ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2016)
65, *Part* 2, *pp.* 273–297

Bayesian group sequential clinical trial design using total toxicity burden and progression-free survival

Brian P. Hobbs, Peter F. Thall and Steven H. Lin

University of Texas MD Anderson Cancer Center, Houston, USA

[Received June 2014. Final revision June 2015]

Summary. Delivering radiation to eradicate a solid tumour while minimizing damage to nearby critical organs remains a challenge. For oesophageal cancer, radiation therapy may damage the heart or lungs, and several qualitatively different, possibly recurrent toxicities that are associated with chemoradiation or surgery may occur, each at two or more possible grades. We describe a Bayesian group sequential clinical trial design, based on total toxicity burden (TTB) and the duration of progression-free survival, for comparing two radiation therapy modalities for oesophageal cancer. Each patient's toxicities are modelled as a multivariate doubly stochastic Poisson point process, with marks identifying toxicity grades. Each grade of each type of toxicity is assigned a severity weight, elicited from clinical oncologists who are familiar with the disease and treatments. TTB is defined as a severity-weighted sum over the different toxicities that may occur up to 12 months from the start of treatment. Latent frailties are used to formulate a multivariate model for all outcomes. Group sequential decision rules are based on posterior mean TTB and progression-free survival time. The design proposed is shown to provide both larger power and smaller mean sample size when compared with a conventional bivariate group sequential design.

Keywords: Bayesian analysis; Co-primary end points; Frailty model; Prior elicitation; Radiation oncology; Sequentially adaptive design; Utilities

1. Introduction

Oesophageal cancer affects over 17000 people per year in the USA, with 5-year survival rates between 20% and 35%. Standard of care is neoadjuvant chemoradiation, consisting of radiation therapy (RT) and concurrent chemotherapy, possibly followed by surgery. The decision of whether surgery may be performed is made adaptively on the basis of early chemoradiation outcomes. Since the oesophagus is nestled between critical organs, dosimetric RT planning is challenging. An ideal RT plan delivers sufficient dose to the tumour while minimizing or avoiding radiation exposure to the heart anteriorly, the spinal cord posteriorly and the lungs on either side.

One recently developed X-ray modality, intensity-modulated radiation therapy (IMRT), uses a computer-controlled multileaf collimator to block the paths of five beams of charged high energy photons partially. The approach enables flexibility for controlling the extent of intensity of radiation over the irradiated volume, and thereby has the potential to irradiate the targeted tumour volume effectively, while limiting radiation exposure to critical organs surrounding the tumour. However, because X-ray beams deposit energy along a path that passes through the targeted volume and beyond, IMRT is incapable of sparing healthy tissues directly behind the tumour. Another modality, proton beam therapy (PBT), delivers radiation by using a beam of

Address for correspondence: Brian P. Hobbs, Department of Biostatistics, University of Texas MD Anderson Cancer Center, Unit 1411, 1515 Holcombe Boulevard, Houston, TX 77030, USA.
E-mail: bphobbs@mdanderson.org

 0035–9254/16/65273

charged protons that have been accelerated through a cyclotron to high energy levels. Unlike photons, protons have a limited range and can be modulated to deposit their maximum intensity at the tumour site, thus sparing the surrounding organs. For example, Zhang *et al*. (2008) have demonstrated that PBT results in better sparing of the lung compared with IMRT.

In this paper, we describe a design that is being used to conduct a randomized, group sequential clinical trial for comparing radiation modalities for stage II–III oesophageal cancer (University of Texas MD Anderson Cancer Center, 2015). The purpose of this trial is to determine whether PBT's dosimetric advantages over IMRT translate into meaningful improvements in clinical outcomes, primarily reduced toxicity and prolonged progression-free survival (PFS), defined as the time to disease progression or recurrence, or death, from the start of RT. Patients with oesophageal cancer undergoing this regime, called 'trimodality therapy', are at risk of several qualitatively different toxicities. These may occur at random times and at varying levels of severity, and some may occur more than once. Toxicities not only impact the patient's quality of life but also may decrease the patient's ability to undergo surgery and thus increase the risk of recurrence. The surgeon's decision of whether a patient may undergo surgery includes consideration of toxicities that have occurred with the chemoradiation. Patients who do undergo surgery are at risk of post-operative complications (POCs), which may be exacerbated by earlier toxicities from the chemoradiation. Although each particular toxicity is unlikely, all are potentially life threatening, necessitating a design with rules that terminate the trial early if the interim data suggest that the trimodality regime is safer with one RT modality *versus* the other. Moreover, although the risk of toxicity often dominates thinking about RT modalities, delaying disease recurrence or progression and thereby prolonging survival remains the therapeutic goal.

The main statistical challenge in designing a trial to compare RT modalities used with the trimodality regime is that the clinical outcomes are very complex. To measure the combined effect of the diverse array of possible toxicities, we define a statistic, the *total toxicity burden* (TTB), which provides a continuous measure of the combined effect of all toxicities experienced by the patient over the course of follow-up. To construct this statistic, numerical weights of each possible grade of each toxicity that quantify their relative severities first must be elicited from the physicians planning the trial. The TTB is defined as a severity-weighted sum over the different toxicities that may occur. Since some toxicities may occur up to 12 months from the start of treatment, each patient's observed TTB is a process that may change over time.

Our trial design treats TTB and PFS as co-primary end points. It relies on a multivariate Bayesian model that accounts for the *incidence and severity* of each type of toxicity, and it accounts for dependence between the toxicity vector, an indicator of whether surgery is performed and PFS. A key feature of the model is that it provides an analytically tractable expression for mean TTB. The two RT modalities are compared by using two group sequential rules, based on the posterior distributions of mean TTB and PFS log-hazard-ratio.

Including a vector of qualitatively different toxicities via TTB in this way is very different from most randomized oncology trials, which are based on PFS or survival, while including toxicities as secondary outcomes. In most trials, toxicity is monitored informally. When formal decision rules based on toxicity are used, they are defined by first reducing the vector of toxicities to a binary indicator of the worst toxicity of any type occurring at or above a given grade, ignoring recurrences entirely. As a basis for comparison, we consider a trial with two sets of conventional group sequential rules with O'Brien–Fleming (O'Brien and Fleming, 1979) boundaries: one based on PFS and the other based on an indicator of any toxicity occurring within 1 year of follow-up. Our simulations, which are given in Section 5, show that our design yields as much as a 66% increase in power and 18% reduction in mean sample size when compared with this conventional design.

The general problem of designing clinical trials to compare multiple end points has been considered by many researchers, predominantly using frequentist approaches for testing composite hypotheses, and relying on large sample normal approximations. O'Brien (1984) considered a generalized linear least squares statistic for composite alternatives that characterizes treatment differences between multiple end points by using a common multiplier. Tang *et al*. (1989a) proposed an approximate likelihood ratio test for multiple treatment effects over all possible directions, with application to group sequential design (Tang *et al*., 1989b). Tang *et al*. (1993) provided group sequential critical values for designs based on several types of frequentist multiple hypothesis testing procedures. Other researches have considered two-stage and multistage designs for monitoring toxicity and response rates in single-arm trials where both end points are binary and observed shortly after treatment (Bryant and Day, 1995; Conaway and Petroni, 1995). Quality-adjusted time without symptoms and toxicity methods were developed to incorporate health-related quality-of-life measures in analysis of time-to-failure end points (Gelber *et al*., 1995). Kosorok *et al*. (2004) provided a group sequential design for multiple primary end points with multiple decision rules that control overall type I error and probabilities of concluding incorrect alternatives. O'Neill (2008) discussed challenges in evaluating the risk *versus* benefit of new therapies in clinical trial design and analysis.

The ideas in this paper are presented in the following sequence. In Section 2, we define TTB for the oesophageal RT trial. The group sequential design is presented in Section 3. In Section 4, we present the probability model, derive the mean TTB and discuss prior specification and elicitation. In Section 5, we present results of a simulation study. Section 6 describes our process and rationale in constructing the model and design for this study and provides general guidelines for practitioners who may wish to use TTB.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Total toxicity burden

Table 1 presents the 11 toxicities that were monitored in the RT trial. Each toxicity is either possibly recurrent at random times over the patient's follow-up period or is a POC evaluated once, approximately 1 week after surgery. Radiation-induced pulmonary and cardio-vascular toxicities may occur up to 12 months following RT and thus require long-term follow-up (Abid *et al*., 2001; Rancati *et al*., 2003; Yusuf *et al*., 2011). Each of the first three possibly recurrent toxicities, pericardial effusion, pleural effusion and pneumonitis, and the POC anastomotic leak, is ordinal with three levels of severity. Each of the remaining seven toxicities has one level of severity. The set of severity weights that were used in the trial are given in Table 1, with a higher weight corresponding to increased severity. The numerical value of each severity weight reflects the relative extent of harm that is associated with experiencing the toxicity at the given level of severity in relation to the other levels of severity of the same toxicity and other toxicities monitored in the trial. These were elicited from the clinical oncologists planning the trial and represent the group's consensus. For example, pericardial effusion requiring medical but not surgical intervention, and the occurrence of a pulmonary embolism, both have elicited weight 60 and thus are considered equally harmful. Weights were elicited in the range 0–100 for convenience, since the oncologists were comfortable with this domain. However, any finite positive domain would work in practice. In our study, $w = 0$ implies no harm to the patient, whereas $w = 100$ represents an extent of harm that is imminently life threatening. We describe the manner in which we elicited the severity weights as well as provide justification for the numerical values in Section 6 and in section A of the on-line supplementary material.

Table 1. Toxicities and elicited severity weights for the 11 toxicities that are monitored in the oesophageal cancer trial†

	Level of severity	*Elicited weight*
Recurrent toxicities		
Pericardial effusion	Non-symptomatic	10
	Medical intervention	60
	Surgical intervention	90
Pleural effusion	Non-symptomatic	10
	Medical intervention	30
	Surgical intervention	60
Radiation pneumonitis	Grade 1–2	20
	Grade 3	60
	Grade 4–5	90
Pneumonia	Occurrence	40
Atrial fibrillation	Occurrence	30
Myocardial infarction	Occurrence	70
POCs		
Anastomotic leak	Radiographic only	30
	Medical intervention	60
	Surgical intervention	90
Acute respiratory distress syndrome	Occurrence	90
Pulmonary embolism	Occurrence	60
Reintubation	Occurrence	70
Stroke	Occurrence	90

†Medical rationale to justify the numerical values is provided in the on-line supplementary material.

Fig. 1 illustrates TTB for a hypothetical patient, computed using each set of weights. Each spike in Fig. 1(b) indicates either a single toxicity or a collection of POCs. The heights of the spikes correspond to the weights, and thus illustrate severities. Fig. 1(a) plots the patient's TTB over time. The hypothetical patient's TTB was 0 until she experienced post-operative reintubation and stroke in week 13, and then successive onsets of pneumonia at weeks 27 and 37. Using the elicited weights, these toxicities contributed severity scores $70 + 90 = 160$ at week 13 and 40 at each of weeks 27 and 37.

