





Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer

Thomas A. Murray, Peter F. Thall, Ying Yuan, Sarah McAvoy & Daniel R. Gomez


To cite this article: Thomas A. Murray, Peter F. Thall, Ying Yuan, Sarah McAvoy & Daniel R. Gomez (2017) Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer, Journal of the American Statistical Association, 112:517, 11-23, DOI: [10.1080/01621459.2016.1176926](https://doi.org/10.1080/01621459.2016.1176926)

To link to this article: <http://dx.doi.org/10.1080/01621459.2016.1176926>


 View supplementary material [↗](#)

 Accepted author version posted online: 13 May 2016.
Published online: 03 May 2017.

 Submit your article to this journal [↗](#)

 Article views: 107

 View Crossmark data [↗](#)

 Citing articles: 1 View citing articles [↗](#)

Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer

Thomas A. Murray^a, Peter F. Thall^a, Ying Yuan^a, Sarah McAvoy^b, and Daniel R. Gomez^b

^aDepartment of Biostatistics, MD Anderson Cancer Center, Houston, TX; ^bDepartment of Radiation Oncology, MD Anderson Cancer Center, Houston, TX

ABSTRACT

A design is presented for a randomized clinical trial comparing two second-line treatments, chemotherapy versus chemotherapy plus reirradiation, for treatment of recurrent non-small-cell lung cancer. The central research question is whether the potential efficacy benefit that adding reirradiation to chemotherapy may provide justifies its potential for increasing the risk of toxicity. The design uses two co-primary outcomes: time to disease progression or death, and time to severe toxicity. Because patients may be given an active third-line treatment at disease progression that confounds second-line treatment effects on toxicity and survival following disease progression, for the purpose of this comparative study follow-up ends at disease progression or death. In contrast, follow-up for disease progression or death continues after severe toxicity, so these are semi-competing risks. A conditionally conjugate Bayesian model that is robust to misspecification is formulated using piecewise exponential distributions. A numerical utility function is elicited from the physicians that characterizes desirabilities of the possible co-primary outcome realizations. A comparative test based on posterior mean utilities is proposed. A simulation study is presented to evaluate test performance for a variety of treatment differences, and a sensitivity assessment to the elicited utility function is performed. General guidelines are given for constructing a design in similar settings, and a computer program for simulation and trial conduct is provided. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2015
Revised January 2016

KEYWORDS

Bayesian analysis; Group sequential; Piecewise exponential model; Radiation oncology; Randomized comparative trial; Utility elicitation

1. Introduction

1.1. Background

This article describes a Bayesian design for a randomized clinical trial to compare chemotherapy alone (C) and chemotherapy plus reirradiation ($C + R$) as second-line treatments for locoregional recurrence of non-small-cell lung cancer (NSCLC). Locoregional disease recurrence, that is, within the original region of disease, after first-line radiation therapy is a leading cause of death in patients with NSCLC. Patients who are not candidates for surgery, due to degraded health status from their first-line treatment or their disease, often are given chemotherapy as second-line treatment. Because many of these patients do not respond to chemotherapy alone, adding reirradiation to chemotherapy recently has been considered as a second-line treatment option (McAvoy et al. 2014). In this trial, reirradiation will be delivered using either intensity modulated radiation therapy (IMRT) or proton beam therapy (PBT). The central research question is whether the potential benefit $C + R$ may provide over C for delaying disease recurrence or death justifies its potential for increasing the risk of toxicity. To address this question, the design will enroll eligible patients, randomize them to either treatment, and follow them for pertinent outcomes. The trial will terminate when either a prespecified trial duration is achieved or there is strong evidence at an interim point in the trial that one treatment is clinically preferable to the other.

A complication in evaluating and comparing these second-line treatments is that each patient's clinical outcome may include some combination of disease progression, toxicity, or death, and all of these possible outcomes are very important. Moreover, an active third-line treatment often is given following disease progression. In this case, for comparing C to $C + R$ as second-line treatments, any toxicity occurring after disease progression is closely related to the third-line treatment, and thus the effects on toxicity of C or $C + R$ as second-line treatments are confounded with the effects of third-line treatment. Similarly, time to death following disease progression is confounded with third-line treatment. To obviate potential confounding from third-line treatment and account for all clinically relevant outcomes, the trial will monitor two co-primary time-to-event outcomes that terminate follow-up at disease progression or death. The first outcome is progression-free survival (PFS) time, defined as the time, Y_{Prog} , to the earliest occurrence of death, distant metastasis, or locoregional disease recurrence. For succinctness, hereafter we will refer to these events collectively as "progression." PFS is considered an acceptable clinical trial end-point in NSCLC (Pilz et al. 2012). The second outcome is time to toxicity, defined as the time, Y_{Tox} , to the earliest occurrence of any severe (grade 3 or 4) National Cancer Institute (NCI) toxicity. Because follow-up ends at progression, a toxicity is observed only if it occurs prior to progression. Hence, these events are semi-competing risks, where toxicity is the

nonterminal event and progression is the terminal event (see, e.g., Fine et al. 2001; Peng and Fine 2007).

Comparing treatments based on the event times Y_{Tox} and Y_{Prog} is challenging because the treatment effects are multidimensional and possibly in opposite directions. For example, relative to C , $C + R$ may improve PFS but increase the probability of toxicity prior to progression. Another possibility is that $C + R$ may only improve PFS given that no toxicities occur, and instead worsen PFS given that a toxicity does occur, possibly because the toxicities that result from $C + R$ compared to C are more severe and thus lead to earlier death. These complexities raise the key issue of whether a particular treatment difference is favorable, where any difference is multidimensional. Because this is central to therapeutic decision making for individual patients, it also plays a central role in scientific comparison of C to $C + R$ in the trial.

The approach often used in such settings is to compare treatments using separate monitoring criteria for PFS or overall survival (OS) and toxicity (Cannistra 2004). This approach is limited by the fact that no formal criteria are specified for deciding whether a particular increase in the risk of toxicity is justified given a particular PFS or OS improvement. If the efficacy and safety outcomes are opposing, physicians must make subjective decisions about whether either treatment is clinically preferable, accounting for these tradeoffs informally. The design described here takes a more transparent approach, based on the above two co-primary outcomes and a utility function that characterizes the clinical desirability for every possible realization of these outcomes. Clinical trial designs for co-primary outcomes have been developed in other oncology settings by Thall et al. (2000), Yuan and Yin (2009), and Hobbs et al. (2015).

1.2. Mean Utility

To combine toxicity and progression information in a practical way that is scientifically and ethically meaningful, we will compare the two treatments based on their mean utilities,

$$\bar{U}(\text{trt } j) = \int U(\mathbf{y})p(\mathbf{y}|\text{trt } j)d\mathbf{y}, \text{ for } j = C, C + R, \quad (1)$$

where $U(\mathbf{y})$ denotes a utility function that characterizes the clinical desirability of all possible realizations $\mathbf{y} = (y_{\text{Tox}}, y_{\text{Prog}})$ of $\mathbf{Y} = (Y_{\text{Tox}}, Y_{\text{Prog}})$, and $p(\mathbf{y}|\text{trt } j)$ denotes the joint probability distribution of the co-primary outcomes for patients given treatment j . For brevity, we use \bar{U}_j to denote $\bar{U}(\text{trt } j)$. Mean utilities have been used in phase I–II trials by Thall et al. (2013) for bivariate time-to-event outcomes, and Lee et al. (2015a) for bivariate binary outcomes. As far as we are aware, this is the first utility-based trial design proposed for semi-competing risks outcomes.

Both $p(\mathbf{y}|\text{trt } j)$ and $U(\mathbf{y})$ in (1) are crucial elements of the design. While $p(\mathbf{y}|\text{trt } j)$ can be estimated via a probability model for the semi-competing risks, $U(\mathbf{y})$ is subjective and must be prespecified to reflect the desirability of \mathbf{y} in this specific NSCLC context. We elicit $U(\mathbf{y})$ from the physicians planning the trial, DG and SM, who are co-authors of this article, so that $U(\mathbf{y})$ reflects their experience treating this disease in this particular clinical setting. While it may be argued that such subjectivity should not be used for treatment evaluation and

comparison, in fact all statistical methods require subjective decisions about what is or is not important, and all physicians have utilities that motivate their therapeutic decisions, whether they have written them down or not. For the methodology described here, the physicians' utilities represent a consensus, and they are given explicitly. A key advantage of our approach is that the utility function makes explicit which treatment is preferred for any efficacy–safety tradeoff, which is in contrast to an approach based on separate criteria for each outcome.

1.3. Outline

In Section 2, we formulate a novel Bayesian model for $p(\mathbf{y}|\text{trt } j)$ in (1) based on piecewise exponential distribution that is robust to model misspecification. We use historical data reported by McAvoy et al. (2014) as a basis for establishing a prior. In Section 3, we describe the utility elicitation process in detail and provide general guidelines for applications of the methodology in other contexts. We propose a parametric function for the specification of $U(\mathbf{y})$ in (1), which satisfies admissibility constraints that are relevant in this context, and elicit interpretable parameter values for this function from the physicians planning the trial. We also discuss efficient calculation of the \bar{U}_j defined in (1). In Section 4, we describe a design that uses a monitoring criterion based on posterior mean utilities of the two treatments. A simulation study is presented to evaluate operating characteristics of the proposed design for a variety of multidimensional treatment differences, and a sensitivity assessment to the utility is performed. In the NSCLC trial design, there are up to two interim tests and a possible third, final test, that is, a group sequential approach is taken, although the methodology can be applied more generally. We provide software to aid implementation of these methods in other contexts (see supplementary material).

2. Probability Model

The possible realizations $(y_{\text{Tox}}, y_{\text{Prog}})$ of $(Y_{\text{Tox}}, Y_{\text{Prog}})$ are contained in the union

$$\begin{aligned} & \{(y_{\text{Tox}}, y_{\text{Prog}}) : y_{\text{Tox}} \in [0, \infty), y_{\text{Prog}} \in (y_{\text{Tox}}, \infty)\} \\ & \cup \{(\text{No Tox}, y_{\text{Prog}}) : y_{\text{Prog}} \in [0, \infty)\}. \end{aligned}$$

The first set in this union contains all progression times with an earlier toxicity time, $y_{\text{Tox}} < y_{\text{Prog}}$, and the second set contains all progression times where no prior toxicity occurs, denoted by ‘‘No Tox.’’ We define the outcomes of the i th patient in terms of potential times to progression, $y_{\text{Prog},i}^*$, toxicity, $y_{\text{Tox},i}^*$, and independent administrative right-censoring, c_i^* , $i = 1, \dots, N$. We refer to these as potential times, indicated by a superscripted asterisk, because they may not be observed. Using conventional time-to-event definitions, the observed data for the i th patient are $y_{\text{Prog},i} = \min\{y_{\text{Prog},i}^*, c_i^*\}$, $\delta_{\text{Prog},i} = I[y_{\text{Prog},i}^* < c_i^*]$, $y_{\text{Tox},i} = \min\{y_{\text{Tox},i}^*, y_{\text{Prog},i}^*, c_i^*\}$, and $\delta_{\text{Tox},i} = I[y_{\text{Tox},i}^* < \min\{y_{\text{Prog},i}^*, c_i^*\}]$, where $I[A] = 1$ if the event A is true and 0 otherwise. We denote the treatment assigned to the i th patient by z_i , which equals C or $C + R$. Therefore, for the i th patient, we observe $(y_{\text{Tox},i}, y_{\text{Prog},i}, \delta_{\text{Tox},i}, \delta_{\text{Prog},i}, z_i)$. Since patients are followed until progression or administrative censoring has occurred,

