JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2020.1821702

Taylor & Francis
Taylor & Francis Group



# Dose-finding design and benchmark for a right censored endpoint

Anaïs Andrillon[a], Sylvie Chevret[a], Shing M Lee[b], and Lucie Biard[a]

[a]INSERM U1153 Team ECSTRRA, Université De Paris, Paris, France; [b]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA

**ABSTRACT**

Dose-finding trials aim to determine a safe dose to be tested in larger trials for efficacy. In oncology, designs generally assume conventional monotonic increasing dose-toxicity relationships, mostly with binary outcomes (e.g., dose-limiting toxicity or not), measured in the first cycle of therapy or for a fixed number of cycles. However, with new anti-cancer agents such as molecularly targeted therapies and immunotherapies, late-onset toxicities have become more frequent. Designs with prolonged observation windows and censored endpoints analyzed using survival models, appear particularly suited to these settings. Moreover, in this setting, the observation of the late-onset toxicity endpoint could be precluded by trial discontinuation due to death, progression, patient withdrawal, or physician discretion, defining a competing event to toxicity. We propose extensions of the Continual Reassessment Method (CRM) dose-finding design using survival working models for right-censored endpoints and for handling treatment discontinuation in the toxicity observation window, namely the Survival-CRM (Surv-CRM) and the informative survival-CRM (iSurv-CRM). We also developed a benchmark approach for its evaluation. In a simulation study, we compared the performance of the Surv-CRM and iSurv-CRM, to those of the Time-to-event (TITE)-CRM and the nonparametric benchmark. The performance of the proposed methods was consistent with the complexity of scenarios as assessed by the nonparametric benchmark. Without treatment discontinuations, the Surv-CRM provides proportions of correct dose selection close to those of the TITE-CRM with fewer observed toxicities and patients assigned to overtoxic dose levels. In the presence of treatment discontinuation, the iSurv-CRM outperforms the TITE-CRM in identifying the correct dose level.

**ARTICLE HISTORY**
Received 1 September 2020
Accepted 2 September 2020

**KEYWORDS**
Dose-finding; survival data; late-onset toxicity; competing risks; treatment discontinuation; benchmark

## 1. Introduction

The general goal of a phase I cancer clinical trial is to define an "acceptable," usually safe, dose level among several candidates. Toxicity is generally assessed in the first cycle of treatment, based on the observation of dose limiting toxicities (DLT), assuming a monotonic increasing dose-DLT relationship. This optimal paradigm does not work when considering the new classes of anti-cancer agents, such as molecularly-targeted therapies (MTAs) and immunotherapies drugs. Indeed, their modes of action, often targeting the immune system or cancer-cell specific pathways, differ from those of cytotoxic chemotherapies. In particular, new agents may confer late-onset toxicities, related to their specific pharmacology or to prolonged administration, contrary to cytotoxic chemotherapies, that are mainly associated with acute severe toxicities and interval administration (Mathijssen et al. 2014; Postel-Vinay et al. 2011; Wages et al. 2018). Thus, new agents require the use of specific designs with prolonged observation windows (Postel-Vinay et al. 2014) and dose optimization for drugs in later phases of development (Lee et al. 2016,

**CONTACT** Anaïs Andrillon. anais.andrillon@u-paris.fr INSERM U1153 Team ECSTRRA, Université De Paris, Paris, France
Supplemental data for this article can be accessed on the publisher's website.

2019). Furthermore, with longer observation windows, treatment discontinuations during the trial, due to death, progression, patient withdrawal or physician discretion may occur. Toxicity assessment may be precluded by termination of the patient's exposure to the drug.

Some designs have been adapted for continuous accrual and incomplete observations during trials, such as the Time-to-Event Continual Reassessment Method (Cheung and Chappell 2000; Wheeler et al. 2018; Yuan and Yin 2009). The TITE-CRM design addresses the issue of late-onset toxicity by allowing staggered patient accrual relaxing the need for complete DLT follow-up of previously included patients. By allowing the enrollment of new patients when previous patients are still under observation, it is suitable for long observation windows and late-onset toxicities. Moreover, it can shorten the duration of a trial compared to the standard CRM. Although the TITE-CRM handles incomplete observations it relies on binary endpoints for toxicity and does not take into account the time to occurrence of toxicity. Furthermore, even if some practical strategies address treatment discontinuation, operating characteristics of the design may be impacted (Biard et al. 2020). To handle late-onset toxicities while allowing for continuous accrual, we propose new dose-finding designs, the survival-CRM (Surv-CRM) and the informative survival CRM (iSurv-CRM), that formally include the information of time to toxicity using a survival working model for right-censored endpoints, thus allowing the toxicity outcomes to be delayed or unobserved due to competing events within the trial observation window.

As for any clinical trial design, it is crucial to have diagnostic tools to evaluate the performance of a newly proposed design in term of false/correct trial conclusion. O'Quigley et al. (2002) first introduced the nonparametric optimal benchmark to provide an upper bound estimate on the performance of a dose-finding design in the setting of binary toxicity endpoints, under a given scenario, and at a fixed sample size. These benchmark values for dose selection provide meaningful information on the complexity of scenario in terms of maximum tolerated dose (MTD) identification and provide a reference for the performance of a dose-finding algorithm with complete potential information, that is, as if outcomes could be observed at all dose levels for all patients. It contrasts with real clinical trials, in which patients are allocated to receive only one dose level, thus resulting in *incomplete information* due to non observation of patient outcomes at all other dose levels. Cheung (2014) developed the benchmark for phase I/II clinical trials simultaneously evaluating binary toxicity and efficacy endpoints, or for phase I trials with multiple toxicity endpoints. Similarly, Mozgunov et al. (2020) generalized a benchmark for dose finding designs to various settings with several discrete and continuous outcomes. To our knowledge, no specific benchmark approaches for survival dose-finding models has been implemented.

In this work, we propose the survival-CRM (Surv-CRM) design building on the CRM dose-finding design, but using survival models for right-censored DLT endpoints allowing the outcomes to be delayed. To handle possible informative censoring, we also propose the informative survival-CRM (iSurv-CRM), that extends the Surv-CRM by considering a competing-risk model. In addition, we develop a nonparametric benchmark approach for evaluation of dose-finding designs with right censored time-to-event endpoints.

## 2. Methods

We consider dose-finding clinical trials where patients are followed-up for an observation window $[0, t^*]$ defined for DLT observation and safety assessment. By time $t^*$, a patient may have two different outcomes: DLT under treatment or not. In the case of novel agents, the need for prolonged observation windows leads us to consider a time-to-event working model which allows for right-censored observations.

Denote $n$ the maximum sample size of the trial. Let $\mathbb{D} = \{d_1, d_2, \ldots, d_m\}$ be the set of numerical labels for the $m$ doses investigated in the trial. We assumed no censoring due to lost to follow-up.

Given our framework of oncology phase I dose-finding trials, where patients are usually in advanced stage of their disease, such an assumption is likely valid.

## 2.1. Problem formulation

### 2.1.1. Survival-CRM design (Surv-CRM)

We first considered censoring occurs during trial accrual only, at the time a dose assignment computation, when a fraction of the toxicity window has been observed for some patients and they have not developed any DLT yet. In this setting, censored observations are administrative, and thus independent from the time-to-toxicity endpoint, and could be considered as noninformative.

Let $T$, be the time to first DLT and define $\lambda(t)$ the instantaneous hazard function, and $\Lambda(t)$ the cumulative hazard function:

$$\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s. \tag{1}$$

The survival function $S(t)$, that is the probability of being free of DLT at time $t$, is given by:

$$S(t) = exp(-\Lambda(t)). \tag{2}$$

and the cumulative incidence function, $F(t) = Prob(T \leq t) = 1 - S(t)$, can be expressed as:

$$F(t) = \int_0^t f(s)\,ds \;=\; \int_0^t \lambda(s)\,S(s)\,ds \tag{3}$$

where $f(\cdot)$ is the density function of the time to DLT, $T$.

At the end of a dose-finding trial, given an observation window $[0, t^\star]$, for each included patient , $i, i = 1, \ldots, n$, observations consist in pairs $(X_i^\star, Y_i^\star)$, where $X_i{}^\star = min(T_i, t^*)$ and $Y_i = \mathbb{I}\,(T_i \leq t^*)$, with $I(\cdot)$ the indicator function ($Y_i = 1$ if the patient experienced a DLT before $t^*$, or $Y_i{}^* = 0$ otherwise).

