Received: 29 June 2021 | Revised: 1 March 2022 | Accepted: 3 May 2022

DOI: 10.1002/pst.2231

WILEY

MAIN PAPER

Two-stage subgroup-specific time-to-event (2S-Sub-TITE): An adaptive two-stage time-to-toxicity design for subgroup-specific dose finding in phase I oncology trials

Alana McGovern[1,2] | **Andrew G. Chapple**[3] | **Clement Ma**[1,4,5,6]

[1]Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, Massachusetts, USA

[2]Department of Statistics, University of Washington, Seattle, Washington, USA

[3]Biostatistics Program, School of Public Health, Louisiana State University Health Science Center, New Orleans, Louisiana, USA

[4]Harvard Medical School, Boston, Massachusetts, USA

[5]Centre for Addiction and Mental Health, Toronto, Canada

[6]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Correspondence
Clement Ma, Centre for Addiction and Mental Health, 1001 Queen Street West, Toronto, Ontario, M6J 1H1, Canada.
Email: clement.ma@camh.ca

Funding information
Biostatistics Program at the Dana-Farber/Boston Children's Cancer and Blood Disorders Center

Abstract

For phase I trials, the subgroup-specific time-to-event (Sub-TITE) design identifies the maximum tolerated dose (MTD) separately in 2+ heterogeneous patient subgroups. Sub-TITE allows borrowing strength and dynamic clustering across subgroups from the trial's start, but delaying the initiation of borrowing and clustering may improve trial accuracy. We propose the 2-stage Sub-TITE (2S-Sub-TITE) design in which the trial starts by estimating separate models per subgroup, and then initiates the Sub-TITE design at some prespecified point of patient accrual. We evaluate the operating characteristics of the 2S-Sub-TITE design using simulations. Nine configurations of the 2S-Sub-TITE design (varying in timing of initiation of borrowing/clustering and prior probability of subgroup heterogeneity, *p_hetero*) and three control methods were compared across 1000 randomly-generated true toxicity probability scenarios. Effects of priors, sample size, escalation rules, target toxicity probability, accrual rate, and number of subgroups were evaluated. Metrics included: proportion of correct selection (*PCS*) of the true MTD, and average number of toxicities incurred. Among the 5 2S-Sub-TITE configurations (out of 9 total) with the highest *PCS* (45%) when the subgroup heterogeneity assumption is correct (all of which out-perform the control methods by 2%–6%), the configuration which enables borrowing and clustering allowance with $p\_hetero = 0.7$ starting at 75% patient accrual best minimizes toxicities as well as losses in accuracy if the heterogeneity assumption is incorrect. For trials with high confidence in subgroup heterogeneity, the 2S-Sub-TITE configuration enabling borrowing/clustering with $p\_hetero = 0.7$ starting at 75% patient accrual exhibits superior dose-finding accuracy compared to existing methods.

KEYWORDS

Bayesian adaptive design, continual reassessment method, phase I clinical trials, subgroup-specific dose finding

Andrew G. Chapple and Clement Ma are joint senior authors.

1 | INTRODUCTION

Phase I clinical trials in oncology aim to find the maximum tolerated dose (MTD) of an experimental agent. While a majority of phase I trials have used conservative rule-based designs,[1] such as the 3 + 3,[2] rolling six,[3] and accelerated titration[4] designs, model-based designs, such as the continual reassessment method (CRM), more frequently find the correct MTD.[5–7] The CRM, uses a Bayesian framework to model the dose-toxicity relationship using the observed dose limiting toxicity (DLT) data to date and a set of prior toxicity assumptions to recommend a dose level for the next patient.[5] A popular extension is the time-to-event CRM (TITE-CRM) which can model partial DLT observations to allow for continuous patient enrollment while maintaining similar accuracy and safety.[8–10]

These designs implicitly assume that enrolled patients are relatively homogeneous to estimate a single MTD across all patients. When there is a possibility of significant patient heterogeneity, numerous published methods can account for subgroup-specific dose-finding.[11–23] One such design, called subgroup-specific time-to-event (Sub-TITE), uses spike-and-slab priors and previous patient DLT information to estimate the dose toxicity curve for each subgroup at each step.[24] The observed similarity between these estimated subgroup-specific dose toxicity curves effectively determines whether the subgroups will be modeled in a two-parameter vector model, or clustered together in a one-parameter vector model. This method allows for trials with subgroup-specific dose-finding to benefit from borrowing strength between subgroups, while allowing subgroup information to be clustered together when necessary.

Our current study is motivated by an ongoing Pacific Pediatric Neuro-Oncology Consortium (PNOC) Phase I trial of DAY 101 (formerly TAK-580, MLN2480), a type II BRAF inhibitor, for children with low-grade gliomas (NCT03429803).[25] Preliminary data suggested differing response by body surface area (BSA), thus motivating a subgroup-based dose-finding approach. The investigators implemented the Sub-TITE design to test five dose levels among two BSA subgroups. The investigators were uncomfortable with the aggressive dose escalation rules in the original Sub-TITE design and implemented a more conservative escalation approach (see Section 2). In addition, the original design allowed borrowing and clustering from the trial's start, but gathering additional DLT observations before making these allowances may increase trial accuracy.

In this article, we introduce a practical extension of the Sub-TITE design, in which the two subgroups are first modeled separately according to the TITE-CRM and then at some pre-specified point of patient accrual, the Sub-TITE design is introduced with varying parameter settings. The 2-Stage Sub-TITE (2S-Sub-TITE) design is a flexible framework that can allow more DLT information to be gathered before determining whether subgroups should be clustered. Our goal is to identify the optimal time at which to initiate the Sub-TITE design, as well as the optimal prior assumptions regarding subgroup heterogeneity, to maximize the true MTD recommendation rate and minimize the average number of toxicities when the assumption of subgroup heterogeneity is correct, while minimizing losses when this assumption is incorrect. We will also investigate whether these optimal conditions vary by investigator-specified prior, escalation rules, sample size, target toxicity probability, accrual rate, and number of subgroups.

