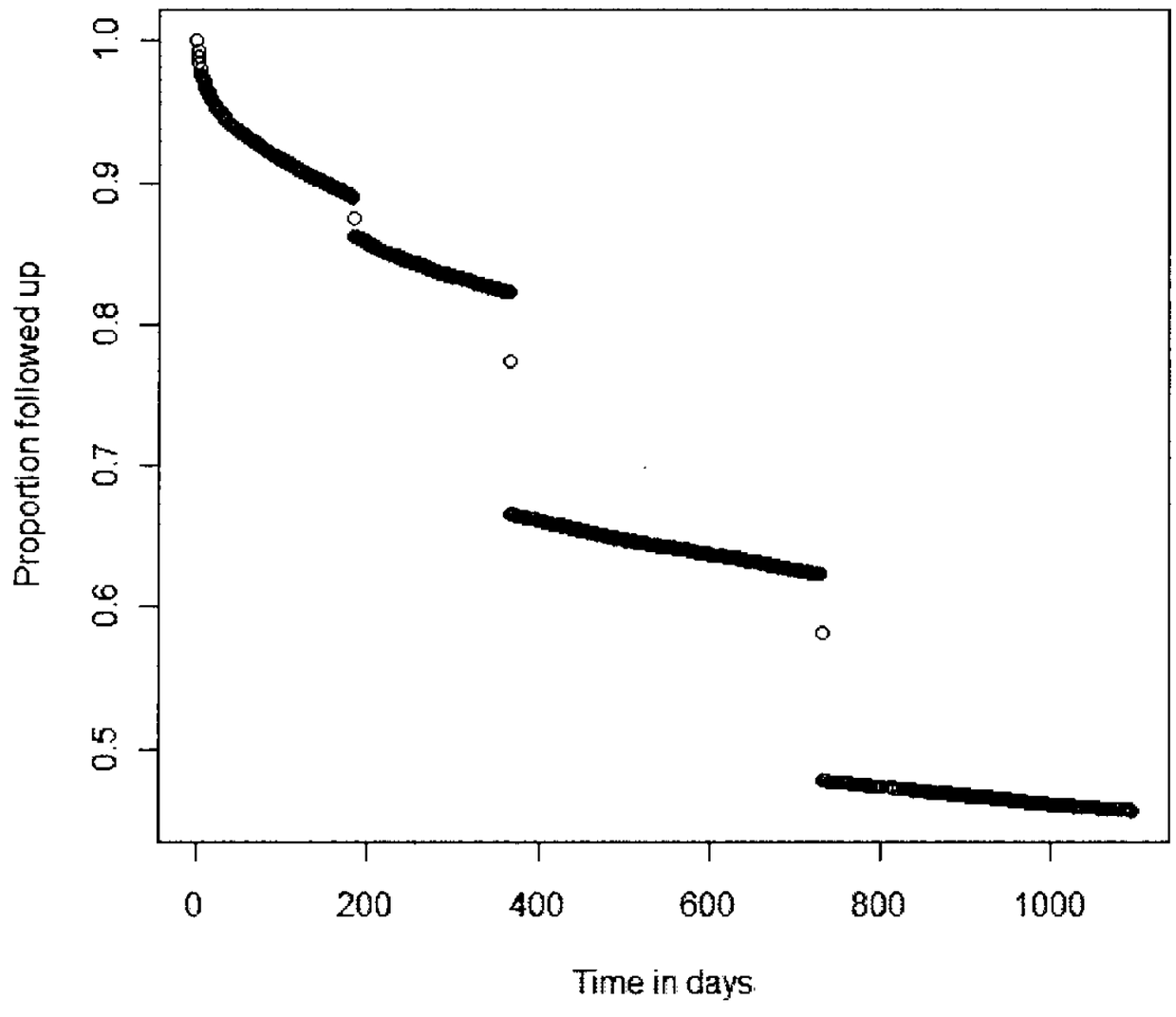
prediction accuracy. For this reason we also excluded patients with a previous transplant of any organ, and patients who received a split liver. During this time frame 28,165 patients who fit our inclusion criteria were transplanted. Complete data on the covariates considered were available on 21,268 patients. A total number of 3473 deaths were observed (16/%).

The endpoint was patient survival (death from any cause) from the time of liver transplant to three years after transplant. Patients who died on the same day of transplant were assigned a survival time of 0.5 days for all models. Patients followed for less than three years were censored at the date of last follow up assessment. Retransplanted patients were censored at the time of retransplant. Censoring would be described as high (Lin, 1997) at 84%. Of the censored observations, 46% were followed for the entire three year period, 6% were censored due to retransplant, and the remaining 48% were censored due to end of follow up time or loss to follow up.

The data (complete cases only) was randomly split into a mutually exclusive training set (n = 14178) and test set (n = 7090). While the survival tree methods can accommodate missing data, the Cox models cannot; complete cases only were used to facilitate comparison between models. The training set was used to build each model, and the test set was used to evaluate the performance of the models built using the training set. The same set of covariates was used in building each model.

Subjects lost to follow up are potentially a large problem with registry data, especially when patients who die are mistakenly classified as lost to follow up. To combat this, the SRTR provides supplemental data from the social security death master file as a reference and we cross checked our data source with this index. Patient follow up forms are turned into the OPTN at set times after transplant, but often forms are turned in early when a patient dies. After choosing your date of data extraction, it is important to avoid bias from forms turned in early, just before the date of data extraction, due to the death of the patient. You will then have follow up information from the patients who died but not from the ones still alive, simply because the due date for their forms has not arrived. The SRTR

Figure 1: Distribution of follow up time in the training set.



recommends using the maximum date for which we expect follow up from the OPTN on a particular patient as the last day at risk and not using information from forms turned in early for patients who died. Median time to censoring in the training set was 451 days, calculated by means of a reverse Kaplan-Meier analysis according to Schemper (1996). Figure 1 shows a plot of the proportion followed up over time. The discrete data collection intervals of the SRTR registry are illustrated in this plot.

4.2 Description of Variables

The variables considered as possible predictors were based on research conducted by experts in the field (Merion, 2005; Rana, 2008; Schaubel, 2009; Watt, 2010). We examined recipient age at transplant, recipient body mass index, recipient serum creatinine pre-transplant, recipient albumin pre-transplant, recipient International Normalized Ratio (INR) pre-transplant, recipient bilirubin pre-transplant, recipient BMI pre-transplant, diagnosis of cholestatic liver disease, diagnosis of Hepatitis C virus (HCV), diagnosis of

acute hepatic necrosis, diagnosis of malignancy of any type, diagnosis of non-cholestatic liver disease, diagnosis of metabolic liver disease, diagnosis of hepatocellular carcinoma (HCC), diagnosis of viral liver disease, diagnosis of alcoholic liver disease, previous portal vein thrombosis in the recipient, whether or not the recipient was on life support, whether or not the recipient had had prior abdominal surgery, whether or not the recipient was hospitalized or in ICU, whether or not the recipient underwent dialysis status in the week prior to transplant, donor age, donor cause of death categorized by trauma, anoxia, cardiovascular disease, and other, donor race categorized by White, Hispanic, Black, Asian, and other, whether or not the donor organ was shared from another transplant centre, cold ischemic time of the transplanted liver, and whether or not the donor was non heart beating. Charts detailing the categorization of any continuous variables are provided in the Appendix. For all models, the following impossible data values were set to NA, thus excluding the records from the analysis: five observations with a pre-transplant creatinine of zero, two observations with donor age of zero, seven observations with a height of zero, 28 observations with a height under 100 cm, and four observations with a cold ischemic time of zero.

4.3 Statistical Models

4.3.1 Cox Proportional Hazards Model: General Remarks

The popularity of the Cox PH model (Cox, 1972) is unparalleled in survival literature and this popularity extends to published research in liver transplantation. The model is expressed in terms of the hazard function as follows:

$$h(t, \mathbf{X}) = h_0(t)\, exp\left(\sum_{i=1}^{p} \beta_i X_i\right)$$

where $\mathbf{X}$ represents a set of explanatory variables and $h_0(t)$ is the unspecified baseline hazard, assumed to be the same for all subjects. The predictors act multiplicatively on the hazard function. Thus the hazard at time t is the product of the baseline hazard h_0 and the

explanatory variables which are independent of time. The baseline hazard is a function of time but not the covariates **X**. It is unspecified, making the model semi-parametric. The ability to leave the baseline hazard unspecified yet still obtain good estimates of regression coefficients is a fundamental reason for the popularity of the model. The key assumption of the model is that the ratio of the hazards for different individuals is constant over time. A meaningful illustration of this assumption would be to say that a person with a risk of death at baseline that was twice as high as that of another person would have a risk of death at all subsequent times to be twice as high (Harrell, 2001). A covariate that does not satisfy the PH assumption can be modelled using stratification or a time-dependent variable.

4.3.2 Cox Proportional Hazards Model 1: Using Cutpoints

We used a model built in a fashion similar to Schaubel et al. (2009) using quartile-based dummy variables for creatinine and albumin. However, unlike Schaubel we did not include re-transplanted subjects, recipients under age 18, diabetic status, or serum sodium pre-transplant, so that we could use a larger data set. Donor age was categorized also according to Schaubel (2009, see Table 2). The model was built using backward elimination selecting covariates with an observed significance level of 5% or smaller. Two interactions identified in Schaubel (2009) as significant were included in the model: donor age > 60 yrs * HCV and recipient age > 55 yrs * recipient age. Continuous variables were centred. Number of events per variable in the training set was approximately 108, with 21 components in the model.

4.3.3 Cox Proportional Hazards Model 2: Using Fractional Polynomials

The categorization of continuous covariates has come under increased scrutiny in recent years. The concern is that dichotomization or categorization could result in a loss of information, leading to a biased model (Royston, 2006). Models may be mis-specified if curved relationships are modelled as linear. The possible advantages gained, such as greater sim-

plicity, have been shown to come at a high cost statistically. Royston (2006) states that the use of dummy variables for a continuous covariate uses up valuable degrees of freedom at the expense of power and precision. Wainer (2006) has shown that cutpoints can be found that will result in either positive or negative associations. Austin et al. (2004) illustrated how categorization may increase the Type I error rate. Most importantly, Royston et al. (2006) note that the impact of categorizing or dichotomizing more than one predictor is unpredictable and could result in a seriously misleading model.

A two term polynomial with powers *p1* and *p2* can be used to model a continuous covariate X, represented as $\beta_1 X^{p1} + \beta_2 X^{p2}$. Benner's (2005) model selection algorithm combines backward elimination with a systematic search for a suitable transformation to represent the influence of each continuous covariate on the outcome. The algorithm selects the 1 or 2 degree fractional polynomial which best predicts the outcome. No transformation is the default if evidence of non-linearity is not found. The powers for the polynomial, p_1 and p_2, are taken from the set S = {-2, -1, -0.5, 0, 0.5, 1, 2, 3} where $x^0 = log(x)$. In a two degree polynomial with equal powers p_1 and p_2, a "repeated powers" function is used where the second term is multiplied by the log of the covariate value, $\beta_1 X^{p_1} + \beta_2 X^{p_2} \cdot log(X)$ where X is a single continuous covariate. The algorithm starts by arranging the explanatory variables in order of decreasing statistical significance for omitting each predictor from a model comprising all the predictors with each term linear (Ambler & Benner, 2008). The best fractional polynomial is determined by fitting each polynomial, starting with the most complex, in the set S and finding the one with the lowest deviance (-2*log-likelihood). Additional details of the variable selection algorithm are provided in Benner (2005) and also Ambler and Royston (2001).

Disadvantages of using fractional polynomials include insufficient power to detect non-linear functions, especially in survival analysis where there are too few events (Royston, 2008). We do not anticipate this to be an issue here given our large data set. Another issue is possible sensitivity to extreme values in a covariate which can be checked through

residual analysis.

We built a model using the `mfp` package (Ambler and Benner, 2008) available in R. All continuous predictors were kept as continuous and were modelled using fractional polynomials. We chose a p-value of 0.05 for the selection of the best functional form for continuous variables in the model. We restricted the variable cold ischemic time to be a polynomial of degree 1 since the effect of this predictor is known to be linear and increasing. There are some probable errors in the registry which would influence the choice of polynomial, e.g. a cold ischemic time of 50 minutes recorded as 50 hours, and if this restriction is not implemented the algorithm chooses a parabolic shaped polynomial showing reduced hazard for patients with the lowest and the highest cold ischemic times which does not make sense clinically.

4.3.4 A Single Survival Tree

Tree-based methods require fewer statistical assumptions compared to methods like the Cox PH model. Extension of tree methods to survival analysis with right censored data was introduced in 1985 (Gordon, 1985).

Survival tree models are grown using a method called recursive partitioning. This is a non-parametric procedure in which the data set is repeatedly subdivided into groups as homogeneous as possible within groups and as heterogeneous as possible between groups by recursively partitioning based on covariates. As described in Therneau (1997), the first step is to identify the single variable which best splits the data into two groups. This process is then repeated separately within each group, until the subgroups reach a specified minimum size, or until no further improvements can be made. There are different splitting criteria available but the `rpart` function, available in the `rpart` package in R (Therneau and Atkinson, 2010), implements the algorithm described by LeBlanc and Crowley (1992). In order to improve stability of the tree, pruning or trimming of the tree is recommended. This is done using cross validation in an attempt to correct for overfitting and to minimize

prediction error.

The top of the tree is called the root. Splits are determined based on the recursive algorithm, each ending in a terminal node of the tree, also called a leaf. There are many tuning parameters available for the recursive partitioning algorithm. We set the minimum split size to 200, specifying the minimum number of observations in a node for which the routine will try to compute a split. We set the threshold complexity parameter to 0.001. This a tuning parameter which regulates the size of the tree. According to Breiman (1984), the complexity parameter is a measure of whether the amount of accuracy that a split adds to a tree warrants the additional complexity. The default is 0.01 which is too high for a large data set such as the one considered here. The tree is then pruned to reduce overfitting, and here we used a complexity parameter of 0.0014 since it gave the best balance between overfitting the data yet still achieving a tree structure large enough to make sense clinically. In survival trees, prediction is carried out based on the Kaplan-Meier curve of the leaf that a new observation falls into.

Advantages of trees include simplicity, interpretability and ease of use. They visually describe the structure of the data and can present a clear picture that is readily interpretable in a clinical setting. They have only one assumption, according to Gordon (1985), namely that the conditional distributions of the covariates are identifiable (each value of the covariate maps onto a unique prediction). A disadvantage of the single tree is instability, particularly if the training set is small (Hothorn, 2004). The tree can depend too much on the data set used to construct it, with the result that small changes in the data set can induce large changes in the tree (Breiman, 1996). Bootstrap aggregation, discussed in the next section, can be used to stabilize the survival tree.

4.3.5 Bootstrap Aggregated (Bagged) Survival Trees

The term bagging is short for bootstrap aggregation and was introduced by Breiman (1996) who applied it to tree-based classification methods. Bagging of survival trees has

been studied extensively by Hothorn et al. (2004) in order to obtain improved predictions from survival trees. Bagging is carried out in two main steps. First, a set of survival trees is generated for B bootstrap samples of the observations. Then the predicted survival probability can be calculated for a set of new patients using the bootstrapped trees. To aggregate the predictions from bagged survival trees, we used Hothorn's "weighted" method of aggregation where observations from each leaf are aggregated directly and one single predictor is computed for the aggregated sample only (Hothorn, 2004). The bootstrap aggregated Kaplan-Meier curve of a new patient is computed by dropping the new observation down each of the B trees successively, combining into one sample, and predicting the Kaplan-Meier curve from this sample.

Other aggregation methods are available, such as by majority voting or by averaging. All three methods are implemented in the `ipred` package in R (Peters, 2002). We chose to leave our complexity parameters the same as in the single model, although in bagging it results in larger trees. Hothorn (2004) notes that the optimal tree size in bagging is still undetermined.

There are some disadvantages in using bagged trees. They can be difficult to interpret compared to single trees. Hothorn describes them as a 'black box' of multiple trees. There are also many tuning parameters available, such that varying results are possible based on selection of parameters. Further, it is possible that bagging of a stable algorithm can actually make it worse. As Breiman (1996) notes, the essential element is the instability of the prediction method. If perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy. Lastly, bagging survival trees on a large data set requires a significant amount of computing power. Initially we restricted the number of bootstrapped trees to 50. We also tested 300 bootstrapped replications in order to confirm Hothorn's findings that more than 50 trees does not lead to further improvement in prediction error (Hothorn, 2004).

