East Bayes
User Manual

v2.0.1

March 4, 2023

Cytel

Cytel East Bayes, 2.0.1, 2023

Contents

Part I

Phase Ia Single-Agent Dose-Finding Designs

1. Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment

1.1 Introduction

This module is about design and conduct of cohort-based phase I dose-finding trials for a single agent. The term "cohort-based" here means that patients are enrolled in cohorts, and dose-escalation decisions are also made in cohorts.

The primary objective of phase I trials is to identify the maximum tolerated dose (MTD), defined as the highest dose with a DLT rate less than or close to a prespecified targeted rate p_T (say, $p_T = 1/6$ or $1/3$). During the past three decades, a large number of designs have been developed for phase I trials. Figure 1.1 lists 12 representative designs over time. The 3+3 design by Storer (1989) has been the most popular design among physicians due to its simplicity in practice. It is a rule-based design and adaptively moves up and down cross doses by assigning three patients per cohort until the MTD is identified. Disadvantages of 3+3 are mainly the lack of reliability to identify the correct MTD (Chen et al., 2009), the lack of flexibility to accommodate patients drop-out or over-enrollment, and the poor statistical operating characteristics in terms of safety and reliability (Ji and Wang, 2013; Nie et al., 2016). Since 1990, many new methods, especially Bayesian methods, have been developed to guide dose escalation. The continual reassessment method (CRM) is the first Bayesian model-based design proposed by O'Quigley et al. (1990). It uses information from all doses to guide decision making. Neuenschwander et al. (2008) extend the CRM and propose the Bayesian logistic regression model (BLRM). Both CRM and BLRM use parametric dose-response curves for statistical modeling and inference. Founded on sound statistical principles, both designs exhibit superior performance when compared with 3+3. However, they are complex and need strong statistical input to safe-guard the practical deployment, which makes them challenging for clinicians to comprehend and implement in practice. In the recent decades, the landscape of phase I dose-

finding designs has been rapidly shifting, noticeably marked by the emergence of interval-based designs, such as the toxicity probability interval (TPI) design (Ji et al., 2007) and two subsequent modifications, the mTPI (Ji et al., 2010; Ji and Wang, 2013) and mTPI-2 (Guo et al., 2017b) designs. In parallel, the cumulative cohort design (CCD) (Ivanova et al., 2007) and the Bayesian optimal interval (BOIN) design (Liu and Yuan, 2015a) further simplify the statistical inference based on a point estimate of toxicity probability and prespecified interval boundaries. BOIN is an overly refined version of CCD, in which the interval boundaries are generated based on an ad-hoc objective function that creates theoretically shaky results. In our East Bayes platform, we decide to adopt and modify the CCD design, following our principle to promote sound methodologies. Finally, in 2019, the evolutionary step of phase I dose-finding designs spirals back to the rule-based approaches in the form of the i3+3 design (Liu et al., 2020), which shows the potential of smart rule-based designs that can achieve comparable operating characteristics to model-based designs.

In this module of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment**, East Bayes performs trial simulation to examine the operating characteristics of eight designs, including i3+3 (Liu et al., 2020), mTPI-2 (Guo et al., 2017b), CRM (O'Quigley et al., 1990), 3+3 (Storer, 1989), mTPI (Ji et al., 2010), modified CCD (mCCD) (Ivanova et al., 2007), BLRM (Neuenschwander et al., 2008) and BOIN (Liu and Yuan, 2015a) designs. Also, the decision table generation and the MTD estimation are incorporated in this module, so that users may generate the decision tables to guide trial conduct and estimate the MTD after trial completion. §1.2 introduces the user interface and tutorial of launching trial simulations and examining results, as well as generating decision tables and estimating MTD. A statistical review of all eight designs are provided in §1.3.

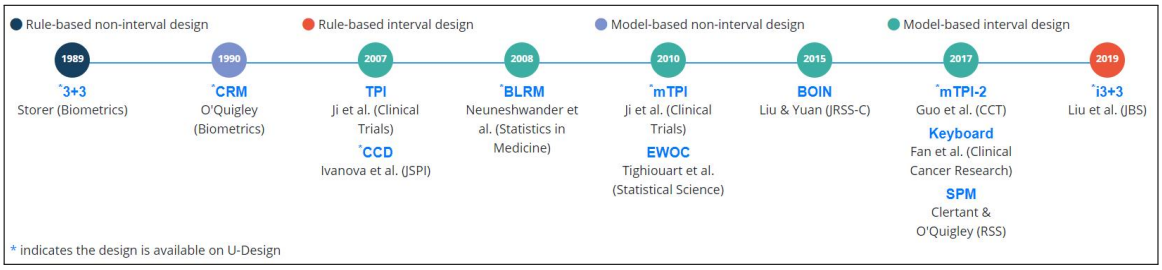


Figure 1.1: The chronicle of phase I dose finding designs (1989-2019).

1.2 User Interface and Tutorial

1.2.1 Overview

Entering the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** page, users will see four main tabs: **Simulation Setup**, **Simulation Results**, **Decision Table** and **MTD Estimation**. The first two tabs allow users to conduct simulations and visualize/download simulation results, and the next two tabs allow users to generate decision tables and estimate the MTD, respectively. In the **Simulation Setup** tab, there are three steps (Figure 1.2): 1) **Set trial parameters**, 2) **Select designs**, and 3) **Generate scenarios**. Users need to complete the steps 1-3 to set up simulations for a single design or multiple designs. Upon completing steps 1-3, users click the "Launch Simulation" button at the bottom of the page. Users may also click the "Reset" button next to **Launch Simulation** to clear all settings. After the simulation is launched, the results of simulations will be displayed in the **Simulation Results** tab. The simulation process can be monitored in real time at the top of the **Simulation Results** tab. Detailed steps of using this module are elaborated next in §1.2.2-§1.2.5.

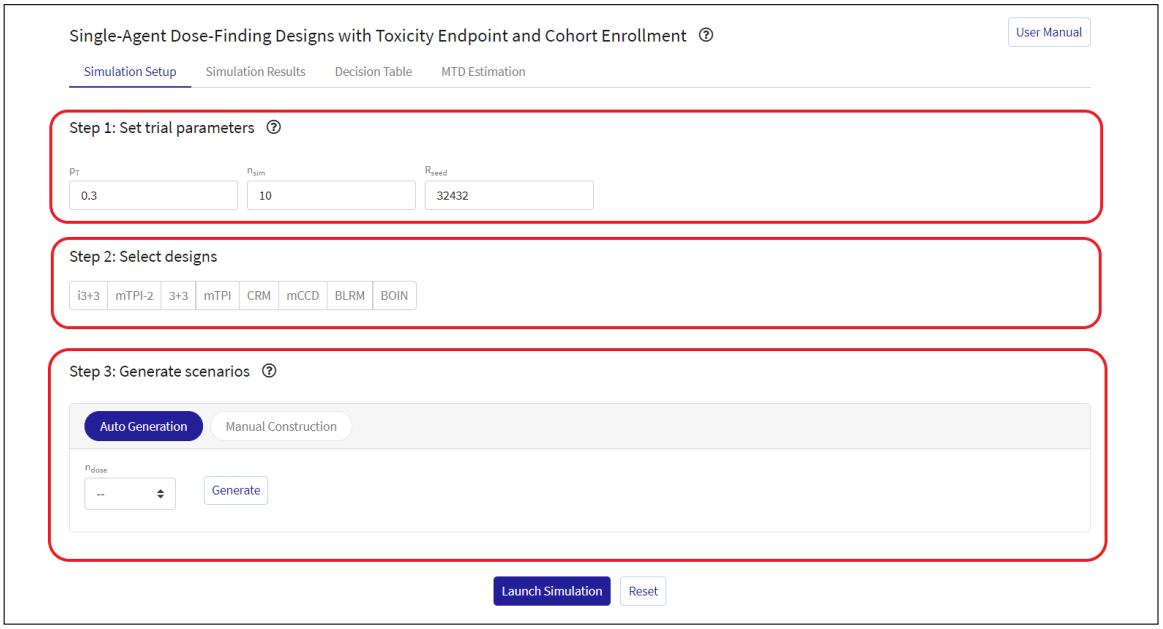


Figure 1.2: Simulation Setup in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

1.2.2 Simulation Setup

In the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module, East Bayes provides eight designs, i3+3, mTPI-2, CRM, 3+3, mTPI, mCCD, BLRM, and BOIN for simulation. Users can choose up to four design configurations for simultaneous comparison in the **Simulation Setup** tab each time. A design configuration means a design such as i3+3, along with the designs settings, such as sample size. Request to allow more than four design configurations by emailing `support@cytel.com`.

1.2.2.1 Step 1: Set trial parameters

Specify the target toxicity probability (p_T), number of simulations (n_{sim}), and random seed of simulation (R_{seed}) for the simulated trials. See Figure 1.3. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameters. The detailed explanation of the above three input arguments is provided in Table 1.1.

Figure 1.3: Set trial parameters in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Table 1.1: Input parameters for trials in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Notation	Parameters	Description
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T. Default value is 0.3.
n_{sim}	The number of simulated trials	The maximum number of simulated trials allowed is 10,000. Default value is 1,000.
R_{seed}	The random seed of simulation	A random seed is a number used to initialize a pseudorandom number generator in the simulation. Default value is 32432.

1.2.2.2 Step 2: Select designs

To select a design, click the button with the design's name on it. Up to four design configurations may be selected for comparison.

Click the "More" link to expand the design list to see all the seven designs and click the "Less" to collapse the list.

Check the "Apply Stopping Rule" box to apply an ad-hoc stopping rule of reaching the maximum number of patients at a dose level during the trial conduct. See the detailed rules in Table 1.2 and §1.3.

Design parameters can be modified in the input box. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameters. See detailed parameter descriptions in Table 1.2.

Figure 1.4: Select designs in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Table 1.2: Input parameters for designs in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Notation	**Parameters**	**Description**
n (all designs)	Sample size	The maximum number of patients to be treated in the trial. The upper limit is set at 100 since the number of patients that are enrolled in phase I clinical trial is typically small. Default value is 30.
d_{start} (all designs)	Starting dose level	The starting dose level in the simulated trials. Default value is 1.
n_{cohort} (except 3+3)	Cohort size	The number of patients in each cohort. Default value is 3.

ϵ_1,ϵ_2 (i3+3, mTPI, mTPI-2, mCCD, BLRM)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence/target interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose MTD. Default values for both are 0.05.
ϵ_1,ϵ_2 (BOIN)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the optimal interval and the target probability. Here, $\epsilon_1 = p_T - \lambda_1$, $\epsilon_2 = \lambda_2 - p_T$ where (λ_1, λ_2) is the optimal interval minimizing the probability of making an erroneous decision based on the initial equivalence interval (ϕ_1, ϕ_2). Default values for ϕ_1, ϕ_2 are $0.6 * p_T$ and $1.4 * p_T$.
p_{EWOC} (BLRM)	Cutoff probability of escalation with overdose control	The threshold of controlling the probability of excessive or unacceptable toxicity. Default value is 0.25.
δ (CRM)	Half-width	The halfwidth of the indifference interval in selecting the skeleton of the model. Default value is 0.05.
K (except 3+3)	Maximum number of patients at a dose level	A number used in the "Stopping Rule" that stops a trial if 1) the dose-assignment decision is to escalate to the next higher dose and there has been K patients enrolled at that dose; or 2) the dose-assignment decision is to stay at the current dose and there has been K patients enrolled at that dose; or 3) if the dose-assignment decision is to de-escalate to the previous lower dose and there has been K patients enrolled at that dose; Default value is 12.

For the BOIN design, click the "Compute" button to compute the initial equivalence interval (ϕ_1, ϕ_2) using the optimal interval $(p_T - \epsilon_1, p_T + \epsilon_2)$. See details in §1.3.8.

Click the "Delete" button to remove the selected designs.

Click the "Apply" button of all the designs before launching simulations to apply all settings.

1.2.2.3 Step 3: Generate scenarios

There are two ways to generate scenarios, automatically (in below **Auto Generation** tab, see Figure 1.5) or through manual construction (in below **Manual Construction** tab, see Figure 1.6). Once scenarios are generated, click the "Launch Simulation" button at the bottom of the page to run n_{sim} (set in step 1) simulations, for each scenario and selected design (set in step 2) combination, assuming p_T (set in step 1).

Auto Generation (Figure 1.5)

Select the number of doses n_{dose} ($3 \leq n_{dose} \leq 10$) from the dropdown box. Upon clicking the "Generate" button, five or six scenarios will be created automatically, each of which contains the true toxicity probabilities for n_{dose} dose levels. These generated scenarios are displayed and editable. The detailed algorithm for scenarios auto generation is provided next.

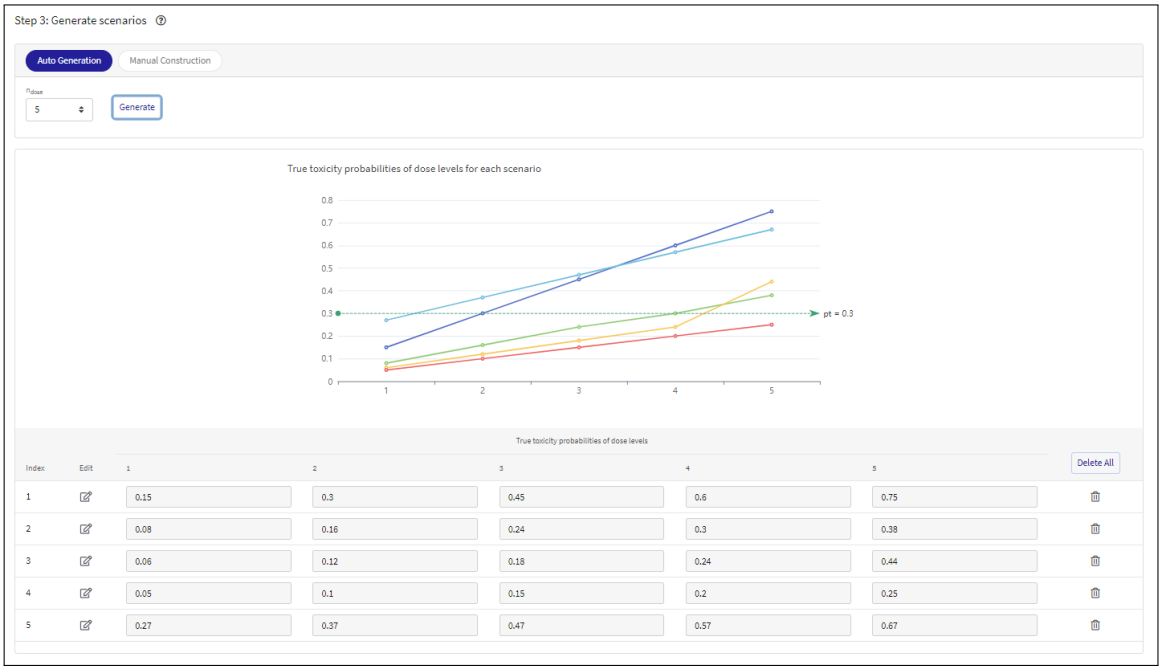


Figure 1.5: Automatically generate scenarios in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Manual Construction (Figure 1.6)

Follow the instructions below to manually construct scenarios. Then click the "Add" button to create these scenarios. The format of input must comply with the following instructions.

- Scenarios should be separated by line breaks;
- Each scenario consists a set of true toxicity probabilities for all dose levels;
- The true toxicity probabilities must be separated by a white space or comma.

For example, by inputting “0.05 0.1 0.15 0.2” or “0.05,0.1,0.15,0.2”, a scenario is presented with true toxicity probabilities of four dose levels, 0.05, 0.1, 0.15 and 0.2.

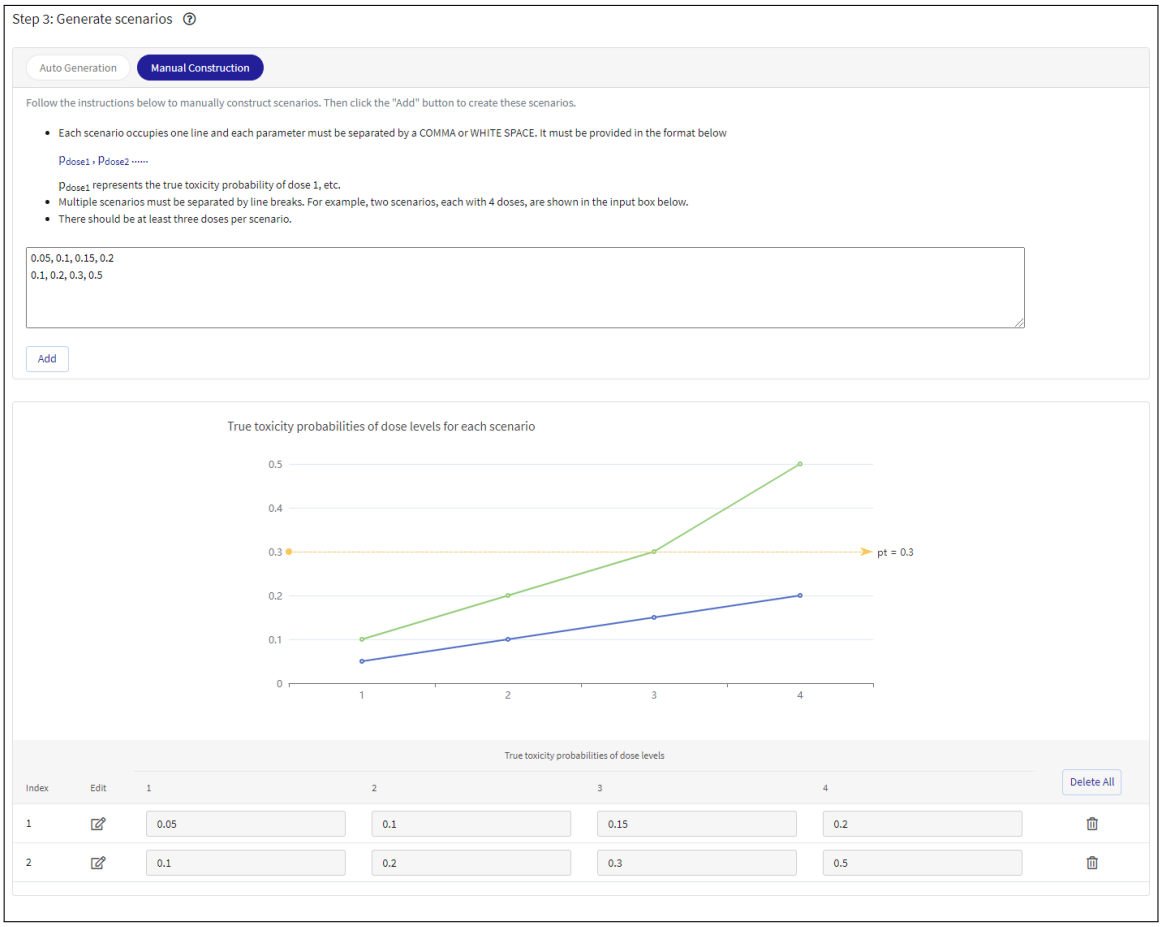


Figure 1.6: Manually generate scenarios in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

The generated scenarios are displayed as a list (Figures 1.5 and 1.6) which appears below the generation section. The generated scenarios are editable by clicking the edit icon . An interactive chart will also be generated to visually display the shape of true toxicity probabilities for each scenario.

Algorithm for Auto Generation

By entering the number of candidate dose levels n_{dose}, five or six scenarios are generated automatically. See Figure 2.7 for an illustration. They represent the four types of dose-response shapes below.

Types	Dose-Response Shape
Ideal	Some doses are tolerable but some are overly toxic, AND there exists at least one dose level close to the target p_T or falling within the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$.
Safe	All doses are safe and tolerable with the true toxicity probabilities lower than the target p_T or the lower bound of the equivalence interval ($p_T - \epsilon_1$).
Toxic	All doses are overly toxic with the true toxicity probabilities higher than the target p_T or the upper bound of the equivalence interval ($p_T + \epsilon_2$).
Steep	Some doses are tolerable but some are overly toxic, AND there is a steep jump in the toxicity probability between two adjacent doses, AND there is no dose close to the target p_T or falling within the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$.

Two "Steep" scenarios are generated, with the toxicity probability steep jump occurring at the first or second half of the doses. Similarly, two "Ideal" scenarios might be generated, with the MTD placed in the first or second half of the doses. This depends on the number of doses. When the number of doses is greater than 6, two scenarios of "Steep" and "Ideal" will be generated.

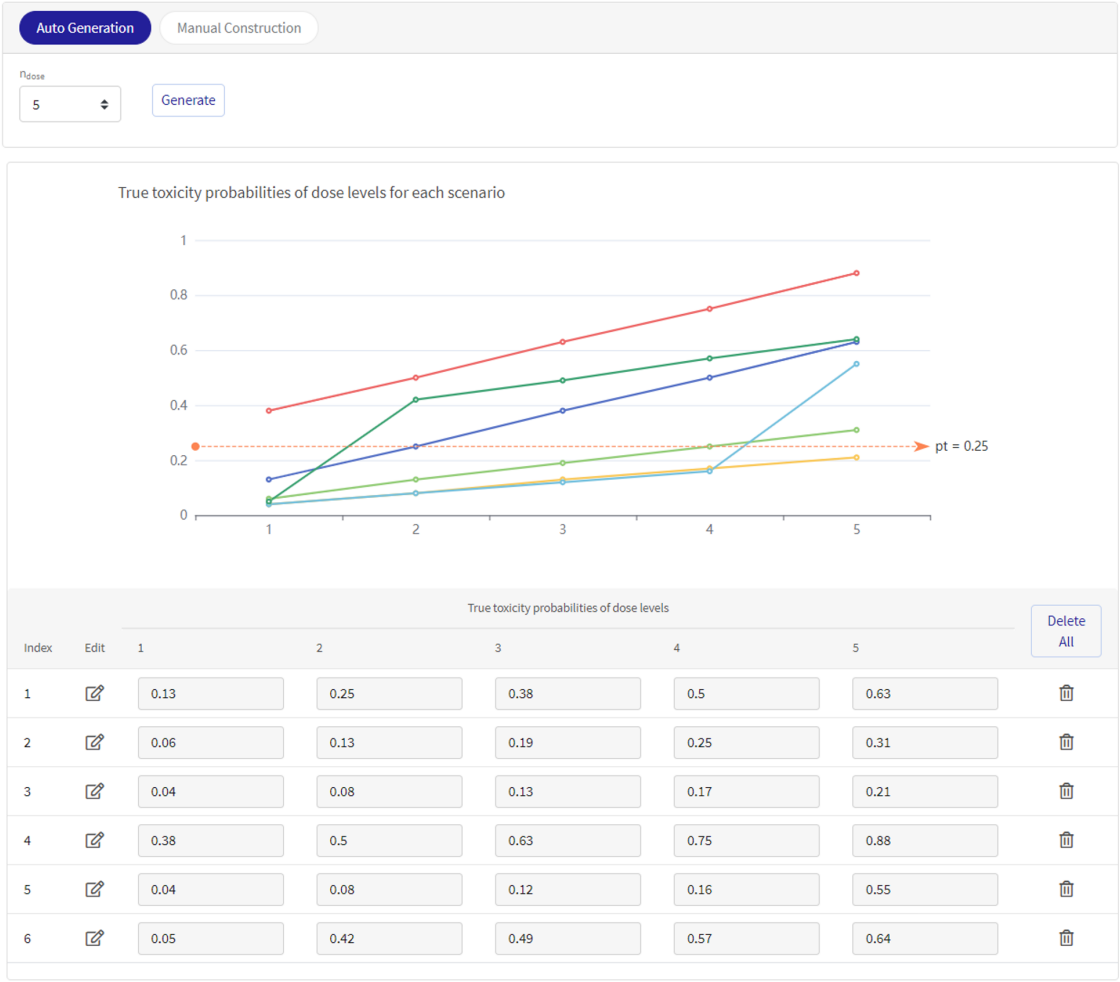


Figure 1.7: An example of automatically generated scenarios. Five dose levels are considered for the trial. The target toxicity probability is $p_T = 0.25$, and the equivalence interval is EI=$[0.2, 0.3]$. The six different lines represents the four types of scenario. In the "Ideal" scenarios (Lines 1 and 2), doses 2 and 4 are the true MTD with toxicity probability falling within the EI, respectively. In the "Safe" scenario (Line 3), all doses are safe with toxicity probabilities lower than the target $p_T = 0.25$. The "Toxic" scenario (Line 4) gives a contrary situation to the "Safe" scenario, where all doses are overly toxic with the toxicity probabilities higher than the target $p_T = 0.25$. The remaining two lines (Lines 5 and 6) are the "Steep" scenarios, in which some doses are tolerable but some are overly toxic, and there is a steep jump in the toxicity probability occurring at the first or second half of the doses (between doses 4 and 5 in Line 5, and doses 1 and 2 in Line 6).

1.2.2.4 Launch Simulation

Once the steps 1-3 are completed, users can conduct simulated clinical trials to examine the operating characteristics of the selected designs using the selected scenarios, by clicking the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 1.5 and 1.6). "Success" message will be displayed on the website as in Figure 1.8 to indicate that the simulation has been successfully launched. Users may click the "OK" button in the pop-up box to track the simulation processing status and simulation results.

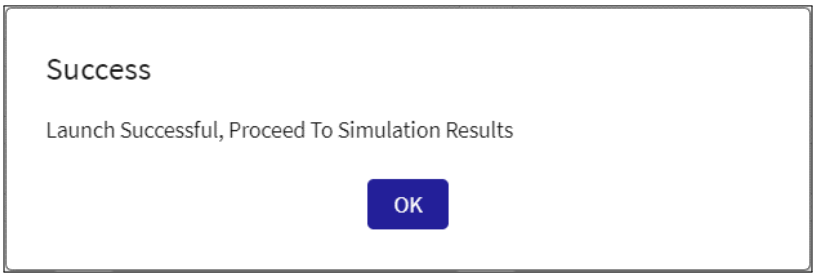


Figure 1.8: "Success" message after launching simulation in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

1.2.3 Simulation Results

In the **Simulation Results** tab, users can view and delete the simulation progress and simulation results (§1.2.3.1), inspect the escalation process in two simulated trials (§1.2.3.2), restore the simulation settings if needed (§1.2.3.3), and download intelligent simulation reports (§1.2.3.4). Specifically, all the simulation results (figures and tables) can be downloaded in Word format, accompanying the statistical sections in a trial protocol. Hereinafter, we use simulation results and operating characteristics interchangeably.

1.2.3.1 View simulation results

In the **Simulation Results** tab, the **Running Simulations** panel exhibits the progress of ongoing simulation (Figure 1.9). The ongoing simulations are displayed in ascending order by the launch time. Click the icon "×" to delete the corresponding simulation.

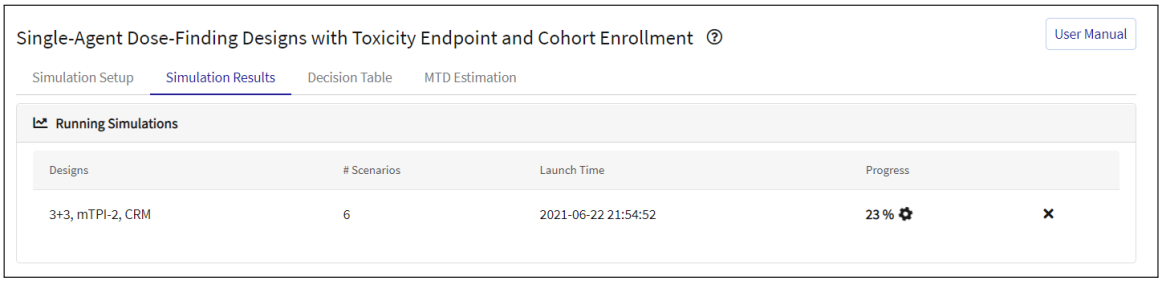


Figure 1.9: Simulation progress in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Once the simulations are completed, the **Running Simulations** panel in Figure 1.9 will disappear, green "*simulation result created*" massages will appear instead and stay at the same place of the **Running Simulations** panel unless explicitly dismissed by clicking the icon "×" at the end of the corresponding row, and the simulation results will be automatically loaded into the **Simulation History** panel (Figure 1.10), with the blue mail icon shown to indicate new results. All the previously completed simulations are also listed in the **Simulation History** panel. Simulation results for other modules can also be viewed under the **Simulation History** by dropping down the "Select a Design Category" button (Figure 1.10). Click the button to delete the selected simulation results.

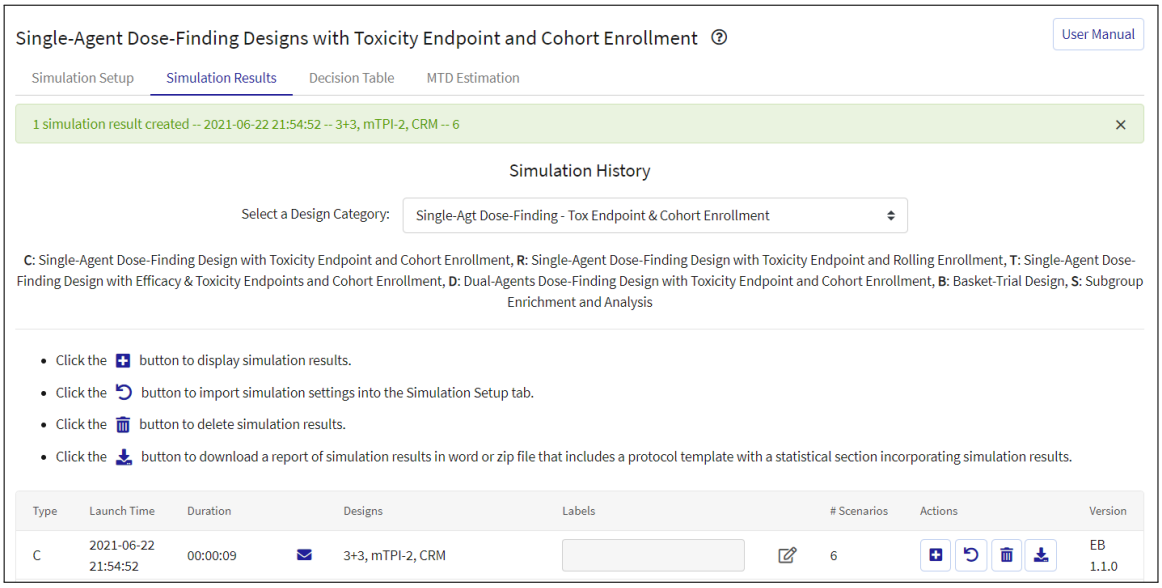


Figure 1.10: Simulation Results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Click the button ⊞ to unfold the simulation results (Figure 1.11). The design settings are firstly displayed at the top of each simulation study (Figure 1.11). Then the results of simulation are shown as plots and tables below.

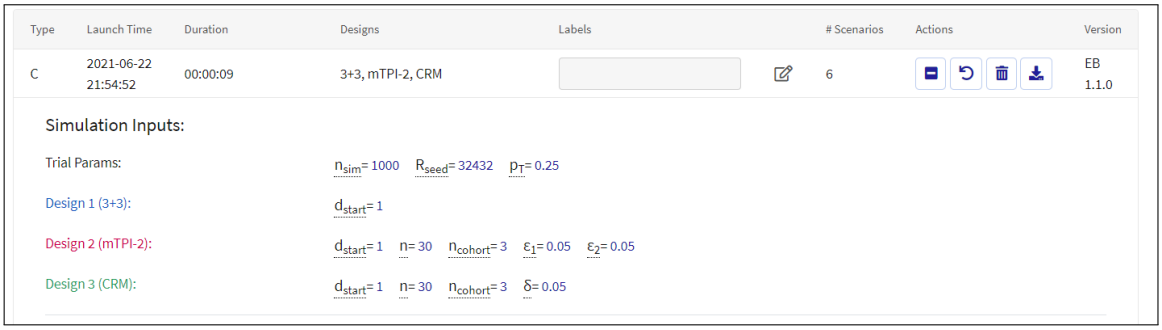
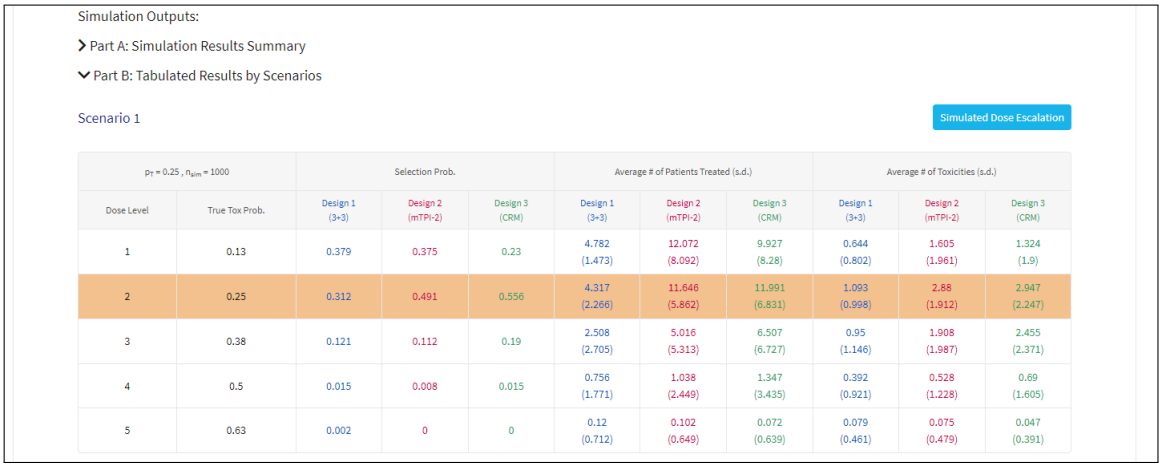


Figure 1.11: View the simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Details of the Simulation Results

The simulation results are divided into two parts, i.e, Simulation Result Summary and Tabulated Results by Scenarios. Each part can be viewed or hidden by clicking the button for that part (Figure 1.12).



$p_T = 0.25$, $n_{sim} = 1000$		Selection Prob.			Average # of Patients Treated (s.d.)			Average # of Toxicities (s.d.)		
Dose Level	True Tox Prob.	Design 1 (3+3)	Design 2 (mTPI-2)	Design 3 (CRM)	Design 1 (3+3)	Design 2 (mTPI-2)	Design 3 (CRM)	Design 1 (3+3)	Design 2 (mTPI-2)	Design 3 (CRM)
1	0.13	0.379	0.375	0.23	4.782 (1.473)	12.072 (8.092)	9.927 (8.28)	0.644 (0.802)	1.605 (1.961)	1.324 (1.9)
2	0.25	0.312	0.491	0.556	4.317 (2.266)	11.646 (5.862)	11.991 (6.831)	1.093 (0.998)	2.88 (1.912)	2.947 (2.247)
3	0.38	0.121	0.112	0.19	2.508 (2.705)	5.016 (5.313)	6.507 (6.727)	0.95 (1.146)	1.908 (1.987)	2.455 (2.371)
4	0.5	0.015	0.008	0.015	0.756 (1.771)	1.038 (2.449)	1.347 (3.435)	0.392 (0.921)	0.528 (1.228)	0.69 (1.605)
5	0.63	0.002	0	0	0.12 (0.712)	0.102 (0.649)	0.072 (0.639)	0.079 (0.461)	0.075 (0.479)	0.047 (0.391)

Figure 1.12: View each part of the simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Part A: Simulation Result Summary

There are four sections in the Simulation Result Plots:

A. Line plots showing five summary statistics of the simulation results for all the designs (Figure 1.13), including **Prob. of Selecting MTD**, **Prob. of Toxicity**, **Prob. of Selecting Does-over-MTD**, **Prob. of Overdosing Allocation**, and **Mean Squared Error**, for each scenario.

B. A table of mean and standard deviation (s.d.) for the five summary statistics (Figure 1.14).

C. [Optional] An empirical CRM decision table if CRM is selected in the simulation (Figure 1.15).

D. [Optional] An empirical BLRM decision table if BLRM is selected in the simulation (Figure 1.16).

A. Line plots:

- The five summary statistics are part of operating characteristics of the designs. They are explained in full detail next.
  - **Prob. of Selecting MTD**: The probability of selecting the true MTD, defined as the proportion of simulated trials that correctly select the true MTD. The higher the value, the better the design.

* For interval-based designs (i3+3, mTPI, mTPI-2, BLRM, & mCCD), the true MTDs are defined as the dose levels of which the true toxicity probabilities fall into the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$; if none of the doses have a toxicity probability that falls into the equivalence interval, the true MTD is defined as the dose with the highest toxicity probability below p_T. For the non-interval-based designs, 3+3 and CRM, the true MTDs is defined as the dose levels with the highest toxicity probabilities lower than or equal to p_T.
* To compare the operating characteristics of multiple designs submitted in a simulation study, the definition of MTD should be unified. If any of interval-based designs (i3+3, mTPI, mTPI-2, BLRM, & mCCD) are used in the simulation, the dose levels of which the true toxicity probabilities fall into the widest equivalence interval $[p_T - \max\{\epsilon_1\}, p_T + \max\{\epsilon_2\}]$ are defined as the true MTDs. Here, $\max\{\cdot\}$ is taken over the designs. If none of the doses fall in, the dose with the highest toxicity probability that is below p_T is the true MTD. For example, consider a case in which users compare four designs, mTPI, mTPI-2, CRM and 3+3, in a simulation study targeting $p_T = 0.3$. Suppose $\epsilon_1 = 0.02$ and $\epsilon_2 = 0.05$ for mTPI, and $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.03$ for mTPI-2. In this case, the true MTD is the dose levels with toxicity probabilities in [0.3-0.05, 0.3+0.05]; if none of the doses have a toxicity probability in [0.3-0.05, 0.3+0.05], the dose with the highest toxicity probability lower than 0.3 is the true MTD.
* If a scenario does not have any MTD (e.g., all doses have toxicity probabilities larger than the target p_T), no selection is the right decision. In this case, the probability of selecting the true MTD is the probability of no selection.

– **Prob. of Toxicity**: The proportion of patients who have experienced DLT across all the simulated trials. The lower the number, the fewer patients having DLTs under the design.
– **Prob. of Selecting Does-over-MTD**: The probability of selecting the dose levels above the true MTD, which is defined by the proportion of simulated trials that select a dose higher than the true MTD at the end of the trial.The lower the value, the better the safety of the design.
– **Prob. of Overdosing Allocation**: The average proportion of patients who are assigned to doses higher than the MTD by the design across all the simulated trials.
– **Mean Squared Error**: The average mean squared error in the toxicity probability of

selected MTD, across all the simulated trials, defined as the average squared distance between the true toxicity probability of the selected dose, and the true toxicity probability of the true MTD for each scenario across the simulations. The scenarios with no true MTDs are excluded.

- For each line plot, the x-axis is the index of scenario and the y-axis is the value of summary statistics. Lines with different colors represent different designs.
- The plots are interactive for better visualization.
  - Hover the mouse on a dot and a box will display the value of each design at the corresponding scenario (e.g. top left plot in Figure 1.13: Prob. of Selecting MTD).
  - Hover the mouse on the design label to highlight the corresponding line and fade the others (e.g. bottom right plot in Figure 1.13: Prob. of Overdosing Allocation).
  - Click the design label to hide the corresponding line and click again to change it back (e.g. top right plot in Figure 1.13: Prob. of Toxicity).

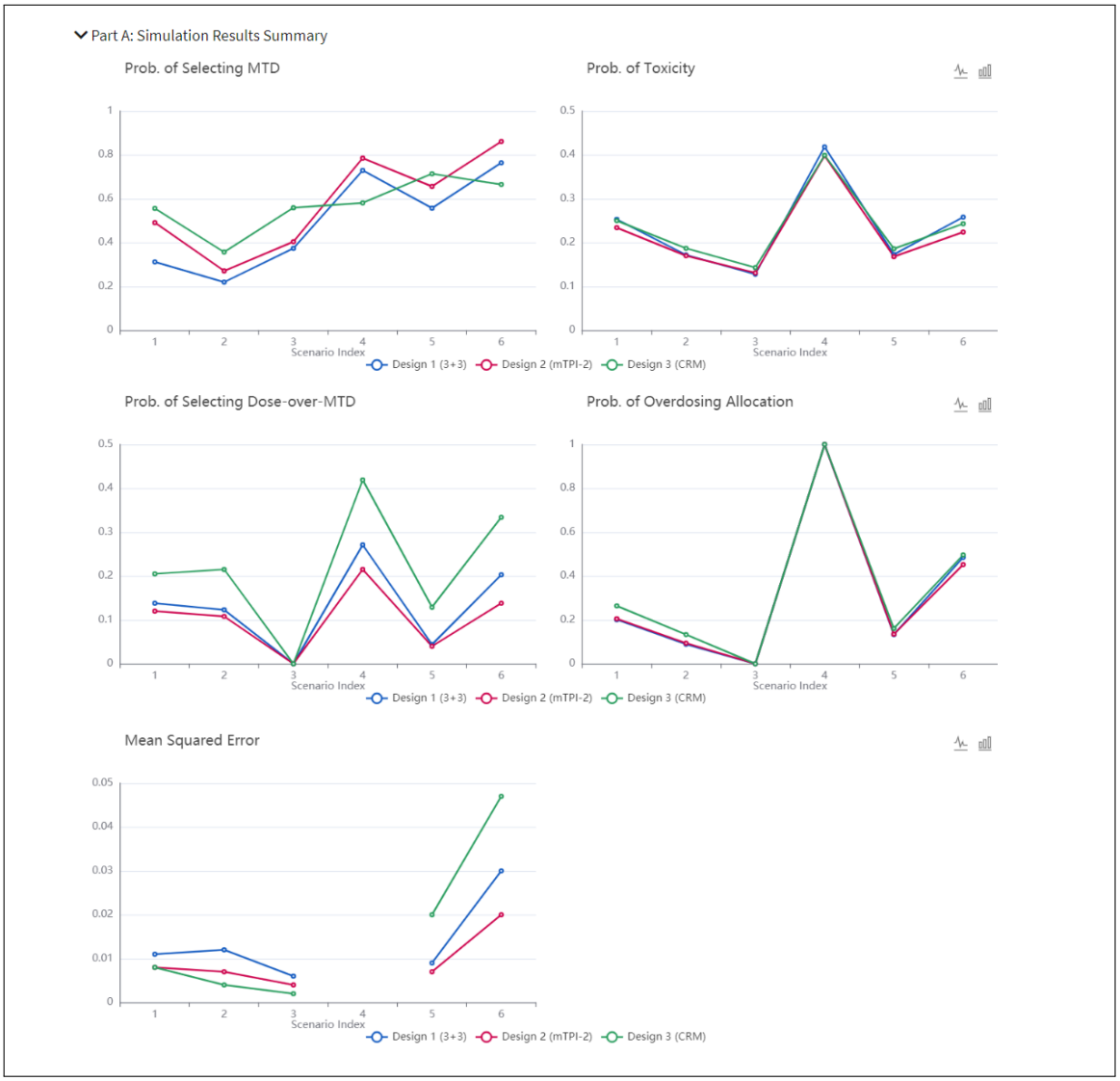


Figure 1.13: Simulation result plots in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

B. Simulation summary table: Figure 1.14 shows the mean±sd of the summary statistics across all scenarios for each design.

C. CRM decision table:

An empirical CRM decision table will be provided in the simulation results if CRM is included in the simulation (Figure 1.15). This table summarizes the frequency of decisions made by CRM across all the simulated trials.

Summary of Performance

	Design 1 (3+3)	Design 2 (mTPI-2)	Design 3 (CRM)
Prob. of Selecting MTD	0.493 ± 0.226	0.578 ± 0.229	0.572 ± 0.123
Prob. of Toxicity	0.234 ± 0.104	0.221 ± 0.095	0.235 ± 0.090
Prob. of Selecting Dose-over-MTD	0.130 ± 0.100	0.104 ± 0.076	0.217 ± 0.148
Prob. of Overdosing Allocation	0.318 ± 0.373	0.314 ± 0.369	0.342 ± 0.362
Mean Squared Error	0.014 ± 0.009	0.009 ± 0.006	0.016 ± 0.019

* Mean ± Standard Deviation

Figure 1.14: Simulation summary table in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

- The lengths of the three colored bars in one cell represent the frequencies of the corresponding dose-finding decisions. The longer the bar, the higher the frequency. For example, the cell in the figure shows that CRM stay at the current dose 31.3% of the times when 2 out of 3 patients experience DLTs at a dose.

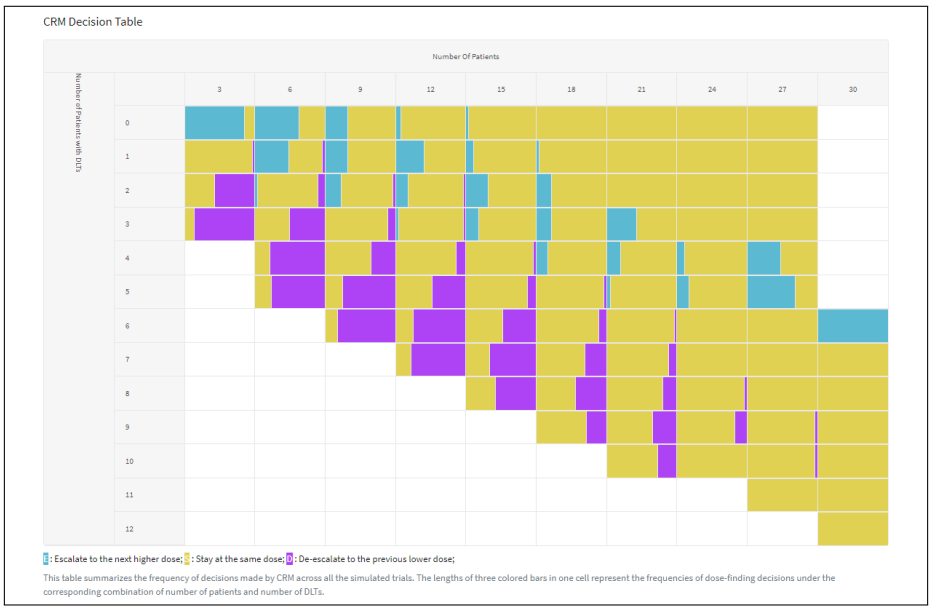


Figure 1.15: CRM decision table in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

D. BLRM decision table:

An empirical BLRM decision table will be provided in the simulation results if BLRM is included

in the simulation (Figure 1.16). This table summarizes the frequency of decisions made by BLRM across all the simulated trials.

- The lengths of the three colored bars in one cell represent the frequencies of the corresponding dose-finding decisions. The longer the bar, the higher the frequency. For example, the cell in the figure shows that BLRM de-escalates to the previous lower doses 26.3% of the times when 1 out of 3 patients experienced DLT.

Figure 1.16: BLRM decision table in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Part B: Tabulated Results by Scenarios

Full simulation results are presented in tabular format arranged by scenarios (Figure 1.17).

In the upper part of Figure 1.17, the first two columns summarize dose levels and their true toxicity probabilities; the remaining columns report three dose-specific summary statistics from the simulations: selection probability, average number of patients treated, and average number of toxicities (i.e. DLTs), along with their standard deviations, at each dose level. Specifically, they are

1) **Selection Prob.**: The proportion of simulated trials that select each dose level as the MTD.
2) **Average** # of Patients Treated (s.d.): The average number of patients treated at each dose level and its standard deviation.
3) **Average** # of Toxicities (s.d.): The average number of patients experienced DLT at each dose level and its standard deviation.

The true MTD(s) of the scenario is(are) highlighted by the orange bar. For the definition of the true MTD in the simulation results, please refer to the definition of **Prob. of Selecting MTD** in the **Simulation Results Plots** above (after Figure 1.11).

In the lower part of Figure 1.17, more trial-specific summary statistics are reported, mainly from five aspects: **MTD Selection**, **Patient Assignment**, **Trial Toxicity**, **Trial Stopping** and **Trial Sample Size**. Click the "More" link to show the summary statistics of **Trial Stopping** and **Trial Sample Size** and click the "Less" to collapse these results. Specifically, they are

- **MTD Selection**
  - **Prob. of Selecting MTD**: The proportion of simulated trials that select the true MTD at the end of the trial.
  - **Prob. of Selecting Does-over-MTD**: The proportion of simulated trials that select the doses higher than the true MTD at the end of the trial.
  - **Prob. of No Selection**: The proportion of simulated trials in which none of the dose levels are selected as the MTD. If a scenario does not have any MTD, this values is treated as the probability of selecting the true MTD.

  For detailed descriptions, please refer to **Simulation Result Plots** section above (after Figure 1.11).
- **Patient Assignment**
  - **Prob. of Correct Allocation (s.d.)** : The average proportion of patients who are correctly assigned to the true MTD by the design across all the simulated trials and its standard deviation.
  - **Prob. of Overdosing Allocation (s.d.)** : The average proportion of patients who are

assigned to doses higher than the MTD by the design across all the simulated trials and its standard deviation.

- **Trial Toxicity**
  - **Prob. of Toxicity**: The proportion of patients experiencing DLT across all the simulated trials. For detailed descriptions, please refer to **Simulation Result Plots** section above (after Figure 1.11).
- **Trial Stopping**
  - **Prob. of Early** Stopping Trial due to Safety Rule: The proportion of simulated trials in which the trial is stopped because the first dose level shows unacceptable toxicity.
  - **Prob. of Early** Stopping Trial due to Reaching K: The proportion of simulated trials in which the trial is stopped because the dose-assignment decision is to escalate/stay/de-escalate to a dose level but that dose has enrolled at least K patients ($K < n$, e.g., $K = 12$) .
  - **Prob. of Stopping Trial due to Reaching** n: The proportion of simulated trials in which the trial is stopped because the total number of patients enrolled and treated in a trial has reached or exceeded the pre-specified maximum sample size n.
- **Trial Sample Size**
  - **Average #** of Patients Treated (s.d.): The average number of patients treated in the simulated trials and its standard deviation. Due to early stopping, this number is lower than or equal to n.
- **Accuracy of Selected MTD**
  - **Mean Squared Error**: The mean squared error is the average squared distance between the true toxicity probability of the selected dose and that of the true MTD across the simulations. If the scenario has no true MTD, N/A is displayed.

When calculating the standard deviation, we use n_{sim} as the denominator instead of $(n_{sim} - 1)$ in East Bayes.

Part B: Tabulated Results by Scenarios

Scenario 1

Simulated Dose Escalation

pT = 0.3 , nsim = 10000		Selection Prob.				Average # of Patients Treated (s.d.)				Average # of Toxicities (s.d.)			
Dose Level	True Tox Prob.	Design 1 (i3+3)	Design 2 (3+3)	Design 3 (mTPI-2)	Design 4 (mCCD)	Design 1 (i3+3)	Design 2 (3+3)	Design 3 (mTPI-2)	Design 4 (mCCD)	Design 1 (i3+3)	Design 2 (3+3)	Design 3 (mTPI-2)	Design 4 (mCCD)
1	0.15	0.335	0.451	0.335	0.218	10.195 (7.855)	5.03 (1.403)	10.195 (7.855)	9.902 (7.582)	1.53 (1.923)	0.745 (0.837)	1.53 (1.923)	1.486 (1.846)
2	0.3	0.507	0.275	0.507	0.561	13.048 (6.868)	4.226 (2.341)	13.048 (6.868)	12.907 (6.56)	3.914 (2.476)	1.286 (1.027)	3.914 (2.476)	3.873 (2.41)
3	0.45	0.136	0.065	0.136	0.195	5.559 (5.997)	1.896 (2.502)	5.559 (5.997)	5.885 (5.927)	2.497 (2.478)	0.844 (1.151)	2.497 (2.478)	2.643 (2.459)
4	0.6	0.01	0.004	0.01	0.016	0.899 (2.417)	0.387 (1.265)	0.899 (2.417)	1 (2.515)	0.544 (1.361)	0.235 (0.75)	0.544 (1.361)	0.603 (1.42)
5	0.75	0	0	0	0	0.048 (0.475)	0.029 (0.339)	0.048 (0.475)	0.056 (0.503)	0.036 (0.341)	0.021 (0.233)	0.036 (0.341)	0.041 (0.363)

		Design 1 (i3+3)	Design 2 (3+3)	Design 3 (mTPI-2)	Design 4 (mCCD)
MTD Selection*	Prob. of Selecting MTD	0.507	0.275	0.507	0.561
	Prob. of Selecting Dose-over-MTD	0.146	0.069	0.146	0.211
	Prob. of No Selection	0.013	0.205	0.013	0.011
Patients Assignment	Prob. of Correct Allocation (s.d.)	0.435 (0.229)	0.33 (0.178)	0.435 (0.229)	0.43 (0.219)
	Prob. of Overdosing Allocation (s.d.)	0.217 (0.244)	0.143 (0.194)	0.217 (0.244)	0.231 (0.244)
Trial Toxicity	Prob. of Toxicity	0.286	0.271	0.286	0.291
Trial Stopping**	Prob. of Early Stopping Trial due to Safety Rule	0.0107	0.2054	0.0107	0.0107
	Prob. of Early Stopping Trial due to Reaching K	0	0	0	0
	Prob. of Stopping Trial due to Reaching n	0.9893	0	0.9893	0.9893
Trial Sample Size	Average # of Patients Treated (s.d.)	29.7489 (2.44329)	11.5689 (4.36503)	29.7489 (2.44329)	29.7489 (2.44329)
Accuracy of Selected MTD	Mean Squared Error	0.012	0.017	0.012	0.011

* The row with orange background color indicates the true MTD.

** For further details concerning Trial Stopping Rule, please refer to section 1.2.2 in the User Manual

Figure 1.17: Simulation result tables in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

1.2.3.2 Simulation trial examples

Users can visualize how a trial is conducted by clicking a "Simulated Dose Escalation" button at the upper right corner of each simulation results table (Figure 1.17). The pop-up box (Figure 1.18) shows the dose escalation process of two simulated trials for each design.

A red or green dot indicates a patient with or without DLT, respectively. Dots within the same region of white or light blue background color indicate patients in the same cohort. The horizontal red line indicates the dose level selected as the MTD at the end of the trial. The absence of the red line indicates none of the dose levels is selected as the MTD.

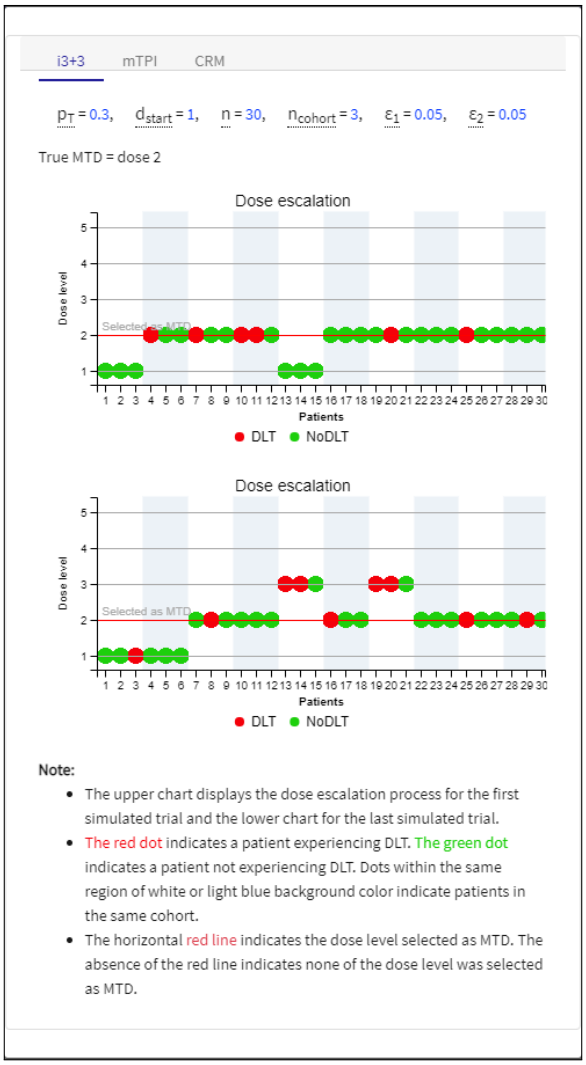


Figure 1.18: Simulation trial examples in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

1.2.3.3 Restore simulation setup

Users can restore the simulation settings from the simulation results by clicking the button at the upper right corner of each simulation results panel (yellow arrow in Figure 1.19). Upon clicking, the display will switch to the **Simulation Setup** page with the same simulation settings restored. This is useful to restore the old simulation settings for reproducible results.

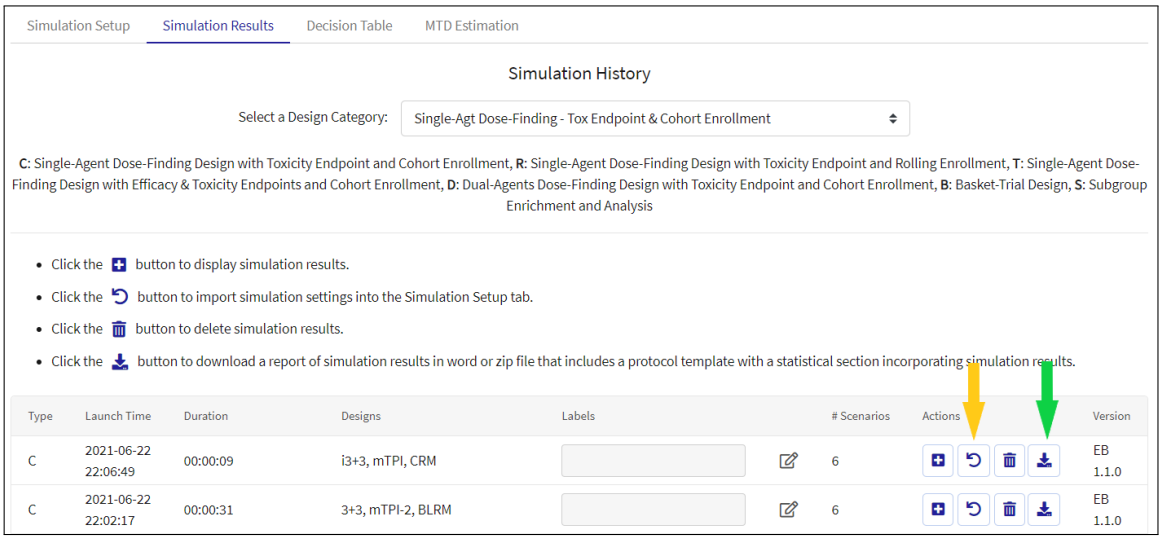


Figure 1.19: Restore simulation setup and download simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

1.2.3.4 Download simulation results

There is a button at the upper right corner of each simulation results panel (green arrow in Figure 1.19). Click it to download a word file, which includes four parts:

- Part A: Complete simulation results under the designs and scenarios users added in the Simulation Setup tab;
- Part B: Intelligent template(s) for the statistical section of i3+3 and/or mTPI-2 design in a trial protocol, if users select i3+3 and/or mTPI-2 in the Simulation Setup tab;
- Part C: Detailed technical descriptions of the designs users added in the Simulation Setup tab;
- Part D: Reference

Users may select the required parts and modify them tailored for their trials or contact us via email (`support@cytel.com`) for consulting services.

1.2.4 Decision Table

In the **Decision Table** tab, users can generate decision tables of five designs, i3+3, mTPI, mTPI-2, mCCD, 3+3, and BOIN designs, to guide the dose escalation/de-escalation during trial conduct. The CRM and BLRM designs do not provide decision tables before the trial is started. However, for both designs, East Bayes provides empirical decision tables after launching simulations (§1.2.3.1).

Manually type in the maximum number of patients at a dose (n), target toxicity probability (p_T) and two small fractions (ϵ_1 and ϵ_2) for decision table generation (Figure 1.21). Hover mouse over each parameter, and a description will be displayed explaining the meaning of the parameter. See detailed parameter descriptions in Table 1.3.

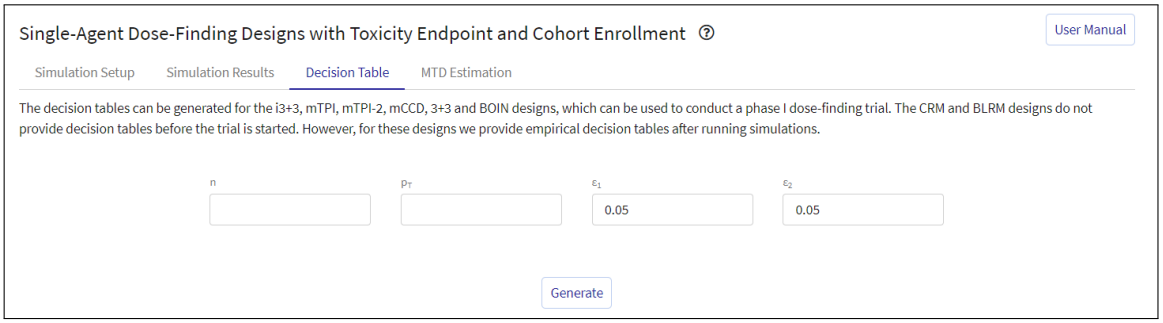


Figure 1.20: Input parameters in the **Decision Table** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Click the "Generate" button to generate five decision tables for five different designs at the same time (Figure 1.21). Users can click the tabs to switch between the tables for the i3+3, mTPI-2, mTPI, mCCD, 3+3, and BOIN designs.

Click the "Download Decision Table" button to save the decision table of the corresponding design in word (.docx).

For each decision table, the column represents the number of patients treated at a dose, which is mostly used for the current dose, the dose currently being used to treat patients in the trial, and the row represents the number of patients among those treated at that dose who have experienced dose-limiting toxicity (DLT) events. Note that these are the counts of patients, not DLT events. For example, column 3 and row 1 means that 3 patients have been treated at the current dose and 1 of them experiences DLT. Each cell in the decision table provides the dose-assignment decision based on the readouts from the corresponding row and column. For example, for column 3 and row 1, i.e., 1 out of 3 patients experiences DLTs, the decision is "S". The letters in the decision table represent

Table 1.3: Input arguments in the **Decision Table** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Notation	Parameters	Description
n	Number of patients at a dose	The maximum number of patients to be treated at a dose. Here, the upper limit is set at 30 since the number of patients that are enrolled at a dose in phase I clinical trial is typically small.
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T.
ϵ_1,ϵ_2 (except BOIN)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence/target interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose MTD. Default values for both are 0.05.
ϵ_1,ϵ_2 (BOIN)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the optimal interval and the target probability. Here, $\epsilon_1 = p_T - \lambda_1$, $\epsilon_2 = \lambda_2 - p_T$ where (λ_1, λ_2) is the optimal interval minimizing the probability of making an erroneous decision based on the initial equivalence interval (ϕ_1, ϕ_2). Default values for both are 0.05.

Figure 1.21: Decision tables generated in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

different dose-assignment decisions as shown below:

- "E" stands for escalating to the next higher dose,
- "S" stands for staying at the current dose,
- "D" stands for de-escalating to the previous lower dose,
- "DU" stands for de-escalating to the previous lower dose and marking the current dose and its higher doses as unacceptably toxic so that they will never be used again in the remainder of the trial.

The 3+3 decision table is fixed regardless of different trial parameters. For CRM (or BLRM), the decision table cannot be easily summarized since the dose-assignment decision under CRM (or BLRM) for a given outcome (say, 1 DLT out of 3 patients) and a given dose are random, depending on existing data in the entire trial including those at other doses. In other words, CRM (or BLRM) could stay, escalate or de-escalate when 1 out of 3 patients having DLT at a dose, which makes it impossible to provide a fixed decision table. Nevertheless, East Bayes provides empirical CRM (or BLRM) decision table in the simulation section when CRM (or BLRM) is implemented in simulation trials (§1.2.3.1).

1.2.5 MTD Estimation

In the **MTD Estimation** tab, users can estimate the MTD for i3+3, mTPI and mTPI-2 designs based on the isotonic regression through Pool Adjacent Violators Algorithm (PAVA), after the dose finding is completed and the DLT outcomes of all patients are collected.

Specify the target toxicity probability (p_T), and two small fractions to define the equivalence interval (ϵ_1 and ϵ_2) in the design. Select the number of doses (n_{dose}) from the dropdown box, then an editable table will be shown below on the page (Figure 1.22). Then manually type in the observed number of toxicities (DLTs) and the number of patients treated at each dose into the table and click the "Estimate" button to estimate the MTD. Finally, the estimated MTD is highlighted in blue background as shown in Figure 1.23.

Hover mouse over each parameter, and a description will be displayed explaining the meaning of the parameter. See detailed parameter descriptions in Table 1.4.

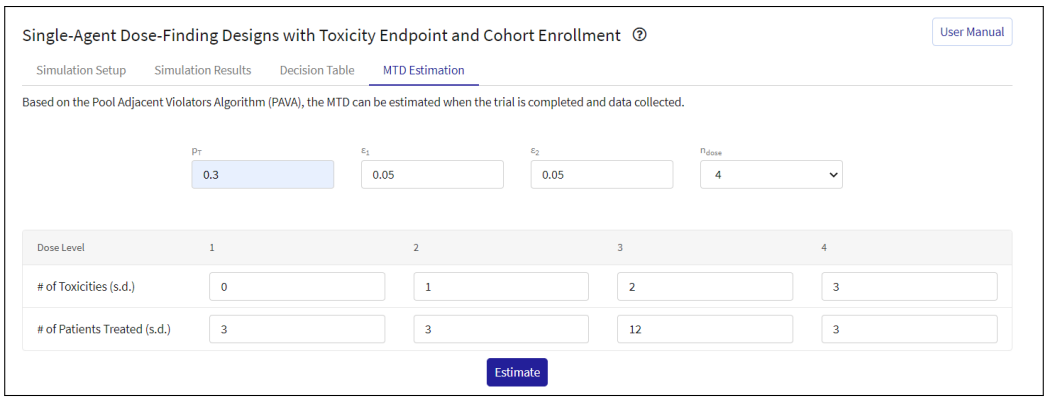


Figure 1.22: Input parameters in the **MTD Estimation** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

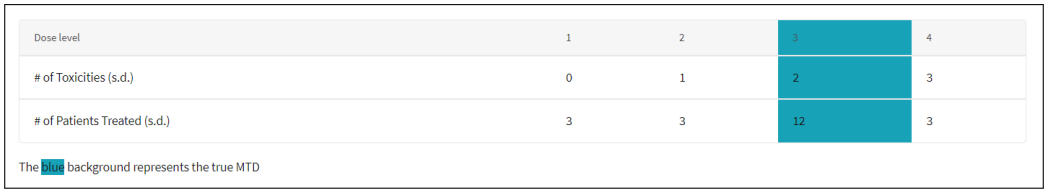


Figure 1.23: MTD estimation in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Table 1.4: Input parameters in the **MTD Estimation** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Notation	Parameters	Description
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T.
ϵ_1,ϵ_2	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence/target interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose MTD. Default values for both are 0.05.
n_{dose}	The number of doses	The number of candidate dose levels for investigation
# of DLTs	The number of patients with DLTs at each dose level	A non-negative integer number of patients with DLT at each dose level
# of patients	The number of patients treated at each dose level	A positive integer number of patients treated at each dose level, which should be no less than the # of DLTs

1.3 Statistical Methods Review

1.3.1 The 3+3 Design

The 3+3 design (Storer, 1989) is a rule-based design which starts by allocating the first cohort of patients to the starting dose (which is often the lowest dose level) and adaptively escalates/de-escalates to the next dose level based on observed number of dose limiting toxicities (DLTs).

1.3.1.1 Design Algorithm

In 3+3, a maximum of six patients are allowed to be treated at any dose level, and the MTD is defined as the highest dose for which one or fewer DLTs occurred in six patients. Its algorithm proceeds as follows:

0. Start the trial by treating three patients at a prespecified starting dose level.
1. Escalate to the next higher dose or de-escalate to the previous lower dose according to the following rules:
   (a) If 0 of 3 patients has a DLT, escalate to next higher dose and treat three patients.
   (b) If 2 or more of 3 patients have DLTs, de-escalate to previous lower dose and treat three patients.
   (c) If 1 of 3 patients has a DLT, treat three more patients at current dose level.
      i. If 1 of 6 has DLT, escalate to next higher dose and treat three patients if the next higher dose has not been tried; otherwise, declare it as the MTD and stop the trial.
      ii. If 2 or more of 6 have DLTs, de-escalate to previous lower dose level and treat three patients.
   (d) If the trial de-escalates to previous lower dose:
      i. If only 3 or less had been treated at the previous lower dose, treat three more patients at that dose.
      ii. If six have already been treated at the previous lower dose, stop the trial and declare the lower dose as the MTD.
2. Escalation never occurs to a dose at which two or more DLTs have already occurred.
3. If de-escalation occurs at the lowest dose, the trial is stopped.
4. Repeat steps 1-3 until either the MTD is identified or the trial is stopped for excessive toxicity.

The above algorithm can be summarized in Figure 1.24 (Yang et al., 2015).

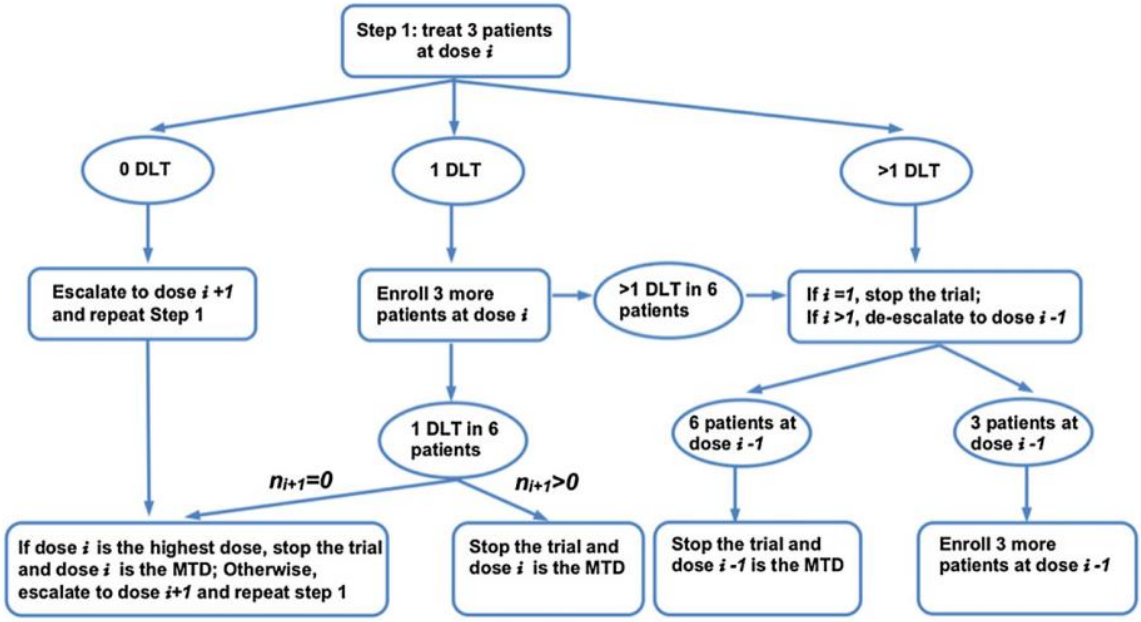


Figure 1.24: Schema of the 3+3 design.

1.3.2 The Continuous Reassessment Method (CRM)

CRM is a Bayesian adaptive model-based design introduced in O'Quigley et al. (1990). It assumes a parametric dose-response model in which the probability of toxicity monotonically increases with dose. The estimated dose-response curve is updated after each patient's toxicity data is observed, and the dose closest to MTD is obtained from the updated dose toxicity curve. In the original CRM (O'Quigley et al., 1990), it is possible to escalate by more than one dose level, which may result in escalation to fairly high doses quite early. Goodman et al. (1995) proposed several practical rules for the original CRM to reduce the risk.

1.3.2.1 Probability Model

Dose-response curve: Denote the dose levels as x_d for $d = 1, ..., D$, and the binary indicator of DLT for the jth patient as Y_j for $j = 1, ..., n$. Let t_j be the dose for patient j, and let $p_d = Pr(Y_j = 1|t_j = x_d)$ be the toxicity probability of dose d. Consider a dose-response function $p_d = \psi(x_d, \theta)$ representing the relationship between p_d and x_d, which includes a single parameter θ. Popular choice of ψ includes the power model, one-parameter logistic model, and hyperbolic tangent model (Cheung, 2011). East Bayes uses a simple one-parameter power model:

$$p_d = \psi(p_{0,d}, \theta) = p_{0,d}^{\exp(\theta)},$$

where $(p_{0,1}, p_{0,2}, \ldots, p_{0,D})$ are pre-specified prior toxicity probabilities ('skeletons'), which monotonically increases with d. The skeletons reflect the initial guess of DLT probabilities.

Prior specification: Let $g(\theta)$ be the prior distribution for θ, which reflects our knowledge of the dose toxicity relationship before the trial begins. In East Bayes, we use the normal density $N(0, 1.16^2)$ by default (Lee and Cheung, 2011). Other choices can be gamma or exponential density.

Estimate the probability of toxicity: Denote the accumulated toxicity data $data \equiv \{(y_d, n_d) : d = 1, 2, \ldots, D\}$, where n_d and y_d are the total number of patients treated at dose d and the corresponding number of patients having DLTs, respectively. Estimate the probability of toxicity p_d for dose level d by

$$\hat{p_d} = \psi(p_{0,d}, E(\theta|data)), \text{ where } E(\theta|data) = \int_{-\infty}^{\infty} \theta f(\theta|data) d\theta, \tag{1.1}$$

for $d = 1, ..., D$, where $f(\theta|data)$ is the posterior of θ given by

$$f(\theta|data) \propto \prod_{d=1}^{D} \psi(p_{0,d}, \theta)^{y_d} \left(1 - \psi(p_{0,d}, \theta)\right)^{n_d - y_d} g(\theta).$$

Calibration of the 'skeleton' values: Lee and Cheung (2011) proposed a fast and systematic approach for selecting the skeleton based on indifference intervals for the MTD. The approach is imbedded in East Bayes by default, and users only need to specify the half-width (δ) of the indifference interval manually to estimate the skeleton.