2 | METHODS

For phase I trials with patient heterogeneity, Figure 1 presents an overview of the different dose-finding strategies considered in this study. Separate TITE-CRM models per subgroup (Figure 1A) allows independent dose finding for each subgroup. A single TITE-CRM model across subgroups (Figure 1B) allows combined dose finding for all patients (thus ignoring potential subgroup heterogeneity). The original Sub-TITE design (Figure 1C) allows for separate dose finding per subgroup while allowing borrowing and dynamic clustering across subgroups from the trial start. The original Sub-TITE design is a hybrid of separate TITE-CRM models for each subgroup and a single TITE-CRM model across all subgroups.

The 2S-Sub-TITE design (Figure 1D) is a 2-stage sequential combination of separate TITE-CRM models for each subgroup, and the original (1-stage) Sub-TITE design. The 2S-Sub-TITE design begins with Stage 1, which uses the TITE-CRM to estimate a separate dose-finding model for each subgroup and proceeds until a pre-specified number of Stage 1 patients have been accrued, denoted N_{sep}. Then, Stage 2 begins and continues until the maximum number of patients is reached, denoted N_{max}. Stage 2 uses a form of the Sub-TITE, determined by prior assumptions of patient heterogeneity.

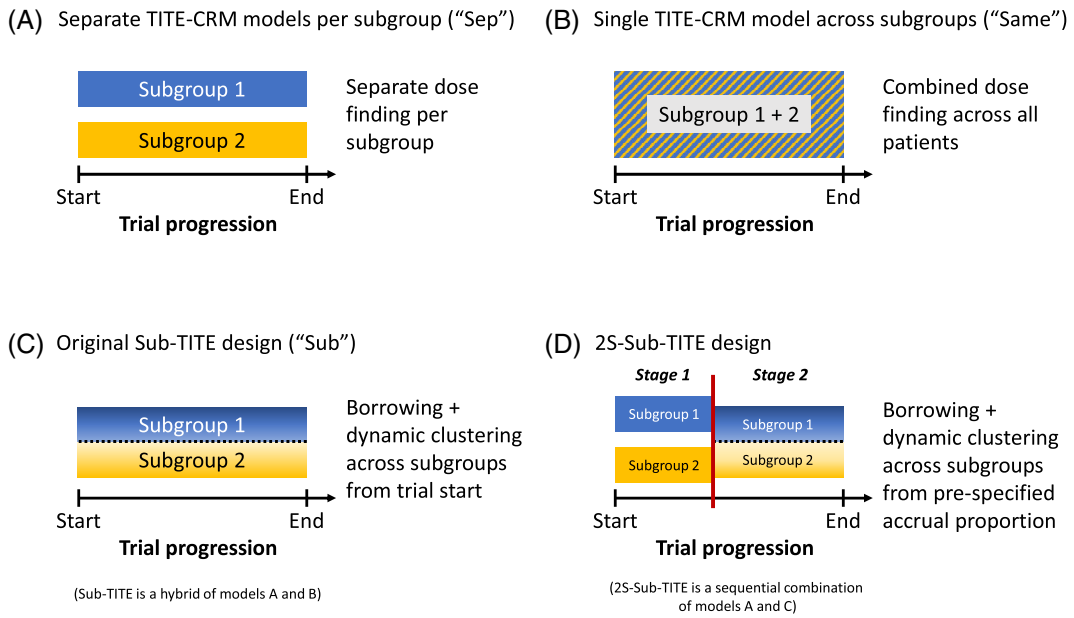

FIGURE 1 Overview of dose-finding strategies for phase I trials with patient heterogeneity: (A) Separate TITE-CRM models per subgroup; (B) Single TITE-CRM model across subgroups; (C) Original Sub-TITE design; and (D) 2S-Sub-TITE design. CRM, continual reassessment method; 2S-Sub-TITE, 2 stage-subgroup-specific time-to-event

2.1 | Stage 1

Slightly modifying the notation of Chapple & Thall,[24] let standardized dose $x_j = \left(x_j^{raw} - \overline{x^{raw}}\right)/SD(x^{raw})$, for $j = 1,...,J$, where x_j^{raw} is dose j. Given $g = 1,2,...,G$ subgroups, let $W_i \in \{1,...,G\}$ denote the subgroup and $x_{[i]}$ the dose level of the ith patient. The probability of toxicity given dose level $x_{[i]}$ and subgroup $W_i = g$ is

$$logit\left\{\pi\left(x_{[i]},\boldsymbol{\theta}_g\right)\right\} = \alpha_g + \exp\left(\beta_g\right)x_{[i]}, \tag{1}$$

where $\pi\left(x_{[i]},\boldsymbol{\theta}_g\right) = P\left(Y_i = 1 | x_{[i]}, W_i = g, \boldsymbol{\theta}_g\right)$ and $\boldsymbol{\theta}_g = \left(\alpha_g, \beta_g\right)$ where α_g and β_g are the intercept and dose effects for subgroup g, respectively. To allow use of partial follow up information, we use a working likelihood similar to Cheung and Chappell.[8] Let T be the fixed DLT observation period. For patient i at trial time t, let $\mu_i(t)$ denote the follow-up time with $\mu_i(t) = T$ for all follow-up times past the observation period. Let weight function $\omega_i(t) = \mu_i(t)/T$, and $Y_i(\mu_i(t))$ be a stochastic binary indicator of patient i experiencing a DLT by trial time t. The likelihood contribution for a patient in subgroup g is

$$L\left(\boldsymbol{\theta_g} | D_{n_t}, W_i = g\right) = \pi\left(x_{[i]},\boldsymbol{\theta}_g\right)^{Y_i(\mu_i(t))}\left\{1 - \omega_i(t)\pi\left(x_{[i]},\boldsymbol{\theta}_g\right)\right\}^{1-Y_i(\mu_i(t))} \tag{2}$$

where $D_{n_t} = \left\{Y_i(\mu_i(t)), W_i, \mu_i(t), x_{[i]}, i = 1,...,n_t\right\}$ denotes the data collected by trial time t and n_t is the number of patients accrued at time t. Unique subgroup-specific normal prior distributions are assumed for the parameter vectors $\boldsymbol{\theta_1},...,\boldsymbol{\theta_G}$ and posterior distributions are estimated via Markov Chain Monte Carlo (MCMC) sampling with Metropolis Hastings proposals (Supplement A). Posterior means of the vectors $\boldsymbol{\theta_1},...,\boldsymbol{\theta_G}$ are entered into the dose-toxicity model along with the values $x_1,...,x_J$ to estimate the subgroup-specific toxicity probabilities at each dose level. Posterior means of these parameters were used to estimate toxicity probabilities instead of computing posterior toxicity probabilities for the entire sample and averaging them. The primary reason was due to reduced computation time; this approach did not result in changes in operating characteristics. Thus, dose-finding estimates are modeled completely separately for each subgroup, effectively conducting G trials under the original TITE-CRM design. We will use this model until a pre-specified point of patient accrual, N_{sep}/N_{max}, and then proceed to Stage 2.