The following notation expresses the TTB as a function of elicited severity weights and event indicators arising from two multivariate marked point processes: one characterizing recurrent toxicity; the other POCs following surgery. Indexing the toxicities in Table 1 by $k = 1, \ldots, 11$, we denote the vectors of elicited ordinal toxicity severity weights by $\mathbf{w}_1, \ldots, \mathbf{w}_{11}$. For example, $\mathbf{w}_1 = (w_{1,1}, w_{1,2}, w_{1,3}) = (10, 60, 90)$ for the three levels of pericardial effusion, $\mathbf{w}_4 = w_{4,1} = 30$ if atrial fibrillation occurs, and so on (Table 1). Without loss of generality, we represent patient follow-up as a proportion of the maximum follow-up duration (52 weeks) required to account for late onset radiation-induced toxicity, $t \in (0, 1]$. Let $\mathbf{N}(t) = \{N_1(t), \ldots, N_6(t)\}$ denote the multivariate counting process characterizing the numbers of toxicities occurring by time t for the six recurrent toxicities. The $1 \times M_k$ vector $\mathbf{Z}_{k,j}(t) = \{Z_{k,j,1}(t), \ldots, Z_{k,j,M_k}(t)\}$ denotes the multinomial point process that marks the severity of the jth occurrence of toxicity type k. Each $Z_{k,j,m}(t)$ is a simple point process with $Z_{k,j,m}(t^*) = 1$, for all $t^* > t$ if the jth incidence of the kth toxicity occurs at the mth level of severity before time t. If the jth event has not occurred by time t, then $Z_{k,j,m}(t^*) = 0$, for all $m = 1, \ldots, M_k$.

We use the elicited weights $\mathbf{w}_k$, the observed occurrence processes $N_k(t)$ and the mark processes $\mathbf{Z}_k(t)$ to define the *toxicity burden for the kth recurrent toxicity at follow-up time t*,

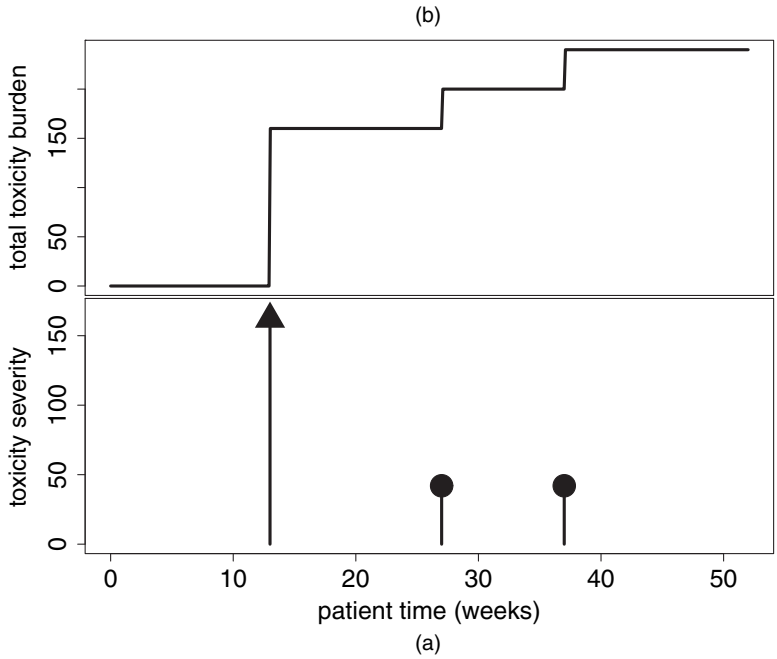

Fig. 1. (a) Example of toxicity severity scores and (b) their sum, the TTB, for a single patient: ▲, POCs; ●, recurrent toxicities

$$B_k^{\mathrm{REC}}(t) = \sum_{j=1}^{N_k(t)} \mathbf{w}_k' \mathbf{Z}_{k,j}(t) = \sum_{j=1}^{N_k(t)} \sum_{m=1}^{M_k} w_{k,m}\, Z_{k,j,m}(t). \tag{1}$$

The jth occurrence of recurrent toxicity type k produces a jump of size $\mathbf{w}_k' \mathbf{Z}_{k,j}(t)$ in the $B_k^{\mathrm{REC}}(t)$ process at the event time. We define the TTB *contributed by the six recurrent toxicities at follow-up time* t as the sum

$$B^{\mathrm{REC}}(t) = \sum_{k=1}^{6} B_k^{\mathrm{REC}}(t).$$

For the subset of patients who undergo surgery, there are five possible POCs in the RT trial. These consist of one ordinal-valued toxicity, anastomotic leak, indexed by $k=7$, with $M_7=3$ levels, and four binary-valued toxicities for which $M_k \equiv 1$, indexed by $k=8,\ldots,11$. The POCs are assessed only once, approximately 1 week following surgery. Let $\{S(t)=0, 1 : 0<t<1\}$ denote an event process with at most a single jump discontinuity of size 1 at the time that surgery is performed. Matching indices for recurrent events (k, j, m) = (toxicity type, recurrence number, severity level), we use $\mathbf{Z}_7(t) = \{Z_{7,1,1}(t), Z_{7,1,2}(t), Z_{7,1,3}(t)\}$ to denote the vector of severity level indicators for the ordinal-valued POC. The occurrence indicators for the four remaining POCs are denoted by $Z_{k,1,1}(t)$, $k=8,\ldots,11$. We define the *toxicity burden contributed by POCs* at follow-up time t as

$$B^{\mathrm{POC}}(t) = S(t)\left\{ \sum_{m=1}^{3} w_{7,m} Z_{7,1,m}(t) + \sum_{k=8}^{11} w_{k,1} Z_{k,1,1}(t) \right\}.$$

We can now define each patient's TTB at follow-up time t as the sum of these two components:

$$B(t) = B^{\mathrm{REC}}(t) + B^{\mathrm{POC}}(t). \tag{2}$$

Because we account for occurrence over time and possible recurrences, this definition generalizes that used for Bayesian phase I dose finding by Bekele and Thall (2004).

3. Group sequential bivariate trial design

Because toxicities that are induced by RT are rare but serious, our trial must include interim monitoring based on TTB. In planning the RT trial, the participating oncologists were unwilling to continue randomizing if the data showed strong evidence of a difference between PBT and IMRT for either TTB or PFS. Consequently, these are co-primary end points, and the group sequential design stops the trial early if one modality is superior with respect to TTB, PFS or both. The trial's decision scheme thus is complex. For each interim decision, there are three possibilities for each of the two outcomes, namely that PBT is superior, IMRT is superior or neither is superior. This yields nine possible joint decisions.

At any trial time, patients will have different follow-up times $t_1, \ldots, t_n$, and some will have undergone surgery whereas others will not, so in general their TTBs will not be comparable. For example, $B^{\mathrm{REC}}(0.5)$ and $B^{\mathrm{REC}}(1)$ correspond to different follow-up periods. Thus, we shall conduct the trial by specifying a joint model for the multivariate patient outcome, deriving the resulting expectation of TTB, $E\{B(1)\}$, and formulating group sequential decision rules for comparing safety by using the intermodality difference in mean TTB. Importantly, $E\{B(1)\}$ is the function of severity weights and model parameters (we shall use $\boldsymbol{\theta}$ to denote model parameters) that characterizes the extent of TTB experienced by a patient on average over the course of the entire at-risk duration for one modality. Our trial's decision rules are based on the resulting posteriors of the difference in mean TTB, $\Delta(\boldsymbol{\theta}, \mathbf{w}) = E\{B(1) \mid \mathrm{IMRT}\} - E\{B(1) \mid \mathrm{PBT}\}$, and the PFS log-hazard-ratio δ^{ξ}. Model specification and derivation of $\Delta(\boldsymbol{\theta}, \mathbf{w})$ are given in Section 4.

Let $\mathcal{D}(\tau)$ denote the observed data for all patients enrolled by trial time τ. Let $\varepsilon^{\mathrm{TTB}}$ denote a small value of $\Delta(\boldsymbol{\theta}, \mathbf{w})$ and $\varepsilon^{\mathrm{PFS}}$ a small value of δ^{ξ} that are considered clinically insignificant. At interim analysis time τ, the safety comparison is based on posterior probabilities that the difference in mean TTB of one modality compared with the other exceeds $\varepsilon^{\mathrm{TTB}}$:

$$\varphi^{\mathrm{PB}}(\varepsilon^{\mathrm{TTB}}, \tau) = \Pr\{\Delta(\boldsymbol{\theta}, \mathbf{w}) > \varepsilon^{\mathrm{TTB}} | \mathcal{D}(\tau)\}$$

in favour of PBT, and

$$\varphi^{\mathrm{IM}}(\varepsilon^{\mathrm{TTB}}, \tau) = \Pr\{-\Delta(\boldsymbol{\theta}, \mathbf{w}) > \varepsilon^{\mathrm{TTB}} | \mathcal{D}(\tau)\}$$

in favour of IMRT.

Similarly, the PFS comparison is based on the posterior probability that the PFS log-hazard-ratio exceeds $\varepsilon^{\mathrm{PFS}}$ in favour of one modality compared with the other:

$$\chi^{\mathrm{PB}}(\varepsilon^{\mathrm{PFS}}, \tau) = \Pr\{\delta^{\xi} > \varepsilon^{\mathrm{PFS}} | \mathcal{D}(\tau)\}$$

in favour of PBT, and

$$\chi^{\mathrm{IM}}(\varepsilon^{\mathrm{PFS}}, \tau) = \Pr\{-\delta^{\xi} > \varepsilon^{\mathrm{PFS}} | \mathcal{D}(\tau)\}$$

in favour of IMRT.

The model, which is described in Section 4, is formulated such that larger values of $\varphi^{\mathrm{PB}}(\varepsilon^{\mathrm{TTB}}, \tau)$ represent stronger *a posteriori* evidence that PBT is the safer modality, whereas larger values of $\chi^{\mathrm{PB}}(\varepsilon^{\mathrm{PFS}}, \tau)$ correspond to PBT being more effective for delaying recurrence or progression

and prolonging survival. Similarly, larger $\varphi^{\mathrm{IM}}(\varepsilon^{\mathrm{TTB}}, \tau)$ or $\chi^{\mathrm{IM}}(\varepsilon^{\mathrm{PFS}}, \tau)$ correspond to superior safety or effectiveness of IMRT.

Numerical values of $\varepsilon^{\mathrm{TTB}}$ and $\varepsilon^{\mathrm{PFS}}$ must be specified in the context of possible values of $\Delta(\boldsymbol{\theta}, \mathbf{w})$ and δ^{ξ}. Whereas the log-hazard-ratio δ^{ξ} is readily interpretable, the magnitude of clinical relevance for $\Delta(\boldsymbol{\theta}, \mathbf{w})$ may be less obvious. For the RT trial, any improvement in mean TTB or PFS was considered clinically relevant by the participating oncologists; therefore posterior probabilities were computed using $\varepsilon^{\mathrm{TTB}} = \varepsilon^{\mathrm{PFS}} = 0$. The resulting decision rules are structurally similar to the multiple-hypothesis test-based method of Kosorok *et al.* (2004) using 'vague' alternative hypotheses.

To conduct a trial with the group sequential comparisons, we must determine both the 'timing' and the minimal extent of 'evidence' that is required to confer each decision at each analysis time. For our trial, we defined posterior thresholds on φ and χ as functions of information statistics, which are referred to hereafter as 'decision boundaries', with which the posterior probabilities will be compared to make decisions during the trial. The information statistics characterize the proportion of total information, $\mathcal{I}(\tau) \in [0, 1]$, that has been observed at the time of analysis in relation to the maximum possible information that could be observed in the trial. Using accumulated information, rather than calendar time or sample size, to decide when to perform the interim comparisons provides robustness to misspecification of the assumed rate of enrolment, which is a common practical issue in group sequential trials.