4.4 Evaluation of Predictive Accuracy

The predicted survival probability for each subject in the test set was determined from each model and the integrated Brier score was calculated in order to evaluate predictive accuracy. Integrated Brier scores were compared to the Brier score of the benchmark Kaplan-Meier model using the Wilcoxon rank sum test. The concordance statistic was also evaluated in the Cox models.

5 Results

5.1 Descriptive Statistics

Table 1 provides a summary of descriptive statistics for the training set as a whole and also for censored vs uncensored observations. Significant differences ($p < 0.002$) were found between the percentage of censored and uncensored patients with a diagnosis of AHN, HCV, HCC, and a donor race of Black. These may simply be due to chance - it is not caused by more patients in these categories being censored for re-transplantation (6.5% overall compared to less than 6.5% for all categories except donor race Black at 6.6%), nor is it caused by more patients in these categories being transplanted later rather than earlier in the period under study.

5.2 Brier Scores and Concordance Statistic

Table 2 shows the integrated Brier score calculated on the test set for each of the four models we evaluated. Also shown is the score for the benchmark Kaplan-Meier model considering no covariate information. In addition, we show the relative gain in predictive power of each prognostic model with respect to the benchmark pooled Kaplan-Meier estimate, calculated as 1 - IBS(model)/IBS(benchmark). For the calculation of the IBS at the maximum three year time point we used the score at day 1092 since all observations are

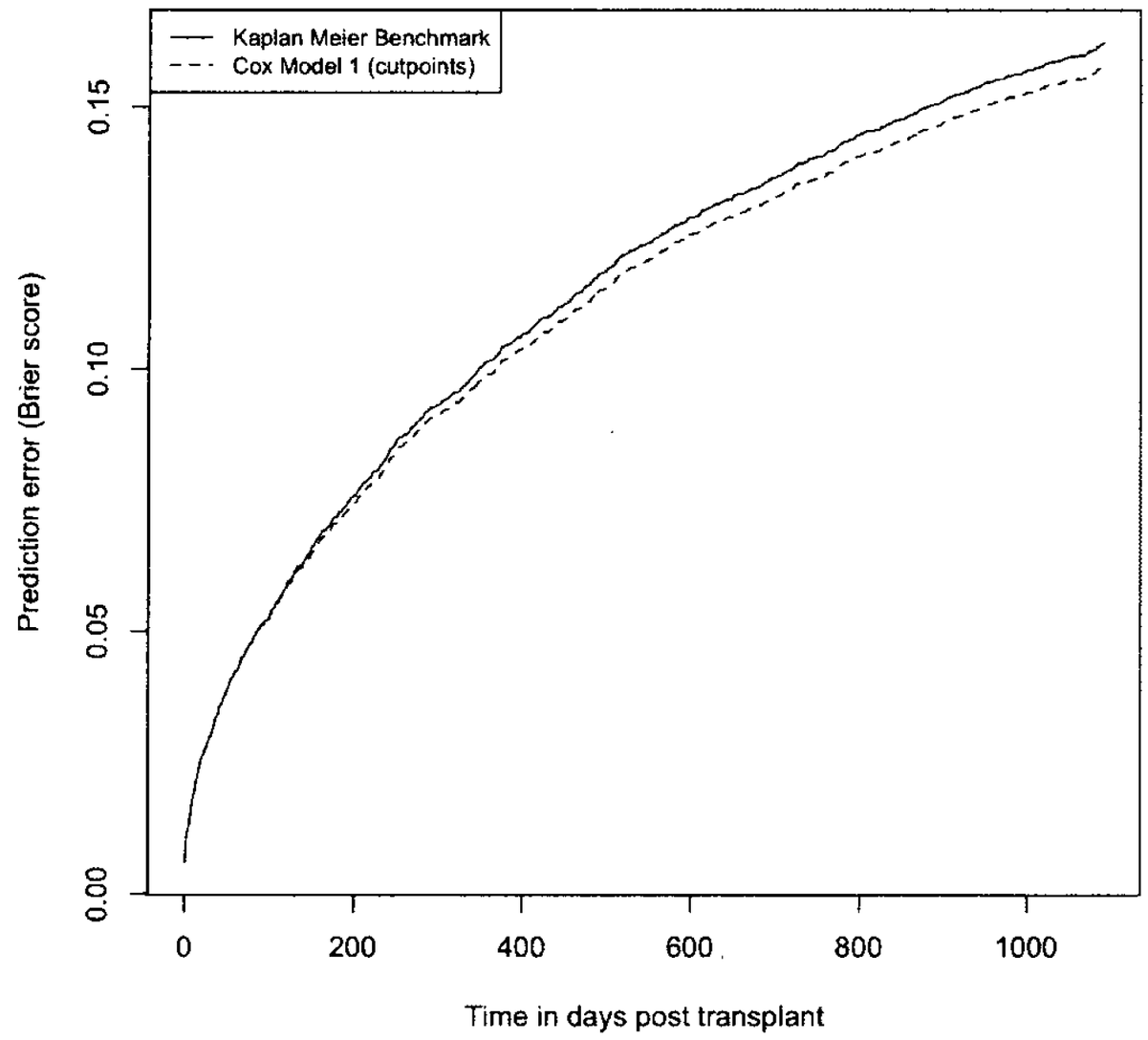
Table 1: Selected characteristics of training data at time of transplant. Note that subjects may fall into more than one diagnosis category.

	All		Censored		Uncensored	
Variable	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
Recipient age (years)	53.7	9.9	53.2	9.9	53.0	9.9
Donor age (years)	42.0	17.5	42.5	17.5	41.7	17.5
	All		**Censored**		**Uncensored**	
	n	**%**	**n**	**%**	**n**	**%**
Sex						
Female	4493	31.7	1698	31.2	2795	32.0
Male	9685	68.3	3740	68.8	5945	68.0
Recipient diagnosis						
Acute hepatic necrosis	896	6.3	300	5.5	596	6.8
Alcoholic cirrhosis	3636	25.6	1441	26.5	2185	25.1
Cholestatic cirrhosis	1319	9.3	509	9.4	810	9.3
Hepatitis C	6076	42.9	2241	41.2	3835	43.9
Hepatocellular carcinoma	1936	13.7	904	16.6	1032	11.8
Metabolic liver disease	391	2.8	151	2.8	240	2.7
Non-cholestatic cirrhosis	10442	73.6	4028	74.0	6414	73.4
Other diagnosis	1246	8.8	465	8.6	781	8.9
Donor race/ethnicity						
Asian	274	1.9	97	1.8	177	2.0
Black	2069	14.6	902	16.6	1167	13.3
Hispanic	1698	12.0	676	12.4	1022	11.7
White	9978	70.4	3699	68.0	6279	71.2
Other	153	1.1	64	1.2	89	1.0

Table 2: Integrated Brier Scores for each model, measured from time of transplant to 3 years after transplant.

Model	IBS	% improvement over benchmark
Benchmark null model	0.1142	–
Cox Model 1 (cutpoints)	0.1114	2.45%
Cox Model 2 (fp)	0.1115	2.36%
50 Bagged Survival Trees	0.1119	2.01%
Single Survival Tree	0.1122	1.75%

Figure 2: Prediction error curves for Cox model 1 (cutpoints) compared to the benchmark Kaplan-Meier curve (covariate information ignored).



censored at the final timepoint of 1095 days. All models score significantly better than the Kaplan-Meier benchmark, a null model that does not employ any covariate information ($p < 0.0001$ for all models). However, the prediction error curves shown in figures 2 to 5 lead one to question the scientific significance of the predictive accuracy gained through the use of any of these models. The comparison of the prediction error curves for each model to the benchmark Kaplan-Meier curve are shown on separate graphs to improve readability. The best integrated Brier score of 0.1114 was obtained using Cox model 1 (with cutpoints) similar to the one used by Schaubel et al. (2009), with a gain in prediction accuaracy of 2.45% compared to the benchmark model.

This was followed closely by Cox model 2 (using fractional polynomials) at 0.1115, and then the bagged survival trees (0.1119). We found that the IBS for the 50 bagged trees and the 300 bagged trees were the same to four decimal places. The single survival tree was the simplest, most parsimonious method and is the most easily understood in a clinical

Figure 3: Prediction error curves for Cox model 2 (fp) compared to the benchmark Kaplan-Meier curve (covariate information ignored).

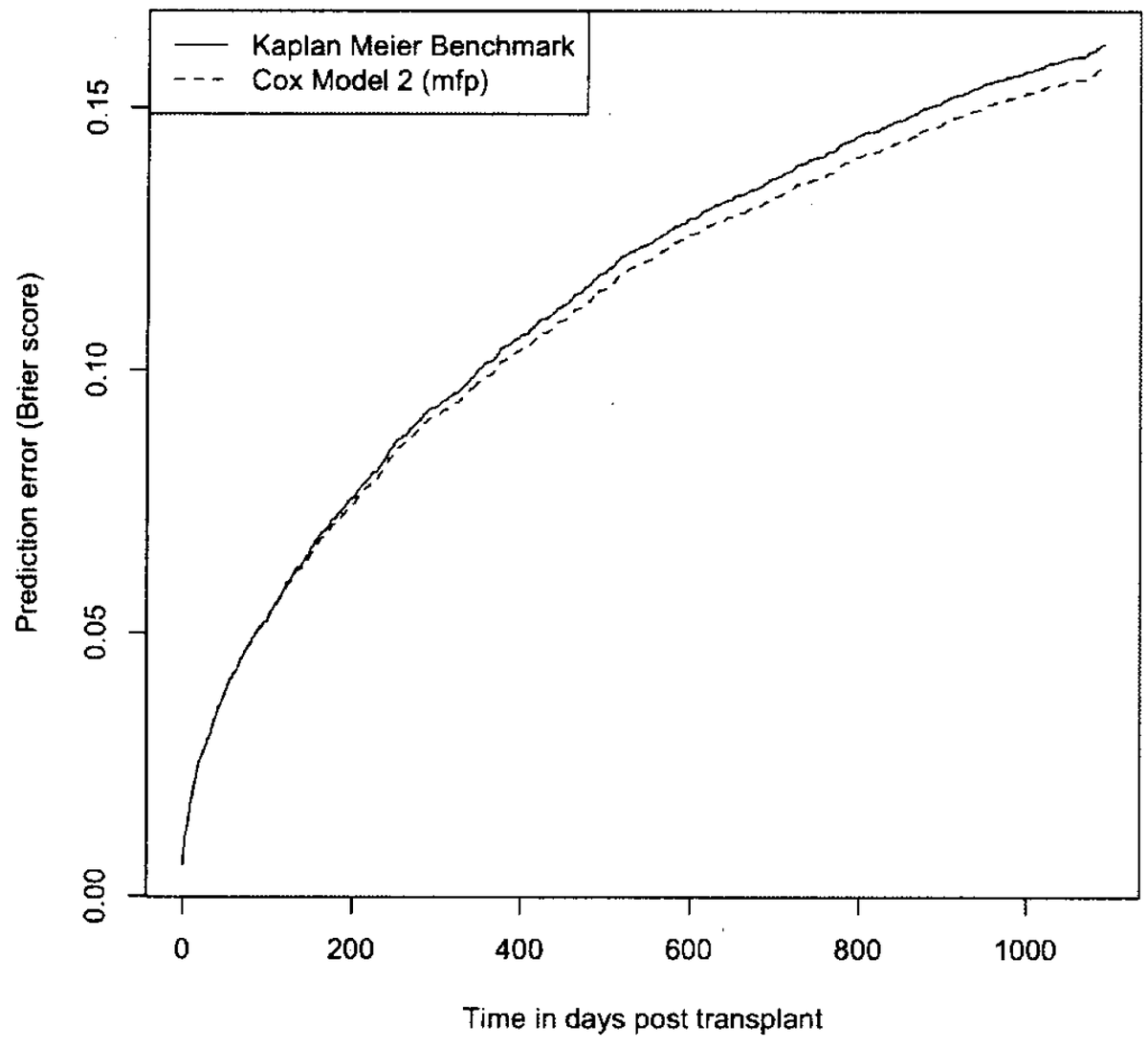


Figure 4: Prediction error curves for 50 bagged survival trees compared to the benchmark Kaplan-Meier curve (covariate information ignored).

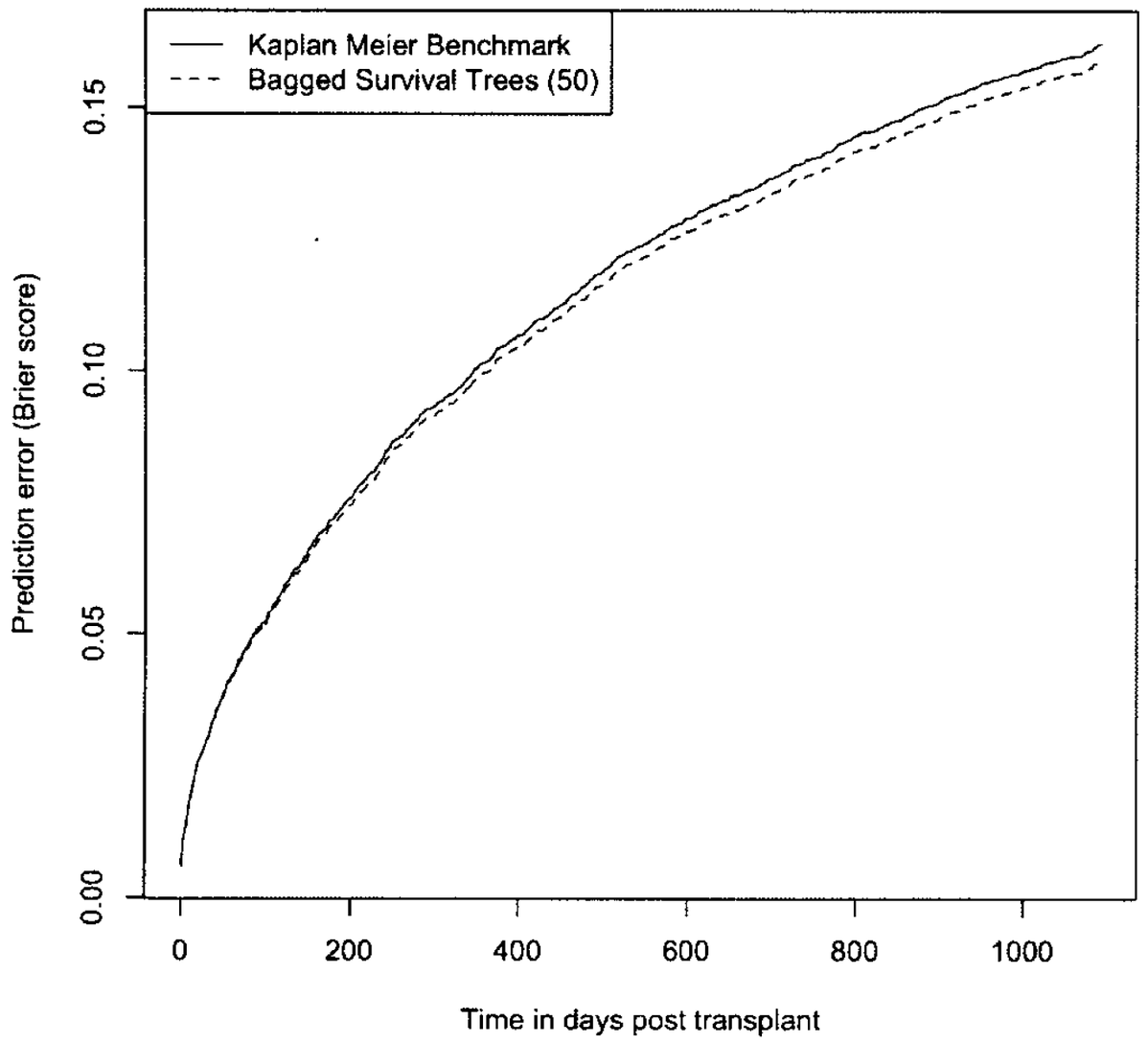
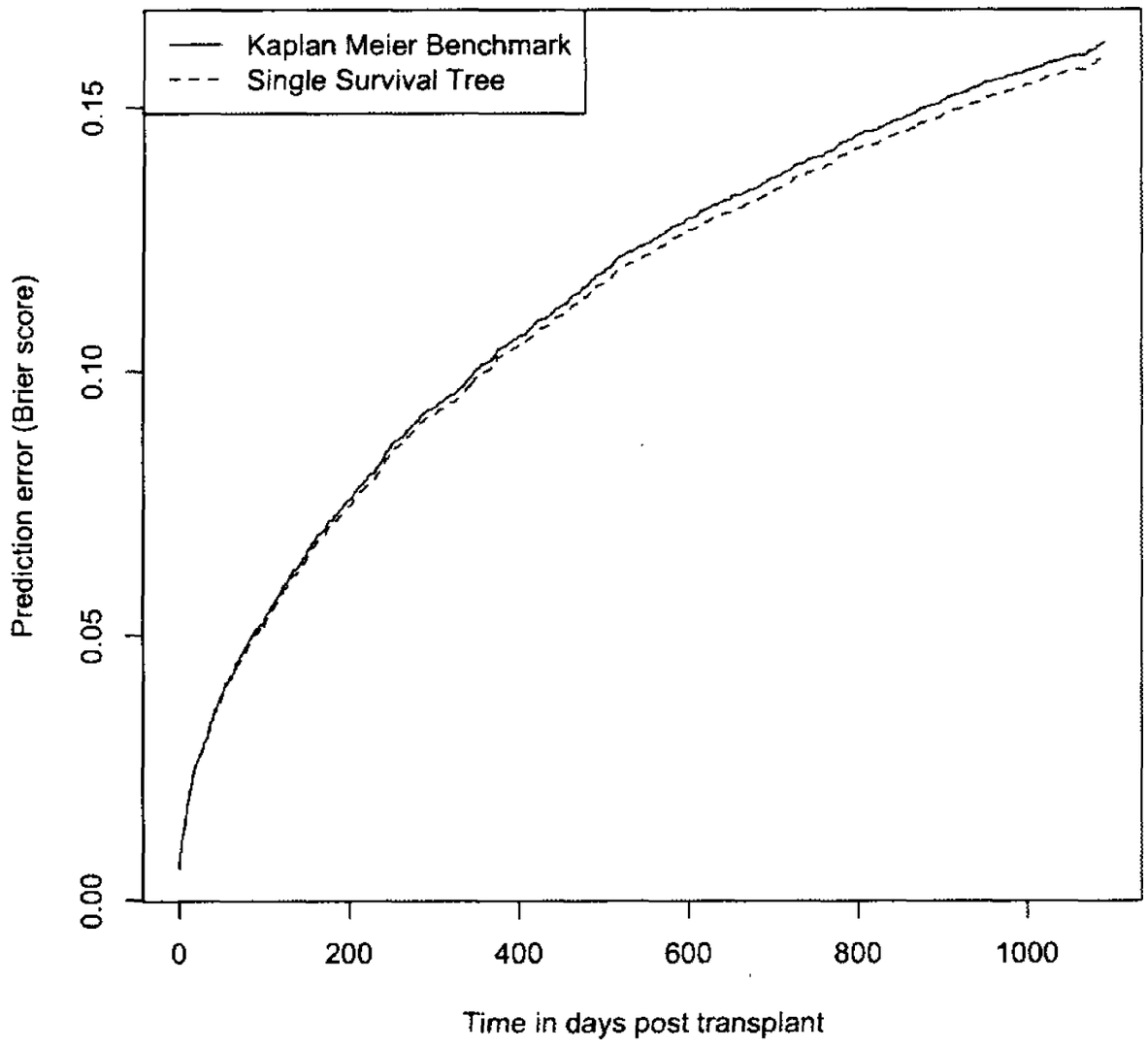


