Published in final edited form as:

Stat Methods Med Res. 2017 October ; 26(5): 2455–2480. doi:10.1177/0962280216688034.

Flexible modeling of the hazard rate and treatment effects in long-term survival studies

Yolanda Hagar¹, James J Dignam², and Vanja Dukic¹

¹Department of Applied Mathematics, University of Colorado, CO, USA

²Department of Public Health Sciences, University of Chicago, IL, USA

Abstract

The effects of predictors on time to failure may be difficult to assess in cancer studies with longer follow-up, as the commonly used assumption of proportionality of hazards holding over an extended period is often questionable. Motivated by a long-term prostate cancer clinical trial, we contrast and compare four powerful methods for estimation of the hazard rate. These four methods allow for varying degrees of smoothness as well as covariates with effects that vary over time. We pay particular attention to an extended multiresolution hazard estimator, which is a flexible, semi-parametric, Bayesian method for joint estimation of predictor effects and the hazard rate. We compare the results of the extended multiresolution hazard model to three other commonly used, comparable models: Aalen's additive model, Kooperberg's hazard regression model, and an extended Cox model. Through simulations and the analysis of a large-scale randomized prostate cancer clinical trial, we use the different methods to examine patterns of biochemical failure and to estimate the time-varying effects of androgen deprivation therapy treatment and other covariates.

Keywords

Biochemical failure; hazard rate; multiresolution hazard; non-proportional hazards; prostate cancer; survival analysis

1 Introduction

Many human cancers today are considered “chronic diseases,” with long-term disease trajectories and multiple co-morbidities. Consequently, long-term cancer outcomes may be affected by numerous factors, ranging from obvious patient and treatment characteristics to secular and health-care trends that affect treatment policy and practice. However, the long-term nature of patient and health-care-related processes and changing complexity of information can make the analysis of long-term patterns in cancer survivor data sets challenging. In addition to sparsely observed failure times, these data often exhibit non-

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Vanja Dukic, Department of Applied Mathematics, University of Colorado, Boulder, ECOT 225, 526 UCB, Boulder, CO 80309, USA. vanja.dukic@colorado.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

proportional effects over time, requiring flexible and computationally efficient statistical methods to characterize the evolving failure hazard.

The motivating problem in this article comes from a series of prostate cancer clinical trials from the Radiation Therapy Oncology Group (RTOG), which were designed to examine the effects of radiation and androgen deprivation (AD) therapy on outcomes in men with localized (i.e. non-metastatic) disease. In this article, we are particularly interested in quantifying the effects of AD therapy on “biochemical failure,” which is a measure frequently used to determine if prostate cancer has returned after treatment. It is known that short-term AD therapy can delay the time until prostate cancer recurrence (and death),^{1,2} and longer duration AD therapy has been proven more beneficial than a shorter duration therapy.³ However, given that AD treatments can have unpleasant side effects, clinicians have been reluctant to assign AD for longer than necessary. Additionally, because prostate cancer is generally slow to progress,⁴ men with early-stage prostate cancer tend to have low mortality rates and are thus observed over extended periods of time, making long-term treatment benefits difficult to precisely ascertain due to a multitude of co-morbidities and sparsity of information at long follow-up times.

The main goal of our analysis is to infer if and how a longer term AD therapy impacts the time to biochemical failure and how the effects of longer term treatment may change over the course of a 10+ year post-diagnosis follow-up period. Several previous studies have used time-aggregated summaries (i.e. survival curves, cumulative incidence) to estimate cumulative biochemical failure risk over time.^{5,6} However, the focus of our paper goes beyond integrated summaries and we instead concentrate on the estimation and inference of the hazard rate of biochemical failure. Analysis of the hazard rate function will allow us to examine, in detail, if the risk of biochemical failure for the long-term treatment subjects increases later in follow-up, indicating that failures in this group were deferred in time rather than avoided. Additionally, because of the long-term follow-up of the patients in this data set, we want to allow the effects of treatment to change over time. While this increases the complexity of our analysis, it is a necessary requirement for quantifying how treatment affects the risk of biochemical failure over longer periods of time.

Previous analyses specifically examining the hazard rate of biochemical failure and the effects of treatment are relatively scarce, and have been limited to the clinical literature, using intuitive summaries to approximate the annual hazard.^{7–10} These analyses provide basic, useful information on the hazard rate of biochemical failure over time, but do not provide joint estimation of the effects of multiple covariates on the hazard rates, and do not perform well when the number of observed failures is small.

An initial examination of the RTOG biochemical failure data show sparsely observed biochemical failures in later follow-up years, and evidence of a treatment effect that changes over time. Thus, we require a model that can jointly produce estimates of the hazard rate of biochemical failure and the effects of covariates (including effects that change over time), accommodate periods with sparsely observed failures, and allow for uncertainty propagation for a variety of parameters.

In this article, we highlight the Bayesian non-proportional multiresolution hazards model (NPMRH), which is an extension of the existing multiresolution hazard (MRH) model.^{11–17} This model has three important features: (1) quantification of the hazard rate, (2) estimated covariates effects that can be constant or changing over time, and (3) robust estimation of the model parameters over periods of sparsely observed failures through the prior and a pruning tool. In addition, we demonstrate that the Bayesian implementation of the model allows for flexible and easy inference on a number of parameters we are interested in.

In addition to the NPMRH model, we present three other commonly used models that provide estimates of the hazard rate, allow for covariate effects that change over time, and that have available software in CRAN-R¹⁸ for implementation: Aalen's additive model,^{19,20} Kooperberg's hazard regression model²¹ (HARE), and a time-varying Cox model.²² We compare these four models through simulations and analysis of the RTOG biochemical failure data and show that while each model can produce estimates of the quantities we are interested in, the results vary depending on the implementation. While each of the four models we compare in this paper has been investigated previously, here we demonstrate how the four models differ in theory, in estimation, and in results. This provides valuable insight for the biostatistics and medical research community and can help guide researchers in determining which models might work best under specific scenarios.

This paper is organized as follows: in the next section, we provide a short overview of the prostate cancer clinical trials data and the statistical issues related to these studies with long-term follow-up. Section 3 presents the methodologies for the different survival models, with a focus on the development of the NPMRH model. Section 4 presents results from a set of simulations with non-proportional treatment hazard rates, and section 5 contrasts the model estimates when analyzing the RTOG biochemical failure data set. This article concludes with a discussion of the clinical and statistical importance and implications of our findings. All MRH models presented have been fitted using the MRH R package.²³

2 AD in prostate cancer

A typical prostate treatment involves surgery or radiation therapy, often combined with some form and duration of hormone treatment, which is known as AD therapy. These treatment interventions, when effective, greatly decrease prostate-specific antigen (PSA) for a period during and after treatment. PSA is a biomarker routinely measured to screen for possible presence of prostate cancer. "Biochemical failure" is defined as a two-unit rise in PSA level following a post-treatment PSA nadir²⁴ and is a measure frequently used to determine if prostate cancer has returned after treatment.^{25–27} Characterization of the biochemical failure hazard over time and the effects of treatment regimens would provide a strong foundation for determining recurrence patterns, and could be of great value for clinical management, design of clinical trials, and biologic insights into prostate cancer progression in different population subgroups.

The data we use to analyze the biochemical failure hazard rate come from the RTOG. The trial we examine in this paper is RTOG 92-02, one of a series of RTOG clinical trials conducted from the 1980s to the present to investigate the treatment for prostate cancer.

Between 1992 and 1995, 1521 eligible participants with locally advanced high-risk prostate cancer were accrued in over 200 treatment centers across North America. All patients were randomized to either receive four months of AD therapy (the +0m AD group), or four months of AD therapy plus an additional 24 months of therapy (the +24m AD group). Further protocol details and study description can also be found in Horwitz et al.³ and the RTOG website.²⁸

For each patient enrolled, several features of the original cancer were recorded at baseline: the Gleason score, T-stage of the tumor, and the PSA level at diagnosis. All three measures are designed to assess the severity of the cancer at diagnosis, with higher scores indicating a possibly more aggressive cancer.^{29–31} In addition to the age at diagnosis, these are considered to be important predictors in the analysis of biochemical failure.

The final data set considered in this analysis comprises 1421 subjects, after the removal of 100 subjects with missing Gleason scores. Of those 1421 subjects, the sample median time to observed biochemical failure or censoring was 4.9 years (2.8 years for observed failure times and 8.5 years for censored times, SD=3.9, range = 0.03–13.65). Biochemical failure was observed for 50.4% of the patients before the end of the study period; however, only 12% of the population had an observed failure occurred after five years, which is an average of less than 2% observed failures per year between five years and the end of the study. Table 1 summarizes the sample characteristics in more detail.

Evidence of a changing long-term treatment effect was provided by examination of the Schoenfeld residuals and methods presented in Grambsch and Therneau³² ($p < 0.01$, Figure 1). The p -value indicates that the null hypothesis that the treatment effect is constant over time should be rejected. Additionally, the examination of the residuals in Figure 1 shows a non-linear weighted least squares line (deviation from linearity indicates that the covariate effect is not constant over time), providing more evidence that the effect is non-proportional. No other covariate effects showed evidence of non-proportionality over time.

3 Non-proportional hazards models

The hazard function for time to failure T is defined as

$$h(t) = \lim_{\Delta \rightarrow 0} P(t \leq T < t + \Delta | T \geq t) / \Delta = f(t) / S(t), \text{ where } S(t) = P(T > t) \text{ is the survival}$$

function and $f(t) = -S'(t)$ is the probability density function of T . While the hazard rate can provide a more detailed pattern over time that is not always visible in aggregate measures such as the survival curve or cumulative hazard function, it may also be more difficult to estimate reliably. This is particularly true if event counts are sparse, as is often the case in studies that follow subjects for extended periods of time.

Various parametric and non-parametric estimators have been developed for the hazard function, some allowing for the inclusion of covariates under the proportionality assumption.^{33–35} However, with longer term follow-up, the assumption of constant relative hazard rates between different patient subgroups defined by covariates may be questionable, so this assumption must be relaxed over time, leading to “non-proportional hazards” models. Initial investigations involved the use of piece-wise constant hazard functions, examples of

which can be found in Holford,^{36,37} Laird and Olivier³⁸ and Taulbee.³⁹ Other methods address non-proportionality issues by pretesting and comparison of two survival or hazard functions through graphics and asymptotic confidence bands,^{40–42} or through asymptotic confidence bands for changes in the predictor effects over time.^{43,44} Some of these methods mentioned require large sample sizes for inference, and their performance can degrade over time in studies with sparser outcomes in later periods. Alternative approaches to handling non-proportionality have been implemented through accelerated failure time (AFT) models (with initial work done by Buckley and James⁴⁵ and a thorough review found in Kalbfleisch and Prentice⁴⁶), which accommodate non-proportionality through specific parameterization of the time to failure and the covariates. As a result, these models can only accommodate certain types of non-proportionality, such as non-proportional hazards that do not cross.