Specifically, assume $\Theta = [b_1, b_{D+1}]$ is the parameter space (i.e. $\theta \in \Theta$) and $H_1 = [b_1, b_2)$, $H_d = [b_d, b_{d+1})$ for $d = 2, \ldots, D-1$ and $H_D = [b_D, b_{D+1})$ where b_d is the solution for $\psi(p_{0,d-1}, b_d) + \psi(p_{0,d}, b_d) = 2p_T$ for $d = 2, \ldots, D$. Based on Lee and Cheung (2011), define the half width of the indifference interval for the MTD (d) as

$$\delta_d = \frac{\psi(p_{0,d+1}, b_{d+1}) - \psi(p_{0,d-1}, b_d)}{2}, d = 2, \ldots, D-1.$$

By specifying a common half-width indifference interval for all dose levels, that is $\delta_d = \delta$, the skeletons $p_{0,1}, \ldots, p_{0,D}$ can be obtained recursively. Given a starting dose ν, a target p_T and a prior mean of $\theta = 0$, $p_{0,\nu}$ can be obtained via backward substitution, i.e. $p_T = \psi(p_{0,\nu}, 0) = p_{0,\nu}$. The remaining skeletons can be obtained by solving the following equations:

$$\begin{cases} \psi(p_{0,d-1}, b_d) + \psi(p_{0,d}, b_d) = 2p_T \\ \psi(p_{0,d-1}, b_d) = p_T - \delta \end{cases} \quad \text{for } d \leq \nu;$$

$$\begin{cases} \psi(p_{0,d}, b_{d+1}) + \psi(p_{0,d+1}, b_{d+1}) = 2p_T \\ \psi(p_{0,d+1}, b_{d+1}) = p_T + \delta \end{cases} \quad \text{for } d > \nu.$$

East Bayes takes $\nu = [D/2]$ as the prior guess of MTD by default.

1.3.2.2 Design Algorithm

Dose Finding Rules: Assume patients are enrolled in cohorts. After each cohort of patients completes the DLT follow-up period, the dose to be assigned is the one that has the posterior mean probability of toxicity closest to the target p_T. In other words, the next cohort of patients is assigned to dose $d^* = \text{argmin}_d |\hat{p}_d - p_T|$ where $\hat{p}_d$ is the posterior mean of toxicity probability.

Additional safety rules: In East Bayes, three additional rules are applied for safety.

[*Rule 1: Dose Exclusion*] If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses will be excluded and never be used again in the remainder of the trial to avoid any other patients receiving treatment at those doses. An exception of Rule 1 is that when there is only 1 DLT observed at a dose, the rule is not enforced.

[*Rule 2: Early Stop*] If the current dose is the lowest dose (first dose) and is considered excessively toxic according to Rule 1, early stop the trial for safety.

[*Rule 3: No-Skipping Escalation*] Dose-escalation cannot happen by more than one level. That is, suppose the current dose is d. If the next dose d^* satisfies $(d^* - d) > 1$, escalate to dose $(d + 1)$ instead.

[*Rule 4: Coherence*] Dose-escalation cannot happen when the empirical toxicity probability of the new cohort of patients is larger than the target p_T.

Here in Rules 1 and 2, $Prob\{p_d > p_T \mid data\}$ is a function of the cumulative distribution of $beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$. In East Bayes, $\alpha_0 = \beta_0 = 1$ is used. Lastly, no escalation is permitted if the empirical rate of DLT for the most recent cohort is higher than p_T, according to the coherence principle (Cheung, 2011).

Trial termination: The trial proceeds until any of the following stopping criteria is met:

1. If the prespecified maximum total sample size n is reached;
2. If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;
3. *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by K $(K < n)$:*
   - If the CRM decision is "S", to stay at the current dose, and the current dose level has enrolled K patients;
   - If the CRM decision is "E", to escalate to the next higher dose, and the next higher dose has enrolled K patients;
   - If the CRM decision is "D", to de-escalate to the previous lower dose, and the previous lower dose has enrolled K patients.

MTD selection: Once all the enrolled patients complete the DLT observation and the trial is not stopped early, the dose level d^{**} is selected as the MTD with the smallest difference of $|\hat{p}_d - p_T|$ among all safe doses d, where $\hat{p_d}$ is the posterior mean of toxicity probability for dose d. For CRM, the MTD can be an untried doses as long as it does not exceed the highest tried dose.

1.3.3 The Bayesian Logistic Regression Method (BLRM)

The Bayesian Logistic Regression Method (BLRM) is a model-based design proposed by Neuenschwander et al. (2008). BLRM improves upon CRM in that it offers a more flexible representation of the dose toxicity relationship and accounts for the uncertainty associated with DLT probability point estimation during dose finding. In BLRM, one classifies the posterior probability of toxicity into four categories: under-dosing, targeted, excessive, and unacceptable toxicity, and calculates the posterior probability of DLT rate falling into four corresponding intervals at each dose. The final dose recommendation aims at maximizing the probability of targeted toxicity while controlling the probability of excessive or unacceptable toxicity at a pre-specified threshold. Besides, BLRM can also accommodate different conservatism for dose finding behavior through specification of a loss function.

1.3.3.1 Probability Model

For a set of candidate doses $d \in \{1, ..., D\}$, where D is the number of doses, BLRM assumes a two-parameter logistic model between dose levels x_d and the probability of DLT p_d, which is given by

$$\text{logit}(p_d) = \log(\alpha) + \beta \log(x_d / x_{d^*}), \quad \alpha > 0, \beta > 0$$

where x_{d^*} is the reference dose, determined so that $\log(\alpha)$ is the log-odds of toxicity when $x_d = x_{d^*}$. East Bayes uses a default set of doses, $x_d = 5 \times d$, and a default reference dose level x_{d^*}, the ceiling of $(D+1)/2$. As a result, users do not need to input the candidate doses and reference doses manually on East Bayes. However, we offer customized service allowing input of these values upon users' requests.

1.3.3.2 Dosing Intervals and Selection

Probability intervals: Suppose the target probability of DLT is p_T and BLRM divides the probability interval $(0, 1)$ into four categories: under-dosing $p_d \in (c_0 = 0, c_1]$, target toxicity $p_d \in (c_1, c_2]$, excessive toxicity $p_d \in (c_2, c_3]$ and unacceptable toxicity $p_d \in (c_3, c_4 = 1)$. After each patient cohort is enrolled and toxicity data are observed, the posterior distribution of p_d is used to calculate the four probabilities of under-dosing, targeted, excessive and unacceptable toxicity. Based on the four probabilities, the next dose will be selected depending on one of the following two methods: minimize the Bayes risk or maximize the distance to the targeted toxicity probability subject to escalation with overdose control (EWOC).

Method 1: Minimize the Bayes risk A formal loss function is introduced to quantify the penalty of ending up in each of the four aforementioned intervals:

$$L(\boldsymbol{\theta}, x_d) = \begin{cases} \ell_1 & \text{if} \quad p_d \in (0, c_1] \\ \ell_2 & \text{if} \quad p_d \in (c_1, c_2] \\ \ell_3 & \text{if} \quad p_d \in (c_2, c_3] \\ \ell_4 & \text{if} \quad p_d \in (c_3, 1) \end{cases}$$

Using the above loss function, one can calculate the Bayes risk $= \ell_1 \times Prob\{p_d \in (0, c_1] \mid Data\} + \ell_2 \times Prob\{p_d \in (c_1, c_2] \mid Data\} + \ell_3 \times Prob\{p_d \in (c_2, c_3] \mid Data\} + \ell_4 \times Prob\{p_d \in (c_3, 1) \mid Data\}$ and the dose minimizing the Bayes risk is selected as the next dose. In Neuenschwander et al. (2008), three different loss functions are compared in terms of dose-escalation behavior: (i) aggressive ('1-0-1-1'), (ii) conservative ('1-0-1-2'), and (iii) very conservative ('1-0-2-4').

Depending on the compound and/or indication under study, the probability interval specification and loss function should be tailored to the specific clinical setting. However, the specification of loss function may be difficult and may complicate the interactions with clinical teams, thus the dose recommendation approach below is often used instead of the actual Bayesian decision analytic framework.

Method 2: Maximize the distance to the target toxicity probability subject to EWOC Babb et al. (1998) proposed to select the dose for each cohort patients as the one that maximizes the probability of targeted toxicity, i.e., $Prob\{p_d \in (c_1, c_2] \mid Data\}$ subject to the constraint that the probability of overdosing (i.e., excessive and unacceptable toxicity) does not exceed a predefined threshold p_{EWOC}. That is, choose the dose level subject to the constraint $Prob\{p_d \in (c_2, 1) \mid Data\} \leq p_{EWOC}$.

East Bayes adopts the second method for dose recommendation by default, except that the targeted interval is defined as $(c_1 = p_T - \epsilon_1, c_2 = p_T + \epsilon_2]$ to make it consistent with settings in mTPI and mTPI-2 designs.

1.3.3.3 Posterior and Prior

Prior Specification: Model parameters $\boldsymbol{\theta} = (\alpha, \beta)'$ follow a multivariate log-normal prior $\pi(\boldsymbol{\theta})$, given by

$$\log(\boldsymbol{\theta}) = \begin{pmatrix} \log(\alpha) \\ \log(\beta) \end{pmatrix} \sim MVN \left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma \right\}, \text{where } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where "MVN" stands for a multivariate normal distribution. Let $\boldsymbol{\eta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ be the hyperparameter set of the model. In East Bayes we use the ***quantile-based non-informative prior*** calculator proposed by Neuenschwander et al. (2008) to obtain the values of $\boldsymbol{\eta}$.

The hyperparameter calculation process is based on a set of quantiles for the probabilities of toxicity that are derived from minimally informative unimodal beta distributions. Here, a beta distribution $X \sim beta(a, b)$ is defined as a minimally informative unimodal distribution, given a prespecified quantile $q(p)$ of the prior distribution, if (i) $Prob\{X < q(p)\} = p$, (ii) $a \geq 1$ or $b \geq 1$ (or both), and (iii) a+b minimal. For a given prior quantile $q(p)$, the parameters and the quantiles of a minimally informative unimodal beta distribution can be easily obtained. If $q(p) > p$, $beta(a, 1)$ is minimally informative unimodal if $a = \ln(p)/\ln\{q(p)\}$. Alternatively, if $q(p) < p$, $beta(1, b)$ is minimally informative unimodal if $b = \ln(1-p)/\ln\{1-q(p)\}$. Specifically, the following steps are used for this prior distribution specification:

1. Obtain the set of prior quantiles Q for the distribution of p_d. In East Bayes, we summarize prior information at a given dose using the median, 2.5%-th and 97.5%-th percentiles, denoted by $q_d = \{q_d(2.5\%), q_d(50\%), q_d(97.5\%)\}$.
   (a) For the lowest dose $d = 1$, the prior probability of exceeding a certain threshold $q_1(\phi_1)$ is ϕ_1. In East Bayes, the following default values will be used: $Prob\{p_1 > 0.4\} = 5\%$, i.e. for the lowest dose the probability of excessive toxicity will be set to be 5 percent.
   (b) For the highest dose $d = D$, the prior probability of falling below a certain threshold $q_D(\phi_2)$ is ϕ_2. In East Bayes, the following default values will be used: $Prob\{p_D \leq 0.2\} = 0.05$, i.e. for the highest dose the probability of under-dosing will be set to be 5 percent.
   (c) Assuming a minimally informative unimodal beta distribution in (a) and (b) leads to prior medians for the probabilities of toxicity p_1 and p_D, say $\mu_1 = q_1(50\%)$ and $\mu_D = q_D(50\%)$.
   (d) Prior medians $\mu_1, \ldots, \mu_D$ are assumed to be linear in log-dose on the logit scale. This decides the minimally informative unimodal beta distributions for each dose d.
   (e) For each dose d, two quantiles (2.5% and 97.5%) is derived using minimally informative unimodal beta distributions with prior medians equal to μ_d.
   (f) Therefore, a set of $D \times 3$ quantiles are obtained, denoted by $Q = \{q_{dk}\}$ with $q_{dk} = q_d(\pi_k)$, $d = 1, 2, \ldots, D$, $k = 1, 2, 3$, where $\pi_1 = 2.5\%$, $\pi_2 = 50\%$ and $\pi_3 = 97.5\%$.
2. For the two-parameter logistic model the above constructed quantiles Q are then compared with the quantiles Q' coming from the bivariate normal prior distribution. We will minimize

the following criteria:

$$C(Q, Q') = \max_{d,k} |q_{dk} - q'_{dk}|, d = 1, 2, \ldots, D, k = 1, 2, 3.$$

The minimization of $C(Q, Q')$ leads to the optimal parameter for the prior distribution $\boldsymbol{\eta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, which can be achieved by a stochastic optimization using a Metropolis algorithm (Robert and Casella, 2013).

Posterior Calculations: The dose selection process described above requires the calculation of the posterior probability $Prob\{p_d \in (c_{i-1}, c_i] \mid Data\}$, for $i = 1, 2, 3, 4$, which is calculated with respect to

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x}) \propto \frac{e^{\sum_{d=1}^{D} y_d(\log(\alpha)+\beta \log(x_d/x_{d^*}))}}{\prod_{d=1}^{D}(1 + e^{\log(\alpha)+\beta \log(x_d/x_{d^*})})^{n_d}} \times \pi_0(\boldsymbol{\theta}).$$

where $\boldsymbol{n} = \{n_1, \ldots, n_D\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_D\}$ are observed toxicity data, n_d and y_d are the number of patients treated and having DLTs at the dose d, respectively. Let $Data \equiv (\boldsymbol{n}, \boldsymbol{y})$, and $\boldsymbol{x} = \{x_1, \ldots, x_D\}$ are candidate dose levels. Using Markov chain Monte Carlo (MCMC) simulation, the posterior inference is made based on the posterior samples drawn for (α, β) via Metropolis-Hastings algorithm.

1.3.3.4 Design Algorithm

Dose Finding Rules: Assume patients are enrolled in cohorts. After each cohort of patients completes the DLT evaluation period, the dose to be assigned by BLRM is the one that has the largest posterior probability being at the targeted interval, i.e., $Prob\{p_d \in (p_T - \epsilon_1, p_T + \epsilon_2] \mid Data\}$ subject to the constraint that the probability of overdosing does not exceed a predefined threshold p_{EWOC}, i.e., $Prob\{p_d \in (p_T + \epsilon_2, 1) \mid Data\} \leq p_{EWOC}$.

Additional safety rules: In East Bayes, three additional rules are also applied for safety.

- [***Rule 1: Dose Exclusion***] If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses will be excluded and never used again in the remainder of the trial.
- [***Rule 2: Early Stop***] If the current dose is the lowest dose (first dose) and is considered excessively toxic, i.e., $Prob\{p_1 > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, stop the trial early for safety.

  Besides, if all doses violate the EWOC rule, the trial will also be terminated early with no MTD selected before the prespecified maximum sample size is reached.

– [***Rule 3: No-Skipping Escalation***] Dose escalation cannot increase by more than one level, although dose de-escalation can (Goodman et al., 1995). That is, suppose the current dose is dose level d. If the next dose d^* satisfies $(d^* - d) > 1$, escalate to dose $(d + 1)$ instead.

Here in Rules 1 and 2, $Prob\{p_d > p_T \mid Data\}$ is a function of the cumulative distribution of $beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$, and $\alpha_0 = \beta_0 = 1$ is used in East Bayes by default, where y_d and n_d are the number of patients treated and the number of DLTs at the dose d.

Trial termination: The trial proceeds until any of the following stopping criteria is met:

1. If the prespecified maximum total sample size is reached;
2. If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is stopped early and the MTD cannot be determined;
3. *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by* K $(K < n)$*:*
   - *If the BLRM decision is "S", to stay at the current dose, and the current dose has enrolled* K *patients;*
   - *If the BLRM decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled* K *patients;*
   - *If the BLRM decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled* K *patients.*

MTD selection: Once all the enrolled patients complete the DLT observation and the trial is not stopped early, the dose level d^{**} is selected as the MTD which maximizes the posterior probability of toxicity rate falling into the targeted interval i.e., $d^{**} = \text{argmax}_{d=1,\ldots,D} Prob\{p_d \in (p_T - \epsilon_1, p_T + \epsilon_2] \mid Data\}$ among all doses that are used and do not violate the EWOC rule.

1.3.4 The Modified Toxicity Probability Interval (mTPI) Design

This section describes the modified toxicity probability interval (mTPI) design proposed by Ji et al. (2010). The mTPI design is an extension of the toxicity probability interval (TPI) method (Ji et al., 2007), which uses a simple Bayesian hierarchical model and a decision framework for dose finding.

The mTPI design starts from the specification of three intervals: the under-dosing interval $(0, p_T - \epsilon_1)$, the proper dosing interval $(p_T - \epsilon_1, p_T + \epsilon_2)$ and the over-dosing interval $(p_T + \epsilon_2, 1)$. Unlike the CRM and BLRM, which assumes a parametric curve to model the dose-toxicity response, the mTPI uses a simple beta-binomial model to estimate the toxicity probability and makes the decisions of dose escalation and de-escalation based on the unit probability mass (UPM) of the three intervals. At the end, mTPI selects the dose of which the isotonic transformed toxicity probability is the closest to the target p_T as the MTD.

1.3.4.1 Probability Model

Consider a phase I trial with D candidate doses for escalation. Let $p_1, \ldots, p_D$ denote the true toxicity probabilities for doses $d = 1, \ldots, D$. The observed data include n_d, the number of patients treated at dose d, and y_d, the number of patients experiencing a toxicity. Let $Data = \{(y_d, n_d); d = 1, 2, \ldots, D\}$.

The mTPI design employs a simple beta-binomial hierarchical model as follow:

$$y_d \mid n_d, p_d \sim binomial(n_d, p_d)$$
$$p_d \sim beta(\alpha, \beta)$$

The posterior distribution of p_d is given by

$$p_d \mid y_d, n_d \sim beta(\alpha + y_d, \beta + n_d - y_d). \tag{1.2}$$

In East Bayes, we adopt the prior $beta(1, 1)$ for p_d, because it would lead to slightly conservative posterior inference as the prior mean is 0.5, which is usually above p_T.

1.3.4.2 Dosing Intervals

The under-dosing interval is defined as $(0, p_T - \epsilon_1)$, the over-dosing interval as $(p_T + \epsilon_2, 1)$, and the equivalence interval as $(p_T - \epsilon_1, p_T + \epsilon_2)$ for proper dosing, where ϵ_1 and ϵ_2 are small fractions, such as 0.05, to account for the uncertainty around the true target toxicity p_T. The three dosing intervals are associated with three different dose-finding decisions. The under-dosing interval corresponds

to a dose **escalation (E)**, the over-dosing interval corresponds to a dose **de-escalation (D)**, and the equivalence interval corresponds to **staying (S)** at the current dose.

1.3.4.3 Dose Finding Rules

Given an interval and a probability distribution, define the UPM of that interval as the probability of the interval divided by the length of the interval. The mTPI design calculates the UPMs for the three dosing intervals, and the one with the largest UPM implies the corresponding dose-finding decision. That decision provides the dose level to be used for future patients. More specifically, given the current dose level d, the mTPI conducts the following steps for dose assignment for the future patients.

1. Compute the UPM for each of the three toxicity probability intervals as follows:

$$\begin{aligned}
\text{UPM(D)}_d &= \frac{Prob\{p_d \in (p_T + \epsilon_2, 1) \mid Data\}}{1 - (p_T + \epsilon_2)}, \\
\text{UPM(S)}_d &= \frac{Prob\{p_d \in (p_T - \epsilon_1, p_T + \epsilon_2) \mid Data\}}{\epsilon_1 + \epsilon_2}, \\
\text{UPM(E)}_d &= \frac{Prob\{p_d \in (0, p_T - \epsilon_1) \mid Data\}}{p_T - \epsilon_1}.
\end{aligned}$$

   Here, the numerator in UPM calculation, $Prob\{\cdot\}$ is calculated according to the beta posterior distribution in (2.1).

2. Select one of the following actions: "E", "S" or "D" corresponding to the highest UPM of each toxicity interval. That is, the dose decision is given by

$$M^* = \underset{M \in \{D,S,E\}}{\text{argmax}} \; UPM(M)_d.$$

   In other words,
   - Escalate to dose $(d+1)$, if $\text{UPM(E)}_d > \text{UPM(S)}_d$ and $\text{UPM(E)}_d > \text{UPM(D)}_d$,
   - Stay at dose d, if $\text{UPM(S)}_d \geq \text{UPM(E)}_d$ and $\text{UPM(S)}_d > \text{UPM(D)}_d$,
   - De-escalate to dose $(d-1)$, if $\text{UPM(D)}_d \geq \text{UPM(E)}_d$ and $\text{UPM(D)}_d \geq \text{UPM(S)}_d$.

   For example, if the under-dosing interval has the largest UPM, decision $M^* = E$ will be executed and the next cohort of patients will be treated at the next higher dose level $(d+1)$.

Ji et al. (2010) and Guo et al. (2017b) have shown that the above UPM-based decision rules correspond to the Bayes' rule under a formal Bayesian decision theoretic framework, if we use the uniform prior for p_d.

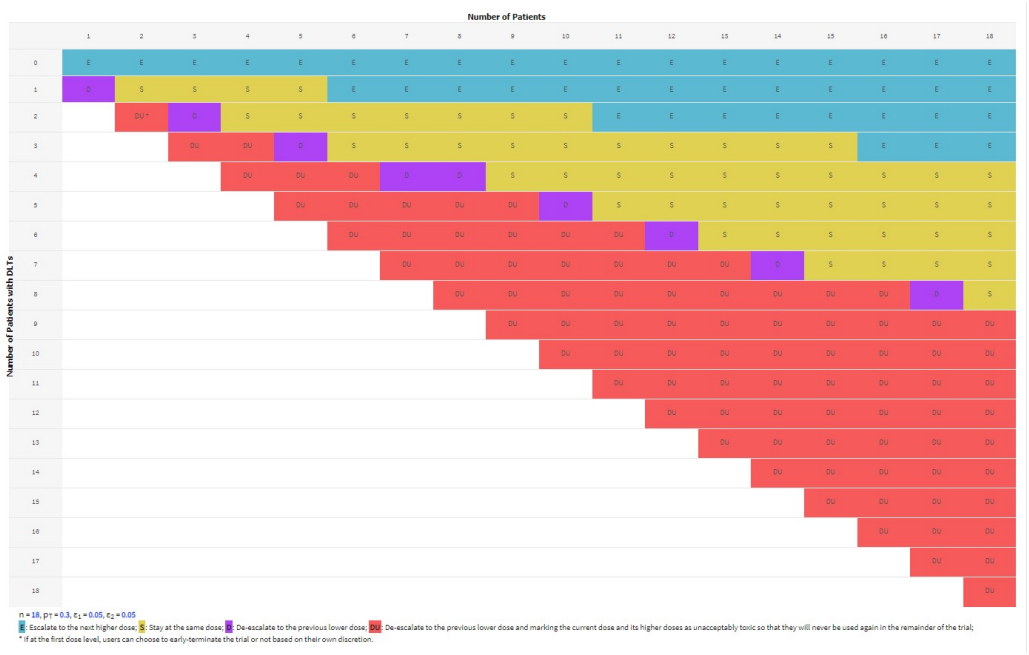


Figure 1.25: An example of mTPI decision table generated via East Bayes. The target toxicity probability $p_T = 0.3$, and the equivalence interval (EI) is $(0.25, 0.35)$ for up to 18 subjects. Each column represents (n) number of subjects treated at the current dose and each row represents (y) number of subjects with DLTs at the dose level. Each cell in the table provides the dose-finding decision based on the readouts from the corresponding row (y) and column (n). The letters in the decision table represent different dose-assignment decisions.

The mTPI design pre-calculates all the dose-finding decisions in advance, allowing investigators to examine the decisions before the trial starts. See Figure 1.25 for an example. Therefore, mTPI exhibits the same simplicity and transparency as rule-based methods like 3+3. The decision table can be generated via East Bayes under module **Decision & MTD**.

1.3.4.4 Design Algorithm

The mTPI algorithm proceeds as follows:

1. At each dose level, treat a cohort of patients, with the first cohort at a prespecified starting dose.
2. After all patients in each cohort complete the DLT evaluation, the dose-finding decision for the next cohort will be determined according to the following rules:

(a) Compute the posterior probability of excessive toxicity at the current tried dose, i.e., $Prob\{p_d > p_T \mid Data\}$ which is a function of the cumulative distribution of $beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$, similar in (2.1). In East Bayes, $\alpha_0 = \beta_0 = 1$ is used.

i. ***[Additional Safety Rule 1: Dose Exclusion]*** If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses will be excluded and never be used again in the remainder of the trial to avoid any other patients receiving treatment at those doses. An exception of Rule 1 is that when there is only 1 DLT observed at a dose, the rule is not enforced.

Also, at that time, the decision is "D", to de-escalate to previous lower dose.

ii. ***[Additional Safety Rule 2: Early Stop]*** If the current dose is the lowest dose and considered excessively toxic according to Rule 1 in i, early stop the trial for safety.

(b) If the trial is not stopped early, assign the next cohort of patients to the dose according to the decision table or the procedures in Section 1.3.4.3.

(c) If the dose-assignment decision is "E" but the next higher dose has been excluded by Rule 1, continue to enroll the next cohort at the current dose instead.

(d) If the dose-assignment decision is "E" and the current dose is the highest dose, continue to enroll the next cohort at the current dose instead.

(e) If the dose-assignment is "D" and the current dose is the lowest dose, continue to enroll the next cohort at the current dose instead.

3. Repeat steps 1-2, stop the trial when any of the following conditions is satisfied:

(a) If the prespecified maximum total sample size is reached;

(b) If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;

(c) *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by* $K (K < n)$*:*

- *If the mTPI decision is "S", to stay at the current dose, and the current dose has enrolled* K *patients;*
- *If the mTPI decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled* K *patients;*
- *If the mTPI decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled* K *patients.*

1.3.4.5 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped early, an isotonic regression (Ji et al., 2010; Ivanova, Anastasia and Wang, Kai, 2006) is used to select the MTD based on the observed DLT data from all the dose levels. Follow the steps below:

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. Here in East Bayes, an independent prior $beta(0.005, 0.005)$ is used to compute the posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior mean by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA) (Robertson, 1988), the estimated DLT probabilities satisfying the order constraint is obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.
2. Among all the tried doses for which $Prob\{p_d > p_T \mid Data\} < \xi$ and $\hat{p}_d \leq p_T + \epsilon_2$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d\, |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference),
   (a) If there is at least one dose lower the target p_T among all the tied doses, choose the highest dose among those as the estimated MTD;
   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

1.3.5 The Modified Toxicity Probability Interval-2 (mTPI-2) Design

This section describes the modified toxicity probability interval-2 (mTPI-2) design proposed by Guo et al. (2017b). The mTPI-2 improves the mTPI design by blunting the Ockhams razor that leads to some statistically sound but ethical challenging decisions in mTPI. For example, when $p_T = 0.3$ and 3 out of 6 patients experience DLTs at a dose, the mTPI decision is "S", stay at the current dose and enroll more patients. Such a decision may be considered too aggressive. To this end, mTPI-2 constructs a series of dosing interval with equal length to guide the dose escalation and de-escalation, mitigating the effect of interval length in the mTPI design. Otherwise, the model, the design algorithm, and the MTD selection are the same as those in mTPI in Section 1.3.4.

1.3.5.1 The effect of Ockham's razor in mTPI

The mTPI design has been shown to be simple, transparent, and superior to the 3+3 design (Ji and Wang, 2013; Yang et al., 2015). However, some decisions in mTPI may be debated in practice. For example, when the target toxicity probability $p_T = 0.3$, and 3 out of 6 patients treated at a dose experience DLT events, mTPI would suggest "S", stay at the current dose and enroll more patients to be treated at the dose. Since the empirical rate is $3/6$, or 50%, oftentimes one would argue that the more desirable decision should be D, de-escalate to the next lower dose level. Another case is when p_T=0.3 and 2 out of 9 patients experience DLT events at a dose, mTPI would suggest S as well. Investigators could argue that the decision should be E, escalation since the empirical rate is 2/9, or 22%. Guo et al. (2017b) noted that these decisions are due to the Ockham's razor (Jefferys and Berger, 1992), which is a Bayesian principle that prefers parsimonious models in model selection. The mTPI design treats the three intervals as three models, and penalizes models based on the model size which is the length of each interval. Figure 1.26 gives an example of the effect of the Ockham's razor in mTPI. Statistically speaking, there is nothing wrong with the Ockham's razor in mTPI as the Bayesian inference takes into account the model complexity when choosing the optimal decision. However, for human clinical trials patient safety often outweighs statistical optimality. To this end, mTPI-2 modifies the decision theoretic framework and blunt the Ockham's razor, which leads to practically desirable decision rules.

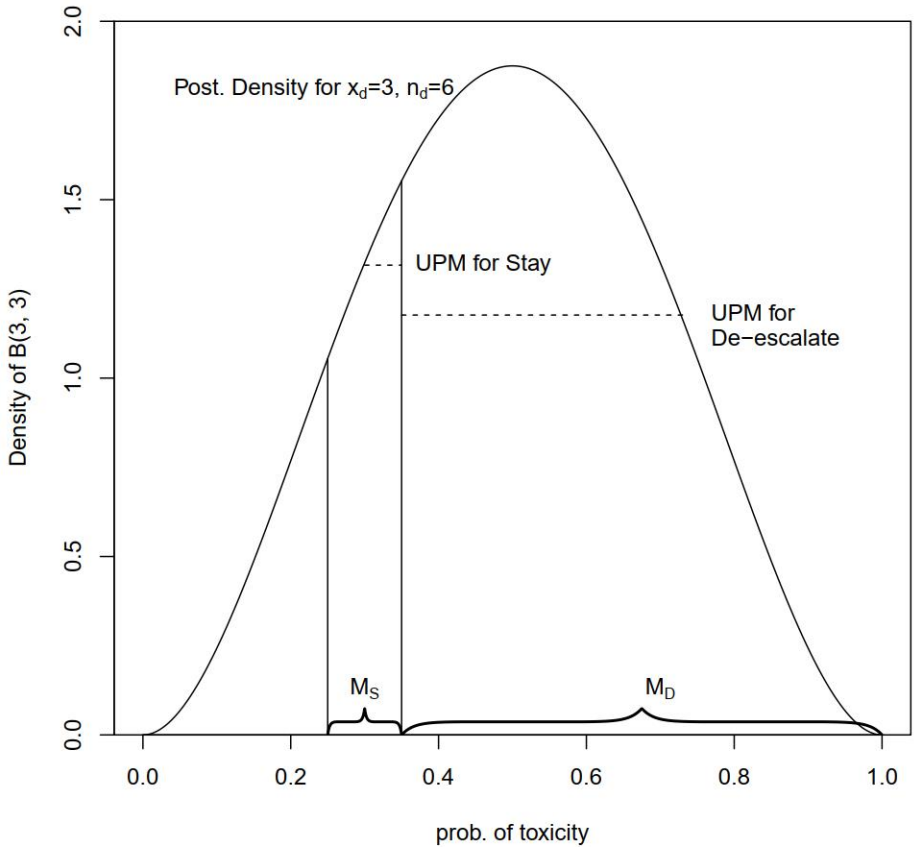


Figure 1.26: An example demonstrating the effect of the Ockham's razor in mTPI. Shown is the posterior density of p_d when $x_d = 3$ and $n_d = 6$. Even though the shape of the density suggests that dose d might be above the MTD, e.g., the posterior mode is to the right of the equivalence interval (shown as the two vertical bars), the UPM for decision "S" (stay) is still larger than that of the UPM for decision D (de-escalate). Therefore, mTPI would still choose to "Stay" despite that the shape of the posterior density of p_d indicates otherwise. This is due to the larger size (longer length) of the interval M_D than M_S and the Ockham's razor, which prefers the smaller model M_S.

1.3.5.2 Dose Finding Rules

The basic idea in mTPI-2 is to divide the unit interval $(0, 1)$ into subintervals with equal length, given by $(\epsilon_1 + \epsilon_2)$. This results in multiple intervals with the same length except for the boundary intervals, see Figure 1.27. For clarify, denote *EI* the equivalence interval $(p_T - \epsilon_1, p_T + \epsilon_2)$, and *LI* a set of intervals below *EI*, and *HI* a set of intervals above *EI*. For example, when $p_T = 0.3$ and $\epsilon_1 = \epsilon_2 = 0.05$, the *EI* $= (0.25, 0.35)$, the *LI* intervals are

$$LI = \{M_1^{LI} = (0.15, 0.25), M_2^{LI} = (0.05, 0.15), M_3^{LI} = (0, 0.05)\},$$

and the *HI* intervals are

$$\begin{aligned} HI = \{&M_1^{HI} = (0.35, 0.45), M_2^{HI} = (0.45, 0.55), M_3^{HI} = (0.55, 0.65), M_4^{HI} = (0.65, 0.75), \\ &M_5^{HI} = (0.75, 0.85), M_6^{HI} = (0.85, 0.95), M_7^{HI} = (0.95, 1)\}. \end{aligned}$$

Other than the boundaries $(0, 0.05)$ and $(0.95, 1)$, all the intervals have the same length. The boundaries do not affect the decision making since they are clearly associated with "E" and "D" decisions, respectively. See Guo et al. (2017b) for details.

The dose finding rules are given as follows:

- If the equivalence interval $M^{EI} = (p_T - \epsilon_1, p_T + \epsilon_2)$ has the largest UPM, it is selected as the winning interval and dose-assignment decision of mTPI-2 is "S", to stay at the current dose.
- If any interval M_j^{LI} in *LI* has the largest UPM, it is selected as the winning interval and dose-assignment decision of mTPI-2 is "E", to escalate to the next higher dose.
- If any interval M_k^{HI} in *HI* has the largest UPM, it is selected as the winning interval and dose-assignment decision of mTPI-2 is "D", to de-escalate to the previous lower dose.

In Figure 1.27, for the same posterior density corresponding to $y_d = 3$ and $n_d = 6$, interval M_2^{HI} exhibits the largest UPM and therefore the decision is now "D". Note that the same decision theoretic framework as mTPI is in place except that now there are multiple intervals corresponding to "D" or "E", and the intervals all have the same length except the boundary ones, thereby blunting the Ockham's razor.

The same as mTPI, all the dose-finding decisions of mTPI-2 can be pre-tabulated in advance, allowing investigators to examine the decisions before the trial starts. see Figure 1.28 for an example. And the decision table can also be generated via East Bayes under module **Decision & MTD**.

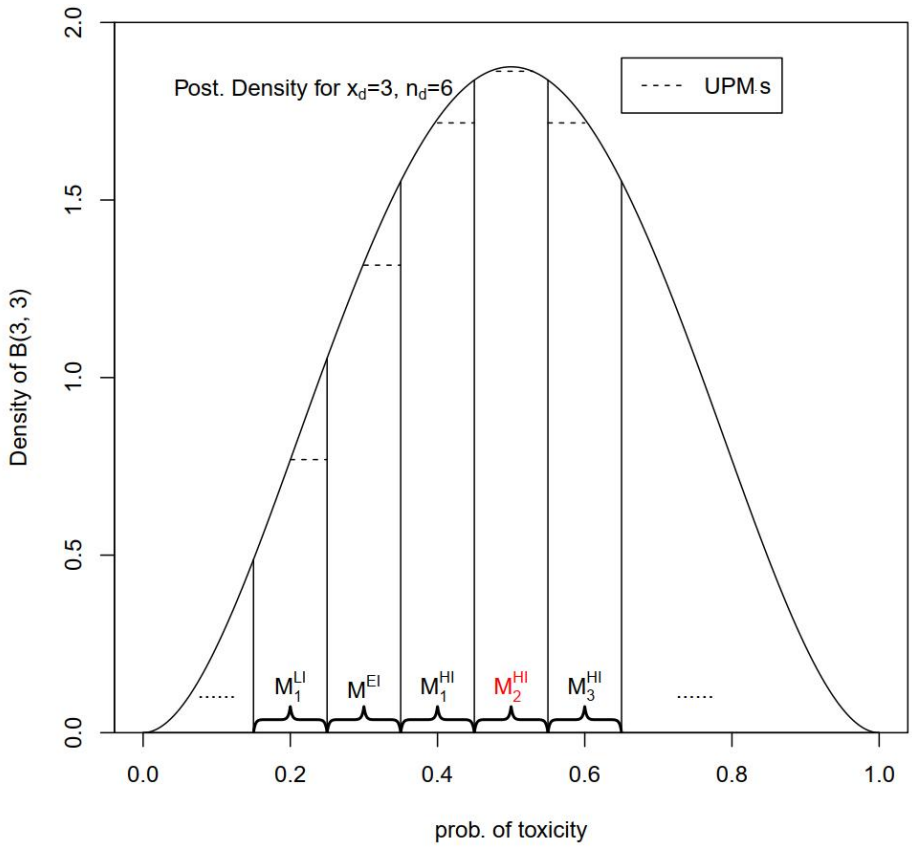


Figure 1.27: An example demonstrating the new framework of mTPI-2. Here, *EI* is the equivalence interval $(p_T - \epsilon_1, p_T + \epsilon_2)$, and *LI* denotes the intervals below *EI*, and *HI* denotes the intervals above *EI*. Interval M_2^{HI} exhibits the largest UPM and therefore the decision is now "D", to de-escalate.

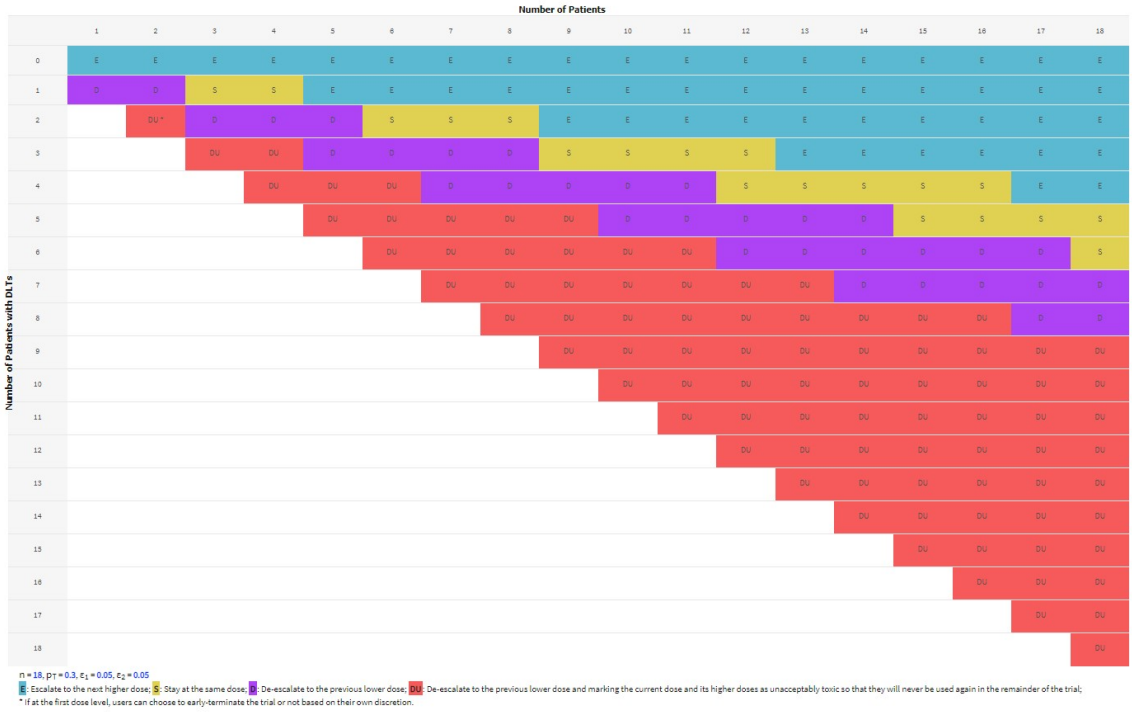


Figure 1.28: An example of mTPI-2 decision table generated via East Bayes. The target toxicity probability $p_T = 0.3$, and the equivalence interval (EI) is $(0.25, 0.35)$ for up to 18 subjects. Each column represents (n) number of subjects treated at the current dose and each row represents (y) number of subjects with DLTs at the dose level. Each cell in the table provides the dose-finding decision based on the readouts from the corresponding row (y) and column (n). The letters in the decision table represent different dose-assignment decisions.

1.3.5.3 The Keyboard Design

The Keyboard design is proposed by Yan et al. (2017), which is based on the same construction as the mTPI-2. In the Keyboard design, the sub-intervals are called ”keys” and the key associated with the largest posterior probability is chosen to guide the dose-assignment decisions. When the intervals are with equal-length, the winning interval with the largest posterior probability is the same as the interval with the largest UPM. Therefore, the keyboard design is the same as the mTPI-2 design.

1.3.6 The i3+3 Design

The i3+3 design is a rule-based design for finding the maximum tolerated dose (MTD) proposed by Liu et al. (2020). The i3+3 design defines an equivalence interval (EI) $[p_T - \epsilon_1, p_T + \epsilon_2]$ with the target probability of toxicity p_T and two small fractions, ϵ_1 and ϵ_2, and allocates the next cohort of patients based on the relationship between toxicity rate observed on the current cohort of patients and the equivalence interval. Similar to the 3+3 design, i3+3 is rule-based but assumes that toxicity increases with dose. It has been demonstrated to perform as good as major model-based designs and is flexible enough to accommodate different target toxicity probability as well as different cohort sizes (Liu et al., 2020).

1.3.6.1 Design Algorithm

Dose finding rules: Suppose dose d is currently used in the trial to treat patients, and y_d patients have experienced dose limiting toxicities (DLTs) out of n_d patients that have been treated. Based on EI, the i3+3 design identifies the appropriate dose for the next cohort of patients according to the following five simple rules, which accounts for the variability in the observed toxicity data (y_d and n_d) for each dose.

Current dose: d, No. enrolled: n_d, No DLTs: y_d		
Condition	Decision	Next dose level
$\frac{y_d}{n_d}$ below EI	Escalation(E)	$d+1$
$\frac{y_d}{n_d}$ inside EI	Stay(S)	d
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ below EI	Stay(S)	d
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ inside EI	De-escalation(D)	$d-1$
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ above EI	De-escalation(D)	$d-1$

Here, a value is below the EI means that the value is smaller than $(p_T - \epsilon_1)$, the lower bound of the EI. A value is inside the EI means that the value is larger than or equal to $(p_T - \epsilon_1)$ but smaller than or equal to $(p_T + \epsilon_2)$. A value is above the EI mean that the value is larger than $(p_T + \epsilon_2)$, the upper bound of the EI. All potential decisions based on the above set of rules could be pre-tabulated in advance via East Bayes under module **Decision & MTD**, allowing investigators for examination before the trial starts. See Figure 1.29 for an illustration. When d is the highest dose or lowest dose, the above rules are modified as special cases:

- If the current dose is the highest dose, and $\frac{y_d}{n_d}$ is below the EI, stay ("S") instead of escalating ("E") because there is no dose to escalate to.

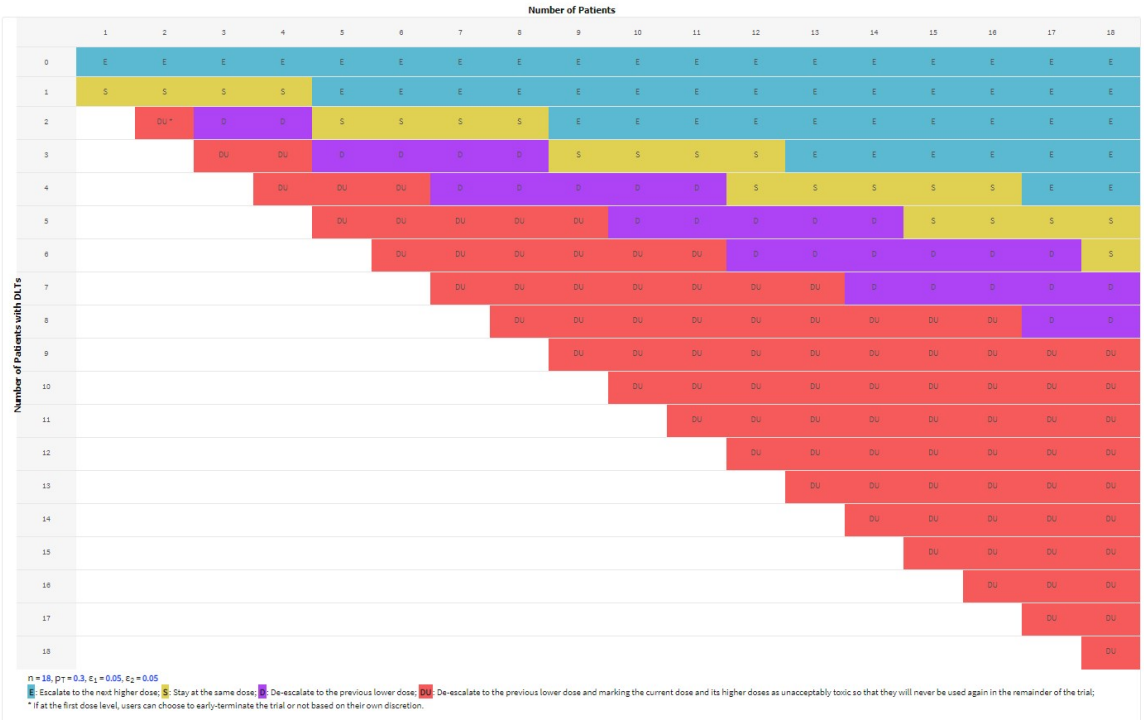


Figure 1.29: An example of i3+3 decision table generated via East Bayes. The target toxicity probability $p_T = 0.3$, and the equivalence interval (EI) is $(0.25, 0.35)$ for 18 subjects. Each column represents (n) number of subjects treated at the current dose and each row represents (y) number of subjects with DLTs at the dose level. Each cell in the table provides the dose-finding decision based on the readouts from the corresponding row (y) and column (n). The letters in the decision table represent different dose-assignment decisions.

– If the current dose is the lowest dose, and $\frac{y_d}{n_d}$ is above the EI, stay ("S") instead of potentially de-escalating ("D") because there is no dose to de-escalate to.

Safety rules: Following the mTPI and mTPI-2 design (Ji et al., 2010; Ji and Wang, 2013; Guo et al., 2017b), two safety rules are added as ethical constraints to avoid excessive toxicity:

– [***Rule 1: Dose Exclusion***] If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses are excluded and never be used again in the remainder of the trial. An exception of Rule 1 is that when there is only 1 DLT observed at a dose, the rule is not enforced.
– [***Rule 2: Early Stop***] If the current dose is the lowest dose (first dose) and is considered excessively toxic according to Rule 1, early stop the trial for safety.

In safety Rules 1 and 2, $Prob\{p_d > p_T \mid Data\}$ is a function of the cumulative beta distribution $beta(\alpha_0+y_d, \beta_0+n_d-y_d)$, and $\alpha_0 = \beta_0 = 1$ is used in East Bayes by default. And if i3+3 decision based on the current dose is "E", i.e., $\frac{y_d}{n_d}$ is below the EI, while the next higher dose level $(d+1)$ has been declared excessive toxicity and been excluded, stay ("S") instead of escalating ("E") because there is no available dose to escalate to.

Trial termination: The trial proceeds until any of the following stopping criteria is met:

1. If the prespecified maximum total sample size n is reached;
2. If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;
3. *Optional: ad-hoc rules of maximum number of patients in one dose, denoted by $K(K < n)$:*
   - *If the i3+3 decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *If the i3+3 decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *If the i3+3 decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients;*

1.3.6.2 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped early, the MTD selection under the i3+3 design follows the same procedure as in the mTPI and mTPI-2 design (Ji et al., 2010; Ji and Wang, 2013; Guo et al., 2017b). Follow the steps below:

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. Here in East Bayes, an independent prior $beta(0.005, 0.005)$ is used to compute the posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior mean by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA) (Robertson, 1988), the estimated DLT probabilities satisfying the order constraint is obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.

2. Among all the tried doses for which $Prob\{p_d > p_T \mid Data\} < \xi$ and $\hat{p}_d \leq p_T + \epsilon_2$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference),
   (a) If there is at least one dose lower the target p_T among all the tied doses, choose the highest dose among those as the estimated MTD;
   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

1.3.7 The Modified Cumulative Cohort Design (mCCD)

The cumulative cohort design (CCD) was formally proposed by Ivanova et al. (2007), which is also an interval-based design. But unlike the mTPI and mTPI-2 designs (in Sections 1.3.4 and 1.3.5, respectively) which calculate the posterior probability that the toxicity rate p_d falls into each interval and decide the decision based on a formal Bayesian decision framework, the CCD design just relies on the point estimate $\hat{p}_d$ and compares it with the equivalence interval boundaries, $(p_T - \epsilon_1)$ and $(p_T + \epsilon_2)$. In East Bayes, we construct a modified CCD (mCCD) design, which follows the same concept for dose finding as CCD, except that we add some other safety rules. The mCCD design is not published and is adopted by Cytel Inc.

1.3.7.1 Probability Model

Consider a phase I trial with D candidate doses for escalation. Let $p_1, \ldots, p_D$ denote the true toxicity probabilities for doses $d = 1, \ldots, D$. The observed data include n_d, the number of patients treated at dose d, and y_d, the number of patients experiencing a toxicity. Let $Data = \{(y_d, n_d); d = 1, 2, \ldots, D\}$.

The CCD design simply uses the empirical point estimate $\hat{p}_d = y_d/n_d$ as the estimation of toxicity rate p_d for dose level d.

1.3.7.2 Dosing Intervals

The mCCD design prespecifies three toxicity probability intervals: the under-dosing interval $(0, p_T - \epsilon_1]$, the equivalence interval $(p_T - \epsilon_1, p_T + \epsilon_2)$, and the over-dosing interval $[p_T + \epsilon_2, 1)$, where ϵ_1 and ϵ_2 are small fractions, such as 0.05, to account for the uncertainty around the true target toxicity. The three dosing intervals are associated with three different dose-escalation decisions. The under-dosing interval corresponds to a dose escalation (E), the over-dosing interval corresponds to a dose de-escalation (D), and the equivalence interval corresponds to staying at the current dose (S).

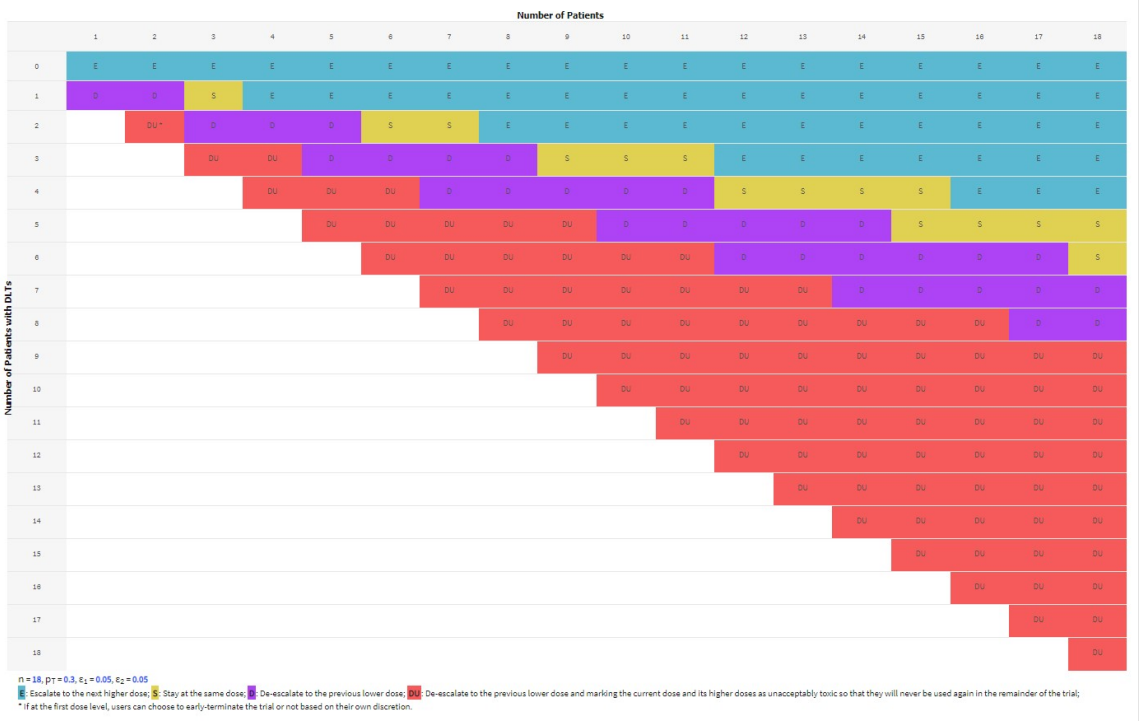


Figure 1.30: An example of mCCD decision table generated via East Bayes. The target toxicity probability $p_T = 0.3$, and the equivalence interval (EI) is $(0.25, 0.35)$ for up to 18 subjects. Each column represents (n) number of subjects treated at the current dose and each row represents (y) number of subjects with DLTs at the dose level. Each cell in the table provides the dose-finding decision based on the readouts from the corresponding row (y) and column (n). The letters in the decision table represent different dose-assignment decisions.

1.3.7.3 Dose Finding Rules

Suppose the current dose level is d, the mCCD applies the same concept for dose finding as CCD, that is, uses the equivalence interval as the boundaries for thresholding the estimate $\hat{p}_d$. Specifically,

1. Escalate to dose $(d+1)$, if $\hat{p}_d \in (0, p_T - \epsilon_1]$, i.e., if $y_d/n_d \leq p_T - \epsilon_1$,
2. Stay at dose d, if $\hat{p}_d \in (p_T - \epsilon_1, p_T + \epsilon_2)$, i.e., if $p_T - \epsilon_1 < y_d/n_d < p_T + \epsilon_2$,
3. De-escalate to dose $(d-1)$, if $\hat{p}_d \in (p_T + \epsilon_2, 1]$, i.e., if $y_d/n_d \geq p_T + \epsilon_2$.

The decision table based on the above rules can be generated via East Bayes before the beginning of the trial for investigators to examine. see Figure 1.30 for an example.

1.3.7.4 Design Algorithm

The mCCD algorithm is similar as mTPI, which proceeds as follows:

1. At each dose level, treat a cohort of patients, with the first cohort at a prespecified starting dose.
2. After all patients in each cohort complete the DLT evaluation, the dose-assignment decision for the next cohort will be determined according to the following rules:
   (a) Compute the posterior probability of excessive toxicity at the current tried dose, i.e., $Prob\{p_d > p_T \mid Data\}$ which is a function of the cumulative Beta distribution $Beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$. In East Bayes, $\alpha_0 = \beta_0 = 1$ is used.
      i. ***[Additional Safety Rule 1: Dose Exclusion]*** If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses are excluded and never be used again in the remainder of the trial. Following the original paper, we only enforce this rule for BOIN and mCCD when the number of subjects assigned to the dose is more than 2.
      Also, at that time, the decision is "D", to de-escalate to previous lower dose.
      ii. ***[Additional Safety Rule 2: Early Stop]*** If the current dose is the lowest dose and considered excessively toxic according to Rule 1 in i, early stop the trial for safety.
   (b) If the trial is not stopped early, assign the next cohort of patients to the dose according to the decision table or the procedures in Section 1.3.7.3.
   (c) If the dose-assignment decision is "E" but the next higher dose has been excluded by Rule 1, continue to enroll the next cohort at the current dose instead.
   (d) If the dose-assignment decision is "E" and the current dose is the highest dose, continue to enroll the next cohort at the current dose instead.
   (e) If the dose-assignment is "D" and the current dose is the lowest dose, continue to enroll the next cohort at the current dose instead.
3. Repeat steps 1-2, stop the trial when any of the following conditions is satisfied:
   (a) If the prespecified maximum total sample size is reached;
   (b) If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;
   (c) *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by* $K (K < n)$*:*

- *If the mCCD decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
- *If the mCCD decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
- *If the mCCD decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

1.3.7.5 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped early, an isotonic regression (Ji et al., 2010; Ivanova, Anastasia and Wang, Kai, 2006) is used to select the MTD based on the observed DLT data from all the dose levels. Follow the steps below:

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. Here in East Bayes, an independent prior $Beta(0.005, 0.005)$ is used to compute the posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior mean by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA) (Robertson, 1988), the estimated DLT probabilities satisfying the order constraint is obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.
2. Among all the tried doses for which $Prob\{p_d > p_T \mid Data\} < \xi$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d\, |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference),
   (a) If there is at least one dose lower the target p_T among all the tied doses, choose the highest dose among those as the estimated MTD;
   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

1.3.8 The Bayesian Optimal Interval Design (BOIN)

Liu and Yuan (2015b) extended CCD and developed the BOIN design, with local and global BOIN as two versions. The authors stated that BOIN is an improvement of CCD since it uses interval boundaries that are optimal based on an objective function. In the local BOIN design, based on a user-provided initial equivalence interval (ϕ_1, ϕ_2), an optimization procedure is defined to minimize an objective function that is claimed to be the probability of making an erroneous decision. Then the optimal values of equivalence interval boundaries are obtained and denoted as λ_1 and λ_2. The BOIN design first examines if $\hat{p}_d$ falls into one of the three intervals $(0, \lambda_1]$, (λ_1, λ_2), and $[\lambda_2, 1)$, and escalates to dose $(d+1)$, stays at dose d, or de-escalates to dose $(d-1)$, accordingly.

In other words, the BOIN design uses the same concept for dose finding as the CCD design, except BOIN changes the original user-provided boundary $\phi_{1,2}$ to $\lambda_{1,2}$ based on an optimization criterion. And Liu and Yuan (2015a) showed that (λ_1, λ_2) is always nested in the original interval (ϕ_1, ϕ_2) under local BOIN. In contrast, the mTPI (mTPI-2) and mCCD designs do not have the λs and use the user-provided ϕ's values for decision making. See Figure 1.31 for an illustration (Ji and Yang, 2017). As the Figure 1.31 shows, there is a gap between the λs and the user-provided ϕ's values in BOIN's decision making. The gap is small enough to be ignorable when the sample size is small. And BOIN performs well in terms of operating characteristics of safety and reliability in phase I trials.

In East Bayes, we implement the local BOIN design as recommended by Liu and Yuan(2015). Instead of specifying the initial equivalence interval (ϕ_1, ϕ_2), we ask users to directly provide ϵ_1, ϵ_2, where $(\lambda_1 = p_T - \epsilon_1, \lambda_2 = p_T + \epsilon_2)$ is the optimal interval for decision making. Therefore, designs like BOIN, mCCD, mTPI (mTPI-2) would use the same equivalence interval for decision making as long as they share the same target probability p_T and the ϵ_1, ϵ_2 values. Users can click the "Compute" button to retrive the ϕ_1 and ϕ_2 values from the specified ϵ_1, ϵ_2.

1.3.8.1 Probability Model

Consider a phase I trial with D candidate doses for escalation. Let $p_1, \ldots, p_D$ denote the true toxicity probabilities for doses $d = 1, \ldots, D$. The observed data include n_d, the number of patients treated at dose d, and y_d, the number of patients experiencing a toxicity. Let $Data = \{(y_d, n_d); d = 1, 2, \ldots, D\}$.

The BOIN design uses the empirical toxicity rate $\hat{p}_d = y_d/n_d$ as the estimation of toxicity rate p_d for dose level d. This is the same as the CCD and i3+3 designs.

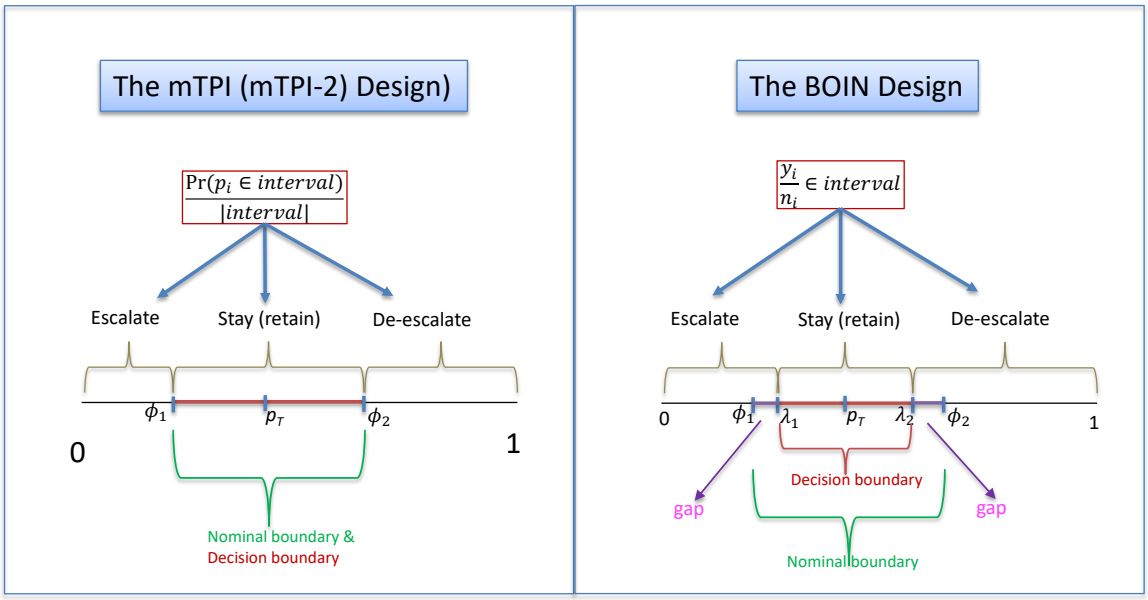


Figure 1.31: A graphical illustration between the decision frameworks under the mTPI (mTPI-2) and the BOIN design. Under mTPI, the two probability boundaries $\phi_1 = p_T - a$ and $\phi_2 = p_T + b$ are elicited from the clinicians and treated as known. Under BOIN, the two ϕ_1 and ϕ_2 values are also elicited from clinicians, but not used for decision making. Instead, two new values, λ_1 and λ_2 are derived based on an optimization procedure and used for decision making. There is a gap on each side of the p_T (right panel) due to the optimization process, and the gap is independent of sample size.

1.3.8.2 Optimal Interval

Denote the initial lower and upper bound of the equivalence interval for p_T as ϕ_1 and ϕ_2, which are elicited from clinicians. Suppose dose d is currently administered in the trial. In the BOIN design, the optimal interval (λ_1, λ_2) minimizing the probability of making an erroneous decision are given by

$$\lambda_1 = log\left(\frac{1-\phi_1}{1-p_T}\right) / log\left(\frac{p_T(1-\phi_1)}{\phi_1(1-p_T)}\right)$$
$$\lambda_2 = log\left(\frac{1-p_T}{1-\phi_2}\right) / log\left(\frac{\phi_2(1-p_T)}{p_T(1-\phi_2)}\right)$$

In East Bayes, we let $\epsilon_1 = p_T - \lambda_1$ and $\epsilon_2 = \lambda_2 - p_T$. Given p_T, ϵ_1, and ϵ_2 , ϕ_1 and ϕ_2 could be conversely computed numerically using the equations above.

The BOIN design uses three toxicity probability intervals: the under-dosing interval $(0, \lambda_1]$, the equivalence interval (λ_1, λ_2), and the over-dosing interval $[\lambda_2, 1)$ for three different dose-escalation decisions. The under-dosing interval corresponds to a dose escalation (E), the over-dosing interval corresponds to a dose de-escalation (D), and the equivalence interval corresponds to staying at the current dose (S).

1.3.8.3 Dose-Finding Rules

Suppose the current dose level is d, BOIN uses the optimal equivalence interval (λ_1, λ_2) as the boundaries for thresholding the estimate $\hat{p}_d$. Specifically,

1. Escalate to dose $(d+1)$, if $\hat{p}_d \in (0, \lambda_1]$, i.e., if $y_d/n_d \leq \lambda_1$,
2. Stay at dose d, if $\hat{p}_d \in (\lambda_1, \lambda_2)$, i.e., if $\lambda_1 < y_d/n_d < \lambda_2$,
3. De-escalate to dose $(d-1)$, if $\hat{p}_d \in [\lambda_2, 1)$, i.e., if $y_d/n_d \geq \lambda_2$.

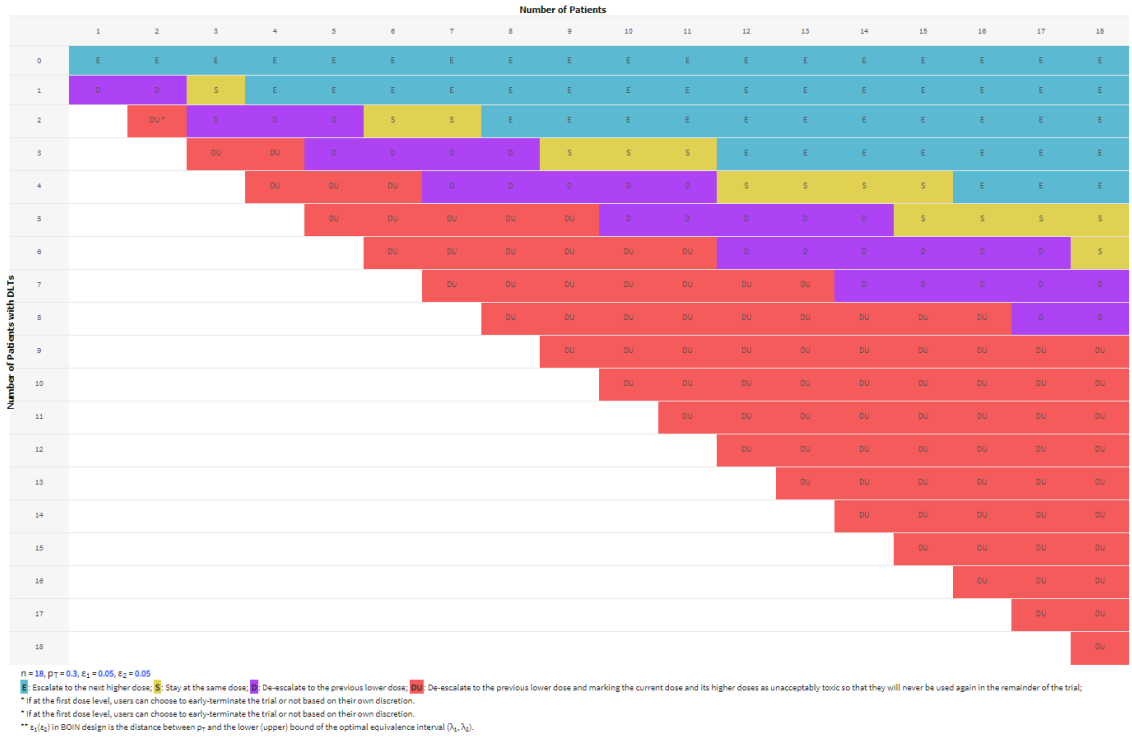


Figure 1.32: An example of BOIN decision table for 18 subjects generated via East Bayes. The target toxicity probability $p_T = 0.3$, and the optimal equivalence interval (λ_1, λ_2) is $(0.25, 0.35)$, and hence $\epsilon_1 = \epsilon_2 = 0.05$. Each column represents (n) number of subjects treated at the current dose and each row represents (y) number of subjects with DLTs at the dose level. Each cell in the table provides the dose-finding decision based on the readouts from the corresponding row (y) and column (n). The letters in the decision table represent different dose-assignment decisions.

The decision table based on the above rules can be generated via East Bayes before the beginning of the trial for investigators to examine. see Figure 1.32 for an example.

1.3.8.4 Design Algorithm

The BOIN algorithm is similar to mCCD, which proceeds as follows:

1. At each dose level, treat a cohort of patients, with the first cohort at a prespecified starting dose.
2. After all patients in each cohort complete the DLT evaluation, the dose-assignment decision for the next cohort will be determined according to the following rules:
   (a) Compute the posterior probability of excessive toxicity at the current tried dose, i.e., $Prob\{p_d > p_T \mid Data\}$ which is a function of the cumulative beta distribution $Beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$. In East Bayes, $\alpha_0 = \beta_0 = 1$ is used.
      i. ***[Additional Safety Rule 1: Dose Exclusion]*** If the current dose is considered excessively toxic, i.e., $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is a probability close to 1, say 0.95, the current and all higher doses are excluded and never be used again in the remainder of the trial. Following the original paper, we only enforce this rule for BOIN and mCCD when the number of subjects assigned to the dose is more than 2.
      Also, at that time, the decision is "D", to de-escalate to previous lower dose.
      ii. ***[Additional Safety Rule 2: Early Stop]*** If the current dose is the lowest dose and considered excessively toxic according to Rule 1, stop the trial for safety.
   (b) If the trial is not stopped in (a), assign the next cohort of patients to the dose according to the decision table or the procedures in Section 1.3.7.3.
   (c) If the dose-assignment decision is "E" but the next higher dose has been excluded by Rule 1 in(a), continue to enroll the next cohort at the current dose instead, i.e., the decision is changed to "S".
   (d) If the dose-assignment decision is "E" and the current dose is the highest dose, continue to enroll the next cohort at the current dose instead, i.e., the decision is changed to "S".
   (e) If the dose-assignment is "D" and the current dose is the lowest dose, continue to enroll the next cohort at the current dose instead, i.e., the decision is changed to "S".
3. Repeat steps 1-2; stop the trial when any of the following conditions is satisfied:
   (a) If the prespecified maximum total sample size is reached;
   (b) If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;
   (c) *Optional: ad-hoc stopping rules of maximum number of patients at a dose, denoted by*

$K (K < n)$: In any of the following cases, stop the trial

- *If the BOIN decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
- *If the BOIN decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
- *If the BOIN decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

1.3.8.5 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped early, an isotonic regression (Ji et al., 2010; Ivanova, Anastasia and Wang, Kai, 2006) is used to select the MTD based on the observed DLT data from all the dose levels. Follow the steps below:

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. Here in East Bayes, an independent prior $Beta(0.005, 0.005)$ is used to compute the posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior mean by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA) (Robertson, 1988), the estimated DLT probabilities satisfying the order constraint is obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.
2. Among all the tried doses for which $Prob\{p_d > p_T \mid Data\} < \xi$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference), denote the trial value as p^*.
   (a) If $p^* <= p_T$, choose the highest dose among the tied doses as the estimated MTD;
   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

2. Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment

2.1 Introduction

Phase I oncology dose-finding trials assign cancer patients to ascending doses of a new investigational drug (or drug combinations) and adaptively decide the dose level of newly enrolled patients based on observed binary dose-limiting toxicity (DLT) outcomes. The goal is to determine the maximum tolerated dose (MTD) of the drug(s), defined as the highest dose that has a toxicity probability less than or close to a prespecified target rate p_T. Popular statistical designs, such as the 3+3 (Storer, 1989), CRM (O'Quigley et al., 1990), mTPI-2 (Guo et al., 2017b), and i3+3 (Liu et al., 2020) designs described in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module, typically enroll patients in cohorts, follow the enrolled cohort for a certain time period (e.g. 28 days), and apply sequential decisions that determine the dose level for each cohort based on the observed toxicity data. Accrual is suspended after enrollment of each cohort of patients until all the patients in the current cohort have been fully followed with definitive DLT or non-DLT outcomes. This type of cohort-based designs can be inefficient, especially if the trial needs to be frequently suspended. See Skolnik et al. (2008) and Doussau et al. (2016) for discussion. For example, subsequent patients can be turned away during accrual suspension, resulting in waste of precious patient resource. In addition, trial duration is prolonged due to the suspensions.

To shorten the study duration of phase I trials and reduce the number of accrual suspensions, this module describes a number of rolling-enrollment designs, which allows concurrent patient enrollment that is faster than cohort-base enrollment.

Besides the operating characteristics in terms of the safety and reliability reported in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module, this module enables users to compare the trial duration based on real-life settings, which are characterized as

three user-input parameters, the mean inter-patient arrival time, the maximum DLT follow-up time, and the probability of inevaluability (such as drop off) of enrollment patients. The procedure of simulating patients enrollment and evaluation is described in details in Section 2.3.1.

Hereinafter, the terms “Enrollment” and “Accrual” are used interchangeably.

2.2 User Interface and Tutorial

2.2.1 Overview

Upon entering the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** page, three main tabs are presented: **Simulation Setup**, **Simulation Results**, and **MTD Estimation**. The first two tabs allow users to conduct simulations and visualize/download simulation results, and the last tab allows users to estimate the MTD. In the **Simulation Setup** tab, there are four steps (Figure 2.1): 1) **Set enrollment parameters**, 2) **Set trial parameters**, 3) **Select designs**, and 4) **Generate scenarios**. Users need to complete all four steps to set up simulations for a single or multiple designs. Upon completion, users click the "Launch Simulation" button at the bottom of the page. Users may also click the "Reset" button next to "Launch Simulation" to clear all settings. After the simulation is launched, the results of simulations will be displayed in the **Simulation Results** tab. Simulation progress can be monitored in real time at the top of the **Simulation Results** tab. Detailed steps of using this module are described in §2.2.2-§2.2.4.