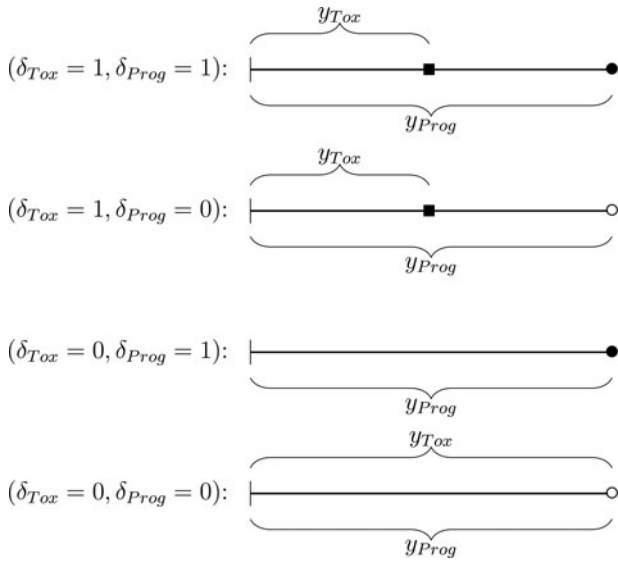


Figure 1. The four possible orderings for the two co-primary outcomes (y_{Tox} , y_{Prog}). We use a dark square to denote toxicity, a dark circle to denote progression, and a clear circle to denote administrative censoring.

$y_{Tox,i} \leq y_{Prog,i}$, and $\delta_{Tox,i} = 1$ only if $y_{Tox,i} < y_{Prog,i}$. **Figure 1** illustrates the four orderings of (y_{Tox}, y_{Prog}) that may be observed.

2.1. Likelihood Specification

We develop a Bayesian hierarchical model for the joint distribution of the co-primary outcomes by conditioning on a binary indicator, ξ , of whether a toxicity will occur prior to progression. We rely on a Bernoulli distribution to model the probability of toxicity prior to progression, and piecewise exponential distributions to model the times to toxicity and progression given that a toxicity *will* occur prior to progression, that is, $\xi = 1$, and the time to progression given that a toxicity *will not* occur prior to progression, that is, $\xi = 0$. Discussions of piecewise exponential models for univariate time-to-event outcomes have been given by Holford (1976) and Friedman (1982) in a frequentist context, and by Ibrahim et al. (2001, Section 3) in a Bayesian context.

Let $\text{Bern}(\pi)$ denote a Bernoulli distribution with probability π , let $\text{Exp}(\lambda; \tilde{\mathbf{t}})$ denote a piecewise exponential distribution with hazard vector $\lambda = (\lambda_1, \dots, \lambda_{K+1})$ corresponding to the partition $\tilde{\mathbf{t}} = (0 = \tilde{t}_0 < \tilde{t}_1 < \dots < \tilde{t}_K < \tilde{t}_{K+1} = \infty)$, and let $\text{Exp}(\lambda; \tilde{\mathbf{t}})_{[L, R]}$ denote a piecewise exponential distribution with domain truncated to $[L, R]$. The piecewise exponential distribution has the hazard λ_k for $t \in [\tilde{t}_{k-1}, \tilde{t}_k)$, $k = 1, \dots, K + 1$. Using this notation, our assumed model for $p(\mathbf{y}|\text{trt } j)$ in (1) may be summarized as follows:

$$\begin{aligned} \xi | \text{trt } j &\sim \text{Bern}(\pi_j) \\ Y_{Tox} | \xi = 1, \text{trt } j &\sim \text{Exp}(\lambda_{T,j}; \tilde{\mathbf{t}}_T) \\ Y_{Prog} | \xi = 1, Y_{Tox}, \text{trt } j &\sim \text{Exp}(\lambda_{P1,j}; \tilde{\mathbf{t}}_{P1})_{[Y_{Tox}, \infty)} \\ Y_{Prog} | \xi = 0, \text{trt } j &\sim \text{Exp}(\lambda_{P2,j}; \tilde{\mathbf{t}}_{P2}). \end{aligned} \quad (2)$$

This models each event time as a piecewise exponential with parameters depending on the observed, relevant preceding

variables. We use the subscripts T , $P1$, and $P2$ to distinguish the parameters characterizing the conditional hazards for toxicity and progression given $\xi = 1$, and for progression given $\xi = 0$, respectively. We will discuss selecting the partitions $\tilde{\mathbf{t}}_\ell = (0 = \tilde{t}_{\ell,0} < \tilde{t}_{\ell,1} < \dots < \tilde{t}_{\ell,K_\ell} < \tilde{t}_{\ell,K_\ell+1} = \infty)$ below, for $\ell = T, P1, P2$. This model reflects historical experience that some proportion of the patients given treatment j will not experience toxicity before progression, either because the treatment did not cause enough damage to result in a grade 3 or higher toxicity or because progression happened too quickly for an earlier toxicity to occur.

The model proposed in (2) is similar to that of Zhang et al. (2014), who do regression for semi-competing risks with treatment switching. Other semi-competing risks models have been proposed by Xu et al. (2010), Conlon et al. (2014), and Lee et al. (2015b). While these alternative models could be applied in this context, the proposed model greatly facilitates posterior estimation and calculation of the mean utilities in (1). Moreover, the proposed model is robust because it does not rely on restrictive parametric assumptions for the joint distribution of the times to toxicity and progression, or the treatment effects on this joint distribution, such as proportional-hazards. The proposed model can be extended to adjust for covariates \mathbf{x} by assuming say, $\log\{\pi_j/(1 - \pi_j)\} = \alpha_j + \mathbf{x}'\boldsymbol{\beta}$, and $\lambda_{\ell,j} = \lambda_{\ell,j}^* \exp\{\mathbf{x}'\boldsymbol{\gamma}_\ell\}$, for $\ell = T, P1, P2$, and $j = C, C + R$. However, since this is a randomized clinical trial, covariate adjustment is not strictly necessary to ensure fair treatment comparisons, and we do not consider it below.

We assume patients in the trial, indexed by $i = 1, \dots, N$, will contribute independent observations. For some patients, ξ_i will be observed, and for other patients, ξ_i will not be observed and it will be an unknown parameter that is updated in our Gibbs sampling algorithm. The observed pair $(\delta_{Tox,i}, \delta_{Prog,i})$ determines whether ξ_i is known, as follows. If $\delta_{Tox,i} = 1$, that is, a toxicity is observed, then $\xi_i = 1$, regardless of whether $\delta_{Prog,i} = 0$ or 1. If $\delta_{Tox,i} = 0$ this alone does not determine ξ_i , but if $\delta_{Tox,i} = 0$ and $\delta_{Prog,i} = 1$, that is, progression is observed with no prior toxicity, then $\xi_i = 0$. Finally, if $\delta_{Tox,i} = 0$ and $\delta_{Prog,i} = 0$, that is, both progression and toxicity are administratively censored, then ξ_i is unknown. However, in this case ξ_i is partially observed, in the sense that the full posterior conditional distribution for ξ_i depends on the administrative censoring time, and the likelihood contribution is a mixture.

Temporarily suppressing the patient index i and treatment index j , the four possible likelihood contributions of an individual observation are as follows:

$$\begin{aligned} (\delta_{Tox} = 1, \delta_{Prog} = 1) &: \pi h(y_{Tox} | \lambda_T; \tilde{\mathbf{t}}_T) S(y_{Tox} | \lambda_T; \tilde{\mathbf{t}}_T) \\ &\quad \times h(y_{Prog} | \lambda_{P1}; \tilde{\mathbf{t}}_{P1}) \left[\frac{S(y_{Prog} | \lambda_{P1}; \tilde{\mathbf{t}}_{P1})}{S(y_{Tox} | \lambda_{P1}; \tilde{\mathbf{t}}_{P1})} \right] \\ (\delta_{Tox} = 1, \delta_{Prog} = 0) &: \pi h(y_{Tox} | \lambda_T; \tilde{\mathbf{t}}_T) \\ &\quad \times S(y_{Tox} | \lambda_T; \tilde{\mathbf{t}}_T) \left[\frac{S(y_{Prog} | \lambda_{P1}; \tilde{\mathbf{t}}_{P1})}{S(y_{Tox} | \lambda_{P1}; \tilde{\mathbf{t}}_{P1})} \right] \\ (\delta_{Tox} = 0, \delta_{Prog} = 1) &: (1 - \pi) h(y_{Prog} | \lambda_{P2}; \tilde{\mathbf{t}}_{P2}) \\ &\quad \times S(y_{Prog} | \lambda_{P2}; \tilde{\mathbf{t}}_{P2}) \end{aligned}$$

$$(\delta_{\text{Tox}} = 0, \delta_{\text{Prog}} = 0) : \pi S(y_{\text{Tox}} | \lambda_T; \tilde{\mathbf{t}}_T) + (1 - \pi) S(y_{\text{Prog}} | \lambda_{P2}; \tilde{\mathbf{t}}_T), \quad (3)$$

where the hazard function at y is

$$h(y | \lambda_\ell; \tilde{\mathbf{t}}_\ell) = \lambda_{\ell,k} \text{ for } y \in [\tilde{t}_{\ell,k-1}, \tilde{t}_{\ell,k}), k = 1, \dots, K_\ell,$$

and the probability that $Y > y$ is

$$S(y | \lambda_\ell; \tilde{\mathbf{t}}_\ell) = \exp \left\{ - \sum_{k=1}^{K_\ell} [\min(y, \tilde{t}_{\ell,k}) - \min(y, \tilde{t}_{\ell,k-1})] \lambda_{\ell,k} \right\},$$

$$\ell = T, P1, P2.$$

Conditioning on ξ , the first three expressions in (3) are unaffected, whereas the fourth expression in (3) can be written as $[\pi S(y_{\text{Tox}} | \lambda_T; \tilde{\mathbf{t}}_T)]^\xi [(1 - \pi) S(y_{\text{Prog}} | \lambda_{P2}; \tilde{\mathbf{t}}_{P2})]^{1-\xi}$, which facilitates computation.

We are able to give this explicit form for the mixture likelihood contribution in the case where ξ is unknown, that is, $(\delta_{\text{Tox}} = 0, \delta_{\text{Prog}} = 0)$ in (3), because the distributions of the event times, Y_{Tox} and Y_{Prog} are defined conditional on ξ in (2). Given $\xi = 0$, we assume that Y_{Prog} is piecewise exponential, and Y_{Tox} does not enter the likelihood in this case because conditioning on $\xi = 0$ implies that a toxicity will not occur prior to disease progression. Given $\xi = 1$, we assume that Y_{Tox} is piecewise exponential, and, given the value of Y_{Tox} , we assume that Y_{Prog} is piecewise exponential with left-truncation time Y_{Tox} because conditioning on $\xi = 1$ implies that a toxicity will occur prior to disease progression. Lastly, we propose using distinct parameters for each treatment group, that is, $(\pi_j, \lambda_{T,j}, \lambda_{P1,j}, \lambda_{P2,j})$ for $j = C, C + R$, thereby avoiding commonly used, yet often unchecked and unjustified parametric assumptions, like proportional hazards, which typically are used to obtain a univariate parameter for treatment comparison. We avoid such parametric assumptions by using mean utility as a one-dimensional basis for constructing a comparative test, described below.