Let $n(\tau)$ denote the number of patients enrolled by τ. We used the cumulative follow-up duration at τ, $\mathcal{I}^{\mathrm{TTB}}(\tau) = \Sigma_{i=1}^{n(\tau)} t_i / N$, to define an information statistic for TTB. Because t_i denotes the ith patient's proportion of total follow-up out of 52 weeks, $\mathcal{I}^{\mathrm{TTB}}$ represents the cumulative follow-up for toxicity as a proportion of total follow-up that would be observed if all N patients were monitored for toxicity for the entire post-RT toxicity at-risk period of 52 weeks. Let $C_i(\tau)$ indicate whether the ith patient's PFS duration is right censored at trial time τ. The appropriate information statistic for PFS is the proportion of events by τ, $\mathcal{I}^{\mathrm{PFS}}(\tau) = \Sigma_{i=1}^{n(\tau)} \{1 - C_i(\tau)\} / N$. To use φ and χ for decision making at trial time τ, we defined boundaries on the posterior probability domain by using the function

$$c^{y}(\tau) = 1 - \beta^{y} (\mathcal{I}^{y})^{\alpha^{y}}(\tau), \qquad \text{for } y \equiv \mathrm{TTB} \text{ or } y \equiv \mathrm{PFS}. \tag{3}$$

The exponents $\alpha^{\mathrm{TTB}}, \alpha^{\mathrm{PFS}} > 0$ and scaling parameters $0 < \beta^{\mathrm{TTB}}, \beta^{\mathrm{PFS}} < 1$ must be calibrated to obtain a design that satisfies prespecified size, power and optimality criteria. The approach is similar to the boundary functions that were proposed by Wathen and Thall (2008). The boundaries (3) are consistent with conventional group sequential designs (O'Brien and Fleming, 1979; Lan and DeMets, 1983; DeMets and Lan, 1994) in the sense that early stopping requires a smaller numerical difference as more information accrues during the trial.

The nine joint decision rules are given in Table 2. Decisions 1, 2 or 4 each would lead to the conclusion that PBT is superior, since it is superior to IMRT for either both end points or at least one end point without evidence of a clinically significant difference for the other. Similarly, Decisions 6, 8 or 9 would lead to the conclusion that IMRT is superior. Decision 5 corresponds to the absence of evidence for a meaningful difference between PBT and IMRT for either end point. Decision 5 may be achieved at the end of the trial; in this case it might be described as failing to reject the global null hypothesis. Decisions 3 and 7 both conclude that, with either modality, it is inferior for one outcome and superior for the other. These conclusions are not made by conventional hypothesis tests but easily could arise in any clinical setting where treatments have both harmful and beneficial effects. At the extremes, decisions 1 and 9 might be called 'win–win' decisions and represent extremely optimistic scenarios that are rarely obtained in practice. For

Table 2. Joint decision rules for monitoring TTB and PFS at trial time τ†

PFS rule	*TTB rule*		
	$\varphi^{\mathrm{PB}} > c^{\mathrm{TTB}}$	$\varphi^{\mathrm{PB}} \vee \varphi^{\mathrm{IM}} < c^{\mathrm{TTB}}$	$\varphi^{\mathrm{IM}} > c^{\mathrm{TTB}}$
$\chi^{\mathrm{PB}} > c^{\mathrm{PFS}}$	1, PBT safer and more effective	2, PBT more effective with indeterminate safety	3, IMRT safer and less effective
$\chi^{\mathrm{PB}} \vee \chi^{\mathrm{IM}} < c^{\mathrm{PFS}}$	4, PBT safer with indeterminate efficacy	5, continue enrolling patients	6, IMRT safer with indeterminate efficacy
$\chi^{\mathrm{IM}} > c^{\mathrm{PFS}}$	7, PBT safer and less effective	8, IMRT more effective with indeterminate safety	9, IMRT safer and more effective

†For brevity, we denote the posterior probabilities $\varphi^{\mathrm{PB}} = \varphi^{\mathrm{PB}}(\varepsilon^{\mathrm{TTB}}, \tau)$, $\varphi^{\mathrm{IM}} = \varphi^{\mathrm{IM}}(\varepsilon^{\mathrm{TTB}}, \tau)$, $\chi^{\mathrm{PB}} = \chi^{\mathrm{PB}}(\varepsilon^{\mathrm{PFS}}, \tau)$ and $\chi^{\mathrm{IM}} = \chi^{\mathrm{IM}}(\varepsilon^{\mathrm{PFS}}, \tau)$ and the monitoring boundaries $c^{\mathrm{TTB}} = c^{\mathrm{TTB}}(\tau)$ and $c^{\mathrm{PFS}} = c^{\mathrm{PFS}}(\tau)$. $u \vee v =$ maximum$\{u, v\}$.

this treatment regime and disease, it is considered unethical to continue randomizing patients if one modality is determined to be inferior for either end point. Thus, the only case where the trial is continued interimly is under decision 5, since it reflects clinical equipoise, or indifference for both PFS and TTB.

For trial conduct, the decision rules are applied at three interim analyses when 33%, 50% and 67% of the expected total information has accrued and utilized at the final analysis 1 year after the patient enrolment period has ended. Patients are expected to be enrolled at a rate of four per month, requiring a total of 3.75 years to reach the targeted maximum enrolment of $N = 180$, and 4.75 years to complete patient follow-up.

4. Probability model

We expect associations between the toxicities, surgery and PFS. Thus, we formulated our model to induce a dependence structure that we believed was qualitatively consistent with the clinical context, for which we made the following assumptions. Radiation delivered to the thoracic cavity may adversely affect critical organs and thus has the potential to reduce survival. Additionally, a patient with early RT-induced toxicity is less likely to undergo surgery, hence experiencing both a higher risk of disease progression and a lower risk of POCs. To reflect these assumptions, we used the following frailty model. Indexing patients by $i = 1, \ldots, n$, let $\{U_i, i = 1, \ldots, n\}$ denote independent and identically distributed random patient *frailties* with $E(U_i) = 1$ and $\mathrm{var}(U_i) = \phi$. To induce positive correlation between the counts of the recurrent toxicities, we employed the common device of assuming that the kth recurrent toxicity process $N_{i,k}(t)$ of patient i in treatment arm x is Poisson distributed with conditional intensity $U_i Z \psi_k(x)$. This is presented in Section 4.1. We used an exponential model for time to surgery (which is presented in Section 4.2) with hazard rate multiplied by U_i^{-1}. The conditional distribution of PFS Y_i given U_i, which is presented below in Section 4.3, is assumed to follow a piecewise exponential distribution with the baseline hazard on each time subinterval multiplied by U_i. Details are given in section B of the on-line supplementary material. Averaging over the distribution of U_i, these assumptions give a model with

(a) association between the recurrent toxicity counts $N_{i,1}(t), \ldots, N_{i,6}(t)$,
(b) each $N_{i,k}(t)$ negatively correlated with Y_i, so that increased toxicity is associated with shorter PFS, and

(c) negative association between surgery and the incidence of each recurrent toxicity, and positive association between surgery and PFS.

Moreover, the model must also yield analytical tractability for expressing the difference in mean TTB between treatment modalities as a function of model parameters and severity weights, since its posterior along with the posterior distribution for the PFS hazard ratio are used in the group sequential procedure. We assumed that the frailties follow an inverse gamma distribution, $U_i \sim \Gamma^{-1}(1/\phi+2, 1/\phi+1)$. This gives a multivariate model that yields an analytically tractable expression for the mean TTB difference. We considered other parametric and non-parametric models, but we did not use them because they required numerical integration to compute mean TTB, which complicated computation of the design's operating characteristics.

Our model represents just one possible set of assumptions. Moreover, alternative methods could have been used to account for interdependence between toxicities, surgery and PFS. For example, instead of using recurrent event processes, one could assume that a transformation $g(\cdot)$, applied to the TTB-statistic and perhaps scaled per follow-up duration (i.e. $g\{B(t)/t\}$) is Gaussian. Then one could conceivably proceed by specifying a multivariate normal model for the transformed TTB statistic in conjunction with transformations of the time-to-surgery and time-to-PFS end points. A perhaps more appealing, but less tractable, solution might use copulas to describe the dependence between toxicities, surgery and PFS.

4.1. Recurrent toxicity processes

Initially, we considered a model with toxicity-specific marked point processes for severities and treatment effects for both event recurrence and severity probabilities. For the six recurrent toxicities, this requires a minimum of 18 model parameters. Similarly, assuming POC-specific marked point processes with a modality effect for the rate of surgery and separate modality effects for the severity probabilities for each of the five POCs requires a minimum of 12 model parameters. We found this model to be too complex to use as a practical basis for trial design.

To simplify the model further and to reduce its dimension, we now exploit the fact that treatment comparisons based on TTB need only consider the incidence of each possible total toxicity severity. Together, the recurrent toxicities in the RT trial may have one of seven unique severities: $\mathbf{w}^* = (10, 20, 30, 40, 60, 70, 90)$. Thus, with a slight abuse of notation, hereafter we formulate the model in terms of the counting processes $N_{i,1}(t)$ for all toxicities of any type giving total severity 10, $N_{i,2}(t)$ for all toxicities of any type with total severity 20, and so on. Hereafter, the toxicity index is $k = 1, \ldots, 7$ rather than $1, \ldots, 11$. This reduces the number of model parameters from 30 to 18. However, to assess robustness in the simulation studies that are described in Section 5, we shall use a saturated 11-dimensional marked point process model to generate the data.

We assume that each patient's risk of radiation-induced toxicity depends on the type of irradiation that is delivered (at the group level) as well as the anatomic location of the tumour (at the patient level). The latter impacts the dosimetric plan and thereby determines the extent of irradiation that is delivered to healthy tissues in neighbouring regions. To accommodate intra-patient dependence, we used doubly stochastic Poisson (Cox) processes to induce association between recurrent toxicity severity (Cox, 1955; Snyder and Miller, 1991; Jacobsen, 2006; Cook and Lawless, 2007).

We shall denote treatment by $x = -0.5$ for IMRT and $x = 0.5$ for PBT. Let $\psi_k(x) = \lambda_k \exp(-x\delta_k^{\psi})$ denote the mean rate of recurrent toxicity severity k for a patient in treatment arm x. Thus, $\lambda_k > 0$ is the baseline rate and δ_k^{ψ} is the real-valued PBT *versus* IMRT RT modality effect on the log-mean-rate. We assume that, given U_i and x_i, the kth recurrent severity process $\{N_{i,k}(t), t \geqslant$

$0|U_i, x_i\}$ for patient i is a Poisson process with conditional intensity $U_i \psi_k(x_i)$. The random frailty U_i acts as a common scalar of the mean rates of $N_{i,1}(t), \ldots, N_{i,7}(t)$ inducing positive association between the event processes. Since $E(U_i) = 1$, after averaging over the distribution of U_i, $N_{i,k}(t)$ has unconditional mean $t\psi_k(x_i)$, variance $t\psi(x_i) + \phi t^2 \psi_k(x_i)^2$ and covariance $\text{cov}\{N_{i,r}(t), N_{i,l}(t)|x_i\} = \phi t^2 \psi_r(x_i)\psi_l(x_i)$. The frailty variance ϕ determines the degree of association between counts over disjoint intervals within each event process as well as the degree of association between different event processes (Cox and Isham, 1980; Breslow, 1984; Lawless, 1987a,b).