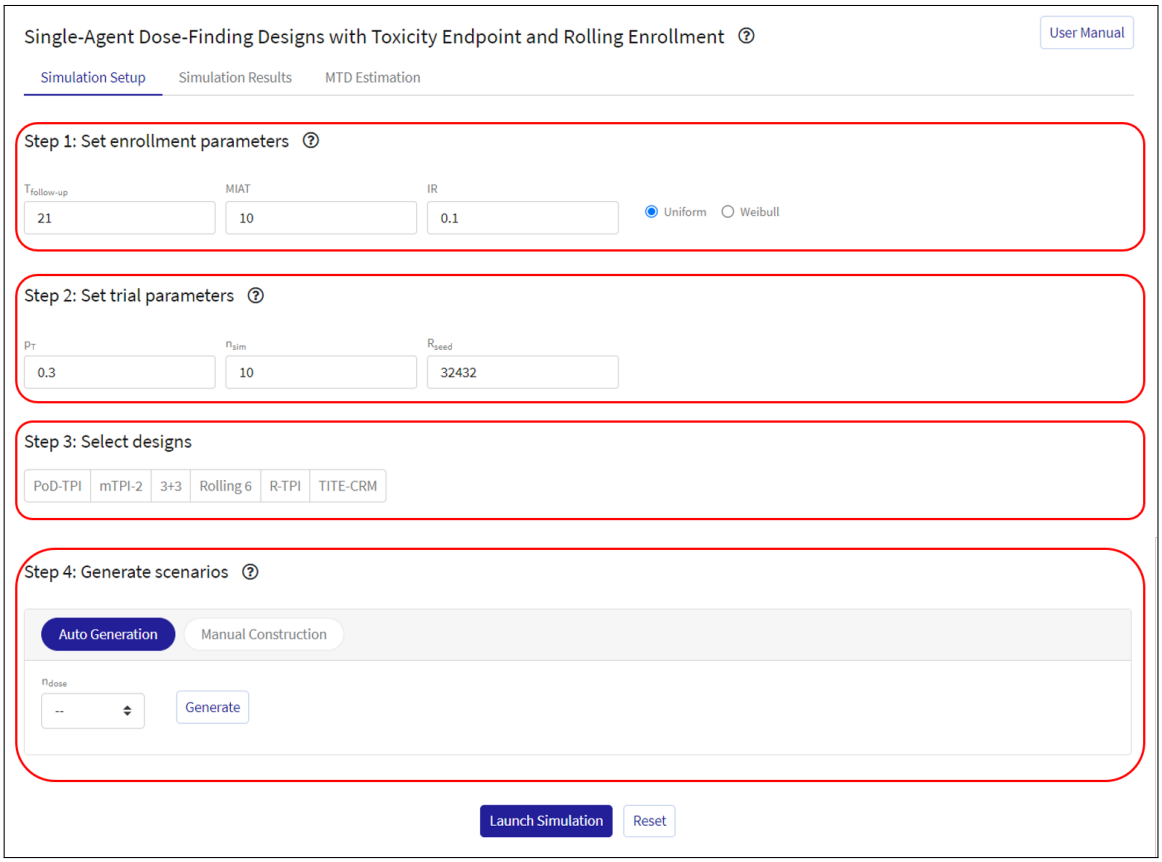


Figure 2.1: Simulation setup in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

2.2.2 Simulation Setup

In the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module, East Bayes provides six designs, 3+3, mTPI-2, Rolling 6, R-TPI, PoD-TPI and TITE-CRM. Users can choose up to four designs for simultaneous comparison in the **Simulation Setup** tab. In application messages will prompt the user if the setup is not complete and more fields require values prior to launching the simulation(s) Requests to allow more than four designs to be simultaneously compared can be made by emailing support@cytel.com.

2.2.2.1 Step 1: Set enrollment parameters

Specify the maximum follow-up time ($T_{follow-up}$), mean interpatient arrival time ($MIAT$), inevaluable rate (IR), and the distribution of time to DLT for the simulation. See Figure 2.2. If the Weibull distribution is selected for the time to DLT, two parameters of the Weibull distribution, α and γ can be specified. The detailed explanation of the above three input arguments is provided in Table 2.1. The technical details of simulating patients enrollment are provided in §2.3.1.

Figure 2.2: Set enrollment parameters in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Table 2.1: Input arguments for enrollment parameters in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Notation	Parameters	Description
$T_{follow-up}$	The maximum follow-up time	The DLT observation period for each patient in the trial (days). The default value is 21 days.
$MIAT$	Mean interpatient arrival time	The mean chronologic time (days) for a patient to arrive in the clinic and be eligible for study. The default value is 10 days.
IR	Inevaluable rate	The proportion of patients who entered the trial and received the treatment, but dropped out due to non-DLT related event when being followed. The default value is 0.1.
α, γ	The two parameters of the Weibull distribution	If a DLT occurs within the assessment window, with probability α it occurs within the last fraction γ of the follow-up time period. The default values are both 0.5.

2.2.2.2 Step 2: Set trial parameters

Specify the target toxic probability (p_T), number of simulations (n_{sim}), and random seed of simulation (R_{seed}) for the simulated trials. See Figure 2.3. A detailed explanation of the above three input arguments is provided in Table 2.2.

Figure 2.3: Set trial parameters in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Table 2.2: Input arguments for trials in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Notation	Parameters	Description
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T. The default value is 0.3.
n_{sim}	The number of simulated trials	The maximum number of simulated trials allowed is 10,000. The default value is 10.
R_{seed}	The random seed of simulation	A random seed is a number used to initialize a pseudorandom number generator in the simulation. The default value is 32432.

2.2.2.3 Step 3: Select designs

To select a design, click the button bearing the design's name on it. Up to four design configurations may be selected for comparison. For example, one could choose a design with four different sample sizes or four designs with the same sample size.

When setting the sample size n for the PoD-TPI, mTPI-2, R-TPI and TITE-CRM designs, two options are provided: 1) match with 3+3, if a 3+3 design is selected; 2) manually input. Check the "Match with 3+3" box to use the average sample size of the selected 3+3 design as the maximum sample size n for any none 3+3 designs. If two or more 3+3 design configurations are selected, East Bayes chooses the first 3+3 design in the design list as the benchmark. A 3+3 design must be selected first in order to check the "Match with 3+3" box. Figure 2.4 presents an example where the mTPI-2, 3+3, Rolling 6 and R-TPI designs are selected, with the sample size of mTPI-2 matching 3+3's, and the sample size of R-TPI being a manually input value, 30.

For the mTPI-2 and 3+3 designs, check the "Apply Decision in Advance" box to apply a modified and faster version of each designs. The modified designs use the following rules in dose finding: If unobserved toxicity responses of any enrolled patients in the current cohort have no influence on the decision of dose escalation, an early dose assignment decision will be made immediately without waiting for the patients that are still being followed without no definitive outcomes.

Design parameters can be modified in the input box. See detailed parameter descriptions in Table 2.3.

Click the "Delete" button to remove the selected designs.

Click the "Apply" button for each designs before launching simulations.

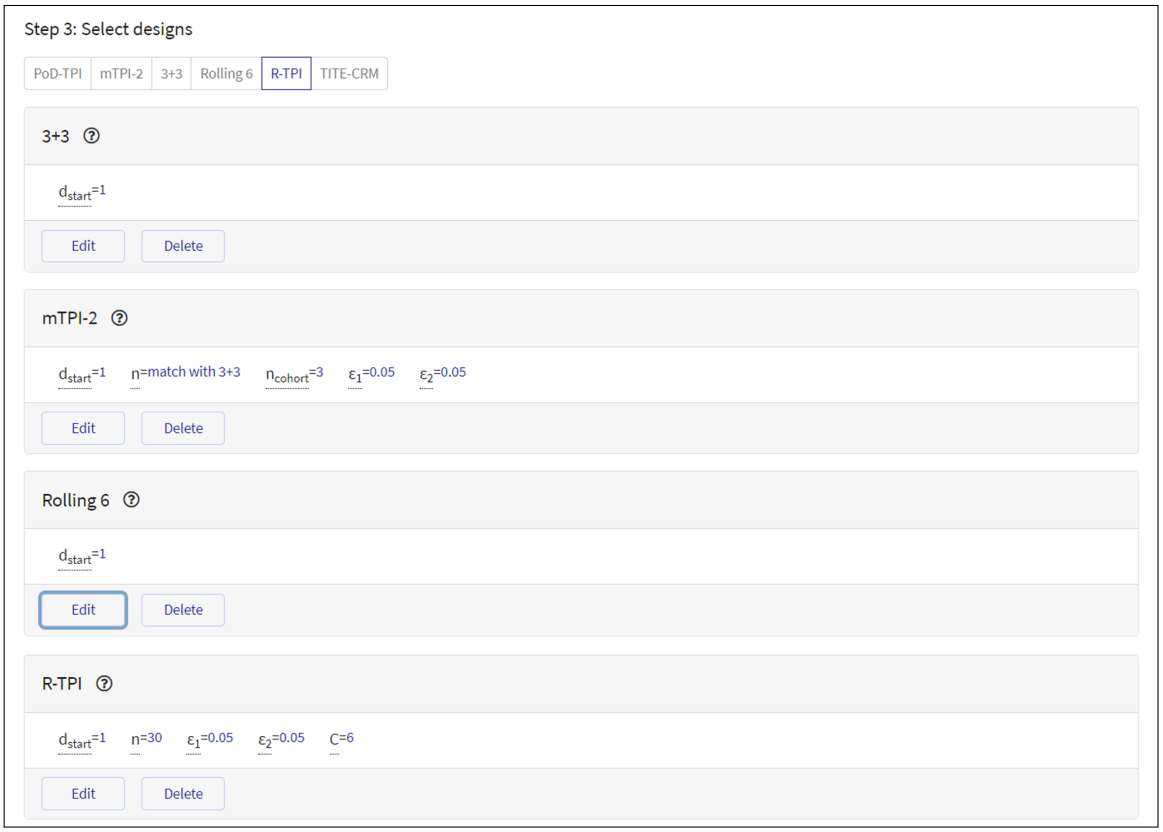


Figure 2.4: Select designs in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Table 2.3: Input parameters for designs in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment**.

Notation	Parameters	Description
d_{start} (all designs)	Starting dose level	The starting dose level in the simulated trials. The default value is 1.
n (mTPI-2, R-TPI, PoD-TPI, TITE-CRM)	Sample size	The maximum number of patients to be treated in the trial. The upper limit is set at 100 since the number of patients that are enrolled in phase I clinical trial is typically small. The default value is 30.
ϵ_1,ϵ_2 (mTPI-2, R-TPI, PoD-TPI)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence/target interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose of MTD. The default values for both are 0.05.
n_{cohort} (mTPI-2, PoD-TPI)	Cohort size	The number of patients in each cohort. The default value is 3. For 3+3, the cohort size is 3 by default, and for the Rolling 6 and R-TPI designs, there is no concept of cohort size and patients are enrolled as needed except if enrollment is suspended by the design .
C (R-TPI, TITE-CRM)	The maximum number of pending patients allowed in the trial	The maximum number of pending patients allowed in the trial without observed outcomes. It can be provided by users to control the enrollment speed. For the Rolling 6 design, C is 6 by default.
π_E, π_D (PoD-TPI)	The thresholds of the decision probabilities in the suspension rules	If the posterior probability of escalation is less than π_E, escalation is not allowed and the trial is suspended. If the posterior probability of de-escalation is higher than π_D, stay is not allowed and the trial is suspended. The default values are 1 and 0.15.
δ (TITE-CRM)	Half-width	The halfwidth of the indifference interval in selecting the skeleton of the model. The default value is 0.05.

2.2.2.4 Step 4: Generate scenarios

There are two ways to generate scenarios, automatically (in below **Auto Generation** tab, see Figure 2.5) or through manual construction (in below **Manual Construction** tab, see Figure 2.6). Once scenarios are generated, click the "Launch Simulation" button at the bottom of the page to run the n_{sim} (set in step 1) simulations, for each scenario and the selected design(s) (set in step 2) combination, assuming p_T (set in step 1) is the target for the MTD.

Auto Generation (Figure 2.5)
Select the number of doses n_{dose} ($3 \leq n_{dose} \leq 10$) from the dropdown box. Upon clicking the "Generate" button, five or six scenarios will be created automatically, each of which contains the true toxicity probabilities for n_{dose} dose levels. These generated scenarios are displayed and editable. The detailed algorithm for scenarios auto generation is provided next.

Manual Construction (Figure 2.6)
Follow the instructions below to manually construct scenarios. Then click the "Add" button to create these scenarios. The format of input must comply with the following instructions.

- Scenarios are separated by line breaks;
- Each scenario consists of a set of true toxicity probabilities for all dose levels;
- The true toxicity probabilities must be separated by a white space or comma.
- There should be at least three doses for each scenario.

For example, by inputting "0.05 0.1 0.15 0.2" or "0.05,0.1,0.15,0.2", a scenario is presented with true toxicity probabilities of four dose levels, 0.05, 0.1, 0.15 and 0.2.

The generated scenarios are displayed as a list (Figures 2.5 and 2.6) which appears below the generation section. The generated scenarios are editable by clicking the edit icon. An interactive chart will also be generated to visually display the shape of true toxicity probabilities for each scenario.

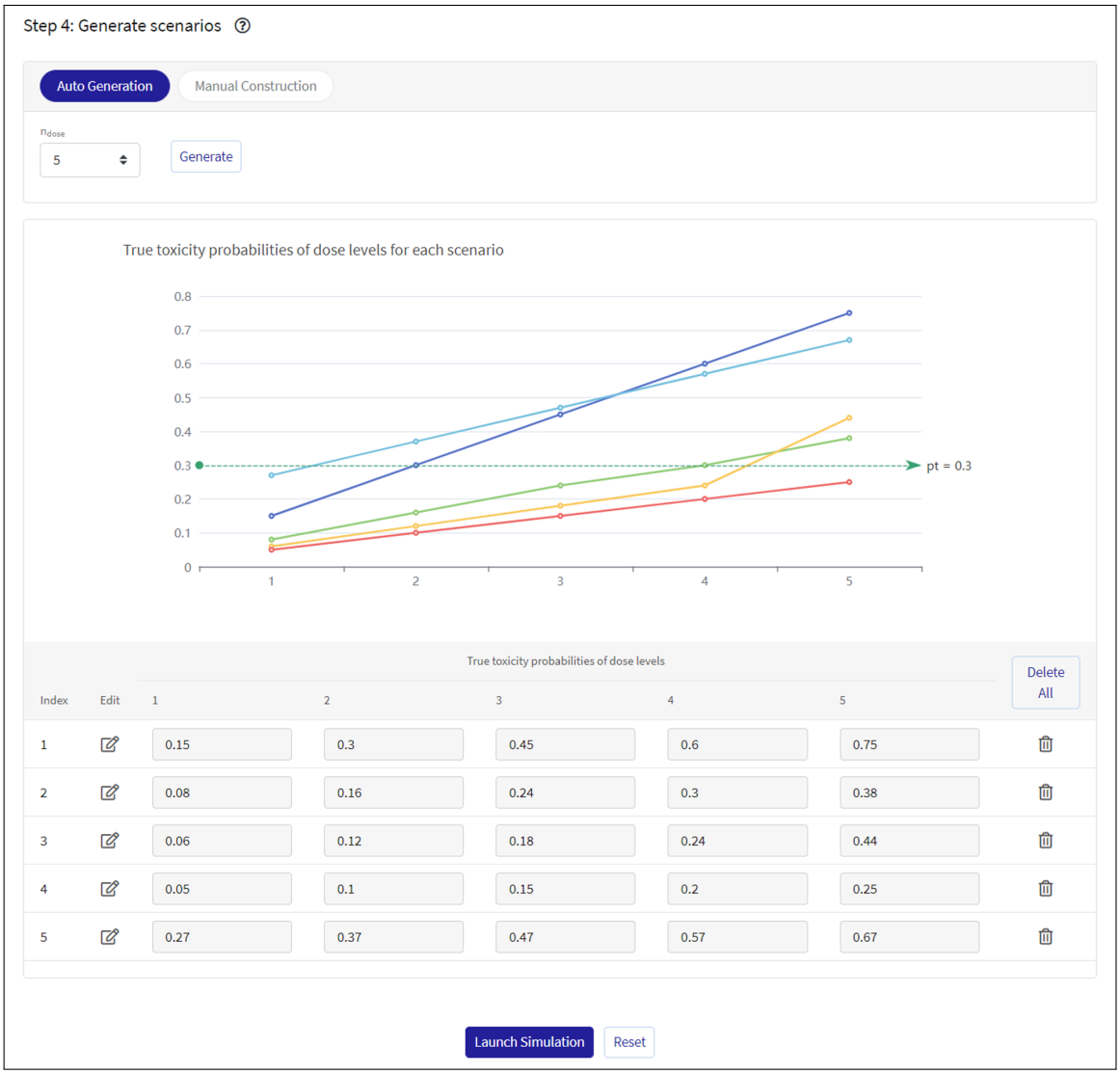


Figure 2.5: Automatically generate scenarios in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

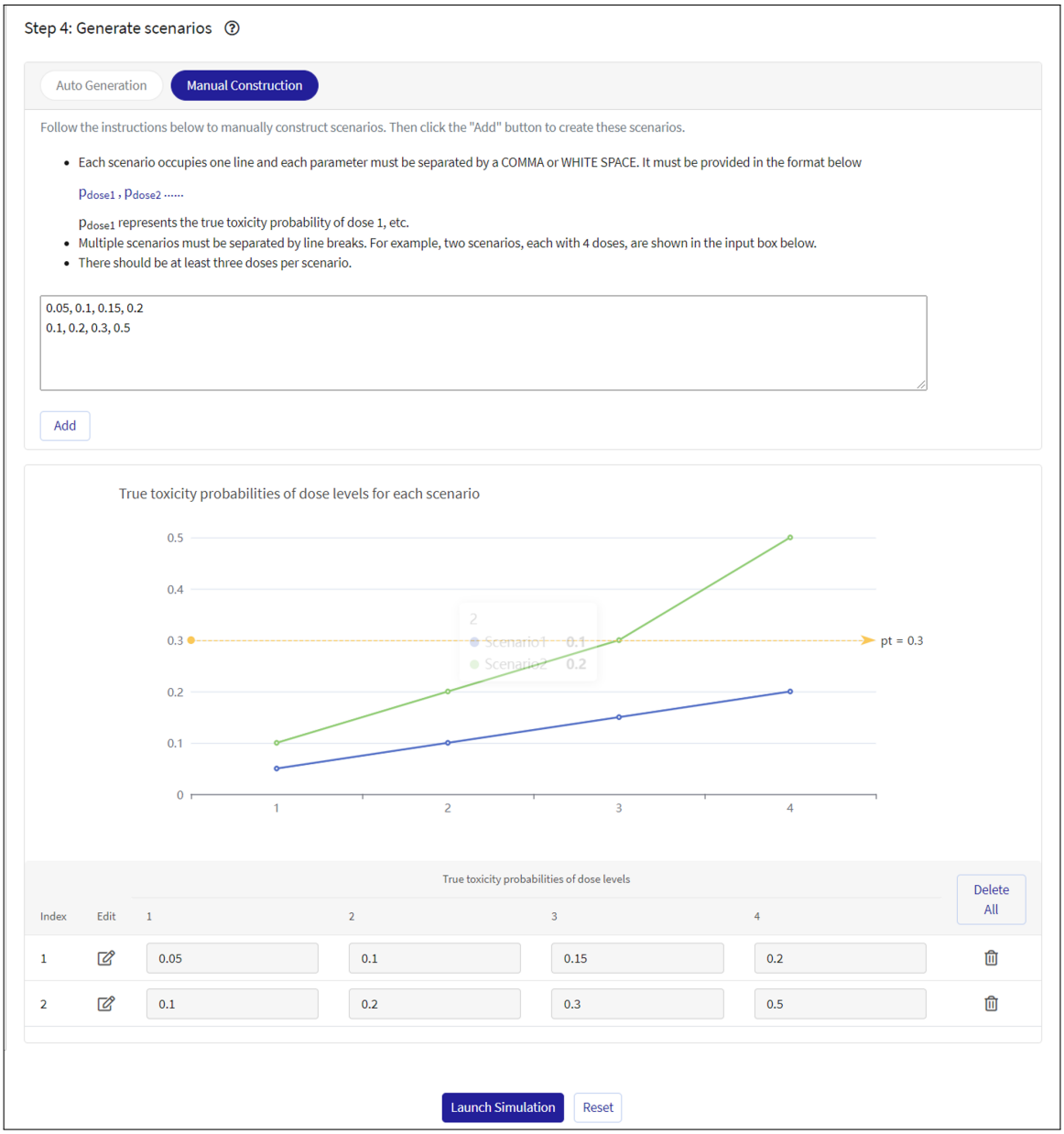


Figure 2.6: Manually generate scenarios in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Algorithm for Auto Generation

By entering the number of candidate dose levels n_{dose}, five or six scenarios are generated automatically. See Figure 2.7 for an illustration. They represent the four types of dose-response shapes below.

Types	Dose-Response Shape
Ideal	Some doses are tolerable but some are overly toxic, AND there exists at least one dose level close to the target p_T or falling within the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$.
Safe	All doses are safe and tolerable with the true toxicity probabilities lower than the target p_T or the lower bound of the equivalence interval ($p_T - \epsilon_1$).
Toxic	All doses are overly toxic with the true toxicity probabilities higher than the target p_T or the upper bound of the equivalence interval ($p_T + \epsilon_2$).
Steep	Some doses are tolerable but some are overly toxic, AND there is a steep jump in the toxicity probability between two adjacent doses, AND there is no dose close to the target p_T or falling within the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$.

Two "Steep" scenarios are generated, with the toxicity probability steep jump occurring at the first or second half of the doses. Similarly, two "Ideal" scenarios might be generated, with the MTD placed in the first or second half of the doses. This depends on the number of doses. When the number of doses is greater than 6, two scenarios of "Steep" and "Ideal" will be generated.

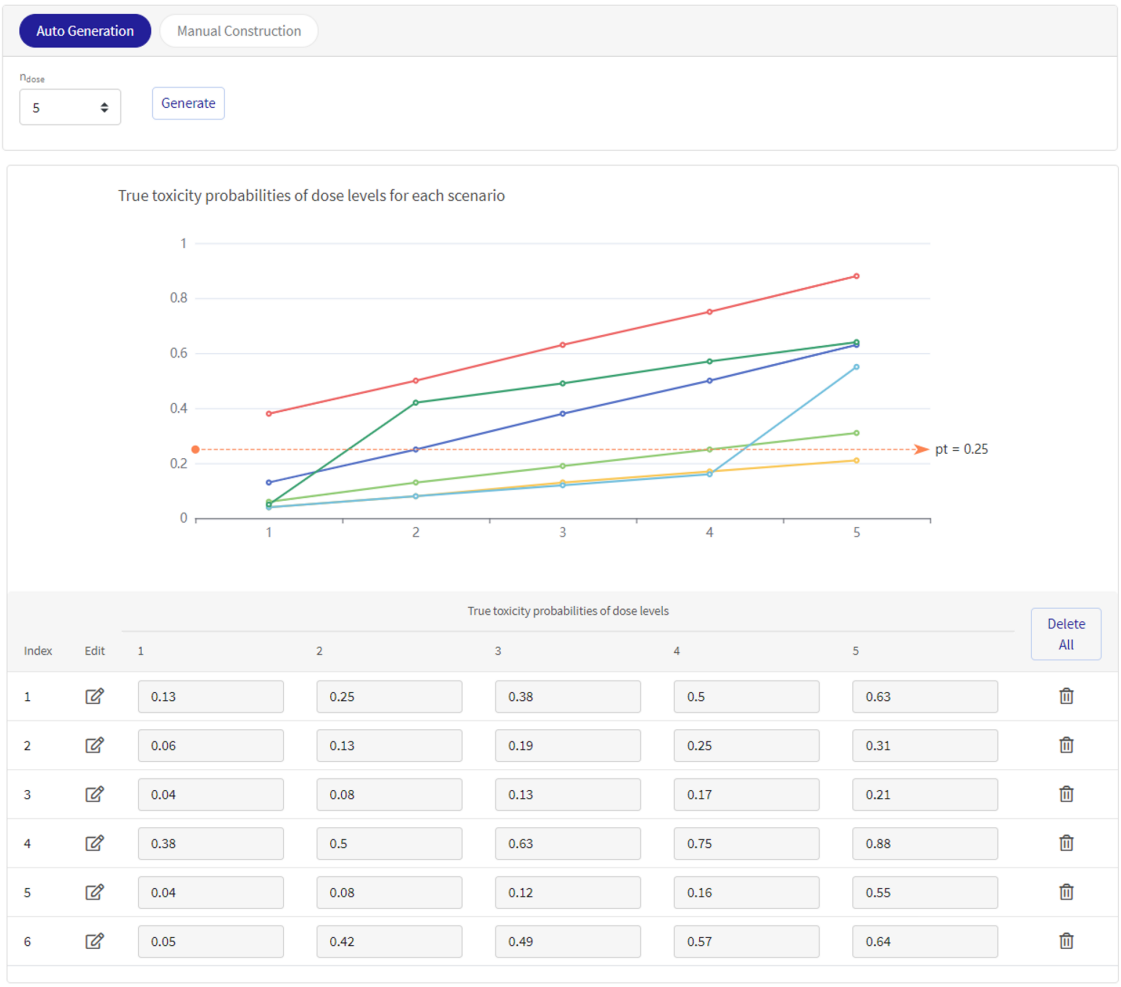


Figure 2.7: An example of automatically generated scenarios. Five dose levels are considered for the trial. The target toxicity probability is $p_T = 0.25$, and the equivalence interval is EI=$[0.2, 0.3]$. The six different lines represents the four types of scenario. In the "Ideal" scenarios (Lines 1 and 2), doses 2 and 4 are the true MTD with toxicity probability falling within the EI, respectively. In the "Safe" scenario (Line 3), all doses are safe with toxicity probabilities lower than the target $p_T = 0.25$. The "Toxic" scenario (Line 4) gives a contrary situation to the "Safe" scenario, where all doses are overly toxic with the toxicity probabilities higher than the target $p_T = 0.25$. The remaining two lines (Lines 5 and 6) are the "Steep" scenarios, in which some doses are tolerable but some are overly toxic, and there is a steep jump in the toxicity probability occurring at the first or second half of the doses (between doses 4 and 5 in Line 5, and doses 1 and 2 in Line 6).

2.2.2.5 Launch simulation

Once steps 1) -4) are completed, users can conduct simulated clinical trials by clicking the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 2.5 and 2.6) to examine the operating characteristics of the selected designs using the selected scenarios. A "**Launch Successful**" message will be displayed on the website (Figure 2.8) to indicate that the simulation has been successfully launched. Users may click the "Proceed To Simulation Results" button in the pop-up box to track the simulation processing status and simulation results.

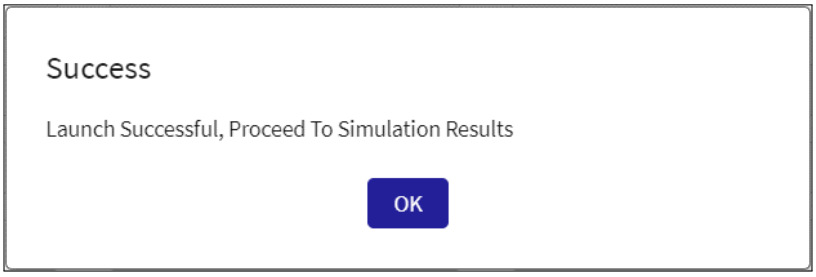


Figure 2.8: A "**Launch Successful**" message appears after launching simulation in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

2.2.2.6 Benchmark

Since the rolling designs may shorten trial duration, it is important to benchmark against designs that do not use rolling enrollment, such as mTPI-2 design. To facilitate the comparison, East Bayes automatically simulate trials based on the mTPI-2 design in this module, regardless if mTPI-2 is selected in step 3 (§2.2.2.3). If mTPI-2 is already selected, East Bayes does nothing additional. If mTPI-2 is not selected, East Bayes will add mTPI-2 to the design list and simulations will be executed based on the mTPI-2 design (with default settings) in addition to the designs selected by users. The sample size of the added mTPI-2 design will be the largest sample size among selected rolling designs.

2.2.3 Simulation Results

In the **Simulation Results** tab, users can view the simulation progress and simulation results (§2.2.3.1), restore the simulation settings (§2.2.3.2), and download intelligent simulation reports (§2.2.3.3). Specifically, all the simulation results (figures and tables) can be downloaded in Word format. Hereinafter, we use the terms "simulation results" and "operating characteristics" interchangeably.

In addition, the mTPI-2 design and its simulation results will be displayed as benchmark to rolling designs, if mTPI-2 is not selected by users. It will be labeled as "**Benchmark**".

2.2.3.1 View simulation results

In the **Simulation Results** tab, the **Running Simulations** panel displays the progress of simulations being computed (Figure 2.9). Simulations are displayed in ascending order by the launch time. Click the icon "×" to cancel a simulation in progress.

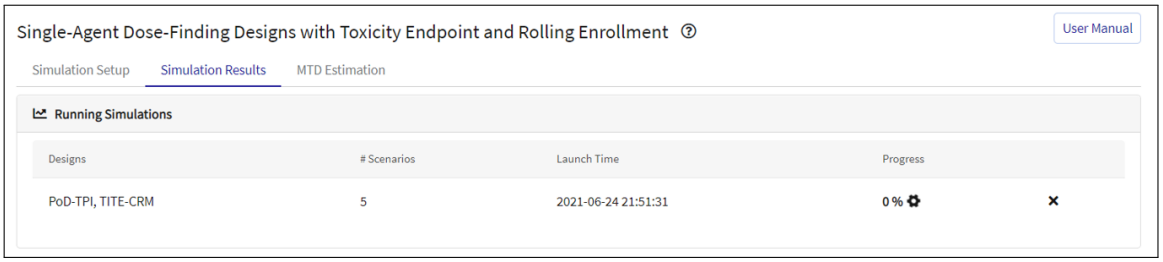


Figure 2.9: Simulation progress in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

When all the simulations are completed, the **Running Simulations** panel in Figure 2.9 is not shown, instead a "*simulation result created*" message is shown. These messages can be dismissed by clicking the icon "×" at the end of the corresponding row. The simulation results are automatically loaded into the **Simulation History** panel (Figure 2.10), with the blue mail icon shown to indicate new results. All previously completed simulations are also listed in the **Simulation History** panel. Simulation results for other modules can also be viewed under the **Simulation History** by dropping down the "Select a module" button (Figure 2.10). Click the button to delete the selected simulation results.

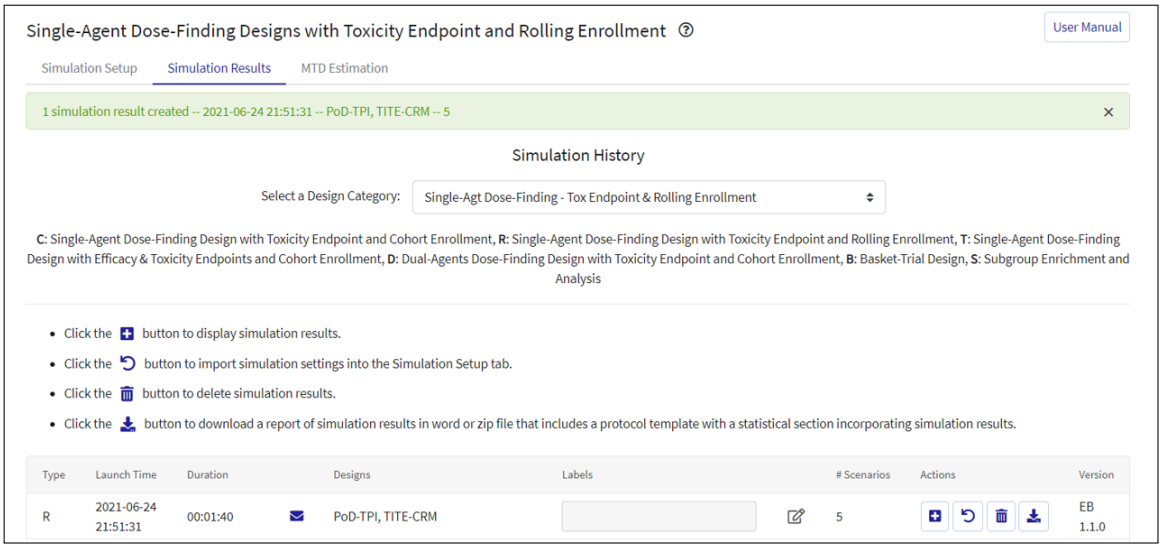


Figure 2.10: Simulation Results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Click the button ⊞ to expand and view the simulation results (Figure 2.10). The design settings are firstly displayed at the top of each simulation study (Figure 2.11). Then the results of simulation are shown as plots and tables below.

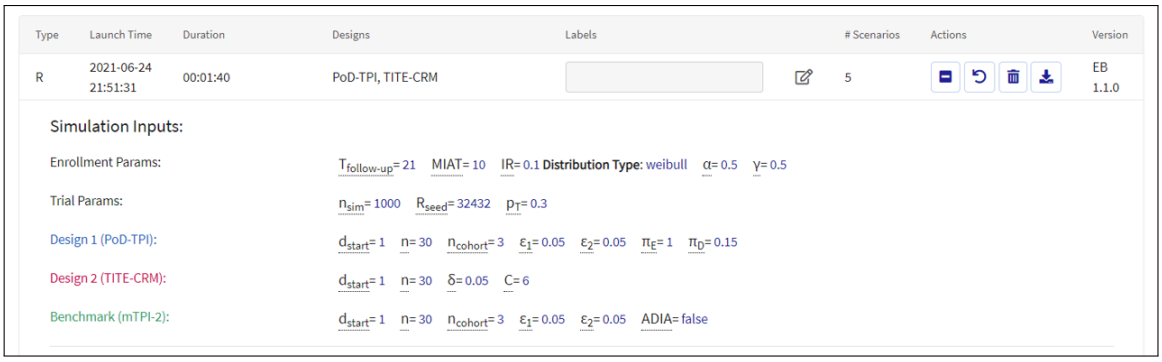


Figure 2.11: View the simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Details of the Simulation Results

The simulation results are divided into four parts, i.e, Duration and Risks, Tabulated Results by Scenarios, Key Metrics for Dose Finding, and Inconsistent Decisions Breakdowns. Each part can be viewed or hidden by clicking the button for that part (Figure 2.12).

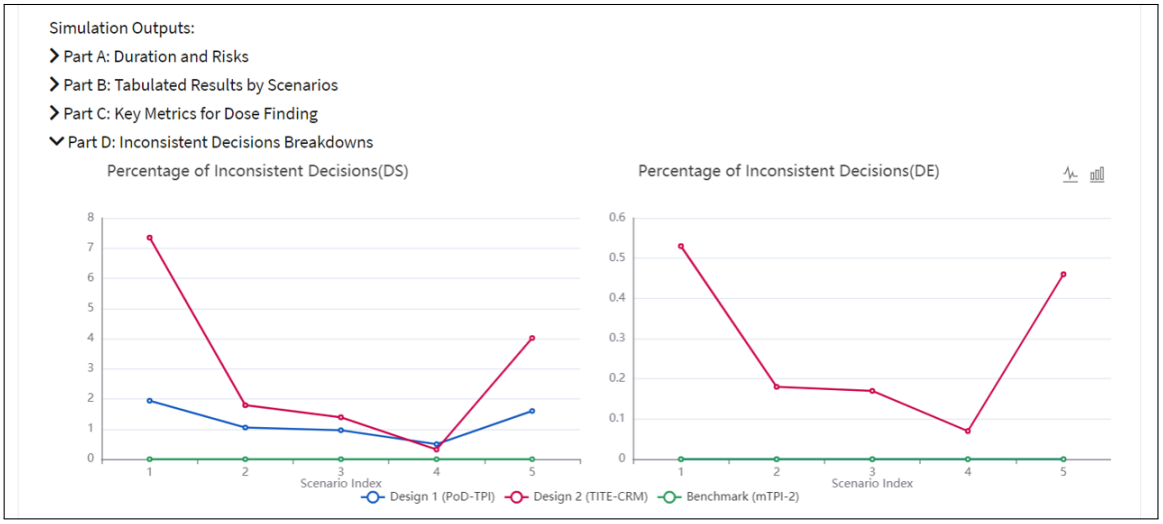


Figure 2.12: View each part of the simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Part A: Duration and Risks

There are three items in Part A:

a. Line plots showing the **Trial Duration (days)** and **Sum of Risky Decisions (%)** for all the designs (Figure 2.13).

b. A table showing the **Average Trial Duration (days) and Risky Decisions (max % across scenarios)** for all designs (Figure 2.14).

c. A table of mean and standard deviation (s.d.) for seven summary statistics, including **Prob. of Selecting MTD**, **Prob. of Toxicity**, **Prob. of Selecting Does-over-MTD**, **Average Number of Enrolled Patients**, **Prob. of Overdosing Allocation**, **Mean Squared Error**, and **Trial Duration**. (Figure 2.15).

Each item is explained next:

a. Line plots:

- The line plots display two summary statsitics, **Trial Duration (days)** and **Sum of Risky Decisions (%)**, for all the designs.
  - **Trial Duration (days)**: The average time (in days) for a trial. The lower the value, the

faster the trial and the more economic of the design.

- **Sum of Risky Decisions (%)**: The sum of maximum percentage of risky decisions across all the scenarios. A risky decision is a decision made by a rolling design that is different and more aggressive than the decision that would have been made by the mTPI-2 design if all the patients in the cohort had been completely followed and their outcomes observed. There are three types of risky decisions:

  1. DS, which refers to the risky decision of S, stay, taken by the rolling design based on incomplete follow-up data, when the mTPI-2 design would decide to D, de-escalate, if patients were to complete follow up and their outcomes were observed;
  2. DE, which refers to the risky decision of E, escalate, taken by the rolling design based on incomplete follow-up data, when the mTPI-2 design would decide to D, de-escalate, if patients were to complete follow up and their outcomes were observed;
  3. SE, which refers to the risky decision of E, escalate, taken by the rolling design based on incomplete follow-up data, when the mTPI-2 design would decide to S, stay, if patients were to complete follow up and their outcomes were observed.

  In all three cases, the rolling design will assign patients to a higher dose than the non-rolling mTPI-2 design, and therefore they are considered risky.

- For each line plot, the x-axis is the index of scenario and the y-axis is the value of summary statistics. Lines with different colors represent different designs.
- Results are interactive:
  - Hover the mouse on a dot and a box will display the value of each design at the corresponding scenario.
  - Hover the mouse on the design label to highlight the corresponding line and fade the others.
  - Click the design label to hide the corresponding line and click again to change it back.
  - Click the line chart or the bar plot icon on the top right to switch between line charts and bar plots.

b. Average trial duration and risky decisions table:

Figure 2.14 shows a screenshot summarizing the average trial duration and the maximum percentage of inconsistent decisions across all scenarios for two designs. An inconsistent decision refers to a decision that is different from what would be made by mTPI-2 if patients were to complete follow up and their outcomes were observed. There are six types of inconsistent decisions.

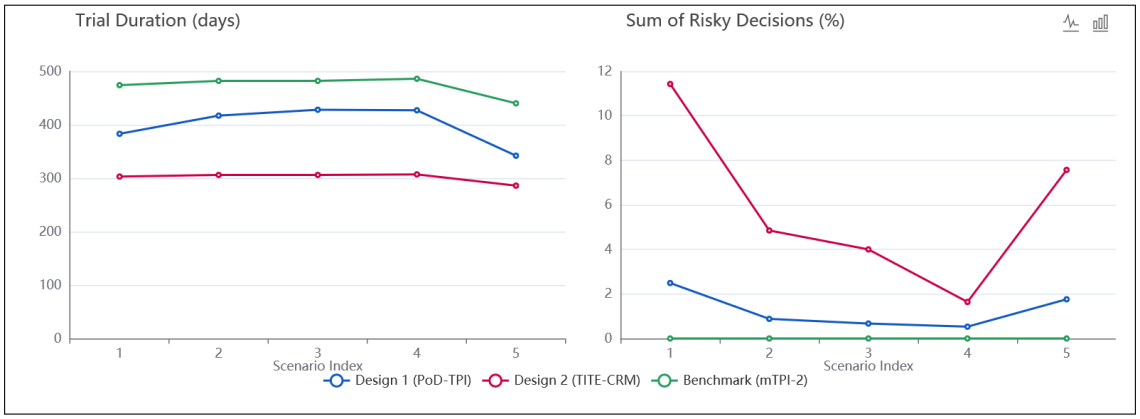


Figure 2.13: Simulation result Part A in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

DS, DE, and SE are the three risky and inconsistent decisions defined above. The column **Sum** in Figure 2.14 refer to sum of maximum percentage of DS, DE and SE decisions across all scenarios. A rolling design with smaller value is safer.

Average Trial Duration (days) and Risky Decisions (max % across scenarios)

Design	Risky Decisions				Duration (days)
	DS	DE	SE	Sum	
Design 1 (PoD-TPI)	2.5	0.0	0.0	2.5	399
Design 2 (TITE-CRM)	6.5	0.4	4.5	11.4	302
Benchmark (mTPI-2)	0.0	0.0	0.0	0.0	473

Figure 2.14: Average trial duration and risky decisions table in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

c. Simulation summary table:

Figure 2.15 shows the mean $\pm$ standard deviation of seven summary statistics across all scenarios for each design, as part of operating characteristics of the designs. They are explained in full detail next.

- **Prob. of Selecting MTD**: The probability of selecting the true MTD, defined as the proportion of simulated trials that correctly select the true MTD. The higher the value, the better the design.

  - For interval-based designs (mTPI-2 & R-TPI), the true MTDs are defined as the dose levels of which the true toxicity probabilities fall into the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$; if none of the doses have a toxicity probability that falls into the equivalence interval, the true MTD is defined as the dose with the highest toxicity probability below p_T. For the non-interval-based designs, 3+3 and Rolling 6, the true MTDs is defined as the dose levels with the highest toxicity probabilities lower than or equal to p_T.
  - To compare the operating characteristics of multiple designs submitted in a simulation study, the definition of MTD should be unified. If any of interval-based designs (mTPI-2 & R-TPI) are used in the simulation, the dose levels of which the true toxicity probabilities fall into the widest equivalence interval $[p_T - \max\{\epsilon_1\}, p_T + \max\{\epsilon_2\}]$ are defined as the true MTDs. Here, $\max\{\cdot\}$ is taken over the designs. If none of the doses fall in, the dose with the highest toxicity probability that is below p_T is the true MTD. For example, consider a case in which users compare four designs, R-TPI, mTPI-2, Rolling 6 and 3+3, in a simulation study targeting $p_T = 0.3$. Suppose $\epsilon_1 = 0.02$ and $\epsilon_2 = 0.05$ for R-TPI, and $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.03$ for mTPI-2. In this case, the true MTD is the dose levels with toxicity probabilities in $[0.3 - 0.05, 0.3 + 0.05]$; if none of the doses have a toxicity probability in $[0.3 - 0.05, 0.3 + 0.05]$, the dose with the highest toxicity probability lower than 0.3 is the true MTD.
  - If a scenario does not have any MTD (e.g., all doses have toxicity probabilities larger than the target p_T), no selection is the right decision. In this case, the probability of selecting the true MTD is the probability of no selection.

- **Prob. of Toxicity**: The proportion of patients who have experienced DLT across all the simulated trials. The lower the number, the fewer patients having DLTs under the design.
- **Prob. of Selecting Does-over-MTD**: The probability of selecting the dose levels above the true MTD, which is defined as the proportion of simulated trials that select a dose higher than the true MTD at the end of the trial. The lower the value, the better the safety of the design.
- **Average # of Enrolled Patients**: The average number of patients enrolled in the trial, including the patients who complete the DLT observation period with DLT or non-DLT, and patients who drop out of the trial and become inevaluable for DLTs.
- **Prob. of Overdosing Allocation**: The average proportion of patients who are assigned to doses higher than the MTD by the design across all the simulated trials.
- **Mean Squared Error**: The average mean squared error (MSE) in the toxicity probability of selected MTD, across all the simulated trials. The MSE is defined as the average squared

distance between the true toxicity probability of the selected dose, and the true toxicity probability of the true MTD for each scenario across the simulated trials. The scenarios with no true MTDs are excluded.

- **Trial Duration (days)**: The average time (in days) for a trial. The lower the value, the faster the trial and the more economic of the design.

Summary of Performance

	Design 1 (PoD-TPI)	Design 2 (TITE-CRM)	Benchmark (mTPI-2)
Prob. of Selecting MTD	0.438 ± 0.087	0.500 ± 0.099	0.452 ± 0.093
Prob. of Toxicity	0.232 ± 0.057	0.223 ± 0.049	0.240 ± 0.056
Prob. of Selecting Dose-over-MTD	0.144 ± 0.098	0.237 ± 0.130	0.140 ± 0.095
Average # of Enrolled Patients	26.659 ± 0.513	29.555 ± 0.725	26.535 ± 0.702
Prob. of Overdosing Allocation	0.151 ± 0.125	0.206 ± 0.116	0.164 ± 0.129
Mean Squared Error	0.008 ± 0.003	0.009 ± 0.005	0.008 ± 0.003
Trial Duration (days)	399 ± 33	302 ± 8	473 ± 17

* Mean ± Standard Deviation

Figure 2.15: Simulation summary table in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Part B: Tabulated Results by Scenarios

Full simulation results are presented in tabular format grouped by scenario(Figure 2.16).

In the upper part of Figure 2.16, the first two columns summarize dose levels and their true toxicity probabilities; the remaining columns report dose-specific summary statistics from the simulations including: selection probability, average number of patients treated, and average number of toxicities (i.e. DLTs), along with their standard deviations (s.d.), at each dose level. Specifically, these are:

1) **Selection Prob.**: The proportion of simulated trials that select each dose level as the MTD.
2) **Average # of Patients Treated (s.d.)**: The average number of patients treated at each dose level.
3) **Average # of Toxicities (s.d.)**: The average number of patients experienced DLT at each dose level.

The true MTD(s) of the scenario is(are) highlighted by the orange bar. For the definition of the true MTD in the simulation results, please refer to the definition of **Prob. of Selecting MTD** in **Part A**.

In the lower part of Figure 2.16, more trial-specific summary statistics are reported, including: **MTD Selection**, **Patient Assignment**, **Trial Toxicity**, **Trial Stopping**, **Trial Duration**, **Trial Sample Size**, **Trial Duration**, and **Accuracy of Selected MTD**.

- **MTD Selection**
  - **Prob. of Selecting MTD**: The proportion of simulated trials that select the true MTD at the end of the trial.
  - **Prob. of Selecting Does-over-MTD**: The proportion of simulated trials that select a dose higher than the true MTD at the end of the trial.
  - **Prob. of No Selection**: The proportion of simulated trials in which none of the dose levels are selected as the MTD. If a scenario does not have any MTD, this values is treated as the probability of selecting the true MTD.

  For detailed descriptions, please refer to **Part A**.
- **Patient Allocation**
  - **Prob. of Correct Allocation (s.d.)**: The average proportion of patients who are correctly assigned to the true MTD by the design across all the simulated trials and its standard deviation.
  - **Prob. of Overdosing Allocation (s.d.)**: The average proportion of patients who are assigned to doses higher than the MTD by the design across all the simulated trials and its standard deviation.

- **Trial Toxicity**
  - **Prob. of Toxicity**: The proportion of patients experiencing DLT across all the simulated trials. For detailed descriptions, please refer to **Part A**.
- **Trial Stopping**
  - **Prob. of Early Stopping Trial due to Safety Rule**: The proportion of simulated trials in which the trial is stopped because the first dose level shows unacceptable toxicity.
- **Trial Sample Size**
  - **Average # of Patients Treated (s.d.)**: The average number of patients treated in the simulated trials and its standard deviation. Due to early stopping, this number may be lower than n.
- **Trial Duration**
  - **Average Trial Duration (s.d.)**: The average time (in days) a trial and its standard deviation.
- **Accuracy of Selected MTD**
  - **Mean Squared Error**: The mean squared error is the average squared distance between the true toxicity probability of the selected dose and that of the true MTD across the simulations. If the scenario has no true MTD, N/A is displayed.
- **Risk**
  - **Sum of Risky Decisions (%)**: The sum of percentage of risky decisions compared to the decisions that would be made by mTPI-2 if complete outcomes were observed in each scenario. For detailed descriptions, please refer to **Part A**.

When calculating the standard deviation, we use n_{sim} as the denominator instead of $(n_{sim} - 1)$ in East Bayes.

Part B: Tabulated Results by Scenarios

Scenario 1

$p_T = 0.3$, $n_{sim} = 1000$		Selection Prob.			Average # of Patients Treated (s.d.)			Average # of Toxicities (s.d.)		
Dose Level	True Tox Prob.	Design 1 (PoD-TPI)	Design 2 (TITE-CRM)	Benchmark (mTPI-2)	Design 1 (PoD-TPI)	Design 2 (TITE-CRM)	Benchmark (mTPI-2)	Design 1 (PoD-TPI)	Design 2 (TITE-CRM)	Benchmark (mTPI-2)
1	0.15	0.364	0.193	0.376	10.431 (7.168)	12.445 (9.162)	9.283 (6.808)	10.431 (1.932)	12.445 (1.964)	9.283 (1.765)
2	0.3	0.486	0.538	0.469	11.527 (6.02)	9.813 (6.503)	11.879 (5.884)	11.527 (2.268)	9.813 (2.197)	11.879 (2.24)
3	0.45	0.137	0.216	0.127	4.215 (4.941)	5.433 (5.831)	4.777 (5.186)	4.215 (2.087)	5.433 (2.154)	4.777 (2.283)
4	0.6	0.007	0.036	0.014	0.614 (1.866)	1.465 (2.672)	0.796 (2.237)	0.614 (1.087)	1.465 (1.361)	0.796 (1.217)
5	0.75	0	0.005	0.001	0.031 (0.352)	0.557 (1.337)	0.041 (0.411)	0.031 (0.284)	0.557 (0.866)	0.041 (0.297)

		Design 1 (PoD-TPI)	Design 2 (TITE-CRM)	Benchmark (mTPI-2)
MTD Selection*	Prob. of Selecting MTD	0.486	0.538	0.469
	Prob. of Selecting Dose-over-MTD	0.144	0.257	0.142
	Prob. of No Selection	0.006	0.012	0.013
Patients Assignment	Prob. of Correct Allocation (s.d.)	0.428 (0.22)	0.327 (0.217)	0.44 (0.215)
	Prob. of Overdosing Allocation (s.d.)	0.181 (0.223)	0.248 (0.257)	0.208 (0.234)
Trial Toxicity	Prob. of Toxicity	0.274	0.258	0.284
Trial Stopping**	Prob. of Early Stopping Trial due to Safety Rule	0.003	0.012	0.008
Trial Sample Size	Average # of Patients Treated (s.d.)	26.818 (1.962)	29.713 (2.653)	26.776 (2.516)
Trial Duration	Average Trial Duration (s.d.)	383 (64)	303 (58)	474 (63)
Accuracy of Selected MTD	Mean Squared Error	0.012	0.014	0.013
Risk	Sum of Risky Decisions (%)	2.5	11.4	0.0

* The row with orange background color indicates the true MTD.

** For further details concerning Trial Stopping Rule, please refer to section 1.3 in the User Manual

Figure 2.16: Simulation result Part B in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Part C: Key Metrics for Dose Finding

Figure 2.17 includes plots showing six key summary statistics, **Prob. of Selecting MTD**, **Prob. of Toxicity**, **Prob. of Selecting Does-over-MTD**, **Average # of Enrolled Patients**, **Prob. of Overdosing Allocation**, **Mean Squared Error** for all designs. Their values are already reported in Part B. However, Part C provides a better visualization for enhanced user experiences.

Part D: Inconsistent Decisions Breakdowns

Simulation Results Part D (Figure 2.18) includes plots showing the percentage of the six inconsistent decisions, **DS**, **DE**, **SE**, **SD**, **ED**, and **ES** for all designs. In addition, SD, ED, ES refer to three overly conservative inconsistent decisions. The first letter is the decision that would be made by the mTPI-2 design if patients were to complete follow up, and the second letter is the decision taken by the rolling design instead.

Figure 2.17: Simulation result Part C in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

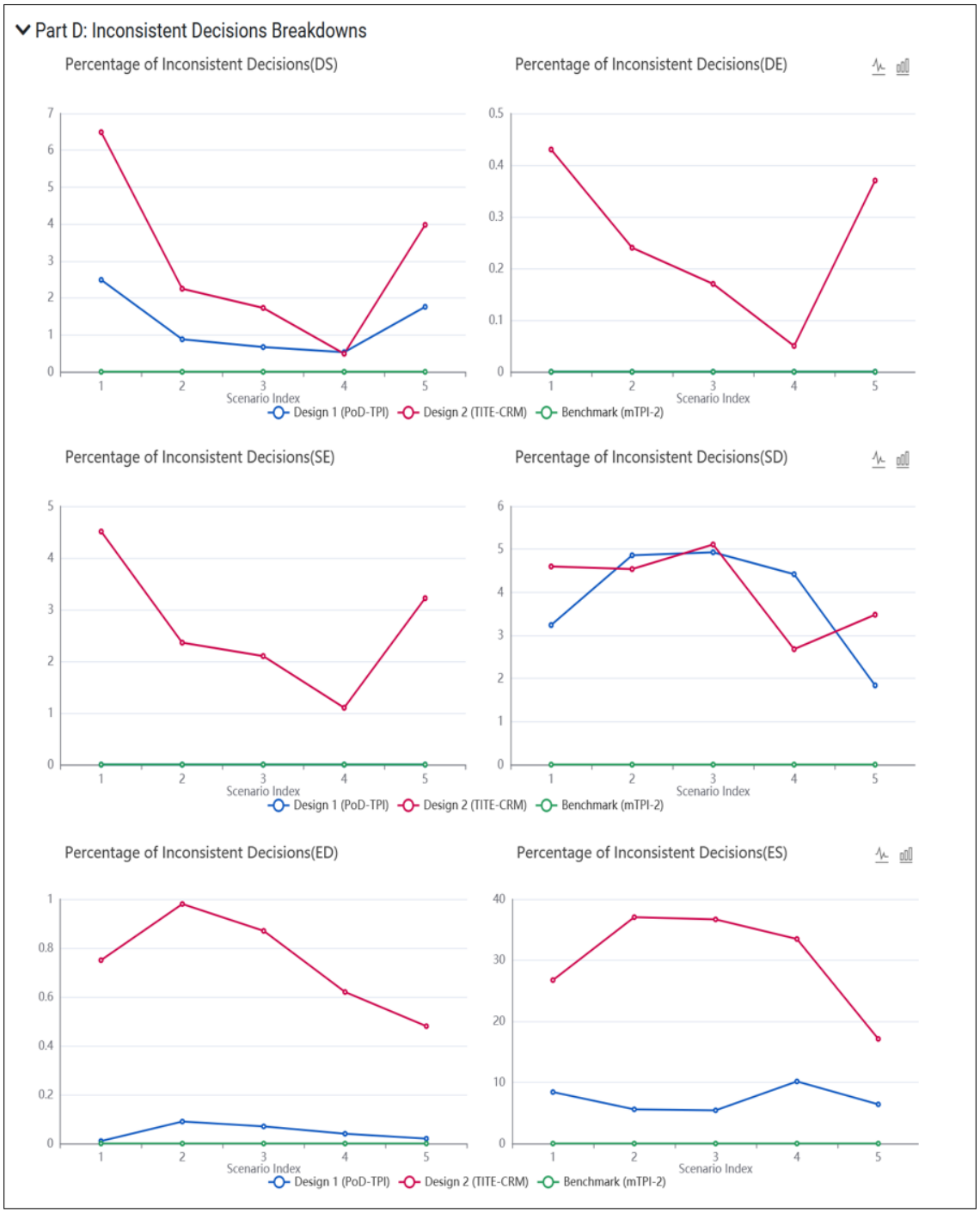


Figure 2.18: Simulation result Part D in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

2.2.3.2 Restore simulation

Users can restore the simulation settings from the simulation results by clicking the "Restore" button found to the right of each simulation in the results panel. Upon clicking, the display will switch to the **Simulation Setup** page with the same simulation settings restored.

2.2.3.3 Download simulation results

There is a "Download Report" button found to the right of each simulation in the results panel. Click it to download a word file, which includes three parts:

- Part A: Complete simulation results under the designs and scenarios users added in the Simulation Setup tab;
- Part B: Detailed technical descriptions of the designs users added in the Simulation Setup tab;
- Part C: Reference

These reports may be used to include in submissions reports or if more detailed work than what is offered is required, please contact us via email (support@cytel.com) for consulting services.

2.2.4 MTD Estimation

In the **MTD Estimation** tab, users can estimate the MTD for mTPI-2, R-TPI and PoD-TPI designs based on the isotonic regression through Pool Adjacent Violators Algorithm (PAVA), after the dose finding is completed and the DLT outcomes of all patients are collected.

Specify the target toxicity probability (p_T), and two small fractions to define the equivalence interval (ϵ_1 and ϵ_2) in the design. Select the number of doses (n_{dose}) from the dropdown box, and an editable table will appear (Figure 2.19). Manually type in the observed number of toxicities (DLTs) and the number of patients treated at each dose into the table, and click the "Estimate" button to estimate the MTD. Finally, the estimated MTD is highlighted in blue background as shown in Figure 2.20.

See Table 2.4 for detailed parameter descriptions.

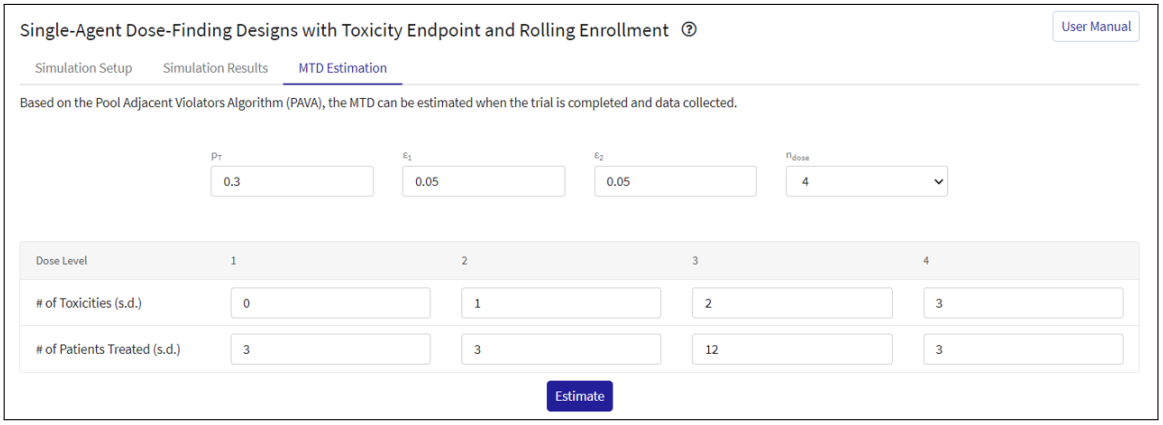


Figure 2.19: Input parameters in the **MTD Estimation** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

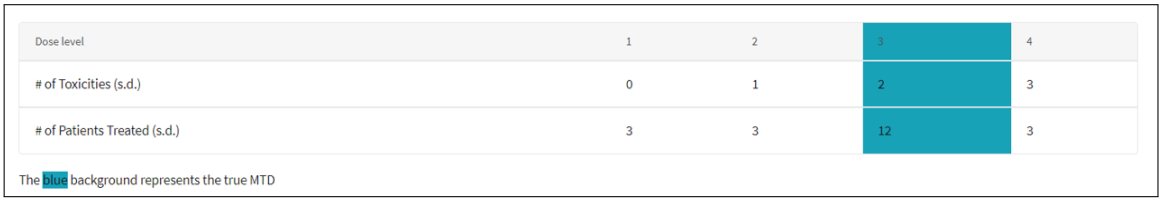


Figure 2.20: MTD estimation in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Rolling Enrollment** module.

Table 2.4: Input parameters in the **MTD Estimation** tab of **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Notation	Parameters	Description
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T.
ϵ_1,ϵ_2	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence/target interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose MTD. The default values for both are 0.05.
n_{dose}	The number of doses	The number of candidate dose levels for investigation.
# of DLTs	The number of patients with DLTs at each dose level	A non-negative integer number of patients with DLT at each dose level.
# of patients	The number of patients treated at each dose level	A positive integer number of patients treated at each dose level, which should be no less than the # of DLTs.

2.3 Statistical Methods Review

2.3.1 Simulating Patients Enrollment and Evaluation

To better demonstrate the benefit of rolling-based designs in accelerating the trial conduct, we assess trial duration in addition to safety and reliability of the designs. This module simulates trials based on practical settings in order to better reflect real-world situations.

In East Bayes, we fix $a = 1$ in sampling the inter-patient arrival time, so the mean inter-patient arrival time is MIAT $= b$. And also, for simplicity, we assume that there is no waiting time between the time of patient arrival in the clinic and the starting time of treatment, so the on-study start time is 0.

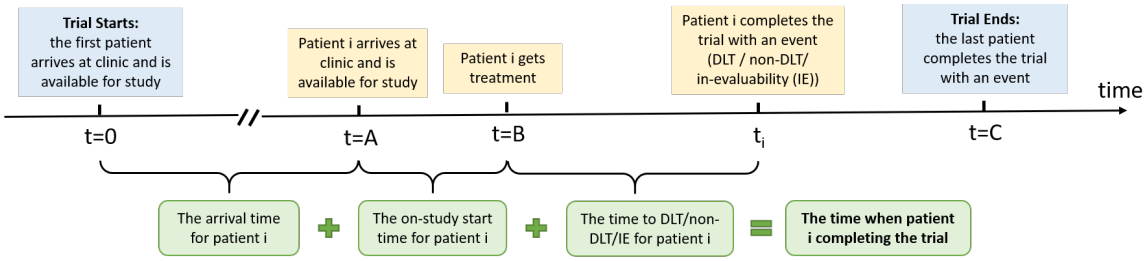


Figure 2.21: Simulating patients enrollment and evaluation in the ***Single Agent – Rolling-Based Designs***.

Figure 2.21 illustrates the simulation process of patients enrollment and evaluation. Specifically:

1. The trial enrollment assumes an inter-patient arrival time which is the average time (in days) between enrollment of two consecutive patients. The inter-patient arrival time is sampled from a gamma distribution, with the shape parameter a and scale parameter b. Therefore, the mean inter-patient arrival time (MIAT) is MIAT $= ab$. For example, a MIAT 10 or 5 days means, on average, every 10 or 5 days a new patient is eligible for enrollment, and hence the trial would enroll three or six patients per month, respectively.
2. To mimic real-life oncology dose-finding trials, each enrolled patient in the simulation study is also assigned an on-study start time (the gap between the time of arrival in the clinic and the starting time of treatment) and an inevaluable rate (such as drop off). Specifically,
   (a) A random binary DLT/non-DLT outcome is generated with the true probability of toxicity for the corresponding dose at which the patient is assigned.

(b) A random binary evaluability/inevaluability outcome is generated with an inevaluable rate (IR) for the enrolled patient.

(c) An on-study start time is sampled from a uniform distribution ranging from 0 to the maximum waiting time, where the maximum waiting time is prespecified.

(d) If a DLT occurs for a patient, the time to DLT is sampled from the uniform distribution ranging from 0 to the maximum DLT follow-up period $T_{\text{follow-up}}$;

(e) If no DLT occurs for a patient, the time to non-DLT is set at $T_{\text{follow-up}}$.

(f) If a patient becomes inevaluable, the time to inevaluability (IE) of that patient is sampled from a uniform distribution ranging from 0 to the sampled time to event (either DLT or non-DLT) of that patient.

Therefore, assume that the trial starts at time $t = 0$ (i.e., the first patient arrives and is available for study at time $t = 0$), a patient completes the trial with one of the three events: DLT, non-DLT, or IE, at time $t_i =$ arrival time $+$ on-study start time $+$ time to DLT or non-DLT or IE.

2.3.2 The 3+3 Design

The 3+3 design (Storer, 1989) is a rule-based design which enrolls patients in a cohort of three. It starts by allocating the first cohort of three patients to the starting dose (which is often the lowest dose level) and adaptively escalates/de-escalates to the next dose level based on observed number of dose limiting toxicities (DLTs). Besides, in this module, the ethics constraint of "decision-in-advance" (§2.3.2.2) is adopted, which is applicable to the real-life trials.

2.3.2.1 Design Algorithm

In 3+3, a maximum of six patients are allowed to be treated at any dose level, and the MTD is defined as the highest dose for which one or fewer DLTs occurred in six patients. Its algorithm proceeds as follows:

0. Start the trial by treating three patients at a prespecified starting dose level.
1. Escalate to the next higher dose or de-escalate to the next lower dose according to the following rules:
   (a) If 0 of 3 patients has a DLT, escalate to next higher dose and treat three patients.
   (b) If 2 or more of 3 patients have DLTs, de-escalate to next lower dose and treat three patients.
   (c) If 1 of 3 patients has a DLT, treat three more patients at current dose level.
      i. If 1 of 6 has DLT, escalate to next higher dose and treat three patients if the next higher dose has not been tried; otherwise, declare it as the MTD and stop the trial.
      ii. If 2 or more of 6 have DLTs, de-escalate to next lower dose level and treat three patients.
   (d) If the trial de-escalates to next lower dose:
      i. If only 3 or less had been treated at the next lower dose, treat three more patients at that dose.
      ii. If six have already been treated at the next lower dose, stop the trial and declare the lower dose as the MTD.
2. Escalation never occurs at a dose at which two or more DLTs have already occurred.
3. If de-escalation occurs at the lowest dose, the trial is stopped.
4. Repeat steps 1-3 until either the MTD is identified or the trial is stopped for excessive toxicity.

2.3.2.2 The "Decision-in-Advance" Rule

When the observed data from existing patients at a dose lead to a definitive decision regardless what happens to the patients to be enrolled or still under follow-up at the dose, the decision is executed immediately without the need to wait. For example, under 3+3, if 2 patients have been enrolled to a newly tested dose d and both of them experience DLTs, stop enrollment at dose d, de-escalate to $(d-1)$ immediately, and start enrolling patients at $(d-1)$. This rule of "decision-in-advance" can accelerate the trial conduct and shorten trial duration.

2.3.3 The Modified Toxicity Probability Interval-2 (mTPI-2) Design

The modified toxicity probability interval-2 (mTPI-2) design (Guo et al., 2017b) is a cohort-based design which enrolls patients according to a pre-planned cohort size. It is also a model-base design, which uses a simple beta-binomial model to estimate the toxicity probability and makes dose escalation/de-escalation decisions based on the unit probability mass (UPM) of a series of dosing interval with equal length. At the end, mTPI-2 selects the dose of which the isotonic transformed toxicity probability is the closest to the target p_T as the MTD. In this module, the "decision-in-advance" (§2.3.3.3) is adopted for mTPI-2 to speed up the trial conduct.

2.3.3.1 Probability Model

Consider a phase I trial with D candidate doses for escalation. Let $p_1, \ldots, p_D$ denote the true toxicity probabilities for doses $d = 1, \ldots, D$. The observed data include n_d, the number of patients treated at dose d, and y_d, the number of patients experiencing a toxicity. Let $Data = \{(y_d, n_d); d = 1, 2, \ldots, D\}$.

The mTPI-2 design employs a simple beta-binomial hierarchical model as follow:

$$y_d \mid n_d, p_d \sim binomial(n_d, p_d)$$
$$p_d \sim beta(\alpha, \beta)$$

The posterior distribution of p_d is given by

$$p_d \mid y_d, n_d \sim beta(\alpha + y_d, \beta + n_d - y_d). \tag{2.1}$$

We adopt the prior $beta(1, 1)$ for p_d, which leads to a slightly conservative posterior inference as the prior mean is 0.5, which is usually above p_T.

2.3.3.2 Dose-Finding Rules

Equal-width Dosing Intervals: The mTPI-2 design improves over the mTPI design (Ji et al., 2010) by blunting the Ockhams razor that leads to some statistically sound but practically debatable decisions in te mTPI design. In mTPI, the unit interval $(0, 1)$ is divided into three subintervals: the under-dosing interval $(0, p_T - \epsilon_1)$, the equivalence interval $(p_T - \epsilon_1, p_T + \epsilon_2)$, and the over-dosing interval $(p_T + \epsilon_2, 1)$. Here, ϵ_1 and ϵ_2 are small fractions, such as 0.05, to account for the uncertainty around the true target toxicity p_T. However, mTPI-2 resolves the Ockhams razor problem fundamentally by dividing the intervals $(0, p_T - \epsilon_1)$ and $(p_T + \epsilon_2, 1)$ into shorter subintervals with

length $(\epsilon_1 + \epsilon_2)$; which is the same as the length of the equivalence interval, to mitigate the effect of interval length in the mTPI design. Formulaically described below, we denote *EI* the equivalence interval $(p_T - \epsilon_1, p_T + \epsilon_2)$, and *LI* a set of intervals *EI* as:

$$\begin{aligned} LI =&\{M_1^{LI} = (p_T - 2\epsilon_1 - \epsilon_2, p_T - \epsilon_1), M_2^{LI} = (p_T - 3\epsilon_1 - 2\epsilon_2, p_T - 2\epsilon_1 - \epsilon_2), \cdots, \\ & M_J^{LI} = (0, p_T - J\epsilon_1 - (J-1)\epsilon_2)\}, \end{aligned}$$

and *HI* a set of intervals above *EI*:

$$\begin{aligned} HI =&\{M_1^{HI} = (p_T + \epsilon_2, p_T + \epsilon_1 + 2\epsilon_2), M_2^{HI} = (p_T + \epsilon_1 + 2\epsilon_2, p_T + 2\epsilon_1 + 3\epsilon_2), \cdots, \\ & M_K^{HI} = (p_T + (K-1)\epsilon_1 + K\epsilon_2, 1)\}. \end{aligned}$$

Therefore, if $p_T = 0.3$ and $\epsilon_1 = \epsilon_2 = 0.05$,

$$\begin{aligned} EI =&(0.25, 0.35) \\ LI =&\{M_1^{LI} = (0.15, 0.25), M_2^{LI} = (0.05, 0.15), M_3^{LI} = (0, 0.05)\}, \\ HI =&\{M_1^{HI} = (0.35, 0.45), M_2^{HI} = (0.45, 0.55), M_3^{HI} = (0.55, 0.65), M_4^{HI} = (0.65, 0.75), \\ & M_5^{HI} = (0.75, 0.85), M_6^{HI} = (0.85, 0.95), M_7^{HI} = (0.95, 1)\}. \end{aligned}$$

Other than the boundaries $(0, 0.05)$ and $(0.95, 1)$, all the intervals have the same length. The boundaries do not affect the decision making since they are clearly associated with "E" and "D" decisions, respectively. See Guo et al. (2017b) for details.

Dose-Finding Rules: Given the interval and a probability distribution like (2.1), define the unit probability mass (UPM) of that interval as the probability of the interval divided by the length of the interval. Mathemetically, the UPM of an interval (a, b) equals to

$$\text{UPM} = \frac{Prob\{p \in (a, b) \mid Data\}}{b - a}$$

The mTPI-2 design selects the (sub-)interval with the largest UPM value as the winning interval and takes the dose-escalation decision corresponding to the winning (sub-)interval. More specifically:

- If the equivalence interval $M^{EI} = (p_T - \epsilon_1, p_T + \epsilon_2)$ has the largest UPM, it is selected as the winning interval and the dose-assignment decision of mTPI-2 is "S", to stay at the current dose.
- If any interval M_j^{LI} in *LI* has the largest UPM, it is selected as the winning interval and the dose-assignment decision of mTPI-2 is "E", to escalate to the next higher dose.
- If any interval M_k^{HI} in *HI* has the largest UPM, it is selected as the winning interval and the dose-assignment decision of mTPI-2 is "D", to de-escalate to the next lower dose.

2.3.3.3 The "Decision-in-Advance" Rule

When the observed data from existing patients at a dose lead to a definitive decision regardless what happens to the patients to be enrolled or still under follow-up at the dose, the decision is executed immediately without the need to wait. For example, under mTPI-2, if 2 patients have been enrolled to a newly tested dose d and both of them experience DLTs, stop enrollment at dose d, de-escalate to $(d-1)$ immediately, and start enrolling patients at $(d-1)$. This rule of "decision-in-advance" can accelerate the trial conduct and shorten trial duration.

2.3.3.4 Safety Rules

For trial safety, two additional rules are applied.

- [*Rule 1: Dose Exclusion*] If the current dose is considered excessively toxic, i.e., $n_d \geq 3$ and $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses are excluded and never used again in the remainder of the trial.
- [*Rule 2: Early Stop*] If the current dose is the lowest dose (first dose) and is considered excessively toxic according to Rule 1, early stop the trial for safety.

In safety Rules 1 and 2, $Prob\{p_d > p_T \mid Data\}$ is calculated under the beta distribution in (2.1).

2.3.3.5 Trial Termination

The trial proceeds until any of the following conditions is satisfied:

1. If the prespecified maximum total sample size is reached;
2. If the lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is early stopped and the MTD cannot be determined;
3. *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by $K(K < n)$:*
   - *If the mTPI-2 decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *If the mTPI-2 decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *If the mTPI-2 decision is "D", to de-escalate to the next lower dose, and that next lower dose has enrolled K patients.*

2.3.3.6 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped early, the mTPI-2 design applies an isotonic regression to select the MTD. Following the steps as below:

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. An independent prior $beta(0.005, 0.005)$ is used to compute the posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior mean by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA), the estimated DLT probabilities satisfying the order constraint is obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.
2. Among all the tried doses for which $Prob\{p_d > p_T \mid Data\} < \xi$ and $\hat{p}_d \leq p_T + \epsilon_2$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d\, |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference),
   (a) If there is at least one dose lower than the target p_T among all the tied doses, choose the highest dose among those as the estimated MTD;
   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

2.3.4 The Rolling 6 Design

The Rolling 6 design (Skolnik et al., 2008) extends the 3+3 design with the aim to reduce the occurrence of accrual suspension after enrolling each set of three patients, thereby accelerating the trial. It allows for accrual of two to six patients concurrently onto a dose level based on the number of patients concurrently enrolled and evaluable (# Enrolled), the number experiencing dose-limiting toxicity (DLT) (# DLTs), and the number still at risk of developing a DLT (# Pending). The Rolling 6 is a rule-based design and all the dose assignment rules for the six patients are pretabulated in Table 2.5.

Table 2.5: The decision table of the Rolling 6 design.

	Observed data at dose d			Decision	
# Enrolled	# DLTs	# Non-DLTs	# Pending	MTD Not Exceeded	MTD Exceeded
2	0, 1	any	any	S	-
2	2	0	0	D	-
3	0	0, 1, 2	3, 2, 1	S	-
3	0	3	0	E	-
3	1	0, 1, 2	2, 1, 0	S	-
3	≥ 2	any	any	D	-
4	0	0,1,2,3	4,3,2,1	S	S
4	0	4	0	E	S
4	1	0,1,2,3	3,2,1,0	S	S
4	≥ 2	any	any	D	D
5	0	0,1,2,3,4	5,4,3,2,1	S	S
5	0	5	0	E	S
5	1	0,1,2,3,4	4,3,2,1,0	S	S
5	≥ 2	any	any	D	D
6	0	0,1,2,3,4	6,5,4,3,2	Suspend	Suspend
6	0	5,6	1,0	E	MTD
6	1	0,1,2,3,4	5,4,3,2,1	Suspend	Suspend
6	1	5	0	E	MTD
6	≥ 2	any	any	D	D

NOTE. 1) This table does not take into account inevaluable patients, such as patients who drop off during the DLT observation period; 2) Escalation never occurs to a dose at which 2 or more DLTs have already occurred, because the dose is considered excessively toxic and should be excluded from the remaining dose finding; 3) If de-escalation occurs at the lowest dose level, then the study is terminated.

ABBREVIATIONS: DLT, dose-limiting toxicity; MTD, maximum tolerated dose; E, escalate to next higher dose level; S, stay at the current dose level; D, de-escalate to previous lower dose level.