2.2 | Stage 2

We implement an over-parameterized version of the Sub-TITE design[24] to remove a baseline group, allowing all combinations of subgroups to cluster freely, by adding a shared intercept and slope parameter to the dose-toxicity curve in the form

$$logit\{\pi(x_{[i]},W_i,\boldsymbol{\theta})\} = \alpha + \sum_{g=1}^{G}\alpha_g I(W_i=g) + \exp\left\{\beta + \sum_{g=1}^{G}\beta_g I(W_i=g)\right\}x_{[i]}. \quad (3)$$

The working likelihood function for n_t patients at trial time t is

$$L(\boldsymbol{\theta_1},...,\boldsymbol{\theta_G}|\,D_{n_t}) = \prod_{i=1}^{n_t}\sum_{g=1}^{G} I[W_i=g]\pi(x_{[i]},W_i,\boldsymbol{\theta}_g)^{Y_i(\mu_i(t))}\{1-\omega_i(t)\pi(x_{[i]},W_i,\boldsymbol{\theta}_g)\}^{1-Y_i(\mu_i(t))}.$$

In this overparameterized version, the normal prior on $\boldsymbol{\theta_1}$ is used for the shared parameter vector (α,β) and priors on $\theta_2,...,\theta_G$ are expressed as deviations from the shared parameters, with $\boldsymbol{\theta_1}$ having a vector mean 0. Similar to in Stage 1, posterior sampling is carried out via MCMC and posterior means of $(\alpha,\beta,\boldsymbol{\theta_1},...,\boldsymbol{\theta_G})$ are entered into Equation (3) with $x_1,...,x_J$ in place of $x_{[i]}$ to estimate the subgroup specific dose-toxicity probabilities at each dose level. This model borrows strength among the subgroup specific dose-toxicity curves.

2.3 | Dynamic clustering of subgroups

We allow for the dynamic clustering of subgroups by introducing the random latent subgroup membership variables $\zeta_1,...,\zeta_G \in \{1,...,G\}$ as defined in Chapple & Thall. These induce clustering through the prior distribution

$$\alpha_g\,|\,\zeta_g \sim I\left[\zeta_g=g\right]N(\widetilde{\alpha}_g,\sigma_\alpha) + \sum_{m\neq g} I\left[\zeta_g=m\right]\delta_{\alpha_m}(\alpha_g),$$

$$\beta_g\,|\,\zeta_g \sim I\left[\zeta_g=g\right]N\left(\widetilde{\beta}_g,\sigma_\beta\right) + \sum_{m\neq g} I\left[\zeta_g=m\right]\delta_{\beta_m}\left(\beta_g\right),$$

where $\delta_{\alpha_m}(\alpha_g)$ denotes the dirac measure at α_m, i.e. $\alpha_g=\alpha_m$ with probability 1, The hyperparameters $\widetilde{\alpha}_g,\widetilde{\beta}_g,\sigma_\alpha$, and σ_β are derived from a clinician-advised table of prior mean probabilities of toxicity at each dose level, for each subgroup. Given that $\zeta_g \neq g$, the random latent subgroup parameter ζ_g takes on values in the set $\{m{:}\,\zeta_m=m\}$ with equal probability.

The parameter vectors $\boldsymbol{\theta_1},...,\boldsymbol{\theta_G}$ are adaptively clustered and unclustered throughout the MCMC based on the random parameters $\zeta_1,...,\zeta_G$ which are also being sampled. It is possible that, for example, $\theta_m=\theta_g$ for 95% of posterior samples, indicating extreme homogeneity, or 5% of the samples, indicating heterogeneity among subgroups g and m.

A key user-specified parameter is $P[\zeta_g=g]$, denoted *P_hetero*, which controls the probability that subgroup g will *not* be dynamically clustered on other subgroups. Effectively, *P_hetero* represents the prior assumption that a subgroup is heterogeneous and is a fixed value chosen at trial start. *P_hetero* = 1 indicates that the subgroup will never be clustered on other subgroups (fully heterogeneous); *P_hetero* = 0 indicates that the subgroup always be clustered on another subgroup (fully homogeneous). *P_hetero* = 0.9 indicates a 0.9 prior probability that the subgroup will not be clustered on another subgroup. We will explore how *P_hetero* affects dose-finding accuracy.

2.4 | Dose escalation rules

Two sets of dose escalation rules will be implemented. First, "aggressive escalation" rules, as in Chapple & Thall,[24] require only one patient to be *enrolled, but not necessarily fully evaluated*, at a dose before allowing escalation to the

next, previously untested, dose. Second, "conservative escalation" rules require that three patients be *fully evaluated* at a dose before allowing escalation to the next, previously untested, dose.

2.5 | 2S-Sub-TITE method configurations

We evaluated 9 2S-Sub-TITE method configurations with varying combinations of *P_hetero* and the accrual point at which borrowing and clustering allowance begins (Table 1). These combinations were chosen factorially so the full range of potential *P_hetero* and accrual rate pairings could be assessed. Configurations A–C, D–F, and G–I start borrowing and allowing for dynamic clustering at 25%, 50%, and 75% of patient accrual, respectively, with varying *P_hetero*. For comparison, three control methods were evaluated: (1) the original Sub-TITE design, denoted "Sub"; (2) two separate TITE-CRM models, denoted "Sep"; and (3) a single TITE-CRM model, denoted "Same." Observe that these control methods are the designs depicted in Figure 1A–C.