In the Bayesian context, non-proportionality has been addressed in a number of ways, including through the extension of frequentist models,^{47,48} using Bayesian splines,^{49,50} or non-parametrically with dependent Dirichlet Process (DDP) priors.⁵¹ Another Bayesian approach includes the use of Polya tree priors.^{52,53} The Polya tree prior is a recursive, dyadic partitioning of a measurable space Ω . They have been adapted for modeling survival data in a number of ways,^{54–57} including the use of stratified Polya tree priors to address non-proportionality,⁵⁸ by combining “bins” of the Polya tree prior through randomization techniques,⁵⁹ or smoothing the prior tree by allowing the tree branches to be dependent on a latent binomial random variable.⁶⁰ While these models exhibit many of the features we are interested in (estimation of the hazard rate, inclusion of covariates under both the proportional and non-proportional hazards assumption, and smoothing the hazard rate through parameter reduction or prior parameters), these models do not simultaneously incorporate all of these components.

3.1 General notation

In the context of the RTOG clinical trial, each patient is observed for biochemical failure from study entry. The biochemical failure time, t_{fail} , is considered right censored if the event is not observed before the end of the trial observation period (which occurs at time t_J) or before removal from the study (e.g. death from other causes), which is denoted as t_{cens} . We let T_i represent the time from which observation begins to the minimum of t_{fail} , t_{cens} , or t_J for subject i . The hazard rate is then denoted as $h(t|X, Z, \beta, \alpha)$, and the cumulative hazard as $H(t|X, Z, \beta, \alpha) = \int_0^t h(s|X, Z, \beta, \alpha) ds$, where X represents the covariate matrix for the proportional hazards, $\beta = (\beta_1, \dots, \beta_{z_1})$ denotes the $z_1 \times 1$ vector of the time-invariant covariate effects, Z represents the covariate matrix with effects that change over time, and $\alpha = (\alpha_1, \dots, \alpha_{z_2})$ denotes the $z_2 \times 1$ vector of the changing covariate effects. (Depending on the model, each, α_k , $k = 1, \dots, z_2$) may be a single number or vector of numbers.) The survival analysis log-likelihood function is typically expressed as

$$\log ; L(T|\beta, \alpha, X, Z) = \prod_{i=1}^n (\delta_i \log(h(t_i|X_i, Z_i, \beta, \alpha)) - H(t_i|X_i, Z_i, \beta, \alpha)) \quad (1)$$

where δ_i is a censoring indicator that denotes if biochemical failure was observed for subject i .

3.2 The MRH model

The MRH model is a Bayesian, semi-parametric model with a prior that is a type of Polya tree: it uses a fixed, pre-specified partition, and controls the hazard level within each bin through a multiresolution parameterization. The NPRMH model is an extended MRH model that combines the hierarchical multiresolution hazard (HMRH) model presented in Dukic and Dignam¹⁴ with the pruning methodology presented in Chen et al.¹⁶ The HMRH model is capable of modeling non-proportional hazard rates in different subgroups jointly with other proportional predictor effects. The pruning methodology¹⁶ detects consecutive time intervals where failure patterns are statistically similar, increasing estimator efficiency and reducing computing time. Below, we outline the combined NPRMH model with the pruning method and give brief details on the priors, with more information available in Appendix 1.

3.2.1 MRH prior—The MRH method uses a piece-wise constant approximation of the hazard function over J time intervals, parametrized by a set of hazard increments $d_j, j = 1, \dots, J$ (i.e. $d_j = \int_{t_{j-1}}^{t_j} h(s) ds \equiv H(t_j) - H(t_{j-1})$). However, the Polya tree prior of the MRH model allows the model to borrow information across the study period: We assume that $J = 2^M$ (because the tree is dyadically partitioned). The cumulative hazard over the entire study period, H , is equal to the sum of all 2^M hazard increments $d_j, j = 1, \dots, 2^M$. The model then recursively splits H at different branches via the “split parameters” $R_{m,p} = H_{m,2^p} / H_{m-1,p}$, $m = 1, 2, \dots, M-1, p = 0, \dots, 2^{m-1} - 1$. Here, $H_{m,q}$ is recursively defined as $H_{m,q} \equiv H_{m+1,2q} + H_{m+1,2q+1}$ (with $H_{0,0} \equiv H$, and $q = 0, \dots, 2^m - 1$). The $R_{m,p}$ split parameters, each between 0 and 1, guide the shape of the a priori hazard rate over time. Parameterization of the model in this way allows for the hazard increments (i.e. the d_j) to be expressed as a function of the total cumulative hazard H and the split parameters that guide the prior tree. The complete hazard rate prior specification is obtained via priors placed on all tree parameters: a Gamma(a, λ) prior is placed on the cumulative hazard H , and Beta prior on each split parameter $R_{m,p}$ of the form $\mathcal{B}e(2\gamma_{m,p}k^m a, 2(1 - \gamma_{m,p})k^m a)$.

3.2.2 Smoothing through pruning—The MRH model can also perform “pruning,” which merges adjacent bins constructed via the same split parameter, $R_{m,p}$, when the hazard increments in these two bins are statistically similar. This is inferred by testing the hypothesis $H_0: R_{m,p} = 0.5$ against the alternative $H_a: R_{m,p} \neq 0.5$ for each split parameter $R_{m,p}$ ($p = 0, \dots, 2^{m-1} - 1$). If the null hypothesis is not rejected, that split $R_{m,p}$ is set to 0.5 and the adjacent hazard increments are considered equal and the time bins declared “fused.” The hypothesis testing can be applied to all M levels of the tree or just a higher resolution subset of the tree. Pruning can increase the computational efficiency by decreasing the parameter dimension a priori, which can greatly speed up analyses of non-proportional hazards.

3.2.3 Smoothing through adjustment of hyperparameter k —The parameter k in the prior for $R_{m,p}$ controls the correlation among the hazard increments within each bin¹¹ (see Appendix 1 for details). The default value for k in most MRH analyses is 0.5, which

implies zero a priori correlation among the hazard increments. However, when $k > 0.5$, the increments are positively correlated a priori, and, similarly, when $k < 0.5$, the hazard increments are negatively correlated a priori. Another way to understand the impact of k is that higher values lead to smoother hazard functions. In practice, different approaches to choosing a hyperprior for k , including empirical Bayes methods, are possible. However, k will in general tend to depend on the resolution level,¹¹ as well as with the significance level used in the pruning algorithm.¹⁶ Both the resolution and the pruning can be also used to imply the desired a priori level of smoothness of the hazard function.

3.2.4 Parameter estimation—The log-likelihood in equation (1) can be re-written for the NPMRH model as

$$\log ; L(T|\boldsymbol{\beta}, H, R_{m,p}, X) \propto \sum_{\ell=1}^{\mathcal{L}} \sum_{i \in \mathcal{I}_{\ell}} \sum_{j=1}^J (\delta_{ij\ell} \log(d_{j\ell} + X'_{i\ell} \boldsymbol{\beta}) - \tau_{ij\ell} d_{j\ell} \exp \{X'_{i\ell} \boldsymbol{\beta}\}) \quad (2)$$

where $\ell = 1, \dots, \mathcal{L}$ are the different strata for the non-proportional hazards group, and

$$\delta_{ij\ell} = \begin{cases} 1 & \text{if subject } i \text{ is in treatment group } \ell \text{ and has a failure in time bin } j \\ 0 & \text{otherwise} \end{cases}$$

$$\tau_{ij\ell} = \begin{cases} (t_{j\ell} - t_{j-1,\ell})/\omega & \text{if subject } i \text{ is in treatment group } \ell \text{ and } T_{i\ell} > t_{j\ell} \\ (T_{i\ell} - t_{j-1,\ell})/\omega & \text{if subject } i \text{ is in treatment group } \ell \text{ and } T_{i\ell} \in [t_{j-1,\ell}, t_{j\ell}] \\ 0 & \text{otherwise} \end{cases}$$

with ω representing the bin width. The non-proportional covariate effect for covariate k is written as $\boldsymbol{a}_k = (a_{k1}, \dots, a_{k\mathcal{L}})$ and can be described as the log of the hazard ratio between the hazard rate for covariate group k and the baseline hazard rate.

Parameters are estimated using Gibbs sampler,⁶¹ with estimation performed two steps: the pruning step and the Gibbs sampler routine. Details are provided in Appendix 1. In this article, we implement the NPMRH model using the estimateMRH() function in the MRH package²³ in R.

3.3 Alternative models

While there are a number of alternative models for addressing non-proportionality between hazards, we focus on three that have available software in CRAN-R¹⁸ and that meet our main criteria: They provide an estimate of the hazard rate and allow for covariate effects to change over time. There are other models that we do not consider here, such as the Bayesian AFT model by Komárek and Lesaffre⁶² and Komárek⁶³ as AFT models do not allow for crossing hazards. We also do not present results for the survival model based on work by De Iorio et al.⁵¹ and Jara et al.,⁶⁴ as it did not converge when analyzing the biochemical failure data (which may be due to its complete non-parametric form). We also do not include

parametric models or models that only allow for the estimation of covariate effects and not the hazard rate, such as that of Yang and Prentice.⁶⁵

3.3.1 Aalen's additive model—While the majority of survival analysis applications are carried out estimating the effects of covariates under the proportional hazards assumption, Aalen's additive model^{19,20} offers an alternative method for estimating the effects of covariates and is designed to estimated covariate effects that change over time. The hazard rate of the Aalen's additive model is expressed as

$$h(t|\phi(t))=h_0(t)+Y'\phi(t)$$

where $\phi(t)=(\phi_1(t), \dots, \phi_{z_1+z_2}(t))$ is a vector of additive covariate effects at time t , which are referred to as “regression functions.” Additionally, Y is a matrix that includes both X and Z , as the covariate effects in the additive model all vary over time (i.e. no effect is constant over the course of a study).

The advantage of this model is that it produces estimates and standard errors non-parametrically using Martingale theory. We note there are other models that implement Martingale theory to produce nonparametric survival estimates in a similar context.^{33,66,67} However, we discuss this model explicitly as it is easily implemented in R and fits the estimation criteria we are interested in.