2.3.5 The Rolling Toxicity Probability Interval (R-TPI) Design

This section describes the rolling toxicity probability interval (R-TPI) design proposed by Guo et al. (2019). R-TPI combines the idea of rolling accrual in the Rolling 6 design (Skolnik et al., 2008) (§2.3.4) with the model-based framework in mTPI-2 (Guo et al., 2017b) (§2.3.3).

2.3.5.1 Notations

Consider a toxicity-driven phase I dose-finding trial. Let p_T be the target DLT probability, and p_d be the true and unknown DLT probabilities of dose level d, $d = 1, \ldots, D$, where D denotes the prespecified number of dose levels to be investigated. Generally, we assume that p_d is non-decreasing with dose level, i.e. $p_1 \leq p_2 \leq \cdots \leq p_D$. Assume at a given moment, dose d is being used to treat enrolled patients and a total of $(n_d + m_d)$ patients have been assigned to dose d, among whom n_d patients have known outcomes (either with or without DLT) and m_d patients are still being followed without outcomes. Let y_d be the number of patients (among n_d) with DLT, therefore $(n_d - y_d)$ without DLT. The table below describes the breakdowns.

# with DLT	# without DLT	# being followed and no outcomes	Total at dose d
y_d	$(n_d - y_d)$	m_d	$(n_d + m_d)$

2.3.5.2 Dose-Finding Rules and Design Algorithm

The R-TPI design consists of two sets of enrollment schemes, namely the run-in enrollment and the rolling enrollment. To begin the trial, R-TPI enrolls the first patient at the starting dose level.

Run-in Enrollment The run-in enrollment is applied to any new dose level when it is first used to treat patients during the trial. Suppose dose d is decided to be the new dose level for treating patients and it has not been used at any time of the trial. R-TPI starts run-in enrollment and keeps enrolling new patients at dose d until either of the two cases below occurs:

(1) $n_d > 0$, i.e. there is at least one outcome at d,

(2) $n_d = 0$ and $m_d = C$, for a pre-determined C value. This occurs when the first C patients have not completed follow-up at d and are without definitive outcomes. Here, C is the maximum number of pending patients without observed outcomes allowed at a dose to keep enrollment open. For example, for the Rolling 6 design, $C = 6$.

Therefore,

– in case (1), R-TPI starts rolling enrollment (specified below).

– in case (2), R-TPI suspends the enrollment until the first outcome at current dose d and then starts the rolling enrollment (specified below).

Rolling Enrollment Supposes at a given moment of the trial a new patient becomes eligible for enrollment, and the current dose used for treating patients is d at which $(n_d + m_d)$ patients have been treated. To fully understand rolling enrollment, there are two additional points to consider.

- k_d: the number of patients at dose d since it most recently becomes the current dose.
  For example, suppose initially three patients were enrolled at dose level d, and based on their DLT outcomes R-TPI escalated to dose level $(d+1)$ and enrolled patients at $(d+1)$; however, many patients had DLT outcomes at $(d+1)$ and R-TPI de-escalated back to dose level d, and enrolled additional 3 patients. At this time $k_d = 3$.
- $\mathcal{D}_{y_d,n_d}$: the mTPI-2 decision based on the toxicity data of y_d out of n_d patients experiencing DLTs at dose d, $\mathcal{D}_{y_d,n_d} \in \{D, E, S\}$. Here, "D" stands for de-escalating to the next lower dose level $(d-1)$, "E" for escalating to the next higher dose level $(d+1)$, and "S" for staying at the current dose level d. For the detailed mTPI-2 dose escalation rules, please refer to §2.3.3.2.

The dose-assignment of R-TPI assesses three potential decisions 1) the mTPI-2 decision, denoted as $\mathcal{D}_{y_d,n_d}$, based on the observed data; 2) the mTPI-2 decision $\mathcal{D}_{y_d+m_d,n_d+m_d}$ of the most

Table 2.6: The R-TPI dose-finding rules applied in the rolling enrollment.

	mTPI-2 decision for current observation $(\mathcal{D}_{y_d,n_d})$	mTPI-2 decision for the most toxic scenario $(\mathcal{D}_{y_d+m_d,n_d+m_d})$	mTPI-2 decision for the safest scenario $(\mathcal{D}_{y_d,n_d+m_d})$	R-TPI Decision
Case 1	D	D	D	D
Case 2	D	D	S or E	S
Case 3	S	S or D	S	S
Case 4	S	S or D	E	S or Suspend*
Case 5	E	E	E	E
Case 6	E	S or D	E	S or Suspend*

* If 3 or more patients have been enrolled at the same dose $(k_d > 3)$, suspend the trial to avoid over-enrolling on the current dose.

Abbreviations: E, escalate to next higher dose level; S, stay at the current dose level; D, de-escalate to previous lower dose level.

toxic possible scenario where all pending patients were to experience DLTs; and 3) the mTPI-2 decision $\mathcal{D}_{y_d,n_d+m_d}$ of the safest scenario where none of pending patients were to experience DLT. See Table 2.6. Specifically, suppose a new patient is eligible for enrollment, the detailed rolling enrollment rules are described below.

I. If $m_d = 0$, i.e., all the patients enrolled at dose level d have completed their followup with definitive outcomes, assign the new patient according to $\mathcal{D}_{y_d,n_d}$, the decision of mTPI-2 when y_d out of n_d patients experience DLT outcomes.

II. If $0 < m_d \leq C$, i.e., some patients are still being followed without outcomes, consider three cases:

1. If $\mathcal{D}_{y_d,n_d}$ is D, consider the following two cases:
   (a) if $\mathcal{D}_{y_d,n_d+m_d}$ is D, de-escalate to dose level $(d-1)$; apply the run-in enrollment if dose $(d-1)$ is a new dose or re-apply **I/II/III** if it has been used before;
   (b) else, the decision is S and continue patient enrollment at dose d.
2. If $\mathcal{D}_{y_d,n_d}$ is S, consider the following two cases:
   (a) if $\mathcal{D}_{y_d,n_d+m_d}$ is S, assign the new patient to d;
   (b) if $\mathcal{D}_{y_d,n_d+m_d}$ is E,
      i. if $k_d < 3$, enroll the next patient at dose d;
      ii. if $k_d \geq 3$, suspend the enrollment until more patients have observed their outcomes at dose d. Then recalculate the m_d value and re-apply **I** or **II.**
3. If $\mathcal{D}_{y_d,n_d}$ is E, consider the following two cases:
   (a) if $\mathcal{D}_{y_d+m_d,n_d+m_d}$ is E, escalate to dose level $(d+1)$; apply the run-in enrollment if dose $(d+1)$ is a new dose or re-apply **I/II/III** if it has been used before.
   (b) else,
      i. if $k_d < 3$, enroll the next patient to dose d;
      ii. if $k_d \geq 3$, suspend the enrollment until more patients have observed their outcomes at dose d. Then recalculate the m_d value and re-apply **I** or **II.**

III. If $m_d > C$, suspend the enrollment until more patients have observed outcomes at dose d.

2.3.5.3 Safety Rules

For trial safety, two additional rules are applied.

- [*Rule 1: Dose Exclusion*] If the current dose is considered excessively toxic, i.e., $(n_d+m_d) \geq 3$ and $Prob\{p_d > p_T \mid Data\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher doses are excluded and never be used again in the remainder of the trial.
- [*Rule 2: Early Stop*] If the current dose is the lowest dose (first dose) and is considered excessively toxic according to Rule 1, stop the trial for safety.

Here, $Prob\{p_d > p_T \mid Data\}$ is calculated under the beta distribution $Beta(\alpha_0+y_d, \beta_0+n_d-y_d)$. In East Bayes, we use $\alpha_0 = \beta_0 = 1$. For the rolling designs, at the time a dose is deemed unsafe and suspended, there may be some patients with pending outcomes at this dose level. Once their data are observed later, if the safety rule is no longer violated given the new data, this dose may be reopened again for further evaluation.

2.3.5.4 Trial Termination

The R-TPI design stops the trial if any of the following conditions is satisfied:

1. The prespecified maximum total sample size is reached;
2. The lowest dose shows excessive toxicity according to Rule 2; In this case, the trial is stopped and the MTD cannot be determined;
3. *Optional: ad-hoc rules of maximum number of patients at a dose, denoted by $K(K < n)$:*
   - *If the mTPI-2 decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *If the mTPI-2 decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *If the mTPI-2 decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

2.3.5.5 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped due to rule 2 in §2.3.5.3, the R-TPI design applies an isotonic regression to select the MTD. The steps below are followed.

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.

(a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. An independent prior $beta(0.005, 0.005)$ is used to compute each posterior mean and variance.

(b) Compute isotonic regression estimates of the posterior means by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA). The estimated posterior mean DLT probabilities satisfying the order constraint are obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.

2. Among all the tried doses ($n_d > 0$) for which $Prob\{p_d > p_T \mid Data\} < \xi$ and $\hat{p}_d \leq p_T + \epsilon_2$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d\, |\hat{p}_d - p_T|$.

3. In case of a tie (i.e., two or more doses have the smallest difference),

   (a) If there is at least one dose lower than the target p_T among all the tied doses, choose the highest dose among the ones lower than p_T as the estimated MTD;

   (b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

2.3.5.6 R-TPI Decision Table

The R-TPI design requires users to provide the value of the target toxicity rate p_T and two small fractions, ϵ_1 and ϵ_2. The p_T value can be easily elicited from the trial clinician. The values of ϵ_1 and ϵ_2 can be set at 0.05 as the default (Ji et al., 2010) or elicited by asking the clinician the lower and higher bound of the DLT rate that would still be considered as close to p_T. Also R-TPI needs to elicit the value of C to control the speed of patient accrual. With the provided values of p_T, ϵ_1, ϵ_2, and C, one can generate the R-TPI decision table prior to the trial. Therefore, R-TPI exhibits the same simplicity and transparency as rule-based methods.

We provide the decision table of up to seven patients for R-TPI with target DLT rate p_T equal to 0.3, $\epsilon_1 = \epsilon_2 = 0.05$, and $C = 3$, as an example. See Table 2.7.

Table 2.7: R-PTI Decision Table with $p_T = 0.3$, $\epsilon_1 = \epsilon_2 = 0.05$, and $C = 3$.

Observed data at dose d				R-TPI
$n_d + m_d$	y_d	n_d	k_d	Decision
1	0	0	1	S
1	0	1	1	E
1	1	1	1	D
2	0	0,1	any	S
2	0	2	any	E
2	> 0	any	any	D
3	0	0,1,2	3	Suspend
3	0	1,2	< 3	S
3	0	3	any	E
3	1	any	any	S
3	> 1	any	any	D
4	0	1,2,3	3	Suspend
4	0	2,3	< 3	S
4	0	4	any	E
4	1	any	any	S
4	> 1	any	any	D
5	0	2,3,4	3	Suspend
5	0	3,4	< 3	S
5	0,1	5	any	E
5	1	3,4	≥ 3	Suspend
5	1	3,4	< 3	S
5	> 1	any	any	D
6	0	3,4	3	Suspend
6	0	4	< 3	S
6	0	5	any	E
6	1	3,4,5	3	Suspend
6	1	4, 5	< 3	S
6	0,1	6	any	E
6	2	any	any	S
6	> 2	any	any	D

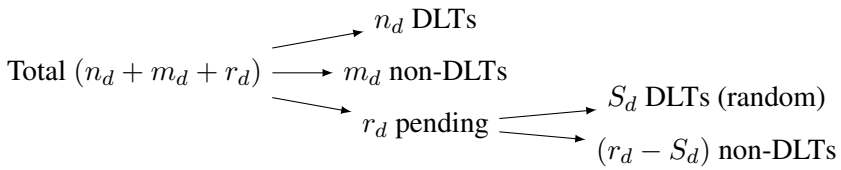
2.3.6 The Probability-of-Decision Toxicity Probability Interval Design (PoD-TPI)

The PoD-TPI design (Zhou et al., 2019a) is motivated by the need to reduce the frequency of enrollment suspension but while maintaining safety. PoD-TPI enables dose assignment in real time in the presence of pending toxicity outcomes. With uncertain outcomes, the dose assignment decisions are treated as a random variable, and the posterior distribution of the decisions can be calculated. The posterior distribution reflects the variability in the pending outcomes and allows a direct and intuitive evaluation of the confidence of all possible decisions. A new and useful feature of PoD-TPI is that it allows investigators and regulators to balance the trade-off between enrollment speed and making risky decisions by tuning a pair of intuitive design parameters.

2.3.6.1 Notations

Consider a toxicity-driven phase I dose-finding trial. Let p_T be the target DLT probability, EI $= (p_T - \epsilon_1, p_T + \epsilon_2)$ be the equivalence interval, and p_d be the true and unknown DLT probabilities of dose level d, $d = 1, \ldots, D$, where D denotes the prespecified number of dose levels to be investigated. Generally, we assume that p_d is non-decreasing with dose level, i.e. $p_1 \leq p_2 \leq \cdots \leq p_D$.

At a given moment of the trial, suppose N patients have been treated, the current dose is d, and the $(N+1)$-th patient is available for enrollment. Recall that Y_i and Z_i denote the DLT outcome and dose assignment of patient i, respectively, $i = 1, \ldots, N$. In particular, $Y_i = 1$ (or 0) represents that patient i experiences (or does not experience) DLT within the assessment window. Since patients enter clinical trials at random time, it is often the case that when the $(N+1)$-th patient is eligible for enrollment, some previously enrolled patients are still being followed without definitive DLT outcomes, and thus their DLT outcomes Y_i's are unknown. Let B_i be the indicator for an unknown DLT outcome, where $B_i = 1$ (or 0) denotes that the DLT outcome of patient i is unknown (or observed). We denote $n_d = \sum_{i=1}^{N} \mathbb{1}(Y_i = 1, Z_i = d, B_i = 0)$ and $m_d = \sum_{i=1}^{N} \mathbb{1}(Y_i = 0, Z_i = d, B_i = 0)$ the numbers of patients with observed DLTs and non-DLTs, respectively. In addition, we use $r_d = \sum_{i=1}^{N} \mathbb{1}(Z_i = d, B_i = 1)$ to denote the number of patients with pending outcomes and write $\mathcal{I}_d = \{i : Z_i = d, B_i = 1\}$ the index set of these patients. Lastly, we denote $S_d = \sum_{i=1}^{N} \mathbb{1}(Y_i = 1, Z_i = d, B_i = 1)$ the number of DLTs among the r_d pending patients that would have been observed had these patients finished their DLT assessment. Since these patients are still being followed, $\{Y_i : i \in \mathcal{I}_d\}$ are not observed and are random variables, and so are S_d. We have $S_d \in \{0, 1, \ldots, r_d\}$. The following figure summarizes the patient statistics at dose d.



2.3.6.2 Dose Assignment Rules

Suppose p_T, ϵ_1 and ϵ_2 are given and fixed. When there are no patients with pending outcomes, i.e. $r_d = 0$, the dose-finding decision A_d of PoD-TPI is the same as that of mTPI-2 (§2.3.3), which only depends on the values n_d and m_d. Let $\mathcal{A}(n, m)$ denote the decision of mTPI-2 if n patients have DLT and m patients have non-DLT. Therefore, $\mathcal{A} = -1, 0$ and 1, denoting de-escalation, stay and escalation, respectively. When $r_d = 0$, $A_d = \mathcal{A}(n_d, m_d)$.

However, in most cases, $r_d \neq 0$ and S_d is not observed later, and the decision A_d becomes a random variable. Through the probability model described in §2.3.6.6, one could calculate the posterior probability $\Pr(S_d = s \mid Data)$, and then the probability of decision $a \in \{-1, 0, 1\}$ based on mTPI-2 can be defined by

$$\Pr(A_d = a \mid Data) = \sum_{s:\mathcal{A}(n_d+s, m_d+r_d-s)=a} \Pr(S_d = s \mid Data). \tag{2.2}$$

Let $A_d^* = \operatorname{argmax}_a \Pr(A_d = a \mid Data)$ denote the decision with the highest PoD. If two decisions tie for the highest PoD, we choose the more conservative one (the smaller value a).

To ensure the safety of the design, we introduce two essential suspension rules.

- If $A_d^* = 1$, i.e. escalation, we suspend the trial if (i) $\Pr(A_d = 1 \mid Data) < \pi_{\mathrm{E}}$ for some threshold $\pi_{\mathrm{E}} \in [0.33, 1]$ or (ii) $m_d = 0$. Condition (i) reflects that escalation is not allowed if the confidence of escalation is less than π_{E}. A larger π_{E} represents more conservative dose escalations. Condition (ii) means escalation is not allowed until at least one patient at the current dose has finished the DLT assessment and does not experience DLT, similar to the rule in Normolle and Lawrence (2006).

- If $A_d^* = 0$, i.e. stay, we suspend the trial if $\Pr(A_d = -1 \mid Data) > \pi_{\mathrm{D}}$ for some threshold $\pi_{\mathrm{D}} \in [0, 0.5]$. This means stay is not allowed if there is a relatively high chance of de-escalation. A smaller π_{D} represents more conservative stays.

If none of the suspension rule is triggered, the optimal decision A_d^* is made. In real applications, the values π_{E} and π_{D} should be chosen according to the desired extent of safety. For example,

$\pi_{\mathrm{E}} = 1$ and $\pi_{\mathrm{D}} = 0.15$ mean eliminating the chance of risky escalations. The dose assignment rules of PoD-TPI is summarized in Algorithm 1.

Algorithm 1 Dose assignment rule of PoD-TPI. Current dose level is d.

if $r_d = 0$ **then**
  Assign the patient to dose $d + \mathcal{A}(n_d, m_d)$
else if $r_d > 0$ **then**
  Calculate $\Pr(A_d = a \mid Data)$ and $A_d^* = \min\{\mathrm{argmax}_a \Pr(A_d = a \mid Data)\}$
  if $A_d^* = 1$ **then**
    if $\Pr(A_d = 1 \mid Data) < \pi_{\mathrm{E}}$ **or** $m_d = 0$ **then**
      Suspend accrual
    else
      Assign the patient to $(d + 1)$
    end if
  else if $A_d^* = 0$ **then**
    if $\Pr(A_d = -1 \mid Data) > \pi_{\mathrm{D}}$ **then**
      Suspend accrual
    else
      Assign the patient to d
    end if
  else if $A_d^* = -1$ **then**
    Assign the patient to $(d - 1)$
  end if
end if

If d is the highest dose, escalation is not possible and continue to enroll patients at the current dose d. Similarly, if d is the lowest dose, de-escalation is not possible and continue to enroll patients at d.

2.3.6.3 Safety Rules

For practical concerns, similar to existing designs (for example, Ji et al., 2010 and Yuan et al., 2018), we include the following two safety rules in PoD-TPI throughout the trial.

– [*Rule 1: Dose Exclusion*] At any moment in the trial, if $(n_d + m_d) \geq 3$ and $\Pr(p_d > p_{\mathrm{T}} \mid n_d, m_d) > 0.95$, exclude dose d and higher doses from the trial.

– [*Rule 2: Early Termination*] If the current dose is the lowest dose (the first dose) and is considered excessively toxic according to Rule 1, terminate the trial due to excessive toxicity.

Here, $Prob\{p_d > p_T \mid Data\}$ is a function of the cumulative beta distribution $Beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$. We use $\alpha_0 = \beta_0 = 1$. For rolling designs, at the time a dose is deemed unsafe and suspended, there may be some patients with pending outcomes at this dose level. Once their data are observed later, if the safety rule is no longer violated, the dose could be reopened again for further evaluation.

2.3.6.4 Trial Termination

The PoD-TPI design stops a trial if any of the following conditions is satisfied:

1. The prespecified maximum total sample size is reached;
2. The lowest dose shows excessive toxicity according to Rule 2 in §2.3.6.3; In this case, the trial is stopped and the MTD cannot be determined;

2.3.6.5 MTD Selection

Once all the enrolled patients complete the DLT observation and the trial is not stopped due to Rule 2 in §2.3.6.4, the PoD-TPI design applies an isotonic regression to select the MTD. Follow the steps below.

1. Compute the isotonically transformed posterior means of DLT probabilities for all the dose levels in the following two steps.
   (a) Using the accumulated safety information about y_d and n_d for $d = 1, \ldots, D$, compute the posterior mean and variance for all the dose levels, $\{\tilde{p}_1, \cdots, \tilde{p}_D\}$ and $\{v_1, \cdots, v_D\}$. An independent prior $beta(0.005, 0.005)$ is used to compute each posterior mean and variance.
   (b) Compute isotonic regression estimates of the posterior means by solving the optimization problem, minimizing $\sum_{d=1}^{D}(\hat{p}_d - \tilde{p}_d)^2/v_d$ subject to $\hat{p}_j \geq \hat{p}_k$, for $j > k$. Such optimization can be done using the pooled adjacent violators algorithm (PAVA), the estimated posterior mean DLT probabilities satisfying the order constraint are obtained, denoted by $\{\hat{p}_1, \cdots, \hat{p}_D\}$.
2. Among all the tried doses ($n_d > 0$) for which $Prob\{p_d > p_T \mid Data\} < \xi$ and $\hat{p}_d \leq p_T + \epsilon_2$, select as the estimated MTD the dose with the smallest difference $|\hat{p}_d - p_T|$. That is, the estimated MTD is $d^* = \text{argmin}_d\, |\hat{p}_d - p_T|$.
3. In case of a tie (i.e., two or more doses have the smallest difference),

(a) If there is at least one dose lower than the target p_T among all the tied doses, choose the highest dose among the ones lower than p_T as the estimated MTD;

(b) Otherwise, choose the lowest dose among the tied doses as the estimated MTD.

2.3.6.6 Probability Model

Likelihood Construction

We construct the likelihood function for the observed data, with which we make inference about the distribution of the time to DLT and calculate the posterior distribution of S_d and A_d (Equation 2.2).

We first introduce some additional notation. Let τ denote the length of the DLT assessment window. In oncology, τ is usually 21 or 28 days, corresponding to a cycle of treatment. Denote by T_i the time to DLT for patient i, $i = 1, \ldots, N$; recall that we assume N patients have been treated. By definition, $Y_i = \mathbb{1}(T_i \leq \tau)$, because Y_i represents whether patient i experiences DLT within the assessment window. Conditional on the dose assignments (Z_i's), the T_i's are assumed to be independent and identically distributed with probability density function $f_{T|Z}$ and survival function $S_{T|Z}$.

Next, the following notations are defined with respect to the time when the $(N+1)$-th patient is available for enrollment. To simplify notation, we do not explicitly write out the dependency on time. Let $U_i = \min\{\tau, e_{N+1} - e_i\}$ denote the potential censoring time for patient i, where e_i is the enrollment time for patient i, and $(e_{N+1} - e_i)$ is the time between the enrollment time of patient i and the time when the new patient $(N+1)$ becomes available. Let $V_i = \min\{T_i, U_i\}$ denote the follow-up time, and let $\delta_i = \mathbb{1}(T_i \leq U_i)$ indicate whether the DLT is observed ($\delta_i = 1$) or censored ($\delta_i = 0$). We note that the case $\{\delta_i = 1\}$ corresponds to $\{Y_i = 1, B_i = 0\}$, and $\{\delta_i = 0\}$ includes $\{Y_i = 0, B_i = 0\}$ and $\{B_i = 1\}$.

Based on survival modeling (see, e.g., Klein and Moeschberger, 2006), patients with observed DLTs ($\delta_i = 1$) contribute $f_{T|Z}$ to the likelihood, and patients with censored observations ($\delta_i = 0$) contribute $S_{T|Z}$ to the likelihood. Therefore, the likelihood function is

$$L = \prod_{i=1}^{N} \left[f_{T|Z}(v_i \mid z_i)^{\mathbb{1}(\delta_i=1)} S_{T|Z}(v_i \mid z_i)^{\mathbb{1}(\delta_i=0)} \right]. \tag{2.3}$$

We define a model for $f_{T|Z}(v_i \mid z_i)$ next.

Sampling Model for Time to Toxicity

We assume a parametric distribution for T_i as follows. First, as in the mTPI-2 design, we assume

$$\Pr(T_i \le \tau \mid Z_i = d, p_d) = \Pr(Y_i = 1 \mid Z_i = d, p_d) = p_d. \tag{2.4}$$

That is, with probability p_d, the DLT for a patient treated by dose d occurs within $(0, \tau]$.

Conditional on $[T_i \le \tau]$ (i.e., $[Y_i = 1]$), we assume a piecewise uniform distribution for $[T_i \mid Y_i = 1, Z_i = d]$ on the interval $(0, \tau]$. That is, we partition $(0, \tau]$ into K sub-intervals $\{(h_{k-1}, h_k], k = 1, \ldots, K\}$, where $0 = h_0 < h_1 < \cdots < h_K = \tau$. For simplicity, we use $K = 3$ sub-intervals with equal length by default, $h_k = k\tau/K$ for $k = 0, 1, 2, 3$. The k-th sub-interval is assigned a weight w_k, and $\sum_{k=1}^{K} w_k = 1$. Conditional on $[Y_i = 1, Z_i = d]$, T_i falls into $(h_{k-1}, h_k]$ with probability w_k and follows a uniform distribution within this interval. The conditional probability density function of $[T_i \mid Y_i = 1, Z_i = d]$ is thus

$$f_{T|Y,Z}(t \mid Y_i = 1, Z_i = d, \boldsymbol{w}) = w_k \cdot \frac{1}{h_k - h_{k-1}}, \quad \text{for } h_{k-1} < t \le h_k. \tag{2.5}$$

Implicitly in (2.5), T_i and Z_i are conditionally independent given $[T_i \le \tau]$, meaning the conditional distribution of the time to DLT is the same across doses. In other words, the parameter $\boldsymbol{w}$ is shared across doses. As toxicity data are typically sparse in phase I trials, the conditional independence assumption allows borrow of information across doses and helps with the estimation of $\boldsymbol{w}$.

Next, according to the law of total probability,

$$\begin{aligned} f_{T|Z}(t \mid Z_i = d, p_d, \boldsymbol{w}) &= \sum_{y \in \{0,1\}} f_{T|Y,Z}(t \mid Y_i = y, Z_i = d, \boldsymbol{w}) \Pr(Y_i = y \mid Z_i = d, p_d) \\ &= p_d \cdot w_k \cdot \frac{1}{h_k - h_{k-1}}, \quad \text{for } h_{k-1} < t \le h_k. \end{aligned}$$

Here, $f_{T|Y,Z}(t \mid Y_i = 0, Z_i = d, \boldsymbol{w}) = 0$ for $t \le \tau$, since $\{Y_i = 0\}$ indicates $\{T_i > \tau\}$. The survival function of T_i is

$$\begin{aligned} S_{T|Z}(t \mid Z_i = d, p_d, \boldsymbol{w}) &= 1 - \int_0^t f_{T|Z}(v \mid Z_i = d, p_d, \boldsymbol{w}) \mathrm{d}v \\ &= 1 - p_d \sum_{k=1}^{K} w_k \beta(t, k), \quad \text{for } t \le \tau, \end{aligned}$$

where

$$\beta(t, k) = \begin{cases} 1, & \text{if } v > h_k; \\ \frac{t - h_{k-1}}{h_k - h_{k-1}}, & \text{if } v \in (h_{k-1}, h_k], k = 1, \ldots, K; \\ 0, & \text{otherwise.} \end{cases}$$

Finally, writing out the parametric forms of $f_{T|Z}$ and $S_{T|Z}$ in Equation (2.3), we obtain the likelihood of $\boldsymbol{p}$ and $\boldsymbol{w}$,

$$L \triangleq L(\boldsymbol{p}, \boldsymbol{w} \mid Data) \propto \prod_{k=1}^{K} w_k^{n_{\cdot k}} \prod_{d=1}^{D} \left\{ p_d^{n_d} (1-p_d)^{m_d} \prod_{i \in \mathcal{I}_d} \left[1 - p_d \sum_{k=1}^{K} w_k \beta(v_i, k) \right] \right\}. \quad (2.6)$$

Here, $n_{\cdot k} = \sum_{i:y_i=1} \mathbb{1}(h_{k-1} < v_i \leq h_k)$ is the number of patients (across all doses) who have experienced DLT in the time interval $(h_{k-1}, h_k]$.

Priors

We complete the probability model with prior models for the parameters $\boldsymbol{p} = (p_1, \ldots, p_D)$ and $\boldsymbol{w} = (w_1, \ldots, w_K)$. We assume

$$p_d \sim \text{Beta}(\theta_{d1}, \theta_{d2}), \quad \text{and} \quad \boldsymbol{w} \sim \text{Dir}(\eta_1, \ldots, \eta_K). \quad (2.7)$$

Here, θ_{d1} and θ_{d2} can be chosen based on prior guess of the DLT probability of each dose, and $(\eta_1, \ldots, \eta_K)$ can be chosen based on prior knowledge of the time to DLT falling into each sub-interval. Here, for simplicity, we use simply setting $\theta_{d1} = \theta_{d2} = 1$ and $\eta_1 = \cdots = \eta_K = 1$ by default.

Posterior of A_d

With the likelihood (2.6) and the prior model (2.7), we can conduct posterior inference on $\boldsymbol{p}$ and $\boldsymbol{w}$. Specifically,

$$\pi(\boldsymbol{p}, \boldsymbol{w} \mid Data) \propto \prod_{d=1}^{D} \pi_0(p_d) \times \pi_0(\boldsymbol{w}) \times L(\boldsymbol{p}, \boldsymbol{w} \mid Data),$$

where $\pi_0(p_d)$ and $\pi_0(\boldsymbol{w})$ are the prior models as in (2.7). Markov chain Monte Carlo simulation is used to draw samples from the posterior distribution $\pi(\boldsymbol{p}, \boldsymbol{w} \mid Data)$.

Based on the sampling models (2.4) and (2.5), we can calculate the probability that a patient experiences DLT within the assessment window given the patient has been followed for v_i ($< \tau$) without DLT, i.e., the conditional probability of $\{Y_i = 1\}$ for $i \in \mathcal{I}_d$. Recall that $\mathcal{I}_d$ contains the indices of the pending patients. For a patient $i \in \mathcal{I}_d$, we have

$$\begin{aligned} q_i(v_i, d, p_d, \boldsymbol{w}) &\triangleq \Pr(Y_i = 1 \mid T_i > v_i, Z_i = d, p_d, \boldsymbol{w}) \\ &= \frac{\Pr(T_i > v_i \mid Y_i = 1, Z_i = d, \boldsymbol{w}) \Pr(Y_i = 1 \mid Z_i = d, p_d)}{\sum_{y=0}^{1} \Pr(T_i > v_i \mid Y_i = y, Z_i = d, \boldsymbol{w}) \Pr(Y_i = y \mid Z_i = d, p_d)} \\ &= \frac{\left[1 - \sum_{k=1}^{K} w_k \beta(v_i, k)\right] p_d}{\left[1 - \sum_{k=1}^{K} w_k \beta(v_i, k)\right] p_d + (1 - p_d)}, \quad (v_i < \tau). \end{aligned}$$

Recall that S_d is the number of patients that will experience DLTs among the pending patients at dose d. Therefore, mathematically $S_d = \sum_{i \in \mathcal{I}_d} Y_i$. By definition, given the observed data (including the pending patients' follow-up times), $[S_d \mid p_d, \boldsymbol{w}, Data]$ follows a Poisson binomial distribution,

$$S_d \mid p_d, \boldsymbol{w}, Data \sim \text{Poisson-binomial}(q_i, i \in \mathcal{I}_d).$$

Here, the Poisson binomial distribution is the distribution of the sum of independent Bernoulli random variables that not necessarily have the same success probabilities. See, for example, Chen and Liu (1997) for an introduction. Furthermore, we have

$$\Pr(S_d = s \mid Data) = \int_{\boldsymbol{w}} \int_{p_d} \Pr(S_d = s \mid p_d, \boldsymbol{w}, Data) \pi(p_d, \boldsymbol{w} \mid Data) \mathrm{d}p_d \, \mathrm{d}\boldsymbol{w}.$$

This integral can be approximated using posterior samples of p_d and $\boldsymbol{w}$. Finally, we can calculate $\Pr(A_d = a \mid Data)$ according to Equation (2.2).

2.3.7 The Time-to-Event Continual Reassessment Method (TITE-CRM)

The TITE-CRM design is a method that incorporates the time-to-DLT of each patient into the CRM design proposed by Cheung and Chappell (2000). The TITE-CRM design enrolls patients as they become available to be studied and has no need to wait until the end of the follow-up window before recruiting the next patient. It accounts for the proportion of the observation period that each currently enrolled patient has been observed and assigns a dose to the next patient at any time given all information available.

2.3.7.1 Notations

Consider a toxicity-driven phase I dose-finding trial. Let p_T be the target DLT probability, and p_d be the true and unknown DLT probabilities of dose level d, $d = 1, \ldots, D$, where D denotes the prespecified number of dose levels to be investigated. Generally, we assume that p_d is non-decreasing with dose level, i.e. $p_1 \leq p_2 \leq \cdots \leq p_D$. Assume at a given moment, dose d is being used to treat enrolled patients and a total number of n patients have been enrolled. Let $Y_{i,n}$ be the indication of toxic response, where $Y_{i,n} = 1$ denotes that prior to the entry time of the $(n+1)$ patient, the ith patient has experienced DLT.

2.3.7.2 Probabilitiy Model

The weighted likelihood: The CRM assumes a parametric model $F(d, \theta)$ to describe the relationship between the dose and the toxicity. The TITE-CRM uses a weighted dose-response model

$$G(d, \omega, \theta) = \omega F(d, \theta),$$

in which weight ω is a function of the time-to-event of a patient. Under this model, the weighted likelihood of θ is

$$L_n(\theta; \boldsymbol{\omega}) = \prod_{i=1}^{n} G(d_{[i]}, \omega_{i,n}, \theta)^{y_{i,n}} \{1 - G(d_{[i]}, \omega_{i,n}, \theta)\}^{1-y_{i,n}}$$

where $y_{i,n}$ and $\omega_{i,n}$ are the indication of toxic response for the ith patient and the weight assigned to this observation just prior to the entry time for the $(n+1)$th patient, respectively, and $d_{[i]}$ is the dose of patient i.

Herein, the weight function is assumed to be

$$\omega_{i,n} = \omega(u_i; T) = \begin{cases} \frac{u_i}{T}, & y_i = 0 \\ 1, & y_i = 1 \end{cases}$$

where u_i is the follow-up time of the ith patient. The simple choice of $\omega_{i,n}$ has been shown to be adequate in many cases via simulation. And for the dose-response curve $F(d, \theta)$, we use a one-parameter power model

$$p_d = \psi(p_{0,d}, \theta) = p_{0,d}^{\exp(\theta)},$$

where $(p_{0,1}, p_{0,2}, \ldots, p_{0,D})$ are pre-specified prior toxicity probabilities ('skeletons'), which monotonically increases with d. The skeletons reflect the initial guess of DLT probabilities.

Prior specification: Let $g(\theta)$ be the prior distribution for θ, which reflects our knowledge of the dose toxicity relationship before the trial begins. We use the normal density $N(0, 1.34)$. Other choices can be gamma or exponential density.

Estimate the probability of toxicity: By the time of the $(n+1)$th patient's arrival, the estimation of parameter θ conditional on the observed data is given by the posterior mean

$$\hat{\theta}_n = \frac{\int \theta L_n(\theta; \boldsymbol{\omega}) g(\theta) d\theta}{\int L_n(\theta; \boldsymbol{\omega}) g(\theta) d\theta}.$$

Using $\hat{\theta}_n$, the estimated probability of toxicity $p_{d,n}$ for dose level d is

$$\hat{p}_d = \psi(p_{0,d}, \hat{\theta}_n).$$

Calibration of the 'skeleton' values: Lee and Cheung (2011) proposed a fast and systematic approach for selecting the skeleton based on indifference intervals for the MTD. The approach is applied by default, and users only need to specify the half-width (δ) of the indifference interval manually to estimate the skeleton.

Specifically, assume $\Theta = [b_1, b_{D+1}]$ is the parameter space (i.e. $\theta \in \Theta$) and $H_1 = [b_1, b_2)$, $H_d = [b_d, b_{d+1})$ for $d = 2, \ldots, D-1$ and $H_D = [b_D, b_{D+1})$ where b_d is the solution for $\psi(p_{0,d-1}, b_d) + \psi(p_{0,d}, b_d) = 2p_T$ for $d = 2, \ldots, D$. Based on Lee and Cheung (2011), define the half width of the indifference interval for the MTD (d) as

$$\delta_d = \frac{\psi(p_{0,d+1}, b_{d+1}) - \psi(p_{0,d-1}, b_d)}{2}, d = 2, \ldots, D-1.$$

By specifying a common half-width indifference interval for all dose levels, that is $\delta_d = \delta$, the skeletons $p_{0,1}, \ldots, p_{0,D}$ can be obtained recursively. Given a starting dose ν, a target p_T and a prior mean of $\theta = 0$, $p_{0,\nu}$ can be obtained via backward substitution, i.e. $p_T = \psi(p_{0,\nu}, 0) = p_{0,\nu}$. The

remaining skeletons can be obtained by solving the following equations:

$$\begin{cases} \psi(p_{0,d-1}, b_d) + \psi(p_{0,d}, b_d) = 2p_T \\ \psi(p_{0,d-1}, b_d) = p_T - \delta \end{cases} \quad \text{for } d \leq \nu;$$

$$\begin{cases} \psi(p_{0,d}, b_{d+1}) + \psi(p_{0,d+1}, b_{d+1}) = 2p_T \\ \psi(p_{0,d+1}, b_{d+1}) = p_T + \delta \end{cases} \quad \text{for } d > \nu.$$

We takes $\nu = [D/2]$ as the prior guess of MTD by default.

2.3.7.3 Design Algorithm

Dose Assignment: Supposes at a given moment of the trial a new $(n+1)$th patient becomes eligible for enrollment, the dose to be assigned is the one that has the posterior mean probability of toxicity closest to the target p_T. In other words, with the first n observations, the estimated $\hat{\theta}_n$ is computed and the next dose level $d_{[n+1]}$ is chosen such that $|F(d_{[n+1]}, \theta_n) - p_T| \leq |F(d_{[k]}, \theta_n) - p_T|$ for $k = 1, ..., D$.

Note: One can replace $F(d, \theta)$ with $F(x_d, \theta)$ where x_d is the actual dosage of the dose level d.

Suspension rule: Suppose the current dose is d. If the number of pending patients is larger than C for a pre-specified threshold C, suspend the enrollment.

2.3.7.4 Safety Rules

For practical concerns, similar to existing designs (for example, Ji et al., 2010 and Yuan et al., 2018), we include the following two safety rules in TITE-CRM throughout the trial.

- [*Rule 1: Dose Exclusion*] At any moment in the trial, if $n_d + m_d \geq 3$ and $\Pr(p_d > p_T \mid n_d, m_d) > 0.95$, exclude dose d and higher doses from the trial.
- [*Rule 2: Early Termination*] If the current dose is the lowest dose (the first dose) and is considered excessively toxic according to Rule 1, terminate the trial due to excessive toxicity.

Here, $Prob\{p_d > p_T \mid Data\}$ is a function of the cumulative beta distribution $Beta(\alpha_0 + y_d, \beta_0 + n_d - y_d)$. We use $\alpha_0 = \beta_0 = 1$. For rolling designs, at the time a dose is deemed unsafe and suspended, there may be some patients with pending outcomes at this dose level. Once their data are observed later, if the safety rule is no longer violated given the new data, this dose could be reopened for further evaluation.

2.3.7.5 Trial Termination

The TITE-CRM design stops a trial if any of the following conditions are satisfied:

1. The prespecified maximum total sample size is reached;
2. The lowest dose shows excessive toxicity according to Rule 2 in §2.3.7.4; in this case, the MTD cannot be determined;

2.3.7.6 MTD Selection

Once all the enrolled patients complete the DLT observation and if the trial is not stopped due to Rule 2 in §2.3.7.5, the dose level d^{**} is selected as the MTD with the smallest difference of $|\hat{p}_d - p_T|$ among all tried and safe doses d, where $\hat{p}_d$ is the posterior mean of toxicity probability for dose d.

3. Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment

3.1 Introduction

Gene therapies and adoptive cell therapies (ACTs), such as the chimeric antigen receptor (CAR) T-cell therapy, have demonstrated promising therapeutic effects in oncology patients. An important and distinct feature of some ACTs is that the probability of response may not increase with dose, which is normally seen for cytotoxic cancer therapeutics. For example, Porter et al. (2011) has shown that increased dose of CAR T-cells does not necessarily lead to higher efficacy. Because of the potential non-monotone relationship between response and dose, traditional phase 1 dose-finding designs searching for the maximum tolerated dose (MTD), like i3+3 (Liu et al., 2020) and mTPI-2 (Guo et al., 2017b) designs, are not suitable to ACTs. For example, the best efficacious dose may be lower than the MTD as higher doses may not lead to higher efficacy.

To this end, the East Bayes introduces the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module which consists of five novel statistical designs for gene and cell therapeutics dose-finding trials. The module performs trial simulations allowing head-to-head comparison of multiple designs, so that users may select the best design for their own clinical trials. The included novel designs are Ji3+3 (Lin and Ji, 2020b), PRINTE (Lin and Ji, 2020a), TEPI (Li et al., 2017), EffTox (Thall and Cook, 2004) and UBOIN (Zhou et al., 2019b), all of which use joint toxicity and efficacy outcomes as endpoints for dose finding. The goal is to identify the optimal biological dose (OBD) that possesses high efficacy and safety simultaneously. As with all other East Bayes modules involving trial simulation, below we provide detailed guid-

ance on setting up simulation for design comparison, and visualising simulation results (operating characteristics). In addition, the decision tables generation and the OBD selection are incorporated in this module so that users may generate the decision tables to guide trial conduct and estimate the OBD after trial completion. All the details are provided next.

3.2 User Interface and Tutorial

3.2.1 Overview

Entering the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** page, users will see four main tabs: **Simulation Setup**, **Simulation Results**, **Decicion table** and **OBD selection**. These four tabs allow user to conduct simulations and visualize/download simulation results, generate decision tables for trial conduct, and select OBD after trial is completed. In the **Simulation Setup** tab, there are three steps (Figure 3.1): 1) **Set trial parameters**, 2) **Select designs**, and 3) **Generate scenarios**. Users need to complete the current step to get access to the next one. Upon completing steps 1-3, users click the "Launch Simulation" button at the bottom of the page to submit the simulations to the cloud for computation. Users may also click the "Reset" button next to **Launch Simulation** to clear all settings. After the simulation is launched, the results of simulations will be displayed in the **Simulation Results** tab. The simulation process can be monitored in real time at the top of the **Simulation Results** tab. Detailed steps of using this module are elaborated next in §3.2.2-§3.2.5.

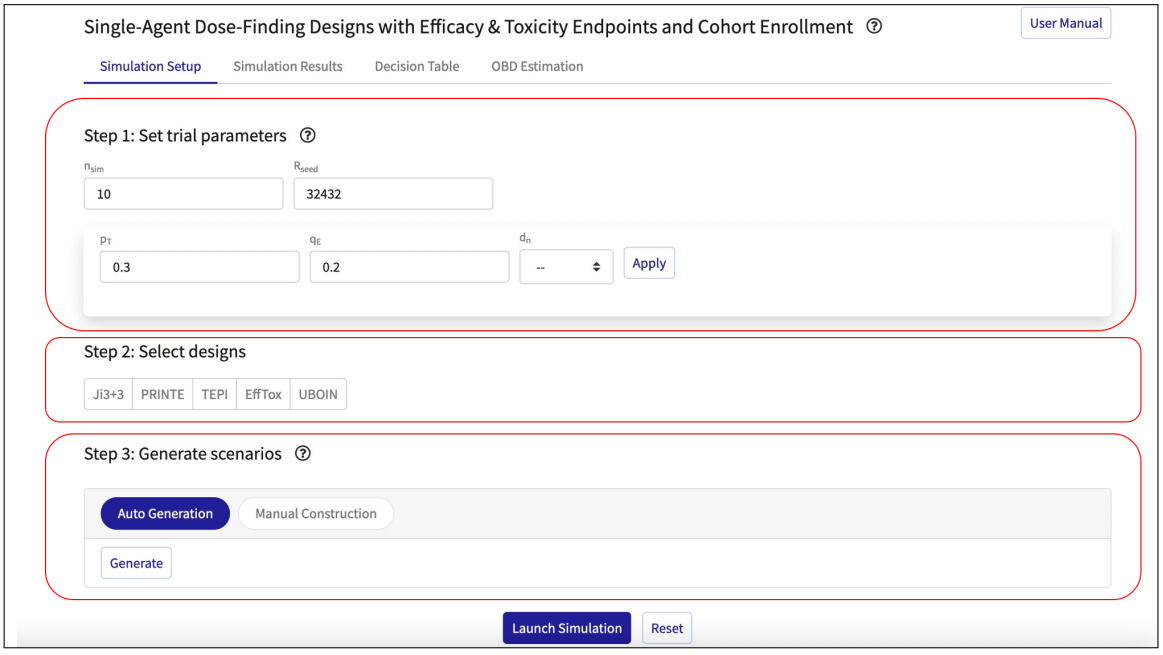


Figure 3.1: Simulation Setup in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.2.2 Simulation Setup

In the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module, East Bayes provides five designs, Ji3+3, PRINTE, TEPI, Efftox, UBOIN, for simulation. Users can choose up to four designs for head-to-head comparison in the **Simulation Setup** tab each time. Three steps of simulation set up are needed.

3.2.2.1 Step 1: Set trial parameters

Specify the number of simulations (n_{sim}) and the random seed of simulation (R_{seed}). Specify the target toxicity probability (p_T) and minimum acceptable efficacy (q_E) for the simulated trials and select a number of doses (n_{dose}) from the dropdown box. Click the "Apply" button to apply the settings. See Figure 3.2. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameters. The detailed explanation on East Bayes interface of the above four input arguments is provided in Table 3.1.

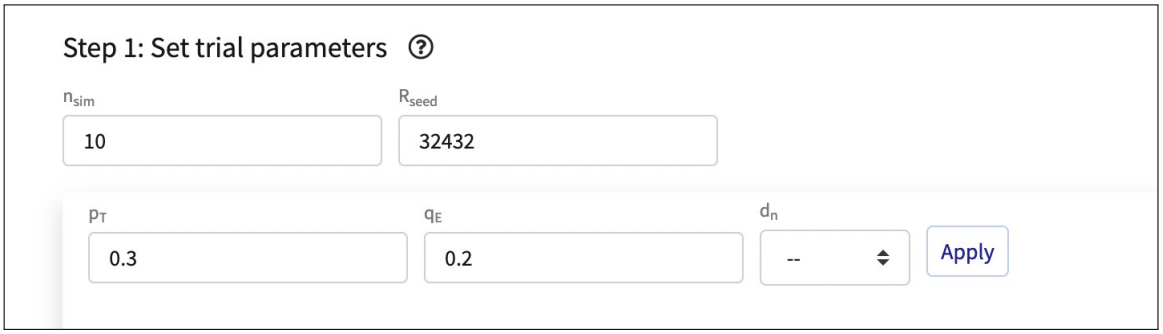


Figure 3.2: Set trial parameters in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Upon clicking the "Apply" button, a table of actual dosage will be displayed. Specify the dosage of each dose level in the table.(Figure 3.3) This is only needed if the EffTox design is selected in Step 2 next. If EffTox is not going to be selected, leave the table unchanged and move to Step 2.

3.2.2.2 Step 2: Select designs

To select a design, click the button with the design's name on it. Up to four designs may be selected for head-to-head comparison.

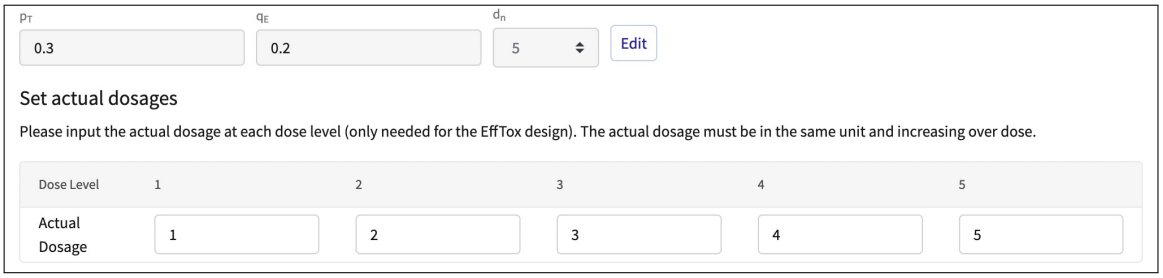


Figure 3.3: Selecting actual dosage in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Check the "Apply Stopping Rule" box to apply an ad-hoc stopping rule that stops the trial if a maximum number of patients has been enrolled at a single dose. See the detailed rules in §3.3.

Click the "Draw" button to plot a contour map of the utility function. The horizontal axis represents efficacy and the vertical axis represents toxicity. See Figure 3.4.

Click the "Apply" button of all the designs before launching simulations to apply all settings.

Click the "Delete" button to remove the selected designs.

Design parameters can be modified in the input box. Hover mouse over each design parameter, and a description will be displayed explaining the meaning of the parameters. See detailed parameter descriptions in Table 3.2.

Table 3.1: Input parameters for trials parameters in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Notation	Parameters	Description
n_{sim}	The number of simulated trials	The maximum number of simulated trials allowed is 10,000. Default value is 1,000.
R_{seed}	The random seed of simulation	A random seed is a number used to initialize a pseudorandom number generator in the simulation. Default value is 32432.
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). Default value is 0.3.
q_E	Minimum acceptable efficacy	The minimum acceptable efficacy used in the futility rule. A dose is considered not promising if the efficacy rate is unlikely to be larger than q_E. Default value is 0.2.
n_{dose}	Number of doses	The number of doses in the trial.

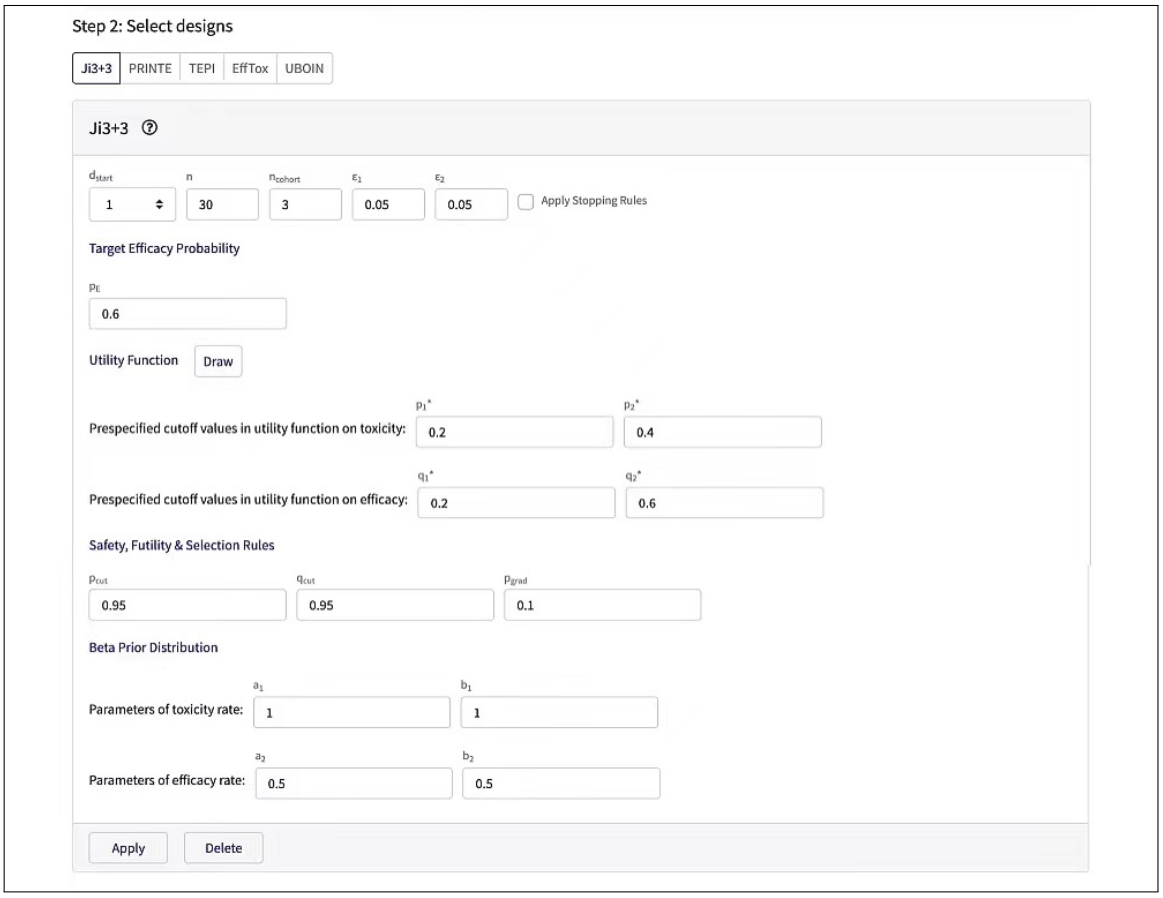


Figure 3.4: Select designs in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Table 3.2: Input parameters for designs in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Notation	Parameters	Description
d_{start} (all designs)	Starting dose level	The starting dose level in the simulated trials. Default value is 1.
n (all designs)	Sample size	The maximum number of patients to be treated in the trial. The upper limit is set at 100 since the number of patients that are enrolled in phase I clinical trial is typically small. Default value is 30.
n_{cohort} (all designs)	Cohort size	The number of patients in each cohort. Default value is 3.
K (all designs)	Maximum number of patients at a dose level	A number used in the "Stopping Rule" that stops a trial if 1) the dose-assignment decision is to escalate to the next higher dose and there has been K patients enrolled at that dose; or 2) the dose-assignment decision is to stay at the current dose and there has been K patients enrolled at that dose; or 3) if the dose-assignment decision is to de-escalate to the previous lower dose and there has been K patients enrolled at that dose; Default value is 12.
p_{cut} (all designs)	Cutoff probability for futility rule	A cutoff probability used in the safety rule. Exclude dose d if $Pr(p_d < p_T \vert Data) > p_{cut}$, where p_T is the target toxicity probability. Default value is 0.95.
q_{cut} (all designs)	Cutoff probability for efficacy rule	A cutoff probability used in the futility rule. Exclude dose d if $Pr(q_d < q_E \vert Data) > q_{cut}$, where q_E is the minimum acceptable efficacy. Default value is 0.7.
p_E (Ji3+3, PRINTE)	Target efficacy probability	The lower bound of the response probability for the treatment to be considered promising and warrant further clinical development. Default value is 0.4.
ϵ_1,ϵ_2 (Ji3+3, PRINTE)	ϵ_1: lower margin ϵ_2: higher margin	Two small fractions used to define the equivalence interval of the MTD. Any dose with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ is considered an acceptable dose MTD. Default values for both are 0.05.

p_1^*,p_2^* (Ji3+3, PRINTE, TEPI)	Prespecified cutoff values in utility function on toxicity	Cutoff values in utility function for toxicity. The toxicity utility score is 1 when $p < p_1^*$, is 0 when $p > p_2^*$ and linearly decreases when p is between (p_1^*, p_2^*). Default values are 0.2 and 0.4.
q_1^*,q_2^* (Ji3+3, PRINTE, TEPI)	Prespecified cutoff values in utility function on efficacy	Cutoff values in utility function for efficacy. The efficacy utility score is 0 when $p < p_1^*$, is 1 when $p > p_2^*$ and linearly increases when p is between (p_1^*, p_2^*). Default values are 0.2 and 0.6.
p_{grad} (Ji3+3, PRINTE)	Cutoff probability for a dose to be considered as OBD	A cutoff value used when choosing OBD. If the posterior probability of utility function lying in the admissible utility region is below p_{grad}, no OBD will be selected and the trial ends without selecting an optimal dose. Default value is 0.2.
a_1, b_1 (Ji3+3, PRINTE, TEPI)	Prior beta distribution parameters of toxicity rate	The parameters in the prior beta distribution of toxicity rate, $Beta(a_1, b_1)$. Default values for both are 1 to be conservative, since a Beta(1,1) prior implies *a prior* a dose has a toxicity rate of 0.5 with effective sample size of 0.5.
a_2, b_2 (Ji3+3, PRINTE, TEPI)	Prior beta distribution parameters of efficacy rate	The parameters in the prior beta distribution of efficacy rate, $Beta(a_2, b_2)$. Default values for both are 0.5, which is Jefferey's prior (Jeffreys, 1946).
s_1 (UBOIN)	Maximum sample size in one dose at stage 1	The maximum number of patients to be treated in one dose at stage 1. Move to stage 2 when the number of patients treated on one of the doses reaches s_1. A value between 9 and 15 generally yields good operating characteristics. Default value is 12.
s_2 (UBOIN)	Maximum sample size at one dose at stage 2	The maximum number of patients to be treated in one dose at stage 2. Stop the trial and choose OBD when the number of patients treated at one of the doses reaches s_2. For most trials, a value between 18 and 24 is a reasonable choice for s_2. Default value is 18.

Pick The Winner, Adaptive Randomization (UBOIN)	Methods to select next dose	Pick The Winner: The pick-the-winner (PW) approach deterministically assigning the next cohort of patients to dose that has the largest posterior mean utility. Adaptive Randomization: The adaptive randomization (AR) approach adaptively randomizes the next cohort of patients to a dose with probability proportional to its posterior mean utility.
$\pi^*_{1,E}, \pi^*_{2,T}, \pi^*_{3,E}, \pi^*_{3,T}$ (EffTox)	Parameters in the desirable trade-off target values	$\pi^*_{1,E}$ is the smallest efficacy probability that the physician would consider desirable if toxicity were impossible. $\pi^*_{2,T}$ is the maximum desirable value of toxicity if the efficacy were 1. Set $\pi^*_{1,E}, \pi^*_{2,T}, \pi^*_{3,E}, \pi^*_{3,T}$ so that $\pi^*_1 = (\pi^*_{1,E}, 0), \pi^*_2 = (1, \pi^*_{2,T}), \pi^*_3 = (\pi^*_{3,E}, \pi^*_{3,T})$ Default values are $0.15, 0.6, 0.25, 0.3$.

3.2.2.3 Step 3: Generate scenarios

There are two ways to generate scenarios, automatically (in the "Auto Generation" tab) or through manual construction (in the "Manual Construction" tab). Users could also manually add or delete scenarios. Once scenarios are generated, click the button "Submit" to notify the software that the scenarios are final, then click the "Launch Simulation" button at the bottom of the page to run n_{sim} (set in step 1) simulations, for each scenario and selected design (set in step 2), using the p_T and q_E values. (set in step 1).

3.2.2.3.1 Auto Generation Click the "Auto Generation" button and six diverse scenarios will be created automatically, each of which contains the true toxicity probabilities for n_{dose} dose levels. These generated scenarios are displayed (Figure 3.5). One can click the 🗑 button to delete any scenario.

3.2.2.3.2 Manual Construction A list of toxicity/efficacy probabilities are displayed. Click "Add" to add an empty, editable row of toxicity or efficacy probabilities. Click the 🗑 button to delete the row. Click "Delete All" to delete all the rows.

Check the "Select" box in the front to select the row of toxicity or efficacy probabilities. Click "Select All" to select all the toxicity or efficacy rows.

Upon selection, click "Generate" to generate scenarios which will combine existing rows of toxicity and efficacy probabilities. The scenarios will be displayed in. (Figure 3.7)

Once the scenarios are generated, clicking the 🗑 button will delete a scenario. Clicking "Delete All" will delete all the scenarios. Click the "Submit" button to notify the software that all the scenarios are final (Figure 3.7). If there are duplicated scenarios in the list, a message will be displayed on the website to indicate that the duplicated scenarios have been removed. Click the "OK" button to proceed to launch simulation. (Figure 3.8)

3.2.2.4 Launch Simulation

Once the above Steps 1-3 are completed, users can conduct simulated clinical trials to examine the operating characteristics of the selected designs using the selected scenarios, by clicking the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 3.7). A "**Success**" message will be displayed on the website (Figure 3.9) to indicate that the simulation has been successfully launched. Users may click the "OK" button in the pop-up box to track the simulation processing status and simulation results.

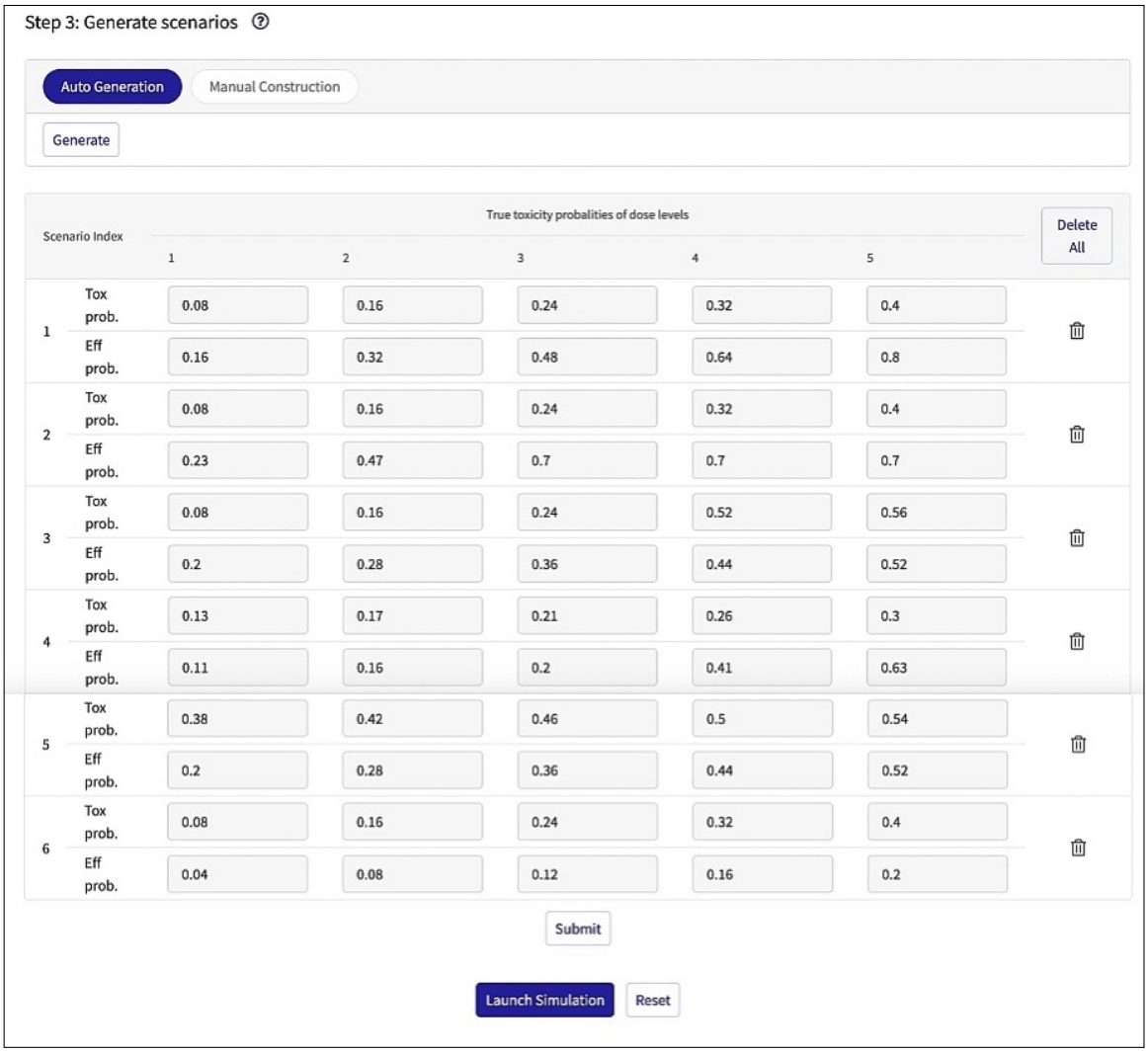


Figure 3.5: Automatically generated scenarios in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

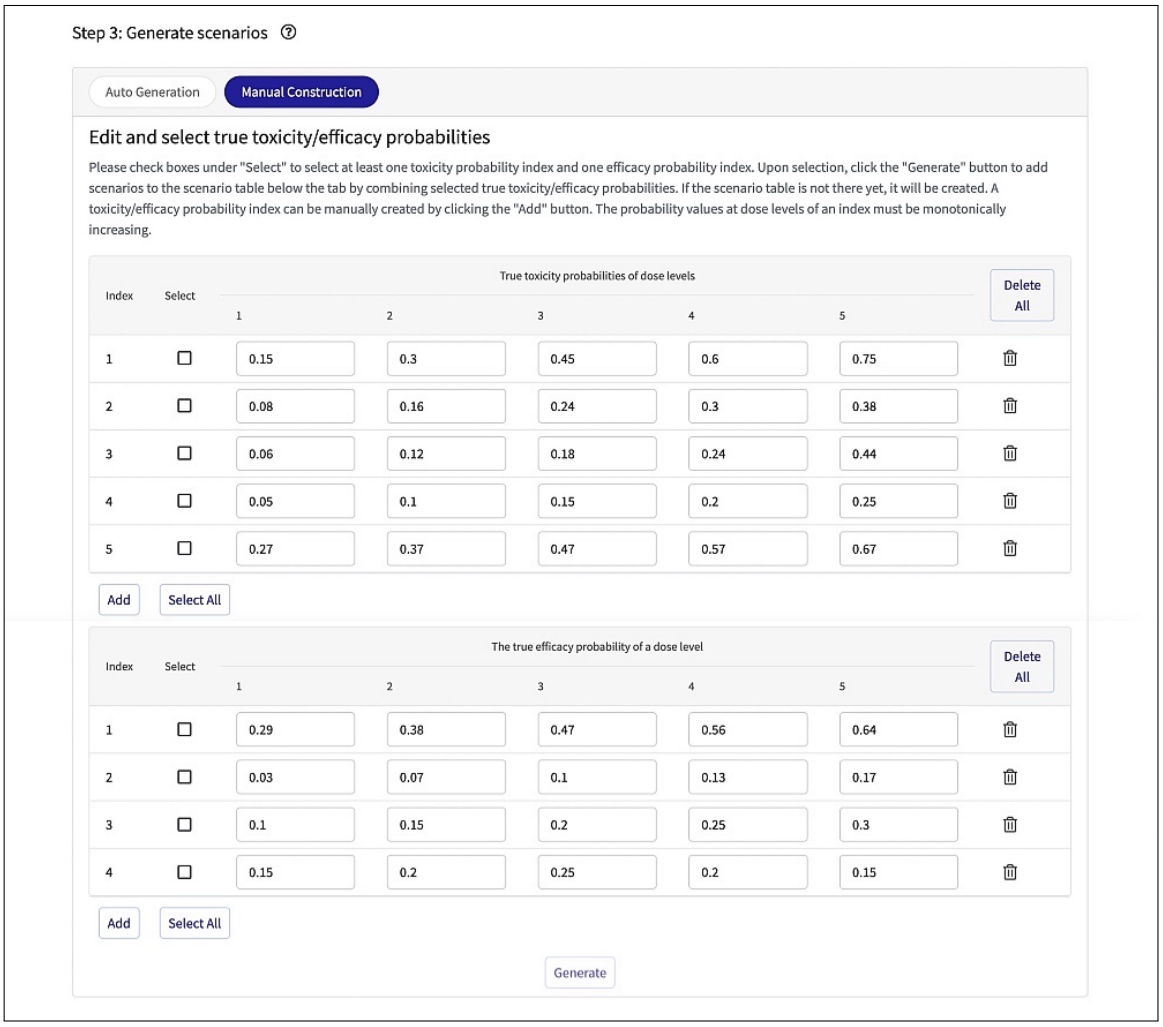


Figure 3.6: Selecting toxicity and efficacy in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

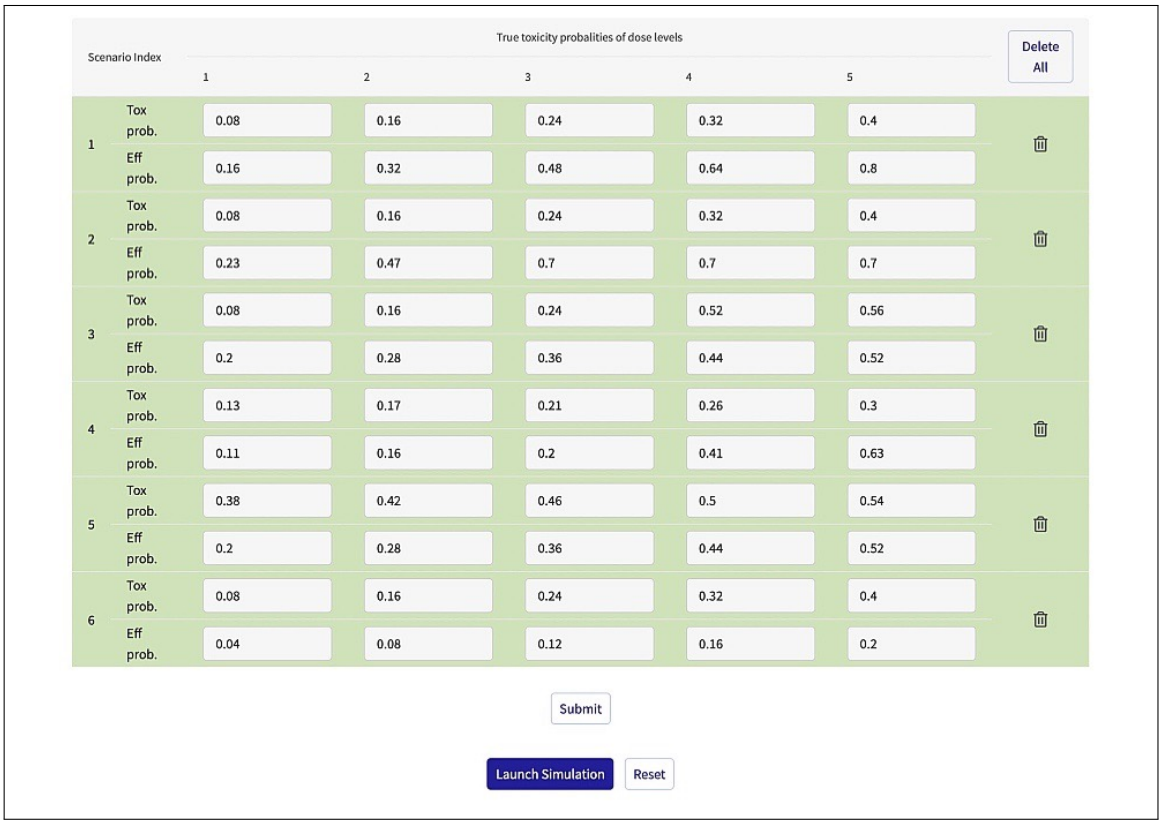


Figure 3.7: Selecting scenarios in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

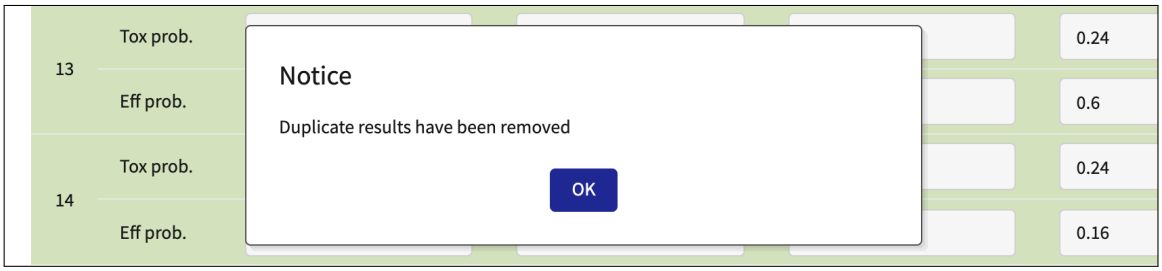


Figure 3.8: Removing the duplicated scenarios in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

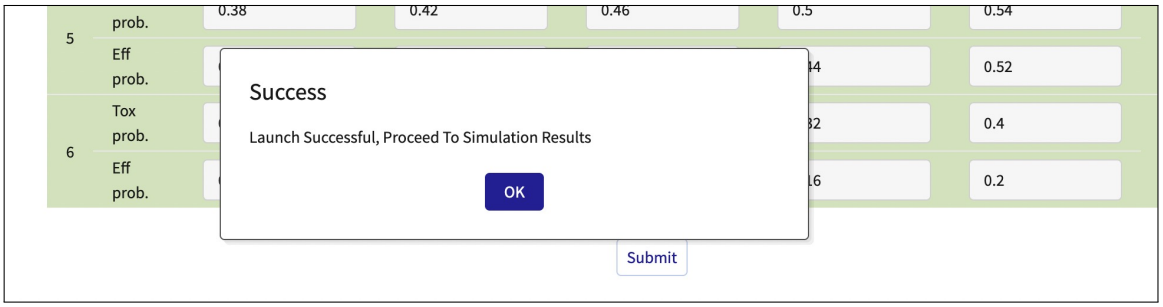


Figure 3.9: "**Launch Successful**" message after launching simulation in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.2.3 Simulation Results

In the **Simulation Results** tab, users can view and delete the simulation progress and simulation results (§3.2.3.1), restore the simulation settings if needed (§3.2.3.2), and download intelligent simulation reports (§3.2.3.3). Specifically, all the simulation results (figures and tables) can be downloaded in Word format, accompanying the statistical sections in a trial protocol. Hereinafter, we use simulation results and operating characteristics interchangeably.

3.2.3.1 View simulation results

In the **Simulation Results** tab, the **Running Simulations** panel exhibits the progress of ongoing simulation (Figure 3.10). The ongoing simulations are displayed in ascending order by the launch time. Click the icon "×" to delete the corresponding simulation.