Reintroducing the individual and treatment indices, i and j , under (2), the full likelihood may be expressed in the following computationally convenient form:

$$L(\lambda, \xi, \pi | \mathbf{y}_{\text{Prog}}, \mathbf{y}_{\text{Tox}}, \delta_{\text{Prog}}, \delta_{\text{Tox}}, \mathbf{z})$$

$$\propto \prod_{j=C, C+R} \pi_j^{\sum_{i:z_i=j} [\delta_{\text{Tox},i} + (1 - \delta_{\text{Tox},i})(1 - \delta_{\text{Prog},i}) \xi_i]}$$

$$\times \prod_{j=C, C+R} (1 - \pi_j)^{\sum_{i:z_i=j} [(1 - \delta_{\text{Tox},i}) \{ \delta_{\text{Prog},i} + (1 - \delta_{\text{Prog},i})(1 - \xi_i) \}]}$$

$$\times \prod_{j=C, C+R} \prod_{k=1}^{K_T+1} \sum_{i:z_i=j} \delta_{T,i,k} \lambda_{T,j,k}$$

$$\times e^{- \left\{ \sum_{i:z_i=j} [\delta_{\text{Tox},i} + (1 - \delta_{\text{Tox},i})(1 - \delta_{\text{Prog},i}) \xi_i] y_{T,i,k} \right\} \lambda_{T,j,k}}$$

$$\times \prod_{j=C, C+R} \prod_{k=1}^{K_{P1}+1} \sum_{i:z_i=j} [\delta_{\text{Tox},i} \delta_{P1,i,k}] e^{- \left\{ \sum_{i:z_i=j} \delta_{\text{Tox},i} (y_{P1,i,k} - y_{T1,i,k}) \right\} \lambda_{P1,j,k}}$$

$$\times \prod_{j=C, C+R} \prod_{k=1}^{K_{P2}+1} \sum_{i:z_i=j} [(1 - \delta_{\text{Tox},i}) \delta_{P2,i,k}] \lambda_{P2,j,k}$$

$$\times e^{- \left\{ \sum_{i:z_i=j} (1 - \delta_{\text{Tox},i}) [\delta_{\text{Prog},i} + (1 - \delta_{\text{Prog},i})(1 - \xi_i)] y_{P2,i,k} \right\} \lambda_{P2,j,k}}, \quad (4)$$

where $\delta_{T,i,k} = 1$, if $\delta_{\text{Tox},i} = 1$ and $y_{\text{Tox},i} \in [\tilde{t}_{T,k-1}, \tilde{t}_{T,k})$, and $\delta_{T,i,k} = 0$, otherwise, that is, a binary indicator for whether a toxicity has occurred in the k th interval of $\tilde{\mathbf{t}}_K$, and $y_{T,i,k} = \min\{y_{\text{Tox},i}, \tilde{t}_{T,k}\} - \min\{y_{\text{Tox},i}, \tilde{t}_{T,k-1}\}$, that is, the overlap with the k th interval of $\tilde{\mathbf{t}}_T$. We define $\delta_{P1,i,k}$, $y_{P1,i,k}$, and $y_{T1,i,k}$ with respect to $\tilde{\mathbf{t}}_{P1}$, and $\delta_{P2,i,k}$ and $y_{P2,i,k}$ with respect to $\tilde{\mathbf{t}}_{P2}$ similarly.

2.2. Prior Specification

To complete the model, we specify independent, conjugate prior beta distributions for π_j , and gamma distributions for $\lambda_{\ell,j,k}$, $k = 1, \dots, K_\ell$, $\ell = T, P1, P2$, and $j = C, C + R$. This prior choice for the λ 's is similar to the independent gamma process (IGP) proposed by Walker and Mallick (1997) for univariate time-to-event analysis using a piecewise exponential model, and results in a conditionally conjugate model structure, thereby greatly facilitating posterior estimation. Let $\text{Beta}(u, v)$ denote a beta distribution with mean $u/(u + v)$, and let $\text{Gamma}(c, d)$ denote a gamma distribution with mean c/d . The explicit prior specification is

$$\pi_j \sim \text{Beta}[a\pi^*, a(1 - \pi^*)],$$

$$\lambda_{\ell,j,k} \sim \text{Gam}(r_{\ell,k}, r_{\ell,k}/\lambda_{\ell,k}^*),$$

$$k = 1, \dots, K_\ell + 1, \ell = T, P1, P2, \quad (5)$$

for $j = C, C + R$, where the a 's, π^* 's, r 's and λ^* 's denote pre-specified hyperparameters. To ensure unbiased comparisons, we assign corresponding treatment parameters the same prior distribution, that is, π_C and π_{C+R} are assigned the same beta prior distribution, and $\lambda_{T,C,1}$ and $\lambda_{T,C+R,1}$ are assigned the same gamma prior distribution, etc.

The prior mean of π_j is π^* and the prior effective sample size is a , which is an intuitive measure for the amount of information provided by the prior (Morita et al. 2008). For example, in a beta-binomial model, as in this model, the prior effective sample size (ESS) of a $\text{Beta}(u, v)$ prior distribution is $u + v$. To ensure that the prior does not provide an inappropriate amount of information, the prior ESS should be small, say 1. Similarly, the prior mean of $\lambda_{\ell,j}$ is $\lambda_{\ell,j}^*$ and the prior effective number of events is $\sum_{k=1}^{K_\ell} r_{\ell,k}$, for $k = 1, \dots, K_\ell + 1$, $\ell = T, P1, P2$, and $j = C, C + R$. To ensure that the priors do not provide an inappropriate amount of information, we use default values of $a = 1$ and $r_\ell = r_{\ell,k} = 1/(K_\ell + 1)$, for $k = 1, \dots, K_\ell + 1$ and $\ell = T, P1, P2$. In contrast, specifying values for π^* and the $\lambda_{\ell,j}^*$'s will depend on the context. These values should reflect the physicians expert knowledge, or historical data, when available. We use historical data from a similar patient population with locoregionally recurrent NSCLC treated with reirradiation therapy, possibly with concurrent chemotherapy (McAvoy et al. 2014). We take $\lambda_{\ell,j}^* = \lambda_{\ell,1}^* = \dots = \lambda_{\ell,K_\ell+1}^*$, so that a priori the hazards are constant at magnitudes seen in the historical data, which

is reasonable in this context. Specifically, we specify $\pi^* = 0.15$, $\lambda_T^* = 0.37$, $\lambda_{P1}^* = 0.10$, and $\lambda_{P2}^* = 0.07$.

In the Web supplement, we consider an alternative prior specification that is motivated by the hierarchical Markov gamma process (HMGP) proposed by Nieto-Barajas and Walker (2002). This alternative specification induces dependence between hazard parameters in successive intervals, while retaining a largely conditionally conjugate model structure. The HMGP tends to give less variable estimates for the hazard parameters than the IGP, but it increases computational complexity and does not substantially affect the design's properties (see Web supplement Table 2). Other options have been proposed by Gamerman (1991), Gray (1994), and Arjas and Gasbarra (1994). However, these alternatives do not result in conditionally conjugate structures like the IGP and HMGP priors that we consider. A review of these prior processes is provided in Ibrahim et al. (2001, Section 3). We conduct posterior estimation using a Gibbs sampler and provide the full conditional distributions in the Web supplement. We also provide R software to estimate the model with either prior specification (see supplementary material).

2.3. Partition Specification

The piecewise exponential distributions in the model rely on three partitions, \tilde{t}_T , \tilde{t}_{P1} , and \tilde{t}_{P2} . The sampling methods proposed by either Arjas and Gasbarra (1994) or Demarqui et al. (2008) may be implemented to facilitate posterior estimation of these partitions within the Gibbs sampler. However, these methods are computationally demanding in our design setting, which requires extensive simulation to assess the operating characteristics of the design. Alternatively, these partitions can be pre-specified to provide an adequately flexible probability model. To greatly facilitate calculation of mean utilities in the sequel, we use identical partitions with $K = 12$ two-month intervals that span the 24-month observation period. We denote the shared partition by \tilde{t} , with $\tilde{t}_0 = 0$, $\tilde{t}_1 = 2$, $\tilde{t}_2 = 4$, ..., $\tilde{t}_{12} = 24$, and $\tilde{t}_{13} = \infty$. As a sensitivity analysis, we report simulation results comparing shared partitions with one, two, and four month intervals (see Web supplement Table 4).

3. Utility Function

The utility function, $U(y)$, in (1) must be specified to reflect the clinical desirability of every possible realization of the two time-to-event outcomes, with larger values indicating greater desirability. For this reason, we recommend that $U(y)$ be elicited from the physicians planning the trial. For these outcomes, which are semi-competing risks, it may seem that eliciting $U(y)$ from the physicians is very challenging; however, we provide guidelines below to carry out this critical task in practice. We establish key admissibility constraints for $U(y)$ and propose a class of parametric functions that satisfy these constraints. Relying on this class of parametric functions and a discretization of the response domain, we then demonstrate how to develop a spreadsheet that may be provided to the physician(s) to facilitate utility elicitation. We also provide an example spreadsheet (see supplementary material). When there are multiple physicians, we suggest that each physician select numerical utilities

using the spreadsheet on their own, and then confer with the other physicians until consensus utilities are obtained. We provide detailed guidelines and an illustration below, in the context of the NSCLC trial.

The utility elicitation approach below is similar to that of Thall et al. (2013), who elicited utilities for bivariate time-to-event outcomes in the context of phase II trials. They suggested that the two-dimensional outcome domain first be discretized, then numerical utilities be elicited for all of the resulting elementary events, and finally a smooth utility surface should be fit to the discrete numerical values. In contrast, the approach below relies on a preemptively established class of admissible utility functions so that the physicians need not select numerical values for all elementary events on the partition. Rather, they select numerical values for two interpretable parameters that characterize the underlying utility function and explicitly determine the numerical utilities on the partition. Another important advantage of the elicitation approach given below is that the resulting numerical utilities on the partition must satisfy the admissibility constraints. Moreover, the mean utilities in (1) can be calculated using either the discrete utilities on the partition or the underlying parametric utility function without concern about over-smoothing the elicited values. Lastly, the utility parameters provide a parsimonious basis for a sensitivity analysis to the elicited numerical utility values (see, e.g., Web supplement Table 3).