2.6 | Simulation design

We performed extensive simulations in R version 4.0.1 to assess the operating characteristics of different configurations of our 2S-Sub-TITE design and compare them to the original (1-Stage) Sub-TITE design[24] and the original TITE-CRM design.[8] Our primary varied parameters were: (1) the true toxicity probability scenarios and (2) the dose-finding methods determining when and how borrowing and clustering take place (including 9 2S-Sub-Tite method configurations and three control methods: the original Sub-TITE method, and combined and separate original TITE-CRM models). This would allow us to evaluate the efficacy of each method whether or not the assumption of subgroup heterogeneity was correct.

2.7 | Randomly-generated true toxicity probability scenarios

We randomly generated 1000 true toxicity probability scenarios where we specified the probabilities π_{gj}^{True} for $g = 1, 2$ and $j = 1, 2, 3, 4, 5$. We first randomly selected a dose level in each subgroup to be the true MTD and generated its true toxicity probability from a 0.1-length uniform distribution around the target (e.g., between 0.15 and 0.25 for a target of 0.2). We then generated the remaining true toxicity probabilities under two conditions: (1) that all toxicity probabilities increased monotonically and none were within 0.01 of each other and (2) that no dose-toxicity probabilities were closer to the target than the randomly chosen true MTD. This was done repeatedly for both subgroups until the average

TABLE 1 Summary of nine 2S-Sub-TITE configurations and three comparison methods

	Method	Start borrowing and clustering at	Probability that subgroups are heterogeneous (P_hetero)
A	Brrw&Clstr@25%|Phet = 0.5	25% Accrual	0.5
B	Brrw&Clstr@25%|Phet = 0.7	25% Accrual	0.7
C	Brrw&Clstr@25%|Phet = 0.9	25% Accrual	0.9
D	Brrw&Clstr@50%|Phet = 0.5	50% Accrual	0.5
E	Brrw&Clstr@50%|Phet = 0.7	50% Accrual	0.7
F	Brrw&Clstr@50%|Phet = 0.9	50% Accrual	0.9
G	Brrw&Clstr@75%|Phet = 0.5	75% Accrual	0.5
H	Brrw&Clstr@75%|Phet = 0.7	75% Accrual	0.7
I	Brrw&Clstr@75%|Phet = 0.9	75% Accrual	0.9
Sub	Brrw&Clstr@0%|Phet = 0.9	0% Accrual	0.9
Sep	Separate TITE-CRM	N/A	1
Same	Combined TITE-CRM	N/A	0

Abbreviations: Brrw, borrrowing; Clstr, clustering; N/A, not applicable; Phet, "P_hetero," probability that subgroups are heterogeneous.

difference in true subgroup-specific toxicity probabilities was less than 0.3. We added this last restriction since it is unlikely that dose 1 (e.g.) would have a toxicity probability of 0.1 in one subgroup and 0.8 in another.

The large number of simulation scenarios was randomly generated to cover a wider, more comprehensive array of potential trial scenarios. Of these scenarios, 45.9% had the same MTD across subgroups and 54.1% had differing MTDs across subgroups. A descriptive summary of the 1000 randomly generated scenarios is shown in Table 2.

2.8 | Simulation parameters

We performed a primary set of simulations using the primary simulation parameters defined below. We performed a series of secondary simulations to vary key parameters individually while fixing other primary simulation parameters.

1. Number of subgroups: Trials have $G = 2$ (primary) or $G = 3$ (secondary) subgroups
2. Number of dose levels per subgroup: All trials have five standardized dose levels per subgroup $(x_1,...,x_5) = (-2, -1, 0, 1, 2)$ with each subgroup starting on dose x_1. Negative dose levels are possible since these are standardized values and not raw dose levels.
3. DLT observation period: All trials assume a 1-month DLT observation period.
4. Accrual rate: Trials assume an accrual rate of two patients (primary) or four patients (secondary) per month per subgroup.
5. Target toxicity probability (π^*): Trials assume $\pi^* = 0.2$ (primary) or $\pi^* = 0.3$ (secondary). A typical toxicity rate ranges between 0.2 and 0.33 in phase I trials.[26] The occurrence of a DLT was assumed to be uniform across the DLT observation period.
6. Prior Toxicity probabilities: For trials with $G = 2$ subgroups, we selected 18 different sets of prior toxicity probabilities (primary) $\pi^e_{g,j}$ for $g = 1, 2$ and $j = 1, 2, 3, 4, 5$. For trials with $G = 3$ subgroups, we selected 2 different sets of prior toxicity probabilities (secondary) $\pi^e_{g,j}$ for $g = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$ (See Supplement B).
7. Escalation rules: Trials used conservative (primary) or aggressive escalation rules (secondary), as defined above.
8. Sample size: Trials accrued a total sample size $N_{max} = 40$ (primary) or $N_{max} = 64$ (secondary) patients. All trials assumed equal accrual to each subgroup.

2.9 | Simulation conduct

We performed trial simulations using the SubTite R package version 4.0.1.[27] First, we calculated the prior means and variances for each of the 18 sets of priors using the *GetPriorMeans()* function with the fixed standardized levels and the hypervariances on the intercepts and slope set to 5 and 1, respectively, as in Chapple & Thall.[24]

TABLE 2 Summary of 1000 randomly generated scenarios for trials with 2 subgroups and 5 dose levels

Characteristic	Category	Value(s)
Proportion of scenarios with true MTD per dose level per subgroup	Subgroup 1	*($dose_1$, $dose_2$, $dose_3$, $dose_4$, $dose_5$)* (0.192, 0.310, 0.247, 0.173, 0.078)
	Subgroup 2	(0.209, 0.306, 0.240, 0.162, 0.083)
Proportion of scenarios with N_{MTD} dose levels between subgroup true MTDs	0	0.459
	1	0.447
	2	0.090
	3	0.004
True toxicity probability per dose level per subgroup		*($dose_1$, $dose_2$, $dose_3$, $dose_4$, $dose_5$)*
Mean	Subgroup 1	(0.092, 0.175, 0.284, 0.446, 0.675)
	Subgroup 2	(0.095, 0.179, 0.286, 0.454, 0.681)
SD	Subgroup 1	(0.071, 0.099, 0.161, 0.223, 0.244)
	Subgroup 2	(0.072, 0.101, 0.158, 0.227, 0.248)
Difference in true toxicity probability between subgroups per dose level		
Mean		(−0.003, −0.005, −0.002, −0.007, −0.006)
SD		(0.076, 0.081, 0.091, 0.101, 0.105)

Next, we simulated trials using the *SimTrial()* function. Our primary set of simulations consisted of 10 simulated trials using the primary parameters specified above, for each combination of 18 priors, 12 method configurations, and 100 scenarios, totaling 21,600,000 simulated trials (See Supplement C for exact function input). We conducted secondary sets of simulations for each combination of 18 priors, 12 method configurations, and a minimum of 200 scenarios for each of the secondary parameters specified above.