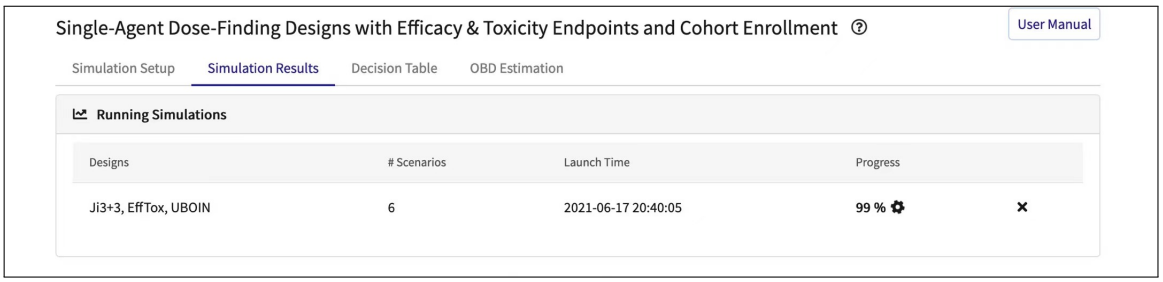


Figure 3.10: Simulation progress in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Once the simulations are completed, the **Running Simulations** panel in Figure 3.10 will disappear, success "*simulation result created*" messages will appear instead and stay at the same place of the **Running Simulations** panel unless explicitly dismissed by clicking the icon "×" at the end of the corresponding row, and the simulation results will be automatically loaded into the **Simulation History** panel (Figure 3.11), with the blue mail icon shown to indicate new results. All the previously completed simulations are also listed in the **Simulation History** panel. Simulation results for other modules can also be viewed under the **Simulation History** by dropping down the "Select a Design Category" button (Figure 3.11).

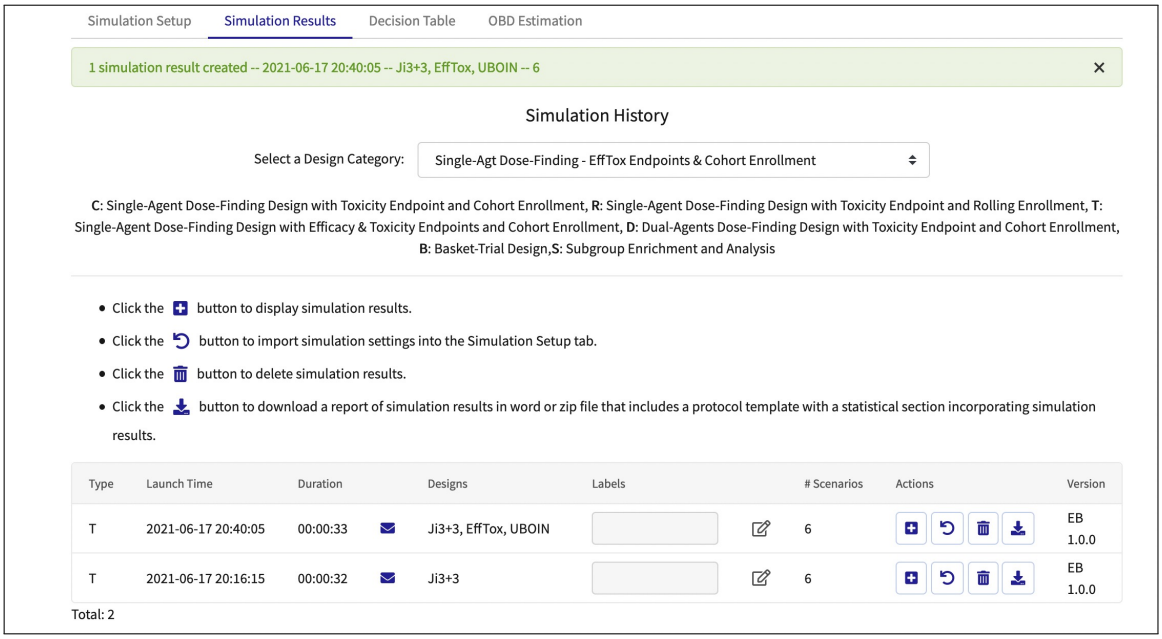


Figure 3.11: Simulation Results in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Click the button ⊞ to unfold the simulation results (Figure 3.12). The design settings are firstly displayed at the top of each simulation study (Figure 3.12). Then the results of simulation are shown as plots and tables below. And one can also click the 🗑 button to delete the selected simulation results.

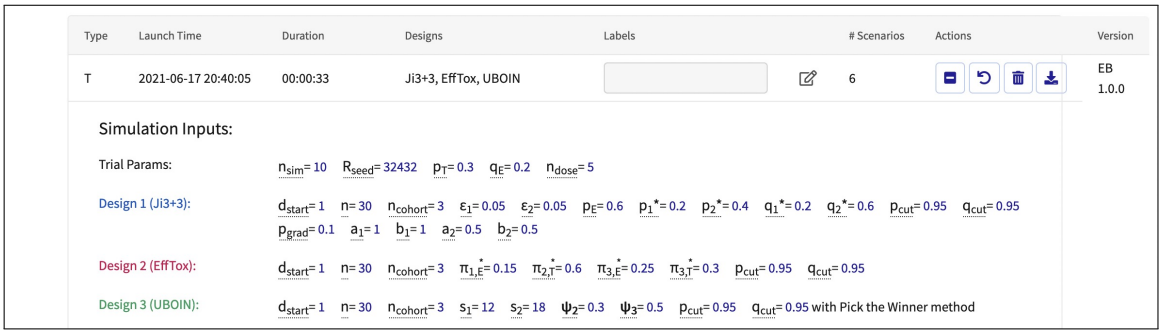


Figure 3.12: View the simulation results in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.2.3.1.1 Tabulated Results by Scenarios

Full simulation results are presented in tabular format arranged by scenarios (Figure 3.13).

In the upper part of Figure 3.13, the first three columns summarize dose levels, their true toxicity and true efficacy probabilities; the remaining columns report four dose-specific summary statistics from the simulations: selection probability, average number of patients treated, average number of toxicities (i.e. DLTs), along with their standard deviations, and average number of responses, along with their standard deviations, at each dose level. Specifically, they are

1) **Selection Prob.**: The proportion of simulated trials that select each dose level as the MTD.
2) **Average # of Patients Treated (s.d.)**: The average number of patients treated at each dose level and its standard deviation.
3) **Average # of Toxicities (s.d.)**: The average number of patients experienced DLT at each dose level and its standard deviation.
4) **Average # of Responses (s.d.)**: The average number of patients observed efficacy response at each dose level and its standard deviation.

The true OBD(s) of the scenario is (are) highlighted by the orange bar. The true OBD is defined as the dose that achieves the highest utility, which could be calculated using true toxicity, efficacy probabilities and the utility function.

In the lower part of Figure 3.13, more trial-specific summary statistics are reported, mainly from five aspects: **OBD Selection**, **Subjects Assignment**, **Trial Toxicity**, **Trial Stopping** and **Trial Sample Size**. Specifically, they are

- **OBD Selection**
  - **Prob. of Selecting OBD**: The proportion of simulated trials that select the true OBD at the end of the trial. The higher the value, the better the design.
  - **Prob. of Selecting Does-over-MTD**: The proportion of simulated trials that select the doses higher than the true MTD at the end of the trial. The lower the value, the better the safety of the design.
  - **Prob. of No Selection**: The proportion of simulated trials in which none of the dose levels are selected as the OBD. If a scenario does not have any OBD, this values is treated as the probability of selecting the true OBD.
- **Subjects Allocation**
  - **Prob. of Correct Allocation (s.d.)**: The average proportion of patients who are correctly assigned to the true OBD by the design across all the simulated trials and its standard deviation. The higher the value, the better the design.

  - **Prob. of Overdosing Allocation (s.d.)**: The average proportion of patients who are assigned to doses higher than the MTD by the design across all the simulated trials and its standard deviation. The lower the number, the better the safety of the design.

- **Trial Toxicity**
  - **Prob. of Toxicity**: The proportion of patients experiencing DLT across all the simulated trials. The lower the number, the fewer patients having DLTs under the design.
- **Trial Stopping**
  - **Prob. of Early Stopping Trial due to No admissible dose**: The proportion of simulated trials in which the trial is stopped because there is no admissible dose left. This means that all the doses have unacceptable toxicity or efficacy and are excluded by safety rule or futility rule.
  - **Prob. of Early Stopping Trial due to Reaching** K: The proportion of simulated trials in which the trial is stopped because the dose-assignment decision is to escalate/stay/de-escalate to a dose level but that dose has enrolled at least K patients ($K < n$, e.g., $K = 12$).
  - **Prob. of Stopping Trial due to Reaching** n: The proportion of simulated trials in which the trial is stopped because the total number of patients enrolled and treated in a trial has reached or exceeded the pre-specified maximum sample size n.
- **Trial Sample Size**
  - **Average # of Patients Treated (s.d.)**: The average number of patients treated in the simulated trials and its standard deviation. Due to early stopping, this number is lower than or equal to n.
- **Statistics of UBOIN**
  - **Prob. of Entering Stage II**: The proportion of simulated trials in which the trial enters Stage II because the number of patients at one dose has reached or exceeded the pre-specified maximum sample size s_1 in Stage I.
  - **Average # of Patients Treated in Stage I**:The average number of patients treated in Stage I in the simulated trials.
  - **Average # of Patients Treated in Stage II**:The average number of patients treated in Stage II in the simulated trials.

When calculating the standard deviation, we use n_{sim} as the denominator instead of $(n_{sim} - 1)$ in East Bayes.

Simulation Outputs:

Tabulated Results by Scenarios

Scenario 1

			Selection Prob.			Average # of Patients Treated (s.d.)			Average # of Toxicities (s.d.)			Average # of Responses (s.d.)		
Dose Level	True Tox Prob.	True Eff Prob.	Design 1 Ji3+3	Design 2 EffTox	Design 3 UBOIN	Design 1 Ji3+3	Design 2 EffTox	Design 3 UBOIN	Design 1 Ji3+3	Design 2 EffTox	Design 3 UBOIN	Design 1 Ji3+3	Design 2 EffTox	Design 3 UBOIN
1	0.08	0.16	0.1	0.1	0	3.3 (0.9)	4.5 (4.5)	8.7 (3.407)	0.1 (0.3)	0.6 (0.917)	0.8 (0.98)	0.8 (0.6)	0.8 (1.166)	1.9 (1.578)
2	0.16	0.32	0.3	0	0.4	9 (7.348)	3.6 (1.2)	10.2 (4.069)	1.2 (1.249)	0.6 (0.663)	2.1 (1.221)	3.3 (2.968)	0.9 (0.539)	3.7 (1.616)
3	0.24	0.48	0.3	0.1	0.3	8.4 (6.264)	8.7 (6.067)	6.3 (3.662)	2.3 (2.238)	1.7 (1.487)	1.8 (1.327)	4.5 (4.653)	4 (4.171)	2.9 (2.3)
4	0.32	0.64	0.3	0.6	0.1	7.8 (6.997)	8.7 (5.274)	3.6 (2.939)	1.8 (1.939)	3.4 (2.653)	1.5 (1.36)	4.5 (4.522)	5.7 (4.383)	2.2 (1.833)
5	0.4	0.8	0	0.2	0.2	1.5 (2.419)	4.5 (5.408)	1.2 (2.4)	0.8 (1.327)	1.9 (2.071)	0.2 (0.4)	1.3 (2.369)	3.6 (4.152)	1.2 (2.4)

* The cell with orange background color indicates the TRUE OBD for each design

		Design 1 Ji3+3	Design 2 EffTox	Design 3 UBOIN
OBD Selection*	Prob. of Selecting OBD	0.3	0.1	0.3
	Prob. of Selecting Dose-over-MTD	0	0.2	0.2
	Prob. of No Selection	0	0	0
Patients Assignment	Prob. of Correct Allocation (s.d.)	0.28 (0.209)	0.56 (0.242)	0.21 (0.122)
	Prob. of Overdosing Allocation (s.d.)	0.05 (0.081)	0.29 (0.176)	0.04 (0.08)
Trial Toxicity	Prob. of Toxicity	0.207	0.273	0.213
Trial Stopping	Prob. of Early Stopping Trial due to No Admissible Dose	0	0	0
	Prob. of Early Stopping Trial due to Reaching K**	0	0	0
	Prob. of Stopping Trial due to Reaching n**	1	1	1
Trial Sample Size	Average # of Patients Treated (s.d.)	30 (0)	30 (0)	30 (0)
Statistics of UBOIN	Prob. of Entering Stage II			0.8
	Average # of Patients Treated in Stage I			22.8
	Average # of Patients Treated in Stage II			7.2

* If there is no OBD, Prob. of Selecting OBD equals to Prob. of No Selection

** K, the maximum sample size at a dose; n, the maximum sample size in a trial.

Figure 3.13: Simulation result tables in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.2.3.2 Restore simulation setup

Users can restore the simulation settings from the simulation results by clicking the button at the upper right corner of each simulation results panel (yellow arrow in Figure 3.14). Upon clicking, the display will switch to the **Simulation Setup** page with the same simulation settings restored. This is useful to restore the old simulation settings for **Reproducible results**.

Figure 3.14: Restore simulation setup and download simulation results in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.2.3.3 Download simulation results

There is a button at the upper right corner of each simulation results panel (green arrow in Figure 3.14). Click it to download a word file, which includes four parts:

- Part A: Complete simulation results under the designs and scenarios users added in the Simulation Setup tab;
- Part B: Detailed technical descriptions of the designs users added in the Simulation Setup tab;
- Part C: Reference.

Users may select the required parts and modify them tailored for their trials or contact us via email (support@cytel.com) for consulting services.

3.2.4 Decision Tables

This function generates decision tables based on the Ji3+3, PRINTE, and TEPI designs, which can be used to conduct a dose-finding trial. Users can click the tabs to switch between the tables for the Ji3+3, PRINTE, and TEPI designs.

Manually type in the design settings for decision table generation (Figure 3.15). The parameters are the same as the ones in Step 2 (3.2.2.2) in the Simualtion Setup tab. See detailed parameter descriptions in Table 3.2.

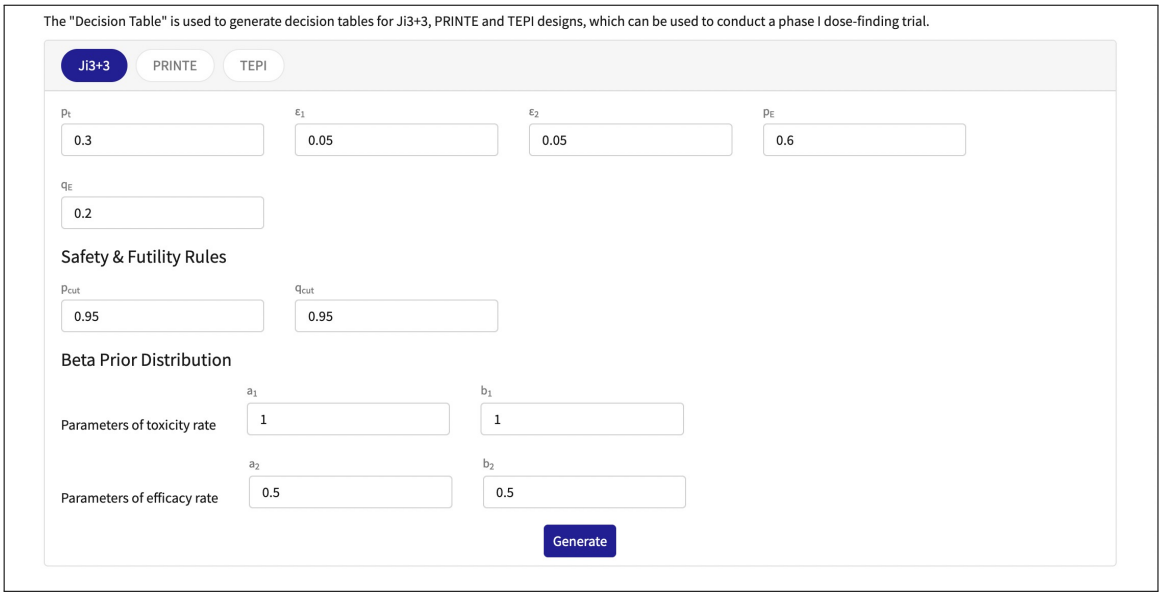


Figure 3.15: Input parameters in the **Generate Decision Table** tab of **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

Click the "Generate" button to generate decision table (Figure 3.16). Decision tables are automatically generated for 3, 6, 9 and 12 patients at a dose in the panel below.

To generate a single decision table by specifying the number of patients treated at a dose d, set n_d in the box and click the button "Add". (Figure 3.16)

For each decision table, the column represents the number of patients responses among those treated at the dose, and the row represents the number of patients who have experienced dose-limiting toxicity (DLT) events. Note that these are the counts of patients, not DLT events or responses. For example, column 3 and row 1 means that among the patients that have been treated at the current dose 3 of them experiences DLT, and 1 of them responses.

Each cell in the decision table provides the dose-assignment decision based on the readouts

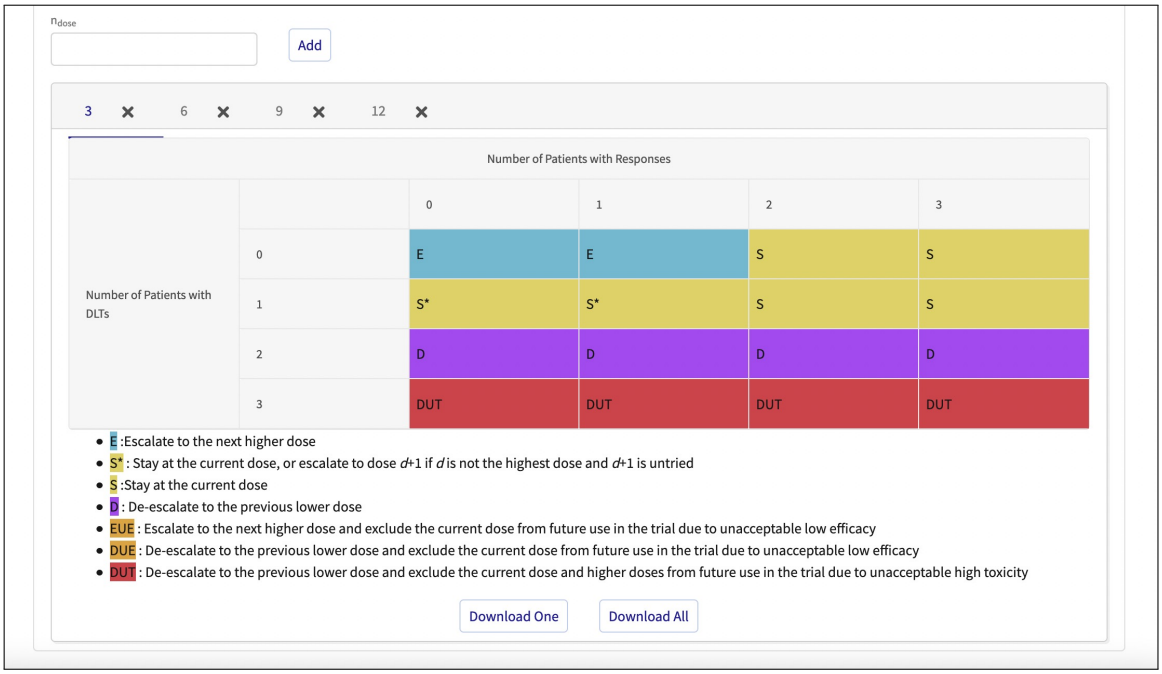


Figure 3.16: Decision tables in the **Generate Decision Table** tab of **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

from the corresponding row and column. For example, for column 3 and row 1, i.e., 3 patients experience DLTs, and 1 patient has efficacy response, the decision is "EU". The letters in the decision table represent different dose-assignment decisions as shown below:

- "E" stands for escalating to the next higher dose,
- "S*" stands for staying at the current dose, or escalate to dose $d + 1$ if d is not the highest dose and $d + 1$ is untried
- "S" stands for staying at the current dose,
- "D" stands for de-escalating to the previous lower dose,
- "DUT" stands for de-escalating to the previous lower dose, and the current dose and its higher doses is deemed unacceptable due to severe toxicity and will not be used again in the study. If at the first dose level, users can choose to early-terminate the trial or not based on their own discretion.
- "EUE" stands for escalating to the higher dose and marking the current dose as unacceptable (due to futility) so that it will never be used again in the remainder of the trial.
- "DUE" stands for de-escalating to the previous lower dose, and the current dose is deemed

unacceptable due to futility and will not be used again in the study.

Some additional detailed explanation of the decisions are provided in the decision table report. The meaning of the notations are shown below:

- The superscript * on DUE indicates that according to the Ji3+3 design, the decision is S and the current dose is deemed unacceptable due to futility. In this case, a decision S indicates a moderate or high toxicity probability, so the only sensible action is to de-escalate to the previous lower dose, and remove the current dose (due to futility) from the study.
- The superscript ** on DUT indicates that if the current dose is the first dose level, users can choose to early-terminate the trial or not based on their own discretion.

Click “DOWNLOAD ONE” to download a word file, which includes the design settings and the single decision table in the tab selected. Click “DOWNLOAD ALL” to download a word file, which includes the design settings and all the decision tables generated.

3.2.5 OBD Estimation

In this module, all designs aim to estimate the OBD when the trial is completed and the data is collected. The detailed statistical models for the included designs are described in §3.3.

First, select a design and provide corresponding model parameters. Second, select the number of doses (n_{dose}) from the dropdown box, and an editable table will be shown on the website (Figure 3.17). For the Ji3+3, PRINTE and TEPI design, provide the the number of patients treated, the observed number of DLT events, and provide the observed number of efficacy events at each dose into the table; For the UBOIN and Efftox design, provide the observed number of patients who has no efficacy but DLT ($(Y_E, Y_T) = (0, 1)$), no efficacy and no DLT ($(Y_E, Y_T) = (0, 0)$), efficacy and DLT ($(Y_E, Y_T) = (1, 1)$) and the number of patients who has no efficacy and no DLT ($(Y_E, Y_T) = (1, 0)$). Click the "Generate" button to estimate the utilities of each dose and estimate the OBD for the trial. The estimated utility will be displayed in a table and the estimated OBD will be highlighted in green color as shown in Figure 3.18.

See detailed parameter descriptions in Table 3.2 in §3.2.2.2.

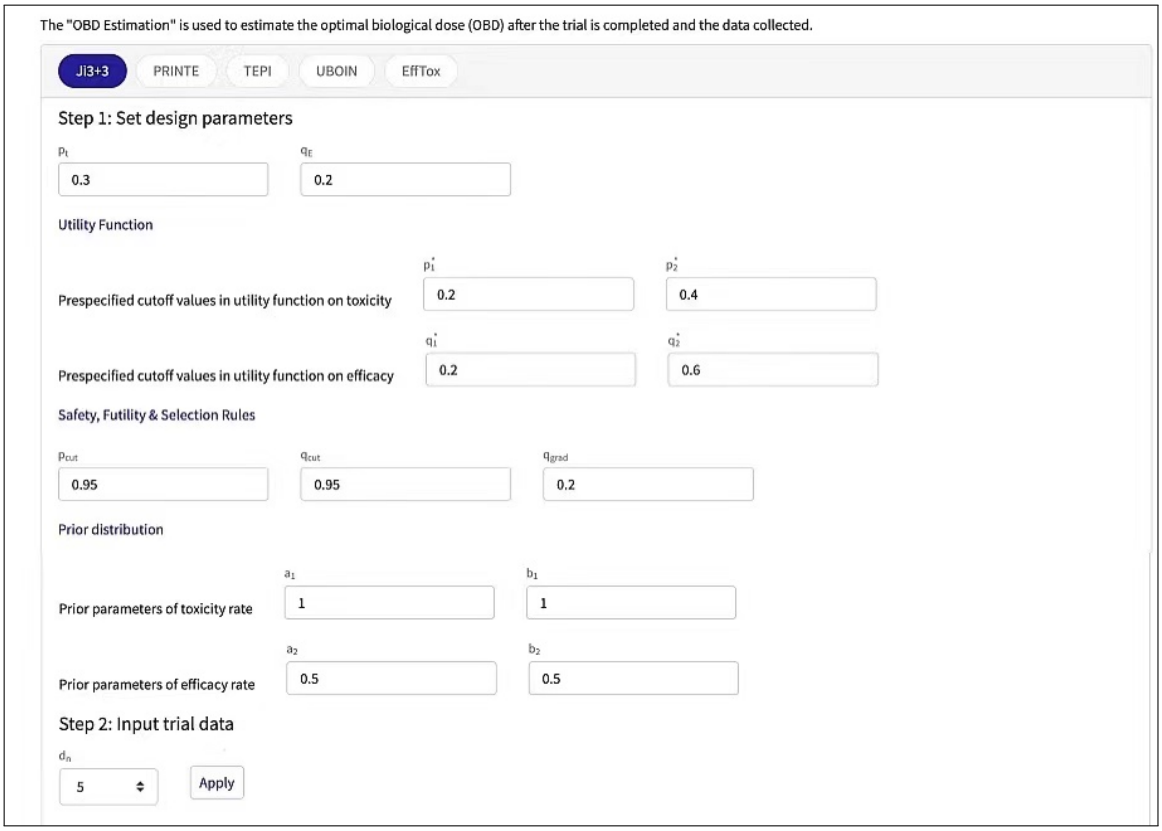


Figure 3.17: Input parameters in the **OBD Estimation** tab of **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

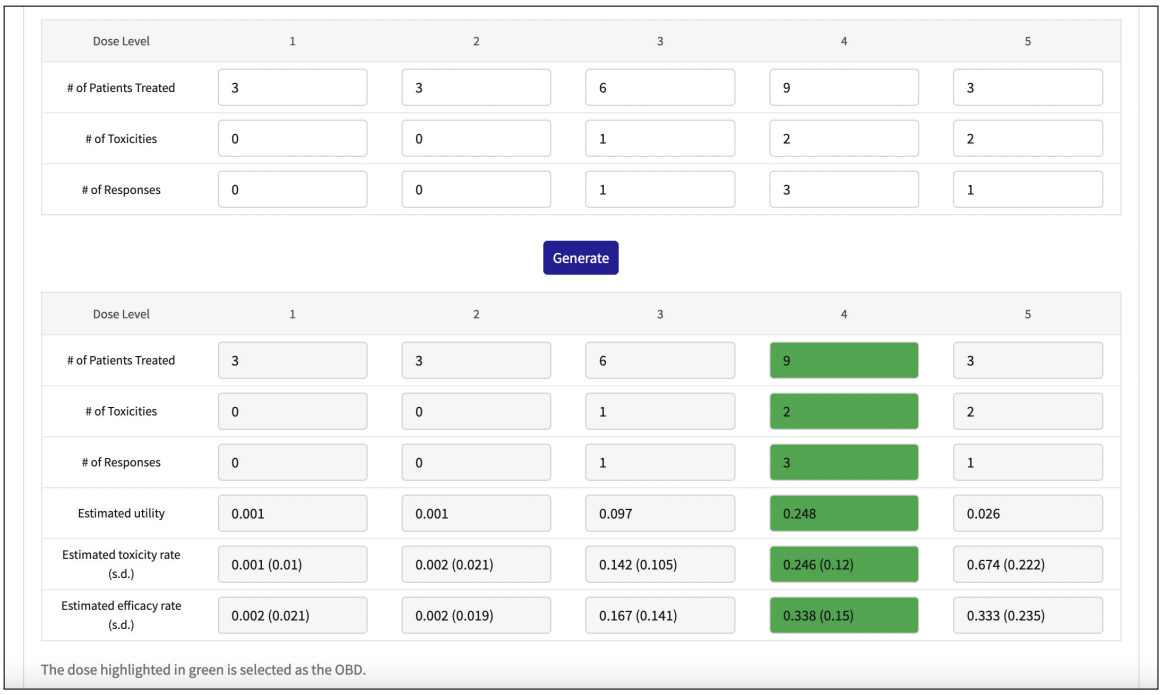


Figure 3.18: Determine the estimated OBD in the **Single-Agent Dose-Finding Designs with Efficacy&Toxicity Endpoints and Cohort Enrollment** module.

3.3 Statistical Methods Review

3.3.1 The Joint i3+3 (Ji3+3) Design

Ji3+3 is a rule-based phase I/II ACT dose-finding design proposed by Lin and Ji (2020b). Building upon i3+3 (Liu et al., 2020), Ji3+3 takes into account of both toxicity and efficacy outcomes in making dosing recommendations. Basically, the decision rules of the Ji3+3 design incorporate and extend the toxicity rules in i3+3 with a set of efficacy rules. Simulation results show that Ji3+3 outperforms existing designs when monotonic dose response assumption is violated, and achieves comparable performance when the assumption holds. Since Ji3+3 is a model-free design, it is transparent to physicians and simple to implement.

3.3.1.1 Dose-Finding Algorithm

Consider D ascending doses in a single-agent ACT phase I trial. Due to ethical considerations, it is always assumed that the toxicity probability p_d increases with dose level d, that is, $p_1 \leq \cdots \leq p_D$. However, the efficacy probability q_d may increase initially and then reach a plateau from which minimal improvement or even decreasing efficacy may be seen with increasing dose. For this reason, we assume that q_d is not monotone with d, and that p_d and q_d are independent. Suppose that dose d is currently used in the trial and n_d patients have already been allocated to dose d, with x_d and y_d patients experiencing toxicity and efficacy outcomes. Aggregating across all the doses, the trial data are denoted as $Data = \{(n_d, x_d, y_d), d = 1, \cdots, D\}$

Denote p_T as the target toxicitiy rate, which is the probability of toxicity at the MTD; denote p_E as the target efficacy rate. In Ji3+3, $[p_T - \epsilon_1, p_T + \epsilon_2]$ is defined as the Equivalence Interval (EI), where (ϵ_1, ϵ_2) are two small fractions that account for the uncertainty around p_T. This allows doses whose toxicity probabilities differ from p_T to be considered as the MTD. Given the observed data $Data$, the dose-finding algorithm of the Ji3+3 design is shown in Table 3.3. The algorithm follows these principles:

1. If there is lack of evidence for efficacy, escalate to achieve higher efficacy; else, stay at the current dose because it is considered to have sufficient efficacy.
2. For toxicitiy, the idea is to compare the observed toxicity rate $\frac{x_d}{n_d}$ with the EI.
   - If $\frac{x_d}{n_d}$ is below the EI, the dose is considered safe; if $\frac{x_d}{n_d}$ is inside the EI, the dose is considered to be close to the MTD; if $\frac{x_d}{n_d}$ is above the EI, the dose is considered not safe except when $\frac{x_d - 1}{n_d}$ is below the EI.

Table 3.3: Schema of the Ji3+3 design.

Current dose d; n_d patients, x_d Tox, y_d Eff		
Eff cond.	Tox cond.	Next dose (Decision)
$\frac{y_d}{n_d} \leq p_E$	$\frac{x_d}{n_d} <$ EI	$d+1$ (E)
	$\frac{x_d}{n_d} \in$ EI	d (S or E^*)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} <$ EI	d (S)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} \in$ EI	$d-1$ (D)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} >$ EI	$d-1$ (D)
$\frac{y_d}{n_d} > p_E$	$\frac{x_d}{n_d} <$ EI	d (S)
	$\frac{x_d}{n_d} \in$ EI	d (S)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} <$ EI	d (S)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} \in$ EI	$d-1$ (D)
	$\frac{x_d}{n_d} >$ EI & $\frac{x_d-1}{n_d} >$ EI	$d-1$ (D)

*: Escalate to dose $d+1$ if $n_{d+1} = 0$.

- When $\frac{x_d-1}{n_d}$ is below the EI and $\frac{x_d}{n_d}$ is above the EI, the data is noisy since increment of one toxicity event renders the observed toxicity rate to jump from below the EI to above the EI. In other words, the observed data is not very informative because change of one toxicity event can greatly influence the toxicity estimate.

Consider an example. Suppose EI $= [0.2, 0.3]$ with $x_d = 1$ and $n_d = 3$. Even though $\frac{x_d}{n_d} = \frac{1}{3}$ is above the EI, $\frac{x_d-1}{n_d} = \frac{0}{3}$ is below the EI. And therefore, dose d should not be considered as above the MTD.

3. Intersecting the two dosing principles for toxicity and efficacy, and taking the more conservative decision between the two, we arrive at the decisions in Table 3.3.
4. When d is the highest dose or lowest dose, the above rules are modified as special cases,

   - If the current dose is the highest dose, decision "E" (escalate and treat the next cohort of patients at the next higher dose) should be replaced with decision "S" (stay and continue to enroll patients at the current dose), since there is no dose to escalate to.

   - Similarly, if the current dose is the lowest dose, decision "D" (de-escalate to the next lower dose) should be replaced with "S" since there is no dose to de-escalate to.

Safety and futility rules

- Safety rule: if $\Pr(p_d > p_T \mid x_d, n_d) > p_{cut}$ for a p_{cut} close to 1 (say, 0.95), exclude doses $d, d+1, \cdots, D$, from future use in the trial; treat the next cohort of patients at dose $(d-1)$.
- Futility rule: if $\Pr(q_d < q_E \mid y_d, n_d) > q_{cut}$ for a q_{cut} close to 1 (say, 0.7), where q_E is the minimum acceptable probability of efficacy, then exclude dose d from future use in the trial. Here, q_E is the reference efficacy rate, e.g., the efficacy rate of standard care.

Note that, here we assume the prior for each p_d follows an independent $beta(a_1, b_1)$, and the prior for each q_d follows an independent $beta(a_2, b_2)$, where $beta(\alpha, \beta)$ denotes a beta distribution with mean $\alpha/(\alpha+\beta)$. The posterior distributions for p_d and q_d in the above rules are $beta(a_1 + x_d, b_1 + n_d - x_d)$ and $beta(a_2 + y_d, b_2 + n_d - y_d)$, respectively.

Stopping rules

The trial is stopped if

1. the prespecified maximum total sample size n is reached; or
2. the lowest dose shows excessive toxicity according to the safety rule; In this case, the trial is early stopped and the MTD cannot be determined; or
3. *optional:*
   - *the Ji3+3 decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *the Ji3+3 decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *the Ji3+3 decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

3.3.1.2 Dose Selection

At the end of the trial, Ji3+3 chooses the OBD using a joint utility score $U(p,q) = f_1(p)f_2(q)$ (suppressing dose d in the notation), which takes the product of toxicity utility $f_1(p)$ in (3.1) and efficacy utility $f_2(q)$ in (3.2).

$$f_1(p) = \begin{cases} 1, & p \in (0, p_1^*), \\ 1 - \frac{p - p_1^*}{p_2^* - p_1^*}, & p \in (p_1^*, p_2^*), \\ 0, & p \in (p_2^*, 1). \end{cases} \tag{3.1}$$

$$f_2(q) = \begin{cases} 0, & q \in (0, q_1^*), \\ \frac{q-q_1^*}{q_2^*-q_1^*}, & q \in (q_1^*, q_2^*), \\ 1, & q \in (q_2^*, 1). \end{cases} \tag{3.2}$$

For toxicity, define two thresholds p_1^* and p_2^* such that the toxicity utility score is 1 when $p < p_1^*$, 0 when $p > p_2^*$, and linearly decreases when p is between (p_1^*, p_2^*). For efficacy, define two thresholds q_1^* and q_2^* such that the efficacy utility score is 0 when $q < q_1^*$, is 1 when $q > q_2^*$, and linearly increases when q is between (q_1^*, q_2^*). The OBD is selected according to the following process.

1. We generate a total of T random samples, $\left\{p_d^{(t)}, t = 1, \cdots, T\right\}$ and $\left\{q_d^{(t)}, t = 1, \cdots, T\right\}$, from the posterior distributions $beta(a_0 + x_d, b_0 + n_d - x_d)$ and $beta(a_0 + y_d, b_0 + n_d - y_d)$ for each dose d, respectively. Here, East Bayes sets $a_0 = b_0 = 0.005$ and $T = 1000$.
2. For toxicity probabilities of all doses in each sample t, $\boldsymbol{p}^{(t)} = (p_1^{(t)}, \cdots, p_D^{(t)})$, we perform isotonic transformation using the pool adjacent violators algorithm (PAVA; Mair et al. 2009) on $\boldsymbol{p}^{(t)}$ to obtain $\widetilde{\boldsymbol{p}}^{(t)} = (\widetilde{p}_1^{(t)}, \cdots, \widetilde{p}_d^{(t)})$, where $\widetilde{p}_i^{(t)} \leq \widetilde{p}_j^{(t)}$ if $i < j$.
3. We propose a probabilistic inference for selecting the OBD and avoid selecting doses with low utility. Define an admissible probability region (APR) $A(p, q) = \{(p, q) \mid p \in (0, p_T], q \in [q_E, 1)\}$. Then the OBD is selected only from the candidate dose set $\mathcal{A}$,

   $$\mathcal{A} = \left\{d \mid p_{\text{in},d} \geq p_{\text{grad}}, n_d > 0, d = 1, \cdots, D\right\},$$

   where $p_{\text{in},d} = \Pr\{(p_d, q_d) \in \text{APR} \mid Data\}$ is the posterior probability that dose d belongs to APR and p_{grad} is a small value (say, 0.1). We use a simple a simple numerical approximation approach to compute $p_{\text{in},d}$ given by

   $$\hat{p}_{\text{in},d} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left\{(\widetilde{p}_d^{(t)}, q_d^{(t)}) \in \text{APR}\right\}.$$

4. The final selected dose d^* is the one that maximizes the utility score $U(p_d, q_d)$. That is, $d^* = \text{argmax}_{d \in \mathcal{A}} \hat{E}[U(p_d, q_d) \mid Data]$, where

   $$\hat{E}[U(p_d, q_d) \mid Data] = \frac{1}{T} \sum_{t=1}^{T} U(\widetilde{p}_d^{(t)}, q_d^{(t)}).$$

3.3.2 The Toxicity and Efficacy Probability Interval (TEPI) Design

TEPI, proposed in Li et al. (2017), is a practical dose-finding design for ACT trials that incorporates both toxicity and efficacy data. It is a natural extension of mTPI by adding the efficacy interval into the dose-finding model. TEPI partitions the unit intervals $(0, 1)$ for both the toxicity probability p_i and efficacy probability q_i into subintervals, denoted as (a, b) and (c, d), respectively. Then it uses beta-binomial models to estimate the efficacy and toxicity probability and makes dosing-decisions based on the joint unit probability mass (JUPM) of the interval combinations $(a, b) \times (c, d)$. TEPI is transparent to clinicians and simple to implement in practice.

3.3.2.1 Elicited decision table

The dose-finding algorithm of TEPI is based on a clinician-elicited decision table in terms of efficacy and toxicity probability intervals. The procedures of ecliting the decision table are as follows.

Consider D ascending doses in a single-agent ACT phase I trial. Due to ethical considerations, it is always assumed that the toxicity probability p_d increases with dose level d, that is, $p_1 \leq \cdots \leq p_D$. However, the efficacy probability q_d may increase initially and then reach a plateau from which minimal improvement or even decreasing efficacy may be seen with increasing dose. For this reason, we assume that q_d is not monotone with d, and that p_d and q_d are independent. Suppose that dose d is currently used in the trial and n_d patients have already been allocated to dose d, with x_d and y_d patients experiencing toxicity and efficacy outcomes. Aggregating across all the doses, the trial data are denoted as $Data = \{(n_d, x_d, y_d), d = 1, \cdots, D\}$

Partition the unit intervals $(0, 1)$ for p_d and q_d into four subintervals. Denoting (a, b) and (c, d) a subinterval in the partition for p_d and q_d respectively, where

$$(a, b) \in \Big\{(0, t_1), (t_1, t_2), (t_2, t_3), (t_3, 1)\Big\},$$

$$(c, d) \in \Big\{(0, e_1), (e_1, e_2), (e_2, e_3), (e_3, 1)\Big\}.$$

The interval combinations $(a, b) \times (c, d)$ form the basis for dose-finding decisions, with each combination corresponding to a specific decision, such as dose escalation or de-escalation. East Bayes uses a default fixed decision for each interval combination, see Table 3.4.

In order to formulate this table, it is required to determine: (i) bounds of efficacy rate interval, e_1, e_2, e_3, and (ii) bounds of toxicity rate interval, t_1, t_2, t_3.

Table 3.4: An default decision table for each interval combination.

			Efficacy Rate			
			Low	Moderate	High	Superb
			$(0, e_1)$	(e_1, e_2)	(e_2, e_3)	$(e_3, 1)$
Toxicity Rate	Low	$(0, t_1)$	E	E	E	E
	Moderate	(t_1, t_2)	E	E	E	S
	High	(t_2, t_3)	D	S	S	S
	Unacceptable	$(t_3, 1)$	D	D	D	D

Note: "E", "S" and "D" denote escalation, stay and de-escalation, respectively.

3.3.2.2 Dose-finding Algorithm

Building upon the preset table, we set up a local decision-theoretic framework and derive a Bayes rule. Here, local means that the framework focuses on the optimal decision to be made for the current dose instead of the trial. We show that the Bayes rule is equivalent to computing the joint unit probability mass (JUPM) for the toxicity and efficacy probability intervals. For a given region A, the JUPM is defined as the ratio between the probability of the region and the size of the region. Considering the two-dimensional unit square $(0, 1) \times (0, 1)$ in the real space, the JUPM for each interval combination $(a, b) \times (c, d)$ is

$$\text{JUPM}_{(a,b)}^{(c,d)} \equiv \frac{\Pr\{p_d \in (a, b), q_d \in (c, d) \mid D\}}{(b - a) \times (d - c)}; 0 < a < b < 1; 0 < c < d < 1. \tag{3.3}$$

Here, the numerator, $\Pr\{p_d \in (a, b), q_d \in (c, d) \mid D\}$, is the posterior probability of p_d and q_d falling in the interval (a, b) and (c, d), respectively.

Assume the prior for each p_d follows an indepedent $beta(a_1, b_1)$, and the prior for each q_d follows an independent $beta(a_2, b_2)$, where $beta(\alpha, \beta)$ denotes a beta distribution with mean $\frac{\alpha}{(\alpha+\beta)}$. The posterior distributions for p_d and q_d are $beta(a_1 + x_d, b_1 + n_d - x_d)$ and $beta(a_2 + y_d, b_2 + n_d - y_d)$, respectively.

Based on the posterior distributions, there exists a winning interval combination $(a^*, b^*) \times (c^*, d^*)$ that achieves the maximum JUPM among all the combinations in Table 3.4, and the corresponding decision for that combination is selected for treating the next cohort of patients.

The basic dose-finding concept of TEPI is as follows. Assume that the current patient cohort is treated at dose d. After the current cohort completes DLT and response evaluation, compute the JUPMs for all the interval combinations in Table 3.4. The TEPI design recommends E," S," or D",

corresponding to the combination with the largest JUPM value according to Table 3.4.

In practice, the TEPI design needs to be calibrated according to physicians' needs. This is transparent and requires some effort. The tuning is for the intervals in Table 3.4 so that the dosing decisions are satisfactory to the clinicians.

To enable ethical constraints, below are two additional rules as part of the dose-finding algorithm to exclude any dose with excessive toxicity and any dose with unacceptable efficacy.

Safety and futility rules

- Safety rule: if $\Pr(p_d > p_T \mid x_d, n_d) > p_{cut}$ for a p_{cut} close to 1 (say, 0.95), exclude doses $d, d+1, \cdots, D$, from future use in the trial; treat the next cohort of patients at dose $(d-1)$.

- Futility rule: if $\Pr(q_d < q_E \mid y_d, n_d) > q_{cut}$ for a q_{cut} close to 1 (say, 0.7), where q_E is the minimum acceptable probability of efficacy, then exclude dose d from future use in the trial. Here, q_E is the reference efficacy rate, e.g., the efficacy rate of standard care.

Note that, here we assume the prior for each p_d follows an independent $beta(a_1, b_1)$, and the prior for each q_d follows an independent $beta(a_2, b_2)$, where $beta(\alpha, \beta)$ denotes a beta distribution with mean $\alpha/(\alpha+\beta)$. The posterior distributions for p_d and q_d in the above rules are $beta(a_1 + x_d, b_1 + n_d - x_d)$ and $beta(a_2 + y_d, b_2 + n_d - y_d)$, respectively.

Stopping rules

The trial is stopped if

1. the prespecified maximum total sample size n is reached; or
2. the lowest dose shows excessive toxicity according to the safety rule; In this case, the trial is early stopped and the MTD cannot be determined; or
3. *optional:*
   - *the TEPI decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *the TEPI decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *the TEPI decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

3.3.2.3 Dose Selection

At the end of the trial, TEPI selects the most desirable dose as the OBD based on a utility score that balances the toxicity and efficacy trade-off. The utility score function is defined as $U(p,q) = f_1(p)f_2(q)$ (suppressing dose d in the notation), where p denotes the toxicity rate, and q denotes the efficacy rate.

Both $f_1(\cdot)$ and $f_2(\cdot)$ are truncated linear functions, given by

$$f_1(p) = \begin{cases} 1, & p \in (0, p_1^*]. \\ 1 - \frac{p-p_1^*}{p_2^*-p_1^*}, & p \in (p_1^*, p_2^*), \\ 0, & p \in [p_2^*, 1) \end{cases} \tag{3.4}$$

$$f_2(q) = \begin{cases} 0, & q \in (0, q_1^*]. \\ \frac{q-q_1^*}{q_2^*-q_1^*}, & q \in (q_1^*, q_2^*), \\ 1, & q \in [q_2^*, 1) \end{cases} \tag{3.5}$$

where p^*'s and q^*'s are prespecified cutoff values. The OBD is selected according to the following process.

1. We generate a total of T random samples, $\left\{p_d^{(t)}, t = 1, \cdots, T\right\}$ and $\left\{q_d^{(t)}, t = 1, \cdots, T\right\}$, from the posterior distributions $beta(a_0 + x_d, b_0 + n_d - x_d)$ and $beta(a_0 + y_d, b_0 + n_d - y_d)$ for each dose d, respectively. Here, East Bayes sets $a_0 = b_0 = 0.005$ and $T = 1000$.
2. For toxicity probabilities of all doses in each sample t, $\boldsymbol{p}^{(t)} = (p_1^{(t)}, \cdots, p_D^{(t)})$, we perform isotonic transformation using the pool adjacent violators algorithm (PAVA; Mair et al. 2009) on $\boldsymbol{p}^{(t)}$ to obtain $\widetilde{\boldsymbol{p}}^{(t)} = (\widetilde{p}_1^{(t)}, \cdots, \widetilde{p}_d^{(t)})$, where $\widetilde{p}_i^{(t)} \leq \widetilde{p}_j^{(t)}$ if $i < j$.
3. Let $\mathcal{A} = \{d \mid n_d > 0, d = 1, \cdots, D\}$ denote the candidate dose set from which doses have been excluded according to safety and futility rules, the final selected dose d^* is the one that maximizes utility scores $U(p_d, q_d)$, that is, $d^* = \text{argmax}_{d \in \mathcal{A}} E[U(p_d, q_d) \mid Data]$, where

$$\hat{E}[U(p_d, q_d) \mid Data] = \frac{1}{T} \sum_{t=1}^{T} U(\widetilde{p}_d^{(t)}, q_d^{(t)}).$$

3.3.3 The Probability Intervals of Toxicity and Efficacy (PRINTE) Design

PRINTE (Lin and Ji, 2020a) builing upon previous work in TEPI (Li et al., 2017), is a dose-finding design which utilizes both toxicity and efficacy in making dosing decisions. Similar to TEPI, PRINTE partitions the unit intervals $(0, 1)$ for both the toxicity probability p_i and efficacy probability q_i into subintervals, and makes dosing-decisions based on the posterior probability of the interval combinations. Compared to TEPI, it does not require a physician-elicited decision table, the choice of which could be arbitrary and difficult, and might be subjective to Ockhams razor (Guo et al., 2017b). Instead, PRINTE utilizes a decision principle that is simple and transparent, and is commonly applied in practice.

3.3.3.1 Probability Model

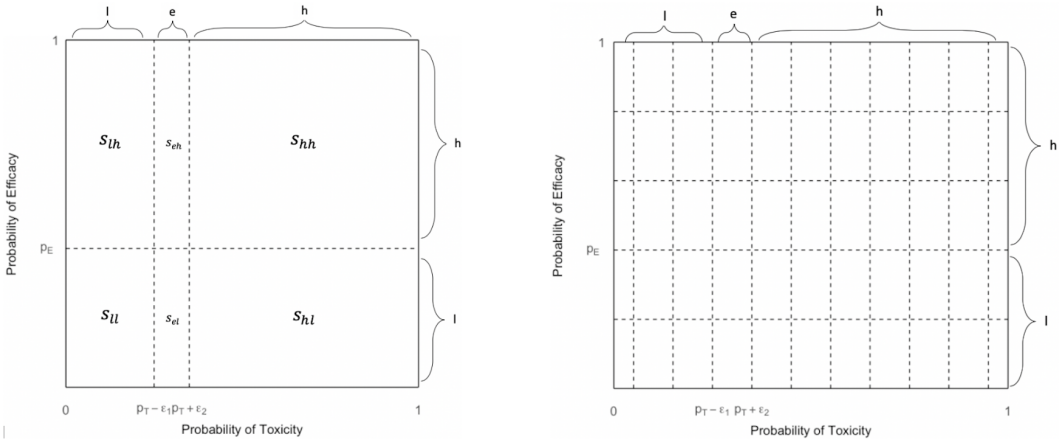
Consider D ascending doses in a single-agent ACT phase I trial. Due to ethical considerations, it is always assumed that the toxicity probability p_d increases with dose level d, that is, $p_1 \leq \cdots \leq p_D$. However, the efficacy probability q_d may increase initially and then reach a plateau from which minimal improvement or even decreasing efficacy may be seen with increasing dose. For this reason, we assume that q_d is not monotone with d, and that p_d and q_d are independent. Suppose that dose d is currently used in the trial and n_d patients have already been allocated to dose d, with x_d and y_d patients experiencing toxicity and efficacy outcomes. Aggregating across all the doses, the trial data are denoted as $Data = \{(n_d, x_d, y_d), d = 1, \cdots, D\}$

Let p_T be the target toxicity probability and p_E be the target efficacy rate. Define the equivalence interval (EI) as $[p_T - \epsilon_1, p_T + \epsilon_2]$ where ϵ_1 and ϵ_2 are two small fractions that allow toxicity probability of MTD to be in a range of values, rather than a single point p_T.

Consider the unit square of $Q = (0, 1) \times (0, 1)$ (here, operation $\times$ represents the Cartesian product) representing the joint probability square of toxicity and efficacy probabilities. For toxicity, there are three probability intervals, $(0, p_T - \epsilon_1)$, $[p_T - \epsilon_1, p_T + \epsilon_2]$, and $(p_T + \epsilon_2, 1)$, which represent the under-dosing, equivalence, and over-dosing intervals. For efficacy, consider two probability intervals, $(0, p_E]$ and $(p_E, 1)$, which corresponds to low and high probability of efficacy. Denote $S_{tox} = \{(0, p_T - \epsilon_1), [p_T - \epsilon_1, p_T + \epsilon_2], (p_T + \epsilon_2, 1)\}$ as the set of three toxicity probability intervals and $S_{eff} = \{(0, p_E], (p_E, 1)\}$ as the set of two efficacy probability intervals. Taking a Cartesian product of the two sets, we obtain a set of six probability rectangles (PRs) in Q, which is given by

$$
\begin{aligned}
S_{joint} &= S_{tox} \times S_{eff} \\
&= \{s_{ll} = (0, p_T - \epsilon_1) \times (0, p_E],\ s_{lh} = (0, p_T - \epsilon_1) \times (p_E, 1), \\
&\quad s_{el} = [p_T - \epsilon_1, p_T + \epsilon_2] \times (0, p_E], s_{eh} = [p_T - \epsilon_1, p_T + \epsilon_2] \times (p_E, 1), \\
&\quad s_{hl} = (p_T + \epsilon_2, 1) \times (0, p_E],\ s_{hh} = (p_T + \epsilon_2, 1) \times (p_E, 1)\},
\end{aligned}
$$

where the two letters l and h denotes low or high, respectively. See Figure 3.19a for a display of the probability rectangles in S_{joint}.



(a) Probability rectangles (PRs)

(b) Probability sub-rectangles (sub-PRs)

Figure 3.19: An example demonstrating the 2-dimensional probability rectangles and sub-rectangles of toxicity and efficacy. (a): The horizontal axis is the probability intervals of toxicity $(0, p_T - \epsilon_1)$, $[p_T - \epsilon_1, p_T + \epsilon_2]$, and $(p_T + \epsilon_2, 1)$. The vertical axis is the probability intervals of efficacy $(0, p_E]$ and $(p_E, 1)$. The Cartesian product of both probability intervals is shown as the 6 probability rectangles (PRs) separated by dashed lines. (b): The horizontal axis is the probability sub-intervals of toxicity, where $(0, p_T - \epsilon_1)$, $[p_T - \epsilon_1, p_T + \epsilon_2]$ and $(p_T + \epsilon_2, 1)$ are further divided into smaller intervals with the same length of $\epsilon_1 + \epsilon_2$. The vertical axis is the probability sub-intervals of efficacy, where $(0, p_E]$ and $(p_E, 1)$ are further divided into multiple smaller intervals with the same length of their maximum common divisor. The Cartesian product of all probability sub-intervals is shown as the probability sub-rectangles (sub-PRs) separated by dashed lines.

Divide the six PRs into sub-PRs with similar area, see Figure 3.19b for an illustration, which is realized by three steps.

1. For the toxicity interval set S_{tox}, divide S_{tox} into sub-intervals given by the length of the

equivalence interval $(\epsilon_1 + \epsilon_2)$. The division is done by keeping the equivalence interval $m_e^t = [p_T - \epsilon_1, p_T + \epsilon_2]$ unchanged, and sub-divide the under-dosing interval $(0, p_T - \epsilon_1)$ and over-dosing interval $(p_T + \epsilon_2, 1)$ into sub-intervals with the length $l_t = \epsilon_1 + \epsilon_2$, except for the sub-intervals on the boundary. Denote the set of all the resulting sub-intervals as $M_{tox} = \{m_{l's}^t, m_e^t, m_{h's}^t\}$, in which $m_{l's}^t$ and $m_{h's}^t$ are the sub-intervals generated by dividing the under-dosing and over-dosing intervals, respectively.

2. For the efficacy interval set S_{eff}, divide two intervals in S_{eff} into sub-intervals with the length l_e,

$$l_e = max\left\{0.10, \frac{\text{gcd}\left(100 * p_E, 100 * (1 - p_E)\right)}{100}\right\},$$

where $\text{gcd}(a, b)$ is the greatest common divisor of a and b. Denote the resulting set of sub-intervals by $M_{eff} = \{m_{l's}^e, m_{h's}^e\}$, where

$$m_{l's}^e = \{(0, p_E - t_1 l_e], \cdots, (p_E - 2l_e, p_E - l_e], (p_E - l_e, p_E]\},$$

$$m_{h's}^e = \{(p_E, p_E + l_e), (p_E + l_e, p_E + 2l_e), \cdots, (p_E + t_2 l_e, 1)\}.$$

Here, t_1 and t_2 are the maximum positive integers such that $p_E - t_1 l_e > 0$ and $p_E + t_2 l_e < 1$, respectively.

3. Take Cartesian product of the set of M_{tox} and M_{eff} to generate a set of two-dimensional sub-PRs of equal area, except for those on the boundary of the toxicity axis next to 0 or 1. These sets are denoted by M_{joint} as illustrated below, where k_{uv}, $u \in \{l, e, h\}$, $v \in \{l, h\}$ denotes the number of sub-PRs in m_{uv}.

$$\begin{aligned} M_{joint} &= M_{tox} \times M_{eff} \\ &= \{m_{l's}^t, m_e^t, m_{h's}^t\} \times \{m_{l's}^e, m_{h's}^e\} \\ &= \{m_{ll} = \{m_{ll}^1, ..., m_{ll}^{k_{ll}}\}, m_{lh} = \{m_{lh}^1, ..., m_{lh}^{k_{lh}}\}, \\ &\quad m_{el} = \{m_{el}^1, ..., m_{el}^{k_{el}}\}, m_{eh} = \{m_{eh}^1, ..., m_{eh}^{k_{eh}}\}, \\ &\quad m_{hl} = \{m_{hl}^1, ..., m_{hl}^{k_{hl}}\}, m_{hh} = \{m_{hh}^1, ..., m_{hh}^{k_{hh}}\}\} \end{aligned}$$

PRINTE treats each sub-PR as a model and considers a model indicator a that takes one of the sub-PRs. Denote m_{uv} as a sub-PR in the set M_{joint}, and define $\{a = m_{uv}\} = \{(p_d, q_d) \in m_{uv}\}$. Embedding the model indicator a into a Bayesian hierarchic model, we compute the posterior probability of each sub-PR given the observed toxicity and efficacy outcomes $\{x_d, y_d\}$, given by $P(a = m_{uv} \mid x_d, y_d, n_d) = Pr((p_d, q_d) \in m_{uv} \mid x_d, y_d, n_d)$. From model selection perspective,

finding the optimal decision is equivalent to selecting the optimal model (sub-PR) that maximizes the marginal posterior model probability.

We further define dose-finding decisions as $a^\star \in \{E, S, D\}$ and maps $a \in \{m_{ll}, m_{lh}, m_{el}, m_{eh}, m_{hl}, m_{hh}\}$ to $a^\star \in \{E, S, D\}$ according to the following rule $\mathscr{R}$.

$$a^\star = \mathscr{R}(a) = \begin{cases} E, & \text{if} \quad a = m_{ll} \\ E, & \text{if} \quad a = m_{el} \quad \text{and} \quad n_{d+1} = 0 \\ S, & \text{if} \quad a = m_{el} \quad \text{and} \quad n_{d+1} > 0 \\ S, & \text{if} \quad a \in \{m_{lh}, m_{eh}\} \\ D, & \text{if} \quad a \in \{m_{hl}, m_{hh}\} \end{cases}$$

The rule $\mathscr{R}$ states that the dosing decisions $\{E, S, D\}$ correspond to the models that describe the toxicity and efficacy probabilities of the dose. According to $\mathscr{R}(a)$, escalation (E) is recommended if toxicity and efficacy are both deemed low; Stay (S) is selected if $n_{d+1} > 0$, toxicity is near the MTD range and efficacy is low, while escalation (E) is recommended if $n_{d+1} = 0$, i.e., dose $(d+1)$ is untried; Stay (S) is selected if either 1) toxicity is low but efficacy is high m_{lh}, or 2) toxicity is near the MTD range and efficacy is high; Lastly, de-escalation D is selected if toxicity is high regardless of efficacy. The goal is to seek an optimal a that leads to an optimal decision $a^\star$.

3.3.3.2 Dose-finding Algorithm

The implementation of PRINTE is simple and transparent. The only required input values are p_T, p_E, and the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$. Once they are provided, optimal decisions $a^{opt\star}$ can be calculated for all possible toxicity and efficacy outcomes at a given dose. Suppose that the current dose is d, $d \in \{1, ..., D\}$. Record $\{x_d, y_d, n_d\}$ and calculate the marginal model posterior probabilities $\Pr(a \mid x_d, y_d, n_d)$, and then the optimal decision $a^{opt\star}$ can be determined. The next cohort of patients is allocated to $\{max(1, d-1), d, min(d+1, D)\}$ according to $a^{opt\star}$.

Safety and futility rules

- Safety rule: if $\Pr(p_d > p_T \mid x_d, n_d) > p_{cut}$ for a p_{cut} close to 1 (say, 0.95), exclude doses $d, d+1, \cdots, D$, from future use in the trial; treat the next cohort of patients at dose $(d-1)$.
- Futility rule: if $\Pr(q_d < q_E \mid y_d, n_d) > q_{cut}$ for a q_{cut} close to 1 (say, 0.7), where q_E is the minimum acceptable probability of efficacy, then exclude dose d from future use in the trial. Here, q_E is the reference efficacy rate, e.g., the efficacy rate of standard care.

Note that, here we assume the prior for each p_d follows an independent $beta(a_1, b_1)$, and the prior for each q_d follows an independent $beta(a_2, b_2)$, where $beta(\alpha, \beta)$ denotes a beta distribution with mean $\alpha/(\alpha+\beta)$. The posterior distributions for p_d and q_d in the above rules are $beta(a_1 + x_d, b_1 + n_d - x_d)$ and $beta(a_2 + y_d, b_2 + n_d - y_d)$, respectively.

Stopping rules

The trial is stopped if

1. the prespecified maximum total sample size n is reached; or
2. the lowest dose shows excessive toxicity according to the safety rule; In this case, the trial is early stopped and the MTD cannot be determined; or
3. *optional:*
   - *the PRINTE decision is "S", to stay at the current dose, and the current dose has enrolled K patients;*
   - *the PRINTE decision is "E", to escalate to the next higher dose, and that next higher dose has enrolled K patients;*
   - *the PRINTE decision is "D", to de-escalate to the previous lower dose, and that previous lower dose has enrolled K patients.*

3.3.3.3 Dose Selection

At the end of the trial, PRINTE chooses the OBD using a joint utility score $U(p, q) = f_1(p)f_2(q)$ (suppressing dose d in the notation), which takes the product of toxicity utility $f_1(p)$ in (3.6) and efficacy utility $f_2(q)$ in (3.7).

$$f_1(p) = \begin{cases} 1, & p \in (0, p_1^*), \\ 1 - \frac{p - p_1^*}{p_2^* - p_1^*}, & p \in (p_1^*, p_2^*), \\ 0, & p \in (p_2^*, 1). \end{cases} \tag{3.6}$$

$$f_2(q) = \begin{cases} 0, & q \in (0, q_1^*), \\ \frac{q - q_1^*}{q_2^* - q_1^*}, & q \in (q_1^*, q_2^*), \\ 1, & q \in (q_2^*, 1). \end{cases} \tag{3.7}$$

For toxicity, define two thresholds p_1^* and p_2^* such that the toxicity utility score is 1 when $p < p_1^*$, 0 when $p > p_2^*$, and linearly decreases when p is between (p_1^*, p_2^*). For efficacy, define

two thresholds q_1^* and q_2^* such that the efficacy utility score is 0 when $q < q_1^*$, is 1 when $q > q_2^*$, and linearly increases when q is between (q_1^*, q_2^*). The OBD is selected according to the following process.

1. We generate a total of T random samples, $\left\{p_d^{(t)}, t = 1, \cdots, T\right\}$ and $\left\{q_d^{(t)}, t = 1, \cdots, T\right\}$, from the posterior distributions $beta(a_0 + x_d, b_0 + n_d - x_d)$ and $beta(a_0 + y_d, b_0 + n_d - y_d)$ for each dose d, respectively. Here, East Bayes sets $a_0 = b_0 = 0.005$ and $T = 1000$.
2. For toxicity probabilities of all doses in each sample t, $\boldsymbol{p}^{(t)} = (p_1^{(t)}, \cdots, p_D^{(t)})$, we perform isotonic transformation using the pool adjacent violators algorithm (PAVA; Mair et al. 2009) on $\boldsymbol{p}^{(t)}$ to obtain $\widetilde{\boldsymbol{p}}^{(t)} = (\widetilde{p}_1^{(t)}, \cdots, \widetilde{p}_d^{(t)})$, where $\widetilde{p}_i^{(t)} \leq \widetilde{p}_j^{(t)}$ if $i < j$.
3. We propose a probabilistic inference for selecting the OBD and avoid selecting doses with low utility. Define an admissible probability region (APR) $A(p, q) = \{(p, q) \mid p \in (0, p_T], q \in [q_E, 1)\}$. Then the OBD is selected only from the candidate dose set $\mathcal{A}$,

$$\mathcal{A} = \left\{d \mid p_{\text{in},d} \geq p_{\text{grad}}, n_d > 0, d = 1, \cdots, D\right\},$$

   where $p_{\text{in},d} = \Pr\{(p_d, q_d) \in \text{APR} \mid Data\}$ is the posterior probability that dose d belongs to APR and p_{grad} is a small value (say, 0.1). We use a simple a simple numerical approximation approach to compute $p_{\text{in},d}$ given by

$$\hat{p}_{\text{in},d} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left\{(\widetilde{p}_d^{(t)}, q_d^{(t)}) \in \text{APR}\right\}.$$

4. The final selected dose d^* is the one that maximizes the utility score $U(p_d, q_d)$. That is, $d^* = \text{argmax}_{d \in \mathcal{A}} \hat{E}[U(p_d, q_d) \mid Data]$, where

$$\hat{E}[U(p_d, q_d) \mid Data] = \frac{1}{T} \sum_{t=1}^{T} U(\widetilde{p}_d^{(t)}, q_d^{(t)}).$$

3.3.4 The EfficacyToxicity (EffTox) Trade-Offs-Based Design

EffTox, proposed in Thall and Cook (2004), is an outcome-adaptive, model-based Bayesian procedure that chooses doses of an experimental agent for successive patient cohorts in a clinical trial based on both efficacy (E) and toxicity (T) outcomes. EffTox models the dose-efficacy and dose-toxicity relationship respectively using two different dose-response curves. Based on accumulating efficacy and toxicity data over the trial, EffTox continuously updates the parameters of the dose-response models. The desirability of each dose x is evaluated by using a family of contours characterizing the trade-off between E and T, and patients are assigned to the most desirable dose in cohorts.

3.3.4.1 Dose-Outcome Models

Assume D dose $s_1, \cdots, s_D$ to be considered in the trial, and code dose as

$$x_d = \log(s_d) - D^{-1} \sum_{k=1}^{D} \log(s_k) \tag{3.8}$$

for use in the regression models. If $0 = s_1 < s_2$, first add s_2 to each s_d before taking logs. Let $\pi(x, \boldsymbol{\theta}) = \{\pi_E(x, \boldsymbol{\theta}), \pi_T(x, \boldsymbol{\theta})\}$ be the probabilities of efficacy and tocixity, where x denotes dose and $\boldsymbol{\theta}$ is the model parameter vector.

Given the current interim trial data $\mathcal{D}$, define x to be an acceptable dose if

$$\Pr\{\pi_E(x, \theta) > q_E \mid \mathcal{D}\} > 1 - q_{cut} \tag{3.9}$$

and

$$\Pr\{\pi_T(x, \theta) < p_T \mid \mathcal{D}\} > 1 - p_{cut}, \tag{3.10}$$

where q_E and p_T are fixed lower and upper limits specified by the physician, and q_{cut} and p_{cut} are fixed probability cutoffs.

For toxicity, assume $\text{logit}(\pi_T(x, \boldsymbol{\theta})) = \mu_T + x\beta_T$, in which we set $\beta_T > 0$ to meet the monotonic dose-toxicity assumption. For efficacy, to allow a wide variety of possible doseresponse relationships, assume $\text{logit}(\pi_E(x, \boldsymbol{\theta})) = \mu_E + x\beta_{E,1} + x^2\beta_{E,2}$. For simplicity, temporarily suppress $(x, \boldsymbol{\theta})$. The joint outcome model is given by

$$\pi_{a,b} = (\pi_E)^a(1-\pi_E)^{1-a}(\pi_T)^b(1-\pi_T)^{1-b} + (-1)^{a+b}\pi_E(1-\pi_E)\pi_T(1-\pi_T)\left(\frac{e^\psi - 1}{e^\psi + 1}\right) \tag{3.11}$$

for $a, b \in \{0, 1\}$ and real-valued ψ. Thus, $\boldsymbol{\theta} = (\mu_T, \beta_T, \mu_E, \beta_{E,1}, \beta_{E,2}, \psi)$. Since β_T should be greater than 0, we assume that β_T is lognormally distributed, with mean $\tilde{\mu}_{\beta_T}$ and standard deviation

$\tilde{\sigma}_{\beta_T}$. Except for β_T, we assume that each component θ_l of $\boldsymbol{\theta}$ is normally distributed with mean $\tilde{\mu}_l$ and standard deviation $\tilde{\sigma}_l$, denoted as $\theta_l \sim N(\tilde{\mu}_l, \tilde{\sigma}_l)$.

The likelihood for a single patient treated at dose x is $\mathcal{L}(\boldsymbol{Y}, x \mid \theta) = \prod_{a=0}^{1}\prod_{b=0}^{1}\{\pi_{a,b}(x,\boldsymbol{\theta})\}^{I\{\boldsymbol{Y}=(a,b)\}}$. Denoting the data for the first n patients in the trial by $\mathcal{D}_n$, for $1 \leq n \leq N$, the likelihood is $\mathcal{L}_n(\mathcal{D}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n}\mathcal{L}(\boldsymbol{Y}_i, x_{(i)} \mid \boldsymbol{\theta})$, where $\boldsymbol{Y}_i$ and $x_{(i)}$ denote the ith patients outcome and dose.

3.3.4.2 EfficacyToxicity Trade-Off Contours

To determine the desirability of each dose, the EffTox design constructs a efficacy-toxicity desirability contour, $\mathcal{C}$, in the two-dimensional domain $\boldsymbol{\Pi} = [0,1]^2$ by fitting a curve to target values of $\boldsymbol{\pi}$ elicited from the physician. The contour $\mathcal{C}$ is then used to construct a family of desirability contours such that all $\boldsymbol{\pi}$ on the same contour are equally desirable. Because the family of contours partitions $\boldsymbol{\Pi}$, this construction provides a basis for comparing doses in terms of their posterior means, $E\{\boldsymbol{\pi}(x,\boldsymbol{\theta}) \mid \mathcal{D}\}$.

To construct $\mathcal{C}$, we first elicit three target values, $\{\boldsymbol{\pi}_1^*, \boldsymbol{\pi}_2^*, \boldsymbol{\pi}_3^*\}$, which the physician considers equally desirable. First, elicit a desirable trade-off target, $\boldsymbol{\pi}_1^* = (\pi_{1,E}^*, \pi_{1,T}^*) = (\pi_{1,E}^*, 0)$, in the case where toxicity has probability 0. That is, elicit the smallest efficacy probability, $\pi_{1,E}^*$, that the physician would consider desirable if toxicity were impossible. Next, elicit $\boldsymbol{\pi}_2^*$ having the same desirability as $\boldsymbol{\pi}_1^*$ by asking the physician what the maximum value of π_T may be if $\pi_E = 1$. Given these two equally desirable extremes, elicit a third pair, $\boldsymbol{\pi}_3^*$, that is equally desirable but is intermediate between $\boldsymbol{\pi}_1^*$ and $\boldsymbol{\pi}_2^*$.

The desirability function of $(\pi_E, \pi_T) = \boldsymbol{\pi} \in [0,1]^2$ is defined to be

$$\begin{aligned}\delta(\pi_E, \pi_T) &= 1 - \|(\pi_E, \pi_T) - (1,0)\|_p \\ &= 1 - \left\{\left(\frac{\pi_E - 1}{\pi_{1,E}^* - 1}\right)^p + \left(\frac{\pi_T - 0}{\pi_{2,T}^* - 0}\right)^p\right\}^{1/p}\end{aligned} \tag{3.12}$$

where $p > 0$. Solve $\delta(\pi_{E,3}^*, \pi_{T,3}^*) = 0$ for p using the bisection method, wherein intervals known to bracket the solution are successively refined (Peter et al., 2014). This gives $\delta(\boldsymbol{\pi}) = 0$ on $\mathcal{C}$ with $\delta(\boldsymbol{\pi})$ increasing as $\boldsymbol{\pi}$ moves along any straight line from a point in $[0,1]^2$ to the ideal pair $(\pi_E, \pi_T) = (1,0)$. After solving for p, the desirability measure can be computed for any point (π_E, π_T) using formula (3.12).

The following definition exploits this structure to induce an ordering on the set of doses.

DEFINITION: *Given $\mathcal{D}$ and x, the desirability, $\delta(x, \mathcal{D})$, of x is the desirability of the posterior mean $E\{\boldsymbol{\pi}(x,\boldsymbol{\theta}) \mid \mathcal{D}\}$.*

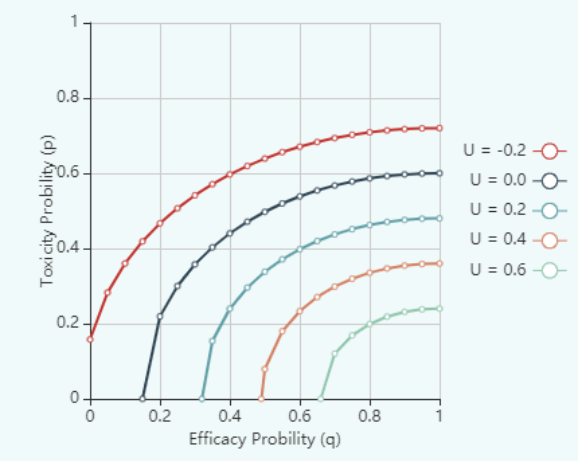


Figure 3.20: Example of efficacy-toxicity desirability contours. The contour $\mathcal{C}$ is the line with desirability equals to 0 (U = 0.0).

To apply this during the trial, after the most recent cohorts data have been incorporated into $\mathcal{D}$, for each x, $(\pi_E, \pi_T) = E\{\boldsymbol{\pi}(x, \boldsymbol{\theta}) \mid \mathcal{D}\}$ is first computed, and then the desirability of x is computed by formula (3.12). Among the doses with acceptable efficacy and toxicity, the dose that maximizes $\delta(x, \mathcal{D})$ is selected.

3.3.4.3 The Trade-Off-Based Algorithm

Initially, the physician must provide a set of doses, a starting dose for the first cohort, N, c, and the limits q_E and p_T used in the acceptability criteria (3.9) and (3.10). The trade-off targets $\{\boldsymbol{\pi}_1^*, \boldsymbol{\pi}_2^*, \boldsymbol{\pi}_3^*\}$ then must be elicited in order to construct $\mathcal{C}$ and the family of trade-off contours. The probability cut-offs q_{cut} and p_{cut} in (3.9) and (3.10) are determined, using preliminary computer simulation results, to obtain a design with desirable operating characteristics. Given this structure, the dose-finding algorithm proceeds as follows:

1. Treat the first cohort at the starting dose specified by the physician.
2. For each cohort after the first, $x \in A(\mathcal{D})$ if x satisfies both (3.9) and (3.10), or if x is the lowest untried dose above the starting dose and it satisfies (3.10).
3. If $A(\mathcal{D}) \neq \phi$, then the next cohort is treated at the most desirable $x \in A(\mathcal{D})$, subject to the constraint that no untried dose may be skipped when escalating.
4. If $A(\mathcal{D}) = \phi$, then the trial is terminated and no dose is selected.
5. If the trial is not stopped early and $A(\mathcal{D}_N) \neq \phi$ at the end of the trial, then the dose $x \in$

$A(\mathcal{D}_N)$ maximizing $\delta(x, \mathcal{D}_N)$ is selected.