3.1. Admissibility Constraints

Establishing constraints for $U(y)$ in (1) substantially reduces the set of admissible specifications for the utility. The utility constraints should be established before eliciting the utilities from the physician(s). This can be done in cooperation with the physicians, although a statistician often can intuit these on their own. In the NSCLC clinical setting, it clearly is desirable for a treatment to delay progression, avoid toxicity, and, if a toxicity occurs, delay its onset. Therefore, $U(y)$ should be nondecreasing in both y_{Tox} and y_{Prog} , and it should not be larger for a progression time with a prior toxicity than for the same progression time with no prior toxicity. In addition, progression, which includes death, is less desirable than toxicity. Therefore, the utility of any outcome with a progression at time y should not be larger than the utility of any outcome with a toxicity at time y or later. This subtle, yet important constraint ensures that the utility of an outcome with death at time y is not larger than the utility of an outcome with toxicity at time y or later. Without this constraint, the utility could be specified so that an early death would be preferable to a toxicity later on, which is not reasonable. Following the above logic and defining $U(\text{No Tox}, Y_{\text{Prog}} = y) = U(Y_{\text{Tox}} = y, Y_{\text{Prog}} = y)$, the utility should satisfy the following constraints:

$$U(y, y') \leq U(y, y'') \leq U(y', y''), \text{ for } y \leq y' \leq y'' \quad (6)$$

After establishing these utility admissibility constraints, we find it helpful to develop a flexible class of parametric functions satisfying these constraints that can be used as a basis for elicitation. For example, a flexible class of utility functions that satisfy

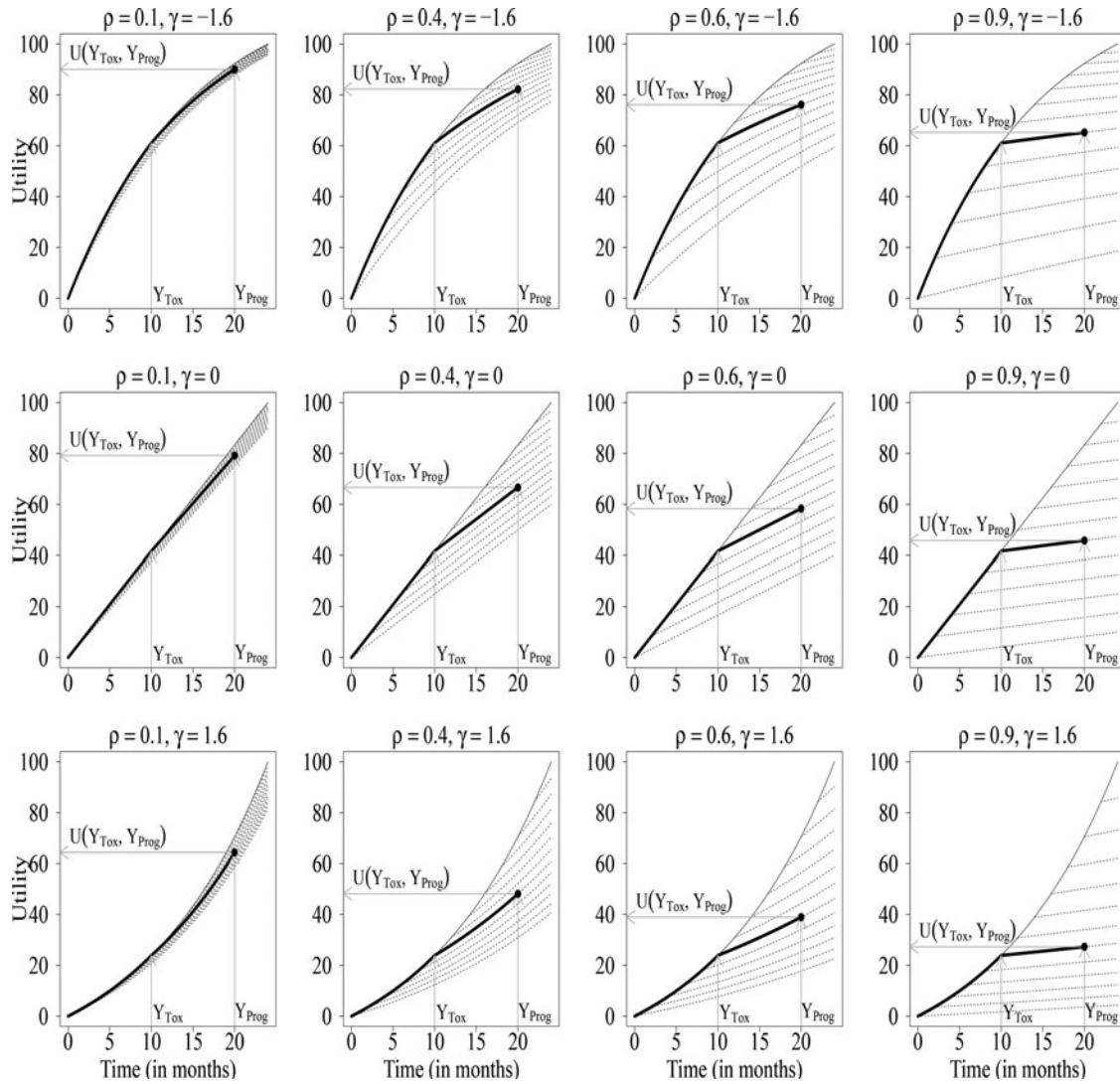


Figure 2. The proposed utility function for various specifications of the toxicity discount parameter, ρ , and the temporal preference parameter, γ , with $\tau = 24$ months. The top line in each panel depicts how the utility increases given no prior toxicity, whereas the lower dashed lines depict how the utility increases following a toxicity at the time-point of departure from the top line. An example case is depicted in each panel where ($Y_{\text{Tox}} = 10$, $Y_{\text{Prog}} = 20$).

the constraints in (6) is defined as follows:

$$U(Y_{\text{Tox}}, Y_{\text{Prog}}) = \begin{cases} \min \{ 100 [Y_{\text{Prog}} - \rho(Y_{\text{Prog}} - Y_{\text{Tox}})] / \tau, 100 \} & \text{if } \gamma = 0, \\ \min \left\{ 100 \left(\frac{\exp\{\gamma [Y_{\text{Prog}} - \rho(Y_{\text{Prog}} - Y_{\text{Tox}})] / \tau\} - 1}{\exp\{\gamma\} - 1} \right), 100 \right\} & \text{otherwise,} \end{cases} \quad (7)$$

where τ is the upper bound on the observation period, $\rho \in [0, 1]$ is the *toxicity discount parameter*, and γ is the *temporal preference parameter*. Figure 2 illustrates $U(y)$ defined in (7) for various specifications of ρ and γ , with $\tau = 24$ months. To determine $U(Y_{\text{Tox}}, Y_{\text{Prog}})$ using the figure for a particular outcome, follow the top line until Y_{Tox} , and then follow parallel to the dashed lines until Y_{Prog} ; projecting the terminal point to the y-axis determines the numerical utility of that outcome. An example is depicted in each panel where ($Y_{\text{Tox}} = 10$, $Y_{\text{Prog}} = 20$). As illustrated by Figure 2, γ controls how rapidly $U(y)$ increases in time and ρ controls the rate at which $U(y)$ increases following a toxicity.

In the sequel, let $y \leq y' \leq y''$. When $\rho = 0$, $U(y, y'') = U(y', y'') = U(y, y')$, that is, toxicity does not affect the utility. When $\rho = 1$, $U(y, y) = U(y, y') = U(y, y'')$, that is, the utility is completely determined by the earliest occurrence of either type of event, and thus, the time to progression after toxicity does not affect the utility. When $0 < \rho < 1$, $U(y)$ increases in y_{Prog} at a diminished rate following a toxicity, so that ρ quantifies the diminished quality of life a patient experiences following toxicity. A formal interpretation for ρ arises from the identity, $U(y, (1 - \rho)^{-1}(y' - y) + y) = U(y', y')$, which implies that $(1 - \rho)^{-1}$ is the factor of additional time that a

patient must be alive and progression-free following a toxicity at time y for the patient's outcome to be equally desirable as that of a progression at time y' with no prior toxicity. For example, using (7), the outcome $(Y_{\text{Tox}} = 0, Y_{\text{Prog}} = y'(1 - \rho)^{-1})$ has the same utility as the outcome (No Tox, $Y_{\text{Prog}} = y'$). From a treatment comparison perspective, if the physician knows that $C + R$ will result in toxicity and C will not, ρ dictates how long $C + R$ would need to delay progression for it to be clinically preferable to C .

Turning to γ in the functional form for $U(y)$ in (7), when $\gamma = 0$, $U(y)$ increases linearly in Y_{Tox} and Y_{Prog} , which implies that all regions of the observation period are equally important. For this reason, we suggest using $\gamma = 0$ as a default value. When $\gamma < 0$, $U(y)$ increases more rapidly during the earlier region of the observation period, which implies that delaying early events is more important, and conversely for $\gamma > 0$, $U(y)$ increases more slowly during the earlier region of the observation period, which implies that delaying early events is less important. A formal interpretation for γ arises from the identity,

$$\exp \left\{ \gamma \left(\frac{\epsilon}{\tau} \right) \right\} = \frac{U(y - \epsilon, y' + \epsilon) - U(y - \epsilon, y')}{U(y - \epsilon, y') - U(y - \epsilon, y' - \epsilon)},$$

which implies that γ controls the relative change in $U(y)$ over successive ϵ -length intervals in the observation period. For example, when $\tau = 24$ months and $\epsilon = 2$ months, the utility increase over successive 2 month intervals in the observation period changes by the factor $\exp\{\frac{\gamma}{12}\}$. From a treatment comparison perspective, γ reflects whether a progression delay from 1 to 3 months has greater clinical importance than a progression delay from 3 to 5 months. Interestingly, some "objective" measures of efficacy arise from certain specifications of $U(y)$ in (7). For example, when $\gamma = 0$ and $\rho = 0$, $U(y) \propto y_{\text{Prog}}$, and mean utility is thus proportional to mean PFS. Similarly, when $\gamma = 0$ and $\rho = 1$, mean utility is proportional to mean progression-and-toxicity-free survival (PTFS).

3.2. Utility Elicitation

Using the functional form for $U(y)$ in (7), we elicited ρ and γ from the physicians as follows: (a) We explained the meaning of ρ and γ to each physician and asked them to select numerical values for ρ and γ individually, and (b) we then asked them to confer to obtain consensus values. To accomplish (a) and (b), we provided each physician with a spreadsheet that takes numerical values for ρ and γ , and populates a discretized outcome domain with numerical utilities based on $U(y)$ (see supplementary material). We asked them to select numerical values using the spreadsheet that reflected their clinical experience treating recurrent NSCLC.

Before creating the spreadsheet, in cooperation with the physicians, we selected an observation period of $[0, \tau = 24]$ months and, following the advice of Thall et al. (2013), we discretized the observation period using the shared partition \tilde{t} defined in Section 2.3. We denote the two-month intervals, which were selected in cooperation with DG and SM, as $I_1 = (0, 2], I_2 = (2, 4], \dots, I_{12} = (22, 24]$. Using these intervals, the semi-competing risks outcome has the elementary events, $(Y_{\text{Tox}} \in I_k, Y_{\text{Prog}} \in I_{k'})$, and (No Tox, $Y_{\text{Prog}} \in I_{k'}$), for $k \leq k' =$

$1, \dots, 12$, and (No Tox, No Prog), where "No Tox" denotes the outcome that no toxicity is observed prior to progression during the 24-month observation period and "No Prog" denotes the outcome that no progression is observed during the 24 month observation period. We found that discretizing time in this way greatly helps the physicians understand the utility, and, as we discuss below, it also facilitates computation of mean utilities in (1).