During the trial, at the time of the dose assignment for a new patient, DLT observations of previously included patients may be right-censored, if only a fraction of the observation window has passed and the patient has not developed any DLT yet. For the MTD estimation during the trial at date $\tau$, we therefore updated data on patient $i$ such as: $(X_i^\tau, Y_i^\tau)$ where $X_i{}^\tau = min(T_i, \tau - \tau_1^0)$ and $Y_i{}^\tau = \mathbb{I}\,(T_i \leq \tau - \tau_i^0)$, where $\tau_i^0$ is the date of arrival of patient $i$.

The proposed survival-CRM is an extension of the CRM with an exponential working model for the cumulative incidence of DLT. Specifically, we assumed $T$ to be exponentially distributed and we assumed the instantaneous hazard of toxicity as an increasing function of the dose:

$$\lambda_j(\beta, d_j) = exp(d_j \times exp(\beta)),\;\; \text{for } j \in \{1, \ldots, m\} \tag{4}$$

where $\beta$ is the unknown model parameter considered a random variable with a prior distribution $\Phi(\beta)$ assumed. Given our setting of right-censored data, the exponential distribution was chosen to model failure times. The cumulative incidence of toxicity at the end of observation window $t^*$ for dose $d_j \in \mathbb{D}$ is then defined as follows:

$$F(t^*, \beta, d_j) \;=\; 1 - exp(t^* \times \exp\,(d_j \times \exp(\beta))) \tag{5}$$

The survival likelihood on $n$ patients can be expressed as follows:

$$L_n(X, Y|\beta) = \prod_{i=1}^{n} f(X_i, \beta, d_i)^{Y_i} S\big(X_j, \beta, d_i\big)^{1-Y_i} \tag{6}$$

with $d_i \in \mathbb{D}$, the dose allocated to patient $i$, $X_i$ the right-censored time-to-toxicity and $Y_i$ the toxicity outcome.

Given the prior distribution of $\beta$, $\Phi(\beta)$, the posterior distribution for $\beta$ is estimated using Bayes formula:

$$\Phi(\beta|\,(X,Y)) \;=\; \frac{L_n(X,Y|\,\beta)\Phi(\beta)d\beta}{\int_{-\infty}^{\infty} L_n(X,Y|\beta)\Phi(\beta)d\beta} \tag{7}$$

and the posterior mean $\hat{\beta}_n$ can be plugged in equation (5) to obtain estimates of the DLT cumulative incidence at each dose level:

$$\hat{\beta}_n \;=\; \frac{\int_{-\infty}^{\infty} \beta\, L_n(X,Y|\,\beta)\Phi(\beta)d\beta}{\int_{-\infty}^{\infty} L_n(X,Y|\beta)\Phi(\beta)d\beta} \tag{8}$$

### 2.1.2. Informative survival-CRM design (iSurv-CRM)

We then considered non negligible occurrences of treatment discontinuation in the observation window $[0, t^*]$ due to disease progression, death, withdrawal, and physician discretion, that preclude complete or reliable DLT assessment. In this setting, DLT and discontinuation of treatment are mutually exclusive events within $[0, t^*]$. We thus extended the survival-CRM (Surv-CRM) to the informative survival-CRM (iSurv-CRM) by considering a competing-risk working model in the dose-finding setting.

In a competing risks framework (Putter et al. 2006), let $k = 1$ denote the occurrence of DLT and $k = 2$ that of treatment discontinuation. Let $\lambda_{jk}(t)$ be the cause-specific hazard and $\Lambda_{jk}(t)$ the cause-specific cumulative hazard of cause $k$ at dose $d_j$. The cause-specific hazard for toxicity $\lambda_{j1}(\cdot)$ was defined according to equation (4) and the cause-specific hazard of discontinuation $\lambda_{j2}(\cdot)$ as follows:

$$\lambda_{2j}(\beta_2,\, d_j) \;=\; \exp(-d_j \times \exp(\beta_2)), for\, j \in \{1, ...., m\} \tag{9}$$

resulting in a decreasing function of the dose. The cumulative incidence of DLT, at time $t^*$ for dose level $j$, corresponding to a subdistribution function, depends on cause 2 through:

$$F_1(t^*, \lambda_{1j},\, \lambda_{2j}) \int_0^{t*} \lambda_{1j} \,\times\, S(s)ds \;=\; \frac{\lambda_{1j} \,\times\, (1 - exp(-(\lambda_{1j} + \lambda_{2j}) \times t^*))}{\lambda_{1j} + \lambda_{2j}} \tag{10}$$

where the event-free survival at time $t$ is defined by:

$$S(t) \;=\; exp(-(\Lambda_1(t) + \,\Lambda_2(t))) \tag{11}$$

Inference on cause–specific hazards may then be performed separately for toxicity and treatment discontinuation, since the log-likelihood factors into two pieces, one involving $\lambda_1$ and the other involving $\lambda_2$ (Jeong and Fine 2006).

## 2.2. Dose-finding algorithm

The objective of the dose-finding trial is to identify the maximum tolerated dose (MTD), $d^*$, defined as the dose which has the estimated cumulative incidence of toxicity at time $t^*$ closest to a pre-specified target $\pi_{DLT}$ among the set $\mathbb{D}$ of candidate dose levels $d_j, j = 1 \ldots, m$:

$$d^* \;=\; \arg\min_{d_j \in\, \mathbb{D}} |\,\hat{F}\,(t^*, \beta, d_j) - \pi_{\,DLT}| \tag{12}$$

In the setting of competing risks, it becomes

$$d^* = \arg\min_{d_j \in \mathbb{D}} |\hat{F}_1(t^*, \lambda_{1j}, \lambda_{2j}) - \pi_{DLT}| \quad (13)$$

#### 2.2.1. Trial design implementation

In the Surv-CRM and iSurv-CRM designs presented above, dose assignments of new patients are performed from the sequentially updated probabilities of DLT, based on available data at the time of analysis. Sequential estimates of the cumulative incidences of DLT were used to assign the estimated dose $\hat{d}^*$ to the next patient cohort. We used a one-stage CRM design, with a Bayesian estimation of the posterior distribution of parameter $\beta$ (equation (7)), with a normal prior for $\beta$, with mean 0 and standard deviation $\sqrt{1.34}$ (O'Quigley and Shen 1996). This prior allowed a reasonable compromise between flexibility and severity for the skeleton, though sensitivity analyses to this choice were also performed. Then, for patient $i$'s dose assignment, initiating a new cohort, we computed the posterior mean $\hat{\beta}_i$ of the model parameter (equation (7)) based on data history $H_i = (X_{i-1}, Y_{i-1})$, and used it as plug-in estimate to obtain estimated cumulative incidences of DLT for each dose level $j$. The new cohort, starting with patient $i$, was allocated to the MTD estimate, defined using (13), given $H_i$. The assignment continued in a sequential fashion until the prespecified sample size $n$ was reached.

#### 2.2.2. Benchmark implementation

For the benchmark, we assumed the whole information about a patient's toxicity outcomes could be summarized in a tolerance profile drawn from a uniform distribution $U(1, 0)$. Precisely, for patient $i$, with profile $u_i$, the quantile transformation $T_{ij} = F^1(u_i, \lambda_j)$ was applied to obtain the time to DLT outcome that this patient would have had at each dose level $j$, with $F$ the assumed cumulative distribution function of time-to-toxicity, in our case an exponential distribution. Independent administrative censoring was then applied at time $t^*$ to obtain trial observations. Basically, for a given patient $i$, this results in a set of outcomes at each of the $m$ candidate dose levels, $(X^*_{i1}, Y^*_{i1}), \ldots, (X^*_{im}, Y^*_{im})$, $i = 1, \ldots, n$, also called the *complete information* about patient $i$.