A likelihood is necessary to conduct posterior inference. Section B.1 of the on-line supplementary material provides the likelihood contribution for the toxicity count vector $\mathbf{N}_i(t)$. Our model is parameterized so that larger treatment effects, denoted by δ with appropriate subscripts and superscripts, correspond to superiority of PBT over IMRT. For example, a larger positive value of δ_k^{ψ} corresponds to a smaller event rate for recurrent severity k with PBT *versus* IMRT.

4.2. Surgery process and post-operative complications

Because TTB is the sum $B(t) = B^{\text{REC}}(t) + B^{\text{POC}}(t)$, treatment comparison based on $E\{B(t)\}$ is influenced by the probability of undergoing surgery following chemoradiation, and thus becoming at risk of POCs. We assumed that a patient experiencing a severe toxicity with chemoradiation is less likely to undergo surgery, whereas surgery reduces the risk of disease recurrence and thus is positively associated with PFS. To reflect these relationships, given frailty U_i and treatment arm x_i, we assume that the time-to-surgery distribution is exponential with conditional hazard rate $\tilde{\lambda} \exp(x_i\tilde{\delta})/U_i$, inducing dependence with $\mathbf{N}_i(t)$ and PFS as per our assumptions when $\phi > 0$. Thus, $\tilde{\lambda}$ is the baseline rate and $\tilde{\delta}$ is the real-valued PBT *versus* IMRT RT modality effect, with $\tilde{\delta} > 0$ corresponding to increased relative rate of surgery in favour of PBT. Section B.2 of the on-line supplementary material provides the likelihood contribution. However, as described in Section 5.1, we evaluate our design's operating characteristics by generating surgery times with an approximation of the baseline hazard function that we expect to observe in the trial by using a piecewise constant model.

The POCs that are monitored in this trial are serious, rare events. A major motivation for the trial is whether the relative incidences or severities of these POCs may differ between the two RT modalities, since chemoradiation may impact the extent to which a patient tolerates surgery. Because all five POCs are assessed at a single time point following surgery, we use a multinomial model for the aggregate severity of the POCs. There are 24 possible values of the total POC severity (computed from Table 1) which we denote in order from least to most severe by $\tilde{\mathbf{w}} = (0, 30, \ldots, 400)$. Let $\mathbf{Z}_i(t) = \{Z_{i,1}(t), \ldots, Z_{i,24}(t)\}$ denote the vector of aggregate severity level indicators for patient i, whereby $\Sigma_{m=1}^{24} Z_{i,m}(t) = 1$ if $S(t) > 0$, and $Z_{i,m}(t) = 0$, for all m otherwise. We assume that $[\mathbf{Z}_i(t)|x_i] \sim \text{multinomial}\{\boldsymbol{\pi}(x_i)\}$, with RT-specific probability vector $\boldsymbol{\pi}(x_i) = \{\pi_1(x_i), \ldots, \pi_{24}(x_i)\}$, where $\Sigma_{m=1}^{24} \pi_m(x_i) = 1$. Section B.3 of the on-line supplementary material provides the likelihood contribution for POC severity.

4.3. Progression-free survival and mixture model

Our trial uses a piecewise constant hazard formulation for PFS (e.g. Ibrahim *et al.* (2001), section 3.1). This model facilitates between-modality comparisons under the typical proportional hazards assumptions that are robust to the actual shape of the underlying baseline hazard. Given the frailty U_i, we assume that the hazard for PFS is piecewise constant over the time axis partition $(0, s_1], (s_1, s_2], \ldots, (s_{G-1}, s_G], (s_G, \infty)$, where $0 < s_1 < s_2 < \ldots < s_G < \infty$. For $[Y_i|U_i]$ in the interval $(s_{g-1}, s_g], g = 1, \ldots, G$, the constant baseline hazard is $\xi_g(x_i, U_i) = U_i \gamma_g \exp(-x_i\delta^{\xi})$,

where $\gamma_g > 0$, and we denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_G)$. This model can approximate any underlying smooth baseline hazard function, providing a robust relative comparison between modalities. We selected the set of hazard discontinuities, $(s_1, \ldots, s_G)$, adaptively using the interim data to obtain a time axis partition that is Akaike information criterion optimal among all sets of equidistant quantiles so that each interval contains at least 10 observed events. Positive values of δ^ξ correspond to longer median PFS for PBT *versus* IMRT. Section B.4 of the on-line supplementary material provides the likelihood contribution for PFS. Additionally, multiplying each interval hazard $\gamma_g \exp(-x_i \delta^\xi)$ by the frailty U_i provides subject-specific perturbations of the baseline hazard function inducing positive or negative correlation with surgery or toxicity respectively.

Collecting terms, the ith patient's observable outcome vector is $\boldsymbol{\mathcal{D}}_i = (\mathbf{N}_i, S_i, \mathbf{Z}_i, Y_i, C_i)$, $i = 1, \ldots, n$, and the overall joint likelihood contribution is obtained by averaging the product of conditional likelihoods of the observables over the frailty distribution. Since these are assumed to be conditionally independent given U_i, this is

$$\mathcal{L}_i(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}_i) = \mathcal{L}_{\mathbf{Z}_i} \int_{u=0}^{\infty} \mathcal{L}_{\mathbf{N}_i}(u)\, \mathcal{L}_{S_i}(u)\, \mathcal{L}_{Y_i}(u)\, \mathrm{d}\Gamma^{-1}\left(u \mid \frac{1}{\phi} + 2, \frac{1}{\phi} + 1\right). \tag{4}$$

The model parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \tilde{\lambda}, \boldsymbol{\pi}, \boldsymbol{\gamma}, \phi, \boldsymbol{\delta}^\psi, \tilde{\delta}, \delta^\xi)$, and we denote the data for n patients by $\boldsymbol{\mathcal{D}} = \cup_{i=1}^n \boldsymbol{\mathcal{D}}_i$.

4.4. Mean total toxicity burden

Having specified a model for the multivariate patient outcome, we can now derive the expectation of equation (2) as well as the intermodality difference in mean TTB, $\Delta(\boldsymbol{\theta}, \mathbf{w}) = E\{B(1)|\mathrm{IMRT}\} - E\{B(1)|\mathrm{PBT}\}$, which provides the basis for comparing safety in our trial. The marginal expected toxicity burden of $N_k(t)$ for a patient who is assigned to treatment x is

$$\mu_k^{\mathrm{REC}}(t, x, \boldsymbol{\theta}) = t\, \psi_k(x) w_k^*, \qquad k = 1, \ldots, 7. \tag{5}$$

The expected severity from POCs over the follow-up period $[0, t)$ for a patient who is assigned to treatment x is a tractable function of $\boldsymbol{\pi}(x)$, the surgery intensity $\tilde{\lambda}$ and frailty variance ϕ, given by

$$\mu^{\mathrm{POC}}(t, x, \boldsymbol{\theta}) = \tilde{\mathbf{w}}' \boldsymbol{\pi}(x) \left[1 - \left\{\frac{\phi t \tilde{\lambda} \exp(x\tilde{\delta})}{\phi + 1} + 1\right\}^{-(1/\phi + 2)}\right]. \tag{6}$$

Details of the derivation are provided in Appendix A. In the absence of frailty dispersion ($\phi = 0$), the mean TTB from POCs would be the multinomial mean $\mu^{\mathrm{POC}}(t, x, \boldsymbol{\theta}) = \tilde{\mathbf{w}}' \boldsymbol{\pi}(x)$. Equation (6) shows that the frailties reduce $\tilde{\mathbf{w}}' \boldsymbol{\pi}(x)$ by a multiplicative factor that depends on $(\phi, \tilde{\lambda}, \tilde{\delta}, x)$ and takes values between 0 and 1. As the frailty dispersion increases, there is a decrease in the probability of undergoing surgery after experiencing recurrent toxicity, thereby attenuating the influence of POCs. The multiplicative term approaches the lower limit $1 - \{t\tilde{\lambda} \exp(x\tilde{\delta}) + 1\}^{-2}$, as $\phi \to \infty$, which is the largest multiplicative amount by which frailty dispersion may reduce $\tilde{\mathbf{w}}' \boldsymbol{\pi}(x)$.

The μs given by equations (5) and (6) quantify risk–severity trade-offs, and their sum,

$$\mu(t, x, \boldsymbol{\theta}) = \sum_{r=1}^{7} \mu_r^{\mathrm{REC}}(t, x, \boldsymbol{\theta}) + \mu^{\mathrm{POC}}(t, x, \boldsymbol{\theta}), \tag{7}$$

is the mean TTB for a patient who is assigned to modality x at follow-up time t. Recalling that $x = -0.5$ for IMRT and $x = 0.5$ for PBT, the design uses the 52-week ($t = 1$) mean difference as the

basis for IMRT *versus* PBT safety comparison, denoted by $\Delta(\boldsymbol{\theta}, \mathbf{w}) = \mu(1, -0.5, \boldsymbol{\theta}) - \mu(1, 0.5, \boldsymbol{\theta})$. Positive values of $\Delta(\boldsymbol{\theta}, \mathbf{w})$ correspond to superior safety for PBT compared with IMRT.

4.5. Establishing priors

To specify priors, we selected distributions to satisfy model constraints or to exploit analytical properties of conditional conjugacy. Let $\boldsymbol{\delta} = (\boldsymbol{\delta}^{\psi}, \tilde{\delta}, \delta^{\xi})$ characterize the respective modality differences for recurrent toxicity, surgery and PFS. The remaining model parameters $\boldsymbol{\theta} - \boldsymbol{\delta}$ characterize baseline features of the recurrent toxicity severity or surgery processes, severity probabilities from POCs or the frailty dispersion.

Given that one modality effectuates a safer plan, we expect reductions in the event rates of each of the radiation-induced toxicities that are monitored in the trial. Under this assumption, parameters that characterize the corresponding 'group level' treatment effects should be positively associated. To effectuate this, we used a hierarchical prior to induce shrinkage among exchangeable treatment effects: $\delta_k^{\psi} \sim N(\delta^{\psi}, \omega^2)$, $k = 1, \ldots, 7$. A non-informative prior was assumed for the hierarchical mean, $\delta^{\psi} \sim N(0, 100)$. Following the recommendations of Gelman (2006), the hierarchical standard deviation ω was assumed to be uniform over $(0, 10]$. For the treatment effects for surgery, $\tilde{\delta}$, and PFS, δ^{ξ}, we assumed $N(0, 100)$ priors. We assumed weakly informative priors for the non-treatment effect parameters, including conditionally conjugate gamma distributions for the λs, and for the piecewise constant PFS hazard parameters $\boldsymbol{\gamma}$ and Dirichlet priors for the POC severity probabilities $\pi(x)$. In the absence of prior knowledge about the extent of interdependence between the observables, we assumed a weakly informative uniform prior distribution over the interval $(0, 10]$ for the frailty variance.