The output of the *SimTrial()* function returns a collection of summary statistics of the simulated trials. The primary endpoint is the probability of correct selection (*PCS*): the proportion of simulations for which the correct dose was chosen as the MTD for each subgroup. Secondary endpoints include: the average number of toxicities in each subgroup (*AvgTox*), and weighted average distance between the target toxicity probability and the toxicity probability at the selected dose (*AvgWgtDist*). Example R code and output for SubTiTE R package functions are available in Supplemental Materials.

3 | RESULTS

3.1 | Primary simulation results

Figure 2 presents the (A–B) Average *PCS* and (C-D) *AvgTox* for 9 2S-Sub-TITE configurations and three control methods averaged across 18 priors x 100 simulated trials for scenarios in which the two true subgroup MTDs were equal ($N = 459$ scenarios) or different ($N = 541$ scenarios). In each heatmap, 2S-Sub-TITE configurations A-I are arranged based on when borrowing and clustering allowance begins (x-axis) and *P_hetero* (y-axis). For example in Figure 2A, S-Sub-TITE configuration C (borrowing/clustering at 25% accrual with $p\_hetero = 0.9$) is listed in the top left cell of the 3x3 grid. The simulated *PCS* is 56.9% (averaged across 459 scenarios with the same true subgroup MTD). Note that when *PCS* for Figure 2A,B, lighter colors indicate higher *PCS*, and therefore better performance, whereas in Figure 2C, D, lighter colors indicate more toxicities, and therefore worse performance.

The "optimal" 2S-Sub-TITE configuration should have relatively high *PCS* (Figure 2A,B) while maintaining low *AvgTox* (Figure 2C,D). When our subgroup heterogeneity assumption is correct (e.g., subgroups have different true MTDs), the 2S-Sub-TITE configurations with highest *PCS* are A, B, E, H, and I (44.6%–45%), which is 2%–6% higher than the 3 control methods (Figure 2B). Among these configurations, configuration H (beginning borrowing and clustering allowance at 75% patient accrual with $p\_hetero = 0.7$) is optimal, with relatively high *PCS* when our subgroup heterogeneity assumption is incorrect (*PCS* = 53.6%; Figure 2A) and the lowest *avgTox* when subgroup MTDs are the same (AvgTox = 7.73; Figure 2C) or different (AvgTox = 7.55; Figure 2D). Heatmaps generated using a third operating characteristic, *AvgWgtDist*, support these results (Supplemental Figure 1).

From Figure 2A we can also observe that if two separate TITE-CRM models (Sep) were used in a trial with truly homogeneous subgroups (where one model should have been used), then an average of 58.7%–45.9% = 12.8% accuracy in MTD selection is lost, but if 2S-Sub-TITE configuration H was used, only an average of 58.7%–53.6% = 5.1% accuracy in MTD selection is lost. Conversely, if the original Sub-TITE method configuration (starting borrowing and clustering allowance at the beginning of the trial with $phetero = 0.9$) were used in the same case, then a mere 2.3% of accuracy in MTD selection is lost on average. However, in the case our subgroup heterogeneity assumption is correct, the original Sub-TITE method results in a smaller 1.3% loss in MTD selection accuracy on average, compared to the 2S-Sub-TITE method with configuration H.

3.2 | Secondary simulation results

In a series of secondary simulations, we assess whether our findings are consistent when we vary different trial parameters (Figure 3 and Supplemental Figures 2–3**)**.

3.2.1 | Effect of prior

In Figure 3A, the *PCS* was averaged across all 1000 scenarios for each of the 18 priors. We observe that although some priors perform better than others, the 18 curves all have a similar shape, indicating that selection of prior does not