Given our setting of right-censored data and in the absence of informative censoring, the non parametric Kaplan Meier estimator was used to estimate the cumulative incidences of DLT at each dose $j$, $\hat{F}(t^*, \lambda_j)$. In the presence of informative censoring, cumulative incidences were estimated as described by Gray (1988). Non parametric estimators for the benchmark were chosen, similarly to the original benchmark proposal for binary endpoints (O'Quigley et al. 2002). In case of a limited sample size $n$, there might be no DLTs at a given dose level; in these cases, $\hat{F}(t^*, \lambda_j)$ was set to 0. Last, the MTD, $d^*$, was estimated from the complete dataset as the dose level associated with the estimated probability of toxicity at time $t^*$, $\hat{F}(t^*, \lambda_j)$ closest to $\pi_{DLT}$ (equation (13)).

The algorithm presented in Table 1 (Supplementary materials) provides step-by-step guidance on how the benchmark can be constructed based on simulated patient data.

### 2.3. Simulation study

We simulated trials to assess the performance of our proposed Surv-CRM and iSurv-CRM designs, in comparison to the TITE-CRM and the proposed benchmark.

#### 2.3.1. Data generation

Assuming that the set of probabilities of DLT at time $t^*$ at each dose level is known, complete trial data were generated according to the tolerance profile procedure presented for the benchmark. Inverse transform sampling was applied on the time to DLT cumulative distribution function. We considered time-constant and time-varying toxicity hazards of DLT, using Weibull distributions. Values of 1, 0.5, and 3 were used for the shape parameter to obtain time-constant, decreasing or increasing hazards respectively, given a standardized timeframe for observation $t^* = 1$ (Figure 1), Supplementary

materials). In case of time-varying hazard, Weibull scale parameters were computed from the cumulative incidence of toxicity at time $t^*$ for each scenario, and according to the desired shape value.

Independent administrative censoring was applied at time $t^*$ to mimic the trial observation window and obtain $X_i = min(T_i, t^*)$ and $Y_i = \mathbb{I}(T_i \leq t^*)$. While the trial design only uses the observed outcomes with the doses assigned following the trial algorithm, the performance of the benchmark is obtained using the *complete information* on patient outcomes at each dose level, hence providing the best performance for the dose-finding objective.

Competing risks data were simulated following Beyersmann et al. (2009). Observations were sampled from an exponential model for the time to first event $T$ (event-free-survival), with all-cause hazard $(\lambda_1 + \lambda_2)$ and the cause of failure at the sampled time $T_i$ determined by a random draw from a Bernoulli distribution with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$ for toxicity (otherwise treatment discontinuation).

### 2.3.2. Simulation setting

Without loss of generality, we set the observation window $t^*$ to 1 (one time unit). We considered $m = 5$ candidate dose levels, and, within the time window $t^*$, a target toxicity probability of DLT at $\pi_{DLT} = 0.25$. The sample size was set at $n = 25$, with cohort size of 1.

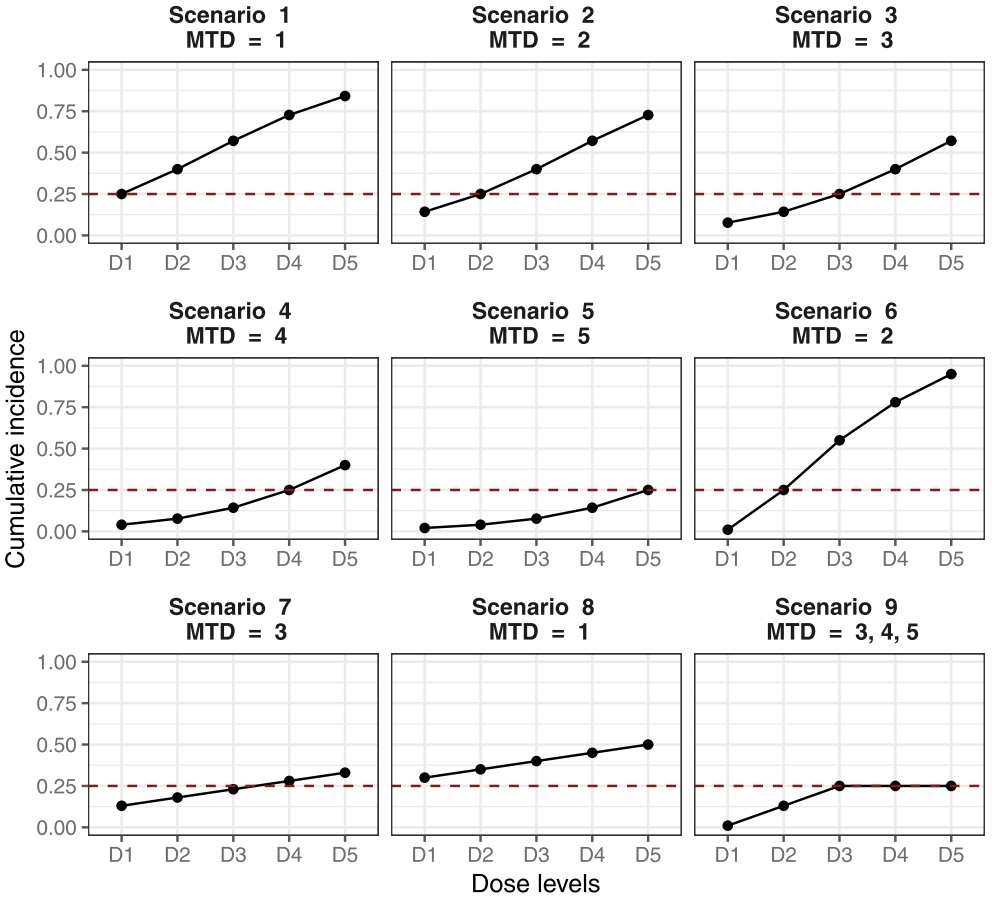


**Figure 1.** Simulation scenarios: cumulative incidence of DLT at time $t^*$ by dose. Simulated trials included $m = 5$ candidate dose levels.

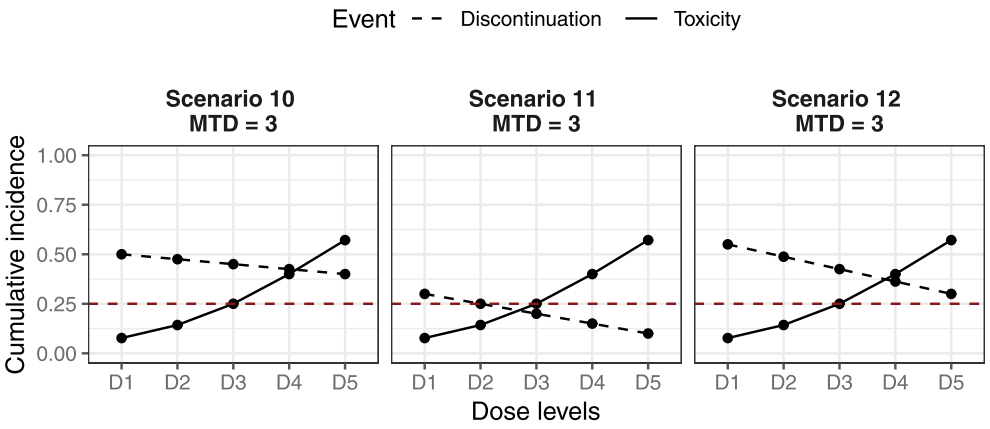


**Figure 2.** Simulation scenarios: cumulative incidence of DLT (solid line) and discontinuation (dashed line) at time $t^*$ by dose. Simulated trials included $m = 5$ candidate dose levels.

Different non-parametric scenarios of dose-toxicity relationships were examined (Figure 1). They were defined based on the cumulative incidence of DLT at time $t^*$ by dose level and they encompass different shapes of dose-DLT relationships, including monotone increasing (Sc1–Sc8, cytotoxicity scenarios) and plateau (Sc9). For the first five scenarios (Sc1–Sc5), the same interval in DLT probability was chosen between dose levels, with the MTD shifted from dose 1 to dose 5, respectively. The remaining four scenarios were sensitivity scenarios. For scenarios 6 (Sc6), a steeper dose-toxicity curve around the MTD was considered. Scenarios 7 and 8 are similar to scenarios 3 and 1, respectively, but with a less steep slope. Simulations with competing events were performed under three additional scenarios, Sc10 to 12 illustrated in Figure 2. We examined three different dose-discontinuation relationships, combined with toxicity scenario Sc3, previously presented. We considered high and moderate risks of discontinuation with cumulative incidence at $t^*$ decreasing from 0.5 to 0.4 respectively for scenario 10, and from 0.3 to 0.1 for scenario 11. In scenario 12, the cumulative incidence ranges from 0.55 to 0.30.