To translate $U(y)$ in (7) to numerical utilities on the partition, we define a discrete version of the utility function, $U_{\text{Discrete}}(y)$, as follows:

$$\begin{aligned} U_{k,k} &= U_{\text{Discrete}}(Y_{\text{Tox}} \in I_k, Y_{\text{Prog}} \in I_k) \\ &= U(\tilde{t}_{k-1}, 0.5[\tilde{t}_k + \tilde{t}_{k-1}]), \\ U_{k,k'} &= U_{\text{Discrete}}(Y_{\text{Tox}} \in I_k, Y_{\text{Prog}} \in I_{k'}) \\ &= U(0.5[\tilde{t}_k + \tilde{t}_{k-1}], 0.5[\tilde{t}_{k'} + \tilde{t}_{k'-1}]), \\ U_{K+1,k} &= U_{\text{Discrete}}(\text{No Tox}, Y_{\text{Prog}} \in I_{k'}) \\ &= U(0.5[\tilde{t}_{k'} + \tilde{t}_{k'-1}], 0.5[\tilde{t}_{k'} + \tilde{t}_{k'-1}]), \\ U_{k,K+1} &= U_{\text{Discrete}}(Y_{\text{Tox}} \in I_k, \text{No Prog}) \\ &= U(0.5[\tilde{t}_k + \tilde{t}_{k-1}], \tilde{t}_K + 0.5[\tilde{t}_1 + \tilde{t}_0]), \text{ and} \\ U_{K+1,K+1} &= U_{\text{Discrete}}(\text{No Tox}, \text{No Prog}) \\ &= U(\tilde{t}_K + 0.5[\tilde{t}_1 + \tilde{t}_0], \tilde{t}_K + 0.5[\tilde{t}_1 + \tilde{t}_0]), \end{aligned}$$

for $k < k' = 1, \dots, K$. To restrict these discrete numerical utilities to the domain $[0, 100]$, we subtracted the minimum, $U(\tilde{t}_0, 0.5[\tilde{t}_1 + \tilde{t}_0])$, from each $U_{k,k'}$ defined above, and then divided by the range, $U(\tilde{t}_K + 0.5[\tilde{t}_1 + \tilde{t}_0], \tilde{t}_K + 0.5[\tilde{t}_1 + \tilde{t}_0]) - U(\tilde{t}_0, 0.5[\tilde{t}_1 + \tilde{t}_0])$. While any compact utility domain could be used, $[0, 100]$ works well in practice when communicating with the physicians, see Thall and Nguyen (2012); Thall et al. (2013). The above translation strategy ensures that $U_{k,k'} \leq U_{k',k'}$, $U_{k,k} \leq U_{k,k'}$, $U_{k,k} \leq U_{K+1,k}$, and $U_{K+1,k} \leq U_{k,k'}$, for $k < k' < k'' = 1, \dots, K$, so the discrete utilities satisfy the previously established admissibility constraints. Crucially, the physicians need only select numerical values for ρ and γ , rather than all $0.5(K + 1) \times (K + 2) + K = 103$ numerical utilities on the partition. In our context, after examining several pairs of (ρ, γ) values and their resulting utilities, the consensus utilities from DG and SM use $\rho = 0.6$ and $\gamma = 0$. The elicited $U(y)$ is illustrated in Figure 2, and the corresponding numerical utilities on the partition are reported in Table 7 of the online supplementary material.

3.3. Mean Utility Calculation

Given the elicited $U(y)$ and the Bayesian model for $p(y|\text{trt } j)$ in Section 2, we discuss how to calculate mean utilities in (1). The mean utility of treatment j , \bar{U}_j , is a function of the model parameters, $(\pi_j, \lambda_{T,j}, \lambda_{P1,j}, \lambda_{P2,j})$, so \bar{U}_j has induced prior and posterior distributions. Because we use G draws from a Gibbs sampler for estimation, to obtain draws from the posterior distribution of \bar{U}_j , we calculate $\bar{U}_j^{(g)}$ at each sampled

value of $(\pi_j^{(g)}, \lambda_{T,j}^{(g)}, \lambda_{P1,j}^{(g)}, \lambda_{P2,j}^{(g)})$ from the Gibbs sampler, for $g = 1, \dots, G$.

The samples from the posterior distribution of \bar{U}_j can be calculated based on either the elicited $U(\mathbf{y})$ or the discrete numerical utilities on the partition. Using the elicited $U(\mathbf{y})$, these samples are given formally as

$$\begin{aligned} \bar{U}_j^{(g)} = & \pi_j^{(g)} \left[100S(\tau | \lambda_{T,j}^{(g)}; \tilde{\mathbf{t}}_T) \right. \\ & \left. + \int_0^\tau \int_u^\tau U(u, v) f(u | \lambda_{T,j}^{(g)}; \tilde{\mathbf{t}}_T) \frac{f(v | \lambda_{P1,j}^{(g)}; \tilde{\mathbf{t}}_{P1})}{S(u | \lambda_{P1,j}^{(g)}; \tilde{\mathbf{t}}_{P1})} dv du \right] \\ & + (1 - \pi_j^{(g)}) \left[100S(\tau | \lambda_{P2,j}^{(g)}; \tilde{\mathbf{t}}_{P2}) \right. \\ & \left. + \int_0^\tau U(v, v) f(v | \lambda_{P2,j}^{(g)}; \tilde{\mathbf{t}}_{P2}) dv \right], \end{aligned} \quad (8)$$

where $f(t) = h(t)S(t)$, for $j = C, C + R$ and $g = 1, \dots, G$. We considered using (8) as the basis for our comparative testing criterion discussed below, however doing so requires a numerical integration routine at each iteration of the Gibbs sampler, and we found this to be too computationally expensive. One evaluation of (8) using numerical integration takes about 4 s in R, so this approach is too computationally expensive for post-processing the G posterior draws from the Gibbs sampler. We instead rely on the elicited numerical utilities on the partition, which provide an excellent approximation to (8) and greatly facilitate computation.

Denote the vector of elicited numerical utilities on the partition by \mathbf{U} , using any convenient ordering. For treatment $j = C, C + R$, we denote

$$\begin{aligned} p_{j,k,k'} &= \Pr\{(Y_{\text{Tox}} \in I_k, Y_{\text{Prog}} \in I_{k'}) | \text{trt } j\}, \\ p_{j,K+1,k'} &= \Pr\{(\text{No Tox}, Y_{\text{Prog}} \in I_{k'}) | \text{trt } j\}, \\ &\quad \text{for } k \leq k' = 1, \dots, K, \text{ and} \\ p_{j,K+1,K+1} &= \Pr\{(\text{No Tox}, \text{No Prog}) | \text{trt } j\}. \end{aligned}$$

Denote the vector of these probabilities with the same ordering as \mathbf{U} by \mathbf{p}_j . Using this notation, given \mathbf{U} and \mathbf{p}_j , the mean utility of treatment j is simply

$$\bar{U}_j = \sum_{k=1}^{K+1} \sum_{k'=k}^{K+1} U_{k,k'} p_{j,k,k'} = \mathbf{U}' \mathbf{p}_j.$$

To facilitate computation of \mathbf{p}_j , we use the same partition $\tilde{\mathbf{t}}$ for all three piecewise exponential distributions in the model defined by (2), and for translating $U(\mathbf{y})$ to \mathbf{U} using the previously described strategy. Given $(\pi_j, \lambda_{T,j}, \lambda_{P1,j}, \lambda_{P2,j})$, the probabilities for the elementary events on the partition are given formally as

$$\begin{aligned} p_{j,k,k} &= \pi_j \left\{ S(\tilde{t}_{k-1} | \lambda_{T,j}; \tilde{\mathbf{t}}) - S(\tilde{t}_k | \lambda_{T,j}; \tilde{\mathbf{t}}) \right. \\ &\quad \left. - S(\tilde{t}_{k-1} | \lambda_{T,j}; \tilde{\mathbf{t}}) \left[\frac{\lambda_{T,j,k}}{\lambda_{T,j,k} - \lambda_{P1,j,k}} \right] \right\} \end{aligned}$$

$$\begin{aligned} & \times \left[\frac{S(\tilde{t}_k | \lambda_{P1,j}; \tilde{\mathbf{t}})}{S(\tilde{t}_{k-1} | \lambda_{P1,j}; \tilde{\mathbf{t}})} - \frac{S(\tilde{t}_k | \lambda_{T,j}; \tilde{\mathbf{t}})}{S(\tilde{t}_{k-1} | \lambda_{T,j}; \tilde{\mathbf{t}})} \right] \Big\}, \\ p_{j,k,k'} &= \pi_j \left[S(\tilde{t}_{k-1} | \lambda_{P1,j}; \tilde{\mathbf{t}}) - S(\tilde{t}_k | \lambda_{P1,j}; \tilde{\mathbf{t}}) \right] \\ & \times \left[\frac{\lambda_{T,j,k}}{\lambda_{T,j,k} - \lambda_{P1,j,k}} \right] \\ & \times \left[\frac{S(\tilde{t}_{k-1} | \lambda_{T,j}; \tilde{\mathbf{t}})}{S(\tilde{t}_{k-1} | \lambda_{P1,j}; \tilde{\mathbf{t}})} - \frac{S(\tilde{t}_k | \lambda_{T,j}; \tilde{\mathbf{t}})}{S(\tilde{t}_k | \lambda_{P1,j}; \tilde{\mathbf{t}})} \right], \end{aligned}$$

$$\begin{aligned} p_{j,K+1,k'} &= (1 - \pi) \left[S(\tilde{t}_{k-1} | \lambda_{P2,j}; \tilde{\mathbf{t}}) - S(\tilde{t}_k | \lambda_{P2,j}; \tilde{\mathbf{t}}) \right], \text{ and} \\ p_{j,K+1,K+1} &= \pi_j S(\tilde{t}_K | \lambda_{T,j}; \tilde{\mathbf{t}}) + (1 - \pi_j) S(\tilde{t}_K | \lambda_{P2,j}; \tilde{\mathbf{t}}), \end{aligned} \quad (9)$$

for $k < k' = 1, \dots, K$. A detailed derivation of the expressions in (9) is provided in the Web supplement. Plugging $(\pi_j^{(g)}, \lambda_{T,j}^{(g)}, \lambda_{P1,j}^{(g)}, \lambda_{P2,j}^{(g)})$ into (9), we obtain $\bar{U}_j^{(g)} = \mathbf{U}' \mathbf{p}_j^{(g)}$, for $g = 1, \dots, G$ and $j = C, C + R$.