To simplify the non-parametric estimation, the parameters of interest are the cumulative regression functions, meaning that that estimates of

$$\Phi_s(t)=\int_0^t \phi_s(u)du, \quad (s=1, \dots, z_1+z_2)$$

are determined instead of an estimates of the $\phi_s(t)$. The estimate of the cumulative hazard $H(t)$ is calculated using empirical methods

$$\hat{H}(t)=\sum_{T_{(k)}\leq t} W^{-1}(T_{(k)}) \mathbb{1}_k$$

where $T_{(k)}$ is the k th ordered observed failure time, $W(t)$ is a generalized inverse of $Y(t)$, and $\mathbb{1}_k$ is an indicator vector with a “1” denoting the subject who had an observed failure at time T_k and “0” for all others. The cumulative estimates of the covariate effects are calculated similarly. Using this method of estimation, the estimates are then stochastic integrals with respect to Martingale, and are therefore asymptotically normally distributed with errors that can be calculated using similar methods to the above.

In this article, we implement the Aalen's additive model using the `aalen()` function in the survival package⁶⁸ in R. Output from this model includes the cumulative hazard increments (we refer to them as $d(t)$ as they are similar to the hazard increments in the NPMRH model) and the decumulated $\hat{\Phi}_s(t)$. Aalens additive model provides estimates of the effects of covariates on the additive scale, which is different from the estimated effects produced by

the multiplicative models (e.g. the proportional hazards model). For this reason, to assure that the different models can still be compared, we estimate the baseline hazard rate and the hazard rate for each covariate, and then calculate the log-hazard ratio. This allows the covariate effects to be reported in the classic log-ratio manner, and compared across models, i.e.

$$\hat{\alpha}_s(t) = \log \left(\left(\hat{H}(t) + \hat{\Phi}_s(t) \right) / \hat{H}(t) \right)$$

To produce reasonable estimates of the hazard rate and covariate effects, smoothing of the functions was required. We show the original estimates produced by the model in Section 4.2.1 and provide details on how these estimates are smoothed in Appendix 2.

3.3.2 HARE model—HARE model is specifically designed to estimate the hazard rate and effects of covariates without requiring the assumption of proportionality between hazards.²¹ The baseline hazard rate is estimated using linear splines and tensor products, such that the log of the hazard function, $\eta(\cdot)$, through a spline basis, and is expressed as

$$\log(h(t|X, Z, \xi)) = \eta(t|\mathbf{X}, \mathbf{Z}, \xi) = \sum_{s=1}^p \xi_s B_s(t|X, Z),$$

where B_s , $s = 1, \dots, p$ is the basis for the p -dimensional linear space of functions used to describe $\eta(t|\mathbf{X}, \mathbf{Z}, \xi)$. The HARE model estimates the parameters using the method of maximum likelihood, with the log-likelihood in equation (1) expressed as

$$L(T|\xi, \mathbf{X}, \mathbf{Z}) = \sum_i \left(\delta_i \eta(T_i|X_i, Z_i, \xi) - \int_0^{T_i} \exp \{ \eta(u|X_i, Z_i, \xi) \} du \right)$$

One of the special features of the HARE model is that the estimation procedure includes a routine for determining the optimal model. This includes finding the best choice for the spline basis, assessing whether a covariate effect changes over time, and whether covariates have a statistically significant effect on the time to failure. This part of the estimation process is done through stepwise addition, stepwise deletion, and comparison of the Bayes Information Criterion⁶⁹ (BIC) measure. This provides an advantage in that the model reported to the user is optimal based on a number of criteria.

In this article, we implement the HARE model using the hare() function in the polyspline package⁷⁰ in R.

3.3.3 A time-varying Cox model—The time-varying Cox model that we examine in this article allows the treatment effect to change over the course of the study.²² The model we examine has a hazard rate with the form

$$h(t|\mathbf{X}, \mathbf{Z}, \beta, \alpha) = \exp \{ \eta_0(t) \} \exp \{ \mathbf{X}'\beta + \mathbf{Z}'\alpha(t) \},$$

where $\exp\{\eta_0(t)\}$ is the baseline hazard rate. Note that in this version of the time-varying Cox model, the covariate effects can be fixed (represented by $X'\beta$) or change over time ($Z'\alpha(t)$) and are specified based on evidence of non-proportionality or biological rationale. The regression coefficients are estimated non-parametrically, using score equations.

The model specification above demonstrates how the time-varying Cox model is similar to other survival models. However, similar to Aalen's additive model, the estimates of $\eta(t)$ and $\alpha(t)$ are reported as cumulative estimates over the study period. For example, the time-varying effects of a parameter $\alpha_s(t)$, $s = 1, \dots, z_2$ are calculated and reported as

$$\mathcal{A}_s(t) = \int_0^t \alpha_s(u) du \quad (3)$$

In order to obtain “classic” estimates of the hazard rate and log-ratio between hazards, the cumulative estimates must be smoothed and then decumulated. The estimates we report in our analyses are the smoothed, decumulated estimates, with smoothing performed using a smoothing spline (implemented with `smooth.spline()` in R), with investigation on the impact of the choice smoothing parameter. We provide details on how these estimates are smoothed in Appendix 2.

In this article, we implement the time-varying Cox model using the `timecox()` function in the `timereg` package⁷¹ in R.

4 Model comparison through simulations

The four different models we discussed in Section 3 have very different implementations and methods for estimation. To compare the performance of the different models, we carried out a set of simulations that allowed us to compare the estimates, accuracy, and robustness of each model. The simulated biochemical failure times for the +0m group were generated from a baseline hazard rate with an initial peak and a steady decline, providing only a small number of observed failures towards the end of the study. The hazard rate for the +24m treatment group was flat and crossed the hazard rate of the +0m group at the mid-study mark. In this way, the +24m treatment was beneficial in the beginning of the study, but then performed worse than the +0m group as time progressed. (The simulated data we use is perhaps more extreme than what is typically observed in biochemical failure patterns. However, the data were generated in this way to really highlight differences among the four models.) Additionally, we included a binary covariate with a time-invariant effect, which for ease of communication of our results we refer to as the “PSA” covariate in this section (with subjects having an initial post-radiation PSA value above the mean representing the baseline group). We generated 100 data sets, each with 2000 subjects split into the covariate groups evenly.

The estimated results from all models were compared through the sample mean, the 2.5% and 97.5% quantiles of the estimates, and the estimated bias and RMSE of the 100 simulations. The estimated bias of the hazard rate or the non-proportional treatment effect

was calculated for each model at each time point t_j (where the time point t_j has an estimate) as

$$\widehat{bias}(\hat{\theta}(t_j)) = 1/100 \sum_{i=1}^{100} (\theta(t_j) - \hat{\theta}(t_{ij})), \quad (4)$$

and the RMSE was calculated at each time point as

$$\widehat{RMSE}(\hat{\theta}(t_j)) = \sqrt{1/100 \sum_{i=1}^{100} (\theta(t_j) - \hat{\theta}(t_{ij}))^2}. \quad (5)$$

For the remainder of the article, for the purpose of notational simplicity, we omit the hats on the estimated bias and RMSE.

4.1 Comparison of NPMRH models

We compared a variety of NPMRH models, all with 32 bins (i.e. $M = 5$). The different models included one with an unpruned prior tree (NPMRH-0), one with the bottom two levels of the tree pruned (NPMRH-2), and one with all five levels of the tree pruned (NPMRH-5). Among the NPMRH-5 models, we compared the results of the model with the default k value ($k = 0.5$) to one with k sampled, and one with $k = 8$, producing very smooth estimates. Each NPMRH model was run for 50,000 MCMC iterations, with the first 10,000 iterations burned, and the remaining thinned by 10 to alleviate autocorrelation. Point estimates for the NPMRH models were calculated as the median of the marginal posterior distribution of each parameter.

The results of the simulations can be observed in Figures 2 through 4. In Figure 2, the 2.5% and 97.5% of the 100 estimates for all models are contrasted against each other, with the true values used to generate the data imposed in black. In the left column, the different levels of pruning are compared (all with $k = 0.5$, the default value). The most obvious feature of these graphs is the bounds of the unpruned model, which is larger than those of the pruned models. There is not much difference between the NPMRH-2 and NPMRH-5 models, as many of the bins in the top three levels of the tree are not fused. The right column of the figure contrasts three NPMRH-5 models: the default ($k = 0.5$, also included on the left), the model where k is sampled, and the model with $k = 8$. There is very little difference between the default model and the model where k is sampled, although this makes sense as the means of the 100 k estimates were 0.491 and 0.439 for the baseline hazard rate and the treatment hazard rate, respectively. A large difference is noted for the models with $k = 8$, with very smooth results and bounds that do not capture the true hazard rate or treatment covariate effect. However, interestingly enough, this model captures the treatment hazard rate perfectly and with the smallest bounds. This indicates possible issues of identifiability with the smoothest model. The 2.5% and 97.5% quantile bounds of the low PSA effect (i.e. time-invariant covariate effects) estimates are very similar across all models (see Figure 3).

Comparison of the estimated bias and RMSE of the baseline hazard rate, log-hazard ratio of the treatment effect over time, and the treatment hazard rate highlight the trade-off in accuracy of estimation between the different models (see Figure 4). While the NPMRH-5 model with $k = 8$ has the lowest integrated absolute estimated bias and RMSE for the treatment hazard rate, its accuracy in estimation of the other parameters is very poor. The models that seem to perform the best when looking at all parameters are the NPMRH-5 model with $k = 0.5$ or k sampled.

4.2 Comparison of all models

We now compare the results of the NPMRH models with those of the other models detailed in Section 3. While multiple versions of each model were considered, we only report the “best” model, with decisions about the best model described in detail below.

4.2.1 Smoothing Aalen’s additive and time-varying Cox models—The Aalen’s additive models and time-varying Cox models both report cumulative estimates that are stochastic. In Figures 5 and 6, we show the form of the original estimates for the simulated data. The estimates have different forms (the hazard rate is log-transformed in the time-varying Cox model, and the constant PSA effect is time-varying in Aalen’s additive model), but are all cumulative.

To report the hazard rates and log-hazard ratio of the treatment effects in the “classic” manner, the estimates must be decumulated and possibly converted further. However, the raw decumulated estimates are highly variable and difficult to interpret, so some type of smoothing must also be implemented. (We outline the specific steps for doing this in Appendix 2.) Because of this, the choice of the degree of smoothing to use can dramatically effect the reported results. Therefore, for each simulated data set, we smooth the estimates using `smooth.spline()` in R, with degrees of freedom ranging from 3 to 25. We then retain the model results with the lowest Akaike Information Criteria (AIC) (which is approximated using the smoothed values plugged into an equation similar to that of equation (1)), and those are the models we report in our results.

4.2.2 Simulation results—All models reported in this section are those with the lowest AIC value for each simulated data set. The average degrees of freedom used to smooth the Aalen’s additive model were 3.1 (sd = 0.3), and the average degrees of freedom used to smooth the time-varying Cox model were 5.1 (sd = 1.4). Among the NPMRH models, 82% of the models with the lowest AIC were the NPMRH-5 models with the default $k = 0.5$. The HARE model did not need smoothing; however, we experienced some difficulties with the HARE model in that some of the estimates were very extreme, and so we contrast the results of the HARE model with all estimates included, and only the reasonable estimates included.