For a comparison purpose, we also applied a TITE-CRM design to the simulated data. We used the Bayesian framework with an empirical dose-toxicity model and a linear weighting scheme. We applied the TITE-CRM design to all scenarios, without (Sc1–9) and with (Sc10–12) competing discontinuation. For Sc10–12, we handled incomplete toxicity observations caused by competing discontinuation by assigning and maintaining weights throughout the trial to patients without DLT who discontinued the trial during the observation window.

Designs skeletons were calibrated using the indifference interval approach (Lee and Cheung 2009; Wheeler et al. 2019), applied on the toxicity working model estimating the cumulative incidence of toxicity at the end of the observation window defined in equation (5) for Surv-CRM, and based on the empirical model for the TITE-CRM. For the Surv-CRM, we implemented the algorithm described by Lee and Cheung (2009) with target toxicity probability $\pi_{DLT} = 0.25$, $m = 5$ dose levels and dose level 3 as anticipated MTD. Based on 2,000 simulations, on a set of scenarios defined following Lee and Cheung (2009), we set the halfwidth of the indifference interval at 0.05, as providing the highest PCS corresponding to the following skeleton {0.069, 0.151, 0.250, 0.346, 0.426}. For the iSurv-CRM, we calibrated the toxicity skeleton using the best guess prior proposed by Polley (2011), i.e. {0.05, 0.10, 0.25, 0.35, 0.50} and we set the prior estimates of the treatment discontinuation probabilities to {0.50, 0.45, 0.40, 0.35, 0.35} based on clinical considerations. For the TITE-CRM, the halfwidth of the indifference interval was set at 0.07, based on the recommendations of Lee and Cheung (2009), corresponding to the following skeleton: {0.043; 0.124; 0.250; 0.398; 0.542}.

Simulated trials were initiated with the first dose level, $d_1$. The cohort size was one patient and no dose skipping was allowed during dose escalation. The trial was terminated either when the maximum sample size was reached, or for safety decisions, i.e. if the posterior probability of $F(t^\star|d_1)$ being above the target $\pi_{DLT}$ was at least 0.95. In the latter, we considered all doses over-toxic and terminated the trial. A modified Surv-CRM with adaptive wait time between two consecutive patients was also examined following the rule proposed by Polley (2011). As a sensitivity analysis, we considered time-constant and time-varying toxicity hazards of DLT and examined different patient accrual schemes by varying the expected number of arrivals per observation window from 1 to 8. Varying the patient-to-window ratio allowed us to simulate short or long observation windows.

### 2.3.3. Performance measures

A total of $N = 10,000$ independent replications of each scenario were run, using either the survival-CRM or the TITE-CRM design, while benchmark performances were obtained from complete data. Based on those $N$ replicates, for each design, we computed the probability of correct selection (PCS) of the true MTD and safety indexes such as the probability of overdose selection (POS), the average number of patients who experienced a DLT during the trial, and the overdose (OD) number, defined as the average number of patients treated at a dose above the true MTD during the trial (Cheung 2011). We also computed the percentage of early trial stopping for safety decisions, ($P_{stop}$).

Finally, we calculated the accuracy index, proposed by Cheung (2011), that incorporates information at all dose levels into a single summary measure:

$$\mathcal{A}_n = 1 - m \times \frac{\sum_{j=1}^{m} |p_j - \pi_{DLT}| \times PS_j}{\sum_{j=1}^{m} |p_j - \pi_{DLT}|}$$

where $p_j$ is the true probability of DLT for dose level $j$ and $PS_j$ is the probability of selecting dose level $j$ at the end of the trial. The index summarizes the distribution of the selected doses through a weighted average, accounting for the discrepancy of the scenario around the target, $\pi_{DLT}$. It ranges from $1 - m$ to 1, reached when the probability of selecting the MTD is equal to 1 and the true probability predefined associated to the MTD is equal to the target. This measure penalizes selecting doses distant from the true MTD: $A_n$ could be negative when the probability of selecting wrong doses is higher than or equal to $1/m$.

## 3. Results

Tables 1 and 2 summarize the performance of both trial designs and the benchmark, for respectively the five main scenarios (Sc1–Sc5) and the four sensitivity scenarios (Sc6–Sc9), when patient accrual is set at four patients per time window, in the absence of competing discontinuations.

First, as expected, the performance of the methods depended on the scenario. In the case of constant hazards of toxicity over time, the highest benchmark accuracy index, 0.97, was reached for the steepest scenario (Sc6), while the lowest, 0.13, was observed with the flatest one (Sc7); corresponding benchmark PCS were 96% and 24%, respectively. The performance of the proposed Surv-CRM design were in line with these findings, with accuracy index of 0.81 for Sc6 and 0.19 for Sc7. Indeed, in case of a flat dose-toxicity relationship with dose level 3 being the MTD (Sc7), the proposed design failed to identify the right dose the majority of time (PCS of 30%), in line with the benchmark (PCS of 24%). However, in scenario 7, $d_3$ and $d_4$ have DLT probability of 0.23 and 0.28 respectively, so that both are reasonable MTD candidates. Indeed, the proportion of correct selection within a 5%-acceptable region, which captures dose candidates lying closely to the true MTD, is 53% (result not shown). In the case of a true plateau relationship (Sc9), the PCS was 83% for the Surv-CRM versus 80% for the benchmark. For scenarios 2 to 4, with the MTD shifted from dose 2 to 4, PCS was 52% for Sc2,

**Table 1.** Simulation results for Sc1 to Sc5 of the survival-CRM, TITE-CRM and benchmark: percent of correct selection (PCS); accuracy index ($A_{25}$); percent of overdose selection (POS); percent of stopped trials for safety ($P_{stop}$); relative bias (R. bias in estimated probability of DLT at the true MTD); average overdose number (OD); number of observed DLT (No. DLT) and number of patients treated with the true MTD (No. MTD) during the trial. $N$=10,000 simulated trials with with $n$=25 and $\pi_{DLT} = 25\%$. Accrual rate of four patients per observation window. (n/a: not applicable).