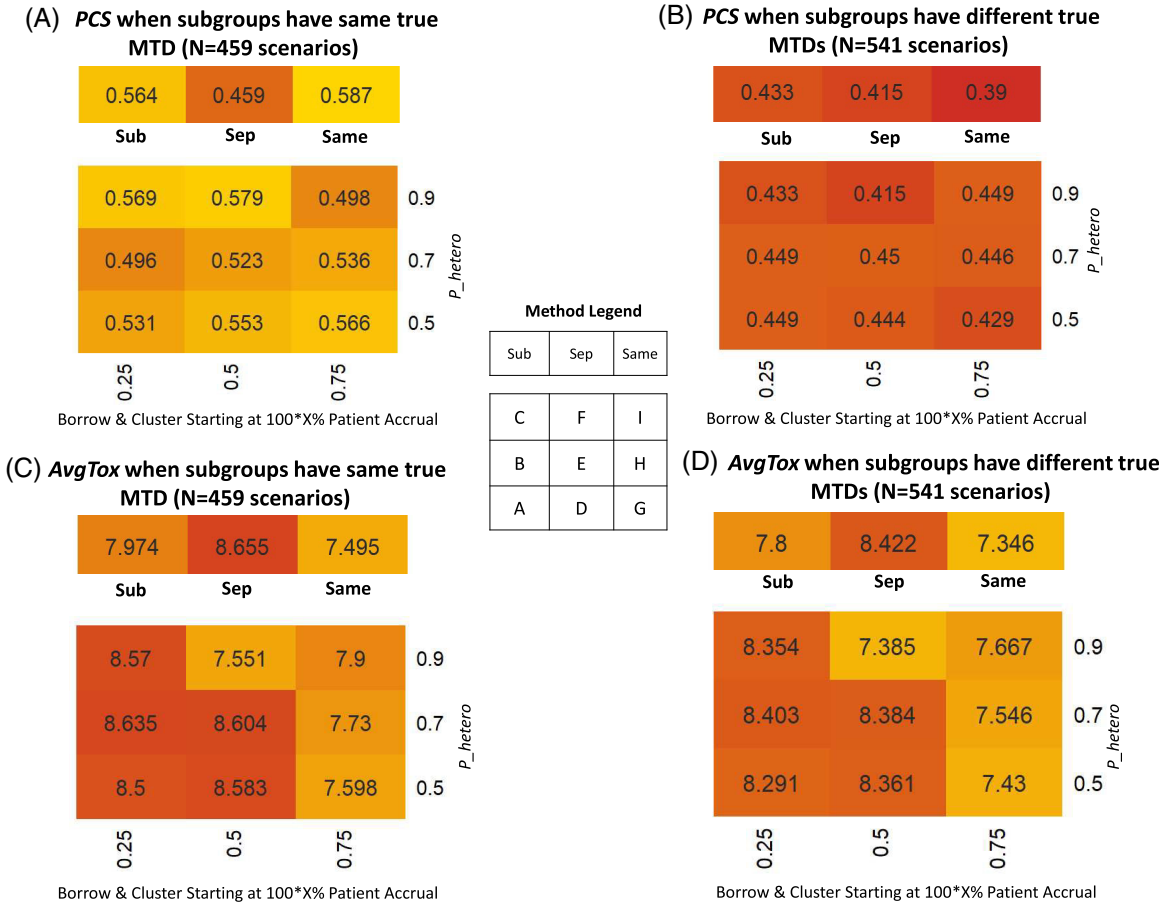

FIGURE 2 Operating characteristics for nine 2S-Sub-TITE configurations and three control methods averaged across 18 priors and 100 simulated trials for: (A) *PCS* in 459 scenarios with the same true subgroup MTDs; (B) *PCS* in 541 scenarios with different true subgroup MTDs; (C) *AvgTox* in 459 scenarios with the same true subgroup MTDs; (D) *AvgTox* in 541 scenarios with different true subgroup MTDs

significantly modify which method configurations have the highest *PCS*. *Effect of Dose Escalation Rules*. In Figure 3B, the *PCS* was averaged across all priors and scenarios for simulations using conservative or aggressive escalation. The relative pattern of *PCS* across configurations is very similar between the conservative and aggressive escalation. There is an average increase in *PCS* of approximately 1.1% (interquartile range [IQR] = 0.7–1.5%) for simulations using aggressive escalation versus conservative escalation. With an interquartile range of 0.7%–1.5%, this increase is both small and consistent across configurations. *Effect of Sample Size*. In Figure 3C, the *PCS* was averaged across all priors and 200 scenarios for simulations using a sample size of N_{max}=40 or N_{max}=64 patients. To ensure a fair comparison, the same combinations of priors and scenarios were simulated under both sample sizes. Increasing the sample size by 24 patients (60%), results in a fairly substantial increase in *PCS*, approximately 5.1% on average, but this increase does not vary significantly by configuration (IQR = 4.7%–5.5%). *Effect of Accrual Rate*. In Figure 3D, the *PCS* was averaged across all priors and 200 scenarios for simulations using an accrual rate of 2 or 4 patients per month. Increasing the accrual rate to 4 patients per month decreases the *PCS* for nearly all configurations.

3.2.2 | Effect of target toxicity probability

In Supplemental Figure 2A, the *PCS* was averaged across all priors and scenarios for simulations using a target toxicity probability of $\pi^* = 0.2$ or 0.3. The *PCS* is uniformly lower for a target of 0.3 versus 0.2. *Effect of number subgroups*. In Supplemental Figure 2B, the *PCS* was averaged across all priors and scenarios for simulations assuming $G = 2$ or $G = 3$ subgroups. The *PCS* is uniformly lower for 3 versus 2 subgroups.