Figure 5: Prediction error curves for the single survival tree compared to the benchmark Kaplan-Meier curve (covariate information ignored).



setting, however it had the highest prediction error (IBS = 0.1122). We also calculated Harrell's concordance statistic on the test set for both Cox models. Cox model 1 (with cutpoints) had a c-statistic of 0.620 while Cox model 2 (fractional polynomials) was 0.619.

First quartile and fourth quartile BMI was a significant predictor in Cox model 1, and the continuous BMI was a significant predictor in Cox model 2 when modelled as a fractional polynomial. However, we chose not include it in either Cox model because of concern over a larger number of data entry errors compared to other variables in the data set. We found that leaving out BMI in the Cox models improved prediction.

We did test the effect of removing probable outliers on the integrated Brier score by removing 96 records with the most likely outliers in the variables BMI and cold ischemic time. BMI less than 10 or greater than 60 were set to NA. Cold ischemic times greater than 30 hours were removed, since most transplant centres limit this time to 24 hours and it is very likely that these times were meant to be entered into the database with a unit of

minutes rather than hours. Unfortunately, it is likely that errors still remain in the database. Removing possible outliers resulted in no overall gain in improvement in Cox Model 1. In Cox Model 2 with outliers removed we saw increased prediction error. In spite of this, removing the outliers improved the c-statistic in both Cox models to 0.624 for Cox model 1 and 0.627 for Cox model 2.

With suspected outliers removed, the predictive accuracy of both survival tree methods declined substantially to achieve an integrated Brier score worse than the corresponding null model.

We also tested prediction accuracy on a reduced data set (n=20,910) in models that included diabetic status as a predictor. While diabetic status was very significant in both Cox models, we did not find that including this variable offered any improvement in the prediction error obtained by the best model without this predictor. Cox model 2 (using fractional polynomials) tied with bagged survival trees to give an improvement in prediction error over the null model of 2.4%, while Cox model 1 (using cutpoints) gave an improvement of 2.3% and the single survival tree only 1.5%. Further details of the model fit can be found in Tables 11 and 12 of the Appendix.

5.3 Cox Model 1: Using Cutpoints

The results of the model fit for the Cox model with cutpoints for selected continuous variables are shown in Table 3 together with the estimated hazard ratios and p-values from the likelihood ratio test. Table 4 shows the results of the model fitted to the entire data set (training and test data). Three variables that were borderline or not significant at a level of 5% were kept in the model because they are known to be important predictors of outcome: diagnosis of hepatocellular carcinoma, receiving a liver from a non heart beating donor, and a recipient having previous incident of portal vein thrombosis. We also tested the performance of this model incorporating stratification by transplant centre; however we found that this did not improve the prediction accuracy of the model.

Table 3: Results of Cox Model 1 (using cutpoints) on the training data.

Variable	β	$exp(\beta)$	p-value
Age at transplant (yrs)	0.0096	1.0096	0.0242
Age > 55 yrs	-0.0551	0.9464	0.4533
Diagnosis: cholestatic cirrhosis	-0.2621	0.7694	0.0034
Diagnosis: hepatocellular carcinoma (HCC)	0.1068	1.1127	0.0895
Diagnosis: hepatitis C virus (HCV)	0.2578	1.2941	< 0.0001
Dialysis in the week prior to transplant	0.2381	1.2688	0.0125
Recipient medical condition: in ICU	0.3669	1.4433	< 0.0001
Recipient medical condition: hospitalized not in ICU	0.2578	1.2941	< 0.0001
Recipient on life support	0.4818	1.6189	< 0.0001
Recipient prior portal vein thrombosis	0.1879	1.2067	0.0600
Recipient prior abdominal surgery	0.1889	1.2080	< 0.0001
Creatinine: 4th quartile	0.2118	1.2358	< 0.0001
Albumin: 1st quartile	0.2477	1.2811	< 0.0001
Non heart beating donor	0.1978	1.2187	0.0701
Donor age (yrs): 40 to 49	0.2345	1.2642	< 0.0001
Donor age (yrs): 50 to 59	0.2869	1.3322	< 0.0001
Donor age (yrs) > 60	0.3384	1.4027	< 0.0001
Donor race: white	-0.1648	0.8480	< 0.0001
Cold ischemic time (hours)	0.0196	1.0198	0.0002
Age at transplant * age > 55 yrs	0.0281	1.0285	0.0002
Recipient HCV status * Donor age (yrs) > 60 years	0.3633	1.4381	0.0004

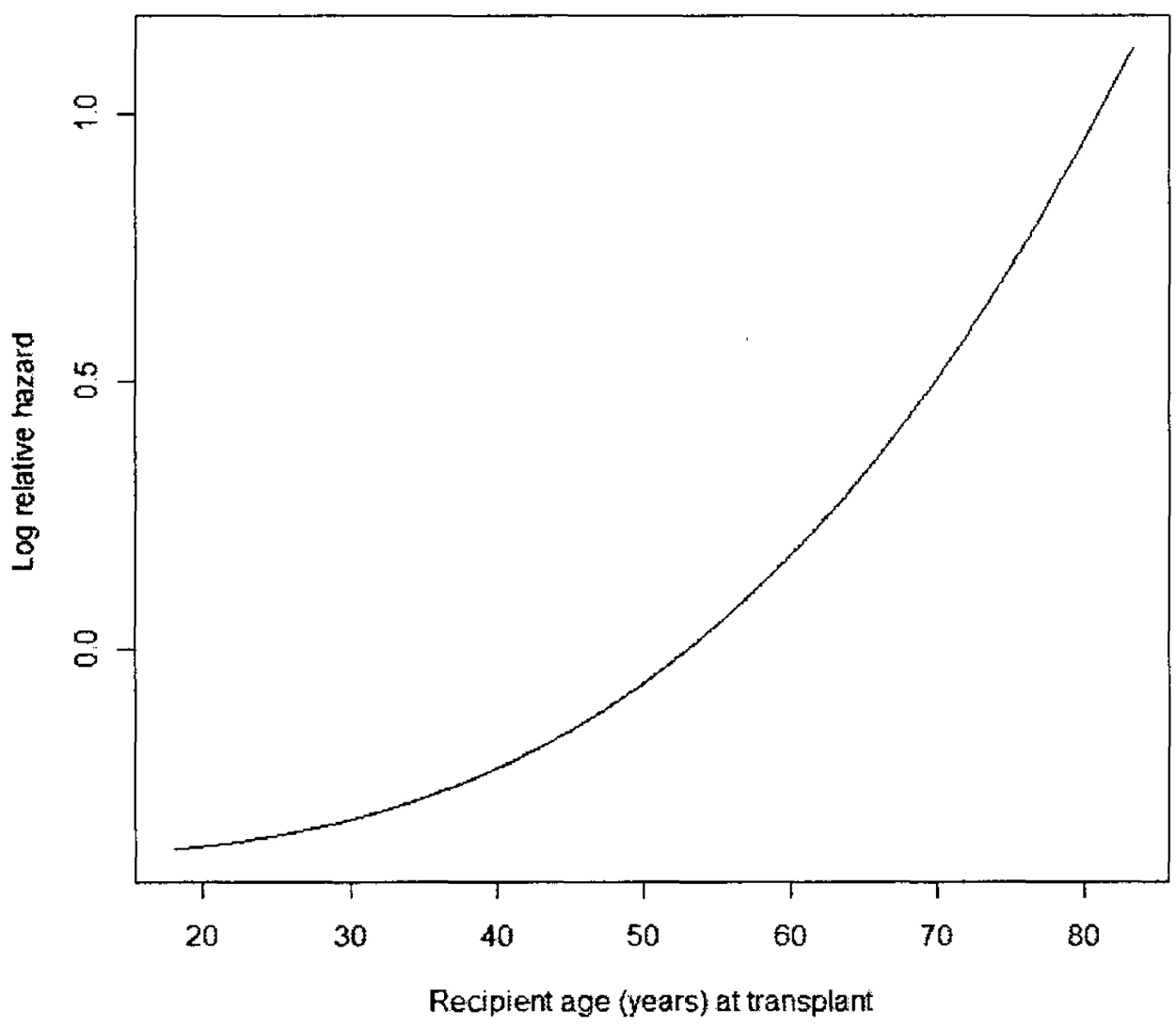
Table 4: Results of Cox Model 1 (using cutpoints) on the entire data set.

Variable	β	$exp(\beta)$	p-value
Age at transplant (yrs)	0.0098	1.0098	0.0041
Age > 55 yrs	-0.0132	0.9869	0.8234
Diagnosis: cholestatic cirrhosis	-0.2465	0.7816	0.0007
Diagnosis: hepatocellular carcinoma (HCC)	0.1522	1.1644	0.0023
Diagnosis: hepatitis C virus (HCV)	0.2233	1.2502	< 0.0001
Dialysis in the week prior to transplant	0.1971	1.2178	0.0115
Recipient medical condition: in ICU	0.3511	1.4206	< 0.0001
Recipient medical condition: hospitalized not in ICU	0.1633	1.1774	0.0014
Recipient on life support	0.4721	1.6034	< 0.0001
Recipient prior portal vein thrombosis	0.2202	1.2463	0.0056
Recipient prior abdominal surgery	0.1745	1.1907	< 0.0001
Creatinine: 4th quartile	0.2463	1.2792	< 0.0001
Albumin: 1st quartile	0.2072	1.2302	< 0.0001
Non heart beating donor	0.1933	1.2132	0.0258
Donor age (yrs): 40 to 49	0.2506	1.2848	< 0.0001
Donor age (yrs): 50 to 59	0.2980	1.3471	< 0.0001
Donor age (yrs) > 60	0.2851	1.3299	< 0.0001
Donor race: white	-0.1307	0.8775	0.0004
Cold ischemic time (hours)	0.0199	1.0201	< 0.0001
Age at transplant * age > 55 yrs	0.0216	1.0218	0.0004
Recipient HCV status * Donor age (yrs) > 60 years	0.4706	1.6010	< 0.0001

5.4 Cox Model 2: Using Fractional Polynomials

Table 5 shows the covariates used in construction of the Cox model using fractional polynomials. The results of the fit on the full data set are shown in Table 6. Note that in the fractional polynomial procedure, all predictors are shifted and re-scaled before being power transformed if non-positive values are encountered or the range of the predictor is reasonably large (Ambler & Benner, 2008). The interaction between the continuous

Figure 6: Effect of recipient age (years) on the log relative hazard of survival.



covariate donor age and the binary variable HCV status was modelled according to the algorithm for multivariable fractional polynomial interaction suggested in Sauerbrei et al. (2007). The fractional polynomial algorithm chose the best fitting polynomial for age at transplant to be $\beta * (age \cdot at \cdot transplant/100)^3$. As noted by Sauerbrei (2006), β and $exp(\beta)$ are not readily interpretable for fractional polynomials. A better approach is to plot the log relative hazard, or fitted function, against X. The log relative hazard function shows the relative effect of the predictor, with the baseline hazard removed. A plot of the log relative hazard vs recipient age is shown in Figure 6. Pre-transplant serum creatinine was

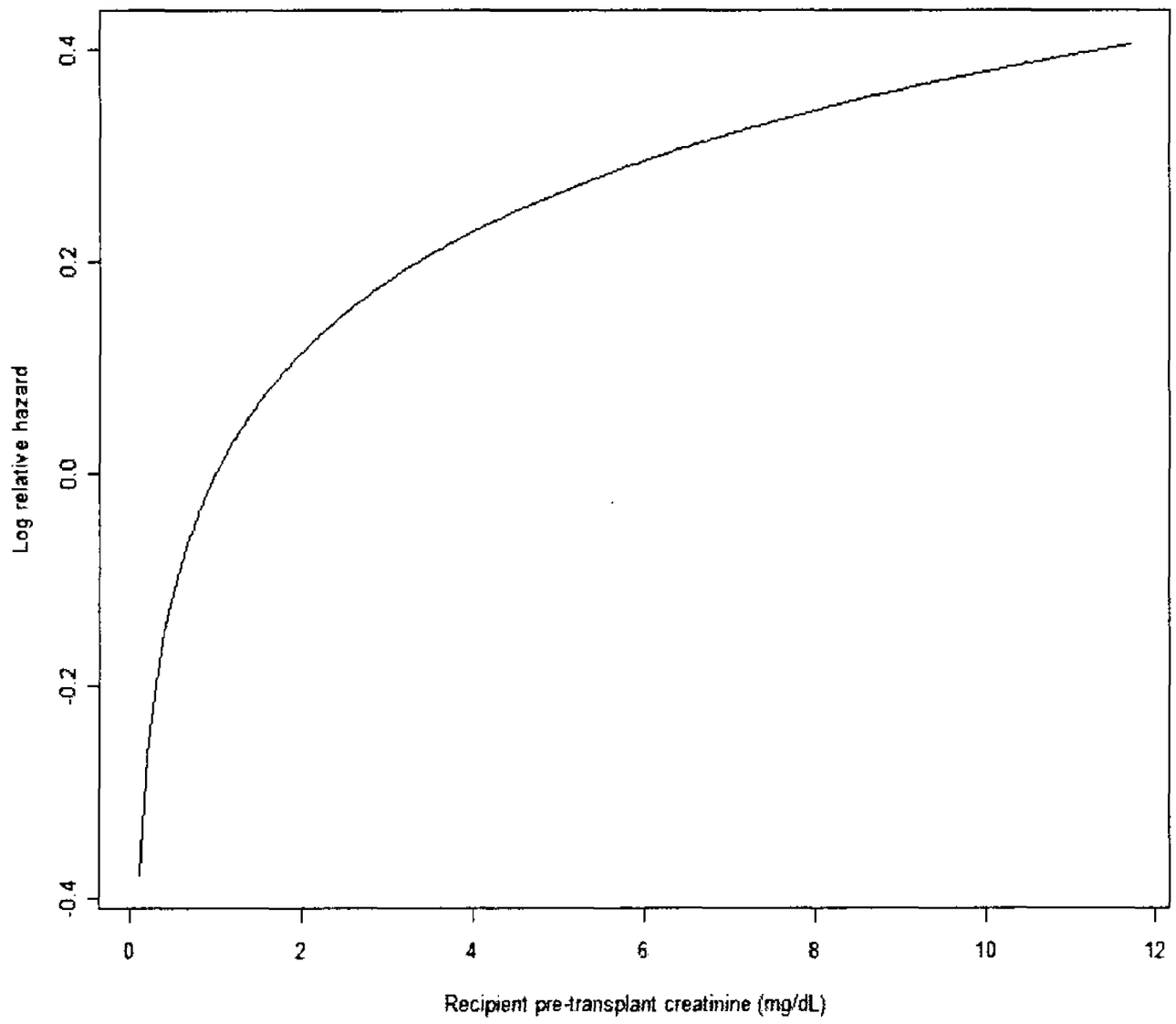
Table 5: Results of Cox Model 2 (using fractional polynomials) on the training data.