| Method | Hazard | PCS | $A_{25}$ | POS | $P_{stop}$ | R. Bias | OD | No. DLT | No. MTD |
|---|---|---|---|---|---|---|---|---|---|
| **S1: 0.25**; 0.40; 0.57; 0.73; 0.84 | | | | | | | | | |
| Surv-CRM | decreasing | 79 | 0.89 | 21 | 18 | 0.044 | 7.52 | 7.28 | 17.48 |
| | constant | 79 | 0.89 | 21 | 12 | 0.032 | 8.57 | 7.78 | 16.43 |
| | increasing | 76 | 0.88 | 24 | 4 | 0.004 | 10.88 | 8.79 | 14.12 |
| TITE-CRM | decreasing | 74 | 0.87 | 26 | 14 | −0.083 | 9.07 | 7.70 | 15.93 |
| | constant | 75 | 0.87 | 25 | 12 | −0.073 | 9.97 | 8.05 | 15.03 |
| | increasing | 76 | 0.88 | 24 | 9 | −0.053 | 11.56 | 8.78 | 13.44 |
| Benchmark | | 80 | 0.90 | 20 | . | 0.000 | . | . | . |
| **S2**: 0.14; **0.25**; 0.40; 0.57; 0.73 | | | | | | | | | |
| Surv-CRM | decreasing | 50 | 0.71 | 18 | 2 | 0.073 | 5.81 | 6.29 | 8.95 |
| | constant | 52 | 0.71 | 19 | 1 | 0.054 | 6.90 | 6.73 | 9.07 |
| | increasing | 54 | 0.71 | 22 | 0 | 0.015 | 9.20 | 7.64 | 9.09 |
| TITE-CRM | decreasing | 57 | 0.73 | 21 | 1 | 0.018 | 6.85 | 6.71 | 10.06 |
| | constant | 57 | 0.73 | 21 | 1 | 0.021 | 7.71 | 7.04 | 9.90 |
| | increasing | 57 | 0.74 | 20 | 1 | 0.027 | 9.35 | 7.71 | 9.49 |
| Benchmark | | 57 | 0.74 | 19 | . | 0.000 | . | . | . |
| **S3**: 0.08; 0.14; **0.25**; 0.40; 0.57 | | | | | | | | | |
| Surv-CRM | decreasing | 49 | 0.56 | 16 | 0 | 0.090 | 4.99 | 5.62 | 8.20 |
| | constant | 50 | 0.57 | 18 | 0 | 0.070 | 5.98 | 6.02 | 8.45 |
| | increasing | 53 | 0.59 | 20 | 0 | 0.031 | 8.06 | 6.80 | 8.59 |
| TITE-CRM | decreasing | 53 | 0.59 | 19 | 0 | 0.072 | 5.74 | 5.99 | 9.04 |
| | constant | 54 | 0.60 | 18 | 0 | 0.070 | 6.50 | 6.28 | 9.08 |
| | increasing | 55 | 0.61 | 18 | 0 | 0.067 | 8.03 | 6.84 | 8.91 |
| Benchmark | | 57 | 0.63 | 19 | . | 0.000 | . | . | . |
| **S4**: 0.04; 0.08; 0.14; **0.25**; 0.40 | | | | | | | | | |
| Surv-CRM | decreasing | 47 | 0.48 | 15 | 0 | 0.106 | 3.94 | 4.98 | 7.91 |
| | constant | 49 | 0.50 | 16 | 0 | 0.084 | 4.76 | 5.29 | 8.26 |
| | increasing | 51 | 0.51 | 20 | 0 | 0.046 | 6.47 | 5.84 | 8.59 |
| TITE-CRM | decreasing | 50 | 0.51 | 17 | 0 | 0.118 | 4.25 | 5.22 | 8.76 |
| | constant | 51 | 0.52 | 17 | 0 | 0.113 | 4.90 | 5.44 | 8.88 |
| | increasing | 52 | 0.53 | 17 | 0 | 0.105 | 6.20 | 5.82 | 8.92 |
| Benchmark | | 57 | 0.57 | 19 | . | 0.000 | . | . | . |
| **S5**: 0.02; 0.04; 0.08; 0.14; **0.25** | | | | | | | | | |
| Surv-CRM | decreasing | 61 | 0.67 | n/a | 0 | 0.105 | n/a | 4.02 | 10.13 |
| | constant | 63 | 0.70 | n/a | 0 | 0.084 | n/a | 4.19 | 11.14 |
| | increasing | 69 | 0.75 | n/a | 0 | 0.049 | n/a | 4.47 | 12.89 |
| TITE-CRM | decreasing | 63 | 0.70 | n/a | 0 | 0.164 | n/a | 4.13 | 10.70 |
| | constant | 64 | 0.71 | n/a | 0 | 0.158 | n/a | 4.25 | 11.39 |
| | increasing | 65 | 0.71 | n/a | 0 | 0.148 | n/a | 4.45 | 12.64 |
| Benchmark | | 75 | 0.80 | n/a | . | 0.000 | . | . | . |

50% for Sc3 and 49% for Sc4 compared to 57% with the benchmark. Of note, with flat scenarios, Sc7 and Sc9, the Surv-CRM design slightly outperformed the benchmark.

Additionally, with constant hazards, the probability of overdose selection (POS) with the survival-CRM was slightly lower to that of the TITE-CRM for most of scenarios. In particular, for Sc1 and Sc8, with MTD at the first dose level, the POS was 21% and 32% with Surv-CRM respectively, versus 25% and 43% with the TITE-CRM. In terms of patient safety during the trials, the average number of patients assigned to a dose above the MTD (OD) during the trial ranged from 4.79 to 11.29 using the survival-CRM versus 4.90 to 13.19 using the TITE-CRM. Finally, the average number of patients who experienced a DLT during the trial (No. DLT) was slightly lower with the Surv-CRM compared to the TITE-CRM for all scenarios. All selection performance measures were not markedly modified when the proportion of right-censored observations decreased, due to a slower accrual rate (Table 2, Supplementary material).

**Table 2.** Simulation results for Sc6 to Sc9 of the survival-CRM, TITE-CRM and benchmark: percent of correct selection (PCS); accuracy index ($A_{25}$); percent of overdose selection (POS); percent of stopped trials for safety ($P_{stop}$); relative bias (R. bias in estimated probability of DLT at the true MTD); average overdose number (OD); number of observed DLT (No. DLT) and number of patients treated with the true MTD (No. MTD) during the trial. $N$=10,000 simulated trials with with $n$=25 and $\pi_{DLT} = 25\%$. Accrual rate of 4 patients per observation window. (n/a: not applicable).

| Method | Hazard | PCS | $A_{25}$ | POS | $P_{stop}$ | R. Bias | OD | No. DLT | No. MTD |
|---|---|---|---|---|---|---|---|---|---|
| **S6**: 0.01; **0.25**; 0.55; 0.78; 0.95 | | | | | | | | | |
| Surv-CRM | decreasing | 75 | 0.81 | 14 | 0 | −0.003 | 4.72 | 6.22 | 13.27 |
| | constant | 76 | 0.81 | 13 | 0 | 0.001 | 5.60 | 6.75 | 12.95 |
| | increasing | 76 | 0.82 | 12 | 0 | 0.006 | 7.24 | 7.85 | 12.42 |
| TITE-CRM | decreasing | 82 | 0.86 | 12 | 0 | −0.021 | 4.91 | 6.67 | 14.66 |
| | constant | 82 | 0.86 | 10 | 0 | −0.001 | 5.51 | 6.99 | 14.12 |
| | increasing | 80 | 0.85 | 7 | 0 | 0.033 | 6.64 | 7.67 | 13.01 |
| Benchmark | | 96 | 0.97 | 3 | . | 0.000 | . | . | . |
| **S7**: 0.13; 0.18; **0.23**; 0.28; 0.33 | | | | | | | | | |
| Surv-CRM | decreasing | 29 | 0.15 | 32 | 1 | 0.116 | 7.46 | 5.36 | 5.64 |
| | constant | 30 | 0.19 | 36 | 0 | 0.061 | 8.58 | 5.57 | 5.80 |
| | increasing | 30 | 0.23 | 44 | 0 | −0.036 | 10.72 | 5.95 | 5.84 |
| TITE-CRM | decreasing | 30 | 0.21 | 37 | 1 | 0.077 | 8.41 | 5.55 | 5.92 |
| | constant | 31 | 0.22 | 39 | 1 | 0.042 | 9.26 | 5.71 | 6.03 |
| | increasing | 31 | 0.25 | 43 | 0 | −0.021 | 10.90 | 5.99 | 6.03 |
| Benchmark | | 24 | 0.13 | 47 | . | 0.001 | . | . | . |
| **S8: 0.30**; 0.35; 0.40; 0.45; 0.50 | | | | | | | | | |
| Surv-CRM | decreasing | 71 | 0.53 | 29 | 27 | −0.158 | 10.20 | 7.47 | 14.80 |
| | constant | 68 | 0.51 | 32 | 18 | −0.184 | 11.29 | 7.86 | 13.71 |
| | increasing | 61 | 0.47 | 39 | 6 | −0.249 | 13.76 | 8.54 | 11.24 |
| TITE-CRM | decreasing | 59 | 0.47 | 41 | 20 | −0.316 | 12.08 | 7.87 | 12.92 |
| | constant | 57 | 0.46 | 43 | 16 | −0.327 | 13.19 | 8.10 | 11.81 |
| | increasing | 53 | 0.43 | 47 | 10 | −0.354 | 15.29 | 8.55 | 9.71 |
| Benchmark | | 69 | 0.51 | 31 | . | −0.002 | . | . | . |
| **S9**: 0.01; 0.13: **0.25; 0.25; 0.25** | | | | | | | | | |
| Surv-CRM | decreasing | 81 | 0.67 | n/a | 0 | 0.148 | n/a | 4.93 | 16.58 |
| | constant | 83 | 0.71 | n/a | 0 | 0.105 | n/a | 5.12 | 17.77 |
| | increasing | 88 | 0.79 | n/a | 0 | 0.027 | n/a | 5.42 | 19.51 |
| TITE-CRM | decreasing | 85 | 0.75 | n/a | 0 | 0.231 | n/a | 5.12 | 17.64 |
| | constant | 86 | 0.76 | n/a | 0 | 0.199 | n/a | 5.25 | 18.49 |
| | increasing | 87 | 0.79 | n/a | 0 | 0.143 | n/a | 5.46 | 19.78 |
| Benchmark | | 80 | 0.66 | n/a | . | 0.000 | . | . | . |