Prior hyperparameters for baseline means were estimated from historical data and/or elicited from the team of oncologists. Each hazard component for PFS was centred at the estimated hazard rate derived from a parametric exponential fit to a cohort of 246 patients with stage II–III oesophageal cancer who underwent the trimodality regime at the MD Anderson Cancer Center with IMRT. Table 3 summarizes elicited information characterizing the non-occurrence rate and severity level probability for each toxicity. Prior means of baseline event rates for recurrent toxicities were derived by combining event rates and severity probabilities among toxicity grades having identical severity weights, assuming independence. For example, the occurrence of an atrial fibrillation or a pleural effusion requiring medical intervention both have severity weight $w_3^* = 30$. The corresponding elicited baseline event rate was obtained by mapping the induced probability that *both* toxicities are absent after 52 weeks of follow-up onto the domain of λ. The prior mean baseline event rate for surgery followed similarly from the expectation that 65% of patients would undergo surgery. For each treatment arm, the mean probability of each POC severity level, which is depicted in Fig. 2, was calibrated by using the elicited POC-specific severity probabilities in Table 3 under the assumption of independence.

Prior variances for toxicity severity and surgery were specified by using the prior effective sample size ESS (Morita *et al.*, 2008, 2012) to characterize prior informativeness relative to the amount of information that is contributed by the likelihood on the basis of the trial's maximum sample size of $N = 180$. The concentration hyperparameters for the probability of aggregate POC severity were scaled to sum to 1, inducing a Dirichlet prior with $\mathrm{ESS} = 1$. We set the gamma rate hyperparameter for each conjugate time homogeneous Poisson process to 5. Thus, conditionally on the frailty, the induced prior distribution for each baseline intensity contained information equivalent to five patients.

Given the frailty variance, priors for the unconditional intensities, $U\lambda_{k*}$s and $U^{-1}\tilde{\lambda}$, are proportional to an intractable mixture of Meijer G-functions (Springer and Thompson, 1970). Therefore, unconditional prior ESS-values were derived by using least squares gamma approx-

Table 3. Elicited prior information based on experience in treating patients with photon radiation therapy†

Toxicity		*52-week absence probability*
Recurrent toxicity event processes		
Pericardial effusion		0.96
Pleural effusion		0.95
Radiation pneumonitis		0.90
Pneumonia		0.85
Atrial fibrillation		0.75
Myocardial infarction		0.95
	Severity level	*Severity probability given occurrence*
Recurrent ordinal toxicity severities		
Pericardial effusion	Non-symptomatic	0.50
	Medical intervention	0.30
	Surgical intervention	0.20
Pleural effusion	Non-symptomatic	0.60
	Medical intervention	0.20
	Surgical intervention	0.20
Radiation pneumonitis	Grade 1–2	0.80
	Grade 3	0.10
	Grade 4	0.10
	Severity level	*Severity occurrence probability*
POCs		
Anastomotic leak	Absence	0.87
	Radiographic only	0.08
	Medical intervention	0.03
	Surgical intervention	0.02
Acute respiratory distress syndrome	Absence	0.97
Pulmonary embolism	Absence	0.97
Reintubation	Absence	0.95
Stroke	Absence	0.98

†For each recurrent toxicity, the probability that a patient will not experience the toxicity over 52 weeks was elicited. Binomial or multinomial severity probabilities were elicited for POCs and recurrent toxicities with ordinal severities. Elicited values for toxicities with identical severity weights were combined to establish prior distributions for baseline model parameters.

imations. Among the recurrent toxicity severities, the resulting ESS-values ranged from a minimum of 0.68 for severity weight $w_{*2} = 20$ (radiation pneumonitis of grade less than 3) to a maximum of 1.8 for $w_{*7} = 90$ (surgical pericardial effusion or radiation pneumonitis of grade 4). The unconditional intensity for surgery had prior ESS = 0.74.

5. Simulation study

5.1. Simulation design

We used simulation as a tool to calibrate the boundary function parameters to obtain a design with desirable operating characteristics, including acceptable overall frequentist size and power. Proper evaluation of the design's frequentist properties required simulation of observables under a reasonable set of true distributions. To ensure robustness, these must include distributions that are substantially different from those in the model that is used to construct the design. We thus simulated the toxicities by using a saturated multivariate marked point process with toxicity-specific parameters for event intensities *and* severities. Baseline model parameters for PFS were

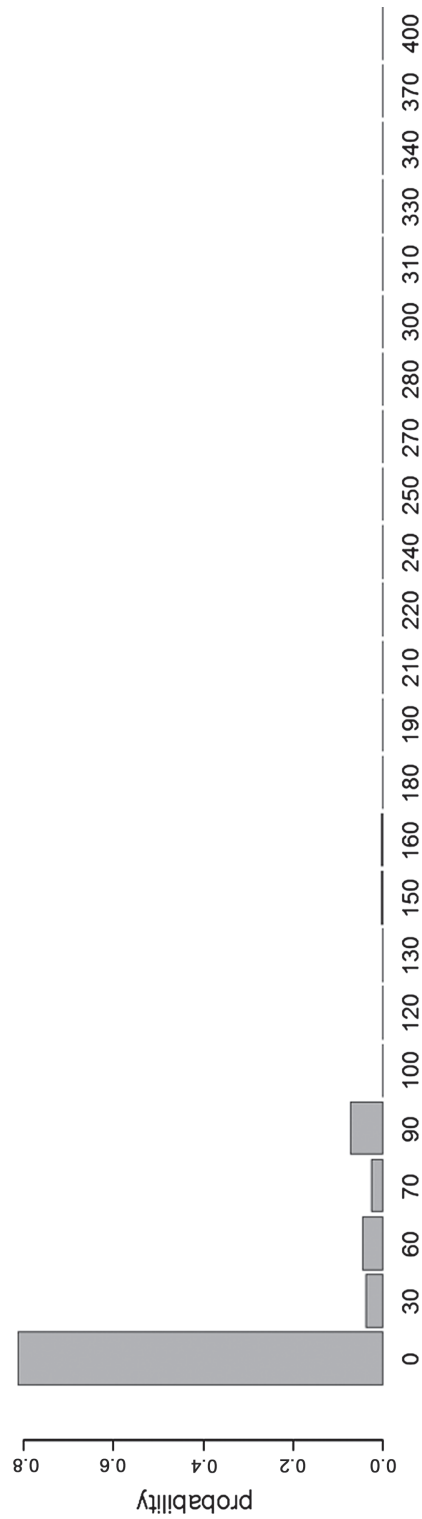

Fig. 2. Elicited Dirichlet prior mean for aggregate POC severity used for posterior inference with prior effective sample size set at 1

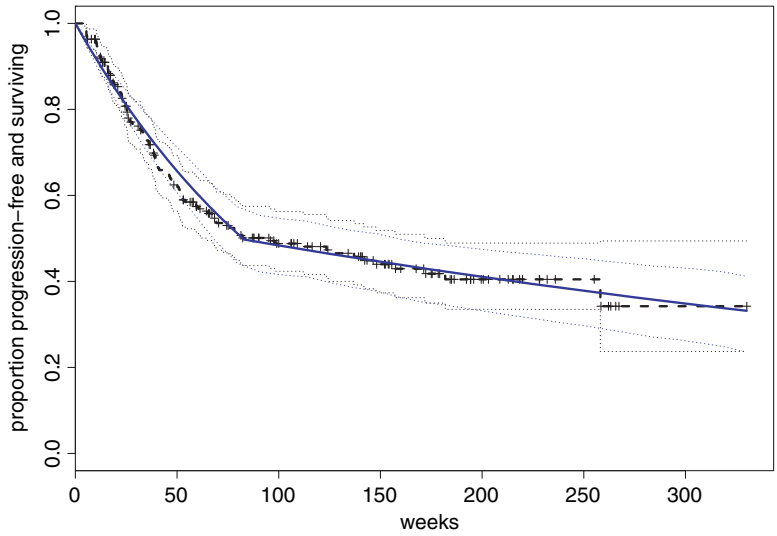

Fig. 3. PFS for a cohort of 246 patients with stage II–III oesophageal cancer who underwent the trimodality regime with IMRT: Kaplan–Meier curve (- - - - - -) and fitted piecewise exponential curve (——) with 95% pointwise confidence intervals

fixed at estimates derived from analysis of a cohort of 246 historical patients treated with IMRT by the participating oncologists. Posterior inference used the model and approach for selecting the time axis partition that was described in Section 4.3, which resulted in an Akaike information criterion optimal partition with two intervals $[0, 83]$ and $(83, \infty)$ and a median of 81.6 weeks. Fig. 3 provides the Kaplan–Meier curve, with 95% log-transformed (Klein and Moeschberger, 2003) pointwise confidence intervals, and fitted survival curve derived from the piecewise exponential analysis (in blue). The corresponding piecewise constant baseline hazard parameters were used to generate random PFS-durations in the simulations.

Surgery event times were generated by using the hazard function that we expect to observe in the trial, which is well approximated by a step function. For example, it is expected that 65% of enrolled patients will undergo surgery following RT, with none undergoing surgery within 4 weeks following RT (therefore within the first 9 weeks of follow-up). For most patients who do undergo surgery it is expected to take place before week 15. Specifically, times to surgery were generated by using a piecewise constant baseline hazard with cumulative distribution probabilities 0.065, 0.49, 0.55 and 0.65 at follow-up durations 10.33, 14.67, 19 and 52 respectively. Both the time axis partition and the cumulative event probabilities were elicited from the surgeon performing the procedure in the trial. The baseline POC severity probability vector was fixed at the elicited prior mean that was used for analysis (Fig. 2). After exploring a range of numerical values, the frailty variance was set equal to 0.20 in the simulations to induce moderate correlation between counts of recurrent events, surgery and PFS.

Table 4 provides the baseline mean burden for each toxicity separately, and their sum or mean TTB. The participating oncologists expect that a typical patient in the trial will experience a TTB score of 33.67, and they believe that atrial fibrillation will contribute the largest component of TTB, followed by pneumonia and anastomotic leak.

A total of 19 scenarios, which are given in Table 5, were simulated to evaluate the design's operating characteristics. Modality effects were induced by adjusting the relative baseline toxicity event rates, recurrent and POC severity probabilities, and difference in PFS hazards. For each

Table 4. Simulation scenarios: baseline mean burden for individual toxicities, and for TTB, obtained from the elicited values in Tables 1 and 3†

	Mean burden for recurrent toxicities						*Mean burden for POCs*					*Total*
	PEF	*PLE*	*RP*	*PNA*	*AFIB*	*MI*	*AL*	*ARDS*	*PEM*	*RI*	*ST*	
μ	1.67	1.23	3.27	6.5	8.63	3.59	4.88	1.17	0.78	1.37	0.59	33.67

†PEF, pericardial effusion; PLE, pleural effusion; RP, radiation pneumonitis; PNA, pneumonia; AFIB, atrial fibrillation; MI, myocardial infarction; AL, anastomotic leak; ARDS, acute respiratory distress syndrome; PEM, pulmonary embolism; RI, reintubation; ST, stroke.