Variable	β	$exp(\beta)$	p-value
$(Recipient \cdot age(yrs) \cdot at \cdot transplant/100)^3$	$2.6 * 10^{-6}$	1.0000	< 0.0001
Donor age (yrs) * recipient HCV positive	0.0168	1.017	< 0.0001
Albumin (g/dL)	-0.1826	0.8331	< 0.0001
Recipient medical condition: in ICU	0.3669	1.443	< 0.0001
Recipient medical condition: hospitalized not in ICU	0.2614	1.299	< 0.0001
Recipient on life support	0.4872	1.628	< 0.0001
Recipient prior abdominal surgery	0.1888	1.208	< 0.0001
Donor age (yrs) * recipient HCV negative	0.0060	1.006	0.0003
Cold ischemic time (hours)	0.0192	1.019	0.0003
Donor race: Hispanic	0.2155	1.240	0.0005
Diagnosis: Cholestatic cirrhosis	-0.2515	0.7776	0.0050
log(Creatinine) (mg/dL)	0.1647	1.179	0.0002
Donor race: Black	0.1386	1.149	0.0215
Dialysis in the week prior to transplant	0.2181	1.244	0.0247
Diagnosis: hepatocellular carcinoma	0.1337	1.143	0.0347
Recipient prior portal vein thrombosis	0.1763	1.193	0.0773
Non heart beating donor	0.1940	1.214	0.0753
Diagnosis: Hepatitis C virus	-0.1677	0.8456	0.1540

Table 6: Results of Cox Model 2 (using fractional polynomials) on the entire data set.

Variable	β	$exp(\beta)$	p-value
$(Recipient \cdot age(yrs) \cdot at \cdot transplant)^3$	$2.4 * 10^{-6}$	1.0000	< 0.0001
Donor age (yrs) * recipient HCV positive	0.0172	1.0170	< 0.0001
Albumin (g/dL)	-0.1630	0.8496	< 0.0001
Recipient medical condition: in ICU	0.3591	1.432	< 0.0001
Recipient medical condition: hospitalized not in ICU	0.1739	1.190	0.0007
Recipient on life support	0.4766	1.611	< 0.0001
Recipient prior abdominal surgery	0.1750	1.191	< 0.0001
Donor age (yrs) * recipient HCV negative	0.0059	1.006	< 0.0001
Cold ischemic time (hours)	0.0196	1.020	< 0.0001
Donor race: Hispanic	0.1824	1.200	0.0003
Diagnosis: Cholestatic cirrhosis	-0.2399	0.7867	0.0010
log(Creatinine) (mg/dL)	0.1690	1.184	< 0.0001
Donor race: Black	0.0801	1.083	0.1039
Dialysis in the week prior to transplant	0.1811	1.199	0.0227
Diagnosis: hepatocellular carcinoma	0.1753	1.192	0.0005
Recipient prior portal vein thrombosis	0.2019	1.224	0.0111
Non heart beating donor	0.1955	1.216	0.0240
Diagnosis: Hepatitis C virus	-0.1961	0.8219	0.0381

Figure 7: Effect of creatinine (mg/dL) on the log relative hazard of survival.



modelled as $\beta * log(creatinine)$. Figure 7 shows a plot of the fitted function (log relative hazard) vs pre-transplant creatinine. The algorithm fitted all other continuous covariates with no transformation since significant evidence of non-linearity was not found. We did not include BMI in the model because of concern over data entry errors. However, when BMI is included, the algorithm chooses a second degree polynomial for BMI that gives a parabolic shape, with dramatically increasing hazard as BMI moves below 20 or above 35.

5.5 Testing the Proportional Hazards Assumption

For the two Cox models we examined whether the assumption of PH was met. The assumption should hold for each covariate in the model. Once potential issues with non-proportionality are identified, Therneau (2000) suggests a strategy which first involves the determination of whether the effect is meaningful. It is often the case that significant non-proportionality may not have a substantial impact, particularly with large sample sizes as is the case here. Therneau suggests for time-fixed categorical variables to plot the Kaplan-

Figure 8: Log-log survival curves for recipient medical condition.

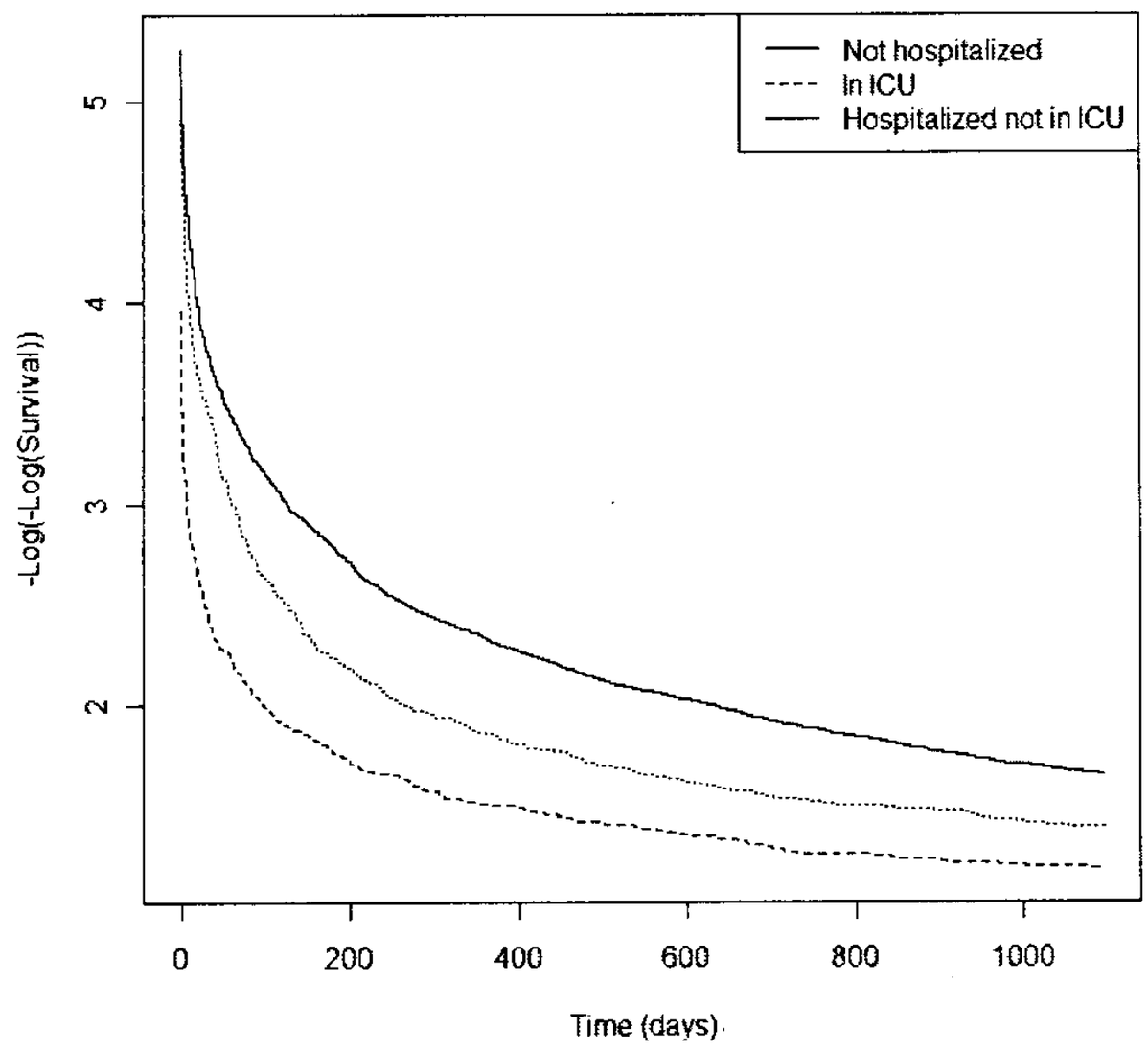


Figure 9: Log-log survival curves for HCV status.

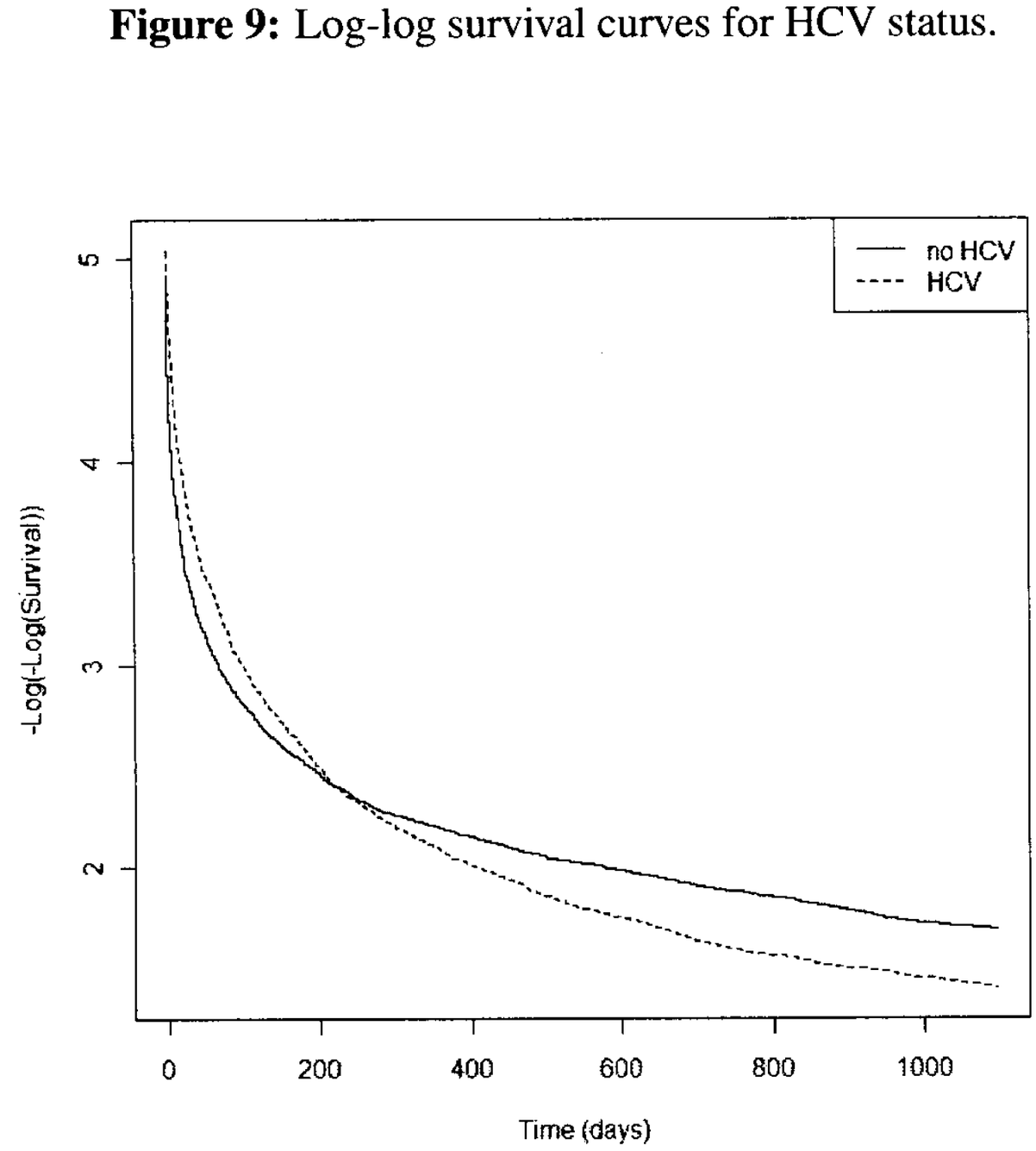
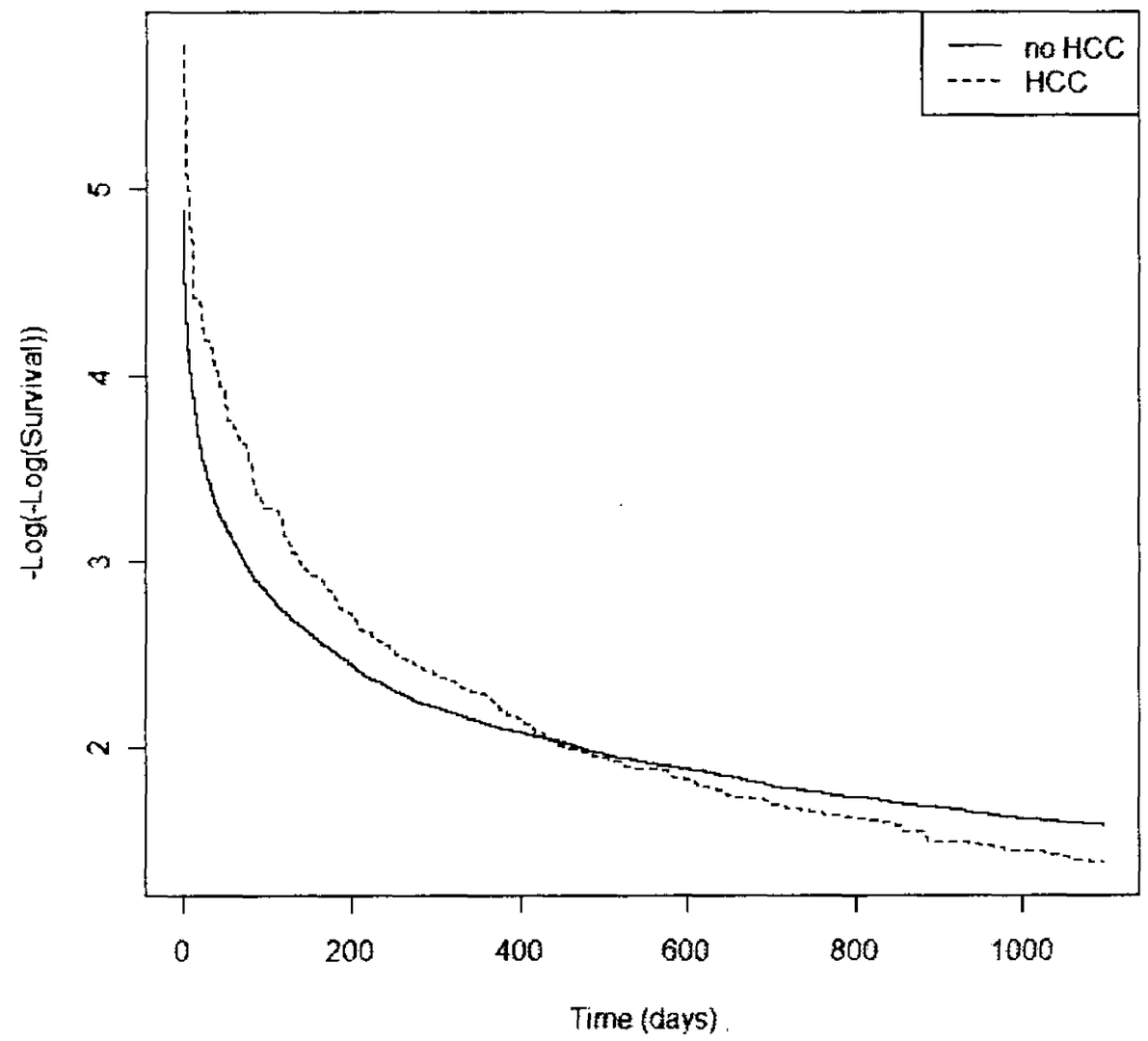


Figure 10: Log-log survival curves for HCC status.