Compared to constant hazard, the performance of the Surv-CRM in terms of correct selection slightly improved in case of an increasing hazard of DLT (late-onset toxicities), for Sc2–Sc5 (Table 1). However, late-onset toxicities resulted in an increased overdose selection, whereas early-onset toxicities (decreasing hazard) resulted in improved (smaller) POS. In case of an increasing hazard of DLT, though the survival-CRM had a somewhat lower PCS compared to the TITE-CRM, it outperformed the later in terms of estimation (with smaller relative biases) for all scenarios. In terms of patients' safety during the trial, consistently across scenarios, late-onset toxicities increased the average number of patients assigned to toxic doses and the average number of observed DLTs and decreased the average number of patients assigned to the MTD; this was even more obvious as the accrual rate increased (Figures 3, 4 and 5). In line with these safety results, the Surv-CRM favored stopping trials for safety more often when all dose levels were considered over-toxic. In particular, in scenarios 1 and 8 with dose level 1 as MTD, and early toxicities (time-decreasing hazard of DLT), the percentage of early trial stopping for safety was 18% and 27% with Surv-CRM, respectively versus 14% and 20% with the TITE-CRM (Tables 1 and 2). Finally, in most scenarios, with time-varying hazards and varying accrual rate, Surv-CRM was slightly safer than the TITE-CRM during the trial.

The performance of all three methods improved with larger sample sizes (Figure 2, Supplementary material). The PCS was 100% with the benchmark in all scenarios but Sc7,

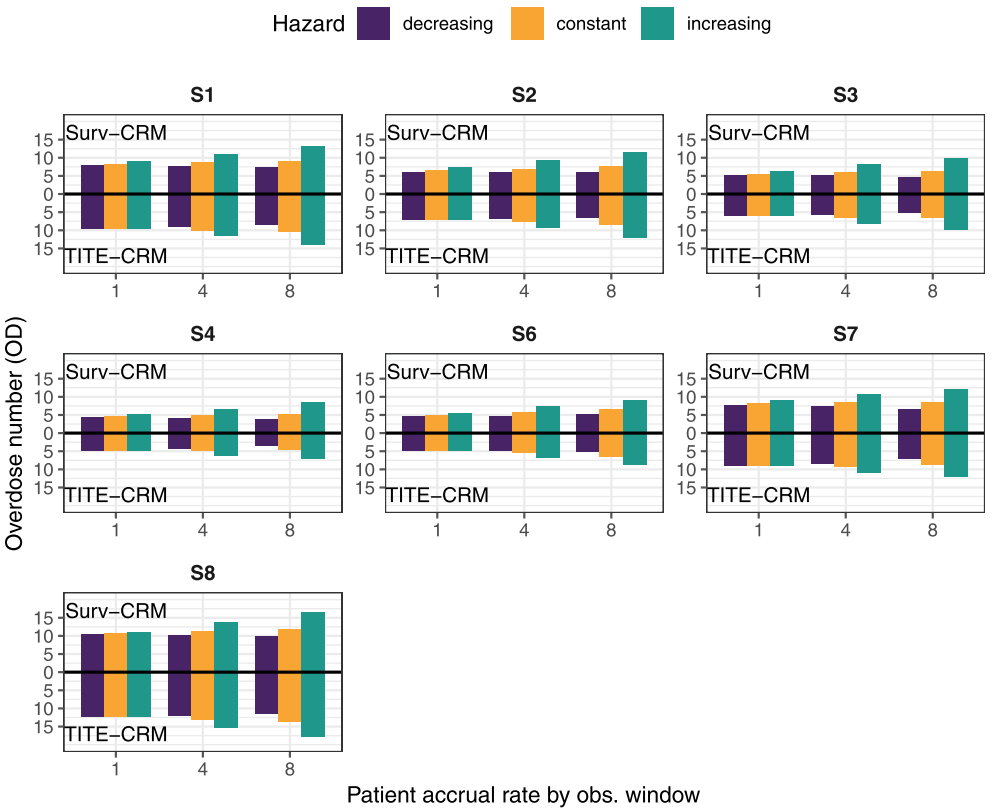


**Figure 3.** Number of patients administered with a dose above the MTD (OD) during the trial with the survival-CRM (barplots above the x-axis) and TITE-CRM (below) designs according to the patient accrual rate by observation window and the shape of hazard. $N$=10,000, $n$=25, and $\pi_{DLT} = 0.25$.

with a minimal sample size varying according to the scenario: from 100 (Sc6) to 300 (Sc4). Consistently, in Scenario 7, using the Surv-CRM and the TITE-CRM, PCS remained low (56% and 58%, respectively), even with $n = 200$.

In the presence of treatment discontinuation precluding complete observation of the toxicity outcomes, simulation results indicated a clear contrast between the informative Surv-CRM and the TITE-CRM (Table 3). The greater the risk of discontinuation, the better the performance of selection with iSurv-CRM compared to TITE-CRM. Indeed, for Sc10 and Sc11 with high and moderate risk of treatment discontinuation, PCS was 60% and 57% with the iSurv-CRM respectively versus 32% and 47% with the TITE-CRM, respectively. For Sc12, the steepest scenario, the iSurv-CRM also outperformed the TITE-CRM (PCS 59% *vs.* 33%). However, the probability of overdose selection ranged from 10% to 18% with the iSurv-CRM, and 3% to 11% with the TITE-CRM. Last, the iSurv-CRM allocated patients to the MTD more frequently than the TITE-CRM.

## 4. Discussion

In this paper, we propose an extension of the CRM that uses a survival working model to handle right-censored outcomes, the survival-CRM design, and the informative survival-CRM design for trials with a non negligible risk of competing event related to treatment discontinuation during the desired observation window, as well as a benchmark for its evaluation. This work was motivated by the need

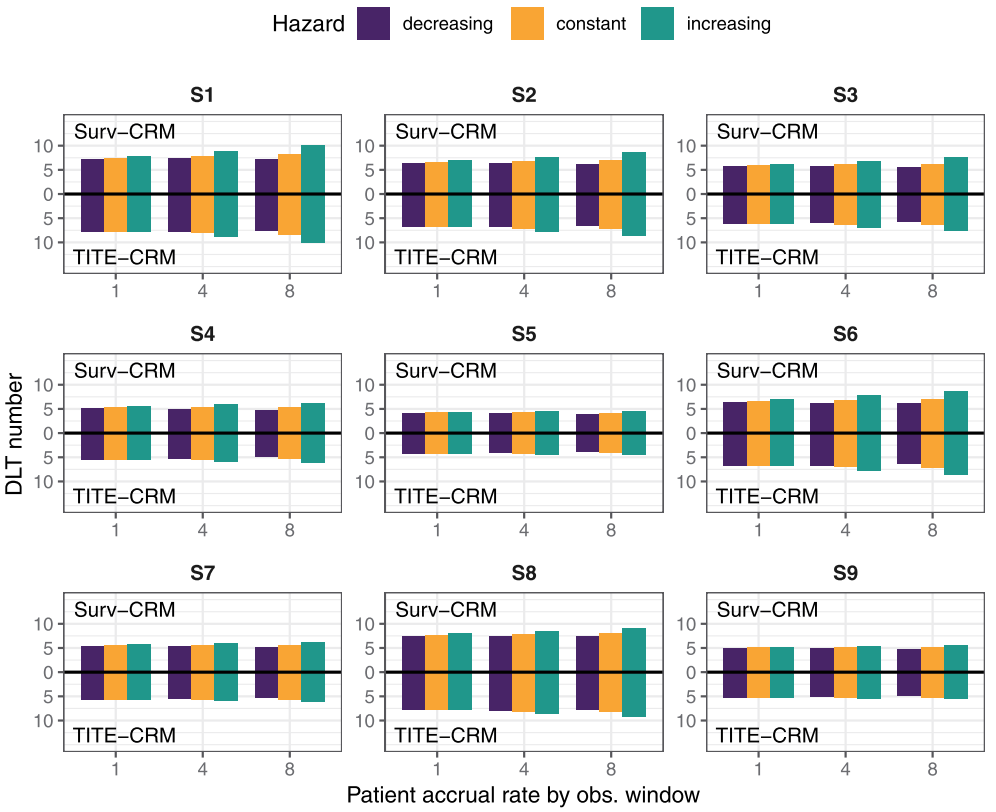


**Figure 4.** Number of observed DLTs during the trial with the survival-CRM (barplots above the x-axis) and TITE-CRM (below) designs according to the patient accrual rate by observation window and the shape of hazard. $N$=10,000, $n$=25, and $\pi_{DLT} = 0.25$.