Table 5. Simulation scenarios combining effects for mean TTB and PFS hazard ratio

Scenario	*% reduction in mean toxicity burden for PBT versus IMRT by toxicity*												*PFS HR*
	PEF	*PLE*	*RP*	*PNA*	*AFIB*	*MI*	*AL*	*ARDS*	*PEM*	*RI*	*ST*	*Total*	
0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	75	50	87	4	34	76	26	70	70	34	3	50	1
2	69	56	32	81	34	8	4	22	70	81	82	50	1
3	67	81	29	7	81	19	6	32	3	67	52	50	1
4	70	48	83	11	14	69	62	71	15	78	50	50	1
5	45	63	7	72	13	71	73	14	17	35	38	50	1
6	56	52	83	64	13	69	40	23	9	1	79	50	1
7	57	50	78	55	14	2	75	74	70	28	58	50	1
8	63	85	57	52	9	79	26	5	79	78	74	50	1
9	59	71	79	39	6	53	75	69	71	6	44	50	1
10	65	24	29	17	64	83	22	25	26	74	70	50	1
11	62	73	75	7	78	45	0	0	0	0	0	50	1
12	60	59	60	41	78	32	0	0	0	0	0	50	1
13	0	0	0	0	0	0	0	0	0	0	0	0	1.35
14	0	0	0	0	0	0	0	0	0	0	0	0	1.6
15	0	0	0	0	0	0	0	0	0	0	0	0	1.8
16	0	0	0	0	0	0	0	0	0	0	0	0	2
17	56	52	83	64	13	69	40	23	9	1	79	50	2
18	56	52	83	64	13	69	40	23	9	1	79	50	0.5

†PEF, pericardial effusion; PLE, pleural effusion; RP, radiation pneumonitis; PNA, pneumonia; AFIB, atrial fibrillation; MI, myocardial infarction; AL, anastomotic leak; ARDS, acute respiratory distress syndrome; PEM, pulmonary embolism; RI, reintubation; ST, stroke. Each scenario is characterized by percentage reductions in mean TTB as well as the HR of PFS, for IMRT *versus* PBT. Scenarios 1–12 were obtained by randomly perturbing toxicity incidences and severity probabilities to give a 50% reduction in mean TTB for PBT, while fixing HR = 1. Scenarios 13–16 were obtained by increasing the IMRT *versus* PBT HR for PFS, while fixing the mean TTB difference to be 0. Scenario 17 combines scenarios 6 and 16 to yield a 'win–win' scenario for PBT, with a 50% reduction in mean TTB for PBT and a twofold increase in PFS hazard for IMRT. Scenario 18 combines scenario 6 with a twofold decrease in IMRT *versus* PBT HR for PFS, to yield a 'win–lose' scenario for PBT.

scenario, Table 5 provides the percentage change from baseline in mean toxicity burden for each toxicity, and the IMRT *versus* PBT hazard ratio HR for PFS. Scenario 0 is the global null hypothesis, where the modalities have identical mean TTB and the PFS HR = 1. Scenarios 1–12 characterize alternatives chosen randomly to yield a 50% reduction in mean TTB for PBT *versus* IMRT, with PFS HR $\equiv$ 1. For example, scenario 12 achieves a 50% reduction in the mean TTB for PBT by adjusting the relative event occurrence rates and severity probabilities for recurrent

toxicities *only* to induce 60% reduction for PBT *versus* IMRT in mean toxicity burden for pericardial effusion and 59% reduction for pleural effusion, as well as 60%, 41%, 78% and 32% reductions for radiation pneumonitis, pneumonia, atrial fibrillation and myocardial infarction respectively, in combination with no difference in mean toxicity burden contributed by POCs and equivalent PFS hazard.

For PFS, we evaluated the sensitivity to four different time invariant hazard ratios, while constraining the mean TTB difference $\Delta(\boldsymbol{\theta}, \mathbf{w}) \equiv 0$ in scenarios 13–16. Two additional scenarios were used to evaluate the design's sensitivity to detecting modality effects for both end points. Scenario 17 combines scenarios 6 and 16 to yield a 'win–win' scenario consisting of a 50% reduction in mean TTB for PBT *and* a twofold increase in PFS hazard for IMRT. Scenario 18 combines scenario 6 with a twofold *decrease* in PFS hazard for IMRT, yielding a 'win–lose' scenario for PBT. Posterior probabilities for the TTB and PFS modality comparisons were calculated by using Markov chain Monte Carlo sampling, details of which are given in section C of the on-line supplementary material.

As a comparator, we used a conventional frequentist bivariate group sequential design based on PFS and a binary indicator Y_T of any toxicity at any level of severity by 52 weeks ($t = 1$). For each patient, Y_T was scored as 1 at the time of toxicity, or as 0 at $t = 1$ if no toxicity occurred. A group sequential log-rank test (see for example Klein and Moeschberger (2003)) was used for PFS, and a normal approximation for a two-sample binomial test for toxicity, both implemented with O'Brien–Fleming monitoring boundaries (O'Brien and Fleming, 1979; Jennison and Turnbull, 2000). Although the information times for the binomial data were not identical to the $\mathcal{I}^{\mathrm{TTB}}(\tau)$ that was used for TTB, we formulated the monitoring schedule for toxicity in the conventional design to be as close as possible to that of TTB in the Bayesian design. Additionally, the conventional design was calibrated so that its familywise type I error rate was 0.07 to match that of the Bayesian design. Appendix B describes the process for selecting optimal monitoring boundaries for our design, and it presents the corresponding optimal group sequential critical values that were used to implement the conventional design.

5.2. Operating characteristics

In this section we present operating characteristics for the RT trial when implemented by using our group sequential design based on TTB and PFS, and for the conventional design. Table 6 presents the marginal decision probabilities for scenarios that characterize true treatment effects for only one end point. Corresponding values for the conventional design are given in parentheses. In scenarios 1–12, all of which have true PFS HR $= 1$, our design provides probability 0.83–0.95 of detecting a 50% reduction in mean TTB, while controlling the false positive rate for TTB at 0.02 or lower. The TTB design has early stopping probabilities 0.56–0.72, resulting in a trial with mean sample size 140–153 in these scenarios. In contrast, the conventional design provides probability 0.50–0.81, and in each scenario it is far less likely than the Bayesian design to conclude correctly that PBT is the safer modality and far less likely to stop early, yielding a larger mean sample size. These large differences may be attributed, in large part, to the conventional practice of combining many toxicities into one binary variable and ignoring their severities.

In scenario 16, where the true difference in mean TTB is $\Delta(\boldsymbol{\theta}, \mathbf{w}) \equiv 0$, both designs result in an identical probability of 0.89 to detect a twofold increase in PFS hazard, while controlling the false positive rate for PFS at 4% or lower. However, the Bayesian design is more likely to stop early and offers smaller mean sample size when compared with the conventional design, i.e. the Bayesian design detects the difference in PFS with the same reliability but offers a shorter trial, with mean sample size 139 *versus* 160 for the conventional design.

Table 6. Marginal probabilities of final comparative decisions and early stopping†

Scenario	*Probabilities for the following final decisions for TTB:*			*Early stopping probability*	*Mean sample size*
	PBT	*Indeterminate*	*IMRT*		
(a) TTB with identical PFS hazard					
0	0.01 (0.01)	0.98 (0.98)	0.01 (0.01)	0.04 (0.01)	178 (180)
1	0.83 (0.50)	0.17 (0.50)	0.00 (0.00)	0.56 (0.18)	153 (173)
2	0.89 (0.66)	0.11 (0.34)	0.00 (0.00)	0.66 (0.28)	145 (168)
3	0.93 (0.80)	0.07 (0.20)	0.00 (0.00)	0.70 (0.38)	143 (163)
4	0.85 (0.54)	0.15 (0.46)	0.00 (0.00)	0.58 (0.22)	150 (172)
5	0.89 (0.62)	0.11 (0.38)	0.00 (0.00)	0.64 (0.25)	149 (170)
6	0.90 (0.67)	0.10 (0.33)	0.00 (0.00)	0.68 (0.30)	144 (167)
7	0.87 (0.61)	0.13 (0.39)	0.00 (0.00)	0.60 (0.27)	150 (169)
8	0.86 (0.53)	0.14 (0.47)	0.00 (0.00)	0.59 (0.19)	151 (172)
9	0.87 (0.61)	0.13 (0.39)	0.00 (0.00)	0.61 (0.27)	149 (170)
10	0.90 (0.67)	0.10 (0.33)	0.00 (0.00)	0.65 (0.27)	148 (169)
11	0.93 (0.79)	0.07 (0.21)	0.00 (0.00)	0.71 (0.41)	143 (164)
12	0.95 (0.81)	0.05 (0.19)	0.00 (0.00)	0.72 (0.40)	140 (162)
	Probabilities for the following final decisions for PFS:				
	PBT	*Indeterminate*	*IMRT*		
(b) PFS with identical mean TTB					
0	0.02 (0.02)	0.96 (0.96)	0.02 (0.02)	0.04 (0.01)	178 (180)
13	0.26 (0.27)	0.74 (0.73)	0.00 (0.00)	0.17 (0.08)	169 (177)
14	0.56 (0.57)	0.44 (0.43)	0.00 (0.00)	0.33 (0.25)	162 (172)
15	0.76 (0.80)	0.24 (0.20)	0.00 (0.00)	0.52 (0.41)	149 (167)
16	0.89 (0.89)	0.11 (0.11)	0.00 (0.00)	0.66 (0.52)	139 (160)

†Part (a) considers decisions for TTB under scenarios 0–12. Part (b) considers decisions for PFS under the null hypothesis (scenario 0) and scenarios 13–16. Operating characteristics for a conventional bivariate sequential design using O'Brien–Fleming monitoring boundaries are provided in parentheses.

Table 7 provides the joint probabilities for each of the nine decisions under scenarios 0, 6, 17 and 18. In the null scenario 0, both designs have a familywise false positive rate of 0.07 or less. Moreover, the four corner decisions have a probability of approximately 0.00, so it is very unlikely that either design results in a trial that yields a false positive result for both end points. Recall that scenario 6 corresponds to a 50% reduction in mean TTB with PBT when the PFS HR $= 1$. The TTB design provides much higher probability for concluding that PBT is superior for TTB and indeterminate PFS when compared with the conventional design, 0.88 *versus* 0.66. For scenario 17, the 'win–win' case for PBT, the probability that both of the Bayesian design's tests, for TTB and for PFS, correctly conclude that PBT is superior is 0.20. However, since PBT will be chosen as the superior modality for any of the three decisions 1, 2 or 4 in Table 3, the design provides probability equal to $0.20 + 0.42 + 0.37 = 0.99$ to detect an improvement for at least one end point. Moreover, the Bayesian design has probability 0.89 of terminating early and mean sample size 125, when compared with 0.65 and 152 for the conventional design respectively. The 0.01 false negative result probability is confined only to the global null decision. In scenario 18, the 'win–lose' case for PBT, the Bayesian design has probability 0.21 of making the correct 'win–lose' conclusion, probability 0.33 of concluding that PBT is superior for TTB and indeterminate for PFS, and probability 0.45 of concluding that IMRT is superior for PFS and indeterminate

Table 7. Joint probabilities of final decisions for TTB and PFS for scenarios 0, 6, 17 and 18 by using three interim analyses†

		Probabilities for the following TTB decisions:		
		PBT better	*Indeterminate*	*IMRT better*
Scenario 0: identical mean TTB and PFS hazard (global null)				
PFS decision	PBT better	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
	Indeterminate	0.01 (0.01)	0.93 (0.93)	0.01 (0.01)
	IMRT better	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
Early stopping probability 0.04 (0.01); mean sample size 178 (180)				
Scenario 6: 50% reduction in mean TTB for PBT and PFS HR = 1 (null case)				
PFS decision	PBT better	0.01 (0.005)	0.01 (0.01)	0.00 (0.00)
	Indeterminate	0.88 (0.66)	0.07 (0.31)	0.00 (0.00)
	IMRT better	0.01 (0.005)	0.02 (0.01)	0.00 (0.00)
Early stopping probability 0.68 (0.30); mean sample size 144 (167)				
Scenario 17: 50% reduction in mean TTB for PBT and PFS HR = 2 (in favour of PBT)				
PFS decision	PBT better	0.20 (0.16)	0.42 (0.65)	0.00 (0.00)
	Indeterminate	0.37 (0.15)	0.01 (0.04)	0.00 (0.00)
	IMRT better	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Early stopping probability 0.89 (0.65); mean sample size 125 (152)				
Scenario 18: 50% reduction in mean TTB for PBT and PFS HR = 0.5 (in favour of IMRT)				
PFS decision	PBT better	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Indeterminate	0.33 (0.35)	0.01 (0.02)	0.00 (0.00)
	IMRT better	0.21 (0.29)	0.45 (0.34)	0.00 (0.00)
Early stopping probability 0.88 (0.73); mean sample size 126 (147)				

†Operating characteristics for a conventional bivariate sequential design using O'Brien–Fleming monitoring boundaries are provided in parentheses.

for TTB. Again, the 0.01 false negative probability is confined only to the global null decision and, similarly to scenario 17, the design has probability 0.88 of terminating early.