3.3.5 The Utility-Based Bayesian Optimal Interval (U-BOIN) Design

U-BOIN (Zhou et al., 2019b) is a model-based design that jointly models toxicity and efficacy using a multinomial-Dirichlet model and employ a utility function to measure dose risk-benefit trade-off. The design consists of two seamless stages. In stage I, the Bayesian optimal interval (BOIN) design (Liu and Yuan, 2015b) is used to quickly explore the dose space and collect preliminary toxicity and efficacy data. In stage II, the posterior estimate of the utility for each dose is continuously updated using accumulating efficacy and toxicity data, and the posterior estimate is used to direct patient allocation and OBD selection.

3.3.5.1 Efficacy-Toxicity Model

Consider a phase I/II trial with J doses under investigation. Let Y_E denote the binary efficacy endpoint, where $Y_E = 1$ denotes response, and 0 otherwise; let Y_T denote the binary toxicity endpoint, where $Y_T = 1$ denotes DLT, and 0 otherwise. The bivariate discrete outcome (Y_E, Y_T) can be equivalently represented by a single variable Y with $2 \times 2 = 4$ levels, with $Y = 1$, if $(Y_E, Y_T) = (0, 1)$; $Y = 2$, if $(Y_E, Y_T) = (0, 0)$; $Y = 3$, if $(Y_E, Y_T) = (1, 1)$; and $Y = 4$, if $(Y_E, Y_T) = (1, 0)$. Here $Y = 1$ is the least favorable clinical outcome (DLT, no efficacy), and $Y = 4$ denotes the most favorable clinical outcome (No DLT, efficacy).

Define $\pi_{jk} = Pr(Y = k \mid d = j)$, $k = 1, \cdots, 4$ and $j = 1, \cdots, J$, with $\sum_{k=1}^{4} \pi_{jk} = 1$, where d denotes the dose level. Assume that Y follows a Dirichlet-multinomial model as follows:

$$Y = k \mid d = j \sim \text{Multinomial}(\pi_{j1}, ..., \pi_{j4}) \tag{3.13}$$

$$(\pi_{j1}, ..., \pi_{j4}) \sim \text{Dirichlet}(a_1, ..., a_4) \tag{3.14}$$

where $a_1, \cdots, a_4 > 0$ are hyperparameters. East Bayes sets $a_k = \frac{1}{4}$, $k = 1, \cdots, 4$, as the default values, such that the prior is vague and equivalent to an effective sample size of 1.

Assume that n_j patients have been treated at dose $d = j$, among whom n_{jk} patients had outcome $Y = k$, where $n_j = \sum_{k=1}^{4} n_{jk}$. Denote $D_j = (n_{j1}, \cdots, n_{j4})$, and the posterior distribution of $\boldsymbol{\pi}_j = (\pi_{j1}, \cdots, \pi_{j4})$ is

$$\boldsymbol{\pi}_j \mid D_j \sim \text{Dirichlet}(a_1 + n_{j1}, \cdots, a_4 + n_{j4}). \tag{3.15}$$

3.3.5.2 Utility

Let ψ_k denote the utility value ascribed to outcome $Y = k$, $k = 1, \cdots, 4$, which can be elicited from physicians to reflect the risk-benefit trade-off underlying their medical decisions using the following

procedures.

- Fix the value of the utility for the least desirable outcome $Y = 1$ as $\psi_1 = 0$, and for the most desirable outcome $Y = 4$ as $\psi_4 = 1$.
- Ask the clinician to use these two utilities as a reference to score the utility values ψ_2, ψ_3 for the other 2 possible outcomes $Y = 2, 3$ to quantify the risk-benefit trade-off under each outcome.

Table 3.5 shows two examples of the utility function.

Table 3.5: Examples of utility.

(a) Example 1

	$Y_T = 1$	$Y_T = 0$
$Y_E = 0$	$\psi_1 = 0$	$\psi_2 = 0.3$
$Y_E = 1$	$\psi_3 = 0.5$	$\psi_4 = 1$

(b) Example 2

	$Y_T = 1$	$Y_T = 0$
$Y_E = 0$	$\psi_1 = 0$	$\psi_2 = 0.3$
$Y_E = 1$	$\psi_3 = 0.65$	$\psi_4 = 1$

Example 1 has utility values $\{\psi_1 = 0, \psi_2 = 0.3, \psi_3 = 0.5, \psi_4 = 1\}$ for the outcomes $\{(Y_E = 0, Y_T = 1), (Y_E = 0, Y_T = 0), (Y_E = 1, Y_T = 1), (Y_E = 1, Y_T = 0)\}$, respectively. Compared to example 1, example 2 rewards the response (i.e., $Y_E = 1$) more, in the presence of DLT (i.e., $Y_T = 1$), by assigning a larger value to ψ_3 (0.65 versus 0.50). This is appropriate for a trial where toxicity can be well managed and efficacy response is highly desirable (e.g., leading to long survival).

Given the values of ψ_k, the true mean utility for dose j is given by

$$U_j = \sum_{k=1}^{4} \psi_k \pi_{jk}. \tag{3.16}$$

Since the true mean utility U_j depends on π_{jk}, which is unknown, it is estimated based on the observed data. Given the interim data $D = \{D_j\}$, the estimate of mean utility is given by

$$\hat{U}_j = \sum_{k=1}^{4} \psi_k E(\pi_{jk} \mid D). \tag{3.17}$$

3.3.5.3 Optimal Biological Dose

Let p_T denote the maximum tolerable DLT rate, and q_E the lowest acceptable response rate. Let $\pi_{T,j} = \pi_{j1} + \pi_{j3} = Pr(Y_T = 1 \mid d = j)$ and $\pi_{E,j} = \pi_{j3} + \pi_{j4} = Pr(Y_E = 1 \mid d = j)$. Define

that dose j is inadmissible, if it meets either one or both of the following two criteria:

$$\Pr(\pi_{T,j} > p_T \mid D) > p_{cut} \tag{3.18}$$

$$\Pr(\pi_{E,j} < q_E \mid D) > q_{cut} \tag{3.19}$$

where p_{cut} and q_{cut} are probability cutoffs. According to (3.13) and (3.14), $\pi_{T,j}$ and $\pi_{E,j}$ follow posterior beta distributions, given by

$$\pi_{T,j} \mid D \sim \text{Beta}(a_1 + a_3 + n_{j1} + n_{j3}, a_2 + a_4 + n_{j2} + n_{j4}),$$

$$\pi_{E,j} \mid D \sim \text{Beta}(a_3 + a_4 + n_{j3} + n_{j4}, a_1 + a_2 + n_{j1} + n_{j2}).$$

The admissible dose is then defined as the dose for which none of the criteria (3.18) and (3.19) is satisfied. Define the OBD as the dose that is admissible and has the highest utility value, i.e.,

$$\text{OBD} = \underset{j \in \mathcal{A}}{\arg\max}(U_j) \tag{3.20}$$

where $\mathcal{A}$ denotes the set of admissible doses.

3.3.5.4 Dose-finding Algorithm

The U-BOIN design consists of two seamless stages (Figure 3.21). The objective of stage I is to quickly explore the dose space to identify a set of admissible doses that are reasonably efficacious and safe for stage II. In stage I, dose escalation is conducted based on only the toxicity outcome. However, efficacy data are also collected and will be used for decision making in stage II. Stage I dose escalation/de-escalation is guided by the BOIN design (Liu and Yuan, 2015b). Due to very limited data and large uncertainty, for patient safety, set the target DLT rate $\phi_T = p_T - 0.05$, slightly lower than the maximum tolerable DLT rate p_T, to ensure that stage I dose exploration concentrates around up to, but not exceeding p_T. Let $\hat{\pi}_{T,j}$ denote the empirical (or maximum likelihood) estimate of $\pi_{T,j}$, given by $\hat{\pi}_{T,j} = \frac{m_j}{n_j}$ where m_j is the number of patients who experienced DLT at the dose level j; and let λ_e and λ_d denote the predetermined optimal escalation boundary and de-escalation boundary. Table 3.6 provides the values of λ_e and λ_d for the commonly used target DLT rate ϕ_T. See the work of Liu and Yuan (2015b) for the derivation and formula to calculate λ_e and λ_d. The dose-finding algorithm in stage I proceeds as follows.

Ia. Patients in the first cohort are treated at dose level 1 or a prespecified starting dose.

Ib. Suppose j is the current dose; use the following rules to assign a dose to the next cohort of patients:

  - Escalate the dose to $j+1$ if $\hat{\pi}_{T,j} \leq \lambda_e$.
  - De-escalate the dose to $j-1$ if $\hat{\pi}_{T,j} \geq \lambda_d$.
  - Otherwise, stay at the current dose j.

Ic. Repeat step Ib until the number of patients treated on one of the doses reaches s_1, and then move to stage II.

In stage I, following the BOIN design, if $\Pr(\pi_{T,j} > p_T \mid m_j, n_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial; the trial is terminated if the lowest dose level is eliminated, where $\Pr(\pi_{T,j} \geq p_T \mid m_j, n_j) > 0.95$ is evaluated based on a beta-binomial model with the uniform prior.

Stage II proceeds as follows.

IIa. Let j^* denote the highest dose level that has been tried. If $\hat{\pi}_{T,j^*} \leq \lambda_e$ and j^* is not the highest dose in the trial, escalate the dose to (j^*+1) for treating the next cohort of patients; otherwise, proceed to step IIb.

IIb. Given the observed interim data D collected in both stages I and II, determine the admissible dose set $\mathcal{A}$ from dose $1, \cdots, j^*$, where none of the criteria (3.18) and (3.19) is satisfied for each dose in $\mathcal{A}$. If no dose is admissible, terminate the trial and no dose should be selected as the OBD. Otherwise, assign the next cohort of patients to a dose in $\mathcal{A}$. In East Bayes, there are two methods to assign the next cohort,

  - Pick The Winner, assigning to dose $j \in \mathcal{A}$ that has the largest posterior mean utility.
  - Adaptive Randomization, adaptively randomizing the next cohort of patients to dose $j \in \mathcal{A}$, with probability ω_j proportional to its posterior mean utility, i.e.,

$$\omega_j = \frac{U_j}{\sum_{j \in \mathcal{A}} U_j}.$$

IIc. Repeat steps IIa and IIb until reaching the prespecified maximum sample size N or the number of patients treated at one of the doses in stage II reach s_2 (Zhou et al. (2019b) recommends that $s_2 > s_1$), and then select the OBD following the rules in §3.3.5.3.

Table 3.6: Dose escalation and de-escalation boundaries of the Bayesian optimal interval design

	Target DLT rate (ϕ_T)					
Boundaries	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
λ_e (escalation)	0.118	0.157	0.197	0.236	0.276	0.316
λ_d (de-escalation)	0.179	0.238	0.298	0.358	0.419	0.480

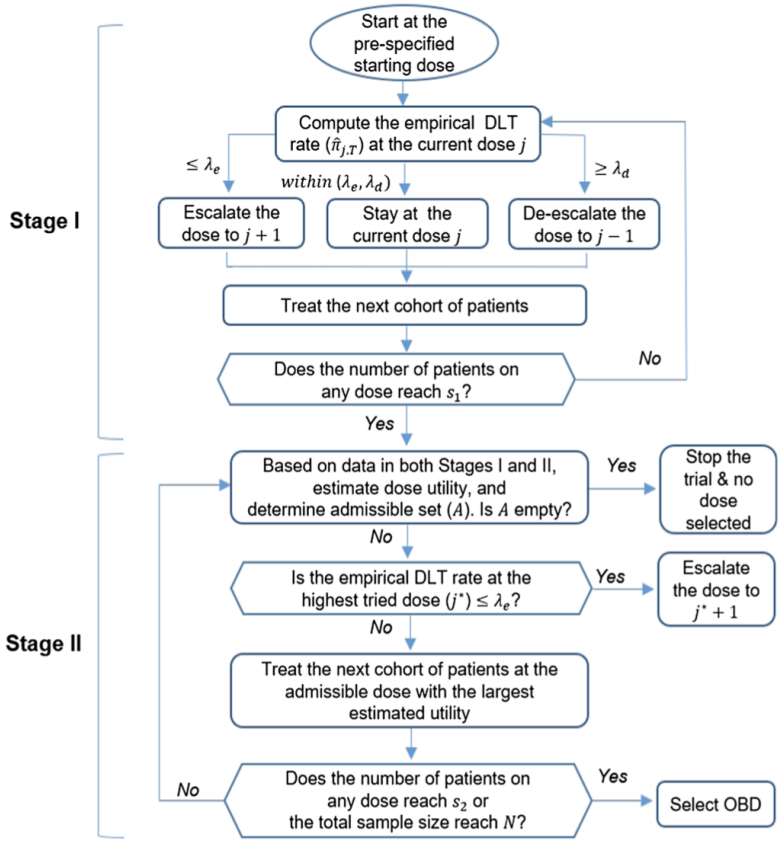


Figure 3.21: Diagram of the utility-based Bayesian optimal interval (U-BOIN) design.

4. Dual-Agents Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment

4.1 Introduction

Combination therapy refers to the use of more than one drug in patient care and is an important therapeutics in many disease settings, including cancer, cardiovascular disease, and infectious disease. In 2013, FDA issued the guidance "Codevelopment of Two or More New Investigational Drugs for Use in Combination" (FDA, 2013), which stated that: "the use of combinations of drugs directed at multiple therapeutic targets can improve treatment response, minimize development of resistance or adverse events". There is growing interest in the development of new investigational drug combinations.

One of the challenges in combination therapy development is to find the optimal dose of each drug when using in combination. Due to the unknown potential interactions between drugs (synergy, antagonism or no interaction), the optimal dose combination might differ from the combination of the optimal dose of each drug when used alone. In this module, we mainly pay attention to the phase I dose-finding trials in oncology, especially dose-finding trials for two agents, with the goal to capture the dose-toxicity relationship for drug combinations and to identify one or more maximum tolerated dose combination (MTDC) or a MTD contour. Only the toxicity outcome, such as dose limiting toxicity (DLT) is considered in this module. A scientific way of characterizing the drug combination-toxicity profile is to test all possible combinations of candidate dose levels of two drugs. However such an approach might be impractical because the number of combinations could be too large for an early-phase trial. For example, if two drugs are to be investigated, each with 3 dose levels, there will be a total of $3 \times 3 = 9$ possible combinations. If more than two drugs are involved, this number grows exponentially to dozens or hundreds. In practice, trialists often escalate the dose level of one drug by holding the dose of another drug at a fixed level. For example, in a

phase I trial of a newly targeted monoclonal antibody (mAb) combined with a PD-1 inhibitor, say pembrolizumab, the dose of PD-1 is often fixed at the approved level (say, 3 mg/kg) and the the dose levels of mAb are varied. If so, some single-agent dose-finding designs, such as mTPI-2 (Guo et al., 2017b) and i3+3 (Liu et al., 2020), could be adopted. However, such an approach may miss the global optimal dose combination since one drug is always at a fixed dose. For example, the optimal dose level of PD-1 when administrated in combination with the mAb might be 1 mg/kg, rather than 3 mg/kg. To this end, "single-agent" dose-finding designs might not be the most scientific way to identify the dual-agents optimal dose.

How to efficiently explore the drug combination-toxicity profile is a statistical problem that requires effective modeling and decision making. In recent years, a large number of designs have been proposed to find one or more maximum tolerated dose combination (MTDC) of two agents, for example, Lyu et al. (2019); Tighiouart et al. (2017); Wages et al. (2017); Lin and Yin (2016); Wages (2017); Mander and Sweeting (2015); Neuenschwander et al. (2015); Cai et al. (2014a); Riviere et al. (2014); Tighiouart et al. (2014); Wages and Conaway (2014); Shi and Yin (2013); Braun and Wang (2010); Yin and Yuan (2009); Conaway et al. (2004) etc. The MTDC is defined as the highest dose combination at which the probability that a patient experiences the DLT is closest to or less than a pre-specified target rate p_T, which is usually determined by physicians or clinical teams, say $p_T = 30\%$. Some of these designs have been applied to real-world trials. For example, a combination dose-finding trial (NCT02366819) uses the CI3+3 design based on the research of our team.

Here, we describe a module in East Bayes, **Dual-Agents Cohort-Based Designs**, which includes the Bayesian logistic regression model (BLRM) for two agents (BLRM-2d) (Neuenschwander et al., 2015), the product of independent beta probabilities dose escalation (PIPE) design (Mander and Sweeting, 2015), and a novel design called Combo i3+3 (CI3+3).

Hereinafter, we use "drug" and "agent", "dose" and "dose combination", interchangeably.

4.2 User Interface and Tutorial

4.2.1 Overview

Entering the **Dual-Agents Cohort-Based Designs** page, users will see two main tabs: **Simulation Setup** and **Simulation Results**. These two tabs allow users to conduct simulations and visualize/download simulation results. The **Simulation Setup** tab requires three steps to set up simulations using one or more designs (Figure 4.1): **Step 1: Set trial parameters**; **Step 2: Select designs**; and **Step 3: Generate scenarios**. Upon completing steps 1-3, users click the "Launch Simulation" button at the bottom of the page. User may also click the "Reset" button to clear all settings. After the simulation is launched, the results of simulations will be displayed in the **Simulation Results** tab. The simulation process can be monitored in real time at the top of the **Simulation Results** tab. Detailed steps of using this module are elaborated in §4.2.2-§4.2.3.

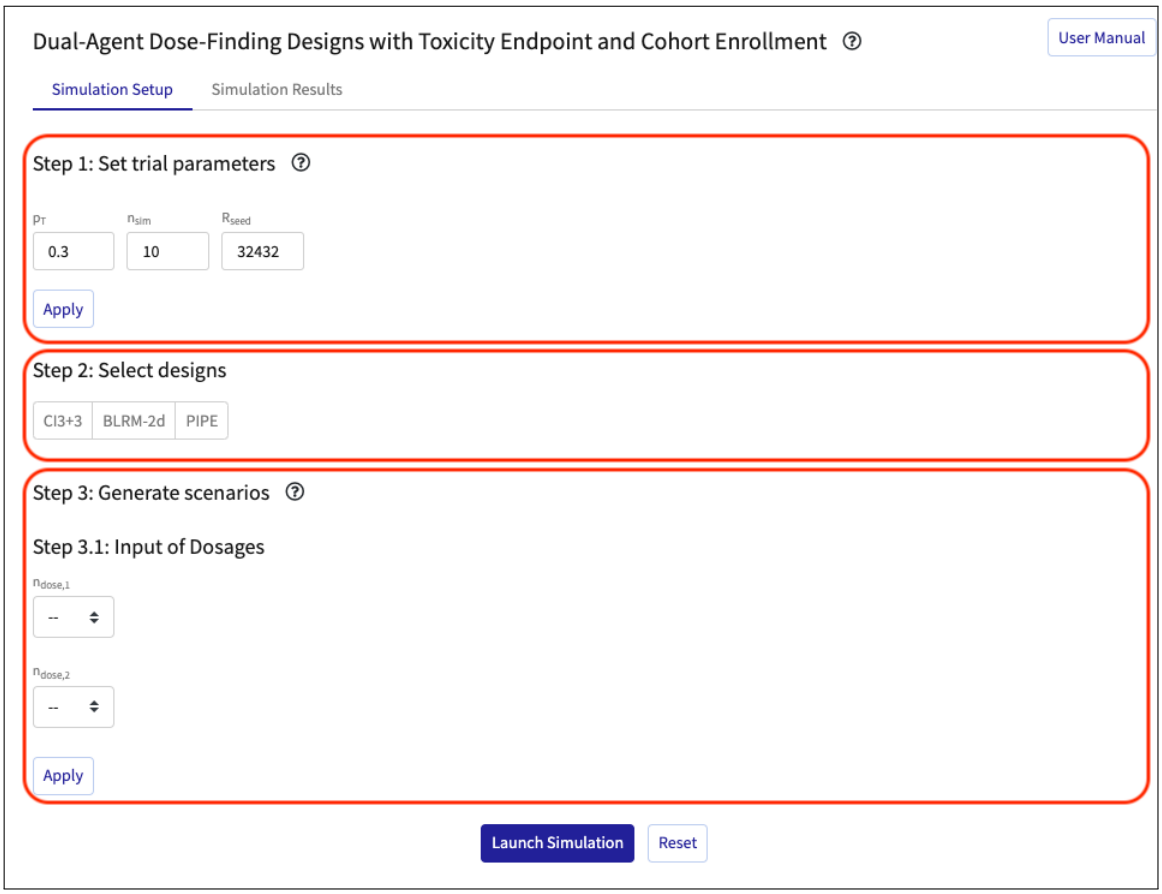


Figure 4.1: Simulation Setup in the **Dual-Agents Cohort-Based Designs** module.

4.2.2 Simulation Setup

In the module of **Dual-Agents Cohort-Based Designs**, East Bayes provides three designs, BLRM-2d, PIPE, and CI3+3, for simulation. Users can choose up to four design configurations for simultaneous comparison in the **Simulation Setup** tab each time. A design configuration means a design such as CI3+3, along with the designs settings, such as sample size. Request to allow more than four design configurations by emailing support@cytel.com.

4.2.2.1 Step 1: Set trial parameters

Specify the target toxicity probability (p_T), number of simulations (n_{sim}) and random seed of simulation (R_{seed}) for the simulation trials. See Figure 4.2. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameter. The detailed description of the above three input parameters is in Table 4.1.

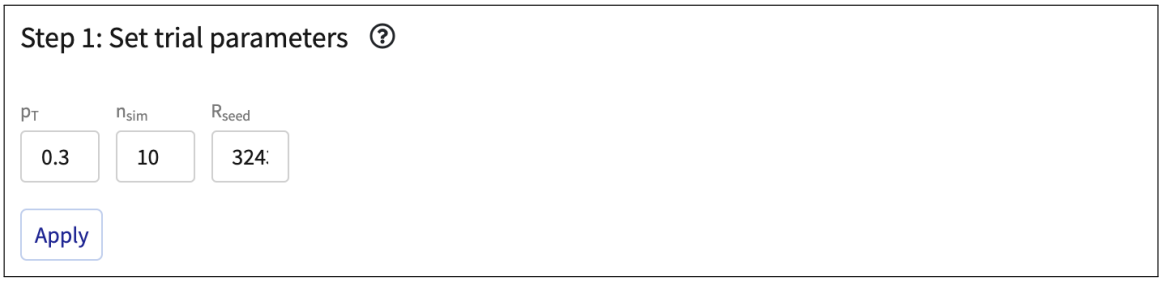


Figure 4.2: Set trial parameters in the **Dual-Agents Cohort-Based Designs** module.

4.2.2.2 Step 2: Select designs

To select a design, click the button with the design's name on it. Up to four design configurations may be selected for comparison.

Click the "Delete" button to remove the selected designs.

Design's parameters can be modified in the input box of corresponding row. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameter. See detailed parameter descriptions in Table 4.2.

Table 4.1: Input parameters for trails parameters in the **Dual-Agents Cohort-Based Designs** module.

Notation	Parameters	Description
p_T	Target toxicity probability	The target toxicity probability of the maximum tolerated dose (MTD). The main objective of phase I clinical trials is to find the highest dose with a toxicity probability closest to or lower than p_T. Default value is 0.3.
n_{sim}	The number of simulated trials	The maximum number of simulated trials allowed is 10,000. Default value is 1,000.
R_{seed}	The random seed of simulation	A random seed is a number used to initialize a pseudorandom number generator in the simulation. Default value is 32432.

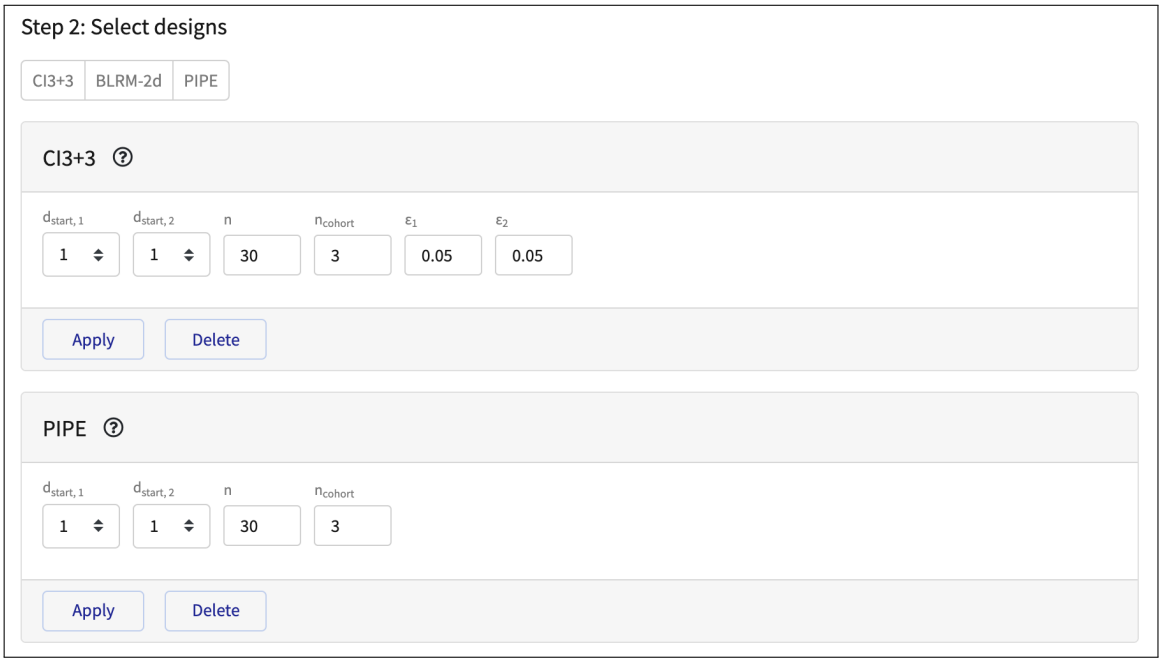


Figure 4.3: Add designs in the **Dual-Agents Cohort-Based Designs** module.

Table 4.2: Input parameters for designs in the **Dual-Agents Cohort-Based Designs** module.

Notation	**Parameters**	**Description**
n (all designs)	Sample size	The maximum number of patients to be treated in the trial. The upper limit is set at 100 since the number of patients that are enrolled in phase I clinical trial is typically small. Default value is 30.
n_{cohort} (all designs)	Cohort size	s in each cohortThe number of patient. Default value is 3.
ϵ_1,ϵ_2 (BLRM, CI3+3)	ϵ_1,ϵ_2	Two small fractions used to define the equivalence/target interval of the MTDC. Any doses with a toxicity probability falling into the interval $[p_T - \epsilon_1, p_T + \epsilon_2]$ will be considered an acceptable dose level as MTDC. Default values for both are 0.05.
p_{EWOC} (BLRM)	Cutoff probability of escalation with overdose control	The threshold of controlling the probability of excessive or unacceptable toxicity. Default value is 0.25
$d_{start,1}$ (all designs)	Starting dose level for agent 1	The starting dose level for agent 1 in the simulation trials. Default value is 1.
$d_{start,2}$ (all designs)	Starting dose level for agent 2	The starting dose level for agent 2 in the simulation trials. Default value is 1.

4.2.2.3 Step 3: Generate scenarios

4.2.2.3.1 Step 3.1: Input of Dosages

Select the number of doses for two agents $n_{dose,1}$ and $n_{dose,2}$ ($2 \leq n_{dose,1}, n_{dose,2} \leq 5$) from the dropdown boxes, and their dose levels, $d_{level,1}$ and $d_{level,2}$. Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameters (Figure 4.4). The default dosages of dose levels of agents 1 and 2 are $\{1.00, 2.00, \cdots, n_{dose,1}\}$ and $\{1.00, 2.00, \cdots, n_{dose,2}\}$, respectively. Once applied, the dosages will be standardized against the dosage of the first dose level. Request to allow more than five dose levels for any agent via email support@cytel.com.

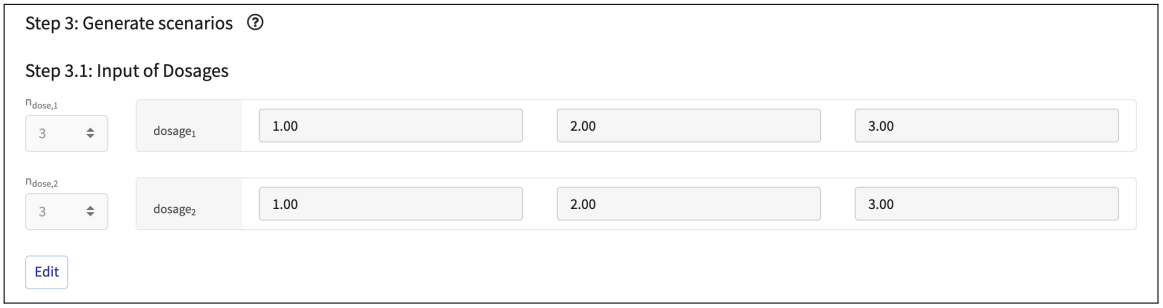


Figure 4.4: Specify input parameters in the **Generate Scenarios** step of the **Dual-Agents Cohort-Based Designs** module.

4.2.2.3.2 Step 3.2: Input of Scenarios

East Bayes provides four ways to generate scenarios. They are described in detail in §4.3.1. Below we provide a quick guidance.

1) automatic construction (**Default Scenarios** tab, see Figure 4.5),
2) logistic regression (**Logistic Regression** tab, see Figure 4.6),
3) specifying marginal toxicity probabilities of each agent and the interaction between two agents (**Mariginals & Interaction** tab, see Figure 4.7),
4) manual construction (**Manual Construction** tab, see Figure 4.8).

1) Default Scenarios (Figure 4.5)

Upon selection of $n_{dose,1}$ and $n_{dose,2}$ and specification of $d_{level,1}$ and $d_{level,2}$, click the "Generate" button to automatically create two default scenarios with diverse dose-toxicity patterns. One is a "Safe" scenario, in which all doses are safe with toxicity probabilities equal to or smaller than the target p_T. The true MTDC locates at the lower right corner of the dose matrix. The other is an

"Ideal" scenario, in which some dose combination are tolerable but some are overly toxic and the true MTDC locates in the middle of the dose matrix. The detailed algorithm for **Default Scenarios** generation is provided in §4.3.1.1.

2) Logistic Regression (Figure 4.6)

Specify the four coefficients of the logistic regression, β_0, β_1, β_2 and β_3, that represent the toxicity probability at the minimum candidate doses of agents 1 and 2 in the logit scale (β_0), the toxicity effect of agent 1 (β_1), the toxicity effect of agent 2 (β_2), and the toxicity effect of the interaction between the two agents (β_3), respectively. Click the "Generate" button to generate the toxicity probabilities for all dose combinations. The detailed algorithm of generating scenarios through **Logistic Regression** is provided in §4.3.1.2.

3) Marginals & Interaction (Figure 4.7)

Specify the marginal true toxicity probabilities of agents 1 and 2 respectively and the interaction effect between the two agents, and click the "Generate" button to generate the toxicity probabilities of all pre-defined dose combinations. The detailed algorithm of generating scenarios through **Marginals & Interaction** is provided in §4.3.1.3.

4) Manual Construction (Figure 4.8)

After clicking the **Manual Construction** tab, an empty dose matrix of two agents ($n_{dose,2} \times n_{dose,1}$) will appear. Users can manually type in the true toxicity probability for each combination. Then click the "Generate" button to generate the scenario.

The generated scenarios will be displayed as a scenario list (Figures 4.5-4.8). Click the "Delete" button to delete the selected scenario.

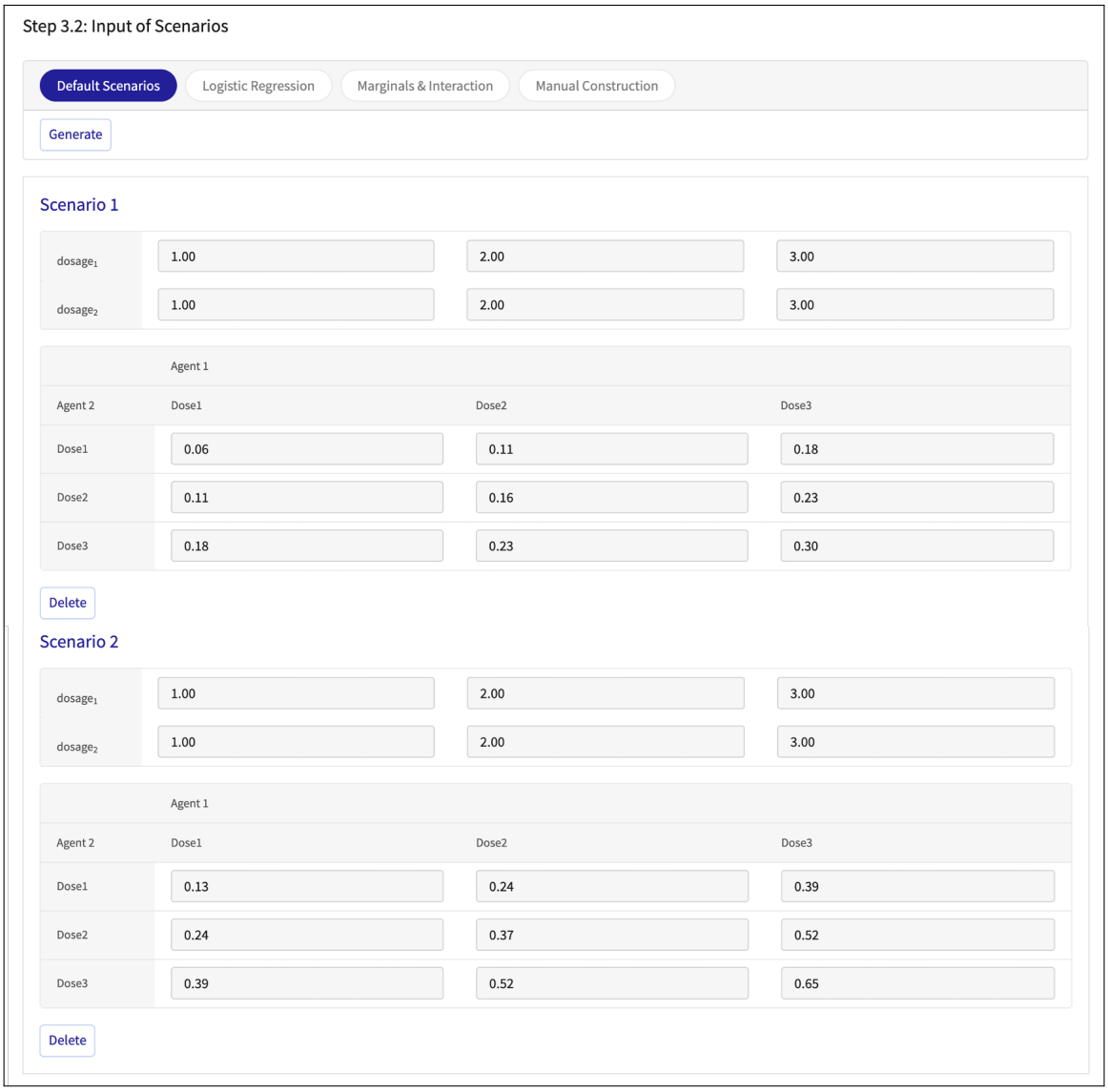


Figure 4.5: Automatically generate scenarios (**Default Scenarios**) in the **Dual-Agents Cohort-Based Designs** module.

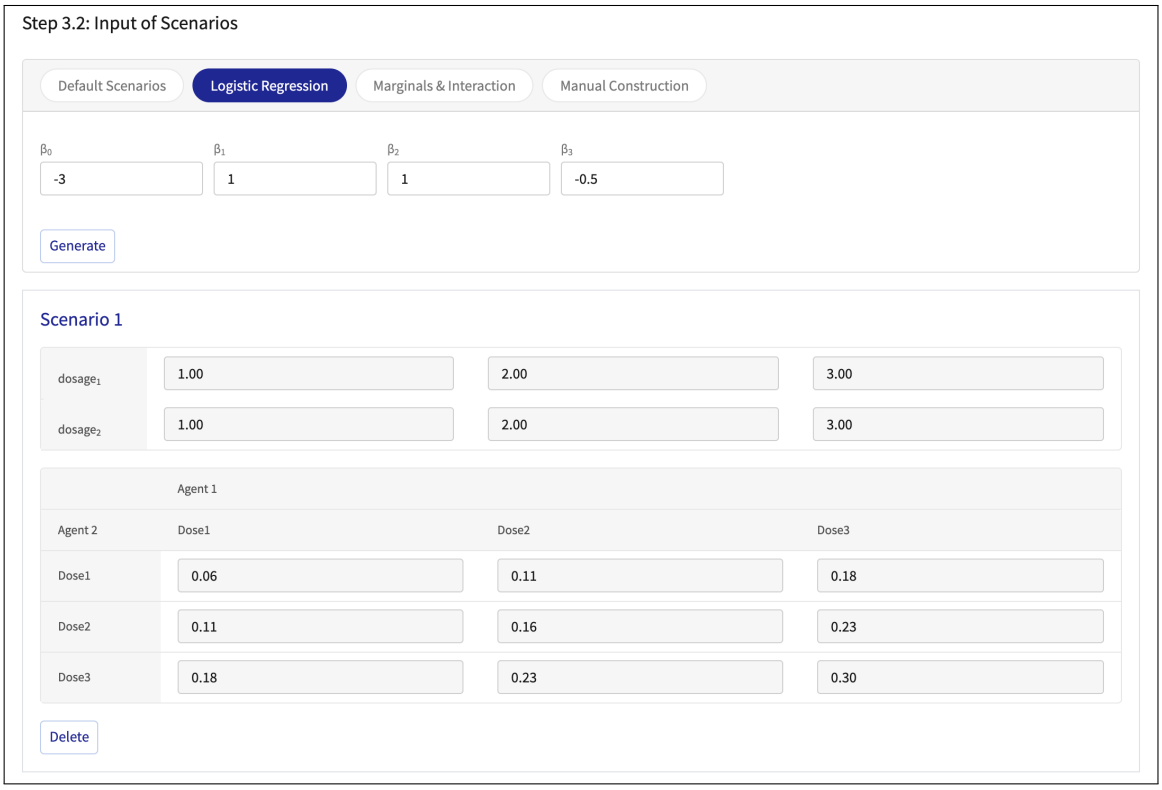


Figure 4.6: Generate scenarios through **Logistic Regression** in the **Dual-Agents Cohort-Based Designs** module.

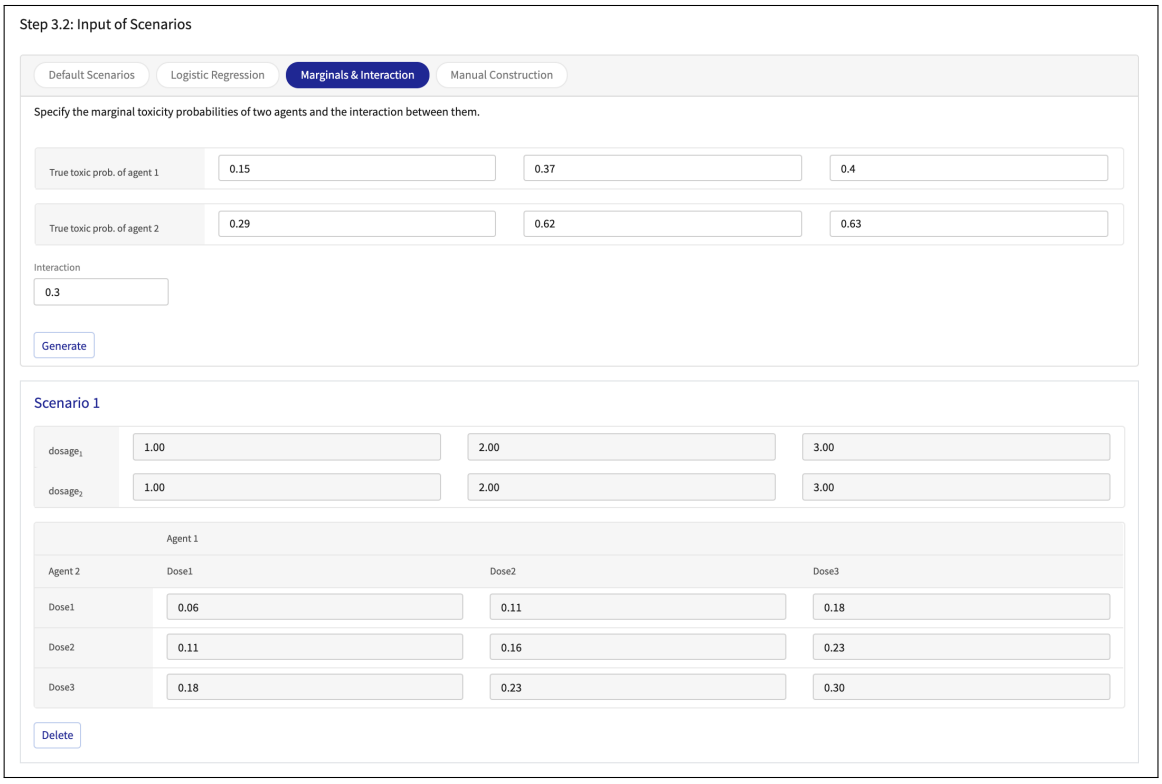


Figure 4.7: Generate scenarios through **Marginals & interactions** in the **Dual-Agents Cohort-Based Designs** module.

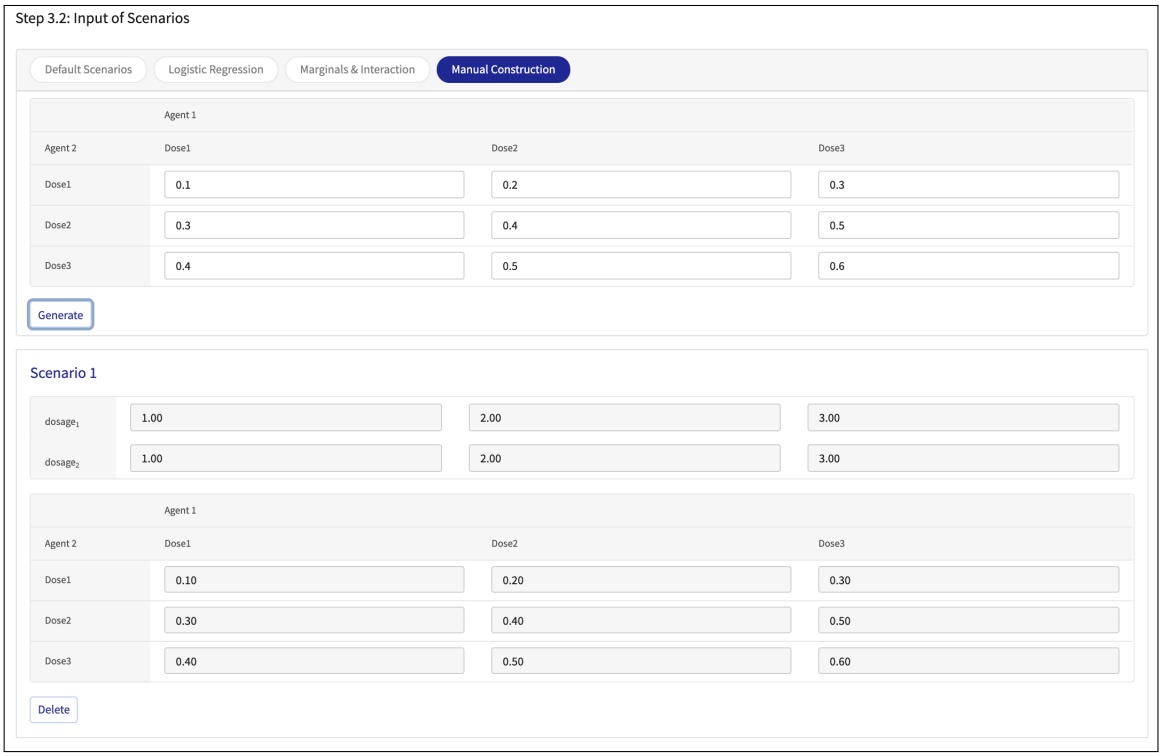


Figure 4.8: Manually generate scenario (**Manual Construction**) in the **Dual-Agents Cohort-Based Designs** module.

4.2.2.4 Launch simulation

Once the steps 1-3 are completed, users can conduct simulated clinical trials to examine the operating characteristics of the selected designs using the selected scenarios, by clicking the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 4.5-4.8). A "**Success**" message will then be displayed on the screen (Figure 4.9) to indicate that the simulation has been successfully launched. Users may click the "OK" button in the pop-up box to track the simulation processing status and simulation results.

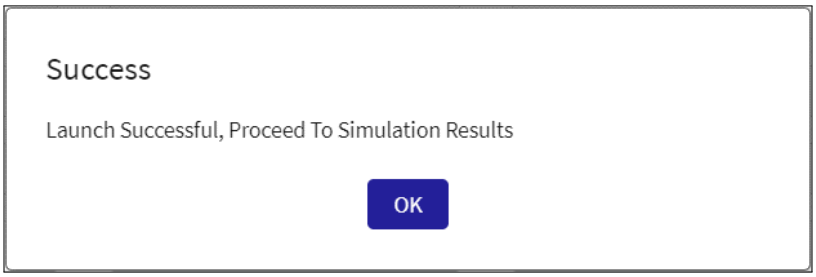


Figure 4.9: "**Success**" message after launching simulation in the **Dual-Agents Cohort-Based Designs** module.

4.2.3 Simulation Results

In the **Simulation Results** tab, users can view the simulation progress and simulation results (§4.2.3.1), restore the simulation settings if needed (§4.2.3.2), and download East Bayes's proprietary report consisting of simulation results in Word format (§4.2.3.3). Hereinafter, we use the terms "simulation results" and "operating characteristics" interchangeably.

4.2.3.1 View simulation results

In the **Simulation Results** tab, the **Running Simulations** panel exhibits the progress of ongoing simulation (Figure 4.10). The ongoing simulations are displayed in ascending order by the launch time. Click the icon "×" to delete the corresponding simulation.

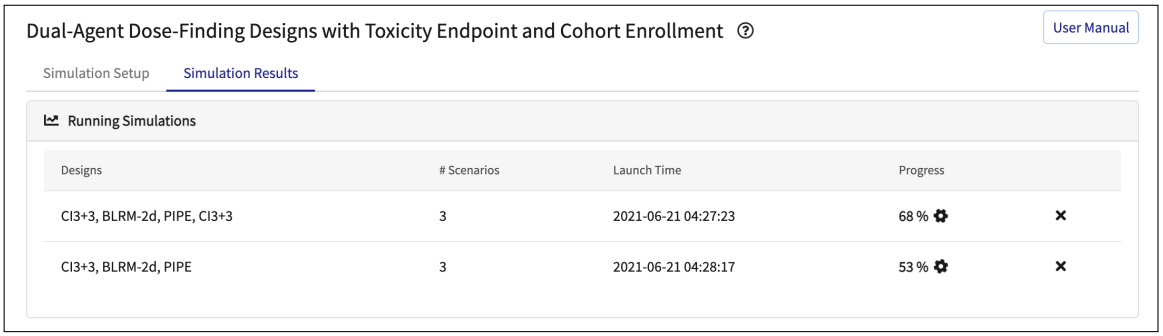


Figure 4.10: Simulation progress in the **Dual-Agents Cohort-Based Designs** module.

Once the simulations are completed, the **Running Simulations** panel in Figure 4.10 will disappear, green "*simulation result created*" massages will appear instead and stay at the same place of the **Running Simulations** panel unless explicitly dismissed by clicking the icon "×" at the end of the corresponding row, and the simulation results will be automatically loaded into the **Simulation History** panel (Figure 4.11), with the blue mail icon to indicate new results. All the previously completed simulations are also listed in the **Simulation History** panel. Simulation results for other modules can also be viewed under the **Simulation History** by dropping down the "Select a module" button (Figure 4.11). Click the 🗑 button to delete the selected simulation results.

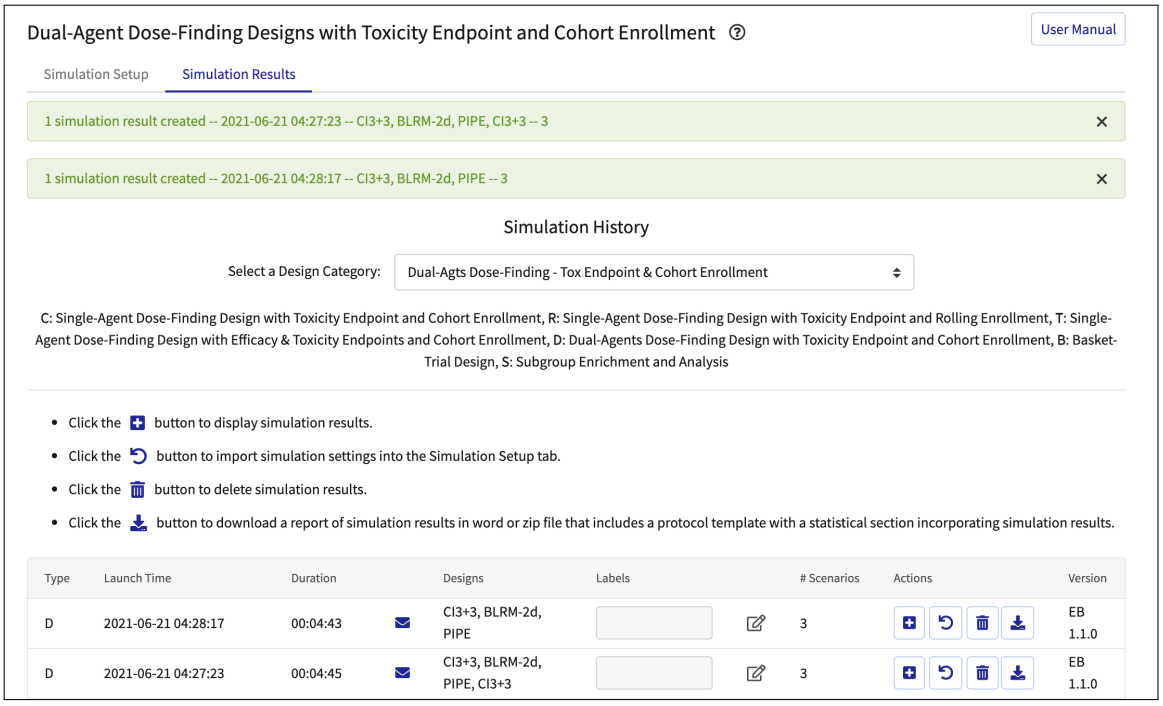


Figure 4.11: Simulation Results in the **Dual-Agents Cohort-Based Designs** module.

Click the ⊞ button to unfold the simulation results (Figure 4.12). The design settings are firstly displayed at the top of each simulation study (Figure 4.12). Then the results of simulation are shown as plots and tables below.

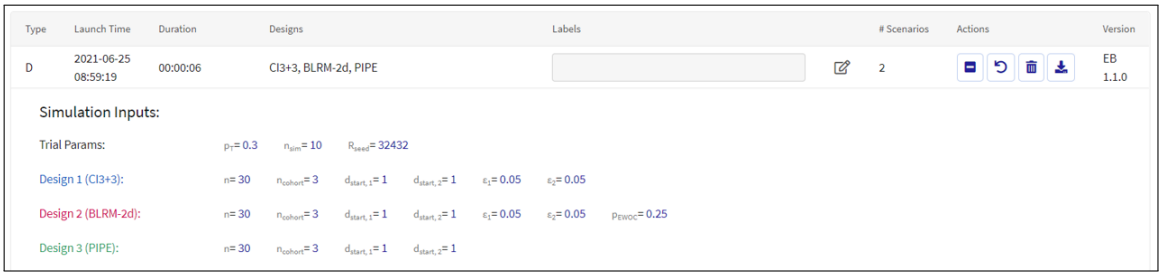


Figure 4.12: View the simulation results in the **Dual-Agents Cohort-Based Designs** module.

Details of the Simulation Results

The simulation results are divided into two parts, i.e, Simulation Result Summary and Tabulated Results by Scenarios. Each part can be viewed or hidden by clicking the button for that part.

Part A: Simulation Result Summary

A. Line plots showing four summary statistics of the simulation results for all the designs (Figure 4.13), including **Prob. of Selecting MTDC**, **Prob. of Toxicity**, **Prob. of Selecting Does-over-MTDC**, and **Prob. of No Selection**.
B. A table of mean and standard deviation (s.d.) for the four summary statistics (Figure 4.14).

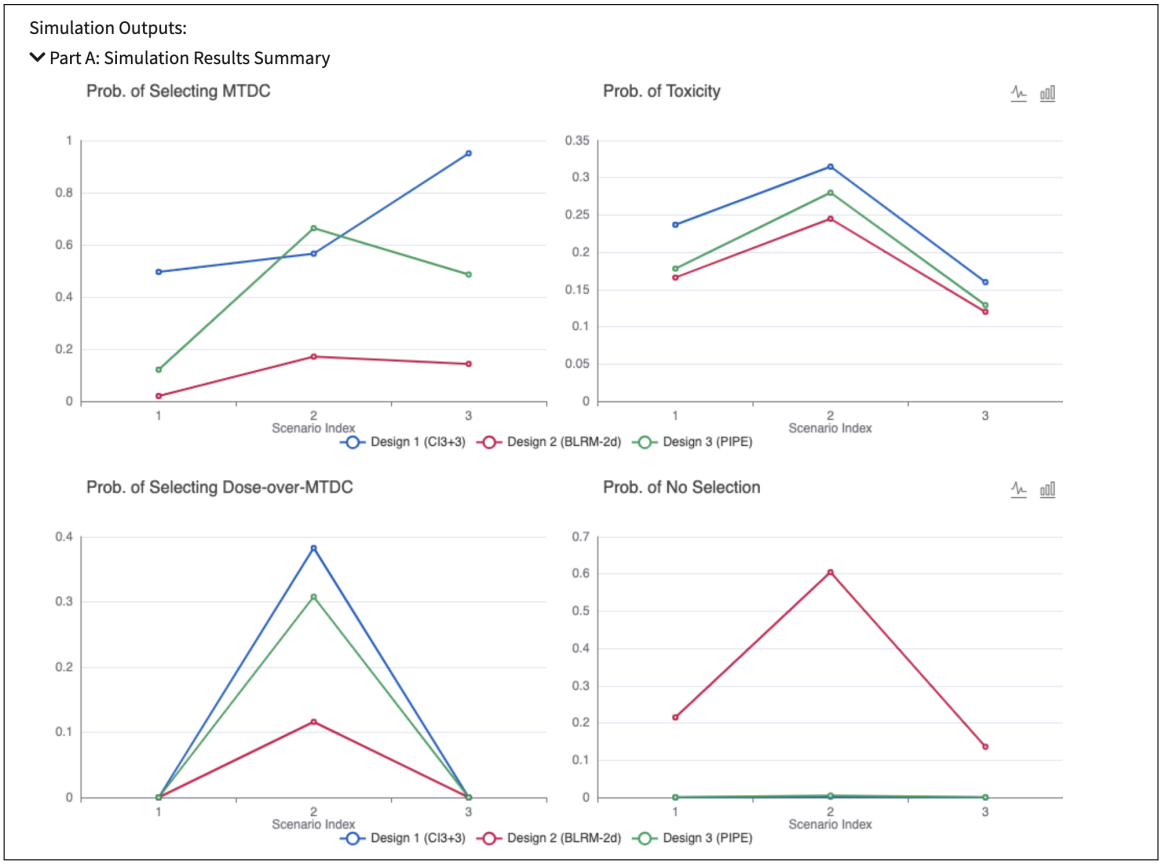
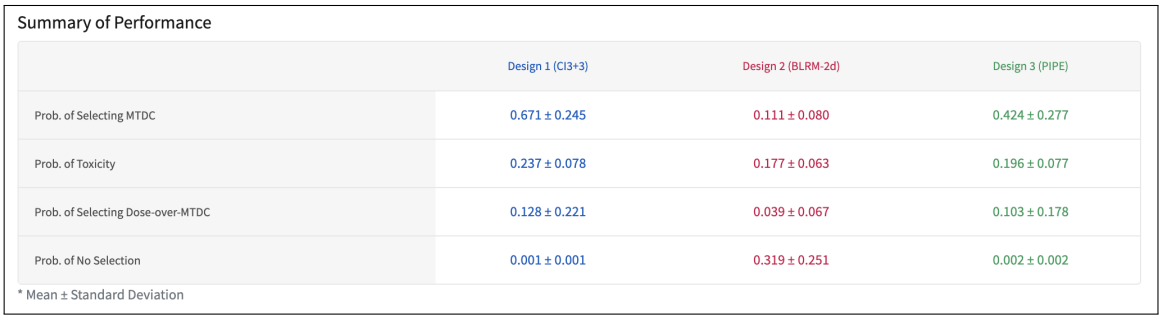


Figure 4.13: Simulation result plots in the **Dual-Agents Cohort-Based Designs** module.



Summary of Performance

	Design 1 (CI3+3)	Design 2 (BLRM-2d)	Design 3 (PIPE)
Prob. of Selecting MTDC	0.671 ± 0.245	0.111 ± 0.080	0.424 ± 0.277
Prob. of Toxicity	0.237 ± 0.078	0.177 ± 0.063	0.196 ± 0.077
Prob. of Selecting Dose-over-MTDC	0.128 ± 0.221	0.039 ± 0.067	0.103 ± 0.178
Prob. of No Selection	0.001 ± 0.001	0.319 ± 0.251	0.002 ± 0.002

* Mean ± Standard Deviation

Figure 4.14: Simulation summary in the **Dual-Agents Cohort-Based Designs** module.

A. Line plots:

- The four summary statistics are part of operating characteristics of the designs. They are explained in full detail next.
  - **Prob. of Selecting MTDC**: The probability of selecting the true MTDC, defined as the proportion of simulated trials that correctly select the true MTDC. The higher the value, the better the design.
    - For CI3+3 & BLRM-2d designs, the true MTDCs are defined as the dose combination levels of which the true toxicity probabilities fall into the equivalence interval $[p_T - \epsilon_1, p_T + \epsilon_2]$; if none of the dose combinations have a toxicity probability that falls into the equivalence interval, the true MTDC is defined as the dose combination with the highest toxicity probability below p_T. For the PIPE design, the true MTDCs are defined as the dose combination levels with the highest toxicity probabilities lower than or equal to p_T.
    - To compare the operating characteristics of multiple designs submitted in a simulation study, the definition of MTDC should be unified. If any of CI3+3 & BLRM-2d designs are used in the simulation, the CI3+3 and BLRM-2d might use different EI's $[p_T - \epsilon_1, p_T + \epsilon_2]$. Then the MTDCs are defined as the dose combination levels of which the true toxicity probabilities fall into the widest equivalence interval $[p_T - \max\{\epsilon_1\}, p_T + \max\{\epsilon_2\}]$. Here, $\max\{\cdot\}$ is taken over the designs. If none of the dose combinations fall in, the dose combination with the highest toxicity probability that is below p_T is the true MTDC. For example, consider a case in which users compare three designs, CI3+3, BLRM-2d and PIPE, in a simulation study targeting $p_T = 0.3$. Suppose $\epsilon_1 = 0.02$ and $\epsilon_2 = 0.05$ for CI3+3, and $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.03$ for BLRM-2d. In this case, the true MTDC is the dose combination levels with toxicity probabilities in $[0.3 - 0.05, 0.3 + 0.05]$; if none of the dose combinations have a toxicity probability in $[0.3 - 0.05, 0.3 + 0.05]$, the dose combination with the highest toxicity probability lower than 0.3 is the true MTDC.
    - For the designs that choose multiple dose combinations as the MTDCs at the end of the trial (PIPE & CI3+3), Prob. of Selecting MTDC is the percentage of simulated trials that correctly select at least one true MTDC.
    - If a scenario does not have any MTDC (e.g., all dose combinations have toxicity probabilities higher than the target p_T), no selection is the right decision. In this

case, the probability of selecting the true MTDC is the probability of no selection.

- **Prob. of Toxicity**: The proportion of patients who have experienced DLT across all the simulated trials. The lower the number, the fewer patients having DLTs under the design.
- **Prob. of Selecting Does-over-MTDC**: The probability of selecting the dose combination levels above the true MTDC, defined as the percentage of simulated trials that select any dose combinations with true toxicity probabilities higher than p_T at the end of the trial. The lower the value, the better the safety of the design.
- **Prob. of No Selection**: The proportion of the simulated trials in which none of the dose combination levels are selected as the MTDC. If a scenario does not have any MTDC, this values is treated as the probability of selecting the true MTDC, i.e., the correct decision.

- For each line plot, the x-axis is the index of scenario and the y-axis is the value of summary statistics. Lines with different colors represent different designs.
- The plots are interactive for better visualization.
  - Hover the mouse on a dot and a box will display the value of each design at the corresponding scenario (top left plot in Figure 4.13: Prob. of Selecting MTDC)
  - Hover the mouse on the design label to highlight the corresponding line and fade the others (bottom right plot in Figure 4.13: Prob. of No Selection).
  - Click the design label to hide the corresponding line and click again to change it back (top right plot in Figure 4.13: Prob. of Toxicity).

B. Simulation summary table: Figure 4.14 shows the mean$\pm$sd of the summary statistics across all scenarios for each design.

Part B: Tabulated Results by Scenarios

Full simulation results are presented mainly in tabular format arranged by scenarios (Figure 4.15), each with five sections (a bubble plot and four tables). The first section is a bubble plot that summarizes the scenario setting with dose levels of two agents and their true toxicity probabilities at each dose combination level. The middle three sections (the first three tables), from top to bottom, report the selection probability, the average number of patients treated, and the average number of toxicities (i.e. DLTs) at each dose combination, respectively. In these four sections, the green, blue and red bubbles (cells) represents doses that are the true MTDC(s), below and above the true MTDC(s), respectively. The last section reports the four trial-specific summary statistics, which are the same as those shown in the **Simulation Result Plots**, mainly from two aspects: MTDC selection and trial toxicity.

The first three tables following the bubble plot (Figure 4.15) present three summary statistics from the simulation.

Selection Prob.: The proportion of simulated trials that select each dose level as the MTDC,
Average # of Patients Treated: The average number of patients treated at each dose level,
Average # of Toxicities: The average number of patients experienced DLT at each dose level.

The last table reports the following summary statistics for the simulation (Figure 4.15).

- **MTDC Selection**
  - **Prob. of Selecting MTDC**: The proportion of simulated trials that select the true MTDC at the end of the trial.
  - **Prob. of Selecting Does-over-MTDC**: The proportion of simulated trials that select the doses higher than the true MTDC at the end of the trial.
  - **Prob. of No Selection**: The proportion of simulated trials in which none of the dose levels are selected as the MTDC.

  For detailed descriptions, please refer to **Simulation Result Plots** section above.
- **Trial Toxicity**
  - **Prob. of Toxicity**: The proportion of patients experiencing DLT across all the simulated trial. For detailed descriptions, please refer to **Simulation Result Plots** section above.

Part B: Tabulated Results by Scenarios

Scenario 1

True Tox Prob.

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	0.060	0.110	0.180
Dose2	0.110	0.160	0.230
Dose3	0.180	0.230	0.300

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

Selection Prob.

Design 1 (CI3+3)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	0.002	0.000	0.016
Dose2	0.006	0.029	0.201
Dose3	0.030	0.220	0.496

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

Design 2 (PIPE)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	0.000	0.038	0.331
Dose2	0.050	0.193	0.357
Dose3	0.331	0.350	0.121

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

# of Patients Treated (s.d.)

Design 1 (CI3+3)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	3.909	0.801	1.350
Dose2	4.221	5.727	9.789
Dose3	2.298	12.132	31.773

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

Design 2 (PIPE)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	3.792	4.068	7.104
Dose2	4.116	7.821	7.815
Dose3	7.374	8.055	3.804

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

# of Toxicities (s.d.)

Design 1 (CI3+3)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	0.235	0.093	0.252
Dose2	0.460	0.862	2.213
Dose3	0.437	2.875	9.611

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

Design 2 (PIPE)

Agent 2 \ Agent 1	Dose1	Dose2	Dose3
Dose1	0.221	0.422	1.265
Dose2	0.436	1.236	1.735
Dose3	1.316	1.840	1.113

Dose below the true MTDC. The true MTDC. Dose above the true MTDC.

Design Performance

	Design 1 (CI3+3)	Design 2 (PIPE)
Prob. of Selecting MTDC	0.496	0.121
Prob. of Toxicity	0.237	0.178
Prob. of Selecting Dose-over-MTDC	0.000	0.000
Prob. of No Selection	0.000	0.001

Figure 4.15: Simulation result tables in the **Dual-Agents Cohort-Based Designs** module.

4.2.3.2 Restore simulation

Users can restore the simulation setting from the simulation results by clicking the button at the upper right corner of each simulation results panel (yellow arrow in Figure 4.16), which will switch the display to the **Simulation Setup** page with the simulation settings restored. This is useful to restore existing simulation settings for reproducible results.

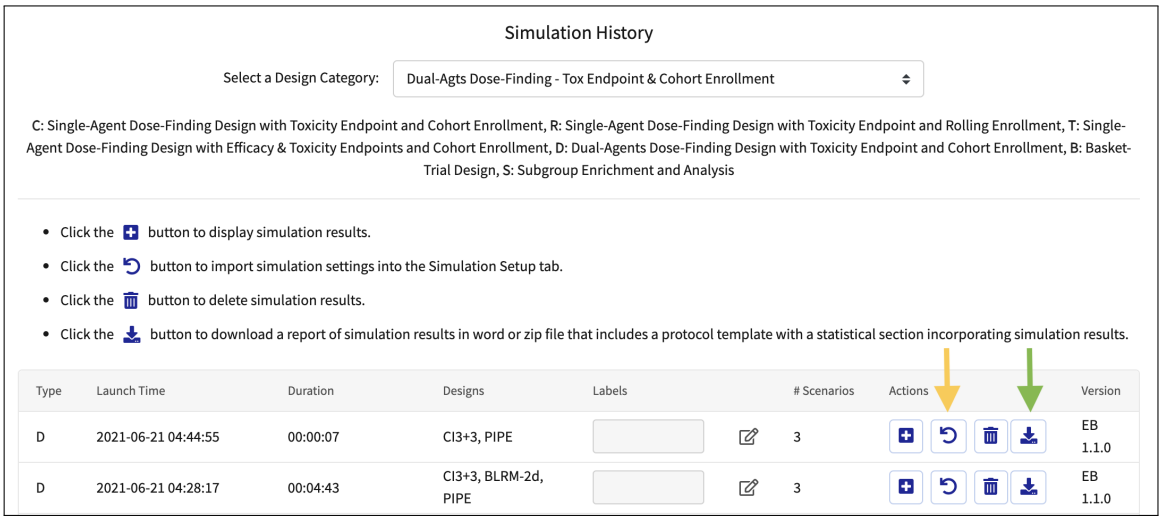


Figure 4.16: Restore simulation setup and download simulation results in the **Dual-Agents Cohort-Based Designs** module.

4.2.3.3 Download simulation results

There is a button at the upper right corner of each simulation results panel (green arrow in Figure 4.16). Click it to download a zip file, which includes a Word file and four line plots of summary statistic shown in Figure 4.13. The Word file is the East Bayes's proprietary report simulation report with complete simulation results under the designs and scenarios users added in the **Simulation Setup** page. Users could update and revise the simulation settings and results tailored for their trials or contact us for consulting services via email `support@cytel.com`.

4.3 Statistical Methods Review

4.3.1 Methods for Scenario Generation

In the **Dual-Agents Cohort-Based Designs** module, East Bayes provides four methods to generate scenarios (different dose-toxicity response patterns) for simulation studies: 1) **Default Scenarios** based on a logistic regression, 2) **Scenarios** through **Logistic Regression**, 3) **Scenarios** through **Marginals & Interactions**, and 4) **Scenarios** through **Manual Construction**. This section describes the detailed methods of the first three methods in details.

Notation

Consider a trial combining $I(I \geq 2)$ dose levels of agent A, denoted by $\{d_{A,1}, d_{A,2}, \ldots, d_{A,I}\}$, and $J(J \geq 2)$ dose levels of agent B, denoted by $\{d_{B,1}, d_{B,2}, \ldots, d_{B,J}\}$, for dose finding. Let $d_{ij} = (d_{A,i}, d_{B,j})$ represent the combination of dose levels i and j for agents A and B respectively and π_{ij} represent its true toxicity probability, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$.

4.3.1.1 Method for Generation of Default Scenarios

In this method, the doses of agents A and B are standardized to be in the interval $[0, 1]$, via $u_i = \frac{d_{A,i} - d_{A,1}}{d_{A,I} - d_{A,1}}$ and $v_j = \frac{d_{B,j} - d_{B,1}}{d_{B,J} - d_{B,1}}$, respectively. Therefore, the lowest dose combination is $(u_1, v_1) = (0, 0)$ and the highest dose combination is $(u_I, v_J) = (1, 1)$. We model the drug combination-toxicity relationship π_{ij} using a four-parameter logistic model:

$$\text{logit}(\pi_{ij}) = \log(\frac{\pi_{ij}}{1 - \pi_{ij}}) = \beta_0 + \beta_1 u_i + \beta_2 v_j + \beta_3 u_i v_j, \tag{4.1}$$

where β_0, β_1, β_2 and β_3 are four unknown parameters that represent the logit of the toxicity probability at the minimum available doses corresponding to $u_1 = v_1 = 0$ (β_0), the toxicity effect of agent A (β_1), the toxicity effect of agent B (β_2), and the toxicity effect of the interaction between two agents (β_3), respectively. Denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ the vector of four unknown parameters in model (4.1).

To specify the unknown values $(\beta_0, \beta_1, \beta_2, \beta_3)$, we follow a procedure as follows. Firstly, we elicit with physicians four "anchor" probabilities π^*_{IJ}, π^*_{1J}, π^*_{I1}, and π^*_{11}, corresponding to the toxicity probabilities of the four dose combinations at $(d_{A,I} = 1, d_{B,J} = 1)$, $(d_{A,I} = 1, d_{B,1} = 0)$,

$(d_{A,1} = 0, d_{B,J} = 1)$, and $(d_{A,1} = 0, d_{B,1} = 0)$. Under (4.1), this means

$$\begin{cases} \beta_0 + \beta_1 + \beta_2 + \beta_3 & = \text{logit}(\pi^*_{IJ}) \\ \beta_0 \qquad\quad + \beta_2 & = \text{logit}(\pi^*_{1J}) \\ \beta_0 + \beta_1 & = \text{logit}(\pi^*_{I1}) \\ \beta_0 & = \text{logit}(\pi^*_{11}) \end{cases}, \tag{4.2}$$

which can be rewritten in matrix format:

$$A\boldsymbol{\beta} = \Pi, \tag{4.3}$$

where

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad \Pi = \begin{pmatrix} \text{logit}(\pi^*_{IJ}) \\ \text{logit}(\pi^*_{1J}) \\ \text{logit}(\pi^*_{I1}) \\ \text{logit}(\pi^*_{11}) \end{pmatrix}.$$

Then the solution of $\boldsymbol{\beta}$ can be easily solved by

$$\hat{\boldsymbol{\beta}} = A^{-1}\Pi, \quad \text{i.e.,} \quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} \text{logit}(\pi^*_{11}) \\ \text{logit}(\pi^*_{I1}) - \text{logit}(\pi^*_{11}) \\ \text{logit}(\pi^*_{1J}) - \text{logit}(\pi^*_{11}) \\ \text{logit}(\pi^*_{IJ}) - \text{logit}(\pi^*_{1J}) - \text{logit}(\pi^*_{I1}) + \text{logit}(\pi^*_{11}) \end{pmatrix}. \tag{4.4}$$

In East Bayes, we assume that the four "anchor" probabilities may take two default choices:

1) $\pi^*_{IJ} = p_T$, $\pi^*_{1J} = \frac{p_T \times J}{I+J-1}$, $\pi^*_{I1} = \frac{p_T \times I}{I+J-1}$ and $\pi^*_{11} = \frac{p_T}{I+J-1}$, in which the MTD is the highest dose combination of the dose matrix; or
2) $\pi^*_{IJ} = p_T + \frac{(1-p_T)(I+J-t-m)}{I+J-t-m+2}$, $\pi^*_{1J} = \frac{\pi_{IJ} \times J}{I+J-1}$, $\pi^*_{I1} = \frac{\pi_{IJ} \times I}{I+J-1}$ and $\pi^*_{11} = \frac{\pi^*_{IJ}}{I+J-1}$, in which the MTD is in the middle of the dose matrix. Here, $t = \frac{I}{2}$, if I is even; otherwise, $t = \frac{I+1}{2}$. Similarly, $m = \frac{J}{2}$, if J is even; otherwise, $m = \frac{J+1}{2}$.

Substitute the estimated $\hat{\boldsymbol{\beta}}$ into equation (4.1) to obtain the probability of toxicity for each dose combinations π^*_{ij}, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. This produces two **Default Scenarios**.

4.3.1.2 Logistic Regression

Using the logistic regression (4.1), users can generate more scenarios by specifying the four parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$. Following §4.3.1.1, one can elicit the "anchor" probabilities to generate scenarios.

4.3.1.3 Marginal & Interactions

In this method, we model the dose-toxicity relationship through marginal toxicity probabilities of each agent when they are used alone and an interaction effect between the two agents.

We start by introducing some additional notation. Let $\pi_{A,i}$ and $\pi_{B,j}$ be two single-agent probabilities of DLT ascribed to i-th level of agent A and j-th level of agent B, respectively, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. In the special case of no interaction (independence), the single-agent toxicities fully determine the toxicity of combinations. For dose combination $(d_{A,i}, d_{B,j})$, the probability of no DLT is $(1 - \pi_{A,i})(1 - \pi_{B,j})$. Under independence, let π^0_{ij} be the probability of no DLT under the combination $(d_{A,i}, d_{B,j})$ when the two drugs are independent; it is true that

$$\pi^0_{ij} = 1 - (1 - \pi_{A,i})(1 - \pi_{B,j}) = \pi_{A,i} + \pi_{B,j} - \pi_{A,i}\pi_{B,j}.$$

On the odds scale this is equivalent to

$$odds^0_{ij} = odds_{A,i} + odds_{B,j} + odds_{A,i} \times odds_{B,j},$$

where $odds^0_{ij} = \pi^0_{ij}/(1 - \pi^0_{ij})$, etc. To allow interaction, one assumes

$$odds_{ij} = odds^0_{ij} \times g(\eta, d_{A,i}, d_{B,j}).$$

In East Bayes, we use the same interaction $g(\cdot)$ for all dose combinations, i.e., $g(\eta, d_{A,i}, d_{B,j}) = \exp(\eta)$. Different values of η represent different relationship between the two agents. Specifically,

- $\eta = 0$: No interaction.
- $\eta < 0$: Protective, i.e., the drug combination produces a toxic effect less than that if the drugs act independently in the body.
- $\eta > 0$: Synergistic, the drug combination produces a toxic effect greater than that if the drugs act independently in the body.