for specific methods for dose finding with novel anti-cancer agents, such as targeted therapies or immunotherapies, which often require prolonged observation windows and result in some patients who do not experience a DLT but have not yet reached the end of the window. Survival inference relies of survival functions, assuming all patients will developed the event under treatment, with valid inference under independent or noninformative censoring. In our setting when the endpoint of interest is the DLT within some observation window of interest, the survival working models can provide unbiased estimates of cumulative incidence of DLT at the end of the observation window, given the choice of this window is likely independent of the toxicity process. Otherwise, when the rhythm of inclusion is short relatively to the observation window, and the observation of toxicity could be incomplete in patients already included, these data could be also considered administratively, i.e., noninformatively right-censored. Survival models have recently gained interest in dose finding due to the need for prolonged observation window with new anti-cancer therapies. Nevertheless, they have been mostly used in settings with complex endpoints, combining toxicity and efficacy endpoints or multiple response endpoints (Guo et al. 2018, 2019; Yuan and Yin 2009). For instance, Yuan and Yin (2009) proposed to jointly model toxicity and efficacy as time-to-event outcomes using a Weibull distribution for the hazard of toxicity, resulting in three model parameters to estimate, potentially less suited to our setting of Phase I clinical with small sample sizes. Thus, we proposed the Surv-CRM phase I dose-finding design using a one-parameter exponential working model for the dose-toxicity relationship, allowing for the use of those censored observations in the sequential process of MTD estimation during the trial. Nevertheless, when the observation window is long relatively to the underlying disease process, the observation of toxicity may be precluded by trial discontinuation

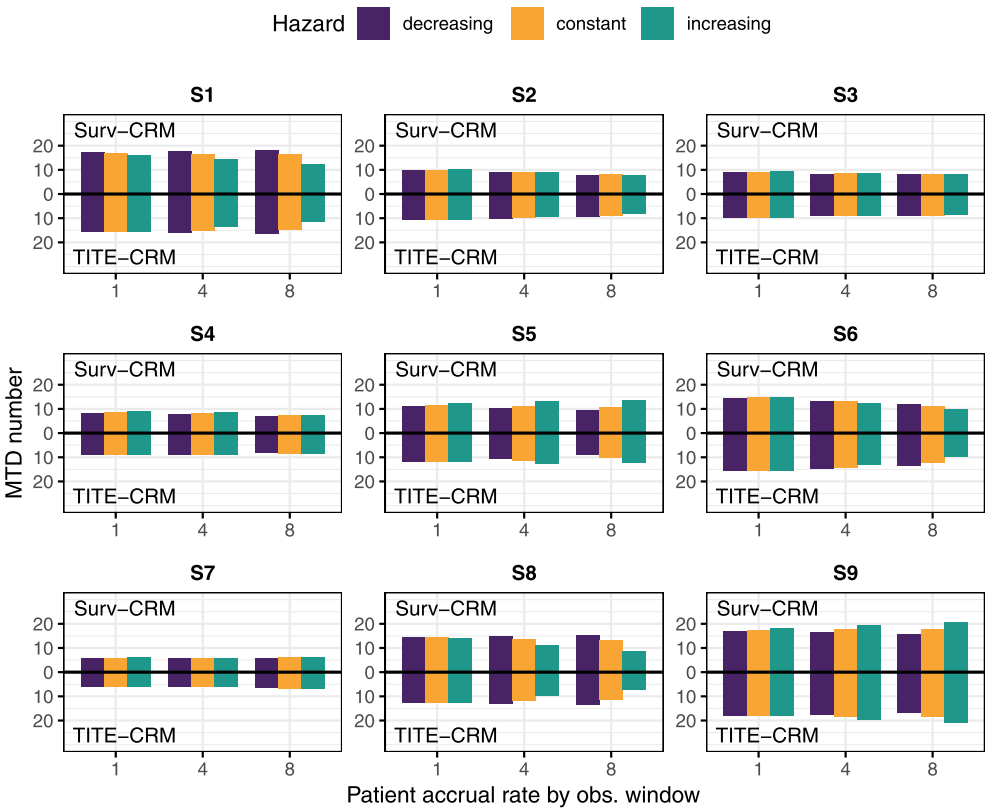


**Figure 5.** Number of patients treated with the true MTD (correct dose number) during the trial with the survival-CRM (barplots above the x-axis) and TITE-CRM (below) designs according to the patient arrival rate by observation window and the shape of hazard. $N$=10,000, $n$=25, and $\pi_{DLT} = 0.25$.

**Table 3.** Simulation results for Sc10–Sc12 of the informative survival-CRM (iSurv-CRM), TITE-CRM and benchmark: percent of correct selection (PCS); accuracy index ($A_{25}$); percent of overdose selection (POS); percent of stopped trials for safety ($P_{stop}$); relative bias (R. bias in estimated probability of DLT at the true MTD); average overdose number (OD); number of observed DLT (No. DLT) and number of patients treated with the true MTD (No. MTD) during the trial. $N$=10,000 simulated trials with with $n$=25 and $\pi_{DLT} = 25\%$. Accrual rate of four patients per observation window; time-constant toxicity hazards. (n/a: not applicable).

| Method | PCS | $A_{25}$ | POS | $P_{stop}$ | R. Bias | OD | No. DLT | No. MTD |
|---|---|---|---|---|---|---|---|---|
| **S10**: $F_1$: 0.08; 0.14; **0.25**; 0.40; 0.57 $F_2$: 0.50; 0.48; 0.45; 0.43; 0.40 | | | | | | | | |
| iSurv-CRM | 60 | 0.68 | 10 | 0.28 | 0.133 | 4.14 | 28.54 | 11.51 |
| TITE-CRM | 32 | 0.44 | 3 | 0.48 | 0.447 | 3.51 | 28.17 | 6.80 |
| Benchmark | 57 | 0.62 | 17 | . | −0.004 | . | . | . |
| **S11**: $F_1$: 0.08; 0.14; **0.25**; 0.40; 0.57 $F_2$: 0.30; 0.25; 0.20; 0.15; 0.10 | | | | | | | | |
| iSurv-CRM | 57 | 0.63 | 18 | 0.08 | 0.110 | 5.75 | 16.56 | 10.49 |
| TITE-CRM | 47 | 0.57 | 11 | 0.15 | 0.214 | 5.42 | 16.65 | 8.16 |
| Benchmark | 56 | 0.62 | 17 | . | 0.000 | . | . | . |
| **S12**: $F_1$: 0.08; 0.14; **0.25**; 0.40; 0.57 $F_2$: 0.55; 0.49; 0.43; 0.36; 0.30 | | | | | | | | |
| iSurv-CRM | 59 | 0.66 | 12 | 0.27 | 0.124 | 4.32 | 27.84 | 11.34 |
| TITE-CRM | 33 | 0.44 | 4 | 0.54 | 0.437 | 3.81 | 28.18 | 6.80 |
| Benchmark | 57 | 0.62 | 16 | . | −0.003 | . | . | . |

related to lack of efficacy of the drug (due to death, progression, withdrawal, etc.). To accommodate

this particular setting, we proposed the iSurv-CRM that uses estimation of the sub-distribution function of toxicity, to define the MTD, rather than that of the survival function.