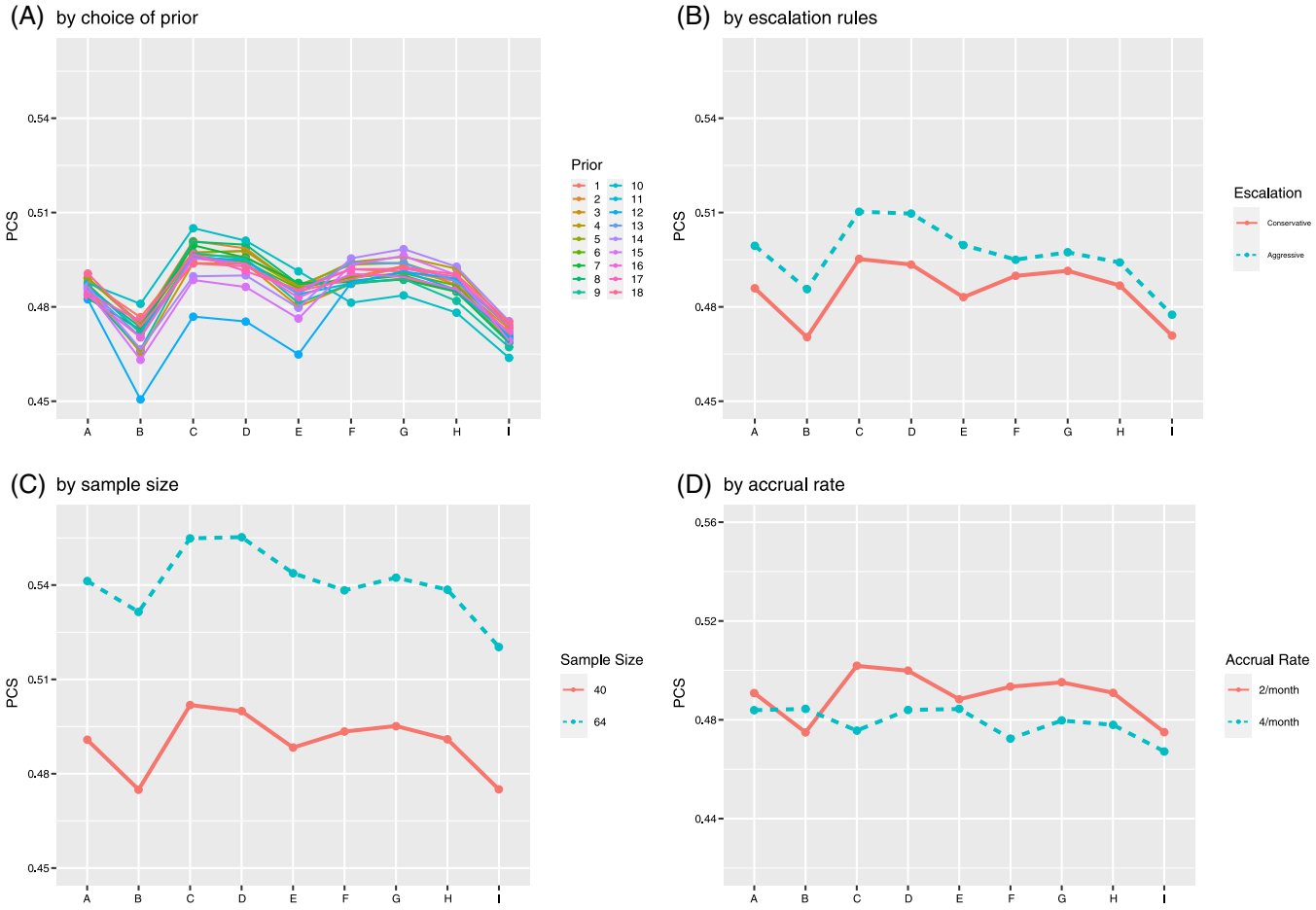

FIGURE 3 Mean probability of correct selection (*PCS*) separated by different: (A) priors; (B) escalation rules; (C) total sample size [N_{max}]; (D) accrual rate

Our secondary simulations suggest that our conclusions regarding *PCS* of the 9 2S-Sub-TITE configurations presented in Figure 2 are generally robust with respect to selection of prior toxicity probabilities, escalation rules, sample size, accrual rate, target toxicity probability, and number of subgroups.

4 | DISCUSSION AND CONCLUSIONS

Based on our simulation results, we recommend the 2S-Sub-TITE method configuration of beginning borrowing and clustering allowance with *p_hetero* = 0.7 at 75% patient accrual in trials when investigators are confident in their assumption of subgroup heterogeneity. For example, the 2S-Sub-TITE design can maximize dose-finding accuracy if preliminary evidence of subgroup heterogeneity in a prior clinical trial is available. This configuration of the 2S-Sub-TITE method increases the dose-finding accuracy by 2% to 6% compared to competing methods, without requiring an increase in the total sample size. For comparison, achieving a similar increase in accuracy of 5.1% requires substantially increasing the sample size by 60% ($N = 40$ to 64 patients). This method configuration also reduces the average number of toxicities by 0.25 to 0.88 compared to competing methods which account for subgroup heterogeneity. The potential trade-off is slightly reduced accuracy when the assumption of subgroup heterogeneity is incorrect. While the original Sub-TITE method configuration (enabling borrowing and clustering allowance with *p_hetero* = 0.9 from trial start) performs quite well, in agreement with the results shown in Chapple & Thall,[24] this recommended 2S-Sub-TITE method configuration is preferable when investigators are confident in subgroup heterogeneity.

While this suggestion is based on simulation results, this recommendation also makes intuitive sense. If a clinician truly believes that patient subgroups are heterogeneous, they may want to do separate dose-finding in those groups for

a large portion of the trial. Later in the trial, we can relax our heterogeneity assumption and allow the dose-finding data gathered for both subgroups to infer whether patient subgroups should be clustered.

While prior toxicity probability assumptions and choice of escalation rules do not have a significant effect on dose-finding accuracy in agreement with the existing literature,[28,29] increasing sample size does increase *PCS*. However, this increase is relatively uniform across all 2S-Sub-TITE configurations, thus not affecting our recommendations. Increasing the accrual rate in the trial generally decreases the *PCS* as a faster accrual rate leads to more participants having incomplete DLT observations when making dose escalation decisions.

Details regarding how these method configurations may be implemented using the *GetSubTite()* function in the SubTite R package can be viewed in Supplement D. We recommend that investigators conduct a thorough comparison of the 2S-Sub-TITE and Sub-TITE designs using trial simulations with parameters pertaining to their trial.

One strength of the paper is our comprehensive set of secondary simulations to assess a variety of secondary simulation parameters. Nonetheless, several parameters are assumed to be fixed including: five dose levels per subgroup starting at the lowest dose, a DLT observation period of one month, uniform occurance of DLT across the observation period. A potential limitation is that only 100 trials were simulated for each setting. However, our focus was to evaluate the designs across a large number ($N = 1000$) randomly-generated scenarios to ensure results are generalizable and not specific to investigator-selected scenarios.

Additionally, this 2-Stage design has the potential to be expanded to three stages: (1) separate TITE-CRM models, (2) Sub-TITE CRM model with only borrowing allowed, and (3) Sub-TITE CRM model with both borrowing and dynamic clustering allowed. While comprehensive simulations were performed with parameters settings varying over subgroup-specific true toxicity probability scenarios, subgroup-specific prior toxicity probabilities, prior probability of subgroup heterogeneity, escalation rules, sample sizes, and nine 2S-Sub-TITE configurations, this study was not feasibly able to address the operating characteristics of such a 3-Stage design because of the increased number of varying factors.

The 2S-Sub-TITE method is novel in its optimally varied use of data throughout the trial. There are several other trial designs, such as accelerated titration[4] and the partial order continual reassessment method (PO-CRM),[22] which effectively use a 2-Stage approach. However, the 2S-Sub-TITE method differs in that its switch to the second stage is based on an optimally chosen point in patient accrual, rather than DLT incidence. Moreover, this method is more flexible than designs such as the PO-CRM because it is not limited by any a priori assumptions regarding ordering of subgroups.[22]

From this simulation study we conclude that when investigators have high confidence in heterogeneity between two subgroups, using the 2S-Sub-TITE method configuration which enables borrowing and clustering allowance with *phetero* = 0.7 starting at 75% patient accrual exhibits superior dose-finding accuracy compared to the TITE-CRM and original Sub-TITE design.