Lastly, we have toxicity probabilities for all dose combinations through $\pi_{ij} = \frac{odds_{ij}}{1+odds_{ij}}$.

4.3.1.4 Manual Construction

We also allow users to manually input scenarios (toxicity probabilities for all dose combinations, π_{ij}). See detailed procedure in §4.2.2.3.

4.3.2 The Product of Independent Beta Probabilities Dose Escalation (PIPE)

The product of independent beta probabilities escalation (PIPE) design is a Bayesian dose finding method for a combination therapy with two active agents, introduced in Mander and Sweeting (2015). The PIPE design aims to target a MTD contour such that the probabilities of toxicity for all dose combinations on this contour equal the prespecified target toxicity level p_T. The dose finding decision process is based on the estimated contour, and multiple dose combinations can be recommended to take forward to phase II.

4.3.2.1 Probability Model

Let $d_{A,i}$ denote the i-th dose level of agent A and $d_{B,j}$ denote the j-th dose level of agent B, $i = 1, 2, \ldots, I(I \geq 2)$ and $j = 1, 2, \ldots, J(J \geq 2)$. Assume $d_{A,i} < d_{A,i+1}$ and $d_{A,j} < d_{A,j+1}$. Let $d_{ij} = (d_{A,i}, d_{B,j})$ represent the combination of dose levels i and j for agents A and B respectively, and π_{ij} represent its true toxicity probability. The toxicity is assumed to be monotonic increasing with increasing dose. That is, $\pi_{ij} \leq \pi_{i+1,j}$, $i = 1, 2, \ldots, I-1$, $\forall j$ and $\pi_{ij} \leq \pi_{i,j+1}$, $j = 1, 2, \ldots, J-1$, $\forall i$.

PIPE assumes π_{ij} follows an independent beta distribution, i.e., $\pi_{ij}|a_{ij}, b_{ij} \sim beta(a_{ij}, b_{ij})$, $\forall i, j$. Here, $(a_{ij} + b_{ij})$ represents a measure of the amount of information contained in the prior, equivalent to the number of patients observed at dose d_{ij} before the trial begins; and $a_{ij}/(a_{ij} + b_{ij})$ and $b_{ij}/(a_{ij} + b_{ij})$ represent the expected prior proportions of DLTs and non-DLTs at dose d_{ij}, respectively. In East Bayes, we use a strong prior $a_{ij} = b_{ij} = 0.5$, $\forall i, j$. The reason we call $beta(0.5, 0.5)$ a strong prior is because we follow the terminology in the PIPE paper (Mander and Sweeting, 2015). Specifically, the authors use the word "strong" to contrast the weak prior in their method which corresponds to $\sum_{ij}(a_{ij} + b_{ij}) = 1$. Request to allow other priors via emailing support@cytel.com.

Patients are recruited into the trial sequentially in cohorts of a pre-specified size with each cohort assigned a dose combination chosen by the design. Suppose after the first m cohorts, $y_{ij}^{(m)}$ patients out of $n_{ij}^{(m)}$ patients have experienced DLT for dose combination d_{ij}; the data up to the end of the m-th cohort are defined by $Data^{(m)} = \left\{y_{ij}^{(m)}, n_{ij}^{(m)}, i = 1, \ldots, I, j = 1, \ldots, J\right\}$. Then because of conjugacy and prior independence of the π_{ij}, the posterior distribution of π_{ij} is also a beta distribution given by

$$\pi_{ij} \mid Data^{(m)}, a_{ij}, b_{ij} \sim beta(a_{ij} + y_{ij}^{(m)}, b_{ij} + n_{ij}^{(m)} - y_{ij}^{(m)}).$$

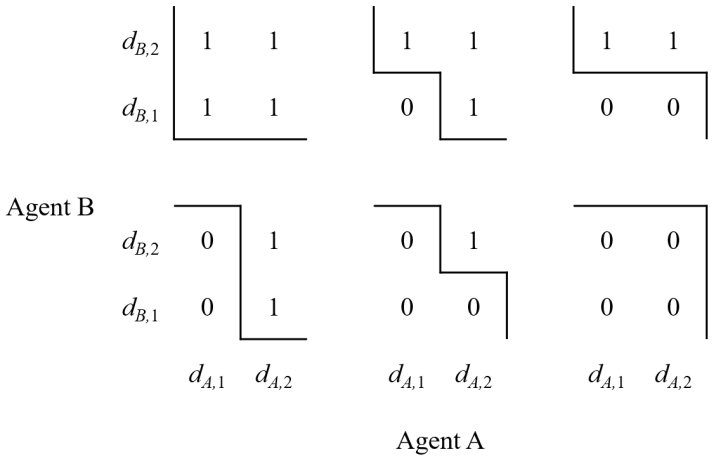


Figure 4.17: Six monotonic MTC examples for two agents (each with two experimental dose levels).

4.3.2.2 Maximum Tolerated Contour (MTC)

The PIPE design aims to locate the MTC$_{p_T}$ corresponding to the pre-specified target probability of toxicity p_T , and uses MTC$_{p_T}$ to recommend the dose level for the next cohorts. The MTC$_{p_T}$ is defined as the boundary in the two-dimensional dose combination space that partition the space into doses with toxicity probabilities above p_T or below p_T. The estimated MTC$_{p_T}$ under PIPE is constrained to follow the monotonicity assumption.

For a discrete set of dose combinations, there are a finite number of locations that a contour can partition the space. And due to monotonicity assumption, only contours that satisfy monotonicity (such contours will be called the "monotonic contours") will be considered. In general, for an $I \times J$ matrix, there are $\binom{I+J}{I}$ monotonic contours in total. For example, consider a situation where each agent has two dose levels of experimentation. There are only six possible monotonic contour choices for the MTC$_{p_T}$, as shown in Figure 4.17. Each contour is represented by a binary matrix indicating whether doses are above the contour (1) or below (0). Define the set of all monotonic contours as $\mathscr{C}$. And let the binary matrices that are members of the set $\mathscr{C}$ be $\mathcal{C}_s$, where $s = 1, \ldots, \binom{I+J}{I}$.

To find the most likely contour for the MTC$_{p_T}$, consider the posterior probability that the toxicity probability is less than or equal to p_T for any dose combination d_{ij}:

$$p_{ij}{}^{(m)} = \text{Prob}\left(\pi_{\text{ij}} \leq \text{p}_\text{T} \mid \text{y}_{\text{ij}}^{(\text{m})}, \text{n}_{\text{ij}}^{(\text{m})}, \text{a}_{\text{ij}}, \text{b}_{\text{ij}}\right).$$

Hence, the probability that the MTC_{p_T} is the contour defined by matrix $\mathcal{C}_s$,

$$\begin{aligned}\alpha_s{}^{(m)} &= \mathbb{P}(\text{MTC} = \mathcal{C}_s \mid Data^{(m)}) \\ &= \prod_{i,j} \left\{1 - p_{ij}{}^{(m)}\right\}^{\mathcal{C}_s[i,j]} \left\{p_{ij}{}^{(m)}\right\}^{1-\mathcal{C}_s[i,j]}, \quad s = 1, 2, \ldots, \binom{I+J}{I},\end{aligned} \tag{4.5}$$

where $\mathcal{C}_s[i,j]$ is the 0-1 indicator for dose combination d_{ij} in the binary matrix as shown in Figure 4.17. The underlying rationale behind the PIPE method is that dose-escalation decisions are based on the most likely $\mathcal{C}_s$ based on $\alpha_s^{(m)}$. In other words, PIPE decides the dose finding based on the contour

$$\mathcal{C}^{*(m)} = \underset{\mathcal{C}_s \in \mathscr{C}}{\operatorname{argmax}}\, \alpha_s^{(m)}. \tag{4.6}$$

4.3.2.3 Dose Finding Rules

PIPE uses $\mathcal{C}^{*(m)}$ as the basis to guide dose finding and to choose from a set of dose combinations that are close to $\mathcal{C}^{*(m)}$. Such set is called the admissible dose set, denoted by $\Omega^{(m)}$. In PIPE, two dose strategies are provided to define $\Omega^{(m)}$: the closest strategy and the adjacent strategy. Let $\Omega^{(m)}_{closest}$ and $\Omega^{(m)}_{adjacent}$ be the two corresponding admissible dose sets, respectively. Here, a dose combination $d_{i'j'}$ is considered closest to $\mathcal{C}^{*(m)}$, if any of the following eight conditions is met,

a1) if $d_{i'j'}$ is above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i',j'] = 1$, and
   i. if $1 < i' \leq I, 1 < j' \leq J$, the dose combinations that are one dose level lower than $d_{i'j'}$ for only agent A or B ($d_{i'-1,j'}$ and $d_{i',j'-1}$) are below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'-1,j'] = \mathcal{C}_s[i',j'-1] = 0$; or
   ii. if $i' = 1, 1 < j' \leq J$, the dose combination that is one dose level lower than $d_{i'j'}$ for agent B ($d_{i',j'-1}$) is below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i',j'-1] = 0$; or
   iii. if $1 < i' \leq I, j' = 1$, the dose combination that is one dose level lower than $d_{i'j'}$ for agent A ($d_{i'-1,j'}$) is below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'-1,j'] = 0$; or
   iv. if $d_{i'j'}$ is the lowest dose combination, i.e., $i' = j' = 1$;

a2) if $d_{i'j'}$ is below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i',j'] = 0$, and
   i. if $1 \leq i' < I, 1 \leq j' < J$, the dose combinations that are one dose level higher than $d_{i'j'}$ for only agent A or B ($d_{i'+1,j'}$ and $d_{i',j'+1}$) are above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'+1,j'] = \mathcal{C}_s[i',j'+1] = 1$.; or
   ii. if $i' = I, 1 \leq j' < J$, the dose combination that is one dose level higher than $d_{i'j'}$ for agent B ($d_{i',j'+1}$) is above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i',j'+1] = 1$; or

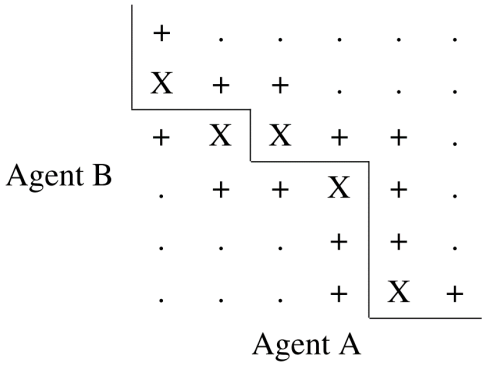


Figure 4.18: The set of admissible doses that are *closest and adjacent* (X) and *adjacent but not closest* (+) to $\mathcal{C}^{*(m)}$.

iii. if $1 \leq i' < I, j' = J$, the dose combination that is one dose level higher than $d_{i'j'}$ for agent A ($d_{i'+1,j'}$) is above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'+1, j'] = 1$; or

iv. if $d_{i'j'}$ is the highest dose combination, i.e., $i' = I$ and $j' = J$.

Similar, a dose combination $d_{i'j'}$ is considered adjacent to $\mathcal{C}^{*(m)}$, if any of the following four conditions is met,

b1) if $d_{i'j'}$ is above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i', j'] = 1$, and

i. if $1 < i' \leq I, 1 < j' \leq J$, among the dose combinations that are a maximum of one dose level lower than $d_{i'j'}$ for both agents A and B, $d_{i-1,j}$, $d_{i,j-1}$ and $d_{i-1,j-1}$, there exits at least one dose combination located below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'-1, j'] = 0$, $\mathcal{C}_s[i', j'-1] = 0$ or $\mathcal{C}_s[i'-1, j'-1] = 0$; or

ii. if the dose level of agent A or B is the lowest, i.e., $i' = 1$ or $j' = 1$;

b2) if $d_{i'j'}$ is below the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i', j'] = 0$, and

i. if $1 \leq i' < I, 1 \leq j' < J$, among the dose combinations that are a maximum of one dose level higher than $d_{i'j'}$ for both agents A and B, $d_{i+1,j}$, $d_{i,j+1}$ and $d_{i+1,j+1}$, there exits at least one dose combination located above the $\mathcal{C}^{*(m)}$, i.e., $\mathcal{C}_s[i'+1, j'] = 1$, $\mathcal{C}_s[i', j'+1] = 1$ or $\mathcal{C}_s[i'+1, j'+1] = 1$; or

ii. if the dose level of agent A or B is the highest, i.e., $i' = I$ or $j' = J$;

Figure 4.18 shows an example for two agents, each with six doses, where the solid line is $\mathcal{C}^{*(m)}$, the sign X's denote the dose combination that are *closest* to $\mathcal{C}^{*(m)}$ and +'s denote the dose combinations that are *adjacent but not closest* to $\mathcal{C}^{*(m)}$. Due to the toxicity monotonicity assumption, all closest doses are adjacent.

In the PIPE paper, Mander and Sweeting (2015) provide two ways to choose one of the admissible dose combinations as the dose for the next cohort,

1) Select the next dose combination to be the admissible dose with the smallest current sample size, where sample size here is defined as both the prior and trial sample size combined, that is, $S_{ij}^{(m)} = n_{ij}^{(m)} + a_{ij} + b_{ij}$. Mathematically, this means to select the dose for the next cohort

$$d_{i^*j^*} = \underset{d_{ij} \in \Omega^{(m)}}{\operatorname{argmin}} \; S_{ij}^{(m)}.$$

   If multiple doses are returned by this function, then the dose combination administered is selected randomly from this set with equal probabilities.

2) Select the next dose combination based on a weighted randomization, where the selection of the admissible doses is weighted by the inverse of their sample size, that is,

$$\mathbb{P}\left(\text{cohort } m+1 \text{ is allocated at } d_{ij} \mid d_{ij} \in \Omega^{(m)}\right) = \frac{S_{ij}^{-1(m)}}{\sum_{d_{ij} \in \Omega^{(m)}} S_{ij}^{-1(m)}}$$

In East Bayes, we take the closest dose strategy 1) to define the admissible dose set, i.e., $\Omega^{(m)} = \Omega_{closest}^{(m)}$, and choose the admissible dose with the the smallest current sample size, i.e., strategy 1) above. Request to apply other dose-escalation rules via email support@cytel.com.

4.3.2.4 Dose Skipping and Safety Rules

In phase I dose-finding trials, dose skipping through the pre-defined levels of agents A and B is often prohibited. Such constraints are accommodated within the PIPE design. In East Bayes, we apply the ***Neighborhood Constraint***, which forces the admissible doses for the next cohort to come from a restricted set of doses that are a maximum of one dose level higher or lower than the current experimented dose both for agents A and B. Besides, East Bayes does not allow diagonal escalation, i.e., escalation from d_{ij} to $d_{i+1,j+1}$ is not allowed. Therefore, the admissible doses can be identified given the adjusted neighborhood constraint, and as an example, are shown in Figure 4.19 for a trial that has its current cohort doses at either (a) d_{11} or (b) d_{33}. In example (a), the dashed box indicates the admissible doses under the current adjusted neighborhood constraint; i.e., doses d_{12} and d_{21}; however, neither is adjacent or closest to the estimated MTD, $\mathcal{C}^{*(m)}$. In this case, PIPE will randomly select one of those two dose combinations to be the next administered dose. In example (b), there are now three dose combinations that are closest, d_{24}, d_{34} and d_{43}, and six adjacent, d_{24}, d_{23}, d_{33}, d_{34}, d_{43} and d_{42}, that could be chosen under the adjacent strategy. Request to apply other constraints, such as the ***Non-neighborhood Constraint*** mentioned in Mander and Sweeting (2015), via emailing support@cytel.com.

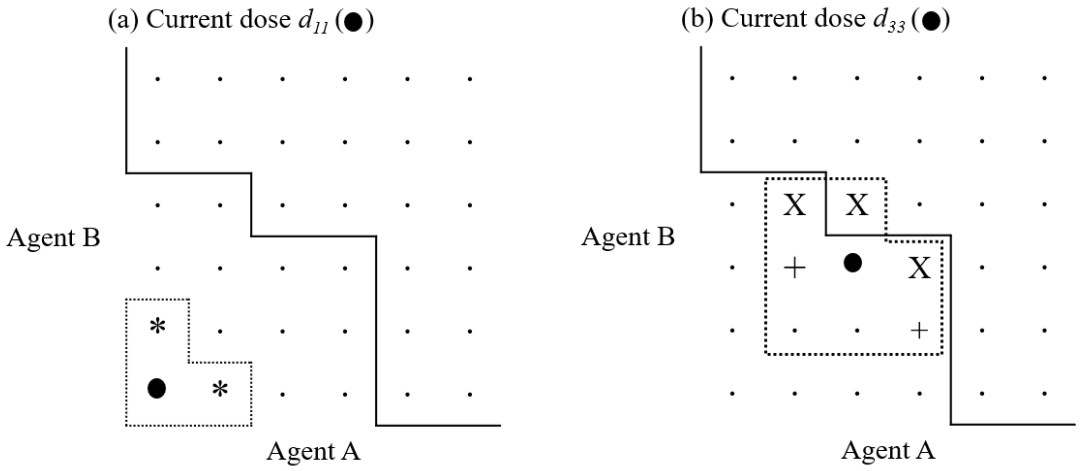


Figure 4.19: The sets of admissible doses that are *closest and adjacent* (X), and *adjacent but not closest* (+) and largest (*) to $\mathcal{C}^{*(m)}$ under a neighborhood constraint without diagonal escalation applied in East Bayes. The dashed line shows the current neighborhood constraint (i.e. only dose combinations within the dashed box are admissible).

Additionally, a ***Safety Constraint*** is imposed to avoid any potential over-dosing. Consider the expected probability of dose combination d_{ij} being above the MTC$_{p_T}$, averaged over the distribution of the monotonic contours. Denote this probability as $q_{ij}^{(m)}$ after m cohorts, which is written as

$$q_{ij}^{(m)} = \sum_{\mathcal{C}_s \in \mathscr{C}} \mathcal{C}_s[i,j]\mathbb{P}(\text{MTC} = \mathcal{C}_s \mid Data^{(m)}).$$

The safety constraint excludes dose combination d_{ij} from the admissible dose set if $q_{ij}^{(m)} > \delta$, where δ is a prespecified constant. Mander and Sweeting (2015) have found that choosing $\delta = 0.8$ gives desired operating characteristics in the simulation studies. East Bayes uses $\delta = 0.8$ by default. The trial is terminated early if there are no available dose combinations that satisfy the safety constraint.

For further safety, two additional safety rules in mTPI-2 and i3+3 are also applied in East Bayes.

- [***Rule 1: Dose Exclusion***] If the current dose combination is considered excessively toxic, i.e., $Prob\{\pi_{ij} > p_T \mid Data^{(m)}\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher dose combinations $\{d_{ml} : i \leq m \leq I, j \leq l \leq J\}$ will be excluded and never be used again in the remainder of the trial.
- [***Rule 2: Early Stop***] If the current dose is lowest dose combination and is considered excessively toxic according to Rule 1 , early stop the trial for safety.

4.3.2.5 The Recommended Phase II Doses

At the end of the trial, multiple doses can be recommended further experimentation at phase II. To do this, after the last cohort M has been enrolled, $C^{*(M)}$ is estimated. Dose combinations that are closest from below to $C^{*(M)}$, have been tried during the trial and do not violate the safety constraint/rules are selected as the recommended phase II doses (RP2Ds).

4.3.3 The Bayesian Logistic Regression Method for Combination of Two Agents (BLRM-2d)

This section describes the Bayesian logistic regression method design for a combination of two active agents (BLRM-2d), proposed by Neuenschwander et al. (2015).

4.3.3.1 Probability Model

Consider a trial combining $I(I \geq 2)$ dose levels of agent A, denoted by $\{d_{A,1}, d_{A,2}, \ldots, d_{A,I}\}$, and $J(J \geq 2)$ dose levels of agent B, denoted by $\{d_{B,1}, d_{B,2}, \ldots, d_{B,J}\}$, for dose finding. Let $d_{ij} = (d_{A,i}, d_{B,j})$ represent the combination of dose levels i and j, and π_{ij} represent the true toxicity probability for dose combination $(d_{A,i}, d_{B,j})$, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Assume $d_{A,i} < d_{A,i+1}$, and $d_{B,j} < d_{B,j+1}$.

The BLRM-2d assumes a logistic model between the marginal toxicity probability of each agent and the dose levels, and the toxicity of probability of the dual agent combination is constructed by the marginal toxicity probability of each agent and the interaction between them, the same as the model in §4.3.1.3. Specifically, the relationship of the marginal toxicity probability of each agent and the dose levels is given by:

$$\text{logit}(\pi_{A,i}) = log(odds_{A,i}) = \log(\alpha_1) + \beta_1 \times \log(d_{A,i}/d_{A,ref}), \quad \alpha_1, \beta_1 > 0, \tag{4.7a}$$

$$\text{logit}(\pi_{B,j}) = log(odds_{B,j}) = \log(\alpha_2) + \beta_2 \times \log(d_{B,j}/d_{B,ref}), \quad \alpha_2, \beta_2 > 0, \tag{4.7b}$$

where α_1, β_1, α_2 and β_2 are the unknown parameters, $\pi_{A,i}$ and $\pi_{B,j}$ are the marginal toxicity probabilities ascribed to i-th level of agent A and j-th level of agent B respectively, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, and $d_{A,ref}$ and $d_{B,ref}$ are the reference doses for agents A and B, respectively. East Bayes uses the (ceiling of $(I + 1)/2$)-th and (ceiling of $(J + 1)/2$)-th level of agents A and B as default reference doses, respectively. This release users from the burden of setting reference doses manually on East Bayes; however, we provide service of customized input of these values upon users requests by emailing us support@cytel.com. In the special case of no interaction, α_1, β_1, α_2, and β_2 fully determine the toxicity probability for a dose combination. For dose combination $(d_{A,i}, d_{B,j})$ the probability of having no DLT is $(1 - \pi_{A,i})(1 - \pi_{B,j})$. Hence, the probability of DLT under no interaction is

$$\pi^0_{ij} = 1 - (1 - \pi_{A,i})(1 - \pi_{B,j}) = \pi_{A,i} + \pi_{B,j} - \pi_{A,i}\pi_{B,j}.$$

On the odds scale, we have

$$\begin{aligned} odds_{ij}^0 = \frac{\pi_{ij}^0}{1-\pi_{ij}^0} &= \frac{\pi_{A,i}+\pi_{B,j}-\pi_{A,i}\pi_{B,j}}{1-(\pi_{A,i}+\pi_{B,j}-\pi_{A,i}\pi_{B,j})} \\ &= \frac{\pi_{A,i}+\pi_{B,j}-\pi_{A,i}\pi_{B,j}}{(1-\pi_{A,i})(1-\pi_{B,j})} \\ &= \frac{\pi_{A,i}}{1-\pi_{A,i}} + \frac{\pi_{B,j}}{1-\pi_{B,j}} + \frac{\pi_{A,i}}{1-\pi_{A,i}} \times \frac{\pi_{B,j}}{1-\pi_{B,j}} \\ &= odds_{A,i} + odds_{B,j} + odds_{A,i} \times odds_{B,j} \end{aligned}$$

Adding an interaction parameter η has the interpretation of an odds-multiplier as follows:

$$odds_{ij} = odds_{ij}^0 \cdot \exp(\eta).$$

Hence, the probability of DLT at dose combination $(d_{A,i}, d_{B,j})$ is given by

$$\pi_{ij} = odds_{ij}/(1+odds_{ij})$$

.

4.3.3.2 Likelihood and Prior Specification

Let n_{ij} and y_{ij} be the number of patients treated at dose combination $(d_{A,i}, d_{B,j})$ and the corresponding number of patients with DLTs, respectively. For observed data, $Data \equiv \{y_{ij}, n_{ij} : i = 1, 2, \ldots, I, j = 1, 2, \ldots, J\}$, the likelihood function is the product of the binomial densities, i.e.,

$$\mathcal{L}(Data \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta) = \prod_i^I \prod_j^J \pi_{ij}^{y_{ij}} \left(1-\pi_{ij}\right)^{n_{ij}-y_{ij}},$$

where $\boldsymbol{\theta_1} = (\alpha_1, \beta_1)$ and $\boldsymbol{\theta_2} = (\alpha_2, \beta_2)$ are vectors of unknown parameters in equations (4.7a) and (4.7b), respectively.

For the prior specification of parameters, α_k and β_k (k=A or B, denoting different agents) follow a multivariate log-normal prior, $\pi(\boldsymbol{\theta_1})$ or $\pi(\boldsymbol{\theta_2})$, given by

$$\begin{pmatrix} \log(\alpha_k) \\ \log(\beta_k) \end{pmatrix} \sim MVN \left\{ \begin{pmatrix} \mu_{k,1} \\ \mu_{k,2} \end{pmatrix}, \quad \Sigma \right\}, \text{where } \Sigma = \begin{pmatrix} \sigma_{k,1}^2 & \rho_k \sigma_{k,1} \sigma_{k,2} \\ \rho_k \sigma_{k,1} \sigma_{k,2} & \sigma_{k,2}^2 \end{pmatrix}, \tag{4.8}$$

where "MVN" stands for a multivariate normal distribution. The interaction parameter η follows a normal distribution as follows $\eta \sim N(\mu_\eta, \sigma_\eta^2)$. In East Bayes, we use the ***quantile-based non-informative prior*** calculator proposed by Neuenschwander et al. (2008) to specify the hyperparameters $(\mu_{k,1}, \mu_{k,2}, \sigma_{k,1}, \sigma_{k,2}, \rho_k)$ in (4.8) for each agent, as described in their Appendix A.1.

The hyperparameter calculation process is based on a set of quantiles for the probabilities of toxicity that are derived from minimally informative unimodal beta distributions. Here, a beta distribution $X \sim beta(a, b)$ is defined as a minimally informative unimodal distribution, given a prespecified quantile $q(p)$ of the prior distribution, if (i) $Prob\{X < q(p)\} = p$, (ii) $a \geq 1$ or $b \geq 1$ (or both), and (iii) a+b minimal. For a given prior quantile $q(p)$, the parameters and the quantiles of a minimally informative unimodal beta distribution can be easily obtained. If $q(p) > p$, $beta(a, 1)$ is minimally informative unimodal if $a = \ln(p)/\ln\{q(p)\}$. Alternatively, if $q(p) < p$, $beta(1, b)$ is minimally informative unimodal if $b = \ln(1-p)/\ln\{1-q(p)\}$. Specifically, the following steps are used for this prior distribution specification for each agent, using agent A as an example:

1. Obtain the set of prior quantiles Q for the distribution of p_d. In East Bayes, we summarize prior information at a given dose using the median, 2.5%-th and 97.5%-th percentiles, denoted by $q_d = \{q_d(2.5\%), q_d(50\%), q_d(97.5\%)\}$.
   (a) For the lowest dose $d = 1$, the prior probability of exceeding a certain threshold $q_1(\phi_1)$ is ϕ_1. In East Bayes, the following default values will be used: $Prob\{p_1 > 0.4\} = 5\%$, i.e. for the lowest dose the probability of excessive toxicity will be set to be 5 percent.
   (b) For the highest dose $d = D$, the prior probability of falling below a certain threshold $q_D(\phi_2)$ is ϕ_2. In East Bayes, the following default values will be used: $Prob\{p_D \leq 0.2\} = 0.05$, i.e. for the highest dose the probability of under-dosing will be set to be 5 percent.
   (c) Assuming a minimally informative unimodal beta distribution in (a) and (b) leads to prior medians for the probabilities of toxicity p_1 and p_D, say $\mu_1 = q_1(50\%)$ and $\mu_D = q_D(50\%)$.
   (d) Prior medians $\mu_1, \ldots, \mu_D$ are assumed to be linear in log-dose on the logit scale. This decides the minimally informative unimodal beta distributions for each dose d.
   (e) For each dose d, two quantiles (2.5% and 97.5%) is derived using minimally informative unimodal beta distributions with prior medians equal to μ_d.
   (f) Therefore, a set of $D \times 3$ quantiles are obtained, denoted by $Q = \{q_{dk}\}$ with $q_{dk} = q_d(\pi_k)$, $d = 1, 2, \ldots, D$, $k = 1, 2, 3$, where $\pi_1 = 2.5\%$, $\pi_2 = 50\%$ and $\pi_3 = 97.5\%$.
2. For the two-parameter logistic model the above constructed quantiles Q are then compared with the quantiles Q' coming from the bivariate normal prior distribution. We will minimize the following criteria:

$$C(Q, Q') = \max_{d,k} |q_{dk} - q'_{dk}|, d = 1, 2, \ldots, D, k = 1, 2, 3.$$

The minimization of $C(Q, Q')$ leads to the optimal parameter for the prior distribution $\boldsymbol{\eta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, which can be achieved by a stochastic optimization using a Metropolis algorithm (Robert and Casella, 2013).

Therefore, the posterior distribution of $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta)$ is given by

$$\begin{aligned} p(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta \mid Data) &\propto \mathcal{L}(Data \mid \boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta)\pi(\boldsymbol{\theta_1})\pi(\boldsymbol{\theta_2})\pi(\eta) \\ &= \prod_{i,j}(\pi_{ij})^{y_{ij}}(1-\pi_{ij})^{n_{ij}-y_{ij}}\pi(\boldsymbol{\theta_1})\pi(\boldsymbol{\theta_2})\pi(\eta), \end{aligned}$$

where $\pi(\boldsymbol{\theta_1})$, $\pi(\boldsymbol{\theta_2})$ and $\pi(\eta)$ are the prior distributions specified above. Using Markov chain Monte Carlo (MCMC) simulation, the posterior samples could be drawn for $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta$ and posterior inference can be made based on the samples.

4.3.3.3 Dose Finding Rules

Suppose the target probability of DLT is p_T, BLRM-2d divides the probability interval $(0, 1)$ into three categories: under-dosing $p_{ij} \in (0, p_T - \epsilon_1]$, target toxicity $p_{ij} \in (p_T - \epsilon_1, p_T + \epsilon_2]$, excessive and unacceptable toxicity $p_{ij} \in (p_T + \epsilon_2, 1)$. After each patient cohort is enrolled and toxicity data are observed, the next dose will be selected depending on ***the Targeted Toxicity Maximization Subject to*** Escalation with Overdose Control (EWOC). That is, select the dose for the next cohort patients as the one that maximizes the posterior probability of falling into the targeted interval, i.e., $\text{argmax}_{i,j}\, Prob\{\pi_{ij} \in (p_T - \epsilon_1, p_T + \epsilon_2] \mid Data\}$ subject to the constraint that the probability of overdosing (i.e., excessive and unacceptable toxicity) does not exceed a predefined threshold p_{EWOC}, i.e., $Prob\{\pi_{ij} \in (p_T + \epsilon_2, 1) \mid Data\} \leq p_{EWOC}$. Here, $Prob\{\cdot\}$ is calculated based on posterior distribution of $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \eta)$.

4.3.3.4 Skipping and Safety Rules

In phase I dose-finding trials, dose skipping and diagonal escalation are often prohibited. To this end, we East Bayes defines the admissible doses for the next cohort as a set of doses that are at most one dose level higher or lower than the current dose for both agents A and B. In addition, East Bayes dose not allow diagonal escalation. See Figure 4.20 for an illustration. In example (a), the current dose combination is d_{11} and the admissible doses are d_{11}, d_{12} and d_{21}; in example (b), the current dose is d_{33} and the admissible doses are d_{34}, d_{43}, d_{33}, d_{24}, d_{42}, d_{23}, d_{32} and d_{22}, a total of eight doses. The trial is terminated early if there are no available doses in the admissible dose set or no doses in the admissible set satisfy the EWOC constraint.

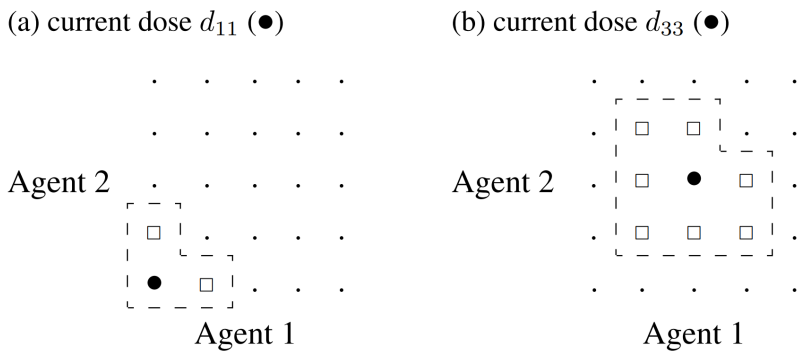


Figure 4.20: The set of admissible doses. Small dots $(\cdot)$ denote the pre-defined dose combinations for the trial, a large dot (●) denotes the current dose, and squares and the large dot (□ and ●) denote the admissible doses for the next cohort patients.

For further safety, two additional safety rules in mTPI-2 and i3+3 are also applied in East Bayes.

- [***Rule 1: Dose Exclusion***] If the current dose combination is considered excessively toxic, i.e., $Prob\{\pi_{ij} > p_T \mid Data^{(m)}\} > \xi$, where the threshold ξ is close to 1, say 0.95, the current and all higher dose combinations $\{d_{ml} : i \leq m \leq I, j \leq l \leq J\}$ are excluded and never used again in the remainder of the trial.
- [***Rule 2: Early Stop***] If the current dose is lowest dose combination and is considered excessively toxic according to Rule 1 , early stop the trial for safety.

In Rules 1 and 2, $Prob\{\pi_{ij} > p_T \mid Data\}$ is a function of the cumulative distribution of $beta(\alpha_0 + y_{ij}, \beta_0 + n_{ij} - y_{ij})$. In East Bayes, $\alpha_0 = \beta_0 = 1$ is used.

4.3.3.5 The MTDC Selection

At the end of the trial, the dose combination $d_{i^*j^*} = (d_{A,i^*}, d_{B,j^*})$ is selected as the MTDC if it maximizes the posterior probability of toxicity rate falling into the targeted interval, i.e., $d_{i^*j^*} = \text{argmax}_{i,j}\, Prob\{\pi_{ij} \in (p_T - \epsilon_1, p_T + \epsilon_2] \mid Data\}$ among all doses that have been used and do not violate the EWOC rule.

4.3.4 The Combo i3+3 Design (CI3+3)

The CI3+3 design is a rule-based design for finding the maximum tolerated dose combination (MTDC) for dual-agent dose-finding trials, proposed by Yuan et al. (2021). It adopts the dose-escalation rules of i3+3 (Liu et al., 2020) and extends them from one dimension to two dimensions.

4.3.4.1 Review of i3+3 Design

We first give a brief review of the i3+3 decision rules (Liu et al., 2020), upon which the CI3+3 design is anchored. The i3+3 design defines an equivalence interval EI $= [p_T - \epsilon_1, p_T + \epsilon_2]$ with the target probability of toxicity p_T and two small fractions, ϵ_1 and ϵ_2, and allocates the next cohort of patients based on the relationship between toxicity probability observed on the current cohort of patients and the equivalence interval. Specifically, suppose dose d is currently used in the trial to treat patients, and y_d patients have experienced dose limiting toxicities (DLTs) out of n_d patients that have been treated. Based on EI, the i3+3 design identifies the appropriate dose for the next cohort of patients according to the following five simple rules.

Current dose: d, No. enrolled: n_d, No. DLTs: y_d		
Condition	Decision	Next dose level
$\frac{y_d}{n_d}$ below EI	Escalation(E)	$d+1$
$\frac{y_d}{n_d}$ inside EI	Stay(S)	d
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ below EI	Stay(S)	d
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ inside EI	De-escalation(D)	$d-1$
$\frac{y_d}{n_d}$ above EI and $\frac{y_d-1}{n_d}$ above EI	De-escalation(D)	$d-1$

Here, a value is below the EI means that the value is smaller than $(p_T - \epsilon_1)$, the lower bound of the EI. A value is inside the EI means that the value is larger than or equal to $(p_T - \epsilon_1)$ but smaller than or equal to $(p_T + \epsilon_2)$. A value is above the EI mean that the value is larger than $(p_T + \epsilon_2)$, the upper bound of the EI.

4.3.4.2 Design Algorithm

For a dual-agent dose-finding trial, suppose I dose levels of agent A, denoted by $\{d_{A,1}, \ldots, d_{A,I}\}$, and J dose levels of agent B, denoted by $\{d_{B,1}, \ldots, d_{B,J}\}$, are to be investigated. Assume $d_{A,i} < d_{A,i+1}$, and $d_{B,j} < d_{B,j+1}$. Let $d_{ij} = (d_{A,i}, d_{B,j})$ denote the combination of i-th dose level for agent A and j-th dose level for agent B, and let π_{ij} denote its true toxicity probability, for

$i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Generally, toxicity is assumed to be monotonic increasing with increasing dose of each agent; That is, $\pi_{ij} \leq \pi_{i+1,j}$, $i = 1, 2, \ldots, I-1$, $\forall j$ and $\pi_{ij} \leq \pi_{i,j+1}$, $j = 1, 2, \ldots, J-1$, $\forall i$. This results in a partial order. Suppose at any moment in the trail, dose combination $d_{ij} = (d_{A,i}, d_{B,j})$ is currently used to treat patients and a total of n_{ij} patients have been assigned to dose combination d_{ij}. Let y_{ij} be the number of patients (among n_{ij}) with DLTs.

The CI3+3 design consists of two stages, with the first stage aiming for rapid escalation through a escalation path (EP), and the second stage for expansive exploration of the dose space. In CI3+3, patients are enrolled in cohorts. To begin the trial, CI3+3 enrolls the first cohort patients at the starting dose combination. For simplicity, suppose the starting dose is the lowest dose combination d_{11}.

Stage I: Run-in Stage

In Stage I, CI3+3 escalates the dose along a prespecified path in order to explore the dose-combination space quickly. Within this path, the doses are fully ordered with monotonic toxicity. Therefore, existing designs for single-agent dose-finding trials can be used. In East Bayes, we use the i3+3 design.

The path can be chosen based on some pre-clinical and clinical information, such as the mechanism of the two agents and the clinical conjecture of MTDC locations. See Figure 4.21 for three possible paths. When we have little information about the path in Stage I, path P_3 in Figure 4.21 might be a good choice. In East Bayes, P_3 is set as the default EP for stage I. In Figure 4.21, P_3 is given by,

$$P_3: \quad \{d_{11} \to d_{21} \to d_{22} \to d_{32} \to d_{33} \to d_{43} \to d_{44} \to d_{54} \to d_{55}\}.$$

If a single path is chosen in stage I, CI3+3 uses the i3+3 design to conduct dose finding along the doses on the path, until

1) a "de-escalation" or a "stay" decision is suggested; or
2) the highest dose along the path is reached.

Stage II: Adaptive Dose-Finding Stage

In Stage II, the full space of dose combinations is explored. Stage II starts at the last dose combination of Stage I and continues to assign the next cohort of patients using an algorithm extending the rules of the i3+3 design.

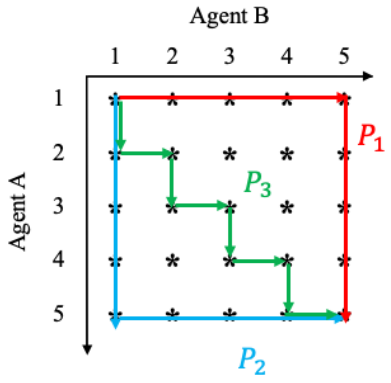


Figure 4.21: Three examples of pathways in the run-in period for CI3+3. In this case, two agents are to be tested, each with five dose levels. The starting dose is the lowest dose combination d_{11}. P_1 represents a pathway in which the dose combination firstly escalates levels of agent B and then levels of agent A when reaching the highest level of agent B; P_2 is the opposite of P_1; Lastly, P_3 alternates the increment of the dose levels of the two agents.

Suppose dose combination d_{ij} is currently used in the trial to treat patients, at which y_{ij} patients have experienced DLT out of n_{ij} enrolled patients. Stage II applies the same up-and-down decisions E, S or D to decide the dose combination for the next cohort of patients.

We now define a distance of two dose combinations. For a dose combination d_{ij}, we call a dose combination d_{kl} a "$M\circ$DC" if the maximum value of differences between i and k, and between j and l, is equal to M, $M = 1, 2, ...$ Mathematically, this means that $M = \max(|i-k|, |j-l|)$. Let $\Omega_{ij}^{(E)}, \Omega_{ij}^{(S)}$, and $\Omega_{ij}^{(D)}$ denote the adjacent candidate sets of dose combinations for the current dose combination $d_{i,j}$ for decision escalation (E), stay (S) and de-escalation (D), respectively. They are defined to be

$$\Omega_{ij}^{(E)} = \left\{ d_{i'j'} \mid 1 \le i' \le I, 1 \le j' \le J, |i'-i| \le 1, |j'-j| \le 1, (i'-i)+(j'-j) = 1 \right\},$$
$$\Omega_{ij}^{(S)} = \left\{ d_{i'j'} \mid 1 \le i' \le I, 1 \le j' \le J, |i'-i| \le 1, |j'-j| \le 1, (i'-i)+(j'-j) = 0 \right\},$$
$$\Omega_{ij}^{(D)} = \left\{ d_{i'j'} \mid 1 \le i' \le I, 1 \le j' \le J, |i'-i| \le 1, |j'-j| \le 1, (i'-i)+(j'-j) = -1 \right\}.$$

The three adjacent candidate sets are the subsets of $1\circ$DCs to d_{ij}. Figure 4.22 gives an example, where the current dose combination is d_{33}, $\Omega_{33}^{(E)} = \{d_{34}, d_{43}\}$, $\Omega_{33}^{(S)} = \{d_{24}, d_{33}, d_{42}\}$, and $\Omega_{33}^{(D)} = \{d_{23}, d_{32}\}$.

We call a dose combination *orderless* to the adjacent candidate set $\Omega_{ij}^{(X)}$ if the order of the toxicity probability between the dose combination and any dose combination in $\Omega_{ij}^{(X)}$ is unknown,

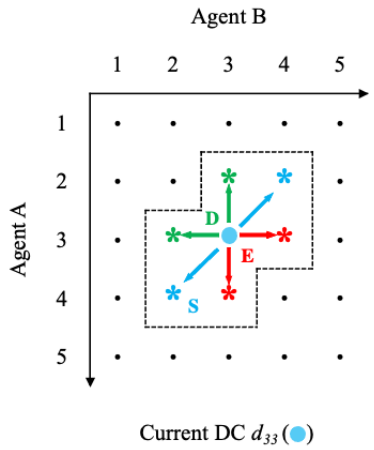


Figure 4.22: An example of the adjacent candidate dose combinations. The dashed box contains the candidate dose combinations, which correspond to three candidate sets of dose combinations for Stage II of CI3+3: red stands for the candidate dose combinations for E, blue for S, and green for D.

$X \in \{E, S, D\}$.

Once the adjacent candidate sets are determined, Stage II of CI3+3 uses a dose-finding algorithm to determine an appropriate dose combination for the next cohort of patients continuously.

Stage II algorithm:

First, determine the up-and-down decisions $\mathcal{A}_{ij} \in \{E, S, D\}$ from the i3+3 design based on the observed data (y_{ij}, n_{ij}) at the current dose combination d_{ij}. The decision $\mathcal{A}_{ij}$ indicates that the next dose combination may be from the adjacent candidate set $\Omega_{ij}^{(\mathcal{A}_{ij})}$. That is, if $\mathcal{A}_{ij}$ equals E, S, or D, the next dose combination will be selected from the adjacent candidate set $\Omega_{ij}^{(E)}$, $\Omega_{ij}^{(S)}$, or $\Omega_{ij}^{(D)}$, respectively.

Second, we consider two special cases to encourage exploration of the dose combination space.

- Let d_{kl} denote a 1∘DC in the adjacent candidate set $\Omega_{ij}^{(\mathcal{A}_{ij})}$ for the current dose combination d_{ij}. A special case is that when

  Condition 1 all the dose combinations d_{kl}'s in the adjacent candidate set $\Omega_{ij}^{(\mathcal{A}_{ij})}$ have already been tested, and

  Condition 2 the corresponding decision is $\mathcal{A}_{kl} = S$ for all $d_{kl} \in \Omega_{ij}^{(\mathcal{A}_{ij})}$.

When conditions 1 & 2 are satisfied, instead of selecting a dose combination from $\Omega_{ij}^{(\mathcal{A}_{ij})}$, we will consider the orderless and untested 1∘DCs to $\Omega_{ij}^{(\mathcal{A}_{ij})}$ (i.e., 1∘DCs to each dose combination in the adjacent candidate set) for future patients. This means assigning patients to potential 2∘DCs.

- Another special case is that when $\mathcal{A}_{ij} = S$ and $n_{ij} >= 12$, i.e., when the current decision is stay and there are more than 12 patients at the current dose combination, we consider assigning pateints to the untested dose combinations in the candidate set $\Omega_{ij}^{(S)}$ first.

Finally, we calculate the posterior probability of belonging to EI of each dose combination in the candidate set $\Omega_{ij}^{(\mathcal{A}_{ij})}$, defined as $\xi_{ij} = Pr\{p_{ij} \in EI \mid y_{ij}, n_{ij}\}$, and select the dose combination d_{ij} with the highest value of ξ_{ij} for the next cohort of patients. The posterior distribution of p_{ij} is $Beta(1 + y_{ij}, 1 + n_{ij} - y_{ij})$ given y_{ij} DLTs out of n_{ij} patients at dose combination d_{ij}.

4.3.4.3 Practical Rules

If dose combination d_{ij} is considered with excessive toxicity, the dose combination and all higher dose combinations with known order $\left\{d_{i'j'} \mid i \leq i' \leq I, j \leq j' \leq J\right\}$ are excluded from the trial and never used again in the remainder of the trial. We deem dose combination d_{ij} overly toxic if

$$Pr\left\{p_{ij} > p_T \mid y_{ij}, n_{ij}\right\} > \xi,$$

where $n_{ij} >= 3$ and the threshold ξ is close to 1, say 0.95. And $Pr\left\{p_{ij} > p_T \mid y_{ij}, n_{ij}\right\}$ is calculated under the beta distribution, $Beta(\alpha_0 + y_{ij}, \beta_0 + n_{ij} - y_{ij})$, with $\alpha_0 = \beta_0 = 1$. If d_{11} is deemed overly toxic, the trial is terminated.

4.3.4.4 MTDC Selection

The trial stops either if d_{11} is overly toxic or when the prespecified maximum sample size N is reached. If d_{11} is overly toxic, no MTDC is selected. Otherwise, we select a MTDC based on the following procedure.

First of all, we assume that the prior for each p_{ij} follows an indepedent $Beta(0.005,\ 0.005)$, and the posterior distribution for each p_{ij} is given by $Beta(0.005 + y_{ij},\ 0.005 + n_{ij} - y_{ij})$. We then estimate p_{ij} by calculating the posterior mean of each dose combination, which is given by $(y_{ij} + 0.005)/(n_{ij} + 0.01)$, and perform a bivariate isotonic regression (Bril et al., 1984) on the

posterior means to meet the monotonic dose-toxicity assumption. Denote the isotonic-transformed posterior means $\hat{p}_{ij}$ for all the dose combinations.

Next, we eliminate dose combinations at which the number of enrolled patients is less than or equal to 3, (i.e, $n_{ij} \leq 3$) and dose combinations that are excessively toxic (i.e, $Pr\{p_{ij} > p_T|y_{ij}, n_{ij}\} > \xi$ or $\hat{p}_{ij} > p_T + \epsilon_2$). These elimination improve the operating characteristics of the designs by weeding out dose combinations with little information or with potential excessive toxicity.

Finally, we select the dose combination for which the $\hat{p}_{ij}$ is the closest to the target rate p_T as the MTDC . When there are ties for $\hat{p}_{ij}$'s with the same index i or j, we select the highest dose combination (largest i or j) among the tied DCs if $\hat{p}_{ij} < p_T$, or the lowest dose combination (smallest i or j) if $\hat{p}_{ij} > p_T$, as the MTDC. If the tied $\hat{p}_{ij}$'s have different i and j, we randomly pick one as the MTDC.

Part II

Phase Ib Expansion Cohort Designs

5. Multiple Cohort Expansion

5.1 Introduction

In modern early-phase clinical trials, often times multiple doses of a new drug are tested in multiple indications to identify the promising doses and arms for phase II or phase III trials. Traditionally, each dose or indication is tested separately in a single trial, resulting in multiple protocols and multiple trials. This module describes a new solution, the multiple cohort expansion (MUCE) design (Lyu et al., 2020). MUCE is a Bayesian solution for cohort expansion trials or the master protocol trials, in which multiple dose(s) and multiple indication(s) are expanded in parallel. It's built on Bayesian hierarchical models with multiplicity control to adaptively borrow information across patient groups from different indications treated with different dose to achieve three major goals:

1. Control the type I error rate (probability of selecting an unpromising drug for further development);
2. Increase the power (probability of selecting a promising drug for further development);
3. Reduce sample size.

As a comprehensive statistical solution, MUCE can be used to calculate the sample size or power, and to conduct interim and final data analyses for making critical decisions. These can be applied in any clinical trials with two or more arms, including:

1. Phase Ib trials with multiple expansion cohorts;
2. Phase II trials with multiple arms;
3. Master protocols including basket, umbrella, and platform trials;

5.2 User Interface and Tutorial

5.2.1 Overview

Entering the **MUltiple Cohort Expansion** page, users will see four main tabs: 1) **Introduction**, 2) **Case Study**, 3) **Quick Demo** and 4) **Data Analysis** (Figure 5.1). In the **Introduction** tab, a general description of MUCE design, its application and benefits is provided 5.1). Then three real-world trials that used MUCE as their trial designs are listed in the the **Case Study** tab, to demonstrate the superiority of MUCE when compared with other designs (§5.2.2). Next, in the **Quick Demo** tab, a demo of the sample size calculation function of MUCE is given, which is based on a simple numerical search algorithm (§5.2.3). Last, in the **Data Analysis** tab (§5.2.4), users could estimate response rates and corresponding posterior probabilities and perform Bayesian hypothesis testing, to conduct interim and final analyses for critical decision-making, such as selecting optimal treatment arm(s).

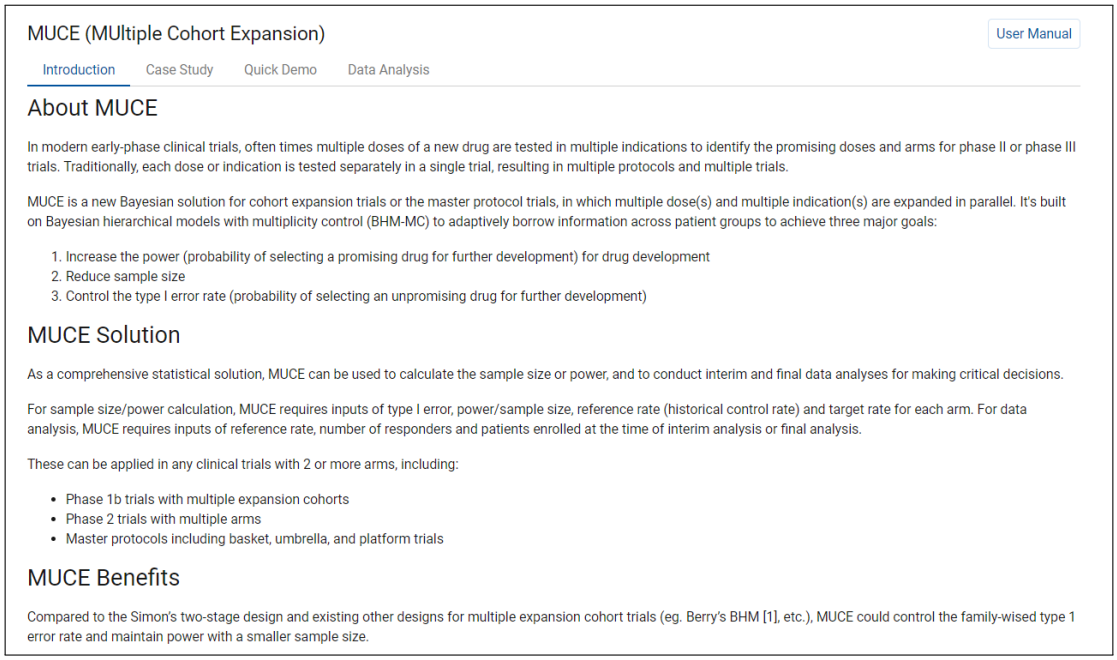


Figure 5.1: Overview of the **Multiple Cohort Expansion** module.

5.2.2 Case Study

The **Case Study** tab lists three real-world cases that apply MUCE (Figure 5.2). In each case study, MUCE is demonstrated to have superior operating characteristics in terms of reducing sample size and controlling the type I error rate (probability of selecting an unpromising drug for further development). Click "Learn More" button in each case box to open and download a PDF file with the detailed descriptions of the case study.

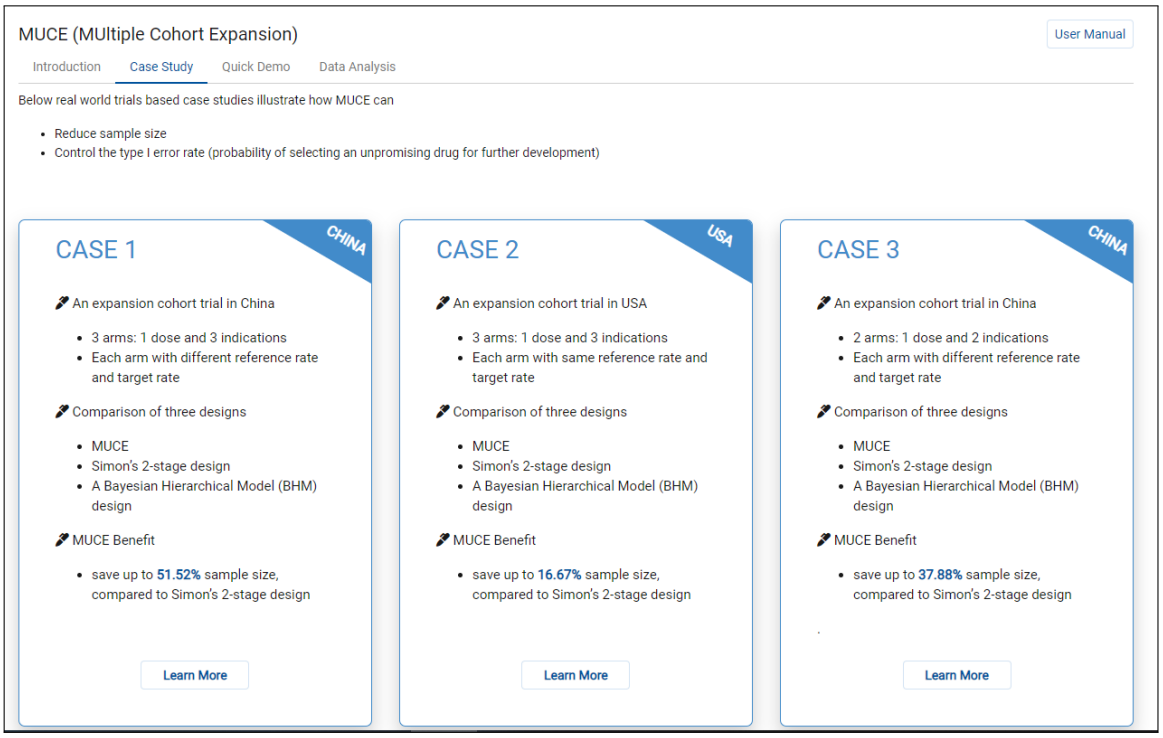


Figure 5.2: Three real world case studies in the **Multiple Cohort Expansion** module.

5.2.3 Quick Demo

This is a demo of the sample size calculation function of the MUCE module on East Bayes. In this demo, all dose-indication arms are assumed to have the same reference response rates and target response rates, therefore all arms should have the same sample size, if the type I error rates and powers are also prespecified the same across all arms. It is a simplified situation and upon these assumption, the sample size of each arm can be easily found through a numerical search algorithm, such as the binary search algorithm. In this quick demo, only limited values are allowed for some input parameters. All limits will be removed in the full version of MUCE module.

5.2.3.0.1 Setup Select the number of doses (n_{dose}) and the number of indications (n_{ind}) from dropdown boxes, resulting in a total of $n_{dose} \times n_{ind}$ dose-indication arms for MUCE designs. Then specify the reference response rate (historical control rate) for each indication (R_{ref}) and the target response rate for each arm (R_{target}). Hover mouse over each trial parameter, and a description will be displayed explaining the meaning of the parameter (Figure 5.3). The detailed explanation of the above four input arguments and their limited values allowed to be selected are provided in Table 5.1. Upon selection, click the "Submit" button to calculate the sample size for each arm using MUCE design to reach the desired type I error rate (α) and power (Figure 5.3).

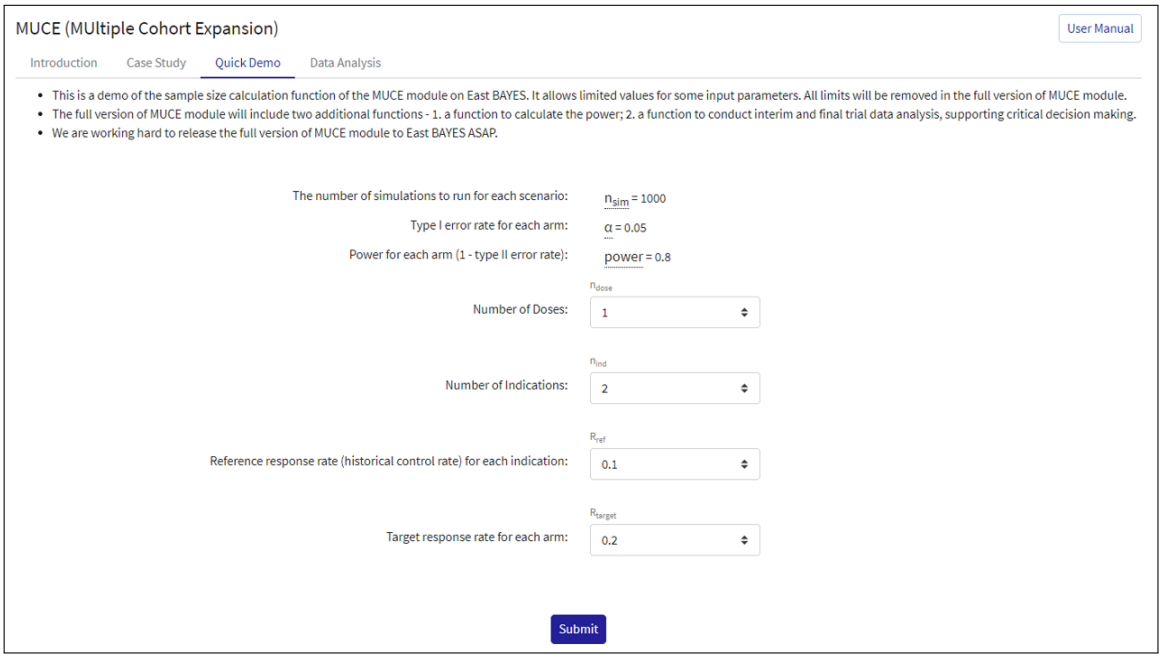


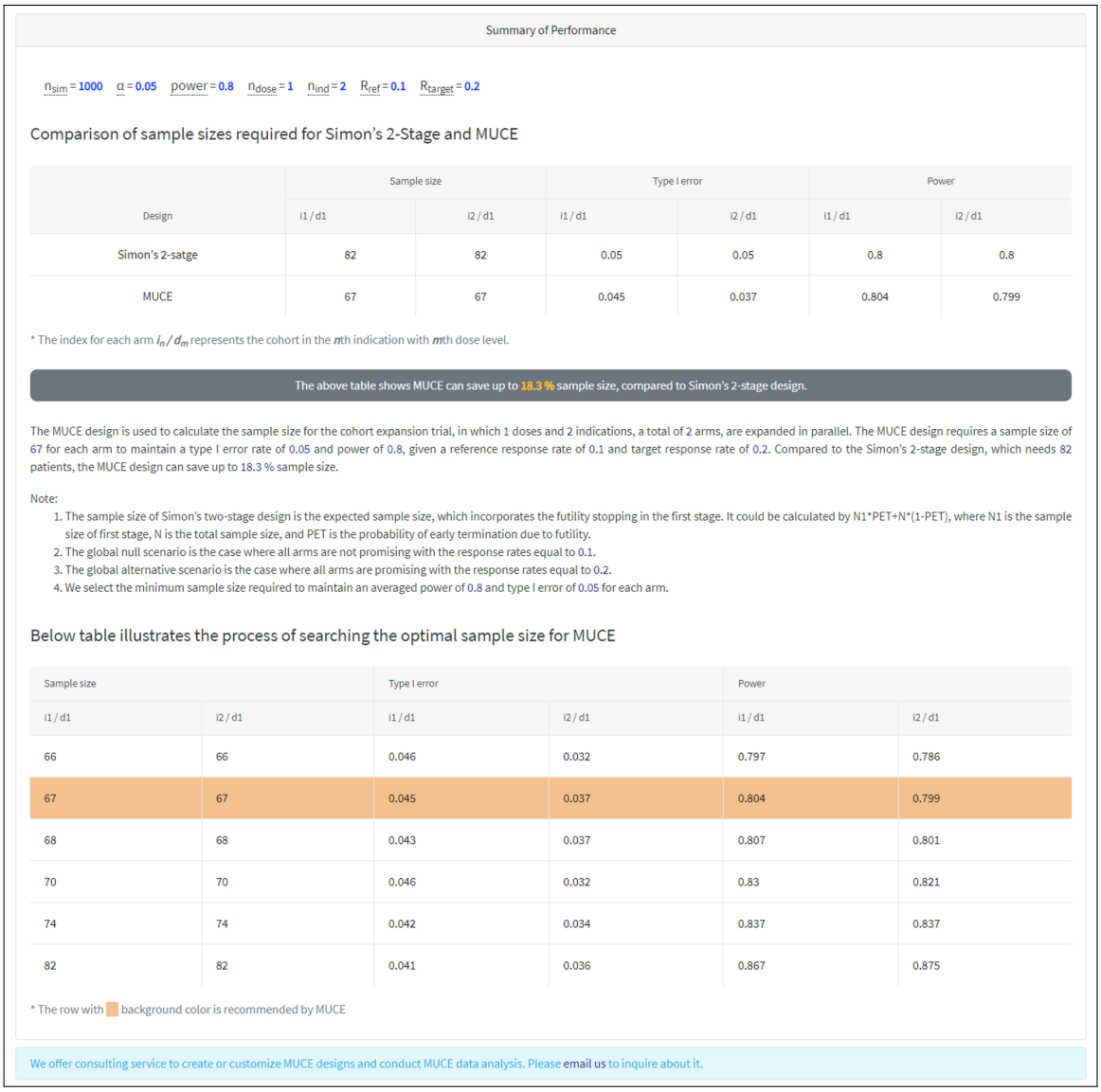
Figure 5.3: Set trial parameters in the **Quick Demo** of the **Multiple Cohort Expansion** module.

Table 5.1: Input trial parameters in the **Quick Demo** of the **Multiple Cohort Expansion** module.

Notation	Parameters	Description
n_{sim}	Number of simulated trials	The number of simulated trials to be conducted for each scenario. In this quick demo, it is fixed at 1,000.
α	Type I error rate	The probability of rejecting null when the null hypothesis is true. In this quick demo, it is fixed at 0.05.
$power$	Power	power= $1-\beta$, where β is the type II error rate, i.e., the probability of rejecting null when the alternative hypothesis is true. In this quick demo, it is fixed at 0.8.
n_{dose}	Number of doses	The number of doses evaluated in the trial. Two values are available for selection In this quick demo, $n_{dose} \in \{1, 2\}$.
n_{ind}	Number of indications	The number of indications expanded in the trial. Two values are available for selection In this quick demo, $n_{ind} \in \{2, 3\}$.
R_{ref}	Reference response rate	The reference response rate (also called the historical control rate) is the largest rate considered to be not promising. Three values are available for selection In this quick demo, $R_{ref} \in \{0.1, 0.2, 0.3\}$.
R_{target}	Target response rate ($R_{target} > R_{ref}$)	The target response rate is the smallest rate considered to be promising. Three values are available for selection in this version,In this quick demo, $R_{target} \in \{0.2, 0.3, 0.4\}$.

5.2.3.0.2 Results The results are displayed in two parts (Figure 5.4):

1. Sample size of MUCE and its comparison with that of Simon's two-stage design.
   - First line lists the values of seven trial parameters in Table 5.1 specified above.
   - A table gives the sample size suggested for MUCE design, to reach the desired type I error and power, using the Simon's two-stage design as benchmark.
   - A description of sample size justification in protocol language.
2. Sample size searching process based on the binary search algorithm.
   - A table lists all the sample size that have been tried in an ascending order, and their corresponding calculated type I error rates and powers.
   - The minimum sample size that reaches the desired type I error rate and power is selected and highlighted in orange background.



Summary of Performance

$n_{sim} = 1000$ $\alpha = 0.05$ power $= 0.8$ $n_{dose} = 1$ $n_{ind} = 2$ $R_{ref} = 0.1$ $R_{target} = 0.2$

Comparison of sample sizes required for Simon's 2-Stage and MUCE

Design	Sample size		Type I error		Power	
	i1 / d1	i2 / d1	i1 / d1	i2 / d1	i1 / d1	i2 / d1
Simon's 2-satge	82	82	0.05	0.05	0.8	0.8
MUCE	67	67	0.045	0.037	0.804	0.799

* The index for each arm i_n / d_m represents the cohort in the nth indication with mth dose level.

The above table shows MUCE can save up to 18.3 % sample size, compared to Simon's 2-stage design.

The MUCE design is used to calculate the sample size for the cohort expansion trial, in which 1 doses and 2 indications, a total of 2 arms, are expanded in parallel. The MUCE design requires a sample size of 67 for each arm to maintain a type I error rate of 0.05 and power of 0.8, given a reference response rate of 0.1 and target response rate of 0.2. Compared to the Simon's 2-stage design, which needs 82 patients, the MUCE design can save up to 18.3 % sample size.

Note:

1. The sample size of Simon's two-stage design is the expected sample size, which incorporates the futility stopping in the first stage. It could be calculated by N1*PET+N*(1-PET), where N1 is the sample size of first stage, N is the total sample size, and PET is the probability of early termination due to futility.
2. The global null scenario is the case where all arms are not promising with the response rates equal to 0.1.
3. The global alternative scenario is the case where all arms are promising with the response rates equal to 0.2.
4. We select the minimum sample size required to maintain an averaged power of 0.8 and type I error of 0.05 for each arm.

Below table illustrates the process of searching the optimal sample size for MUCE

Sample size		Type I error		Power	
i1 / d1	i2 / d1	i1 / d1	i2 / d1	i1 / d1	i2 / d1
66	66	0.046	0.032	0.797	0.786
67	67	0.045	0.037	0.804	0.799
68	68	0.043	0.037	0.807	0.801
70	70	0.046	0.032	0.83	0.821
74	74	0.042	0.034	0.837	0.837
82	82	0.041	0.036	0.867	0.875

* The row with background color is recommended by MUCE

We offer consulting service to create or customize MUCE designs and conduct MUCE data analysis. Please email us to inquire about it.

Figure 5.4: MUCE sample size calculation results in the **Quick Demo** of the **Multiple Cohort Expansion** module.

5.2.4 Data Analysis

The data analysis includes response rates estimation, Bayesian hypothesis tests, and optimal selection of treatment arms, for interim or final analyses, all based on the MUCE design.

5.2.4.0.1 Setup In the Step 1, select numbers of doses and indications ($n_{dose} \in \{1, 2, 3, 4, 5\}$ and $n_{ind} \in \{1, 2, \ldots, 20\}$) from the drop-down boxes, respectively. Upon selection, an input table of the observation data will be automatically generated below the Step 2. And users could manually type in the reference response rate (R_{ref}) for each indication, and the observed numbers of responses and patients for each dose-indication arm. See 5.5 for illustration. Click the "Submit" button to launch the analysis.

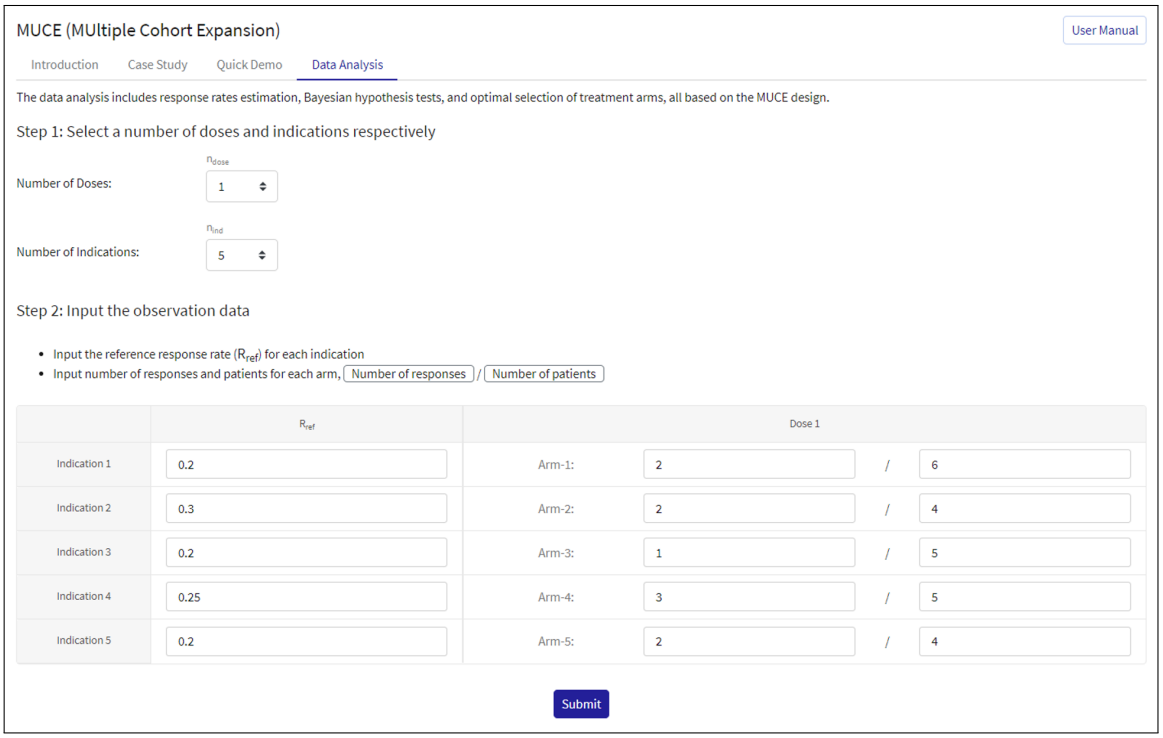


Figure 5.5: Set parameters in the **Data Analysis** of the **Multiple Cohort Expansion** module.

5.2.4.0.2 Results The analysis results are displayed in tables (Figure 5.6).

- The first three columns demonstrate the label-name, and the indexes of dose level and indication, for each arm, respectively.
- The next two columns demonstrate the inputted reference response rate (R_{ref}), the observed numbers of responds and patients (r/n), for each arm, respectively. Also, the response rate of each arm is calculated in ratio.
- The last four columns demonstrate the data analysis results based on the MUCE design, including
  - P_{H_1}: Posterior probability of the alternative hypothesis that the true response rate is larger than the reference response rate. If P_{H_1} is large enough, such as $P_{H_1} > 0.95$, this arm is selected for further investigation (The arm with orange background color); otherwise, it is not selected.
  - P_{mean}: Estimated posterior mean of response rate for each arm.
  - P_{lower} and P_{upper}: The lower and upper boundaries of the interval of the response rate for each arm based on MUCE.

Result table

	Arm	I_{dose}	$I_{indication}$	R_{ref}	r/n (ratio)	P_{H1}	P_{mean}	P_{lower}	P_{upper}
1	Arm-1	1	1	0.2	2 / 6 = 0.33	0.9	0.36	0.08	0.65
2	Arm-2	1	2	0.3	2 / 4 = 0.5	0.92	0.52	0.2	0.9
3	Arm-3	1	3	0.2	1 / 5 = 0.2	0.79	0.29	0.02	0.57
4	Arm-4	1	4	0.25	3 / 5 = 0.6	0.97	0.57	0.25	0.9
5	Arm-5	1	5	0.2	2 / 4 = 0.5	0.96	0.47	0.19	0.86

Note:

- If P_{H1} is large enough, such as $P_{H1} > 0.95$, this arm is selected for future trial considerations (The arm with background color); otherwise, it is not selected.
- Hover the mouse over the table header to see the description of each column.
  - Arm : Name of each arm
  - I_{dose} : The index of dose level
  - $I_{indication}$: The index of indication
  - R_{ref} : The reference response rate for each indication
  - r/n (ratio) : Number of responses / Number of patients (Response rate)
  - P_{H1} : posterior probability of the alternative hypothesis that the true response rate is larger than the reference rate
  - P_{mean} : Estimated posterior mean of response rate for each arm
  - P_{lower} : The lower bound of the credible interval of the response rate for each arm based on MUCE
  - P_{upper} : The upper bound of the credible interval of the response rate for each arm based on MUCE

Figure 5.6: Results in the **Data Analysis** of the **Multiple Cohort Expansion** module.

5.3 Statistical Methods Review

5.3.1 Multiple Cohort Expansion (MUCE) Method

The multiple cohort expansion (MUCE) (Lyu et al., 2020) approach was proposed as a design or analysis method for phase 1b multiple expansion cohort trials, which investigate one or more doses of a new investigational drug in patients from with different indications (cancer types and/or biomarker status). The MUCE design is based on a class of Bayesian hierarchical models that adaptively borrow information across different dose-indication arms. Statistical inference is directly based on the posterior probability of each arm being efficacious, facilitating the decision making that decides which arm to select for further testing.

5.3.1.1 Probability Model

Consider a phase Ib trial that evaluates J different dose levels of a new drug in I different indications. Let (i, j) denote the cohort arm for indication i and dose level j, $i = 1, \ldots, I$, $j = 1, \ldots, J$. The total number of arms is $K = I \times J$. Suppose n_{ij} patients have been treated in arm (i, j), and y_{ij} of them are responders. Let p_{ij} denote the true and unknown response rate for the arm (i, j). We assume y_{ij} follows a binomial distribution conditional on n_{ij} and p_{ij}, $y_{ij} \mid n_{ij}, p_{ij} \sim$ Binomial(n_{ij}, p_{ij}). Whether dose level j is effective for indication i can be examined by the following hypothesis test:

$$H_{0,ij} : p_{ij} \leq \pi_{i0} \quad \text{versus} \quad H_{1,ij} : p_{ij} > \pi_{i0}, \tag{5.1}$$

where π_{i0} is the reference response rate for indication i.