4. Group Sequential Design

For the NSCLC trial, we propose a design with up to two interim tests and one final test, that is, a group sequential procedure (see for example, Jennison and Turnbull 2000). Enrollment is expected to be two patients per month. Due to this logistical constraint and power considerations, we plan a five-year (60 month) trial that will enroll patients until either the trial has been terminated or a maximum sample size of $N_{\max} = 100$ is achieved. We will perform interim tests at 20 and 40 months into the trial, at which points 40 and 80 patients are expected to have been enrolled, respectively. Our comparative test criteria are as follows, wherein we use t to denote the proportion of the trial's maximum duration that has passed at the time of the analysis, \mathbf{D} to denote the observed data at any point in the trial, and $\eta_{\text{Tox},j} = \Pr(Y_{\text{Tox}} < 24 | j)$ to denote the probability of toxicity within the observation period, for treatment $j = C + R, C$. For the probability model in (2), $\eta_{\text{Tox},j} = \pi_j [1 - S(\tau | \lambda_{T,j}; \tilde{\mathbf{t}})]$. Given a test cutoff $p_{\text{cut}}(t)$, and maximum acceptable toxicity probability during the observation period $\bar{\eta}_{\text{Tox}}$, if

$$\Pr(\bar{U}_C > \bar{U}_{C+R} \text{ or } \eta_{\text{Tox},C+R} > \bar{\eta}_{\text{Tox}} | \mathbf{D}) > p_{\text{cut}}(t), \quad (10)$$

then we terminate the trial and conclude that C is superior to $C + R$. If

$$\Pr(\bar{U}_{C+R} > \bar{U}_C \text{ and } \eta_{\text{Tox},C+R} < \bar{\eta}_{\text{Tox}} | \mathbf{D}) > p_{\text{cut}}(t), \quad (11)$$

then we terminate the trial and conclude that $C + R$ is superior to C . If neither (10) nor (11) holds, then there is not sufficient evidence in the current data to conclude that either treatment is superior and we continue the trial, up to the final analysis. Given the set of $\eta_{\text{Tox},j}^{(g)}$'s and $\bar{U}_j^{(g)}$'s from the Gibbs sampler, where the $\bar{U}_j^{(g)}$'s are calculated using the methods in Section 3.3, we estimate the posterior probabilities in (10) and (11) empirically from the posterior sample. Because $U(\mathbf{y})$ defines the desirability of all possible outcomes, given that $\eta_{\text{Tox},C+R} < \bar{\eta}_{\text{Tox}}$, the decision rules are based on the idea that, if $\bar{U}_C > \bar{U}_{C+R}$, then on average C will result in clinically superior outcomes compared to $C + R$, and conversely.

Even though we rely on Bayesian methods, it is important to ensure that the proposed method controls Type I error and has adequate power at the planned sample size for identifying the anticipated treatment differences. To control Type I error, and account for the practical issue that the interim looks may not occur at planned times or sample sizes, we rely on an α -spending function (Lan and Demets 1983). Specifically, we use the α -spending function αt^3 so that our method spends 4%, 26% and 70% of the Type I error at the first interim, second interim and final analysis, respectively. We use simulation to determine $p_{\text{cut}}(t)$ in (10) and (11) at each analysis so that Type I error is spent in this manner. That is, the test cut-off $p_{\text{cut}}(t)$ varies with the α -spending function.

Because we are monitoring two outcomes that are semi-competing risks, treatment differences are more complex than for a univariate outcome. In this context, treatment differences are defined with respect to the joint distribution of $(Y_{\text{Tox}}, Y_{\text{Prog}})$ for each treatment. For example, the irradiation component of $C + R$ may delay progression compared to C , but simultaneously cause additional late-onset toxicities. To elucidate treatment differences, we focus on four interpretable measures of these joint distributions, (a) $\eta_{\text{Tox},j}$ = the probabilities of toxicity during the observation period, (b) $T50_j$ = the median times to toxicity, (c) $\eta_{\text{Prog},j}$ = the probabilities of progression during the observation period, and (d) $P50_j$ = the median times to progression, for $j = C, C + R$. To assess power, we also use simulation, wherein we iteratively generate data for each treatment arm from joint distributions of $(Y_{\text{Tox}}, Y_{\text{Prog}})$ that we specify to exhibit a plausible treatment difference in the NSCLC setting, and we calculate the proportion of simulation runs that lead to each conclusion. Because it is challenging to specify joint distributions of $(Y_{\text{Tox}}, Y_{\text{Prog}})$ for each treatment that exhibit plausible differences, we provide guidelines below.

We recommend generating data from the same joint distribution for the standard of care, that is, treatment C , throughout the simulation study. This joint distribution should be based on historical data, when available, and the clinician’s expertise. We use the structure of the probability model in (2) to specify the joint distribution for C . For the NSCLC trial, we specify $\pi_C, h_T(t|C), h_{P1}(t|C),$ and $h_{P2}(t|C)$ to reflect the historical data reported by McAvoy et al. (2014). Specifically, we specify $\pi_C = 0.15$ and hazard functions that are illustrated in Figure 3, along with

the induced distributions for the times to toxicity and progression. This joint distribution has $\eta_{\text{Tox},C} = 0.15$, where $T50_C = 3$ months, and $\eta_{\text{Prog},C} = 0.79$, where $P50_C = 7.3$ months. Moreover, the hazard for progression after a toxicity is greater during the initial 18 months, but equivalent thereafter. The analytical definitions for the hazard functions and the time to progression distribution are provided in the Web Supplement. We emphasize that the hazard functions are not piecewise constant, so the joint distribution we use to generate data for the simulation study deviates from the underlying probability model that we use for posterior inference.

In contrast, in the simulation study we recommend generating data from various plausible joint distributions for the experimental therapy, that is, treatment $C + R$. To determine a plausible joint distribution for $C + R$, we asked DG and SM to hypothesize values for $\eta_{\text{Tox},C+R}$ and $P50_{C+R}$, which were 0.15 and 14 months, respectively. For the NSCLC trial, we consider joint distributions for $C + R$ that exhibit a range of $\eta_{\text{Tox},C+R}$ around 0.15 and $P50_{C+R}$ around 14 months. To specify these joint distributions, it is practical, and reasonable in this context, to use a proportional-hazards (PH) model with baseline hazard functions defined above for C . As a sensitivity assessment, we also generate data from joint distributions that are specified using an accelerated failure time (AFT) model with baseline hazard functions defined above for C . To further illustrate the robustness of the proposed method, we generate data for each treatment arm from joint distributions that are specified using a much different structure than the underlying model in (2). We provide further details about the specification of these joint distributions for $C + R$ when we report the results of an investigation below. In general, a simulation study should be used to assess the proposed method’s power for the anticipated treatment difference, and then to decide whether the trial should be conducted at the planned sample size. If power is inadequate, the statistician should not recommend running the trial at the planned sample size.

4.1. Simulation Comparator

In our simulation study, we compare the proposed method to one that is based on separate tests for safety and efficacy. Specifically, the comparator assesses efficacy using a log-rank test for

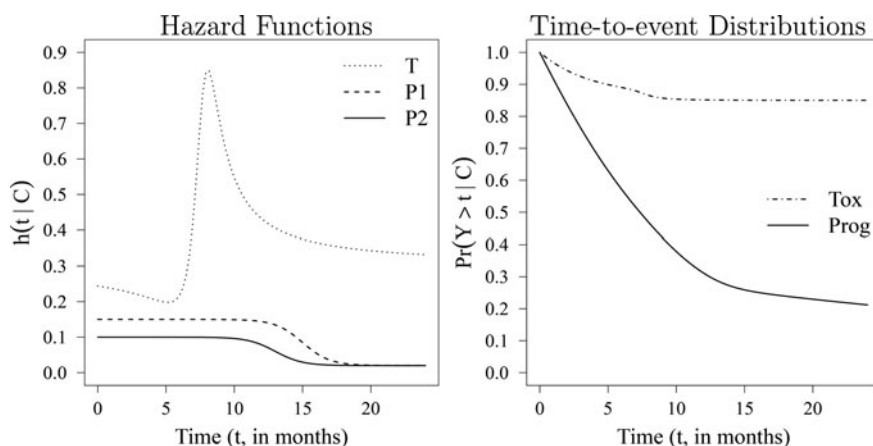


Figure 3. True hazard functions for C in our simulation study, along with the induced distributions of time to toxicity (“Tox”) and progression (“Prog”) when $\pi_C = 0.15$. T denotes the hazard for toxicity given that a toxicity will occur prior to progression, and $P1$ ($P2$) denotes the hazard for progression given prior (no prior) toxicity.

whether the PFS distributions differ between the two treatment arms, and assesses safety using a proportion test of the hypotheses,

$$H_0 : \eta_{\text{Tox},C+R} = \bar{\eta}_{\text{Tox}} \quad \text{versus} \\ H_A : \eta_{\text{Tox},C+R} < \bar{\eta}_{\text{Tox}} \quad \text{or} \quad \eta_{\text{Tox},C+R} > \bar{\eta}_{\text{Tox}}.$$

The above safety test only uses the data from the $C + R$ arm, and assesses whether $C + R$ is safe, which we define as $\eta_{\text{Tox},C+R} < \bar{\eta}_{\text{Tox}}$, or unsafe, which we define as $\eta_{\text{Tox},C+R} > \bar{\eta}_{\text{Tox}}$. For the safety test, we define a binary indicator, x_i , as follows. If the i th patient has 24 months follow-up and had a toxicity prior to progression, we define $x_i = 1$. If the i th patient has 24 months follow-up and did not have a toxicity prior to progression, we define $x_i = 0$. If the i th patient does not have 24-months follow-up, we consider x_i not evaluable. We conduct the safety test using the evaluable x_i 's in the $C + R$ arm.

The decision criteria for the comparator are as follows. Let Z_{Tox} denote the safety test statistic, Z_{Prog} denote the log-rank test statistic, and $c_{\text{Tox}}(t)$ and $c_{\text{Prog}}(t)$ denote prespecified cut-offs for an analysis at time t . If either $Z_{\text{Tox}} > c_{\text{Tox}}(t)$ or $Z_{\text{Prog}} > c_{\text{Prog}}(t)$, we conclude that C is superior to $C + R$. If both $Z_{\text{Tox}} < -c_{\text{Tox}}(t)$ and $Z_{\text{Prog}} < -c_{\text{Prog}}(t)$, we conclude that $C + R$ is superior to C . Otherwise, the trial continues until the next analysis, unless the maximum sample size has been achieved, in which case the trial is inconclusive. In contrast to the proposed method, the comparator requires different cutoffs for the safety and efficacy tests. Following Jennison and Turnbull (2000), we specify O'Brien & Fleming cut-offs that require greater evidence to stop the trial at earlier analyses and control Type I error at the 0.10 level, like the proposed method. Using this decision criteria, the comparator will select $C + R$ when there is sufficient evidence that $C + R$ is more efficacious than C and $C + R$ is safe, and it will select C when either C is more efficacious than $C + R$ or $C + R$ is unsafe. Hence, the comparator does not account for the relative safety of the two regimes, and does not make explicit whether an efficacy improvement combined with a decline in safety is favorable. In contrast, the proposed method will select $C + R$ when there is sufficient evidence that $C + R$ offers a favorable tradeoff for the two outcomes compared to C and $C + R$ is safe, and it will select C when either C offers a favorable tradeoff for the two outcomes compared to $C + R$ or $C + R$ is unsafe.

4.2. Simulation Conduct

In each simulation run, we assign half of the $N_{\text{max}} = 100$ patients to each treatment group, and generate potential event times from the true model for each patient's assigned treatment. We use inversion sampling to generate data, which is a well-known technique for generating data from non-uniform distributions by inverting the cumulative distribution function (Devroye 1986). To determine the observed outcomes at each analysis, we distinguish between *trial time*, defined as the time in months since the start of the trial, and an individual patient's *follow-up time*, defined as the time in months since the patient initiated treatment. We assume that one patient is assigned to each treatment at the beginning of every month, so that the first two patients are enrolled at trial time 0, the second two at trial time 1, the third two at trial time 2, and so on until the last two patients

are enrolled 49 months after the start of the trial. We perform the two interim analyses at 20 and 40 months, and a final analysis at $T_{\text{max}} = 60$ months. Therefore, at the first interim analysis, 40 of the planned 100 patients are enrolled, wherein the first two enrollees have accrued a maximum of 20 months follow-up, the second pair a maximum of 19 months follow-up, and so on. These logistical calculations for the second interim and final analyses follow similarly. For example, at the second interim analysis, 80 of the planned 100 patients are enrolled, and, at the final analysis, the last two patients enrolled have accrued a maximum of 11 months of follow-up.