ACKNOWLEDGMENTS

This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy 's Office of Science.[30,31] We give special thanks to Dr. Karen Wright and Dr. Wendy London for their thoughtful discussion and contributions. Funding for this project was provided by the Biostatistics Program at the Dana-Farber/Boston Children's Cancer and Blood Disorders Center.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Andrew G. Chapple https://orcid.org/0000-0001-5332-2730
Clement Ma https://orcid.org/0000-0002-8188-9600

REFERENCES

1. Nader JH, Neel DV, Shulman DS, Ma C, Bourgeois F, DuBois SG. Landscape of phase 1 clinical trials for minors with cancer in the United States. *Pediatr Blood Cancer*. 2020;67(11):e28694. PMID: 32886429; PMCID: PMC7896417. doi:10.1002/pbc.28694
2. Storer BE. Design and analysis of phase I clinical trials. *Biometrics*. 1989;45:925-937.
3. Skolnik JM, Barrett JS, Jayaraman B, Patel D, Adamson PC. Shortening the timeline of pediatric phase I trials: the rolling six design. *J Clin Oncol*. 2008;26:190-195. doi:10.1200/JCO.2007.12.7712

4. Simon R, Freidlin B, Rubinstein L, et al. Accelerated titration designs for phase I clinical trials in oncology. *J Natl Cancer Inst*. 1997;89:1138-1147. doi:10.1093/jnci/89.15.1138
5. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*. 1990;46:33-48.
6. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in phase I dose-finding studies. *Clin Trials*. 2008;5:465-477. doi:10.1177/1740774508096474
7. Ananthakrishnan R, Green S, Chang M, Doros G, Massaro J, LaValley M. Systematic comparison of the statistical operating characteristics of various phase I oncology designs. *Contemp Clin Trials Commun*. 2017;5:34-48. doi:10.1016/j.conctc.2016.11.006
8. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*. 2000;56:1177-1182. doi:10.1111/j.0006-341x.2000.01177.x
9. Normolle D, Lawrence T. Designing dose-escalation trials with late-onset toxicities using the time-to-event continual reassessment method. *J Clin Oncol*. 2006;24:4426-4433. doi:10.1200/JCO.2005.04.3844
10. Zhao L, Lee J, Mody R, Braun TM. The superiority of the time-to-event continual reassessment method to the rolling six design in pediatric oncology phase I trials. *Clin Trials*. 2011;8:361-369. doi:10.1177/1740774511407533
11. Salter A, O'Quigley J, Cutter GR, Aban IB. Two-group time-to-event continual reassessment method using likelihood estimation. *Contemp Clin Trials*. 2015;45:340-345. doi:10.1016/j.cct.2015.09.016
12. Cunanan KM, Koopmeiners JS. Efficacy/toxicity dose-finding using hierarchical modeling for multiple populations. *Contemp Clin Trials*. 2018;71:162-172. doi:10.1016/j.cct.2018.06.012
13. Cunanan KM, Koopmeiners JS. Hierarchical models for sharing information across populations in phase I dose-escalation studies. *Stat Methods Med Res*. 2018;27:3447-3459. doi:10.1177/0962280217703812
14. Morita S, Thall PF, Takeda K. A simulation study of methods for selecting subgroup-specific doses in phase 1 trials. *Pharm Stat*. 2017;16:143-156. doi:10.1002/pst.1797
15. O'Quigley J, Paoletti X. Continual reassessment method for ordered groups. *Biometrics*. 2003;59:430-440.
16. O'Quigley J, Shen LZ, Gamst A. Two-sample continual reassessment method. *J Biopharm Stat*. 1999;9:17-44. doi:10.1081/BIP-100100998
17. Yuan Z, Chappell R. Isotonic designs for phase I cancer clinical trials with multiple risk groups. *Clin Trials*. 2004;1:499-508. doi:10.1191/1740774504cn058oa
18. Thall PF, Nguyen HQ, Estey EH. Patient-specific dose finding based on bivariate outcomes and covariates. *Biometrics*. 2008;64:1126-1136. doi:10.1111/j.1541-0420.2008.01009.x
19. Thall PF, Wathen JK. Bayesian designs to account for patient heterogeneity in phase II clinical trials. *Curr Opin Oncol*. 2008;20:407-411. doi:10.1097/CCO.0b013e328302163c
20. Liu S, Pan H, Xia J, Huang Q, Yuan Y. Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Stat Med*. 2015;34:1681-1694. doi:10.1002/sim.6442
21. Horton BJ, Wages NA, Conaway MR. Shift models for dose-finding in partially ordered groups. *Clin Trials*. 2019;16:32-40. doi:10.1177/1740774518801599
22. Conaway MR, Wages NA. Designs for phase I trials in ordered groups. *Stat Med*. 2017;36(2):254-265. doi:10.1002/sim.7133
23. Cotterill A, Jaki T. Dose-escalation strategies which use subgroup information. *Pharm Stat*. 2018;17:414-436. doi:10.1002/pst.1860
24. Chapple AG, Thall PF. Subgroup-specific dose finding in phase I clinical trials based on time to toxicity allowing adaptive subgroup combination. *Pharm Stat*. 2018;17:734-749. doi:10.1002/pst.1891
25. TAK-580 Gliomas and other tumors https://ClinicalTrials.gov/show/NCT03429803.
26. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. *J Natl Cancer Inst*. 2009;101(10):708-720. doi:10.1093/jnci/djp079
27. Chapple AG. SubTite: Subgroup specific optimal dose assignment. R Package Version 4.0.0. R package version 4.0.1 ed. 2020.
28. Chevret S. The continual reassessment method in cancer phase I clinical trials: a simulation study. *Stat Med*. 1993;12:1093-1108. doi:10.1002/sim.4780121201
29. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med*. 1998;17:1103-1120. doi:10.1002/(sici)1097-0258(19980530)17:103.0.co;2–9
30. Rea P. The open science grid. *J Phys: Conf Series*. 2007;119:012057. doi:10.1088/1742-6596/78/1/012057
31. Sfiligoi I. The Pilot Way to Grid Resources Using glide in WMS. *WRI World Congress Comput Sci Inform Eng*. 2009;2:428-432. doi:10.1109/CSIE.2009.950

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: McGovern A, Chapple AG, Ma C. Two-stage subgroup-specific time-to-event (2S-SubTITE): An adaptive two-stage time-to-toxicity design for subgroup-specific dose finding in phase I oncology trials. *Pharmaceutical Statistics*. 2022;21(6):1138-1148. doi:10.1002/pst.2231