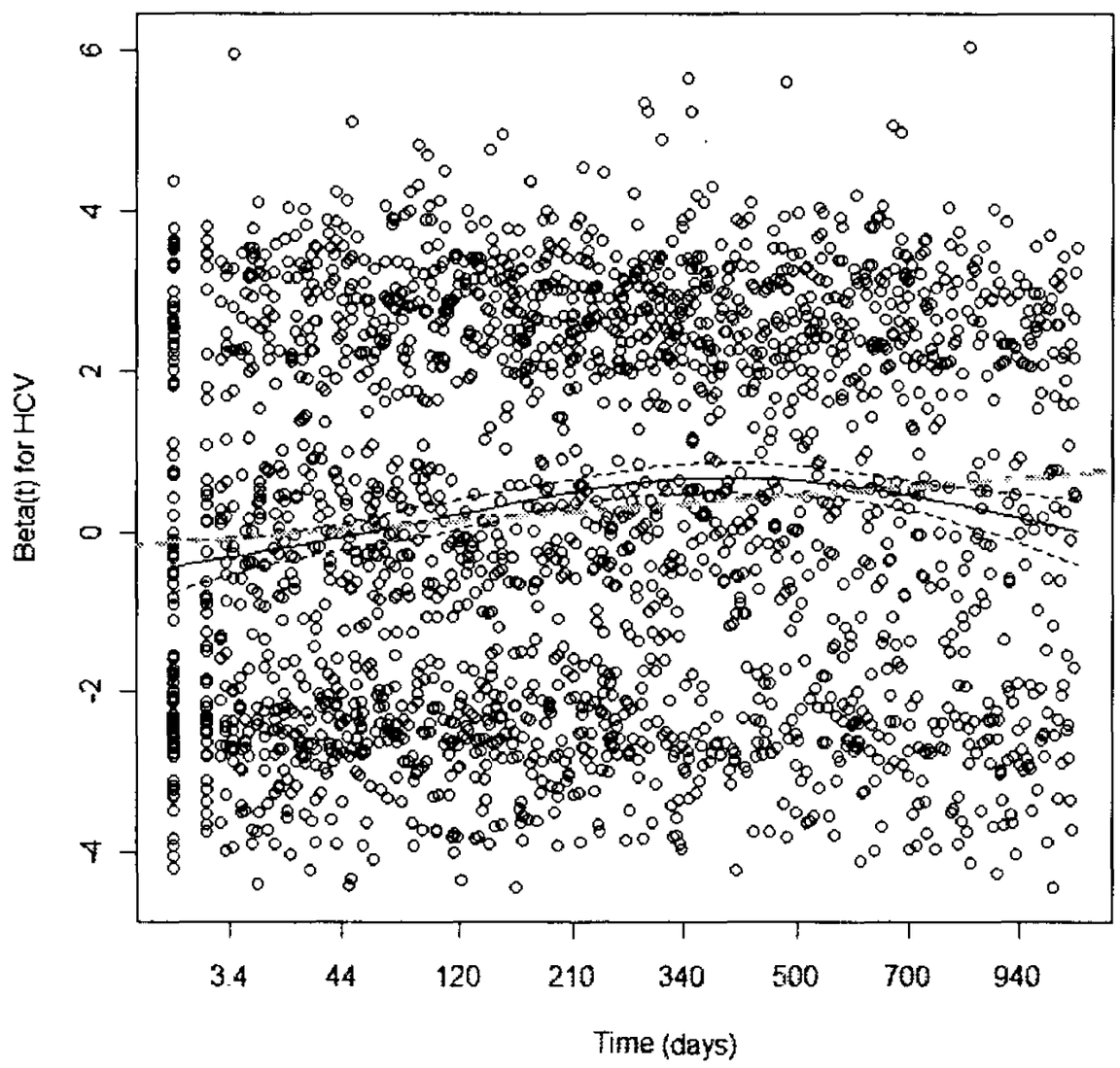
Meier curves for the levels on a log-log scale. Figure 8 shows the Kaplan-Meier survival curves for the three levels of the variable "recipient medical condition." If the PH assumption holds, the three curves should be roughly parallel and we can see no cause for concern with this variable, except perhaps near the time of transplant. However, with HCV status we can see crossing survival curves (see Figure 9). The effect of HCV status on the hazard appears to vary over time, with a slightly protective effect early on for HCV positive recipients and a serious decline beginning approximately 220 days after transplant. This could be due to avoidance by physicians of transplanting marginal livers into patients who are HCV positive, thus giving these subjects an initial advantage. A search of the medical literature reveals evidence that HCV positive patients have a poorer prognosis when receiving higher risk livers (e.g. livers from older donors or fatty livers, see Feng, 2006). Receiving a higher quality liver has a protective effect early on but a decline in survival is seen once the recurrence of the Hepatitis C virus begins to have an effect on the new liver.

Similarly, crossing hazards are also seen with a diagnosis of hepatocellular carcinoma,

with curves crossing after about 420 days after transplant (see Figure 10). Patients with HCC are transplanted earlier in order to avoid spreading of the cancer beyond the liver. However, they are known to have problems with late occurring recurrence of cancer. The same crossing hazards are seen if dummy variables for viral liver disease (including viral liver disease other than Hepatitis C) or malignancy (including cancers other than HCC) are used.

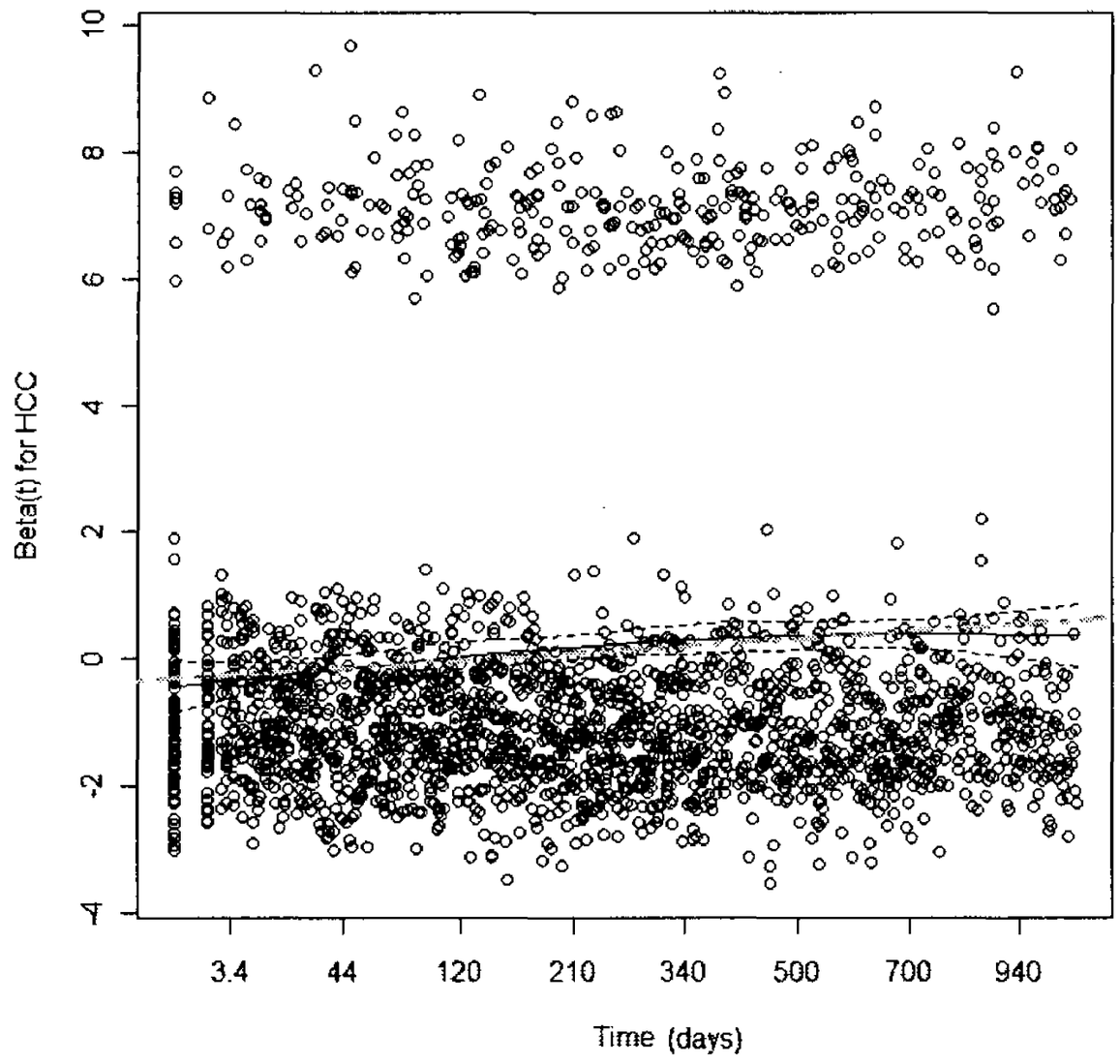
5.5.1 Testing Proportional Hazards: Cox Model 1

Figure 11: Scaled Schoenfeld residuals for HCV status, Cox Model 1.



The Cox model with cutpoints had five variables potentially violating the PH assumption based on significant p-values obtained from the Grambsch-Therneau test: two were as expected, HCV status and HCC status. The three other variables were recipient medical condition, recipient previous abdominal surgery, and the dummy variable for first quartile albumin. These variables are among the most significant predictors in the model. Model-specific plots of scaled Schoenfeld residuals are useful for detecting non-

Figure 12: Scaled Schoenfeld residuals for HCC status, Cox Model 1.

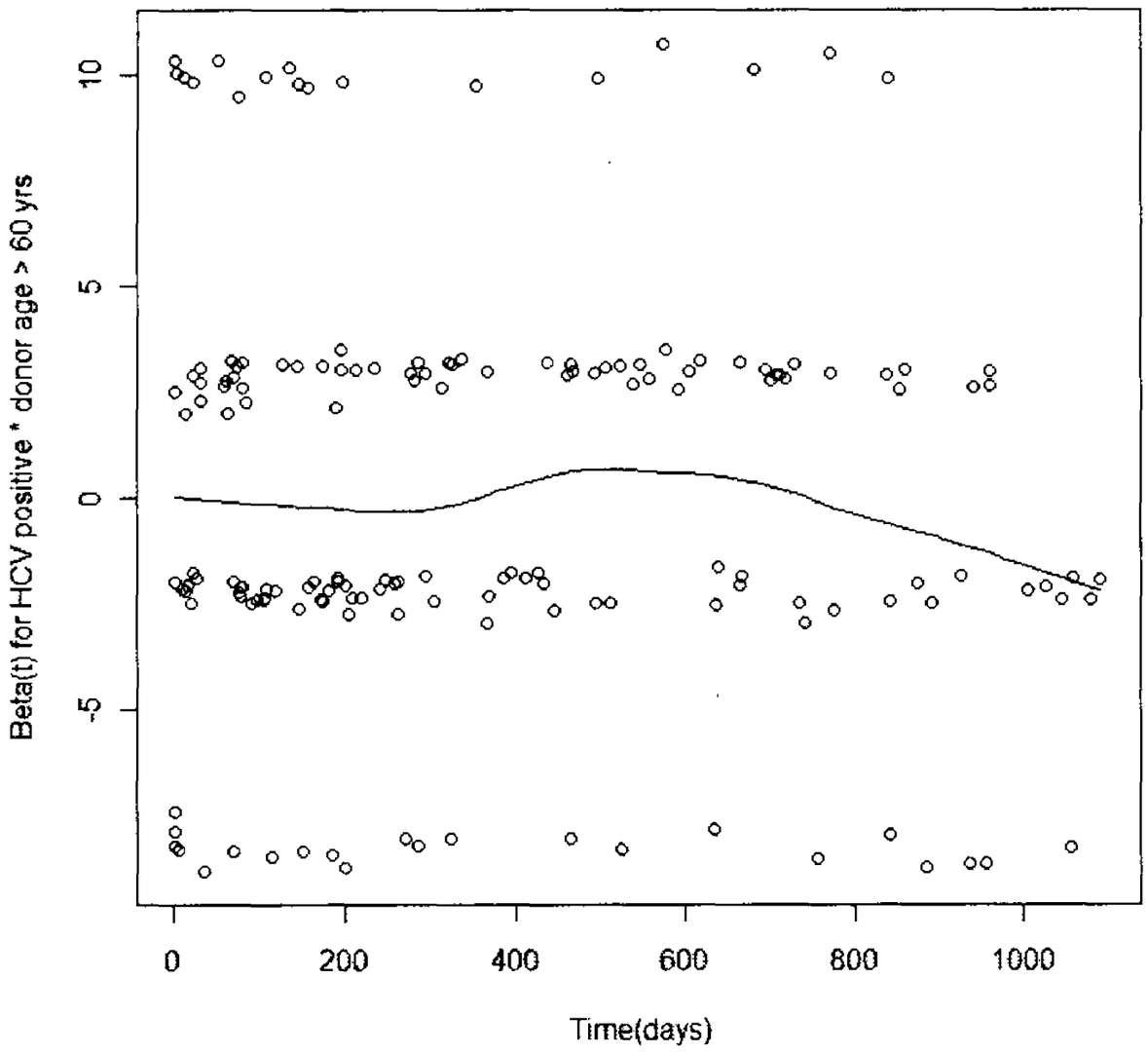


proportionality. A smoothed line is added to the plot and any evidence of non-zero slope is suggestive of non-proportionality (Therneau, 2000) . The residual plots for recipient medical condition show a slight decreasing pattern which may suggest that the negative effect of being in ICU at time of transplant gradually decreases over time - a reasonable clinical conclusion.

Figure 11 shows the scaled Schoenfeld residuals for HCV under Cox model 1, along with the fitted least squares line. The y-axis shows the time-dependent variable beta(t), which gives an estimate of the correlation of HCV status with time. If a covariate is not correlated with time, the plot of the partial residuals against time should have a zero slope. These residual plots can show a banding effect, caused by censoring or by categorical covariates, which can be ignored (Grambsch, 1994). The plot shows that a positive HCV status is protective initially but the effect declines over time. Then, it appears to become protective again after two years if the subject has survived that long. The interaction with donor age > 60 years is not significantly non-proportional and the residual plots for this

interaction are unremarkable. However, if the residuals for the HCV positive and HCV negative patients are separated in this interaction, the resulting plots are still suggestive of non-proportionality, illustrated in Figures 13 and 14. Venables & Ripley (2002) sug-

Figure 13: Scaled Schoenfeld residuals for HCV positive * donor age > 60 years interaction.



gest a search for interactions such as the one used here as one possible way to deal with non-proportionality in a covariate. Kleinbaum (2005) suggests using an adjusted log-log survival curve to assess the effect of the interaction. Figure 16 shows the fit of a Cox PH model stratified by HCV status and adjusted for donor age > 60 years. The plot shows the survival curves still cross and is almost unchanged from the unadjusted version shown in Figure 9. It is unlikely that the interaction used here addresses the non-proportional hazards in HCV status.

A plot of the scaled Schoenfeld residuals for HCC status (Figure 12) shows the same effect as HCV status although the protective effect at the beginning and end of the time period is less pronounced. An important strategy for handling non-proportionality in categorical covariates is stratification, however, we found that stratification by either HCV

Figure 14: Scaled Schoenfeld residuals for HCV negative * donor age > 60 years interaction.