6. Guidelines for constructing a design with total toxicity burden

When patients are at risk of multiple, qualitatively different toxicities, a Bayesian design based on TTB and PFS may offer a powerful tool for comparing safety and effectiveness between competing treatments. However, many decisions must be made when choosing the set of toxicities, features of the Bayesian model and sequential decision rules. These decision are inherently subjective and require close collaboration between the statisticians and physicians. In this section, we briefly explain the process for constructing a design and evaluating its operating characteristics.

6.1. Eliciting toxicities and severity weights

The two essential components of the TTB statistic are the toxicities and severity weights. The toxicities, and when each may occur in the treatment regime, are identified by the physicians. In essence, there are three fundamental types of toxicities: toxicities that occur at random times

(a) with and

(b) without the possibility of recurrence and
(c) toxicities that are observed only at prespecified evaluation times, such as POCs following surgery.

The next step is to ask the physicians to define the ordinal categories of each toxicity that matter clinically, e.g. in Table 1 the three levels consisting of grade 1–2, 3 and 4–5 for radiation pneumonitis, or simply whether pneumonia occurs. Once this structure has been established, the numerical severity weights must be elicited.

Section A of the on-line supplementary material describes the process that we used to elicit the severity weights in Table 1 and provides the medical rationale put forth by the participating oncologists to justify the resultant numerical values. Our approach should be considered informal in the sense that we did not use an established method (e.g. Hunink *et al.* (2014)), which would have been preferable. For example, a structured communication technique known as the 'Delphi method' (Dalkey and Helmer, 1963; Dalkey, 1969; Brook *et al.*, 1986) could have been used to quantify the relative severity of each possible grade of each toxicity. Additional techniques for elicitation, characterization and use of expert opinion were examined systematically by Cooke (1991). A few researchers have effectuated implementations of such utility-based approaches to clinical cancer studies in recent years. Swinburn *et al.* (2010) conducted detailed interviews with clinical experts to establish the relative burden of a variety of toxicities that are commonly encountered from first-line therapies for metastatic renal cell carcinoma when each is experienced in conjunction with stable *versus* progressive disease. Wong *et al.* (2012) conducted sequential semi-structured interviews with cancer patients to attempt to establish patient priorities for weighing prolonged PFS *versus* inflated risk of multiple types of toxicities that may occur from second-line therapy for renal cell carcinoma.

6.2. Modelling decisions

The model should provide a reasonable representation of the process of treatment and outcome observation, but it must be tractable. One must decide whether the toxicities and other outcomes are positively associated, negatively associated or independent. We characterized the incidence of toxicity by using a multivariate Poisson process and formulated a joint model for the toxicities, surgery and PFS duration by using independently and identically distributed patient frailties. A simple device is to invert the patient frailty, i.e. to use $1/U$, to accommodate negative association, or to omit it for independence. In the RT trial, radiation-induced toxicities were assumed to be influenced by the RT modality (at the group level) and the anatomic location of each patient's tumour (at the patient level). Because the tumour's location within the oesophagus influences how a dosimetric plan is formulated and implemented, incidences of all radiation-induced toxicities were assumed to be positively associated. Because radiation-induced toxicities may decrease survival, and PFS includes death as an event, the incidence of toxicity was assumed to be negatively associated with PFS.

To specify prior hyperparameters, a general approach that works well in practice is to use elicited values to establish means and then to calibrate the variances by using ESS. One should avoid priors that are either excessively informative or unrealistically dispersed. Taking this approach, we specified priors that characterized the expected incidence of each toxicity and marginally contained the amount of information that would be contributed by one or two patients. One also must decide whether treatment effects that determine the relative incidences of each toxicity severity are independent or *a priori* dependent for the therapeutic regimes. We decided that it was appropriate to assume that the treatment effects for the radiation-induced toxicities were exchangeable, and thus we used hyperpriors to induce shrinkage for the oesophageal

RT trial. Finally, the model requires specification of a prior for the frailty variance in relation to an assumed degree of association between the end points on the scale of the Poisson intensity domain. We used a uniform prior with lower bound at 0 (independence) and selected the upper bound 10 to restrict the prior to an interval that excluded an unrealistically high probability of toxicity recurrence.

6.3. Design considerations

Several design features must be considered when planning a TTB-based trial. One first must determine the maximum follow-up duration for which a patient will be monitored for toxicity following treatment. The design requires a schedule for the group sequential tests, based on the information statistics that determine the number and 'timing' of the interim analyses. The schedule should be specified in relation to the trial's expected rate of enrolment and the assumed event rates. In addition, the investigators must fix ε to reflect a difference in mean TTB that should be considered too small to be clinically relevant.

We next describe the process of using simulation to calibrate tuning parameters to obtain targeted operating characteristics. To construct alternative simulation scenarios from combinations of the treatment effects $\boldsymbol{\delta}$, one first must ascertain a 'baseline' value for mean TTB, hereafter denoted by μ_0 (e.g. Table 4) that reflects the expected TTB for a typical patient receiving the control therapy. This can be achieved by accounting for or neglecting prior uncertainty for the model parameters, depending on the investigators' preference. The former approach requires Monte Carlo simulation, whereby one generates model parameters from the priors for the baseline model parameters (treatment effects omitted), obtains a prior distribution for mean TTB and fixes μ_0 at the resulting mean. The latter, more practical approach simply fixes the model parameters at their respective prior means and uses equation (7) to determine μ_0 where $\boldsymbol{\delta} = \mathbf{0}$.

When considering simulation scenarios, one uses μ_0 to identify targeted 'effect sizes' for the power computations by considering scenarios that induce varying degrees of relative difference in true mean TTB between treatments. For example, μ_0 was determined to be 33.67 for the control modality, IMRT, in the oesophageal trial. The sample size was selected to target a 50% reduction in mean TTB for PBT, which we denote by $\%\Delta^*$. Thereafter, one can identify alternative simulation scenarios (e.g. scenarios 1–12, 17 and 18 in Table 5) through computation by selecting treatment effect vectors at random, $\boldsymbol{\delta} = \boldsymbol{\delta}^*$, that achieve the target $\%\Delta^* = 100(\mu^*/\mu_0^* - 1)$, where μ^* and μ_0^* denote true values of mean TTB determined by $\boldsymbol{\delta}^*$ for treatment and control respectively. For our model μ_0^* attains the baseline value $\mu_0^* = \mu_0$ when $\boldsymbol{\delta}^* = \mathbf{0}$.

After establishing the alternative scenarios, one must simulate trials under each alternative and the null hypothesis, $\boldsymbol{\delta}^* = \mathbf{0}$, storing the resulting posterior probabilities that are obtained for each sequential interim analysis. After determining the design's false positive rate, optimal monitoring boundaries can be selected by using the process that is described in Appendix B. In the presence of co-primary end points, an objective function characterizing the relative importance of power for each end point must be defined with respect to at least one alternative scenario before selecting an optimal design (note that Appendix B uses equal costs). Finally, operating characteristics can be obtained through post-processing of the replicate trials by using the optimal monitoring boundaries. The entire simulation process may be repeated multiple times to determine a minimal sample size that detects $\%\Delta^*$ with acceptable power for all alternative scenarios.

7. Discussion

We have proposed a Bayesian design for a randomized clinical trial with group sequential treatment comparisons based on posterior mean TTB and PFS. The complexity of the underlying

probability model reflects the complexity of the disease and therapeutic outcomes. TTB, constructed from subjective severity weights, provides a practical continuous statistic for measuring the extent to which a patient tolerates a therapeutic intervention. The statistical model and corresponding trial design offer powerful tools for sequential safety monitoring in settings wherein patients are at risk of many types of toxicities that may result from each single treatment or stem from the combined effects of multiple types of therapy when used concurrently or administered over a sequence of intervention periods. The trial design reflects the fact that, for many diseases and therapeutic regimes, it is essential to account for both toxicity and efficacy in treatment evaluation.

The problem of handling multiple co-primary end points is quite general and does not pertain specifically to whether Bayesian or conventional frequentist decision rules are used. There is an extensive literature on testing for multiple end points. Some useful references are O'Brien (1984), Cook and Farewell (1996), Wassmer *et al.* (1999) and Jennison and Turnbull (2000). For our comparator in the simulation study, we assumed that PFS and the indicator of toxicity were independent. Alternative approaches that account for association between multiple end points have been proposed. Pocock *et al.* (1987) extended the multiple-end-point testing methods of O'Brien (1984) to address the bivariate problem of combining survival and binary end points by using linear combinations of asymptotically multivariate normal test statistics. Chang *et al.* (1997) considered sequential analysis of paired survival end points by using multivariate counting processes arising from a time-dependent frailty model. Murray (2000) discussed a method for two-sample sequential monitoring of paired censored survival end points based on weighted log-rank statistics. The last two methods could have been used as the comparator in our simulation study by replacing the toxicity indicator with a time-to-event end point. Application of any of the three methods aforementioned may have yielded a more powerful comparator.

An important aspect of our decision rules is that they allow the conclusion that one modality is superior in terms of TTB but yields shorter PFS. Such scenarios are not unlikely in many oncology settings, where qualitatively different or more aggressive treatments may improve PFS or prolong survival, but at the cost of increased severity or incidence of adverse events. The model and joint decision rule that are used in our sequentially adaptive design provide a formal method for treatment comparison based on both safety and efficacy.

Acknowledgements

This research was supported by grants R01-CA083932 (BPH and PFT) and P30-CA016672 (all authors) funded by the National Cancer Institute of the US Department of Health and Human Services. We thank three reviewers for their detailed reviews as well as the Joint Editor whose constructive comments effectuated an improved manuscript. In addition, we thank Allen Chang, Mike Palla and Mark Ford of the MD Anderson Cancer Center for facilitating computational resources as well as Dr Jaffer Ajani and Dr Wayne Hofstetter for contributing to the elicitation process.