Although Surv-CRM and TITE-CRM working models are different, the designs performed closely in most simulations in terms of correct dose selection at the end of the trial. Both designs were mostly robust to the rate of right-censored observations, either due to increased patient accrual or to time-varying hazards of DLT (that is, in case of either early or late-onset toxicities). Furthermore, the proposed design outperformed the TITE-CRM in measures of patient safety during the trial; notably, it allowed a reduced number of patients treated at an overtoxic dose level during the trial, and of those experiencing a DLT. Although the main objective of a dose-finding trial is to identify the MTD at the end of the trial, safety for patients enrolled in the trial is a matter of concern, for obvious ethical reasons, in these early phase trials. As we observed in the simulation study, inclusion of an adaptive wait time during the trial as proposed by Polley (2011) in our designs further reduces toxic doses selection without compromising the performances in correctly selecting the true MTD and keeping the trial length within a reasonable constraint. In the presence of competing discontinuations precluding complete observation of the toxicity outcomes, iSurv-CRM clearly outperformed the TITE-CRM for selecting the correct dose, expectedly when informative censoring was high. Additional performance comparisons with designs other than the TITE-CRM also assuming a time-to-event endpoint such as the PoD-TPI recently developed by Zhou et al. (2020) could also be considered in further studies.

To assess the operating characteristics of our proposed survival-CRM design, we developed a non-parametric benchmark approach (O'Quigley et al. 2002), adapted to a working model for censored data using the cumulative incidence of toxicity as the dose-finding endpoint, relying on Kaplan-Meier estimator. This provides an useful tool in assessing the design performance in various toxicity scenarios given a pre-specified sample size. The surv-CRM, as well as the TITE-CRM, outperformed the benchmark in scenarios with flat dose-toxicity relationship, illustrating how parametric methods might, in complex scenarios, combined with small samples sizes, outperform the nonparametric benchmark (Cheung 2011).

In summary, we proposed extensions of the CRM using survival working models for dose finding trials, which fit the natural censoring of data collected in these early clinical trials, and can handle an underlying competing risks framework, together with the corresponding benchmark assessment method. The Surv-CRM design, which provides the flexibility of a survival framework, showed desirable operating characteristics, close to those of the TITE-CRM, while providing a slightly safer profile for patients enrolled in these trials. When trial discontinuations within the observation window are likely to preclude the observation of DLTs in a non-negligible proportion of patients, the iSurv-CRM achieved the best performance in selecting the correct dose and allocating patients to the MTD. It should be considered when designing dose-findings trials of targeted therapies with long observation window in advanced cancer patients where treatment discontinuations are likely to occur.

## Acknowledgments

The authors thank the guest editors of the special issue of the journal and two referees for valuable comments.

## Declaration of interest statement

The authors declare that they have no competing interests.

## Data availability statement

R code for the use of the Surv-CRM, the iSurv-CRM and the associated benchmark is freely available on GitHub (https://github.com/SurvivalCRM).

## References


Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher. 2009. Simulating competing risks data in survival analysis. *Statistics in Medicine* 28 (6):0 956–971. doi:10.1002/sim.3516.

Biard, L., B. Cheng, G. A. Manji, and S. M. Lee. 2020. A simulation study of methods for handling disease progression in dose-finding clinical trials. *Journal of Biopharmaceutical Statistics*1–12. doi:10.1080/10543406.2020.1814796.

Cheung, Y. K. 2011. *Dose finding by the continual reassessment method*. Chapman & Hall/CRC Biostatistics Series. New York.

Cheung, Y. K. 2014. Simple benchmark for complex dose finding studies. *Biometrics* 70 (2):0 389–397. doi:10.1111/biom.12158.

Cheung, Y. K., and R. Chappell. 2000. Sequential designs for phase I clinical trials with late–onset toxicities. *Biometrics* 56 (4):82–1177. doi:10.1111/j.0006-341x.2000.01177.x.

Gray, R. 1988. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* 16 (3):0 1141–1154. doi:10.1214/aos/1176350951.

Guo, B., D. Li, and Y. Yuan. 2018. SPIRIT: A seamless phase I/II randomized design for immunotherapy trials. *Pharmaceutical Statistics* 170 (5):0 527–540. doi:10.1002/pst.1869.

Guo, B., Y. Park, and S. Liu. 2019. A utility-based bayesian phase i–ii design for immunotherapy trials with progression-free survival end point. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68 (2):0 411–425. doi:10.1111/rssc.12288.

Jeong, J.-H., and J. Fine. 2006. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55 (2):0 187–200. doi:10.1111/j.1467-9876.2006.00532.x.

Lee, S. M., D. Backenroth, Y. K. Cheung, D. L. Hershman, D. Vulih, B. Anderson, P. Ivy, and L. Minasian. 2016. Case example of dose optimization using data from bortezomib dose-finding clinical trials. *Journal of Clinical Oncology* 34 (12):0 1395–1401. doi:10.1200/JCO.2015.66.0662.

Lee, S. M., M. Ursino, Y. K. Cheung, and S. Zohar. 2019. Dose-finding designs for cumulative toxicities using multiple constraints. *Biostatistics* 20 (1):0 17–29. doi:10.1093/biostatistics/kxx059.

Lee, S. M., and Y. K. Cheung. 2009. Model calibration in the continual reassessment method. *Clinical Trials* 6 (3):38–227. doi:10.1177/1740774509105076.

Mathijssen, R. H. J., A. Sparreboom, and J. Verweij. 2014. Determining the optimal dose in the development of anticancer agents. *Nature Reviews Clinical Oncology* 11 (5):0 272–281. doi:10.1038/nrclinonc.2014.40.

Mozgunov, P., T. Jaki, and X. Paoletti. 2020. A benchmark for dose finding studies with continuous outcomes. *Biostatistics* 210 (2):0 189–201. doi:10.1093/biostatistics/kxy045.

O'Quigley, J., and L. Z. Shen. 1996. Continual reassessment method: A likelihood approach. *Biometrics* 52 (2):0 673–84. doi:10.2307/2532905.

O'Quigley, J., X. Paoletti, and J. Maccario. 2002. Non-parametric optimal design in dose finding studies. *Biostatistics* 3 (1):0 51–56. doi:10.1093/biostatistics/3.1.51.

Polley, M.-Y. 2011. Practical modifications to the time-to-event continual reassessment method for phase I cancer trials with fast patient accrual and late-onset toxicities. *Statistics in Medicine* 30 (17):0 2130–2143. doi:10.1002/sim.4255.

Postel-Vinay, S., C. Gomez-Roca, L. R. Molife, B. Anghan, A. Levy, I. Judson, J. De Bono, J.-C. Soria, S. Kaye, and X. Paoletti. 2011. Phase I trials of molecularly targeted agents: Should we pay more attention to late toxicities? *Journal of Clinical Oncology* 29 (13):01728–1735. doi:10.1200/JCO.2010.31.9236.

Postel-Vinay, S., L. Collette, X. Paoletti, E. Rizzo, C. Massard, D. Olmos, C. Fowst, B. Levy, P. Mancini, P. Lacombe, et al. 2014. Towards new methods for the determination of dose limiting toxicities and the assessment of the recommended dose for further studies of molecularly targeted agents–dose-limiting toxicity and toxicity assessment recommendation group for early trials of targeted therapies, an European organisation for research and treatment of cancer-led study. *Eur J Cancer* 500 (12):0 2040–9. doi:10.1016/j.ejca.2014.04.031.

Putter, H., M. Fiocco, and R. B. Geskus. 2006. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 260 (11):0 2389–2430. doi:10.1002/sim.2712.

Wages, N. A., C. Chiuzan, and K. S. Panageas. 2018. Design considerations for early-phase clinical trials of immune-oncology agents. *Journal for ImmunoTherapy of Cancer* 60 (8). doi: 10.1186/s40425-018-0389-8.

Wheeler, G. M., A. P. Mander, A. Bedding, K. Brock, V. Cornelius, A. P. Grieve, T. Jaki, S. B. Love, L. Odondi, C. J. Weir, et al. 2019. How to design a dose-finding study using the continual reassessment method. *BMC Medical Research Methodology* 190 (1):0 1–15. doi:10.1186/s12874-018-0638-z.

Wheeler, G. M., M. J. Sweeting, and A. P. Mander. 2018. A Bayesian model-free approach to combination therapy phase I trials using censored time-to-toxicity data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 680 (2):0 309–329. doi:10.1111/rssc.12323.

Yuan, Y., and G. Yin. 2009. Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58 (5):0 719–736. doi:10.1111/j.1467-9876.2009.00674.x.

Zhou, T., W. Guo, and Y. Ji. 2020. Pod-tpi: Probability-of-decision toxicity probability interval design to accelerate phase I trials. *Statistics in Biosciences* 120 (1):0 1–22. doi:10.1007/s12561-019-09264-0.