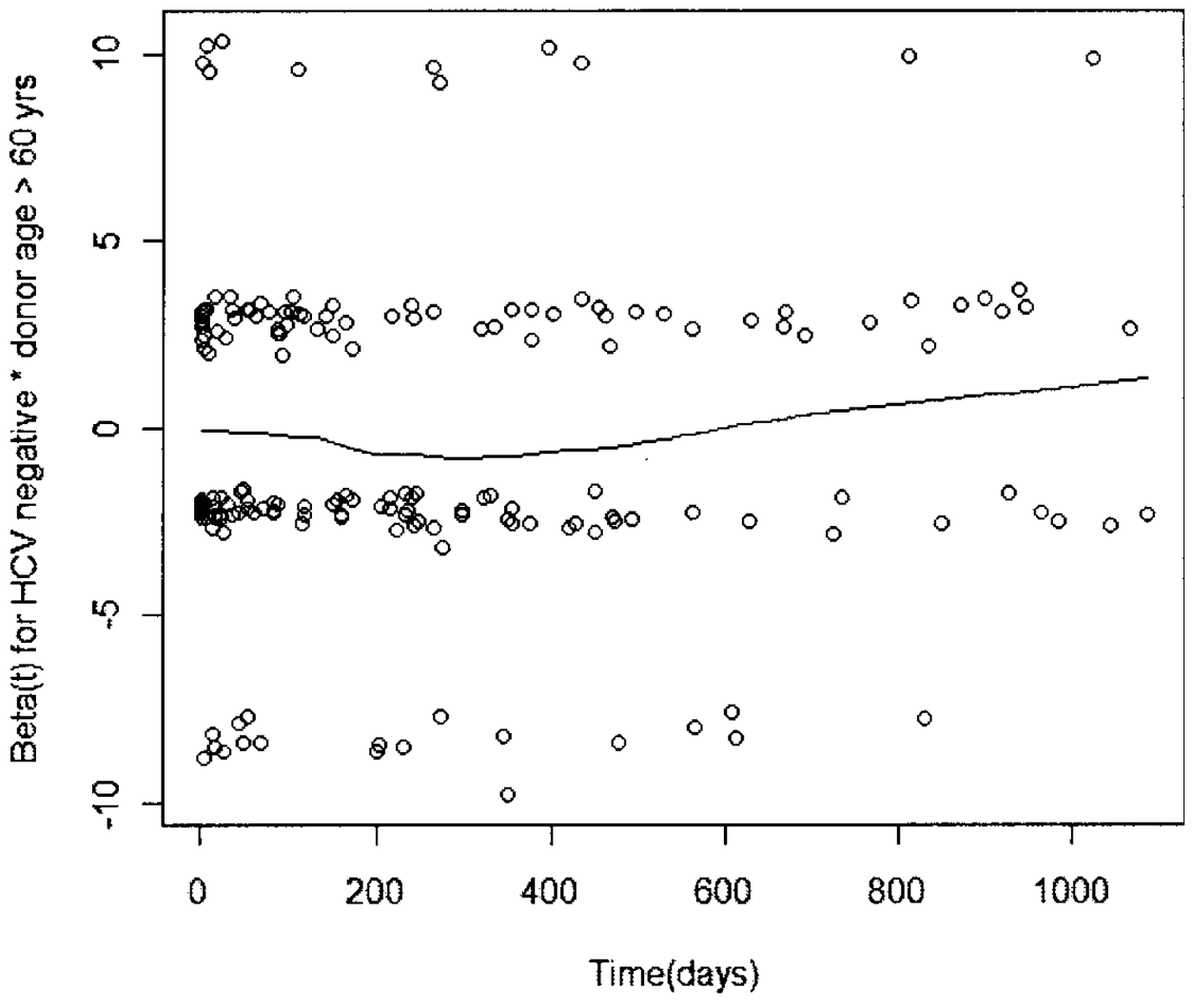
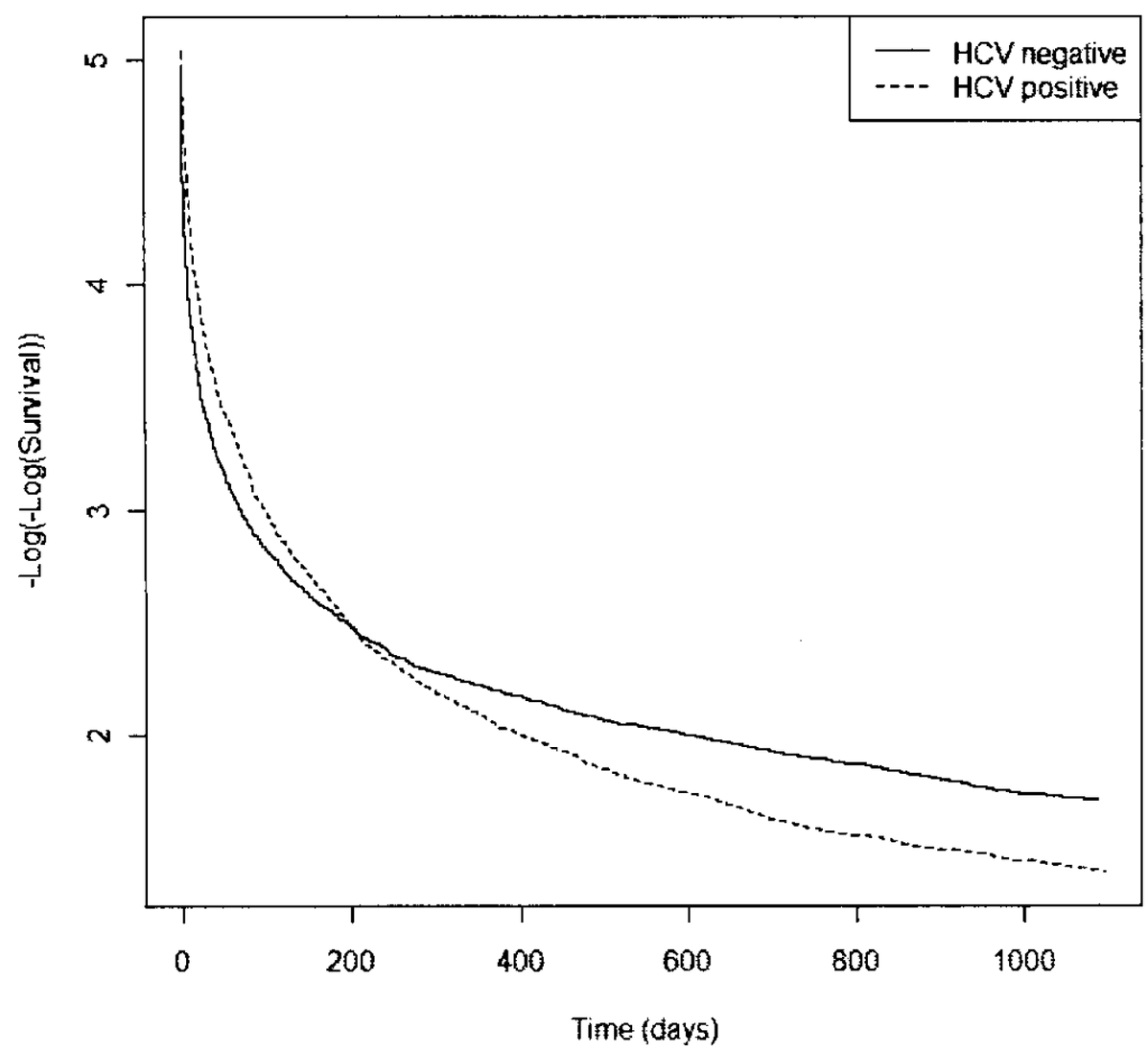


Figure 15: Log-log survival curves for HCV status using the Cox PH model adjusted for donor age > 60 years.



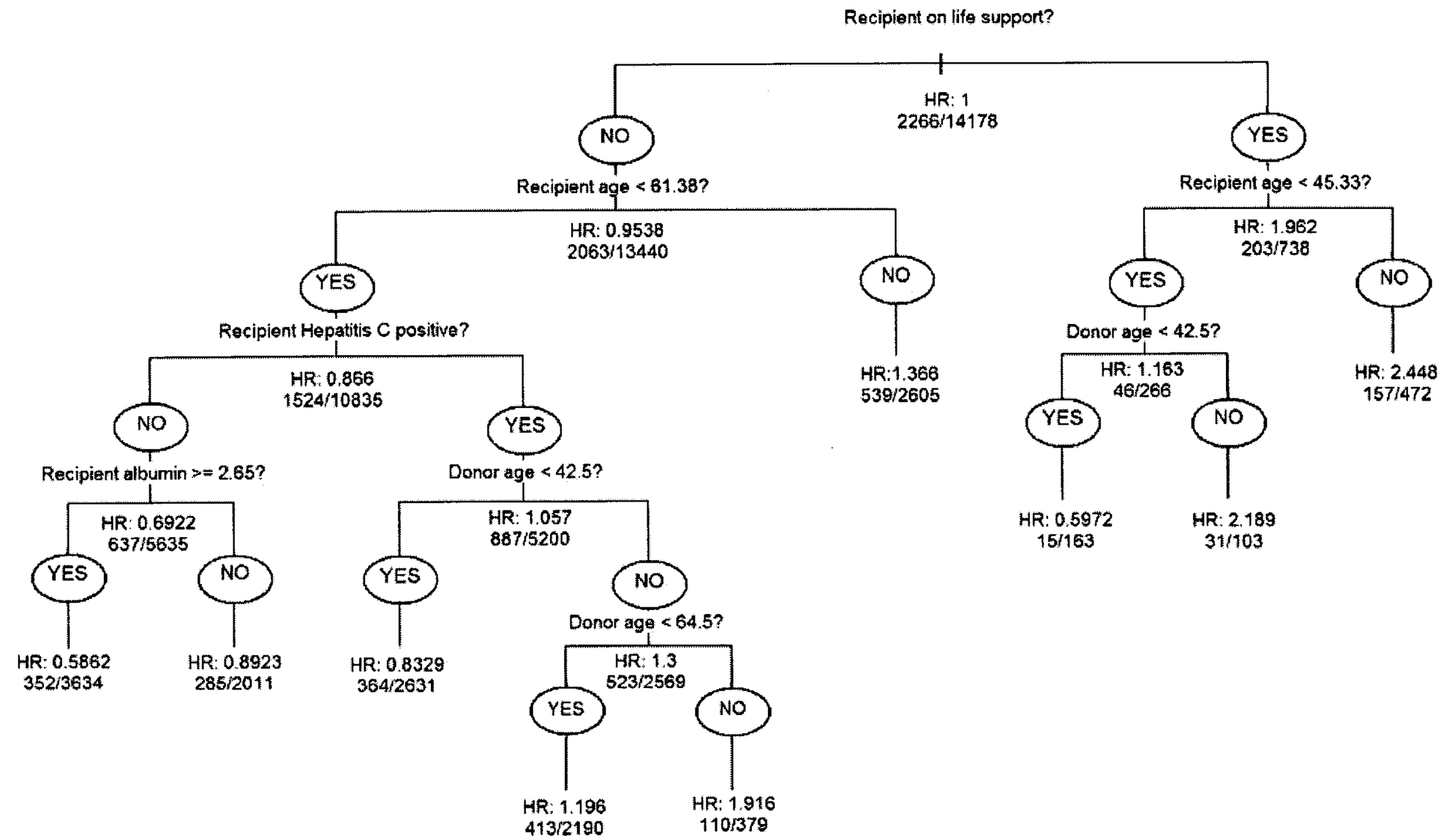
or HCC status did not improve predictive performance. Another alternative is to consider a piecewise PH model or time-by-covariate interactions, but these can be difficult to implement in such a large data set. When we test a model which removed the 96 outliers mentioned above, the result was that first quartile albumin was no longer significant in the Grambsch-Therneau test for non-proportionality, although it remained borderline at p=0.065. HCV, HCC, recipient previous abdominal surgery and recipient medical condition (in ICU) remained significantly non-proportional. The model which included diabetic status as a predictor had similar issues with non-proportionality.

5.5.2 Testing Proportional Hazards: Cox Model 2

The Grambsch-Therneau test for non-proportionality identified four covariates in Cox model 2 with potential issues: recipient medical condition, previous abdominal surgery, creatinine, and diagnosis of hepatocellular carcinoma. Neither HCV nor its interactions were identified as significantly non-proportional, although the residual plot for HCV status showed the same pattern as in Cox Model 1. Log-log survival curves for HCV status using the Cox PH model adjusted for donor age as a continuous variable look very similar to Figure 16. HCC status has already been identified as problematic. We have already noted that the log-log survival plots for recipient medical condition and previous abdominal surgery showed no major cause for concern. The residual plots for recipient medical condition were similar to those seen in Cox model 1. The significant finding of non-proportionality for creatinine could indicate that the functional form chosen is not optimal.

When we remove the 96 outliers in BMI (BMI < 10 or > 60) and cold ischemic time (CIT > 30), all variables identified as significantly non-proportional still remain so in the new model. Prediction error is not improved. The model which included diabetic status as a predictor had similar issues with non-proportionality.

Figure 16: Single survival tree constructed from the training data set.



5.6 The Single Survival Tree

The single survival tree used 5 covariates in tree construction: donor age, recipient age, albumin at time of transplant, Hepatitis C status, and for the initial split, whether or not the recipient was on life support at the time of transplant. Figure 16 shows the single survival tree constructed using the training set. Under each node, the hazard ratio is provided. Below that is the number of deaths divided by the total number in the node at each point in the tree.

After removing the 96 outliers, the survival tree generated by the recursive partitioning algorithm changes the structure of the tree substantially, with the main split now based on recipient creatinine, a variable which was not included in the original tree. Cold ischemic time now appears in the tree, and albumin is no longer present. Recipient age, donor age, recipient HCV status and whether the recipient was on life support are still used in tree construction. However, we found that the predictive accuracy has suffered, with the model achieving an integrated Brier score higher than the corresponding null model with outliers removed. This could be due to possible errors remaining in the variable cold ischemic time, now used in prediction.

5.7 Bagged Survival Trees

A disadvantage of using bagged trees is that we can no longer present the model in a simple tree form. Predictions are made based on each of the bootstrapped trees and aggregated according to the method chosen. Each bootstrapped tree has different characteristics. The median number of nodes in the 50 bagged trees was 21. The median number of nodes in the 300 bagged trees was 21.5.

6 Discussion

6.1 Brier Scores and Concordance Statistic

The prediction error curves provided by the Brier score are a valuable tool for describing the uncertainty of a model over time. These curves can also be drawn for an individual patient. The IBS provides an important description of the overall uncertainty of a predictive model.

The concordance statistics for the two Cox models were similar and were for the most part slightly below the optimistic c-statistics reported in liver transplant survival literature where concordance was calculated on the same data set used to build the model. With heavy censoring, it is likely that the c-statistics found here are inflated. We have also shown that the PH assumption was not satisfied for either Cox model tested here which means the c-statistic is not valid as a performance measure.

The data-splitting approach taken here is not ideal since the indices of accuracy may vary with different splits (Harrell, 1996). Indeed, with the single survival tree we see how minor perturbations in the data influenced tree construction and corresponding Brier score. It is therefore possible that the integrated Brier scores found here are overly optimistic. Van Wieringen et al (2009), in a consideration of evaluation methods for predictive models using gene expression data, showed that the best prediction method can depend on the data set used. Prognostic models for liver transplantation should ideally be tested on similarly large transplant registry data sets such as those from Canada or Europe. Here we randomly split our data set in order to obtain a training set and a test set. However, a problem with this method is that the data used for testing are statistically homogeneous with the training data. May et al. (2004) write that assessment based on independently collected data would provide more realistic results. Formal validation in a prospective study is recommended by Altman et al. (2009). Further, it is recommended by Moons et al. (2009) that any prognostic model be adjusted and updated over time to accommodate changes in practice. A

better approach than the one taken here would be to compare a model built on the U.S. data to a completely separate data set such as the Canadian Organ Replacement Registry or the European Liver Transplant Registry. Transplant registries worldwide offer a unique opportunity to carry out this validation, adjustment and prospective evaluation, once a promising model is developed.

In 1986, Box and Draper wrote, "Remember that all models are wrong: the practical question is how wrong do they have to be to not be useful." How low a Brier score is required before a model can be considered a good predictor is a question that must be answered by consensus of physicians and scientists. Minimally the score should be significantly better than the score of the null model with no covariate information, however, we have shown that sometimes a significant statistical difference does not result in meaningful scientific significance. A plot of the prediction error curves comparing the prediction error to a benchmark model is an essential tool.