Appendix A: Mean total toxicity burden derivation

Here we provide additional details pertaining to the derivation of mean TTB in Section 4.4. Denoting $v = 1/u$ and

$$c = \left\{ \Gamma\left(\frac{1}{\phi} + 2\right) \left(\frac{\phi}{\phi + 1}\right)^{1/\phi + 2} \right\}^{-1},$$

by iterated expectation,

$$\begin{aligned}\mu^{\mathrm{POC}}(t,x,\boldsymbol{\theta}) &= \tilde{\mathbf{w}}'\boldsymbol{\pi}(x)\, E_U[\Pr\{S(t)>0|u,x,\tilde{\lambda},\tilde{\delta}\}|\phi] \\ &= \tilde{\mathbf{w}}'\boldsymbol{\pi}(x)\int_0^\infty \Pr\{S(t)>0|u,x,\tilde{\lambda},\tilde{\delta}\}\,\mathrm{d}\Gamma^{-1}\left(u|\frac{1}{\phi}+2,\frac{1}{\phi}+1\right) \\ &= \tilde{\mathbf{w}}'\boldsymbol{\pi}(x)\int_0^\infty [1-\exp\{-tv\tilde{\lambda}\exp(x\tilde{\delta})\}]v^{1/\phi+1}\exp\left\{-\frac{v(\phi+1)}{\phi}\right\}\mathrm{d}v \\ &= \tilde{\mathbf{w}}'\boldsymbol{\pi}(x)\left(1-c\int_0^\infty v^{1/\phi+1}\exp\left[-v\left\{t\tilde{\lambda}\exp(x\tilde{\delta})+\frac{\phi+1}{\phi}\right\}\right]\mathrm{d}v\right) \\ &= \tilde{\mathbf{w}}'\boldsymbol{\pi}(x)\left[1-\left\{\frac{\phi t\tilde{\lambda}\exp(x\tilde{\delta})}{\phi+1}+1\right\}^{-(1/\phi+2)}\right]. \end{aligned} \tag{8}$$

Appendix B: Selecting optimal sequential monitoring boundaries

A set of optimal sequential monitoring boundaries was selected by using the following process. Initially, we simulated 3000 replications for each of scenarios 0, 6 and 16. For each simulated sequential trial, $g=1,\ldots,3000$, we saved the set of posterior probabilities corresponding to each of the four sequential analyses $\tau_1,\ldots,\tau_4$:

$$\begin{aligned}&\{\varphi^{\mathrm{PB}}(\varepsilon^{\mathrm{TTB}},\tau_1),\varphi^{\mathrm{IM}}(\varepsilon^{\mathrm{TTB}},\tau_1),\chi^{\mathrm{PB}}(\varepsilon^{\mathrm{PFS}},\tau_1),\chi^{\mathrm{IM}}(\varepsilon^{\mathrm{PFS}},\tau_1),\ldots,\varphi^{\mathrm{PB}}(\varepsilon^{\mathrm{TTB}},\tau_4),\\ &\quad\varphi^{\mathrm{IM}}(\varepsilon^{\mathrm{TTB}},\tau_4),\chi^{\mathrm{PB}}(\varepsilon^{\mathrm{PFS}},\tau_4),\chi^{\mathrm{IM}}(\varepsilon^{\mathrm{PFS}},\tau_4)\}^{(g)}.\end{aligned}$$

Any improvement in mean TTB or PFS was considered clinically relevant by the participating oncologists; therefore posterior probabilities were computed by using $\varepsilon^{\mathrm{TTB}}=\varepsilon^{\mathrm{PFS}}=0$.

A gradient optimization method was implemented to select the set of values for the posterior boundary parameters, α^{TTB}, α^{PFS}, β^{TTB} and β^{PFS}, that yielded maximum *total power* for scenarios 1 and 16 among all choices that controlled the familywise type I error at 0.07 or less under scenario 0. We defined the total power to be the sum of the marginal probability that the sequential procedure concluded that PBT was superior to IMRT for TTB in scenario 1 *plus* the marginal probability that the sequential procedure concluded that PBT was superior to IMRT for PFS in scenario 16. The final design used $\alpha^{\mathrm{TTB}}=3.92$ and $\beta^{\mathrm{TTB}}=0.030$, and $\alpha^{\mathrm{PFS}}=0.965$ and $\beta^{\mathrm{PFS}}=0.028$, which produced the operating characteristics that are provided in Tables 6 and 7.

In contrast, the conventional design used O'Brien–Fleming boundaries for both group sequential rules. The operating characteristics in Tables 6 and 7 were computed by using the following critical values for comparing toxicity rates between modalities using sequential z-tests for a difference in proportions: (3.60563, 3.08866, 2.74145, 2.13505). The sequential log-rank testing procedure used the following critical values for z-tests for a difference in PFS corresponding to each of the four analyses: (3.675, 3.217, 2.529, 2.167). The O'Brien–Fleming group sequential critical values were obtained through a two-step process that involved generating a set of candidate boundaries by using statistical software *PASS version 11* (Hintze, 2011) using the closest approximation of the actual interim monitoring schedule, then simulating the TTB design under each candidate boundary and selecting the one that yielded the largest total power among those that controlled the familywise type I error rate at 0.07 or less.

References

Abid, S. H., Malhotra, V. and Perry, M. C. (2001) Radiation-induced and chemotherapy-induced pulmonary injury. *Curr. Opin. Oncol.*, **4**, 242–248.

Bekele, B. N. and Thall, P. F. (2004) Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J. Am. Statist. Ass.*, **99**, 26–34.

Breslow, N. E. (1984) Extra-Poisson variation in log-linear models. *Appl. Statist.*, **33**, 38–44.

Brook, R. H., Chassin, M. R., Fink, A., Solomon, D. H., Kosecoff, J. and Park, R. E. (1986) A method for the detailed assessment of the appropriateness of medical technologies. *Int. J. Technol. Assessmnt Hlth Care*, **2**, 53–63.

Bryant, J. and Day, R. (1995) Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, **51**, 1372–1383.

Chang, I.-S., Hsiung, C. A. and Chuang, Y.-C. (1997) Applications of a frailty model to sequential survival analysis. *Statist. Sin.*, **7**, 127–138.
Conaway, M. R. and Petroni, G. R. (1995) Bivariate sequential designs for phase II trials. *Biometrics*, **51**, 656–664.
Cook, R. J. and Farewell, V. T. (1996) Multiplicity considerations in the design and analysis of clinical trials. *J. R. Statist. Soc.* A, **159**, 93–110.
Cook, R. J. and Lawless, J. F. (2007) *The Statistical Analysis of Recurrent Events.* New York: Springer.
Cooke, R. M. (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York: Oxford University Press.
Cox, D. R. (1955) Some statistical methods connected with series of events (with discussion). *J. R. Statist. Soc.* B, **17**, 129–164.
Cox, D. R. and Isham, V. (1980) *Point Processes.* New York: Chapman and Hall.
Dalkey, N. C. (1969) An experimental study of group opinion. *Futures*, **1**, 408–426.
Dalkey, N. and Helmer, O. (1963) An experimental application of the Delphi method to the use of experts. *Mangmnt Sci.*, **9**, 458–467.
DeMets, D. L. and Lan, K. K. G. (1994) Interim analyses: the alpha spending function approach. *Statist. Med.*, **13**, 1341–1352.
Gelber, R. D., Cole, B. F., Gelber, S. and Goldhirsch, A. (1995) Comparing treatments using quality-adjusted survival: the Q-Twist method. *Am. Statistn*, **49**, 161–169.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–534.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–511.
Hintze, J. (2011) *PASS 11*. Kaysville: NCSS.
Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Pliskin, J. S., Drummond, M. F., Glasziou, P. P. and Wong, J. B. (2014) *Decision Making in Health and Medicine: Integrating Evidence and Values.* Cambridge: Cambridge University Press.
Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer.
Jacobsen, M. (2006) *Point Process Theory and Applications.* Boston: Birkhäuser.
Jennison, C. and Turnbull, B. W. (2000) *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton: Chapman and Hall–CRC.
Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edn. New York: Springer.
Kosorok, M. R., Shi, Y. and DeMets, D. L. (2004) Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics*, **60**, 134–145.
Lan, K. K. G. and DeMets, D. L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
Lawless, J. F. (1987a) Negative binomial and mixed Poisson regression. *Can. J. Statist.*, **15**, 209–225.
Lawless, J. F. (1987b) Regression methods for Poisson process data. *J. Am. Statist. Ass.*, **82**, 808–815.
Morita, S., Thall, P. F. and Müller, P. (2008) Determining the effective sample size of a parametric prior. *Biometrics*, **64**, 595–602.
Morita, S., Thall, P. F. and Müller, P. (2012) Prior effective sample size in conditionally independent hierarchical models. *Baysn Anal.*, **7**, 591–614.
Murray, S. (2000) Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics*, **56**, 984–990.
O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.
O'Brien, P. C. and Fleming, T. R. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
O'Neill, R. T. (2008) A perspective on characterizing benefits and risks derived from clinical trials: can we do more? *Therptc Innovn Regultry Sci.*, **42**, 235–245.
Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987) The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487–498.
Rancati, T., Ceresoli, G. L., Gagliardi, G., Schipani, S. and Cattaneo, G. M. (2003) Factors predicting radiation pneumonitis in lung cancer patients: a retrospective study. *Radtherpy Oncol.*, **67**, 275–283.
Snyder, D. and Miller, M. (1991) *Random Point Processes in Time and Space.* New York: Springer.
Springer, M. D. and Thompson, W. E. (1970) The distribution of products of beta, gamma and Gaussian random variables. *SIAM J. Appl. Math.*, **18**, 721–737.
Swinburn, P., Lloyd, A., Nathan, P., Choueiri, T. K., Cella, D. and Neary, M. P. (2010) Elicitation of health state utilities in metastatic renal cell carcinoma. *Curr. Med. Res. Opin.*, **26**, 1091–1096.
Tang, D. I., Geller, N. L. and Pocock, S. J. (1993) On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, **49**, 23–30.
Tang, D. I., Gnecco, C. and Geller, N. L. (1989a) An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, **76**, 577–583.
Tang, D. I., Gnecco, C. and Geller, N. L. (1989b) Design of group sequential clinical trials with multiple endpoints. *J. Am. Statist. Ass.*, **84**, 776–779.

University of Texas MD Anderson Cancer Center (2015) Phase III randomized trial of proton beam therapy versus intensity-modulated radiation therapy for the treatment of esophageal cancer. In NLM identifier: NCT01512589. *ClinicalTrials.gov*. Bethesda: National Library of Medicine.

Wassmer, G., Reitmer, P., Kieser, M. and Lehmacher, W. (1999) Procedures for testing multiple endpoints in clinical trials: an overview. *J. Statist. Planng Inf.*, **82**, 69–81.

Wathen, J. K. and Thall, P. F. (2008) Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statist. Med.*, **27**, 5586–5604.

Wong, M. K., Mohamed, A. F., Hauber, A. B., Yang, J.-C., Liu, Z., Rogerio, J. and Garay, C. A. (2012) Patients rank toxicity against progression free survival in second-line treatment of advanced renal cell carcinoma. *J. Med. Econ.*, **15**, 1139–1148.

Yusuf, S. W., Sami, S. and Daher, I. N. (2011) Radiation-induced heart disease: a clinical update. *Cardiol. Res. Prac.*, article 317659.

Zhang, X., Zhao, K., Guerrero, T. M., Mcguire, S. E., Yaremko, B., Komaki, R., Cox, J. D., Hui, Z., Li, Y., Newhauser, W. D., Mohan, R. and Liao, Z. (2008) Four-dimensional computed tomography-based treatment planning for intensity-modulated radiation therapy and proton therapy for distal esophageal cancer. *Int. J. Radn Oncol. Biol. Phys.*, **72**, 278–287.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary material for "Bayesian group sequential clinical trial design using total toxicity burden and progression-free survival"'.