Estimates from the simulations are shown in Figure 7, which contrasts the four model types and the estimates of the baseline hazard rate, log-hazard ratio, and treatment hazard rate. Each of the 100 estimates is shown in lighter lines, with the mean superimposed with a darker line. The true value of each of the parameters is shown with a solid black line. The results show a number of interesting things. The HARE model generally performs quite well, although because some of the estimates were not “reliable” (we found eight simulation

results with estimated treatment effects equal to infinity), the mean of the simulations was unreasonable. The Aalen's additive model and the time-varying Cox models both have trouble capturing the curves of the baseline hazard rate and the log-hazard ratio of the treatment effect. Unlike the Aalen's additive model, however, the time-varying Cox model recovers the treatment hazard rate very well, which is similar to the pattern we observed with the NPMRH-5 model with $k = 8$.

Aalen's additive model is additionally hindered by producing time-varying estimates for the effects of the PSA covariate, which should have a constant effect over time. This is clearly observed in Figure 8, where Aalen's additive model has estimates and 2.5% and 97.5% bounds the change dramatically over time. The estimated PSA effect for the HARE model is also displayed as a time-varying effect because in 56% of the simulations, the optimal HARE model calculated a time-varying effect for PSA. However, the HARE model estimates and bounds match closely with the true value of the PSA effect as well as the bounds produced by the NPMRH and time-varying Cox models, which are shown on the left.

The estimated integrated absolute bias and the integrated RMSE for each of the models can be seen in Figure 9, which tells a story similar to that seen in Figure 7. Among all functions (baseline hazard rate, treatment log-hazard ratio, and treatment hazard rate), the HARE performs best if we remove the unreasonable results (we defined "unreasonable" as those with estimated treatment effects equal to infinity). However, when all simulation results from the HARE model are included, this model performs the worst by far. We also point out that in analyzing real data, it may be difficult to define "unreasonable," and subsetting results would not be possible. While the time-varying Cox model captures the treatment hazard rate well, it does a poor job in capturing the true shape of the other two functions. The NPMRH model performs best (after the HARE subset) in estimating both the baseline hazard rate and the treatment log-ratio, while the Aalen's additive model does not perform well in any of the measures. The total estimated integrated absolute bias and RMSE values for all models and all measures can be found in Table 2.

5 Analysis of RTOG prostate cancer clinical trial data

The analyses here investigate the effects of treatment (+0m vs. 24m AD therapy), age (categorized: less than 60 years, 60 to 70 years, 70 to 80 years, and 80 years or older), Gleason + scores (categorized: low grade, intermediate grade, and high grade, corresponding to scores between 2–4, 5–7, and 8–10, respectively), PSA levels (log transformed and then centered at the mean log value equal to 3), and T-stage at diagnosis (binary: stage 2 or stage 3/4) on the time to biochemical failure, allowing for possible non-proportional treatment effects. In order to provide a thorough investigation of the treatment hazard ratio over time and to determine the effects of different modeling and smoothness assumptions on the estimate, we present results for a variety of NPMRH models and compare them to the other models outlines in Section 3.3. The baseline (reference) group comprises subjects who received no additional AD therapy (+0 m), had an intermediate Gleason score, were below age 60 at study entry, and had a T-stage equal to 2. As shown previously, Schoenfeld residuals and methods presented in Held-Warmkessel³² showed evidence that the treatment

effect (long-term vs. short-term therapy) was not proportional over the entire study period (see Figure 1). No other covariate effects showed evidence of non-proportionality over time.

5.1 MRH results

A time resolution $M = 6$ was chosen for the NPMRH analysis in order to provide a fine grain examination of biochemical failure patterns over the course of more than 13 years. Prior trees with more than six levels were not explored as they are clinically impractical (with very few failures in smaller bins of time). Additionally, we did not explore prior trees with fewer levels (i.e. $M = 5$ or fewer) as we used the pruning tool to dictate the fusing of bins within each level of the tree. The resulting $J = 64$ time intervals, partitioning the time axis into bins of length 2.5 months, allowed us to investigate in detail about the biochemical failure hazard rate that is useful to clinical practice. The full 64-bin MRH model with non-proportional treatment hazards (the “NPMRH-0” model) was compared to two pruned MRH models with non-proportional treatment hazard rates, one with all six levels of the MRH tree pruned (“NPMRH-6”), and one with only the bottom three levels of the MRH tree pruned (“NPMRH-3”). Each model was fit using the default k value and by sampling k .

Five separate Markov chain Monte Carlo (MCMC) chains were run for each model, each with the burn-in of 50,000, leaving a total of 150,000 thinned iterations in each chain for analysis. Convergence was determined through the Geweke diagnostic,⁷² graphical diagnostics, and Gelman-Rubin tests.^{73,74} Point estimates for the MRH models were calculated as the median of the marginal posterior distribution of each parameter. Central credible intervals for each parameter were calculated as the 2.5% and 97.5% of each marginal posterior distribution.

The results of the estimated +0m and +24m treatment hazard rates and the treatment log-hazard ratio are shown in Figure 10 (these results are from the NPMRH-6, k sampled model, which we present because it has the lowest AIC value among the different models). These “caterpillar plots,” which are boxplots of the MCMC chains for each parameter, show that the biochemical failure hazard rate for the +0m treatment group (left graph) increases steadily for the first two years, and then generally declines afterward. However, the subjects who received an additional 24 months of AD therapy (middle graph) had a much flatter hazard rate that persisted through most of the study. The exception to this is a bump in the hazard rate at two years, immediately following the end of the AD therapy regimen. This close examination of the hazard rate allows us to observe that while long-term treatment effects diminished over time, biochemical failure was not simply postponed for the +24m group, but the risk was in fact reduced even over a longer period of time.

The estimated treatment effect produced by the NPMRH-6 (k sampled) model can also be seen in Figure 10 in the right graph. Notably, we observe the interval-specific differences, including time periods where the treatment effects were: (1) proportional (constant) or non-proportional (changing), (2) statistically significantly different from previous periods, and (3) statistically significantly different from zero or from the proportional hazards model estimate of the treatment effect. The proportional hazards estimate is shown with a solid blue line and is calculated using the standard Cox proportional hazards model, implemented with `coxph()` in R. (We determine if the intervals are heuristically statistically significantly

different from each other by examining if the 95% boxplots overlap.) For example, between six months and two years, long-term AD therapy had an estimated 75% improvement over short-term AD therapy, which is different than the log-ratio estimate from the proportional hazards model of $\widehat{\beta}_{tx} = -0.59$, which translates to 45% improvement for the +24m group. Additionally, the estimated log-ratio in this time period is statistically significantly different than the estimated treatment effects between 5–6.5 and 8–10 years, which only showed an estimated 26% improvement for subjects on long-term therapy. This model shows that the treatment effect held steady for a certain number of years, then diminished slightly, and held steady for another number of years before diminishing in effectiveness again.

The estimates and their 95% credible intervals for the time-invariant effects (effects of age, Gleason scores, PSA measures, and T-stage) are almost identical among all the MRH models. These time-invariant effect estimates show that an increase in Gleason scores was associated with an increased hazard rate, with a statistically significant difference between baseline subjects and subjects with scores greater than 8. The hazard of biochemical failure decreased with age, although significant differences were only observed for subjects between 70 and 80 years old and baseline subjects. As expected, subjects with a T-stage of 3 or 4 had a higher hazard of biochemical failure compared to subjects with a T-stage equal to 2. Similarly, for every point increase in PSA scores on the log scale (a 2.7 factor increase in PSA measures on the standard PSA scale), there was a statistically significant increase in the hazard rate. The results in Table 3 can also be seen in the model comparison section (Figure 14).

5.1.1 Effects of Smoothing the NPMRH Model—A contrast of the different NPMRH models (with k fixed and sampled) can be seen in Figure 11. Both NPMRH-0 models (left) show evidence of overfitting, and therefore graphically do not appear to be the best model for the data. As we observed in the simulations, the NPMRH-0 models have the widest credible interval bands, while the other models have the narrowest credible interval bands, which is due to the smaller number of estimated parameters and higher failure counts per bin in the pruned model. The NPMRH-3 and NPMRH-6 models do not show many differences, which is because only five additional pairs of bins are fused in the upper three levels of the tree. The credible interval bounds between the models with k fixed and the models with k sampled are very similar, although not identical across the entire study.

5.1.2 Uncertainty propagation of the probability of biochemical failure—The Bayesian implementation of the NPMRH model allows for easy calculation of the estimated probability of biochemical failure and the uncertainty of the estimator through the marginal distributions of the parameters. (These are calculated using the MCMC chains.) The probability of biochemical failure at 1, 5, and 10 years can be observed in Figure 12, which shows the smooth posterior predictive probability densities of biochemical failure, stratified by treatment type for hypothetical subjects with a “worst” or “best” covariate profile. A subject with a “worst” profile had a Gleason score 8, a T-stage 3 or 4 tumor, and a PSA score equal to 1 standard deviation greater than the mean (PSA $\approx$ 52). A subject with a “best” profile had a Gleason score 4, a T-stage equal to 2, and a PSA score equal to 1 standard deviation below the mean (PSA $\approx$ 8).

The advantage of examining these predictive densities is that we can easily assess if the probability of biochemical failure is statistically significantly different than one another. We can examine both the densities (top graph) and the 95% credible intervals (bottom graph) to check for overlapping. If the credible intervals do not overlap, then there is evidence that the probability of failure for the non-overlapping groups is statistically significantly different. Based on these criteria, we see that at the one-year mark, all groups are relatively similar, with the probability of biochemical failure concentrated between 0 and 20%. However, by the five-year mark, the +0m worst group has a significantly higher probability of failure than all other groups, with a worst profile subject on +0m AD therapy having a failure probability centering around 80%, and a best profile subject on +24m of AD therapy had failure probability centering around 20%. Additionally, we can see that for both treatments, the best and the worst profile subjects are significantly different from each other. This same pattern can be observed at the 10-year mark, but with additional spread. A worst profile subject on +0m AD therapy had a failure probability ranging from 80 to 100%, while a best profile subject on +24m therapy had failure rates centering around 40%. We note that these probability estimates are from a survival model treating deaths preceding biochemical failure as independent censoring. To model the observed probability of biochemical failure in the presence of these deaths (due to non-prostate cancer causes) that remove patients from biochemical failure risk, a competing risks modeling approach could be developed.^{46,75}

5.2 Model comparison

We compared the NPMRH models with the results of the three other models we examined in the simulations. This allowed us to assess the models using a real data set, particularly in examining differences in the standard errors of the estimates. The NPMRH model, the Aalen's additive model, and the time-varying Cox models can all be smoothed using different techniques, which can have a significant impact on the resulting estimates. Therefore, in this section, we present the models that have the lowest AIC.⁷⁶ For the Aalen's additive and time-varying Cox models, the smoothed estimates were calculated using a smoothing spline with degrees of freedom ranging from 3 to 120. For each of these 118 smoothed estimates, the AIC values were approximated by calculating the log-likelihood function in equation (1), with the number of parameters equal to the degrees of freedom in the smoother plus the number of covariates included in the model. The model with the lowest AIC was retained. We also present the NPMRH model from Section 5.1 with the lowest AIC. The HARE model did not require any model comparison on our part, as the optimal AIC-based model is chosen internally by the estimation routine.