6.2 Simple Models vs Complex Models

It has often been noted that simple models can have better prediction accuracy than more complex models (Vittinghoff, 2005). Haibe-Kains et al. (2008), in an evaluation of the predictive performance of breast cancer survival models using large microarray data sets comprising more than 1000 patients, found no evidence that complex methods outperform the simplest prognostication techniques. They write that the result suggests that the "loss of interpretability deriving from the use of overcomplex data analysis strategies may not be sufficiently counterbalanced by an improvement in the quality of prediction." In this analysis, however, the number of events per variable was large for all models (> 100), and, as others have shown using simulation studies, there is little difference in predictive accuracy between models when events per variable are large (> 20) (Ambler 2002). In this study we thought we might find that tree models, with fewer assumptions, would predict better than a Cox PH model where assumptions are not satisfied. We discovered, however,

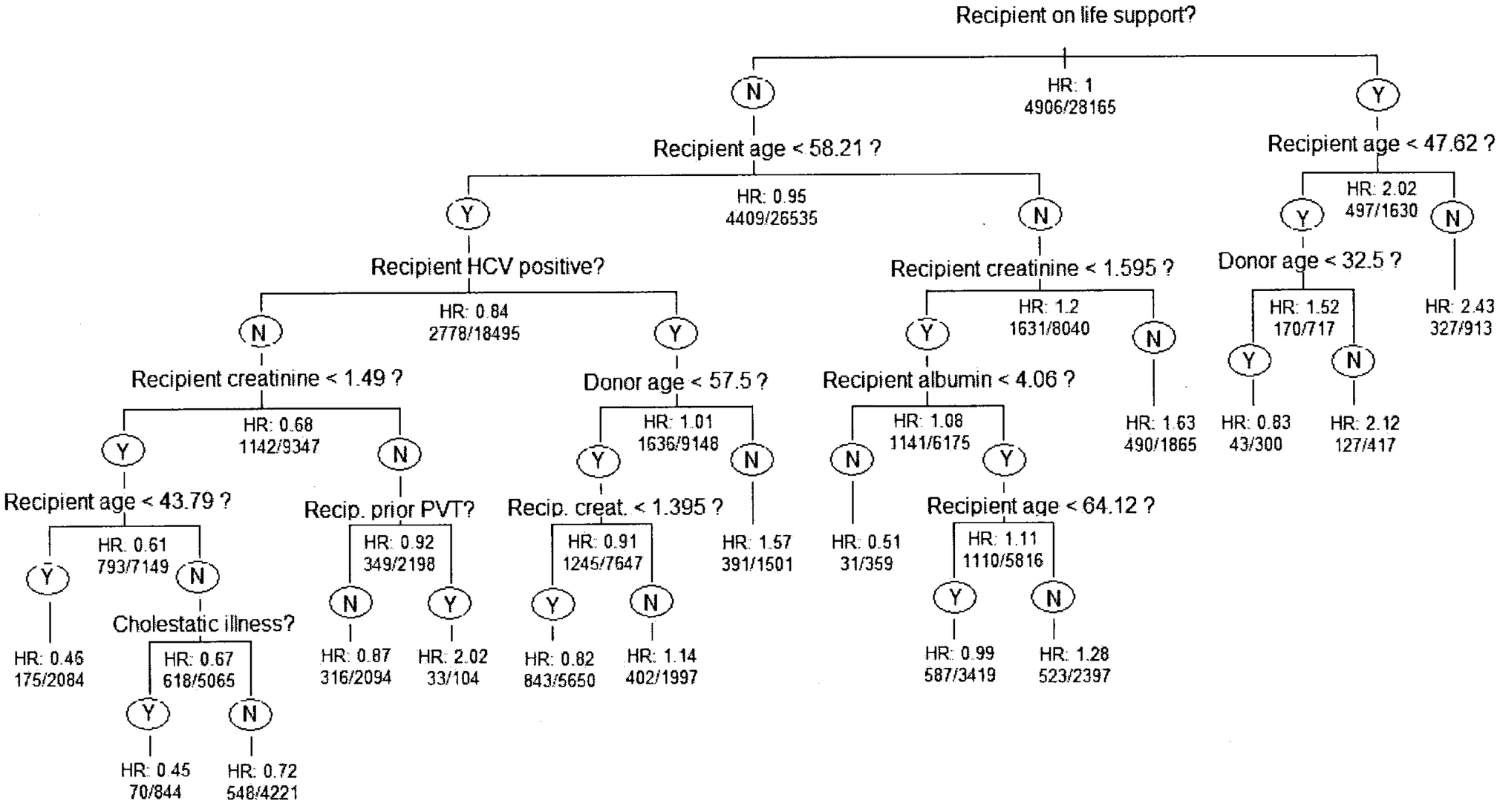
that the survival trees do not perform better, and the Cox models are the best of a group of models which are all somewhat poor predictors. The best model we tested offers only a 2.5% improvement over the benchmark model ignoring covariate information, whereas the simplest tree method offered a 1.75% improvement. The difference between the best and the worst models tested here did reach statistical significance ($p < 0.0001$).

The appealing aspects of a simple model, if one could be determined, are many - all of the variables in a model such as the single survival tree described here are well defined and unambiguous. A single survival tree provides an easily interpretable decision rule. The model does not require complex software to run and with a print copy of the tree, a transplant physician could make a decision. However, in the case of transplant registry data, we found that the single survival tree had the highest prediction error and the tree was unstable. The exercise of removing outliers in the tree models illustrated how the predictive accuracy of single survival trees is influenced by small changes in the data set. While the variables chosen by the recursive partitioning algorithm made sense in terms of known predictors, other criticisms of survival trees are readily at hand. The decision thresholds seem arbitrary and may not be clinically relevant. It is difficult to explain why a donor age of 42.5 is much worse than a donor age of 42. Wyatt and Altman (1995) write that "Model builders should try to avoid arbitrary thresholds for continuous variables." In the case of transplant registry data, a survival tree may still be useful in exploratory analysis, for example in suggesting prognostic groups that are not obvious in a PH analysis or in suggesting topics worthy of more formal research. A single tree constructed from the whole data set (including observations with missing values in the predictors) is shown in Figure 17 as an example.

6.3 Cox Model: Cutpoints vs Fractional Polynomials

The two Cox models tested here found similar covariates to be significant predictors, and even though the two continuous covariates of recipient age and creatinine were mod-

Figure 17: Single survival tree constructed from the entire data set (n=28,165) including records with missing data in any of the predictors.



elled differently, the predictive performance of both models was very similar. This could be because of the large data set where it is possible the problems seen with dichotomizing continuous variables cause fewer problems. It could also be that the quality of the data, e.g. data entry errors in cold ischemic time, influenced the predictive accuracy of one or both models. Note that the comparison presented here is not a fair one since the proportional hazards assumption was not met. In future work, the author intends to examine whether a model using fractional polynomials will perform better in a situation where the assumptions of the model are not violated.

Whether BMI is a significant predictor of survival is a matter of debate among physicians (Nair, 2002; Pelletier, 2007). The evidence from Cox Model 2, when BMI was included, suggests that BMI is best modelled with a non-linear approach. Using the OPTN data set from 1988 to 1996, Nair (2002) categorized BMI into 5 groups and using Cox regression analysis determined that a BMI > 40 kg/m^2 was a significant predictor of increased mortality risk. The smallest group used by Nair et al. included all subjects with BMI < 25 kg/m^2 which may have masked the increased risk associated with very low BMI seen here and also by Pelletier et al. (2007) in their examination of BMI and its effect on the survival benefit of liver transplantation. It is also possible that an interaction between BMI and another covariate such as age or diabetic status is further complicating the picture.

Pre-transplant diabetic status has been identified as a possible predictor of poor outcome after liver transplant (Schaubel, 2009). Recent research has identified a greater risk of the development of serious liver disease in patients already diagnosed with Type II diabetes (Porepa et al., 2010). The authors postulate that the increased risk is caused either by insulin resistance, which can result in damage to the liver through fatty deposits, or by direct glycemic injury to the liver. Insulin resistance has also been associated with a more rapid progression of Hepatitis C virus recurrence after liver transplant (Lornado et al., 2008). Further complicating the picture is that development of new-onset diabetes after liver transplant is a common problem and this has also been identified as a risk factor for

poor outcome after transplant (Watt et al., 2010). When we included pre-transplant diabetic status as a covariate in our models, we did not find it improved prediction accuracy. This could be caused by interaction with the covariate for HCV status. It is also possible that further research into the association between liver disease, insulin resistance and glycemic injury will identify an important predictor that is not currently included in any models.

In this study we did not find any donor cause of death to be a significant predictor of survival, which differs from the results found in Schaubel (2009) who examined patient survival after transplant and Feng (2006) who studied graft survival. It is possible this is a result of different coding used for donor cause of death. The coding used here is available in the Appendix.

In Cox model 1 we included a donor race of White as a predictor of outcome because it gave a lower AIC and more parsimonious model than including both donor race of Black or Hispanic as predictors. In Cox model 2 we found that including donor race of Black and Hispanic gave a lower model deviance compared to using donor race of White only. Changing either model to include the same donor race predictors made no difference in predictive accuracy. Donor race is included in the donor risk index proposed by Feng in 2006. However, the topic is a matter of debate among transplant physicians. Some suggest that stratifying by centre removes donor race as a predictor of survival. Asrani et al. (2010) tested such a model on the SRTR data and still found that their donor race "other" category, which included Hispanic status, was a significant predictor of survival. We also tested a model with stratification by centre and with donor race of Hispanic origin in its own category using dummy variables. We found that a donor race of Hispanic origin remained a significant predictor of poorer outcome while the significance of donor race equal to Black disappeared. We also found that stratification by centre did not improve prediction error.

6.4 Is the Cox Proportional Hazards Model Appropriate for Transplant Data?

The violation of the primary assumption of the Cox PH model by important predictors of survival indicate that it is not the best choice for transplant data. However, the issue of non-proportionality is often not addressed. The non-proportionality seen in HCV and HCC status in particular make the PH model inappropriate for transplant data. As seen in Table 1, patients with a diagnosis of Hepatitis C comprise close to half of the records in the data set. The Cox model can handle non-proportional baseline hazards by stratification on the offending covariate, but we found that stratification by HCV or HCC did not improve the predictive accuracy of the models. This emphasizes the need to look beyond the Cox PH model for more suitable methods which do not transgress the assumptions of the model. A model such as the one proposed by Zeng and Lin (2007) that can handle crossing hazards may be a better choice.

Schaubel and Wei (2007) considered an additive hazards model for survival on the waiting list for liver transplantation. This is another possible alternative to the PH model that could be applied to survival after transplantation.

Poor model fit is not confined to liver transplant models. Schemper (2003) writes that scientists cannot rely on statistical significance of covariates as evidence of good model fit, since "even strong and highly significant covariates of a study may not automatically translate into sufficiently accurate prediction or close determination of individual outcome values." There are also issues specific to liver transplant that need to be addressed in any model of survival, specifically the non-proportional hazards mentioned already, early failure times, newly identificed predictors not available or not collected for an extended time (e.g. pre-transplant sodium), and data quality. Transplant registry data is an evolving organism. What flexibility is available in an organ allocation scheme is influenced by current research, as is seen in the examination of HCV status. The decision making process in organ allocation is influenced by other considerations that may not be collected. The final

conclusion might be that it is simply too difficult to predict a result 3 years in the future from baseline information - there are too many important variables occurring after the time of transplant such as immunosuppressive regime, infection, rejection, and technical issues, which will influence outcome. Christensen (2004) suggests that currently used prognostic variables are not sufficiently informative and he looks to advances in molecular biology for variables giving more information about a disease process. An example offered is the interleukin-10 GG genotype in Hepatitis C patients that is associated with persistent infection (Knapp, 2003). Scientific advances such as this can be combined with properly applied statistical models where assumptions are satisfied in order to improve prediction error. Coordination of registry standards and increased data validation by the large transplant registries worldwide would be of great benefit to research in the field.

7 Conclusion

With this paper we have shown that the non-proportionality found in some covariates used to assess survival after liver transplant make the Cox PH regression model a sub-optimal choice. We found that the predictive accuracy gained through use of the Cox PH model is limited compared to a null model without covariate information. In the hope of finding a better model fit, we tested simpler models generated by survival trees but found they have higher prediction error than models developed using Cox PH regression. The nature of the data requires a complex statistical model, and the search must continue for one that is appropriate. It is critical that new prognostic models undergo a valid assessment of model adequacy. Here we have established a performance benchmark that future prognostic models may be measured against, using the integrated Brier score as a valid assessment of model performance.

8 References

Altman, D.G., Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19,* 453-473.

Altman, D.G., Vergouwe, Y., Royston, P., Moons, K.G.M. (2009). Prognosis and prognostic research: validating a prognostic model. *British Medical Journal, 338,* b605.

Ambler, G., Royston, P. (2001). Fractional polynomial model selection procedures: investigation of Type I error rate. *Journal of Statistical Simulation and Computation, 69,* 89-108.

Ambler, G., Benner, A. (2008) Multivariable Fractional Polynomials. R package version 1.4.6. http://CRAN.R-project.org/package=mfp.

Asrani, S.K., Lim, Y.S., Therneau, T.M., Pedersen, R.A., Heimbach, J., Kim, W.R. (2010). Donor race does not predict graft failure after liver transplant. *Gastroenterology 138(7),* 2341-2347.

Austin, P.C., Brunner, L.J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine, 23,* 1159-1178.

Benner, A. (2005). mfp: Multivariable fractional polynomials. *R News, 5(2),* 2023.

Box, G.E., Draper, N.R. (1986). Empirical Model-Building and Response Surfaces. New York: John Wiley and Sons Inc.

Breiman, L., Freidman, J., Stone, C., Olshen, R. (1984). Classification and Regression Trees. Belmpnt, CA: Wadsworth.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1-3.

Christensen, E., (2004). Prognostic models including the Child-Pugh, MELD and Mayo risk scores - where are we and where should we go? *Journal of Hepatology, 41,* 344-350.

Conover, W.J. (1999). Practical Nonparametric Statistics, 3rd ed. NJ: Wiley.

Cox, D.R. (1972). Regression Models and life tables. *Journal of the Royal Statistical Society, Ser. B, 34,* 187-220.

Feng, S., Goodrich, N.P., Bragg-Gresham, J.L., Dykstra, D.M., Punch, J.D., DebRoy, M.A. (2006). Characteristics associated with liver graft failure: The concept of a donor risk in-

dex. *American Journal of Transplantation, 6,* 783790.

Forster, L.A. Lynn, J. (1988). Predicting life span for applicants to inpatient hospice. *Archives of Internal Medicine, 148,* 2540-2543.

Gerds, A., Schumacher, M. (2006). Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal, 48,* 1029-1040.

Ghobrial, R.M., Gornbein, J., Steadman, R., Danino, N., Markmann, J.F., Holt, C., Anselmo, D., Amersi, F., Chen, P., Farmer, D.G., Han, S., Derazo, F., Saab, S., Goldstein, L.I., McDiarmid, S.V., Busuttil, R.W. (2002). Pretransplant model to predict posttransplant survival in liver transplant patients. *Annals of Surgery, 236(3),* 315-22.

Gonen, M., Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika, 92,* 965970.

Gordon, L., Olshen, R. (1985). Tree-Structured Survival Analysis. *Cancer Treatment Reports, 69,* 1065-1069.

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M. (1999). Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Statistics in Medicine, 18,* 2529-2545.

Grambsch, P.M., Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika Trust, 81,* 515-526.

Habib, S., Berk, B., Chung-Chou, H.C., Demetris, A.J., Fontes, P., Dvorchik, I., Eghstesad, B., Marcos, A., Shakil, O. (2006). MELD and Prediction of Post-Liver Transplantation Survival. *Liver Transplantation, 12,* 440-447.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics, 24* 22002208.

Haibe-Kains, B., Sotiriou, C., Bontempi, G. (2009) Performance Assessment and Comparison for Survival Analysis. R package version 1.1.3.
http://CRAN.R-project.org/package=survcomp.

Hand DJ. (1997) Construction and Assessment of Classification Rules. NY: John Wiley & Sons.

Hari, P.N., Zhang, M.J., Roy, V., Prez, W.S., Bashey, A. et al. (2009). Is the International Staging System superior to the Durie-Salmon staging system? A comparison in multiple myeloma patients undergoing autologous transplant. *Leukemia, 23(8),* 1528-34.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247,* 2543-6.

Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A. (1983). Regression modeling strategies for improved prognostics. *Statistics in Medicine, 3,* 143-152.

Harrell, F.E., Lee, K., Mark, D. (1996). Tutorials in Biostatistics. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine, 15,* 361-387.

Harrell, F.E. (2001). Regression Modeling Strategies. NY: Springer.

Henderson, R., Jones, M. (1995). Prediction in survival analysis: model or medic? in Jewell NP, Kimber AC, Ting-Lee ML and Withmore GA (eds) Lifetime Data: Models in Reliability and Survival Analysis. Dordecht: Kluwer Academic Publishers.

Henderson, R., Jones, M., Stare, J. (2001). Accuracy of point predictions in survival analysis. *Statistics in Medicine, 20,* 3083-96.

Hothorn, T., Lausen, B., Benner, A., Radespiel-Troger, M. (2004). Bagging Survival Trees. *Statistics in Medicine, 23,* 77-91.