Since the potential events can only be observed if they occur during the presently accrued follow-up, in our simulation study, the individual follow-up times at each analysis are the administrative right-censoring times. We assume that this is the only censoring mechanism. In the actual trial, we anticipate very few losses to follow-up for other reasons. Combining the potential event times with the logistical calculations discussed above, we determine the observed data at each analysis, that is, $\mathbf{D} = (\mathbf{y}_{\text{Tox}}, \mathbf{y}_{\text{Prog}}, \boldsymbol{\delta}_{\text{Tox}}, \boldsymbol{\delta}_{\text{Prog}}, \mathbf{z})$. Using the observed data, we apply the stopping criteria for the proposed method and comparator, and thereby determine each method's stopping time and decision for the run. By iterating this entire process, we compare the operating characteristics of the two designs. One run takes about 30 s in R. We used HTCondor, high-performance computing software, to conduct runs in parallel across 200 computational nodes.

4.3. Simulation Results

We set the maximum acceptable toxicity probability within the observation period at $\bar{\eta}_{\text{Tox}} = 0.4$, which is the value that will be used in the actual trial. For the proposed method, we estimate the posterior using the probability model described in Section 2, and calculate mean utilities using the computationally efficient method described in Section 3.3, based on the elicited utilities with $\rho = 0.6$ and $\gamma = 0$. For our main simulation study, we generate data for $C + R$ from joint distributions that are specified using PH models as follows:

$$h_{\ell}(t|C+R) = h_{\ell}(t|C) \exp\{\beta_{\ell}\}, \quad \text{for } \ell = \{T, P1, P2\}.$$

When we specify $\pi_{C+R} = 0.15$ and $\beta_T = \beta_{P1} = \beta_{P2} = 0$, $C + R$ and C have identical joint distributions of $(Y_{\text{Tox}}, Y_{\text{Prog}})$. By adjusting π_{C+R} , β_T , β_{P1} , and β_{P2} , we can specify a joint distribution for $C + R$ with the desired values of $\eta_{\text{Tox},C+R}$, $T50_{C+R}$, $\eta_{\text{Prog},C+R}$, and $P50_{C+R}$. To further simplify specification, we use the same coefficient, $\beta_P = \beta_{P1} = \beta_{P2}$, to adjust both h_{P1} and h_{P2} for $C + R$. Increasing π_{C+R} causes $\eta_{\text{Tox},C+R}$ to increase, increasing β_T causes $T50_{C+R}$ to increase, and increasing β_P causes both $\eta_{\text{Prog},C+R}$ and $P50_{C+R}$ to increase. We compare the two methods for different sets of π_{C+R} , β_T and β_P that result in joint distributions for $C + R$ with a range of plausible values for $\eta_{\text{Tox},C+R}$, $T50_{C+R}$, $\eta_{\text{Prog},C+R}$, and $P50_{C+R}$. We focus on scenarios where $\eta_{\text{Tox},C+R}$ ranges from 0.05 to 0.45, and $P50_{C+R}$ ranges from 5 months to 15 months.

Before presenting the simulation results, we emphasize that in this semi-competing risks context the conventional notion of power is inadequate as it relies on a one-dimensional treatment effect. Here, the treatments may differ for both toxicity and

Table 1. Main simulation study results. $\Delta_U = \bar{U}_{C+R} - \bar{U}_C$ is the mean utility difference, where $\Delta_U > 0$ indicates $C + R$ is more desirable than C ; η_{Tox} and η_{Prog} are the probabilities of toxicity and progression during the 24 month observation period for $C + R$, where $\eta_{Tox} = 0.15$ and $\eta_{Prog} = 0.79$ for C throughout; $T50$ and $P50$ are the median times to toxicity and progression for $C + R$, where $T50 = 3.0$ and $P50 = 7.3$ for C throughout; \bar{N} and \overline{Dur} denote mean sample size and mean duration in months; and " $C + R$ " (" C ") denotes the proportion of runs where the method concluded that $C + R$ (C) is superior. The maximum allowable toxicity probability during the 24 month observation period is $\bar{\eta}_{Tox} = 0.4$.

Scenario	Specifications					Proposed			Comparator		
	Δ_U	η_{Tox}	$T50$	η_{Prog}	$P50$	$C + R$	C	$\bar{N} (\overline{Dur})$	$C + R$	C	$\bar{N} (\overline{Dur})$
1.1	1.4	0.05	3.0	0.78	7.1	0.08	0.03	99.2 (59.2)	0.05	0.06	99.4 (59.4)
1.2	0.0	0.15	3.0	0.79	7.3	0.05	0.05	99.6 (59.6)	0.04	0.05	99.6 (59.6)
1.3	-1.4	0.25	3.0	0.79	7.6	0.02	0.08	98.8 (59.0)	0.02	0.05	99.6 (59.6)
1.4	-2.9	0.35	3.0	0.80	7.8	0.00	0.18	97.0 (57.4)	0.01	0.06	99.6 (59.6)
1.5	-4.3	0.45	3.0	0.80	8.0	0.00	0.46	92.8 (53.6)	0.00	0.21	99.2 (59.2)
2.1	-10.4	0.05	3.0	0.90	4.8	0.00	0.48	96.0 (56.4)	0.00	0.59	91.8 (52.6)
2.2	-11.0	0.15	3.0	0.90	5.1	0.00	0.52	93.8 (54.4)	0.00	0.58	92.8 (53.4)
2.3	-11.7	0.25	3.0	0.90	5.4	0.00	0.60	90.4 (51.6)	0.00	0.51	94.4 (54.8)
2.4	-12.3	0.35	3.0	0.91	5.7	0.00	0.69	86.8 (48.6)	0.00	0.50	95.2 (55.4)
2.5	-12.9	0.45	3.0	0.91	5.9	0.00	0.85	81.2 (44.0)	0.00	0.55	97.2 (55.4)
3.1	20.9	0.05	3.0	0.55	14.8	0.86	0.00	88.8 (48.8)	0.80	0.00	92.0 (52.0)
3.2	18.2	0.15	3.0	0.56	14.4	0.76	0.00	93.6 (53.6)	0.79	0.00	97.4 (57.4)
3.3	15.5	0.25	3.0	0.56	14.1	0.41	0.00	98.2 (58.2)	0.46	0.00	99.4 (59.4)
3.4	12.7	0.35	3.0	0.57	13.9	0.06	0.01	100.0 (60.0)	0.09	0.02	99.8 (59.8)
3.5	10.0	0.45	3.0	0.57	13.7	0.00	0.17	98.6 (58.6)	0.01	0.17	99.4 (59.4)

progression, and these differences may be in opposite directions for these outcomes. In such a case, it may not be clear which treatment is superior. For example, when $\eta_{Tox,C+R} = 0.05$ versus $\eta_{Tox,C} = 0.15$ and $P50_{C+R} = 15$ versus $P50_C = 7$ months, $C + R$ improves both toxicity and progression compared to C , a so-called "win-win" scenario, so $C + R$ is clearly superior to C . In contrast, when $\eta_{Tox,C+R} = 0.25$ versus $\eta_{Tox,C} = 0.15$ and $P50_{C+R} = 15$ versus $P50_C = 7$ months, $C + R$ improves progression but worsens toxicity compared to C , a so-called "win-lose" scenario, so it is not at all clear which treatment is superior, if either. The proposed method, which is based on mean utilities, offers an explicit solution for this problem, whereas the comparator does not.

The results of our main simulation study are given in Table 1. To describe each scenario in the table, we report $\Delta_U = \bar{U}_{C+R} - \bar{U}_C$, which is the mean utility difference, where $\Delta_U > 0$ indicates $C + R$ provides a favorable tradeoff for the two outcomes compared to C ; η_{Tox} and η_{Prog} , which are the probabilities of toxicity and progression during the 24 month observation period for $C + R$, where $\eta_{Tox} = 0.15$ and $\eta_{Prog} = 0.79$ for C throughout; and $T50$ and $P50$, which are the median times to toxicity and progression for $C + R$, where $T50 = 3.0$ and $P50 = 7.3$ months for C throughout. In each scenario block, that is, 1.1–1.5, 2.1–2.5, and 3.1–3.5, $\eta_{Tox,C+R}$ ranges from 0.05 to 0.45, while $T50_{C+R} = T50_C = 3$ months throughout. In contrast, $\eta_{Prog,C+R}$ ($P50_{C+R}$) is relatively invariant in each block near 0.79, 0.90, and 0.56 (7, 5, and 14 months), respectively. Scenario 1.2 is the null case where the joint distribution of (Y_{Tox}, Y_{Prog}) is identical for $C + R$ and C . We used its results to select the probability cut-offs for the proposed method, so it is based on 25,000 runs, which ensures these cut-offs are accurate to three digits. All other scenarios are based on 2500 runs, which ensures the corresponding power figures are accurate to two digits. By design, both methods control Type I error at the $\alpha = 0.10$ level, and each method erroneously concludes that $C + R$ is superior to C , and C is superior to $C + R$ with probability at most 0.05.

In Scenarios 1.1–1.4, $|\Delta_U|$ is quite small, $\eta_{Tox,C+R} < \bar{\eta}_{Tox}$, and both methods are unlikely to conclude that either treatment is

superior. That said, Scenario 1.1 is a "win-lose" case that slightly favors $C + R$ with $\Delta_U = 1.4$, where $C + R$ is safer but less efficacious than C , and the proposed method is more likely to select $C + R$ whereas the comparator is more likely to select C . Scenarios 1.3 and 1.4 are "win-lose" cases that slightly favor C with Δ_U of -1.4 and -2.9 , where $C + R$ is less safe but more efficacious than C , and the proposed method selects C with probabilities 0.08 and 0.18 compared to 0.05 and 0.06 for the comparator. Scenario 1.5 is a "win-lose" case that favors C where $C + R$ is also too toxic, and the proposed method is far more likely to select C than the comparator, with probability 0.46 versus 0.21. Moreover, the proposed method has consistently has smaller mean sample size and duration than the comparator.

Scenarios 2.1–2.5 increasingly favor C as $\eta_{Tox,C+R}$ increases, with Δ_U between -10.4 and -12.9 , and the proposed method accordingly selects C with probability between 0.48 and 0.85. In contrast, the comparator is insensitive to increases in $\eta_{Tox,C+R}$, selecting C with probability between 0.50 and 0.59. More specifically, in Scenarios 2.1 and 2.2, $C + R$ is at least as safe as C but it is less efficacious than C , and the proposed method is less likely to select C than the comparator with respective probabilities 0.48 and 0.52 compared to 0.59 and 0.58. However, in "lose-lose" Scenarios 2.3–2.5, $C + R$ is less safe and less efficacious than C , and the proposed method is far more likely to select C than the comparator with respective probabilities 0.60, 0.69, and 0.85 compared to 0.51, 0.50, and 0.55.