Based on these criteria, the final models we present are Aalen's additive model with four degrees of freedom in the smoothing spline, the NPMRH-6 (k sampled) model (which we refer to as "NPMRH" below), a HARE model, and the time-varying Cox model with 16 degrees of freedom in the smoothing spline. We examined the estimates and the standard errors produced by the different models, as well as the AIC values, the BIC⁶⁹ values, and the Kullback-Leibler divergence measure,⁷⁷ which is a method for measuring the discrepancy between two probability measures.

5.2.1 Comparison of estimates and standard errors—Estimates of the hazard rate produced by the four different models can be observed in Figure 13, with the estimated baseline hazard rate for the +0m treatment group on the top row, and the log-hazard ratio of the long-term treatment effect on the bottom row. On the top row, we see that the NPMRH model has the narrowest 95% CI (credible interval for the NPMRH model, confidence intervals for the remaining three models) for the hazard rate, while the HARE model has the largest bounds. Additionally, the NPMRH model is the only model that has a lower CI bound that does not go below zero. The HARE baseline hazard rate estimate is the smoothest and follows a pattern generally similar to that of the NPMRH model. The time-varying Cox model also shows an initial increase in the baseline hazard rate, followed by a slight decline, but overall the hazard rate is lower than that of the HARE and NPMRH-5 models. Aalen's additive model is the only model that shows the baseline hazard rate increasing over time.

On the bottom row of Figure 13, we observe the differences in the log-hazard ratio of the treatment effect. Most notably, the HARE model estimates a constant effect of treatment over time and did not allow the estimate to vary over the course of the study (thus producing very narrow 95% CI bounds). The NPMRH model, Aalen's model, and the time-varying Cox model show a similar pattern of a decreasing long-term treatment effect (although the Aalen's additive model produces a stochastic estimate in the first six months of the study), with high amounts of variation at the end of the study where the number of observed failures is small.

The estimated fixed covariate effects are compared visually in Figure 14. The NPMRH and time-varying Cox model both produced fixed estimates that are similar, including the 95% CIs. The HARE model did not produce estimates for covariates that were not significant predictors of biochemical failure, and so only estimates for the effects of a high Gleason score, age from 70 to 80 years, and PSA are shown with some similarities and some differences (particularly with PSA) in relation to the NPMRH and time-varying Cox models. As shown in the simulations, the Aalen's additive model produces highly variable estimates that change over time. While the estimates and their 95% CIs are relatively similar to those of the other models throughout most of the study, the first six months and last one to two years have dramatic departures from the other estimates.

5.2.2 Model performance comparison—Table 4 shows several information criteria (AIC and BIC) for all the models considered. Among the NPMRH models, those where k was sampled generally had lower AIC and BIC values (with the exception of the NPMRH-0 models). The NPMRH-6 model, which also has the smallest number of parameters, also had the lowest AIC and BIC values. Among all the models, the HARE model had the lowest AIC value, and the time-varying Cox model had the lowest BIC value. Aalen's additive model had the higher AIC and BIC values when compared to all models except the NPMRH-0 models.

In addition to information criteria comparison, the set of models were also compared based on their Kullback-Leiber (KL) divergence measures,⁷⁷ which is written as

$$D_{KL}(P||Q_j)=\sum_t P(t)\log\left(\frac{P(t)}{Q_j(t)}\right)$$

where $P(t)$ is the “true” probability measure (in this instance, it is calculated as the Kaplan-Meier survival function at time t^{78}), and $Q_j(t)$ is the estimated probability measure for model j , which in this instance is the survival function at time t calculated using the estimate from model j . These calculations can be seen in Figure 15, which shows the KL measures at 1, 2, ..., 13 years on the left and the cumulative KL measures on the right. The NPMRH and time-varying Cox models have similar and lower KL scores throughout most the entire study, with the exception of the first two years. The HARE model is third throughout most of the study, and the Aalen’s model is last. The cumulative values in year 13 can also be observed in Table 4.

6 Discussion

As we have shown, there are additional models that can provide estimates of the hazard rate and the effects of covariates that may change over time and have available software for implementation. The results of the analysis of the biochemical failure data set were varied, although most models generally showed an upward initial peak in the baseline hazard followed by a slow decline, and a beneficial but decreasing treatment effect over time. In this analysis, the time-varying Cox model seemed to perform well: it had the lowest AIC value, as well as the smallest cumulative KL divergence measures. However, the results of the simulations do not show the time-varying Cox model to provide accurate estimates of the baseline hazard rate or log-hazard ratio of the treatment effect. The HARE model also performs well in analyzing the biochemical failure data set in that it has the lowest BIC score. However, it does not provide a time-varying estimated effect for long-term treatment when most evidence points to this being the case. Additionally, the results from the simulations make it difficult to know if the results from the time-varying Cox or HARE on the biochemical failure data set are reliable, given the impact of smoothing on the time-varying Cox estimates, and the variability of the HARE estimates in producing the target quantities.

On the other hand, the NPMRH model performs well in the simulations and has lower AIC and BIC scores in the analysis of the biochemical failure data, particularly when pruning is implemented. We do note, however, that the NPMRH model requires a longer time to run and fit a model (as is the case with most Bayesian models). For example, the NPMRH-6 model takes about 7 h to complete 100,000 MCMC iterations, while the time-varying Cox and Aalen’s Additive models take less than 10 min and the HARE model takes seconds. Because of this longer run-time, the MRH package has features that accommodate running the MRH models as a background job. However, the length of time it takes the MRH model to run poses an advantage for the other modeling strategies.

Use of the NPMRH provided insight into the effects of the duration of AD therapy on biochemical failure, and in particular into how the effects of AD therapy changed throughout

the course of the study. While it was already apparent that 24 months of additional AD therapy is beneficial (relative to the 0 additional months of AD therapy) in that it prolongs the time until biochemical failure and other failure endpoints,³ our investigation has revealed additional insights through examination of the hazard rate of each treatment group. During and immediately after active therapy, the peak in the hazard rate around two years is much flatter for the +24m treatment group. In addition, the +24-month group continued to have a lower hazard rate throughout most of the observation period (over 10 years), although smaller due to the non-proportionality of the treatment effect. Thus, it does appear that the benefits of the additional months of AD therapy, while diminishing over time, are persistent, which suggests that failure in the longer AD duration group are not simply deferred but possibly avoided. We also illustrate how the Bayesian approach can allow the use of posterior predictive failure probabilities, such as in Figure 12, as aids in clinical contexts.

Acknowledgments

The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver, and the National Center for Atmospheric Research. Janus is operated by the University of Colorado – Boulder.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors would like to acknowledge National Institute of Health grants R21 DA027624-01 and U01 GM087729-03, and the National Science Foundation grants NSF-DEB 1316334 and NSF-GEO 1211668 that partially supported Dr. Dukic and Dr. Hagar. National Institutes of Health grants U10 CA21661, U10 CA37422, U10 CA180822, UG1CA 189867 and U10 CA180868 (from the National Cancer Institute) supported conduct of the clinical trial and Dr. Dignam's work. Additional support was provided by Pennsylvania Department of Health 2011 Formula Grant (the Department specifically disclaims responsibility for any analyses, interpretations or conclusions.) The project utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder.

Appendix 1. Sampling and pruning steps for the NPMRH model

The estimation algorithm is performed two steps: the pruning step and the Gibbs sampler routine. The details are listed below.

Pruning step

The pruning step is run only once for each of the $\mathcal{L}$ hazard rates at the beginning of the algorithm as a pre-processing step in order to finalize the MRH tree priors. The $R_{m,p,\ell}$ parameters for which the null hypothesis is not rejected are set to 0.5 with probability 1, while the rest are estimated in the MCMC routine.

Gibbs sampler steps

After the pruning step, the Gibbs sampler algorithm is performed to obtain the approximate posterior distribution of $H_\ell, a_\ell, \lambda_\ell, k_\ell, \gamma_\ell$ and the $R_{m,p,\ell}$ that have not been set to 0.5 for each stratum ($\ell = 1, \dots, \mathcal{L}$) as well as β .

The algorithm is as follows, with steps repeated until convergence:

1. For each of the $\mathcal{L}$ treatment hazard rates ($\ell = 1, \dots, \mathcal{L}$):

- a. Sample H_{ℓ} from the posterior for H_{ℓ} which is a Gamma density with the shape parameter $a_{\ell} + \sum_{i \in \text{tx}_{\ell}} \delta_{i,\ell}$ and rate parameter $\lambda_{\ell}^{-1} + \sum_{i \in \text{tx}_{\ell}} \exp(X'_{i,\ell} \beta) F_{\ell}(T_{i,\ell})$, where $F_{\ell}(T_{i,\ell}) = H_{\ell}(\min(T_{i,\ell}, t_j)) / H_{\ell}(t_j)$.
- b. Sample a_{ℓ} , λ_{ℓ} from their respective posterior distributions (see below).
- c. Sample each $R_{m,p,\ell}$ for which the null hypothesis was rejected from the full conditional

$$R_{m,p,\ell}^{2\gamma_{m,p;\ell} k_{\ell}^m a_{\ell} - 1} (1 - R_{m,p;\ell})^{2(1-\gamma_{m,p;\ell}) k_{\ell}^m a_{\ell} - 1} \prod_{i \in \text{tx}_{\ell}} \left\{ [h_{\ell 0}(T_{i,\ell})]^{\delta_{i,\ell}} \exp(-\exp(X'_{i,\ell} \beta) H_{\ell} F_{\ell}(T_{i,\ell})) \right\}.$$

- d. Sample k_{ℓ} , $\gamma_{m,p,\ell}$ from their respective posterior distributions (see below).
2. With a $\mathcal{N}(0, \sigma_{\beta_s}^2)$ prior (with a known variance) on each covariate effect modeled under the proportional hazards assumption, β_s ($s = 1, \dots, z$), each has the following full conditional distribution

$$\pi(\beta_s | \beta_s^-) \propto \left(\prod_{\ell=1}^{\mathcal{L}} \prod_{i \in \text{tx}_{\ell}} [\exp\{X_{ij,\ell} \beta_s\}]^{\delta_{i,\ell}} \exp\left\{-\exp(X'_{i,\ell} \beta) H_{\ell} F_{\ell}(T_{i,\ell})\right\} \right) \exp\left\{-\frac{\beta_s^2}{2\sigma_{\beta_s}^2}\right\}$$

Note that this posterior distribution includes the full set of observations and covariates, from all strata jointly.