Ikeda, M., Itoh, S., Ishigaki, T., Yamauchi, K. (2001). Application of Resampling Techniques to the Statistical Analysis of the Brier Score. *Methods of Information in Medicine, 40,* 259-64.

Kleinbaum, D.G., Klein, M. (2005). Survival Analysis: A Self-Learning Text. NY: Springer.

Knapp, S., Hennig, B.J., Frodsham, A.J. (2003). Interleukin-10 promoter polymorphisms and the outcome of Hepatitis C virus infection. *Immunogenetics, 55,* 362-369.

Kronek, L-P., Reddy, A. (2009). Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics, 24,* i248-i253.

LeBlanc, M., Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics, 48,* 411-425.

Lin, D.Y., Feuer, E.J., Etzioni, R., Wax, Y. (1997). Estimating Medical Costs from Incomplete Follow-Up Data. *Biometrics, 53*, 419-434.

Lonardo, A., Loria, P. (2008). The hepatitis C virus-associated dysmetabolic syndrome. *Hepatology, 48,* 10181019.

May, M., Royston, P., Egger, M., Justice, A.C., Sterne, J. (2004). Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy. *Statistics in Medicine, 23,* 2375-2398.

Merion, R.M., Schaubel, D.E., Dykstra, D.M., Freeman, R.B., Port, F.K., Wolfe, R.A. (2005). The survival benefit of liver transplantation. *American Journal of Transplantation, 5,* 307-13.

Moons, K.G.M., Altman, D.G., Vergouwe, Y., Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal, 338,* b606.

Nair, S., Verma, S., Thuluvath, P.J. (2002). Obesity and its effect on survival in patients undergoing orthotopic liver transplantation in the United States. *Hepatology, 35,* 105-109.

Orbe, J., Ferreira, E., Nunez-Anton, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in Medicine, 21,* 3493-3510.

Parkes, M.C. (1972). Accuracy of predictions of survival in later stages of cancer. *British Medical Journal, 264,* 29-31.

Pelletier, S.J., Schaubel, D.E., Wei, G., et al. (2007). Effect of Body Mass Index on the Survival Benefit of Liver Transplantation. *Liver Transplantation, 13,* 1678-1683.

Peters, A., Hothorn, T., Lausen, B. (2002). ipred: improved predictors. *R news, 2,* 33-36.

Peters, A., Hothorn, T. (2009) Improved Predictors. R package version 0.8-8. http://CRAN.R-project.org/package=ipred.

Porepa, L., Ray, J.G., Sanchez-Romeu, P., Booth, G.L. (2010) Newly diagnosed diabetes mellitus as a risk factor for serious liver disease. *Canadian Medical Association Journal,* published online ahead of print, June 21, 2010.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rana, A., Hardy, M.A., Halazun, K.J., Woodland, D.C., Ratner, L.E., Samstein, B., Guarrera, J.V., Brown Jr, R.S., Emond, J.C. (2008). Survival Outcomes Following Liver Transplantation (SOFT) Score: A Novel Method to Predict Patient Survival Following Liver Transplantation. *American Journal of Transplantation, 8,* 25372546.

Ravaioli, M., Grazi, G.L., Dazzi, A., Bertuzzo, V., Ercolani, G., Cescon, M., Cucchetti, A., Masetti, M., Ramacciato, G., Pinna, A.D. (2009). Survival benefit after liver transplan-

tation: a single European center experience. *Transplantation, 88(6),* 826-34.

Ricci, P., Therneau, T.M., Malinchoc, M., Benson, J.T., Petz, J.L., Klintmalm, G.B., Crippin, J.S., Wiesner, R.H., Steers, J.L., Rakela, J., Starzl, T.E., Dickson, E.R. (1997). A prognostic model for the outcome of liver transplantation in patients with cholestatic liver disease. *Hepatology, 25,* 672-7.

Rosen, H.R. (2000). Disease recurrence following liver transplantation. *Clin Liver Dis, 4,* 675-689.

Rosen, H.R., Prieto, M., et al. (2003). Validation and refinement of survival models for liver retransplantation. *Hepatology, 38,* 460.

Royston, P., Sauerbrei, W. (2003). Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomial for modelling continuous covariates. NJ: John Wiley and Sons.

Royston, P., Altman, D.G., Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine, 25,* 127-141.

Sauerbrei, W., Meier-Hirmer, C., Benner, A., Royston, P. (2006). Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics & Data Analysis, 50,* 3464-3485.

Sauerbrei W, Royston P, Zapien K. (2007). Detecting and interaction between treatment and a continuous covariate: A comparison of two approaches. *Computational Statistics and Data Analysis, 51,* 4054-4063.

Schaubel, D.E., Wei, G. (2007) Fitting semiparametric additive hazards models using standard statistical software. *Biometrical Journal, 49,* 719-730.

Schaubel, D.E., Guidinger, M.K., Biggins, S.W., Kalbfleisch, J.D., Pomfret, E.A., Sharma, P., Merion, R.M. (2009). Survival benefit-based deceased-donor liver allocation. *American Journal of Transplantation, 9(Part 2),* 970-981.

Schemper, M., Smith, T.L. (1996), A note on quantifying follow-up in studies of failure time. *Control Clin Trials, 17,* 34346.

Schemper, M. (2003). Predictive accuracy and explained variation. *Statistics in Medicine, 22,* 2299-2308.

Schumacher, M., Graf, E., Gerds, T. (2003). How to Assess Prognostic Models for Survival Data: A Case Study in Oncology. *Methods of Information in Medicine, 5,* 564-571.

Stute, W. (1993). Consistent estimation under random censorship when covariables are

present. *Journal of Multivariate Analysis, 45,* 89-103.

Therneau, T.M., Atkinson, E.J. (1997). An Introduction to recursive partitioning using the RPART routines. Technical report, Mayo Clinic Section of Biostatistics.

Therneau, T.M., Grambsch, P. (2000). Modelling Survival Data: Extending the Cox Model. NY: Springer.

Therneau, T., Atkinson, E. (2010). Recursive Partitioning. R package version 3.1-46. http://CRAN.R-project.org/package=rpart.

van Wieringen, W.N., Kun, D., Hampel, R., Boulesteix, A.L. (2009). Survival prediction using gene expression data: a review and comparison. *Computational Statistics and Data Analysis, 53,* 1590-1603.

Venables, W.N., Ripley, B.D., (2002) Modern Applied Statistics with S. NY: Springer.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., McCulloch, C.E. (2005). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. NY: Springer.

Wainer, H. (2006). Finding what is not there through the unfortunate binning of results: The Mendel effect. *Chance, 19(1),* 49-56.

Watt, K.D., Pedersen, R.A., Kremers, W.K., Heimbach, J.K., Charlton, M.R. (2010). Evolution of causes and risk factors for mortality post-liver transplant: results of the NIDDK long-term follow-up study. *American Journal of Transplantation, 10(6),* 1420-1427.

Weismuller, T.J., Prokein, J., Becker, T., Barg-Hock, H., Klempnauer, J., Manns, M.P., Strassburg, C.P. (2008). Prediction of survival after liver transplantation by pre-transplant parameters. *Scand J Gastroenterol., 43(6),* 736-46.

Wyatt, J.C., Altman, D.G. (1995). Prognostic models: clinically useful or quickly forgotten? *British Medical Journal, 311,* 1539-1541.

Zeng, D., Lin, D.Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist. Soc. B, 69* 507-564.

A Appendix

Table 7: Coding for donor cause of death.

Registry term	Cardiovascular disease	Anoxia	Trauma	Other
Death from natural causes				X
Drowning		X		
Intracranial hemorrhage/stroke	X			
Seizure		X		
Drug intoxication				X
Asphyxiation		X		
Cardiovascular	X			
None of the above				X
Gun shot wound			X	
Stab			X	
Blunt injury			X	
Electrical				X

Table 8: Coding for donor race.

Registry term	Asian	Black	Hispanic	White	Other
Native Hawaiian or Other Pacific Islander					X
Black or African American		X			
Hispanic/Latino			X		
American Indian or Alaska Native					X
Asian	X				
White				X	
Multi-Racial					X

Table 9: Coding for recipient diagnosis categories.

Registry term	AHN	Alcohol	BA	Chole.	HCC	HCV	Malig.	Metabol.	Non-Chole.	Other	Viral
AHN: DRUG OTHER SPECIFY	X										
AHN: TYPE A	X										X
AHN: TYPE B- HBSAG+	X										X
AHN: TYPE C	X					X					X
AHN: TYPE D	X										X
AHN: TYPE B AND C	X					X					X
AHN: TYPE B AND D	X										X
AHN: ETIOLOGY UNKNOWN	X										
AHN: OTHER, SPECIFY	X										
CIRRHOSIS: DRUG/INDUST SPECIFY									X		
CIRRHOSIS: TYPE A									X		X
CIRRHOSIS: TYPE B- HBSAG+									X		X
CIRRHOSIS: TYPE C						X			X		X
CIRRHOSIS: TYPE D									X		X
CIRRHOSIS: TYPE B AND C						X			X		X
CIRRHOSIS: TYPE B AND D									X		X
CIRRHOSIS: CRYPTOGENIC									X		
CIRRHOSIS: CHRONIC ACTIVE HEPATITIS: UNK									X		
CIRRHOSIS: OTHER, SPECIFY									X		
CIRRHOSIS: AUTOIMMUNE									X		
CIRRHOSIS: CRYPTOGENIC									X		
CIRRHOSIS: FATTY LIVER (NASH)									X		
ALCOHOLIC CIRRHOSIS		X							X		
ALCOHOLIC CIRR WITH HCV		X				X			X		X
ACUTE ALCOHOLIC HEPATITIS	X	X									
PRIMARY BILIARY CIRR (PBC)				X							
SEC BILIARY CIRR: CAROLI'S DISEASE				X							
SEC BILIARY CIRR: CHOLEDOCHOL CYST				X							
SEC BILIARY CIRR: OTHER SP				X							
PSC: CROHN'S DISEASE				X							
PSC: ULCERATIVE COLITIS				X							
PSC: NO BOWEL DISEASE				X							
PSC: OTHER SPECIFY				X							

Table 10: Coding for recipient diagnosis categories (continued).

Registry term	AHN	Alcohol	BA	Chole.	HCC	HCV	Malig.	Metabol.	Non-Chole.	Other	Viral
FAMILIAL CHOLESTASIS: BYLER'S DISEASE				X							
FAMILIAL CHOLESTASIS: OTHER SPECIFY				X							
CHOLES LIVER DISEASE: OTHER SPECIFY				X							
NEONATAL CHOLESTATIC LIVER DISEASE				X							
NEONATAL HEPATITIS OTHER SPECIFY	X										
BILIARY ATRESIA: EXTRAHEPATIC			X								
BILIARY HYPOPLASIA: NONSYNDROMIC PAUCITY IBD			X								
BILIARY HYPOPLASIA: ALAGILLES SYNDROME			X								
BILIARY ATRESIA OR HYPOPLASIA: OTHER, SPECIFY			X								
CONGENITAL HEPATIC FIBROSIS										X	
CYSTIC FIBROSIS										X	
BUDD-CHIARI SYNDROME										X	
METDIS: ALPHA-1-ANTITRYPSIN DEFIC A-1-A								X			
METDIS: WILSON'S DISEASE, OTHER COPPER								X			
METDIS: HEMOCHROMATOSIS - HEMOSIDEROSIS								X			
METDIS: GLYC STOR DIS TYPE I (GSD-I)								X			
METDIS: GLYC STOR DIS TYPE II (GSD-IV)								X			
METDIS: HYPERLIPIDEMIA-II, HOMOZYGOUS HYPERCHOL.								X			
METDIS: TYROSINEMIA								X			
METDIS: PRIMARY OXALOSIS/OXALURIA, HYPEROXALURIA								X			
METDIS: MAPLE SYRUP URINE DISEASE								X			
METDIS: OTHER SPECIFY								X			
PLM: HEPATOMA - HEPATOCELLULAR CARCINOMA		X		X							
PLM: HEPATOMA (HCC) AND CIRRHOSIS		X		X							
PLM: FIBROLAMELLAR (FL-HC)		X		X							
PLM: CHOLANGIOCARCINOMA (CH-CA)				X							
PLM: HEPATOBLASTOMA (HBL)				X							
PLM: HEMANGIOENDOTHELIOMA, HEMANGIOSARCOMA, ANGIOSARCOMA				X							
PLM: OTHER SPECIFY (I.E., KLATZKIN TUMOR, LEIOMYSARCOMA)				X							
BILE DUCT CANCER: (CHOLANGIOMA, BILIARY TRACT CARCINOMA)				X							
SECONDARY HEPATIC MALIGNANCY OTHER SPECIFY				X							
BENIGN TUMOR: HEPATIC ADENOMA										X	
BENIGN TUMOR: POLYCYSTIC LIVER DISEASE										X	
BENIGN TUMOR: OTHER SPECIFY										X	
TPN/HYPERALIMENTATION IND LIVER DISEASE										X	
GRAFT VS. HOST DIS SEC TO NON-LI TX										X	
TRAUMA OTHER SPECIFY										X	
Hepatitis B: Chronic or Acute											X
Hepatitis C: Chronic or Acute					X						X
OTHER SPECIFY Other										X	

Table 11: Results of Cox Model 1 (using cutpoints) on the training data - including DIABETIC status.

Variable	β	$exp(\beta)$	p-value
Age at transplant (yrs)	0.0052	1.0053	0.2013
Age > 55 yrs	-0.01779	0.9824	0.8086
Diagnosis: non-cholestatic cirrhosis	-0.1279	0.8800	0.0373
Diagnosis: cholestatic cirrhosis	-0.2978	0.7425	0.0023
Diagnosis: metabolic liver disease	-0.3894	0.6775	0.0168
Diagnosis: hepatocellular carcinoma (HCC)	0.1619	1.1757	0.0080
Diagnosis: hepatitis C virus (HCV)	0.2497	1.2837	< 0.0001
Recipient medical condition: in ICU	0.3102	1.3638	0.0003
Recipient medical condition: hospitalized not in ICU	0.2096	1.2331	0.0008
Recipient on life support	0.5820	1.7896	< 0.0001
Recipient prior portal vein thrombosis	0.2125	1.2367	0.0292
Recipient prior abdominal surgery	0.1704	1.1858	< 0.0001
Creatinine: 4th quartile	0.2122	1.2729	< 0.0001
Albumin: 1st quartile	0.2267	1.2544	< 0.0001
Donor age (yrs): 40 to 49	0.2413	1.2642	< 0.0001
Donor age (yrs): 50 to 59	0.3273	1.3872	< 0.0001
Donor age (yrs) > 60	0.3380	1.4021	< 0.0001
Donor race: hispanic	0.2278	1.2558	< 0.0001
Cold ischemic time (hours)	0.0151	1.0152	0.0041
Age at transplant * age > 55 yrs	0.0260	1.0263	0.0005
Recipient HCV status * Donor age (yrs) > 60 years	0.4759	1.6095	< 0.0001

Table 12: Results of Cox Model 2 (using fractional polynomials) on the training data - including DIABETIC status.

Variable	β	$exp(\beta)$	p-value
$(Recipient \cdot age(yrs) \cdot at \cdot transplant/100)^3$	$2.181 * 10^{-6}$	1.0000	< 0.0001
Donor age (yrs) * recipient HCV positive	0.01764	1.018	< 0.0001
Albumin (g/dL)	-0.1592	0.8528	< 0.0001
Recipient medical condition: in ICU	0.3223	1.380	0.0001
Recipient medical condition: hospitalized not in ICU	0.2202	1.246	0.0004
Recipient on life support	0.5811	1.788	< 0.0001
Recipient prior abdominal surgery	0.1714	1.1870	< 0.0001
Donor age (yrs) * recipient HCV negative	0.0070	1.007	< 0.0001
Cold ischemic time (hours)	0.0147	1.0150	0.0054
Donor race: Hispanic	0.2343	1.2640	< 0.0001
Diagnosis: Cholestatic cirrhosis	-0.3045	0.7375	0.0019
Diagnosis: Non-cholestatis cirrhosis	-0.1391	0.8702	0.0225
Creatinine (mg/dL)	0.0735	1.0760	< 0.0001
Diabetic status	0.2226	1.249	< 0.0001
Diagnosis: metabolic liver disease	-0.3928	0.6752	0.0158
Diagnosis: hepatocellular carcinoma	0.1768	1.1930	0.0038
Recipient prior portal vein thrombosis	0.2013	1.2230	0.03860
Diagnosis: Hepatitis C virus	-0.1492	0.8614	0.2108