Scenarios 3.1–3.5 present five cases where $C + R$ has much better efficacy than C but $\eta_{Tox,C+R}$ varies between 0.05 and 0.45. In Scenarios 3.1 and 3.2, $C + R$ is at least as safe as C , both methods are likely to select $C + R$, but the proposed method benefits from a smaller mean sample size and duration than the comparator. Scenario 3.2 reflects the anticipated difference between $C + R$ and C , and the proposed method has adequate power, 0.76, for selecting $C + R$ at the planned sample size. In "win-lose" Scenarios 3.3 and 3.4, $C + R$ is less safe than C but offers a favorable tradeoff as $C + R$ is highly efficacious, whereas in Scenario 3.5 $C + R$ is too toxic despite its large efficacy advantage. These are challenging cases for both methods, as there is low power to

distinguish whether $\eta_{\text{Tox},C+R} < \bar{\eta}_{\text{Tox}}$ or $\eta_{\text{Tox},C+R} > \bar{\eta}_{\text{Tox}}$, and both methods are likely to be inconclusive.

In the online supplementary material, we report results from additional simulation studies, including where $N_{\text{max}} = 200$ with an enrollment rate of 4 patients per month rate. The results show the same general patterns as the main investigation, but with larger power figures. We also report comparisons of the probability model with the IGP prior versus the alternative HMGP prior, utilities with $\rho = 0.1$ and $\rho = 0.9$, and shared partitions with one and four month intervals. The results show that the HMGP prior tends to slightly increase power, but the shared partition negligibly affects the proposed method's operating characteristics. In contrast, for utilities with $\rho = 0.1$ ($\rho = 0.9$) compared to $\rho = 0.6$, the results show that the proposed method is less (more) sensitive to changes in $\eta_{\text{Tox},C+R}$. We also report a simulation study where we generate data for treatment C + R from AFT models, rather than PH models. The results show that the comparator has diminished power, which is not surprising as it relies on the log-rank test, whereas the proposed method is less affected. Finally, we illustrate the flexibility of the proposed method by generating data for both treatments from joint distributions defined using latent outcomes that follow mixture distributions, and thus these joint distributions have a different structure than the underlying probability model for the proposed method.

5. Conclusion

Conventional methods based on separate tests for each clinically important outcome do not reflect the implicit tradeoff between outcomes, so when the treatment affects these outcomes in opposite ways, that is, a “win-lose” scenario, it is not clear which treatment is preferred. The proposed method compares treatments accounting for toxicity and efficacy outcomes via posterior mean utility, which explicitly reflects the physicians' clinical experience with these tradeoffs. The elicited utilities provide a practical basis for transforming complex outcomes, like the two NSCLC semi-competing risk outcomes, into a one-dimensional criterion for comparing treatments. Moreover, the main simulation study in Section 4 shows that, compared to the proposed method, an approach based on separate tests can have much lower power when a treatment provides a modest advantage for both outcomes.

One potential limitation of the proposed design is that follow-up terminates at non-fatal progression. An alternative would be a sequential, multiple assignment, randomized trial (SMART) (see Almirall et al. 2014; Murphy 2003; Murphy and McKay 2004) that has an additional randomization for third-line treatment at disease progression, see, for example, Thall et al. (2000, 2007) or Wang et al. (2012). We chose not to implement a SMART design for practical reasons; the proposed design is already very complex, and third-line treatment options are numerous. Another potential limitation is the piecewise constant hazard assumption, which could be relaxed by assuming continuous hazard models (see, e.g., Sharef et al. 2010) that might provide better model fit to the realized data and greater power than the proposed method. These enhancements are natural extensions to the methods developed in this article, although computation will be a challenge.

Supplementary Material

Web Supplement: This document provides the full conditional distributions for Gibbs sampling, derivations of the elementary event probabilities, simulation details and additional results, and the table of elicited numerical utilities on the partition. (Web-Supplement.pdf)

Software: R software to implement the models discussed in Section 2 using a Gibbs sampler, calculate mean utility discussed in Section 3, and conduct the simulation study detailed in Section 4. (SCRUBB-Design.R)

Utility Elicitation Spreadsheet: Spreadsheet mentioned in Section 3 that was used for utility elicitation. (Utility-Elicitation.xls)

Funding

The work of the first three authors was partially funded by NIH/NCI grant 5-R01-CA083932. Yuan's and Murray's research was partially supported by Award Number R01-CA154591 from the National Cancer Institute.

References

- Almirall, D., Nahum-Shani, I., Sherwood, N., and Murphy, S. (2014), “Introduction to SMART Designs for the Development of Adaptive Interventions: With Application to Weight Loss Research,” *Translational Behavioral Medicine*, 4, 260–274. [22]
- Arjas, E., and Gasbarra, D. (1994), “Nonparametric Bayesian Inference from Right Censored Survival Data, Using the Gibbs Sampler,” *Statistica Sinica*, 4, 505–524. [15]
- Cannistra, S. A. (2004), “The Ethics of Early Stopping Rules: Who is Protecting Whom?” *Journal of Clinical Oncology*, 22, 1542–1545. [12]
- Conlon, A., Taylor, J., and Sargent, D. (2014), “Multi-State Models for Colon Cancer Recurrence and Death With a Cured Fraction,” *Statistics in Medicine*, 33, 1750–1766. [13]
- Demarqui, F. N., Loschi, R. H., and Colosimo, E. A. (2008), “Estimating the Grid of Time-Points for the Piecewise Exponential Model,” *Lifetime Data Analysis*, 14, 333–356. [15]
- Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag. [20]
- Fine, J. P., Jiang, H., and Chappell, R. (2001), “On Semi-Competing Risks Data,” *Biometrika*, 88, 907–919. [12]
- Friedman, M. (1982), “Piecewise Exponential Models for Survival Data With Covariates,” *The Annals of Statistics*, 10, 101–113. [13]
- Gamerman, D. (1991), “Dynamic Bayesian Models for Survival Data,” *Journal of the Royal Statistical Society, Series C*, 40, 63–79. [15]
- Gray, R. J. (1994), “A Bayesian Analysis of Institutional Effects in a Multi-center Cancer Clinical Trial,” *Biometrics*, 50, 244–253. [15]
- Hobbs, B. P., Thall, P. F., and Lin, S. H. (2016), “Bayesian Group Sequential Clinical Trial Design Using Total Toxicity Burden and Progression-Free Survival,” *Journal of the Royal Statistical Society, Series C*, 65, 273–297. [12]
- Holford, T. R. (1976), “Life Tables With Concomitant Information,” *Biometrics*, 32, 587–597. [13]
- Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer. [13,15]
- Jennison, C., and Turnbull, B. W. (2000), *Group Sequential Methods Applications to Clinical Trials*, Boca-Raton, FL: Chapman & Hall/CRC Press. [18,20]
- Lan, K. K. G., and Demets, D. L. (1983), “Discrete Sequential Boundaries for Clinical Trials,” *Biometrika*, 70, 659–663. [19]
- Lee, J., Thall, P. F., Ji, Y., and Müller, P. (2015a), “Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity,” *Journal of the American Statistical Association*, 110, 711–722. [12]
- Lee, K. H., Haneuse, S., Schrag, D., and Dominici, F. (2015b), “Bayesian Semiparametric Analysis of Semicompeting Risks Data: Investigating Hospital Readmission After a Pancreatic Cancer Diagnosis,” *Journal of the Royal Statistical Society, Series C*, 62, 253–273. [13]
- McAvoy, S., Ciura, K., Wei, C., Rineer, J., Liao, Z., Chang, J. Y., Palmer, M. B., Cox, J. D., Komaki, R., and Gomez, D. R. (2014), “Definitive Reirradiation for Locoregionally Recurrent Non-Small Cell Lung Cancer With Proton Beam Therapy or Intensity Modulated Radiation

- Therapy: Predictors of High-Grade Toxicity and Survival Outcomes,” *International Journal of Radiation Oncology, Biology and Physics*, 90, 819–827. [11,12,14,19]
- Morita, S., Thall, P. F., and Müller, P. (2008), “Determining the Effective Sample Size of a Parametric Prior,” *Biometrics*, 64, 595–602. [14]
- Murphy, S. A. (2003), “Optimal Dynamic Treatment Regimes,” *Journal of the Royal Statistical Society, Series B*, 65, 331–355. [22]
- Murphy, S. A., and McKay, J. R. (2004), “Adaptive Treatment Strategies: An Emerging Approach for Improving Treatment Effectiveness,” *Clinical Science*, 12, 7–13. [22]
- Nieto-Barajas, L. E., and Walker, S. G. (2002), “Markov Beta and Gamma Processes for Modelling Hazard Rates,” *Scandinavian Journal of Statistics*, 29, 413–424. [15]
- Peng, L., and Fine, J. P. (2007), “Regression Modeling of Semicompeting Risks Data,” *Biometrics*, 63, 96–108. [12]
- Pilz, L. R., Manegold, C., and Schmid-Bindert, G. (2012), “Statistical Considerations and Endpoints for Clinical Lung Cancer Studies: Can Progression Free Survival (pfs) Substitute Overall Survival (os) as a Valid Endpoint in Clinical Trials for Advanced Non-Small-Cell Lung Cancer?” *Translational Lung Cancer Research*, 1, 26–35. [11]
- Sharef, E., Strawderman, R. L., Ruppert, D., Cowen, M., and Halasyamani, L. (2010), “Bayesian Adaptive B-Spline Estimation in Proportional Hazards Frailty Models,” *Electronic Journal of Statistics*, 4, 606–642. [22]
- Thall, P. F., Millikan, R. E., and Sung, H.-G. (2000), “Evaluating Multiple Treatment Courses in Clinical Trials,” *Statistics in Medicine*, 19, 1011–1028. [12,22]
- Thall, P. F., and Nguyen, H. Q. (2012), “Adaptive Randomization to Improve Utility-Based Dose-Finding With Bivariate Ordinal Outcomes,” *Journal of Biopharmaceutical Statistics*, 22, 785–802. [17]
- Thall, P. F., Nguyen, H. Q., Braun, T. M., and Qazilbash, M. H. (2013), “Using Joint Utilities of the Times to Response and Toxicity to Adaptively Optimize Schedule-Dose Regimes,” *Biometrics*, 69, 673–682. [12,15,17]
- Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007), “Bayesian and Frequentist Two-Stage Treatment Strategies Based on Sequential Failure Times Subject to Interval Censoring,” *Statistics in Medicine*, 26, 4687–4702. [22]
- Walker, S. G., and Mallick, B. K. (1997), “Hierarchical Generalized Linear Models and Frailty Models With Bayesian Nonparametric Mixing,” *Journal of the Royal Statistical Society, Series B*, 59, 845–860. [14]
- Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012), “Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer,” *Journal of the American Statistical Association*, 107, 493–508. [22]
- Xu, J., Kalbfleisch, J. D., and Tai, B. (2010), “Statistical Analysis of Illness-death Processes and Semicompeting Risks Data,” *Biometrics*, 66, 716–725. [13]
- Yuan, Y., and Yin, G. (2009), “Bayesian Dose Finding by Jointly Modelling Toxicity and Efficacy as Time-to-Event Outcomes,” *Journal of the Royal Statistical Society, Series C*, 58, 719–736. [12]
- Zhang, Y., Chen, M.-H., Ibrahim, J. G., Zeng, D., Chen, Q., Pan, Z., and Xue, X. (2014), “Bayesian Gamma Frailty Models for Survival Data With Semi-Competing Risks and Treatment Switching,” *Journal of the Royal Statistical Society, Series B*, 20, 76–105. [13]