Full conditionals for the hyperparameters α , λ , k , and $\gamma_{m,p}$

The parameters in the prior distributions of H and all $R_{m,p}$ s for each covariate stratum ($\ell = 1, \dots, \mathcal{L}$), a_{ℓ} , λ_{ℓ} , k_{ℓ} and $\gamma_{\ell m,p}$ can either be fixed at desired values, or treated as random variables with their own set of hyperpriors. In the case of the latter, they would be sampled within the Gibbs sampler separately for each stratum, according to their own full conditional distributions. Below are the forms of these full conditional distributions for a specific set of hyperpriors we chose.

For notational simplicity, the stratum-specific index is suppressed below. The notation η^- will be used to denote the set of all data and all parameters except for the parameter η itself. The full conditionals are as follows:

- If a is given a zero-truncated Poisson prior, $\frac{e^{-\mu_a} \mu_a^a}{a!(1 - e^{-\mu_a})}$ (chosen for computational convenience), the full conditional distribution for a is

$$\pi(a | a^-) \propto \frac{H^a \mu_a^a}{\lambda^a (a-1)! a!} \cdot \prod_{m=1}^M \prod_{p=0}^{2^m-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p} k^m a} (1 - R_{m,p})^{2(1-\gamma_{m,p}) k^m a}}{B(2\gamma_{m,p} k^m a, 2(1 - \gamma_{m,p}) k^m a)} \right\}$$

- If the scale parameter λ in the Gamma prior for the cumulative hazard function H is given an exponential prior with mean μ_λ , the resulting full conditional is

$$\pi(\lambda|\lambda^-) \propto \frac{1}{\lambda^a} \exp \left\{ - \left(\frac{H}{\lambda} + \frac{\lambda}{\mu_\lambda} \right) \right\}$$

- If k is given an exponential prior distribution with mean μ_k , the full conditional distribution for k is as follows

$$\pi(k|k^-) \propto \prod_{m=1}^M \prod_{p=0}^{2^m-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p}k^ma} (1 - R_{m,p})^{2(1-\gamma_{m,p})k^ma}}{B(2\gamma_{m,p}k^ma, 2(1-\gamma_{m,p})k^ma)} \right\} e^{-\frac{k}{\mu_k}}$$

- If a Beta(u , w) prior is placed on each $\gamma_{m,p}$, the full conditional distribution for each $\gamma_{m,p}$ is proportional to

$$\frac{R_{m,p}^{2\gamma_{m,p}k^ma} (1 - R_{m,p})^{2(1-\gamma_{m,p})k^ma}}{B(2\gamma_{m,p}k^ma, 2(1-\gamma_{m,p})k^ma)} \gamma_{m,p}^{u-1} (1 - \gamma_{m,p})^{w-1}$$

Appendix 2. Smoothing and de-cumulating estimates

As mentioned in the article, the reported estimates for the Aalen's additive model and the time-varying Cox model are cumulative and highly variable. Because of this, the estimates need to be converted before they are reported in the classical manner. Here, we provide step-by-step details on how this was done.

Aalen's additive model

1. Fit the model using `aareg()`, which provides the estimated baseline hazard increments $\hat{d}(t)$ and the estimated additive covariate effects $\hat{\phi}_s(t)$, $s = 1, \dots, z_1 + z_2$ at each time point t . (Note that the hazard increments are not the same as the hazard rate, they are the cumulative hazard within each small bin of time.)
2. Calculate the estimated cumulative baseline hazard by adding the estimated hazard increments: $\hat{H}_0(t) = \sum_{y \leq t} \hat{d}(y)$.
3. Calculate the estimated cumulative hazard for each covariate by adding the estimated baseline hazard increments to the covariate effect increments:
 $\hat{H}_s(t) = \sum_{u \leq t} (\hat{d}(u) + \hat{\phi}_s(u))$ for each covariate s , $s = 1, \dots, z_1 + z_2$. (The Aalen's additive model estimates all covariate effects to change over time, so this is done for every covariate in the model.)
4. Find the estimated log-hazard ratio for each covariate by taking the log of the estimated cumulative hazard for each covariate, divided by the estimated cumulative baseline hazard: $\hat{\alpha}_s(t) = \log(\hat{H}_s(t)/\hat{H}_0(t))$.

5. Calculate the smoothed baseline hazard rate by smoothing the baseline hazard increments (we used `smooth.spline()` in our article) and then divide by the bin width.
6. Calculate the smoothed log-hazard ratio for each covariate by smoothing the log-hazard ratio, $\tilde{\alpha}_s(t)$, using the smoothing spline.

Time-varying Cox model

1. Fit the model using `timecox()`, which provides estimates of the cumulative estimates of the log-hazard ($\hat{\mathcal{N}}_s(t)$) and of the time-varying effects ($\hat{\mathcal{A}}_s(t)$), where $\mathcal{N}(t) = \int_0^t \eta(s) ds$, and $\mathcal{A}_s(t) = \int_0^t \alpha_s(u) du$, $s = 1, \dots, z_2$, at each time point t . Also provided are the estimates of the time-invariant effects, $\hat{\beta}_s$, $s = 1, \dots, z_1$.
2. Find the smoothed baseline hazard:
 - a. Calculate $\tilde{A}(t)$, the smoothed version of $\hat{A}(t)$, $t \in (0, t_j)$ using a smoothing method (we use a smoothing spline in this article, implemented using `smooth.spline()`).
 - b. Calculate $\tilde{\eta}(t)$, the smoothed version of $\hat{\eta}(t)$, by decumulating $\tilde{A}(t)$ (i.e. calculate the difference between each two time points) and then dividing these differences by the bin width.
 - c. Exponentiate $\tilde{\eta}(t)$. This is $\tilde{h}(t)$, the smoothed estimate of the baseline hazard rate.
3. Find the smoothed time-varying effects, $\tilde{\alpha}_s(t)$, $s = 1, \dots, z_1$, for each applicable covariate using the same techniques as in Step (2), but do not perform part 2 c.

References

1. Pilepich M, Winter K, John M, et al. Phase III Radiation Therapy Oncology Group (RTOG) trial 86-10 of androgen deprivation adjuvant to definitive radiotherapy in locally advanced carcinoma of the prostate. *Int J Radiat Oncol Biol Phys*. 2001; 50:1243–1252. [PubMed: 11483335]
2. Pilepich M, Winter K, Lawton C, et al. Androgen suppression adjuvant to definitive radio-therapy in prostate carcinoma: long-term results of phase III RTOG 85-31. *Int J Radiat Oncol Biol Phys*. 2005; 61:1285–1290. [PubMed: 15817329]
3. Horwitz EM, et al. Ten-year follow-up of radiation therapy oncology group protocol 92-02: a phase III trial of the duration of elective androgen deprivation in locally advanced prostate cancer. *Journal of Clinical Oncology*. 2008; 26(15):2497–2504. [PubMed: 18413638]
4. Albertsen P, Hanley J, Fine J. 20-year outcomes following conservative management of clinically localized prostate cancer. *J Am Med Assoc*. 2005; 293:2095–2101.
5. Nguyen Q, Levy L, Lee A, et al. Long-term outcomes for men with high-risk prostate cancer treated definitively with external beam radiotherapy with or without androgen deprivation. *Cancer*. 2013; 119:3265–3271. [PubMed: 23798338]
6. Taira A, Merrick G, Butler W, et al. Time to failure after definitive therapy for prostate cancer: implications for importance of aggressive local treatment. *J Contemp Brachytherapy*. 2013; 5:215–221. [PubMed: 24474970]

7. Amling C, Blute M, Bergstralh E, et al. Long-term hazard of progression after radical prostatectomy for clinically localized prostate cancer: continued risk of biochemical failure after 5 years. *J Urol*. 2000; 164:101–105. [PubMed: 10840432]
8. Dilliogluligil O, Leibman B, Kattan M, et al. Hazard rates for progression after radical prostatectomy for clinically localized prostate cancer. *Adult Urol*. 1997; 50:93–99. [PubMed: 9218025]
9. Hanlon A, Hanks G. Failure patterns and hazard rates for failure suggest the cure of prostate cancer by external beam radiation. *Adult Urol*. 2000; 55:725–729. [PubMed: 10792090]
10. Walz J, Chun F, Klein E, et al. Risk-adjusted hazard rates of biochemical recurrence for prostate cancer patients after radical prostatectomy. *Eur Urol*. 2008; 55:412–421. [PubMed: 19027223]
11. Bouman P, Dukic V, Meng X. Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*. 2005; 15:325–357.
12. Bouman P, Dignam J, Dukic V, et al. A multiresolution hazard model for multi-center survival studies: application to tamoxifen treatment in early stage breast cancer. *J Am Stat Assoc*. 2007; 102:1145–1157. [PubMed: 25620824]
13. Dignam J, Dukic V, Anderson S, et al. Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. *Breast Cancer Res Treat*. 2009; 116:595–602. [PubMed: 18830816]
14. Dukic V, Dignam J. Bayesian hierarchical multiresolution hazard model for the study of time-dependent failure patterns in early stage breast cancer. *Bayesian Anal*. 2007; 2:591–610. [PubMed: 21412443]
15. Dignam J, Dukic V, Anderson S, et al. Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. *Breast Cancer Res Treat*. 2009; 116:595–602. [PubMed: 18830816]
16. Chen Y, Hagar Y, Dignam J, et al. Pruned Multiresolution Hazard (PMRH) models for time-to-event data. *Bayesian Analysis*. 2014 In Review.
17. Hagar Y, Albers D, Pivovarov R, et al. Survival analysis with Electronic Health Record data: experiments with chronic kidney disease. *Stat Anal Data Mining*. 2014; 7:385–403.
18. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
19. Aalen O. A linear regression model for the analysis of life times. *Stat Med*. 1989; 8:907–925. [PubMed: 2678347]
20. Aalen O. Further results on the non-parametric linear model in survival analysis. *Stat Med*. 1993; 12:1569–1588. [PubMed: 8235179]
21. Kooperberg C, Stone C, Truong Y. Hazard regression. *J Am Stat Assoc*. 1995; 90:78–94.
22. Martinussen, T., Scheike, T. Dynamic regression models for survival data. New York, NY: Springer; 2006.
23. Hagar, Y., Chen, Y., Dukic, V. MRH package in R. <http://cran.r-project.org/web/packages/MRH/index.html> (2014, accessed 4 January 2017)
24. Roach M III, Hanks G, Thames H Jr, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix consensus conference. *Int J Radiat Oncol Biol Phys*. 2006; 65:965–974. [PubMed: 16798415]
25. Cooner W, Mosley B, Rutherford C, et al. Prostate cancer detection in clinical urological practice by ultrasonography, digital rectal examination and prostate specific antigen. *J Urol*. 1990; 143:1146–1152. [PubMed: 1692885]
26. Catalona W, Smith D, Ratliff T, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med*. 1991; 324:1156–1161. [PubMed: 1707140]
27. Brawer M, Chetner M, Beatie J, et al. Screening for prostatic carcinoma with prostate specific antigen. *J Urol*. 1992; 147:841–845. [PubMed: 1371559]
28. Radiation Therapy Oncology Group. RTOG 9202 protocol information. www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=9202, 2014 (accessed 11 March 2014)

29. Epstein JI, Alsbrook WC Jr, Amin M, et al. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma. *Am J Surg Pathol*. 2005; 29:1228–1242. [PubMed: 16096414]
30. American Joint Committee on Cancer. What is cancer staging?. <http://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx> (accessed 1 October 2014)
31. Held-Warmkessel, J. Contemporary issues in prostate cancer. Sudbury, MA: Jones and Bartlett Publishers; 2006.
32. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994; 81:515–526.
33. Andersen, P., Borgan, O., Gill, R., et al. Statistical methods based on counting processes. Berlin: Springer-Verlag; 1993.
34. Sinha D, Dey DK. Semiparametric Bayesian analysis of survival data. *J Am Stat Assoc*. 1997; 92:1195–1212.
35. Müller, P., Rodriguez, A. Nonparametric Bayesian inference. Beachwood, Ohio and Alexandria, Virginia, USA: Institute of Mathematical Statistics and American Statistical Association; 2013.
36. Holford T. Life tables with concomitant information. *Biometrics*. 1976; 32:587–597. [PubMed: 963172]
37. Holford T. The analysis of rates and of survivorship using log-linear models. *Biometrics*. 1980; 36:299–305. [PubMed: 7407317]
38. Laird N, Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques. *J Am Stat Assoc*. 1981; 76:231–240.
39. Taulbee J. A general model for the hazard rate with covariables. *Biometrics*. 1979; 35:439–450.
40. Dabrowska D, Doksum K, Song T. Graphical comparison of cumulative hazards for two populations. *Biometrika*. 1989; 76:763–773.
41. Parzen M, Wei L, Ying Z. Simultaneous confidence intervals for the difference of two survival functions. *Scand J Stat*. 1997; 24:309–314.
42. McKeague I, Zhao Y. Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Stat Probab Lett*. 2002; 60:405–415.
43. Wei G, Schaubel D. Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics*. 2008; 64:724–732. [PubMed: 18162108]
44. Dong B, Matthews D. Empirical likelihood for cumulative hazard ratio estimation with covariate adjustment. *Biometrics*. 2012; 68:408–418. [PubMed: 22059527]
45. Buckley J, James I. Linear regression with censored data. *Biometrika*. 1979; 66:429–436.
46. Kalbfleisch, J., Prentice, R. The statistical analysis of failure time data. Chichester: Wiley; 2002.
47. Berry S, Berry D, Natarajan K, et al. Bayesian survival analysis with nonproportional hazards. *J Am Stat Assoc*. 2004; 99:515–526.
48. Nieto-Barajas L. Bayesian semiparametric analysis of short- and long-term hazard ratios with covariates. *Comput Stat Data Anal*. 2014; 71:477–490.
49. Hennerfeind A, Brezger A, Fahrmeir L. Geoadditive survival models. *J Am Stat Assoc*. 2006; 101:1065–1075.
50. Cai B, Meyer R. Bayesian semi parametric modeling of survival data based on mixtures of B-spline distributions. *Comput Stat Data Anal*. 2011; 55:1260–1272.
51. De Iorio M, Johnson W, Müller P, et al. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*. 2009; 65:762–771. [PubMed: 19210742]
52. Ferguson T. Prior distributions on spaces of probability measures. *Ann Stat*. 1974; 2:615–629.
53. Lavine M. Some aspects of Polya tree distributions for statistical modelling. *Ann Stat*. 1992; 20:1222–1235.
54. Muliere P, Walker S. A Bayesian non-parametric approach to survival analysis using Polya trees. *Scand J Stat*. 1997; 24:331–340.
55. Hanson T, Johnson W. Modeling regression error with a mixture of Polya trees. *J Am Stat Assoc*. 2002; 97:1020–1033.

56. Hanson T. Inference for mixtures of finite Polya tree models. *J Am Stat Assoc.* 2006; 101:1548–1565.
57. Zhao L, Hanson T, Carlin B. Mixtures of Polya trees for flexible spatial frailty survival modelling. *Biometrika.* 2009; 96:263–276. [PubMed: 19779579]
58. Zhao L, Hanson T. Spatially dependent Polya tree modeling for survival data. *Biometrics.* 2011; 67:391–403. [PubMed: 20731644]
59. Wong W, Ma L. Optional Polya tree and Bayesian inference. *Ann Stat.* 2010; 38:1433–1459.
60. Nieto-Barajas L, Müller P. Rubbery Polya tree. *Scand J Stat.* 2012; 39:166–184.
61. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell.* 1984; 6:721–741. [PubMed: 22499653]
62. Komárek A, Lesaffre E. Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica.* 2007; 17:549–569.
63. Komárek A. bayessurv package in R. 2015. <http://cran.r-project.org/web/packages/bayesSurv/bayesSurv.pdf> accessed 4 January 2017
64. Jara, A., Hanson, T., Quintana, F., et al. Dppackage package in R. 2012. cran.r-project.org/web/packages/Dppackage/Dppackage.pdf accessed 4 January 2017
65. Yang S, Prentice R. Semiparametric analysis of short-term and long-term hazard ratios with two sample survival data. *Biometrika.* 2005; 92:1–17.
66. Fleming, T., Harrington, D. Counting processes and survival analysis. Hoboken, NJ: John Wiley & Sons; 2011.
67. Andersen P, Borgan O, Hjord N, et al. Counting process models for life history data: a review with discussion and reply. *Scand J Stat.* 1985; 12:97–158.
68. Therneau, TM. A package for survival analysis in s. <http://CRAN.R-project.org/package=survival> (accessed 4 January 2017)
69. Schwarz GE. Estimating the dimension of a model. *Ann Stat.* 1978; 6:461–464.
70. Kooperberg, C. polyspline: polynomial spline routines. 2016. <https://cran.r-project.org/web/packages/polspline/index.html> (accessed 4 January 2017)
71. Scheike, T. timereg package in R. 2014. <http://cran.r-project.org/web/packages/timereg/timereg.pdf> accessed 4 January 2017
72. Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments in Bayesian Statistics 4. Oxford, UK: Clarendon Press; 1992.
73. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Stat Sci.* 1992; 7:457–511.
74. Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat.* 1998; 7:434–455.
75. Prentice R, Kalbfleisch J, Peterson AV Jr, et al. The analysis of failure times in the presence of competing risks. *Biometrics.* 1978; 34:541–554. [PubMed: 373811]
76. Akaike H. A new look at the statistical model identification. *IEEE Trans Pattern Anal Mach Intell.* 1974; 19:716–723.
77. Kullback S, Leibler R. On information and sufficiency. *Ann Math Stat.* 1951; 22:79–86.
78. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958; 53:457–481.

Figure 1. Schoenfeld residuals for the treatment effect, with a weighted least squares line. Deviation from linearity indicates that the treatment effect does not fit the proportional hazards assumption. This method does not, however, provide an estimate of the changing treatment effect.

Figure 2.

The simulation results for the baseline hazard rate, treatment log-hazard ratio, and treatment hazard rate, with the 2.5% and 97.5% quantiles of the 100 estimates shown in contrast with the true value (in black). The left column displays different types of pruning, with the prior tree unpruned (NPMRH-0), the bottom two levels of the prior tree pruned (NPMRH-2), and all levels of the tree pruned (NPMRH-5), all using the default k value of 0.5. Among these models, the most notable feature is the slightly wider bounds in the NPMRH-0 model, which makes sense since there are more parameters in those models. The results of the NPMRH-2 and NPMRH-5 models are similar, as there are not many pruned bins in the upper three levels of the prior tree. The right column of the figure contrasts changes in k for the NPMRH-5 model, with $k=0.5$, k sampled, and k set to 8 (a value that will induce very smooth estimates). There is not much difference between the model where k is sampled and $k=0.5$, which makes sense as the means of the $100k$ estimates were close to 0.50 for both hazard rates. The model with $k=8$ seems to oversmooth the data and fails to capture some of the shape of the baseline hazard rate and the treatment log-ratio. However, the treatment hazard rate is captured almost perfectly with small quantile bounds.

Figure 3.

The simulation results for the PSA covariate (included under the proportional hazards assumption) with the bounds calculated as the 2.5% and 97.5% quantiles of the 100 estimates. The true value of the covariate, which is equal to -0.5 , is denoted with a solid grey line. We see very little difference in the estimates across all models.

Figure 4.

The integrated absolute estimated bias and integrated RMSE of the five NPMRH models. The trade-off between the models is clear: while the NPMRH-5 model with $k=8$ has the lowest integrated values for the treatment hazard rate, it performs poorly when examining the baseline hazard rate and the treatment log-hazard ratio. The models with some pruning (NPMRH-2 and NPMRH-5) generally perform better than the model with no pruning. There is little difference between the NPMRH-5 model with $k=0.5$ and k sampled.

Figure 5.

The cumulative estimates as they are reported by Aalen's additive model. To obtain the hazard rate and the classical time-varying estimate treatment effect, the cumulative estimates must be decumulated and smoothed. We also note that the effects of PSA, which were generated under the proportional hazards assumption, are time-varying in the Aalen's additive model.

Figure 6.

The cumulative estimates as they are reported by the time-varying Cox model. To obtain the hazard rate, the cumulative estimate must be decumulated, smoothed, and exponentiated. To obtain the classical estimate of the effects of treatment over time, the treatment effect must also be decumulated and smoothed.

Figure 7.

The 100 estimates for the baseline hazard rate, the treatment log-hazard ratio, and the +24m treatment hazard rate, as well as, with the mean of the 100 estimates included as a darker line. The solid black line is the true value of the function. The simulations show that the Aalen's additive model and the time-varying Cox model do not perform as accurately as the other two models in estimating the baseline hazard rate and treatment effect. However, unlike the Aalen's additive model, the time-varying Cox model is able to capture the treatment hazard rate very accurately. The HARE model and NPMRH model perform better than the other two, with fairly accurate estimates for all three function. However, the HARE model has eight estimates that are unreasonable, and so the mean of the simulations is erratic.

Figure 8.

The estimated effects of PSA (which has a time-invariant effect) with the true value of the parameter (-0.50) shown with a solid grey line. The mean and 2.5% and 97.5% quantiles of the estimates are included, with the NPMRH and time-varying Cox models on the left, as the PSA effect is constant in those models, and the Aalen's additive and HARE models on the right as the PSA effect can vary. The HARE model uses the spline basis to determine if a covariate should be included under the constant effects assumption, and in these simulations, 44% were included this way, and the remaining were allowed to vary over time. We observe that the lower and upper quantiles of the estimates are generally similar to those of the NPMRH and time-varying Cox model. The Aalen's additive model, which by design includes all covariates under the time-varying assumption, does not capture the true value of the covariate well, particularly towards the end of the study when the sample sizes are smaller.

Figure 9.

The estimated integrated absolute bias and integrated RMSE of the baseline hazard rate, treatment log-hazard ratio, and treatment hazard rate for the four models. The solid blue line represents the HARE model results for all simulations, and the dashed blue line represents the HARE model on the subset of reasonable results. The HARE results for the subset of models produces the best results among all measures, although when using all simulations, it produces the worst results. The NPMRH model performs second best in estimation of the baseline hazard rate and the log-hazard ratio, but the time-varying Cox model performs second best in estimation of the +24m treatment hazard rate. Aalen's additive model does not perform very well in examining any of the measures for any of the functions.

Figure 10.

Caterpillar plots of the hazard rate for the +0m treatment group, the +24m treatment group, and the log-hazard ratio of the two for the NPMRH-6 (k sampled) model. The plots display the 95% boxplots of the MCMC chains in each bin of time, which allows us to observe the median of the chains as well as the 95% bounds of the functions. Additionally, the log-hazard ratio is contrasted against the proportional hazards estimate, which was calculated using a standard Cox proportional hazards model and implemented with `coxph()` in R. We can observe that the hazard rate for the +0m treatment group is higher overall when compared to the hazard rate for the +24m treatment group, with the exception of the tail end of the study where the number of failures is sparse. We also note that the +24m hazard rate is fairly steady throughout the study, although there is an increase in risk for the first five months after long-term AD therapy has ended. Examination of the caterpillar plots for the log-hazard ratio allows us to see that the treatment effect changes over time, with certain periods significantly below zero or the proportional hazards estimate of -0.59 (denoted with the blue line).

Figure 11.

Comparison of the effects of smoothing (via pruning and by fixing or sampling k) on the log-hazard ratio of the estimated treatment effect produced by the different NPMRH models via caterpillar plots. The caterpillar plots are box plots of the MCMC chains for the estimated log-hazard ratio of the treatment effect within each time bin. The models on the left have no pruning (NPMRH-0) and show the most variation between and within bins, with reduced variability as the level of pruning increases (NPMRH-3 and NPMRH-6). Among the NPMRH-3 and NPMRH-6 models, there are only minor difference in those with k fixed and those with k sampled. All caterpillar plots have a horizontal reference line (in blue) at -0.59 , which is the estimate of the treatment effects under the proportional hazards assumption, calculated using a standard Cox proportional hazards model.

Figure 12.

TOP: Smoothed posterior predictive densities of biochemical failure at 1, 5 and 10 years post-diagnosis, stratified by treatment type and hypothetical patient covariate profile. A subject with a “worst” profile had a Gleason score 8, a T-stage equal to 3 or 4, and a PSA score equal to 1 standard deviation greater than the mean (PSA $\approx$ 52). A subject with a “best” profile had a Gleason score 4, a T-stage equal to 2, and a PSA score equal to 1 standard deviation below the mean (PSA $\approx$ 8). As time post-diagnosis increases, the predictive densities of biochemical failure became more spread out, with a worst profile subject on +0m of AD therapy having the highest predictive probability of biochemical failure, while a best profile subject on +24m AD therapy had the lowest predictive probability of biochemical failure. Smoothed density estimates were calculated using `density()` in R. BOTTOM: 95% credible intervals of the failure densities. Credible intervals that do not overlap show evidence that the difference in the probability of failure are statistically significantly different.

Figure 13.

Comparison of the estimated hazard rates for the +0m AD therapy group (top) and the log-hazard ratio of the +24m treatment group, with 95% CI bounds (credible intervals for the MRH model and point-wise confidence intervals for the remaining models) shown in lighter shades or with dashed lines. The HARE model has the widest baseline hazard rate CI bounds and estimates a constant treatment effect over the course of the study. The HARE, NPMRH, and time-varying Cox models also all estimate a baseline hazard rate with an initial increase the first two years, followed by a steady decline. Aalen's additive model shows an increasing baseline hazard rate. The CI bounds for all models are wider towards the end of the study where the number of observed biochemical failures was sparse.

Figure 14.

The estimates for the remaining covariate effects and corresponding 95% intervals (credible intervals for NPMRH, points-wise confidence intervals for the remaining models). Differences in estimates and 95% CIs are very minor between the NPMRH and time-varying Cox models, which estimate constant effects for the remaining covariates. The HARE model also produces constant estimates over time for significant predictors, although the estimates and their bounds can differ when compared to the NPMRH and time-varying Cox models. Aalen's additive model produces estimates that vary over time for all covariate effects. The estimates and bounds are generally similar to the NPMRH and time-varying Cox models, with the exception of the first six months and last year of the study, where the estimates jump dramatically.

Figure 15.

Kullback-Leibler divergence measures for the four models, with within-bin measures on the left and cumulative measures on the right.

Table 1

Sample characteristics of 1421 patients in the final RTOG 92-02 trial data set, by treatment group, with age at diagnosis in 10-year increments, Gleason scores categorized by grade, and T-stage categorized into levels 2 to 4.

	+0m AD		+24m AD		Total Sample	
	N	%	N	%	N	%
Patients	705	49.6	716	50.4	1421	100.0
Age						
Less than 60 years	44	3.1	51	3.6	95	6.7
60 to 70 years	274	19.3	260	18.3	534	37.6
70 to 80 years	363	25.5	371	26.1	734	51.7
80 or more years	24	1.7	34	2.4	58	4.1
Gleason score						
Low grade (2–4)	56	3.9	47	3.3	103	7.2
Intermediate grade (5–7)	462	32.5	495	34.8	957	67.3
High grade (8–10)	187	13.2	174	12.2	361	25.4
T-stage						
T2	325	22.9	331	23.3	656	46.2
T3	360	25.3	353	24.8	713	50.2
T4	20	1.4	32	2.3	52	3.7

The estimated total integrated absolute bias and RMSE values for each of the models and each of the measures.

Table 2

Model	Baseline Hazard rate		Treatment Hazard rate		Treatment (NPH) Log-hazard ratio		PSA (PH) Log-hazard ratio	
	Bias	(RMSE)	Bias	(RMSE)	Bias	(RMSE)	Bias	(RMSE)
Aalen's additive	0.273	(0.097)	0.755	(0.258)	14.950	(5.646)	-0.129	(0.336)
HARE								
(All)	2.2×10^{36}	4.0×10^{37}	∞	(∞)	∞	(∞)	-0.015	(0.357)
(Subset)	0.127	(0.0624)	0.186	(0.076)	3.679	(1.543)	0.012	(0.126)
NPMRH								
(0)	0.199	(0.089)	0.569	(0.234)	8.977	(3.779)	-0.008	(0.106)
(2)	0.177	(0.079)	0.376	(0.172)	6.880	(3.041)	-0.008	(0.107)
(5, k sampled)	0.193	(0.086)	0.289	(0.158)	6.949	(3.108)	-0.007	(0.106)
(5, k = 0.5)	0.192	(0.085)	0.278	(0.149)	6.709	(2.931)	-0.007	(0.106)
(5, k = 8)	0.397	(0.141)	0.105	(0.043)	10.008	(3.758)	-0.018	(0.109)
(Lowest AIC)	0.193	(0.085)	0.281	(0.151)	6.763	(2.973)	-0.007	(0.106)
Time-varying Cox	0.376	(0.131)	0.215	(0.115)	10.571	(4.164)	0.005	(0.106)

Note: Among the NPMRH models, the models based on the lowest AIC tend to have the lowest values, although there are exceptions, such as the estimated bias and RMSE for the NPMRH-5 model with $k = 8$. The other conclusions match those observed in Figure 9.

Table 3

Parameter estimates of the covariate effects (in the $\hat{\beta}$ column) and the 95% credible intervals.

Covariate	$\hat{\beta}$	95% CI for β
Gleason Score		
(Low, 2–4)	−0.11	(−0.44, −0.18)
(High, 8–10)	0.29	(0.12, 0.45)
Age		
(60 to 70 years)	−0.13	(−0.40, 0.17)
(70 to 80 years)	−0.44	(−0.70, −0.15)
(80 or more years)	−0.08	(−0.54, 0.36)
T-stage		
(3 or 4)	0.14	(−0.01, 0.28)
PSA	0.29	(0.21, 0.37)

Note: The intervals for a high Gleason score, age between 70 and 80 years, and PSA do not contain zero, implying that these covariates have a statistically significant effect on time to biochemical failure.

Table 4

Information criteria (AIC and BIC) for the NPMRH models with k fixed and k sampled, the Aalen's additive model, the HARE model, and the time-varying Cox model.

Model	$-2 \times \log(L)$	Effective # of parameters	BIC	AIC	K-L
NPMRH					
0 (k sampled)	6268.8	141	7292.3	6550.8	11.1
0 ($k = 0.5$)	5095.3	139	6104.3	5373.3	22.9
3 (k sampled)	5168.8	45	5495.4	5258.8	19.5
3 ($k = 0.5$)	5190.5	43	5502.6	5276.5	20.4
6 (k sampled) ^a	5104.9	40	5395.3	5184.9	19.6
6 ($k = 0.5$)	5182.8	38	5458.6	5258.8	20.2
Aalen's Additive ^b	5823.8	32	6056.1	5887.8	39.6
HARE	4411.2	91	5071.8	4593.2	25.5
Time-Varying Cox ^b	4633.9	39	4917.0	4711.9	19.0

Note: The values of twice the negative log-likelihood ($-2 * \log(L)$) and the effective number of parameters are shown. Lower BIC and AIC values represent models better supported by the data.

^aThis is the NPMRH model presented in the model comparison section.

^bThe Aalen's additive and time-varying Cox model log-likelihoods, number of estimated parameters, and BIC/AIC values are approximated using the smoothed fitted estimates, not the original estimates.