We perform the hypothesis test (15.6) under a formal Bayesian testing framework. Let λ_{ij} be a binary and random indicator of the hypothesis, such that $\lambda_{ij} = 0$ (or 1) represents that hypothesis $H_{0,ij}$ (or $H_{1,ij}$) is true. Firstly, a prior model for p_{ij} is built under each hypothesis. Let $\theta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ denote the log-odds of the response rate. The null hypothesis $p_{ij} \leq \pi_{i0}$ is equivalent to $\theta_{ij} \leq \theta_{i0}$, and the alternative hypothesis is equivalent to $\theta_{ij} > \theta_{i0}$, where $\theta_{i0} = \log\left(\frac{\pi_{i0}}{1-\pi_{i0}}\right)$. Conditional on λ_{ij}, MUCE assume

$$\theta_{ij} \mid \lambda_{ij} = 0 \sim \text{Trunc-Cauchy}(\theta_{i0}, \gamma; (-\infty, \theta_{i0}]),$$
$$\theta_{ij} \mid \lambda_{ij} = 1 \sim \text{Trunc-Cauchy}(\theta_{i0}, \gamma; (\theta_{i0}, \infty)),$$

where Trunc-Cauchy$(\theta, \gamma; A)$ denotes a Cauchy distribution with location θ and scale γ truncated to interval A.

Secondly, prior models for the probabilities of the hypotheses, $\Pr(\lambda_{ij} = 1)$ and $\Pr(\lambda_{ij} = 0)$, are constructed. To borrow strength across dose levels and indications, we construct a hierarchical prior model for λ_{ij}. A natural and conventional Bayesian approach is to impose a common prior for the probability of $\{\lambda_{ij} = 1\}$, which shrinks the probabilities to a common value. To better exploit the data structure in multiple expansion cohort trials, we propose to differentiate the borrowing strength from two factors: dose and indication. To better exploit the data structure in multiple expansion cohort trials, we propose to differentiate the borrowing strength from two factors: dose and indication. For example, two arms with the same indication or dose might exhibit more similar treatment effects than two arms with different indications and doses. To achieve this, we use a probit model as the prior model for λ_{ij}. Let Z_{ij} be a latent random variable, and $\lambda_{ij} = I(Z_{ij} < 0)$, where $I(\cdot)$ is an indicator function. We model

$$Z_{ij} \sim N(\xi_i + \eta_j, \sigma_0^2).$$

Here, $\mathrm{E}(Z_{ij}) = \xi_i + \eta_j$, in which ξ_i characterizes the effect of indication i and η_j of dose j. The indication-specific effects and dose-specific effects are then separately modeled by common priors,

$$\xi_i \mid \xi_0, \sigma_\xi \overset{iid}{N}(\xi_0, \sigma_\xi^2), \quad \text{and} \quad \eta_j \mid \eta_0, \sigma_\eta \overset{iid}{N}(\eta_0, \sigma_\eta^2).$$

Lastly, we put hyperpriors on ξ_0 and η_0, $\xi_0 \sim N(\mu_{\xi_0}, \sigma_{\xi_0}^2)$ and $\eta_0 \sim N(\mu_{\eta_0}, \sigma_{\eta_0}^2)$.

In brief, the entire hierarchical models are summarized in the following equations:

$$\begin{aligned}
&\text{Likelihood:} && y_{ij} \mid n_{ij}, p_{ij} \sim \text{Binomial}(n_{ij}, p_{ij});\\
&\text{Transformation:} && \theta_{ij} = \text{logit}(p_{ij}), \theta_{i0} = \text{logit}(\pi_{i0});\\
&\text{Prior for } (\theta_{ij} \mid \lambda_{ij}): && \theta_{ij} \mid \lambda_{ij} = 0 \sim \text{Trunc-Cauchy}(\theta_{i0}, \gamma; (-\infty, \theta_{i0}]),\\
&&& \theta_{ij} \mid \lambda_{ij} = 1 \sim \text{Trunc-Cauchy}(\theta_{i0}, \gamma; (\theta_{i0}, \infty));\\
&\text{Prior for } \lambda_{ij}: && \lambda_{ij} = \begin{cases} 0, & \text{if } Z_{ij} < 0,\\ 1, & \text{if } Z_{ij} \geq 0; \end{cases}\\
&\text{Latent probit regression:} && Z_{ij} \mid \xi_i, \eta_j, \sigma_0^2 \sim \text{N}(\xi_i + \eta_j, \sigma_0^2);\\
&\text{Indication-specific effects:} && \xi_i \mid \xi_0, \sigma_\xi^2 \sim \text{N}(\xi_0, \sigma_\xi^2);\\
&\text{Dose-specific effects:} && \eta_j \mid \eta_0, \sigma_\eta^2 \sim \text{N}(\eta_0, \sigma_\eta^2);\\
&\text{Hyperpriors:} && \xi_0 \mid \mu_{\xi_0}, \sigma_{\xi_0}^2 \sim \text{N}(\mu_{\xi_0}, \sigma_{\xi_0}^2),\\
&&& \eta_0 \mid \mu_{\eta_0}, \sigma_{\eta_0}^2 \sim \text{N}(\mu_{\eta_0}, \sigma_{\eta_0}^2).
\end{aligned} \tag{5.2}$$

In East Bayes, the values of the hyperparameters $\gamma = 2.5$, $\mu_{\xi_0} = 0$, $\mu_{\eta_0} = 0$, $\sigma_0^2 = 1$, $\sigma_\xi^2 = 1$, $\sigma_\eta^2 = 1$, $\sigma_{\xi_0}^2 = 1$ and $\sigma_{\eta_0}^2 = 1$ are used by default.

5.3.1.2 Trial Design

Suppose $L(\geq 0)$ interim looks are planned, and the l-th interim analysis is conducted after n_{ij}^l patients have been enrolled in arm k. Let $\mathcal{D}^l \equiv \{(n_{ij}^l, y_{ij}^l) : i = 1, 2, \ldots, I; j = 1, 2, \ldots, J\}$ denote the observed data at interim analysis l, where y_{ij}^l is the number of responders among the n_{ij}^l patients. Denote $\mathcal{D}^{L+1} \equiv \{(n_{ij}^{L+1}, y_{ij}^{L+1}) : i = 1, 2, \ldots, I; j = 1, 2, \ldots, J\}$ the observed data at the end of the trial, where n_{ij}^{L+1} is the prespecified maximum sample size for arm (i, j) and y_{ij}^{L+1} is the total number of responders. The proposed MUCE design with L interim looks is describe as follows:

1. Enroll n_{ij}^1 patients in (i, j)-th arm, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$.
2. Given the data $\mathcal{D}^l$ at the l-th interim look, $l = 1, 2, \ldots, L$,
   (a) *[Futility stopping]* If the posterior probability that the hypothesis of arm (i, j), $H_{1,ij}$, is true (i.e., $\lambda_{ij} = 1$) is small, i.e.,

   $$P_{H_1} = Pr\{\lambda_{ij} = 1 \mid \mathcal{D}^l\} < P_{\textit{futility}},$$

   stop the accrual to the k-th arm for futility;
   (b) *[Efficacy stopping]* If the posterior probability that the hypothesis of arm (i, j), $H_{1,ij}$, is true (i.e., $\lambda_{ij} = 1$) is large, i.e.,

   $$P_{H_1} = Pr\{\lambda_{ij} = 1 \mid \mathcal{D}^l\} < P_{\textit{efficacy}},$$

   stop the accrual to the k-th arm for efficacy;
   (c) Otherwise, continue to enroll patients until reaching the next interim analysis.
3. Once the maximum sample size is reached or all the arms have stopped, evaluate the efficacy for each arm based on all the observed data. If the posterior probability that that the hypothesis of arm k, $H_{1,ij}$, is true (i.e., $\lambda_{ij} = 1$) is large, i.e.,

   $$P_{H_1} = Pr\{\lambda_{ij} = 1 \mid \mathcal{D}^{L+1}\} > \phi_{ij},$$

   arm (i, j) is declared efficacious and promising; otherwise, it is considered not promising.

Part III

Phase II Designs

6. Subgroup Enrichment and Subgroup Analysis (SCUBA)

6.1 Introduction

Patient heterogeneity is common across many diseases, for an example in the cancer therapeutic area see Catenacci (2015). Inter-tumor heterogeneity refers to differences in the basic biology, development, and response to a treatment across different tumors. Different from one-size-fits-all approaches such as the traditional chemo or radiation therapies, precision medicine treats subgroups of patients with targeted therapies based on the heterogeneity in their molecular profiles and baseline characteristics. Some molecularly targeted medications have been successfully developed for subgroups of patients. For example, trastuzumab induces better response in breast cancer patients with over-expressed HER2 than those who do not over-express this growth factor. The pairing of specific treatment (trastuzumab) to subgroup of patients (HER2+ breast cancer) with an identifiable biomarker is a simple example of precision medicine applied to produce better patient outcomes. Increasing efforts to identify more biomarker-drug pairs is an active area of on-going research (Mullard, 2015).

Only a few dozen subgroup treatment pairs (STPs) like (HER2+ breast cancer, Trastuzumab) have been discovered and marketed for cancer care. Many more effective STPs are yet to be identified, partly due to lack of statistical methods for subgroup discovery and analysis. Early work in Simon and Maitournam (2004) and Maitournam and Simon (2005) include theoretical discussions of the efficiency and sample size of targeted trials compared to randomized clinical trials (RCT). Sargent et al. (2005) presented a biomarker-by-treatment interaction design and biomarker-based-strategy design, where the former was an extension of RCT with biomarkers as stratification, and the latter used the biomarker as the identifier of whether to use a particular treatment. Freidlin et al. (2010) compared different biomarker-RCT schemes. All of these methods assume that a fixed num-

ber of prespecified subgroups is available, and test if treatments would exhibit varying therapeutic effects on different subgroups. There is no notion of learning new subgroups as all subgroups are predetermined. This could be problematic if predefined subgroups are not predictive of outcomes or treatment selection. An example is the BATTLE trial (Kim et al., 2011). BATTLE is a pioneering study to test treatment and biomarker interactions using a fixed-subgroup design. The design prespecified five subgroups based on eleven selected biomarkers, and randomized patients within each subgroup to different treatments using response-adaptive randomization. Kim et al. (2011) concluded that the biomarker groups used in BATTLE were less predictive than were individual biomarkers, making them clinically less appealing.

In light of the lessons learned from previous studies, the field has shifted to methods that allow new subgroups to be discovered during and after the clinical trial. Sivaganesan et al. (2011) cast the subgroup identification problem as a model selection problem among different partition models. Ruberg et al. (2010), Foster et al. (2011), and Lipkovich et al. (2011) conducted subgroup analysis by looking for regions in covariate space that have significantly different response rates compared to the average response rate. Zhao et al. (2013) presented a scoring function of multiple baseline covariates to estimate subject-specific treatment differences, based on a working response-covariate model. Berger et al. (2014) proposed a Bayesian model selection approach based on random trees for subgroup identification, in which a continuous response variable and binary covariates are considered. Shen and He (2015) proposed a confirmatory statistical test for the existence of subgroups by using a structured logistic-normal mixture model. Green and Kern (2012) used Bayesian additive regression trees (BART) (Chipman et al., 2010) to identify treatment effect heterogeneity among different subgroups. Lastly, Xu et al. (2016) proposed a subgroup enrichment design, SUBA, aiming to allocate patients to subgroup-specific treatments. Their approach uses a tree-type of random clustering model that splits the biomarker space using the median of observed values for each biomarker.

Here, we describe a module in East Bayes, **Subgroup Enrichment and Analysis**, which performs trial simulation to examine the operating characteristics of the subgroup cluster-based Bayesian adaptive (SCUBA) design (Guo et al., 2017a). The SCUBA design is applicable to phase II randomized and controlled trials with one or more treatment arm and a common control arm. Baseline continuous biomarkers are measured for each patient, based on which subgroups will be estimated. The main problem SCUBA solves is to identify proper subgroups defined as patients whose biomarker values fall into specific ranges. SCUBA can be used as an enrichment design that allows patients to be enrolled in predicted optimal treatment arm or a data analysis method that estimates subgroups of patients at the end of the trial. SCUBA, as a design, consists of two

stages, the run-in stage, and the adaptive stage. During the run-in stage, patients are randomized with fixed ratios between the treatment arms and control; and during the adaptive stage, patients are either assigned to the predicted best treatment arm or adaptively randomized based on the predictive probability of response for each treatment.

§6.2 introduces the user interface and a tutorial on launching trial simulations and examining results. Statistical details of the SCUBA design are provided in §6.3.

6.2 User Interface and Tutorial

6.2.1 Overview

Entering the **Subgroup Enrichment and Analysis — SCUBA** page, users will see two main tabs: **Simulation Setup** and **Simulation Results**. The first tab allows users to conduct simulations and the second to visualize/download simulation results. In the **Simulation Setup** tab, there are three steps (Figure 6.1): 1) **Set trial parameters**, 2) **Select designs**, and 3) **Generate scenarios**. Upon completing steps 1-3, click the "Launch Simulation" button at the bottom of the page to begin the simulation using the current parameters, or click the "Reset" to clear all settings and enter new parameters. After the simulation completes, the results will be displayed in the **Simulation Results** tab. Step-by-step instructions are shown in §6.2.2-§6.2.3. Depending on the number of trials to be simulated, the simulation may take minutes to hours to complete.

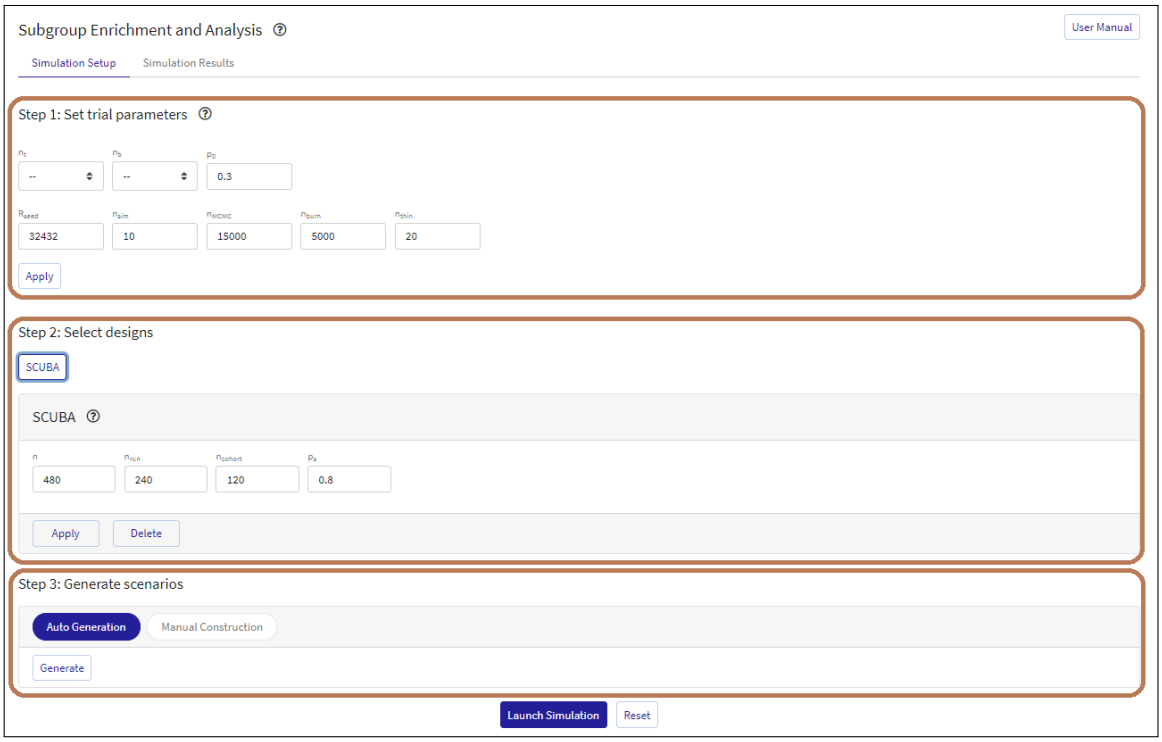


Figure 6.1: Simulation Setup in the **Subgroup Enrichment and Analysis—SCUBA** module.

6.2.2 Simulation Setup

In the **Subgroup Enrichment and Analysis** module, Cytel currently offers only the SCUBA design type. When hovering over the question mark icons a description of parameters used in a section is displayed. If there are parameters you would like to change which are not currently accessible, or designs you would like to see added to this module please contact us by emailing support@cytel.com. More designs and methods for subgroup enrichment and analysis will be added in the future.

6.2.2.1 Step 1: Set trial parameters

Recall that SCUBA is applicable to randomized and controlled phase II trials with potentially more than one treatment arms and a single control arm. First select the number of treatments (n_t) and number of biomarkers (n_b). Then specify the response rate of the control arm (p_0), simulation seed value (R_{seed}), number of MCMC iterations (n_{MCMC}), etc. See Figure 6.2. A detailed explanation of these input arguments is provided in Table 6.1.

Click the "Apply" button in Figure 6.2 to confirm the trial parameters. The "Apply" button changes to "Edit" and can be clicked to change trial parameters as needed.

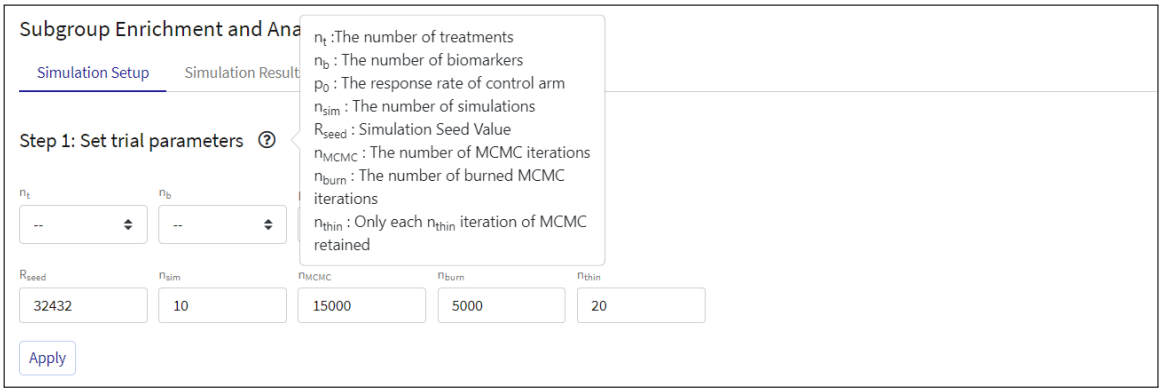


Figure 6.2: Set trial parameters in the **Subgroup Enrichment and Analysis** module.

Table 6.1: The input parameters for a trial in the **Subgroup Enrichment and Analysis** module.

Notation	Parameters	Description
n_t	The number of treatments	The number of treatments in the trial. The range is $[1, 3]$.
n_b	The number of biomarkers	The number of treatments in the trial. The range is $[1, 2]$.
p_0	The response rate of control arm	The assumed response rate of the control arm. In SCUBA, a control arm is assumed to be present by default. Responses of the patients allocated to the control will be sampled from a binomial distribution with probability p_0. The default value is 0.3.
n_{sim}	The number of simulations	The maximum number of simulations allowed is 1000. The default value is 10.
R_{seed}	Simulation seed value	A number used to initialize a pseudorandom number generator in the simulation. The default value is 32432.
n_{MCMC}	The number of MCMC iterations	The maximum number of MCMC iterations allowed is 15,000. The default value is 15,000.
n_{burn}	The number of burned MCMC iterations	The number of initial MCMC iterations n_{burn} $(\leq n_{MCMC})$ which are discarded. The default value is 5000.
n_{thin}	The thinning number of MCMC	After the burn-in, only each n_{thin} iteration of MCMC iterations is retained. The default value is 20.

6.2.2.2 Step 2: Select designs

Click the "SCUBA" design button to select it. Enter the desired design parameters in their respecctive entry fields. For a detailed parameter description list see in Table 6.2.

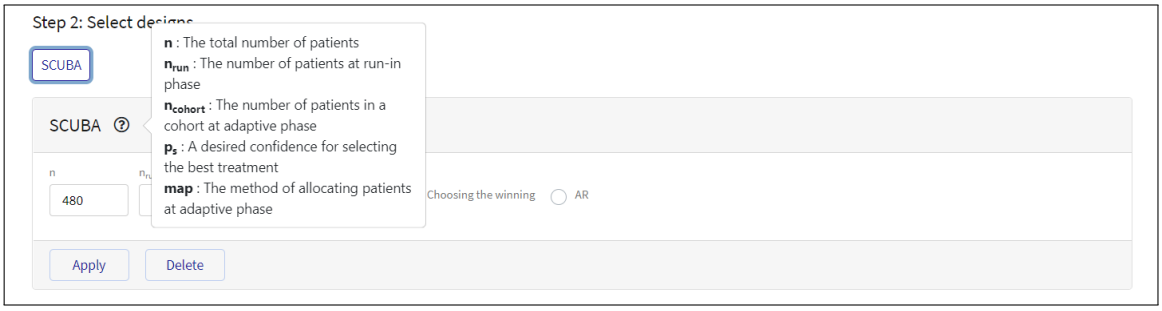


Figure 6.3: Select designs in the **Subgroup Enrichment and Analysis** module.

Table 6.2: Input parameters for designs in the **Subgroup Enrichment and Analysis** module.

Notation	Parameters	Description
n	The total number of patients	The total number of patients to be treated in the trial. The range is $[100, 1000]$ and the default value is 480.
n_{run}	The number of patients in the run-in phase	The patients in the run-in phase are randomized equally to n_t treatments and the control arm. Range is $[n/2, n-1]$ and default value is 240.
n_{cohort}	The number of patients in a cohort in the adaptive phase	The patients in the adaptive phase are assigned in cohorts. The range is $[10, n - n_{run}]$ and the default value is 120. In East Bayes, we set a limit of the number of interim analysis to 4, so $(n - n_{run})/n_{cohort}$ should be less than or equal to 4.
p_s	A desired confidence for selecting the best treatment	A threshold for selecting the best treatment based posterior probability. The range is (0,1) and the default value is 0.8.
map	The method of allocating patients in the adaptive phase	There are two methods in East Bayes, "Choosing the Winner " and "Adaptive Randomization". See details of patient allocation in §6.3.1.3.

6.2.2.3 Step 3: Generate scenarios

There are two ways to generate scenarios, automatically (see Figure 6.4) or through manual construction (see Figure 6.5). In East Bayes, we assume the true response rate is associated with biomarker values based on a probit regression (6.1). Let θ_j be the true response rate for patient j under one treatment. We assume:

$$\theta_j = \Phi_{0,1}(a_{11}x_{j1} + a_{12}x_{j2} + a_{21}x_{j1}^2 + a_{22}x_{j2}^2 + b_{12}x_{j1}x_{j2}), \tag{6.1}$$

where x_{j1} and x_{j2} denote the values of two biomarkers for patient j, and $\Phi_{0,1}(x)$ is the cumulative distribution function (CDF) of a standard Gaussian distribution evaluated at point x. For each treatment within each scenario, there are coefficients which need to be specified. For a single biomarker there will be 2 coefficients (a_{11}, a_{21}), while for 2 biomarkers there will be 5 coefficients $(a_{11}, a_{12}, a_{21}, a_{22}, b_{12})$.

Auto Generation (Figure 6.4)
Upon clicking the "Generate" button, two scenarios will be created automatically, each of which contains the true coefficients of biomarkers or the interaction between biomarkers. These generated scenarios are displayed under different tabs.

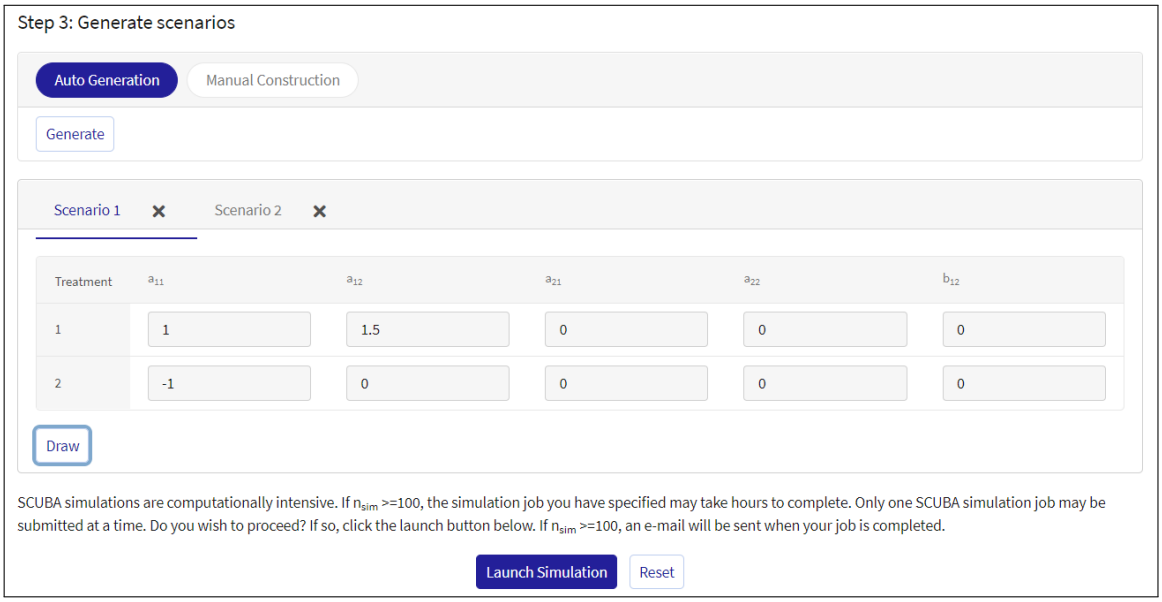


Figure 6.4: Automatically generate scenarios in the **Subgroup Enrichment and Analysis** module.

Manual Construction (Figure 6.5)

Manually input coefficients for each treatment, then click the "Add" button to create these scenarios.

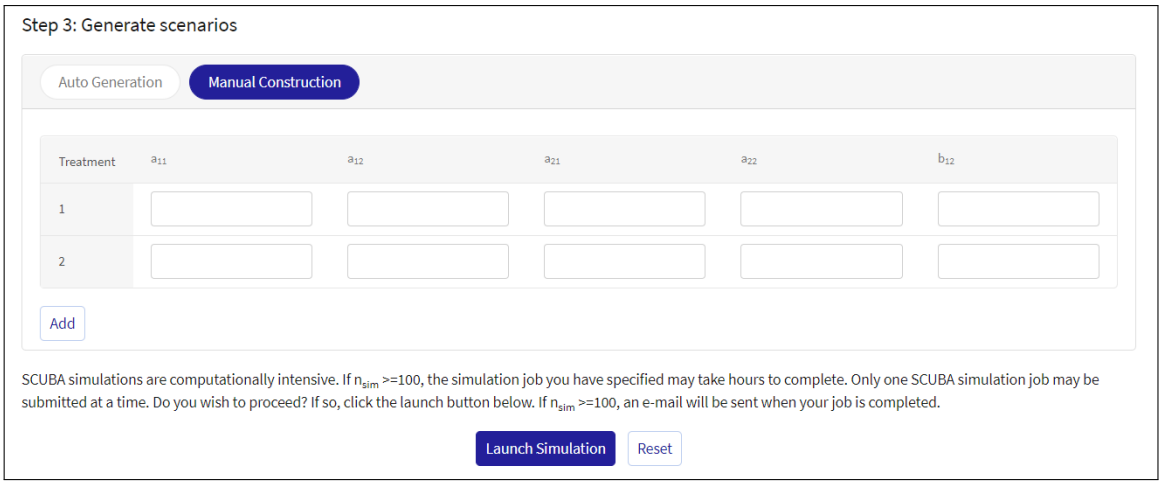


Figure 6.5: Manually generate scenarios in the **Subgroup Enrichment and Analysis** module.

The generated scenarios are displayed under different tabs (Figures 6.4 and 6.5) which appears below the generation section. For each scenario, click "Draw" to visually display the shape of true response rate versus biomarker(s) under each treatment (Figure 6.6). In Figure 6.6, the x-axis and y-axis represent the values of biomarker 1 and 2, respectively, and the color represents the true response rate. The darker the color, the larger the true response rate.

Once scenarios are generated, click the "Launch Simulation" button at the bottom of the page to run n_{sim} (set in step 1) simulations, for each scenario and selected design (set in step 2) combination.

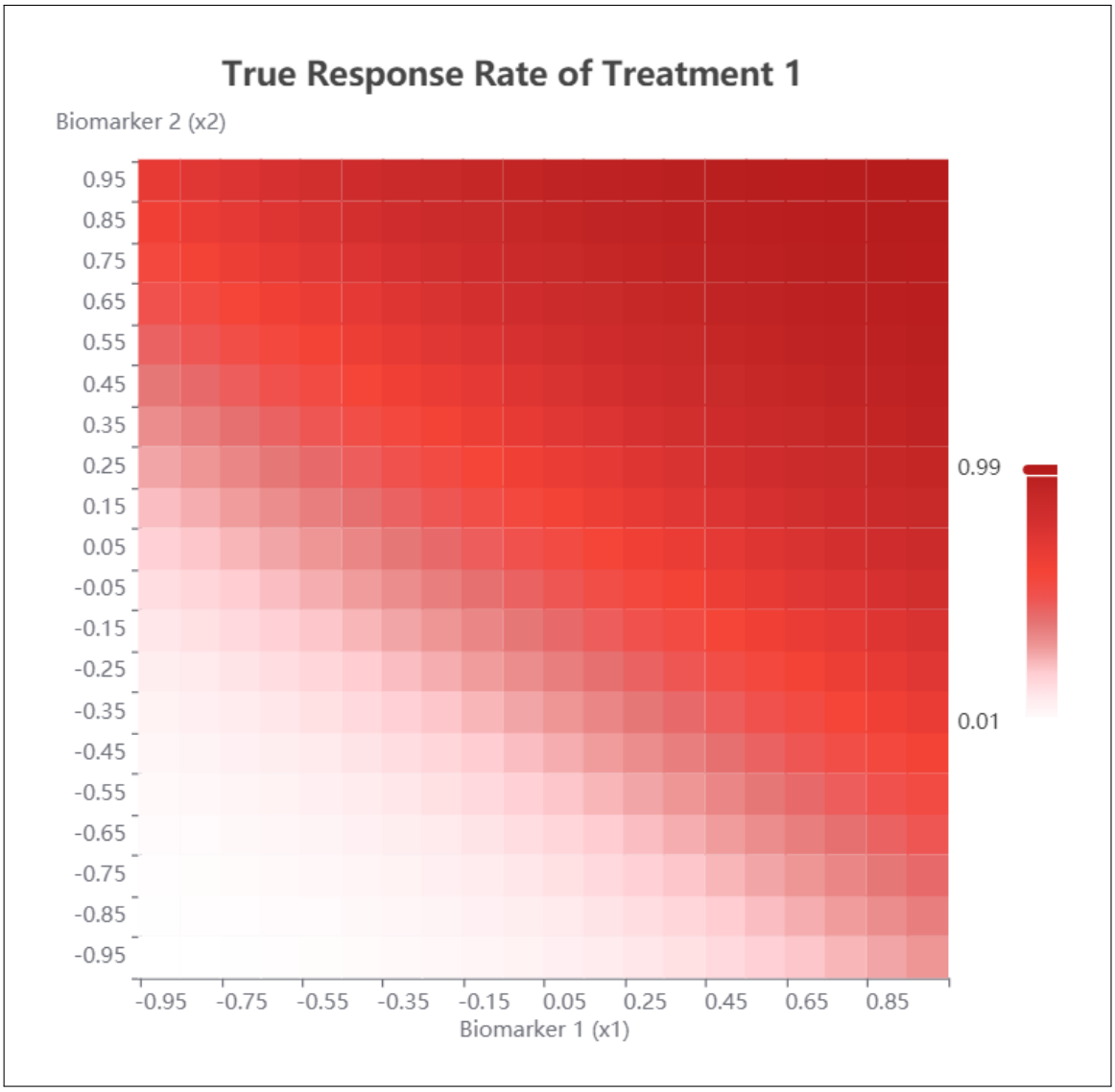


Figure 6.6: An example of the shape of true response rate versus two biomarkers under one treatment in the **Subgroup Enrichment and Analysis** module.

6.2.2.4 Launch Simulation

Once the steps 1-3 are completed, click the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 6.4 and 6.5) to submit the job. A "**Success**" message will be displayed as in Figure 6.7 to indicate the simulation has been successfully launched. Users may click the "Ok" button in the pop-up box to proceed to **Simulation Results** tab and track the simulation processing status and simulation results.

The simulations of this module are computationally intensive. If $n_{sim} \geq 100$, the simulation may take hours to complete and an e-mail will be sent to users when the simulation is finished. Only one SCUBA simulation job may be submitted at a time by a user.

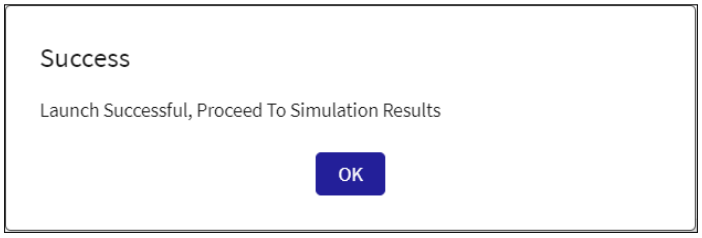


Figure 6.7: "**Success**" message after launching simulation in the **Subgroup Enrichment and Analysis** module.

6.2.3 Simulation Results

The **Simulation Results** tab is primarily used for viewing the simulation jobs and simulation results (§6.2.3.1), restoring simulation settings to make variations in a simulation set as needed (§6.2.3.2), and for downloading simulation reports (§6.2.3.3). Simulation results (figures and tables) can be downloaded in Word format, with accompanying statistical sections in a trial protocol. Hereinafter, we use simulation results and operating characteristics interchangeably.

6.2.3.1 View simulation results

Once simulations are completed, a message appears in the **Running Simulations** panel, and the simulation results are automatically loaded into the **Simulation History** panel (Figure 6.8), a mail icon is used to indicate new results which have not been viewed. The duration displayed depends on the availability of computing resources, and includes the waiting time after submitting the simulation.

Simulation results for other modules can be viewed by using the "Select a Design Category" drop-down box (Figure 6.8).

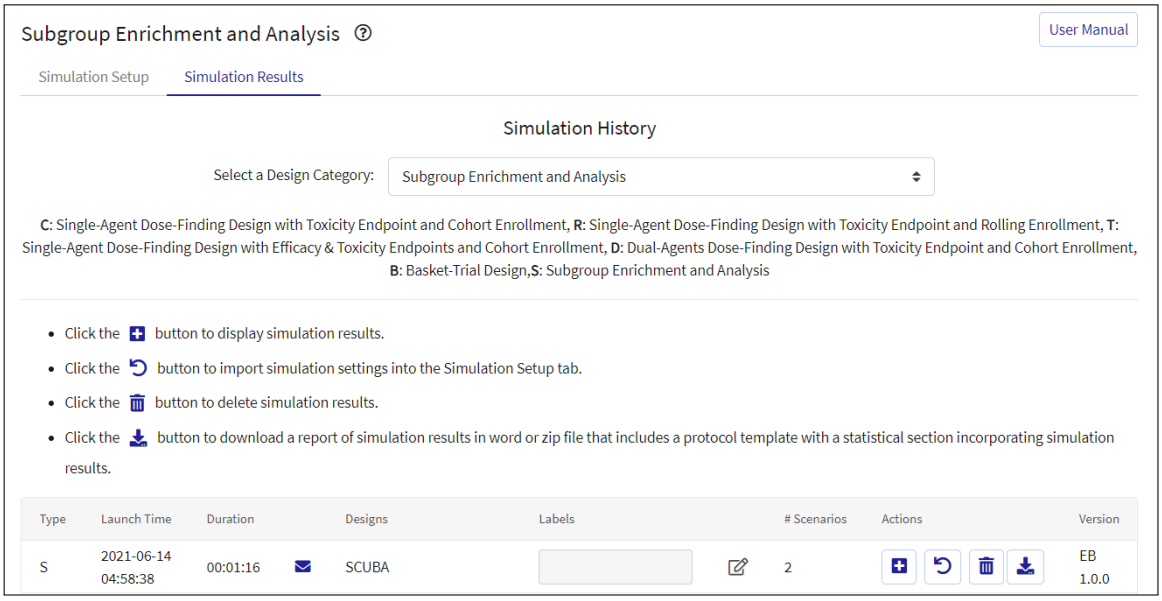


Figure 6.8: Simulation Results in the **Subgroup Enrichment and Analysis** module.

Click the button to expand the pane and view simulation results (Figure 6.9). The design settings are displayed at the top of each simulation study (Figure 6.9) followed by the results in both

tabular and graphical form.

If a set of simulation is no longer needed, click the button to delete the selected simulation results. There is no un-delete option.

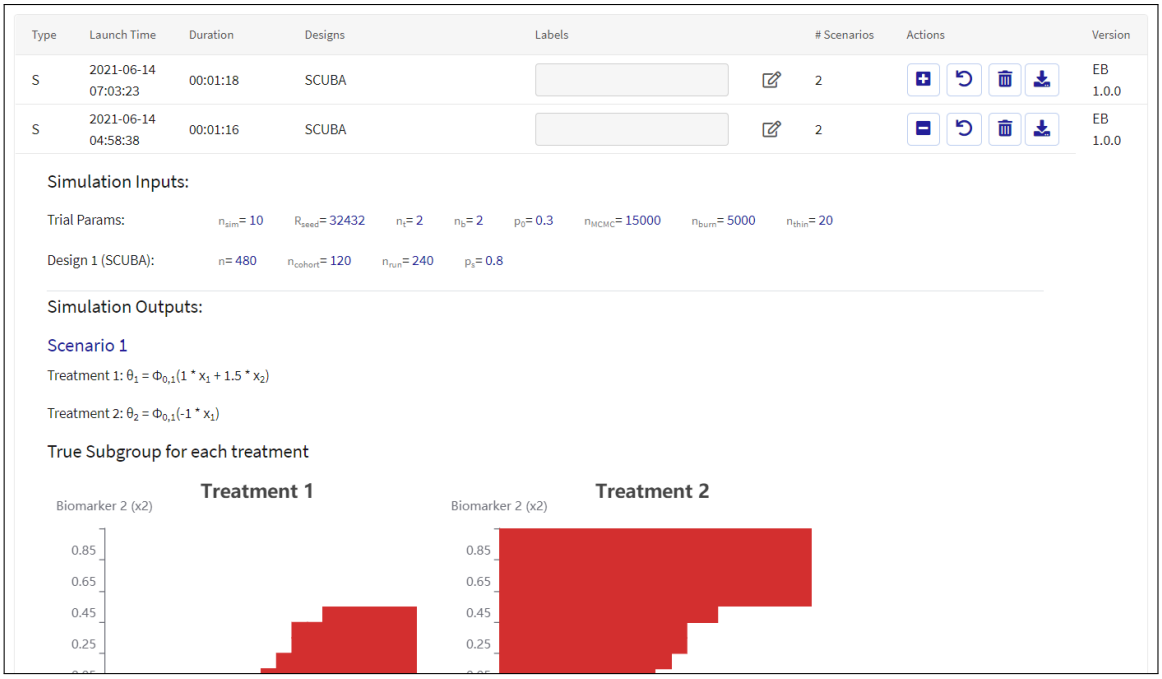


Figure 6.9: View the simulation results in the **Subgroup Enrichment and Analysis** module.

Details of the Simulation Results

Simulation results are presented and arranged by scenarios. There are four sections of simulation result for each scenario:

A. True subgroups for each treatment. (Figure 6.10).

B. Estimated subgroups for each treatment. (Figure 6.11).

C. Table of STP-FDR. (Figure 6.12).

D. Table of Patient Allocation. (Figure 6.13).

A. True Subgroup for each treatment.

These plots shows the true STPs for each arm. The red color represents the true subgroup in which patients have a higher response rate under the treatment or control arm than all other arms. The red pixel is labeled "W" meaning "Winner". The white (blank) region denotes that the arm is not the winner.

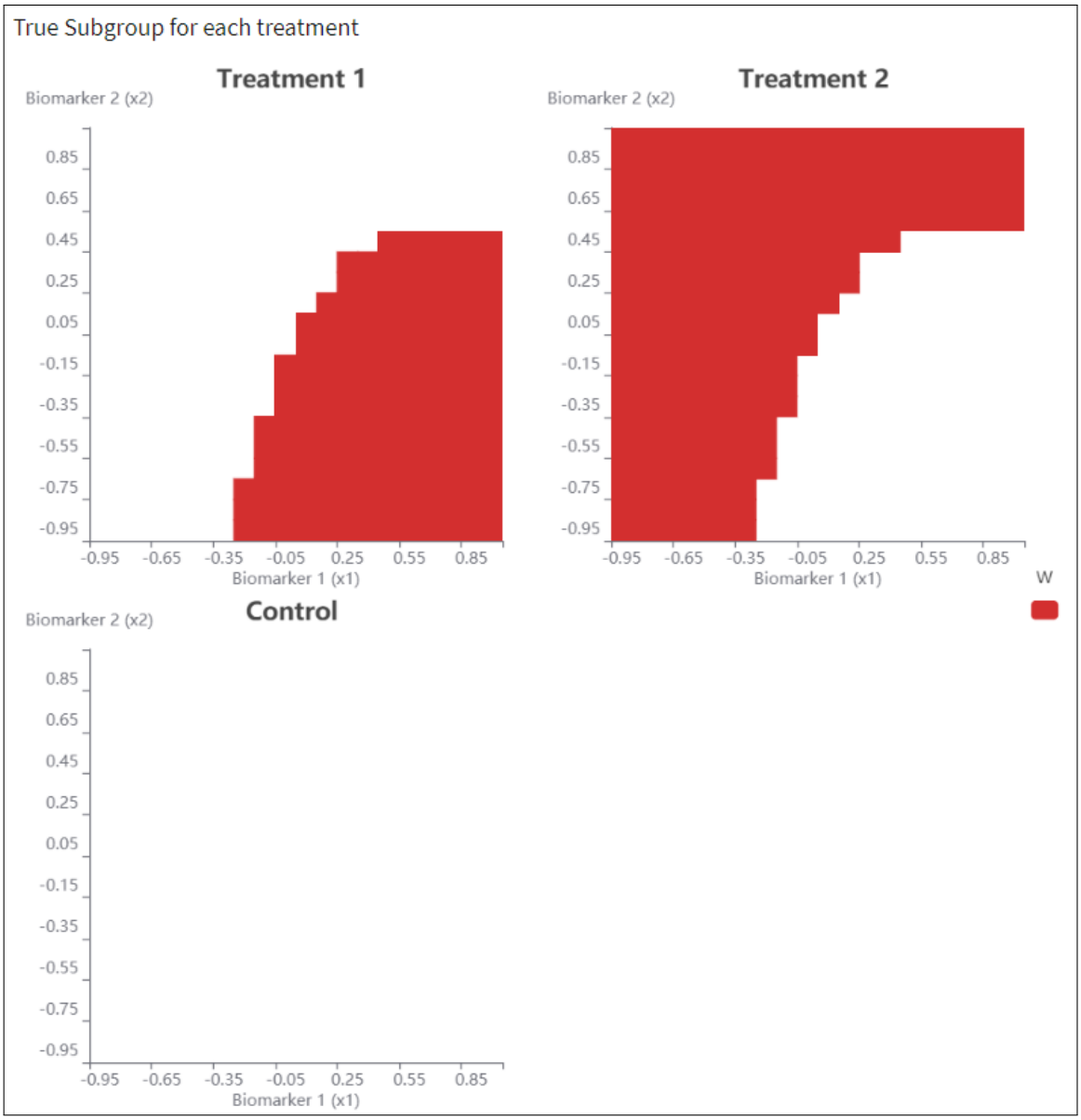


Figure 6.10: True Subgroup for each treatment in the **Subgroup Enrichment and Analysis** module.

B. Estimated Subgroup for each treatment.

These plots show estimated subgroups with the red color representing the frequency of simulated trials in which the arm is selected as the winner in the subgroup. The darker the color, the larger the probability.

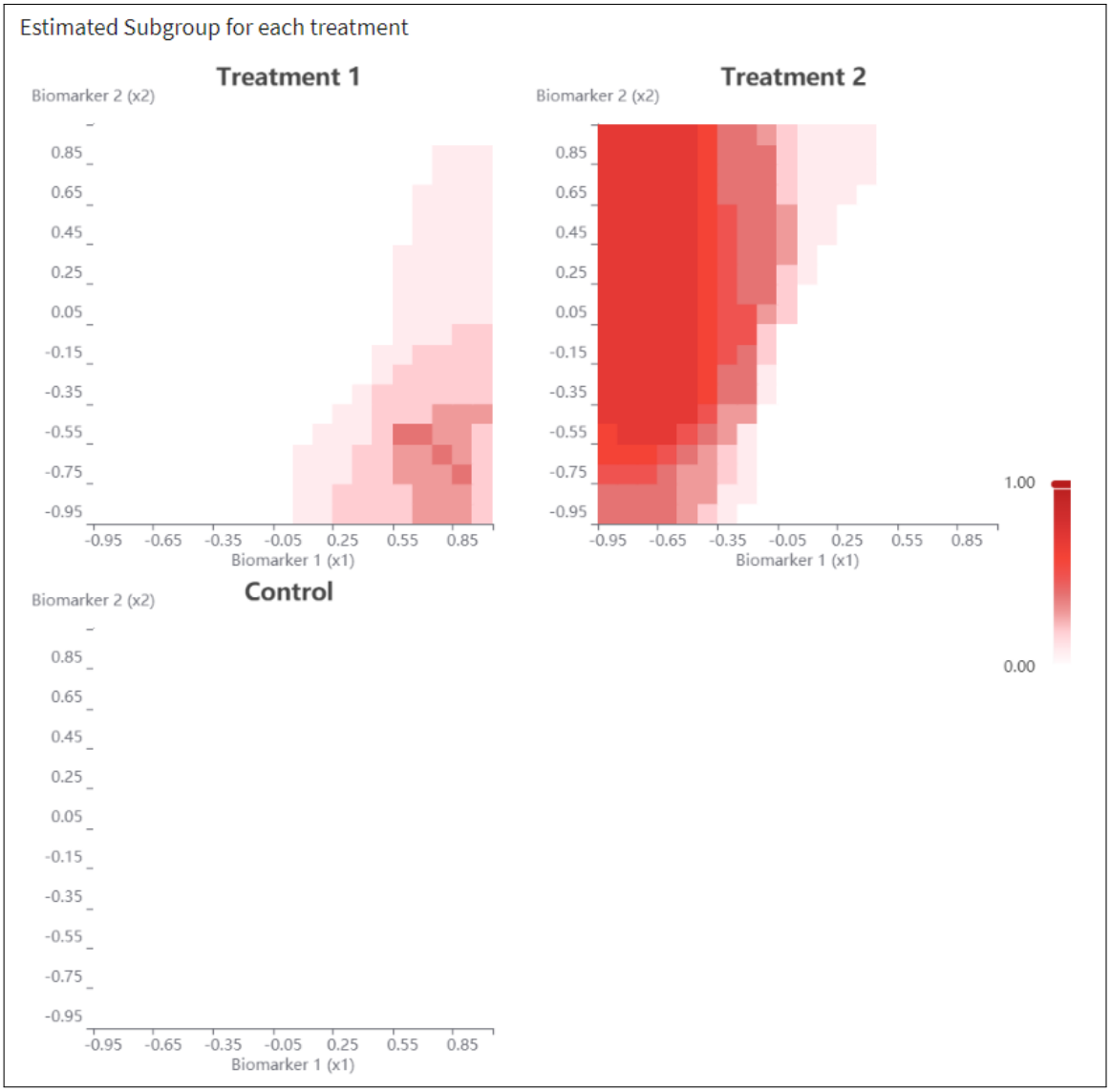


Figure 6.11: Estimated Subgroup for each treatment in the **Subgroup Enrichment and Analysis** module.

C. Table of STP-FDR.

Defining the STP false discovery rate (STP-FDR) as the fraction of the grid points (pixels) that report the wrong winning arm or are outside the true subgroups among estimated STPs, East Bayes reports the mean and standard deviation of STP-FDRs across all the simulated trials. See §6.3.1.4 for details of STP and STP-FDR.

Table of STP-FDR

Treatment 1	Treatment 2	Control	Total
0.012 (0.038)	0.001 (0.004)	0 (0)	0.006 (0.014)

Figure 6.12: Table of STP-FDR in the **Subgroup Enrichment and Analysis** module.

D. Table of Patient Allocation.

According to the simulation truth, SCUBA denotes S_t, a subset of the biomarker space, as the true subgroup in the biomarker space in which patients have a higher response rate under arm t than all the other arms. In other words, S_t is the true optimal subgroup for treatment t, while S_0 denotes the true optimal subgroup for control. The larger the number of patients in S_t assigned to arm t, the better the design.

Table of Patient Allocation

Subgroup	Treatment 1	Treatment 2	Control
S_1	79.6 (10.047)	14.1 (7.279)	1.1 (1.853)
S_2	57.9 (19.496)	86.5 (20.845)	0.8 (1.317)
S_0	0 (0)	0 (0)	0 (0)

Figure 6.13: Table of Patient Allocation in the **Subgroup Enrichment and Analysis** module.

6.2.3.2 Restore simulation setup

If users wish to make a variation of a current design, they can either enter new inputs as they did when creating the current design set, or they can "restore" the simulation input settings from the simulation results by clicking the button (yellow arrow in Figure 6.14). When clicked, this button navigates to the **Simulation Setup** page and pre-populates the input fields.

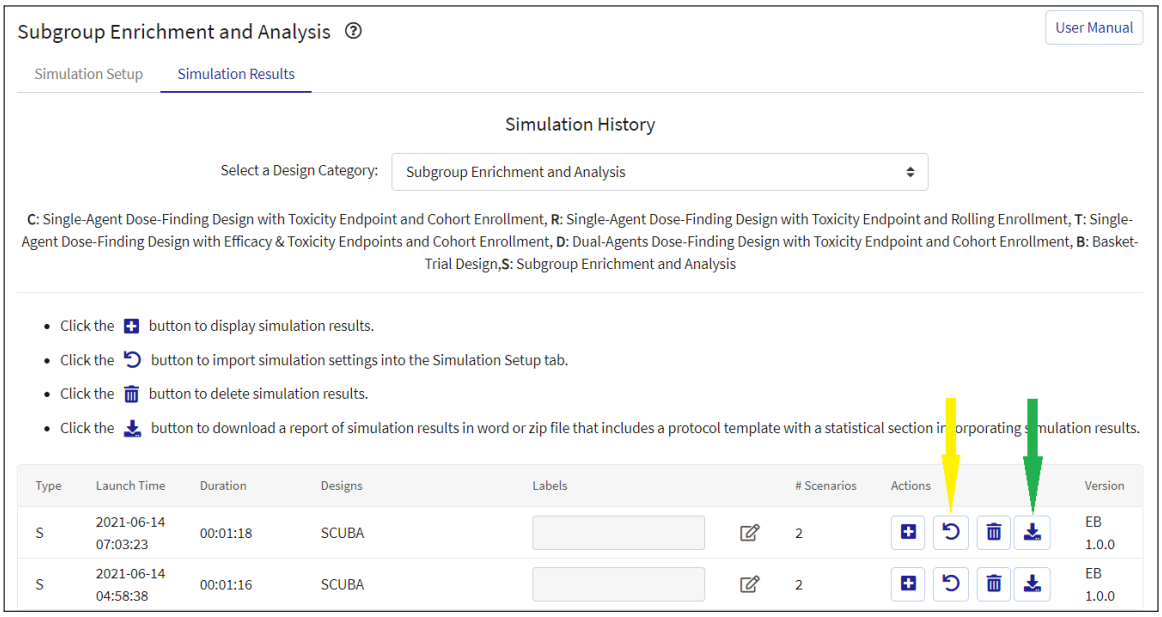


Figure 6.14: Restore simulation setup and download simulation results in the **Subgroup Enrichment and Analysis** module.

6.2.3.3 Download simulation results

The download button button (green arrow in Figure 6.14) creates a Word document, which includes three parts:

- Part A: Complete simulation results for the designs and scenarios users selected
- Part B: Detailed technical descriptions of the designs
- Part C: References

6.3 Statistical Methods Review

6.3.1 The Subgroup Cluster-based Bayesian Adaptive (SCUBA) Design

This section describes the subgroup cluster-based Bayesian adaptive (SCUBA) design proposed by Guo et al. (2017a). The SCUBA design uses lines or planes to partition the continuous biomarker space and define patient subgroups as polygons bounded by these lines or planes. SCUBA also allows a different subset partition for each treatment. The number of linear boundaries in the biomarker space is assumed random, which allows data-driven inference. To borrow strength across subsets, SCUBA assumes a Dirichlet process prior (Ferguson, 1973; Neal, 2000; Hjort et al., 2010) for the response rates across subsets. Therefore, subsets that are geographically distant in the biomarker space can still share the same response rate.

A clinical trial based on SCUBA achieves two goals: 1) enriching the allocation of patients to their precise treatments during the course of the trial and 2) reporting subgroup treatment pairs (STPs) at the end of the trial for future confirmatory studies.

6.3.1.1 Probability Model

Linear Boundary: Suppose under consideration is a total of B biomarkers and T candidate treatments, indexed by $b = 1, \cdots, B$ and $t = 1, \cdots, T$, respectively. Let x_b denote a continuous measurement of biomarker b, such as protein expression. For mathematical convenience, we assume that $x_b \in [-1, 1]$ has been standardized. In the SCUBA design, we assume that the biomarker space may be partitioned differently for different treatments. This would require the partition-related parameters having the treatment index t for mathematical symbols. For simplicity, we suppress the index t in this subsection and will put it back later.

SCUBA uses lines or planes as linear boundaries in the biomarker space $\Omega = [-1, 1]^B$ to define patient subgroups. A linear boundary in the B-dimensional biomarker space can be written as a linear equation, $\sum_{b=1}^{B} \beta_b x_b = c$, where β_b's and c are real values. This general format does not give a unique solution as multiple β's and c's can give the same boundary. To get the unique solution, we impose a constraint, $\sum_{b=1}^{B} \beta_b^2 = 1$, on β_b's. Therefore, a linear boundary s in the B-dimensional biomarker space Ω when $B > 1$ can be written as a standardized linear equation, given by

$$\sum_{b=1}^{B-1} \Big(\prod_{s'=1}^{b-1} r_{s,s'} \Big) \sqrt{1 - r_{s,b}^2} \cdot x_b + \prod_{s'=1}^{B-1} r_{s,s'} \cdot x_B = c_s, \quad s = 1, \cdots, S, \tag{6.2}$$

where $r_{s,b} \in (-1, 1]$ and S is the number of boundaries. When $B = 1$, the boundary can be written $x_1 = c_s$, $s = 1, \cdots, S$ without slope parameters. We assume hereinafter $B > 1$, and remedy to the case when $B = 1$ can be easily made by ignoring the slope r. Since $x_b \in [-1, 1]$ and $c_s \in [-\sqrt{B}, \sqrt{B}]$ there is a 1-to-1 mapping between a linear boundary in Ω and $(r_{s,1}, \cdots, r_{s,B-1}, c_s)$. This facilitates the prior construction for $r_{s,b}$ and c_s later. According to (6.2), the tuple $\boldsymbol{r}_s = (r_{s,1}, \cdots, r_{s,B-1})$ decides the "direction" of the sth boundary and c_s affects the "intercept" of the boundary.

For each direction, we allow up to two parallel linear boundaries to give more flexibility in modeling the biomarker-response surfaces. For example, sometimes response to a treatment is associated with a complex interaction of multiple biomarkers, resulting in a nonlinear biomarker-response surface for both biomarkers (Ala et al., 2013). In other words, we allow $0 \leq M_s \leq 2$ boundaries with the same direction $\boldsymbol{r}_s$, $s = 1, \cdots, B$. This is realized by having M_s number of intercepts $c_{s,a}$, where subscript a index the intercepts with the same direction. Altogether, we allow up to $2 \times B$ lines or planes as subgroup boundaries. Therefore, changing c_s to $c_{s,a}$ we rewrite (6.2) as

$$\sum_{b=1}^{B-1} \Big(\prod_{s'=1}^{b-1} r_{s,s'} \Big) \sqrt{1 - r_{s,b}^2} \cdot x_b + \prod_{s'=1}^{B-1} r_{s,s'} \cdot x_B = c_{s,a}, \quad a = 1, \cdots, M_s, \quad s = 1, \cdots, S, \quad (6.3)$$

Figure 6.15 gives an example of boundaries in the case of two biomarkers. There are two directions in Figure 6.15, with one direction having two lines (dashed) and the other direction having only one line (dotted). Without loss of generality, assume the intercept parameter $c_{s,a}$ is increasing with respect to the index a, that is, $c_{s,a_1} > c_{s,a_2}$ when $a_1 > a_2$. This construction avoids label switching in the posterior inference (McLachlan and Peel, 2004).

Likelihood Function: Hereinafter, we add subscript t to all parameters to allow treatment-specific partitions. For treatment t, define $\boldsymbol{r}_t = \{\boldsymbol{r}_{t,s}, s = 1, \cdots, B\}$ when $\boldsymbol{r}_{t,s} = \{r_{t,s,b}, b = 1, \cdots, B-1\}$ represents the coefficients of the linear boundary for the sth direction, $\boldsymbol{c}_t = \{\boldsymbol{c}_{t,s}, s = 1, \cdots, B\}$ where

$$\boldsymbol{c}_{t,s} = \begin{cases} \varnothing & \text{if } M_{t,s} = 0 \\ \{c_{t,s,1}\} & \text{if } M_{t,s} = 1 \\ \{c_{t,s,1}, c_{t,s,2}\} & \text{if } M_{t,s} = 2 \end{cases}$$

is the sth intercept set for direction $\boldsymbol{r}_{t,s}$, and $\boldsymbol{M}_t = \{M_{t,s}, s = 1, \cdots, B\}$ with $M_{t,s}$ denoting the number of boundaries for direction $\boldsymbol{r}_{t,s}$.

Parameters $(\boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t)$ and their priors induce a random partition $\boldsymbol{\Pi}_t$ for treatment t on the biomarker space Ω. We write the partition $\boldsymbol{\Pi}_t = \{A_{t,1}, \cdots, A_{t,I_t}\}$, where $A_{t,i}$ is the ith partition

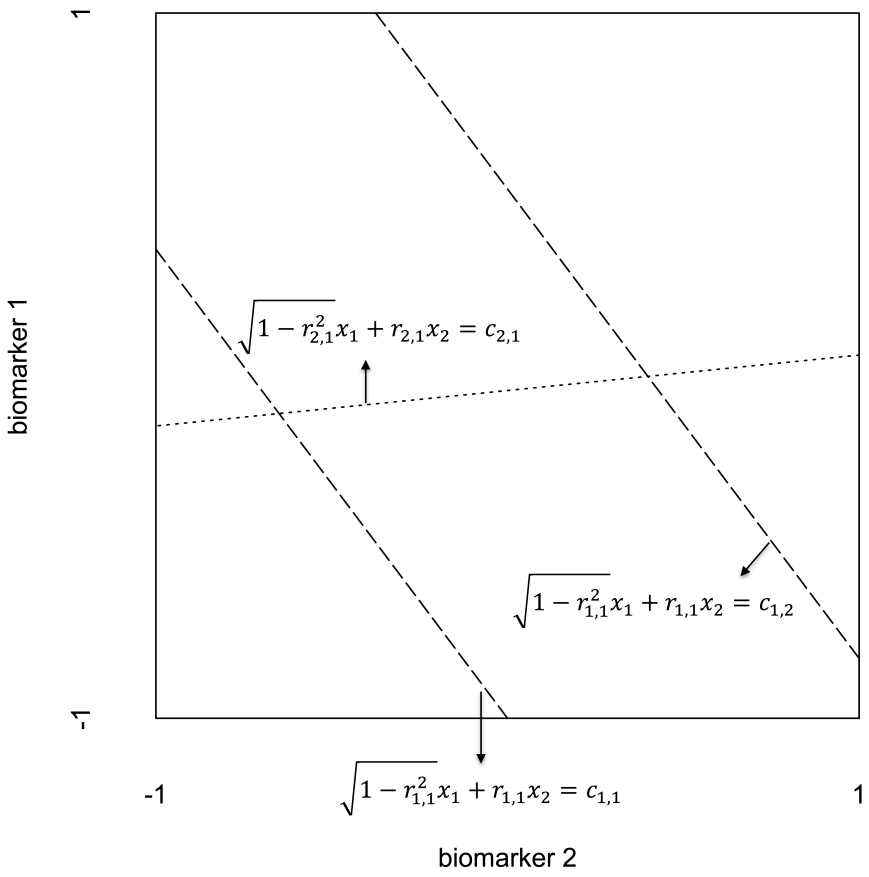


Figure 6.15: (Adopted from Guo et al. (2017a)) An example of partition of a 2-d biomarker space. There are $B = 2$ directions, with $M_1 = 2$ and $M_2 = 1$ linear boundaries for each direction, respectively.

set for treatment t, $i = 1, \cdots, I_t$, and I_t is the random number of partition sets for treatment t. A saturated partition has $M_{t,s} = 2$ boundaries for all directions $s \in \{1, \cdots, B\}$, and every pair of boundaries for one direction intersects the pair of boundaries for another direction. In such a case, there are $I_t = 3^B$ partition sets for treatment t. In general, $I_t \leq \prod_{s=1}^{B}(M_{t,s}+1)$.

Let us consider y_j, the binary outcome for patient j. Let $\boldsymbol{x}_j = (x_{j,1}, \cdots, x_{j,B})$ be the observed biomarker profile and t_j the treatment assignment for patient j, $j = 1, \cdots, n$, respectively. Define $\theta_{t,i} = Pr(y_j = 1 \mid t_j = t, \boldsymbol{x}_j \in A_{t,i})$, the response probability for patients in partition set $A_{t,i}$ for treatment t. The observed data consists of $(y_j, \boldsymbol{x}_j, t_j)$ for all the patients that have been enrolled in the trial. Define $y = (y_1, \cdots, y_n)$, $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, $t = (t_1, \cdots, t_n)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_t)$ and $\boldsymbol{\theta}_t = (\theta_{t,1}, \cdots, \theta_{t,I_t})$, $\boldsymbol{c} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T)$, $\boldsymbol{M} = (\boldsymbol{M}_1, \cdots, \boldsymbol{M}_T)$, $\boldsymbol{r} = (\boldsymbol{r}_1, \cdots, \boldsymbol{r}_T)$, and $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_1, \cdots, \boldsymbol{\Pi}_T)$. The likelihood function is given by

$$\boldsymbol{L}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Pi}) = \prod_j \left\{ \sum_{i=1}^{I_{t_j}} \theta_{t_j,i} \times \mathbf{1}(\boldsymbol{x}_j \in A_{t_j,i}) \right\}^{y_j} \times \left\{ 1 - \sum_{i=1}^{I_{t_j}} \theta_{t_j,i} \times \mathbf{1}(\boldsymbol{x}_j \in A_{t_j,i}) \right\}^{1-y_j}, \tag{6.4}$$

where only one indicator $\mathbf{1}(\boldsymbol{x}_j \in A_{t_j,i})$ equals 1 for patient j across all partition sets i, and the remaining indicators are 0 for the patient.

Prior Models: The joint Bayesian hierarchical model can be written as

$$\boldsymbol{L}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Pi}) p(\boldsymbol{\theta} \mid \boldsymbol{\Pi}) \prod_{t=1}^{T} p(\boldsymbol{\Pi}_t \mid \boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t) p(\boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t). \tag{6.5}$$

In (6.5), $p(\boldsymbol{\Pi}_t \mid \boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t) \equiv 1$ since $(\boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t)$ deterministically decides the partition $\boldsymbol{\Pi}_t$. We only need to specify the prior $p(\boldsymbol{r}_t, \boldsymbol{c}_t, \boldsymbol{M}_t)$.

Assuming the intercept $\boldsymbol{c}_t$ and the slope $\boldsymbol{r}_t$ are independent given $\boldsymbol{M}_t$, we have $p(\boldsymbol{c}_t, \boldsymbol{r}_t, \boldsymbol{M}_t) = p(\boldsymbol{c}_t \mid \boldsymbol{M}_t) p(\boldsymbol{r}_t \mid \boldsymbol{M}_t) p(\boldsymbol{M}_t)$. We allow $M_{t,s}$ to be 0, 1 or 2, and assume a discrete uniform prior with $Pr(M_{t,s} = 0) = Pr(M_{t,s} = 1) = Pr(M_{t,s} = 2) = 1/3$. We construct priors for $\boldsymbol{r}_{t,s}$ and $\boldsymbol{c}_{t,s}$ conditional on $M_{t,s}$. The dimension of $\boldsymbol{c}_{t,s}$ is $M_{t,s}$. When $M_{t,s} = 0$, let $\boldsymbol{c}_{t,s} = \emptyset$. Since $|\boldsymbol{c}_{t,s,a}| \leq \sqrt{B}$, we assume uniform priors as below:

$$c_{t,s,1} \mid M_{t,s} > 0 \sim unif(-\sqrt{B}, \sqrt{B}),$$
$$c_{t,s,2} \mid c_{t,s,1}, M_{t,s} = 2 \sim unif(c_{t,s,1}, \sqrt{B}).$$

Note that the prior model forces $c_{t,s,2} > c_{t,s,1}$ to avoid label switching. Similarly, we take $unif[-1, 1]$ for priors of direction parameters r's.

To complete the prior construction in the model, we propose a Dirichlet process (DP) prior as $p(\boldsymbol{\theta} \mid \boldsymbol{\Pi})$:

$$\begin{aligned} \theta_{t,i}|\boldsymbol{\Pi} &\overset{iid}{\sim} G, \quad t = 1, \ldots, T, \quad i = 1, \ldots, I_t \\ G &\sim DP(\alpha_0, Beta(a_0, b_0)). \end{aligned}$$

We set $\alpha_0 = a_0 = b_0 = 1$. The base measure is then $Beta(1, 1)$, a uniform distribution. The natural clustering characteristic of DP induces possible clusters for the response rates $\{\theta_{t,i}\}$ across treatments and partition sets. This allows borrowing strength using data from all patients.

Posterior Inference:

Based on the joint model (6.5), posterior samples for the parameters are obtained using MCMC simulations. Sampling $M_{t,s}$ among values in 0, 1, or 2 might change the dimension of $\boldsymbol{c}_{t,s}$, $\boldsymbol{r}_{t,s}$, and $\boldsymbol{\theta}_t$. Hence, we make use of reversible jumps (Green, 1995; Richardson and Green, 1997). We make a random choice between changing the value of $M_{t,s}$ to an adjacent status or keeping $M_{t,s}$ at the current value.

Detailed description of the MCMC moves can be found in the web-based supplementary materials for SCUBA (Guo et al., 2017a). Using the posterior samples for all the parameters, we infer the posterior predictive probability described in §6.3.1.3 and the estimated subgroup-treatment pairs (STPs) in §6.3.1.4.

6.3.1.2 Trial Design

The proposed SCUBA design consists of two phases, a run-in phase during which patients are equally randomized to treatments, and an adaptive phase during which patients are allocated using one of two methods defined in §6.3.1.3. After the initial run-in phase, we update the posterior distributions once n_{cohort} new patients' responses are obtained. The trial continues until the specified total sample size is reached.

6.3.1.3 Patient Allocation

For new patients enrolled during adaptive phase in the trial, SCUBA calculates the posterior predictive probability of response under each arm to guide the treatment assignment. Suppose the trial has accumulated data for n patients, including their biomarker profiles, treatment allocations, and responses, denoted by $\boldsymbol{x}(n)$, $\boldsymbol{t}(n)$, and $\boldsymbol{y}(n)$, respectively. Based on the MCMC samples, $\left\{\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{c}^{(k)}, \boldsymbol{M}^{(k)}, \boldsymbol{r}^{(k)}\right), k = 1, \cdots, K\right\}$, the posterior predictive probability of response under

arm t for patient j with biomarker profile $\boldsymbol{x}_j$ among the next n_{cohort} patients is given by

$$\begin{aligned} q_j(t) &= Pr(y_j = 1|\boldsymbol{x}_j, t_j = t, \boldsymbol{y}(n), \boldsymbol{x}(n), \boldsymbol{t}(n)) \\ &= \sum_{\boldsymbol{M}} \int Pr(y_j = 1|\boldsymbol{x}_j, t_j = t, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{M}, \boldsymbol{r}) p(\boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{M}, \boldsymbol{r}|\boldsymbol{y}(n), \boldsymbol{x}(n), \boldsymbol{t}(n)) d\boldsymbol{r} d\boldsymbol{c} d\boldsymbol{\theta} \\ &\approx \frac{1}{K} \sum_{k=1}^{K} Pr\left(y_j = 1|\boldsymbol{x}_j, t_j = t, \boldsymbol{\theta}^{(k)}, \boldsymbol{c}^{(k)}, \boldsymbol{M}^{(k)}, \boldsymbol{r}^{(k)}\right), \\ &\approx \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{i=1}^{I_t^{(k)}} \theta_{t,i} \times \mathbf{1}\left(\boldsymbol{x}_j \in A_{t,i}^{(k)}\right) \right), \end{aligned}$$

where $I_t^{(k)}$ is the number of partition sets and $A_{t,i}^{(k)}$ is the ith partition set of the partition $\Pi_t^{(k)} = \left\{A_{t,1}^{(k)}, \cdots, A_{t,I_t}^{(k)}\right\}$ for treatment t based on the kth MCMC sample, $\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{c}^{(k)}, \boldsymbol{M}^{(k)}, \boldsymbol{r}^{(k)}\right)$. And only one indicator $\mathbf{1}(\boldsymbol{x}_j \in A_{t,i})$ equals 1 for patient j across all partition sets i, and the remaining indicators are 0 for the patient.

Depending on the purpose of the trial, East Bayes allows the two following approaches to allocate patients,

- **Choosing the Winner** : this approach assigns the next cohort of patients to the arm $\hat{t}$ with the highest posterior predictive probability, i.e.,

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, q_j(t). \tag{6.6}$$

- **Adaptive Randomization** (AR): this approach use adaptive randomization based on the frequency of arm t having the highest posterior predicted probability $p_j(t)$, given by

$$p_j(t) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left(\underset{t'}{\operatorname{argmax}}\, q_j^{(k)}(t') = t\right), \tag{6.7}$$

  where $q_j^{(k)}(t') = Pr(y_j = 1|\boldsymbol{x}_j, t_j = t', \boldsymbol{\theta}^{(k)}, \boldsymbol{c}^{(k)}, \boldsymbol{M}^{(k)}, \boldsymbol{r}^{(k)})$. Then AR allocates patient j to arm t with a probability proportional to $p_j(t)$.

6.3.1.4 Report Subgroup-Treatment Pair (STP)

One unique feature of SCUBA is its ability to report multiple STPs in multi-arm clinical trials. This approach works quite well in finding the true STPs with low false discovery rates.

Reporting STPs hinges on the discovery of regions in the biomarker space Ω in which one treatment outperforms all the others. We define an equally spaced grid of H values $\{x_{b,1}, \ldots, x_{b,H}\}$ for the biomarker b where each $x_{b,h} \in [-1, 1]$. Taking the Cartesian product of the grids across all B biomarkers, we obtain a $B-$dimensional grid $\tilde{\boldsymbol{x}}$ of size H^B points. In the MCMC samples, the k^{th} iteration generates a set of boundaries on Ω for each treatment t, denoted by $(\boldsymbol{M}_t^{(k)}, \boldsymbol{c}_t^{(k)}, \boldsymbol{r}_t^{(k)})$. These boundaries subsequently define partition sets $\boldsymbol{\Pi}_t^{(k)} = \{A_{t,1}^{(k)}, \ldots, A_{t,I_t^{(k)}}^{(k)}\}$. For each grid point $\tilde{x}_h$, $h = 1, \ldots, H^B$, we can find the partition set $A_{t,i}^{(k)}$ so that $\tilde{x}_h \in A_{t,i}^{(k)}$. Knowing now the partition set $A_{t,i}^{(k)}$ we record the vector of response rates as $\hat{\boldsymbol{\theta}}_h^{(k)} = (\hat{\theta}_{1,h}^{(k)}, \ldots, \hat{\theta}_{T,h}^{(k)})$ from the same MCMC iteration, which consists of response rates under all different treatments. The collection over all the MCMC iterations, $\{\hat{\boldsymbol{\theta}}_h^{(k)}, k = 1, \ldots, K\}$ can be used to report the best treatment for the hth grid point. For example, given a desired confidence p_s, $p_s \in (0, 1)$, we select the "winning" treatment $\hat{t}_h$ for the hth grid point if

$$\hat{Pr}\left(\theta_{\hat{t}_h,h} > \max_{t \neq \hat{t}_h} \theta_{t,h}\right) = \frac{1}{K} \sum_k \mathbf{1}\left(\hat{\theta}_{\hat{t}_h,h}^{(k)} > \max_{t \neq \hat{t}_h} \hat{\theta}_{t,h}^{(k)}\right) > p_s. \tag{6.8}$$

If $\hat{Pr}\left(\theta_{\tilde{t},h} > \max_{t \neq \tilde{t}} \theta_{t,h}\right) \leq p_s, \forall \tilde{t} \in \{1, \cdots, T\}$, we do not report any winning treatment $\hat{t}_h$ for the grid point $\tilde{x}_h$ and set $\hat{t}_h = \emptyset$. Then over all the grid points, the collection $\{(\hat{t}_h, \tilde{x}_h),\ h = 1, \ldots, H\}$ provides a map of STPs on the biomarker space, allowing blank space to indicate undecided regions.

6.4 Summary

The SCUBA design is capable of handling a trial with multiple treatment arms and providing desirable subgroups for each arm. It can take in more than one continuous biomarker and discover novel subgroups by thresholding the biomarkers adaptively. It is a true precision medicine approach with power statistical modeling and inference.

SCUBA can also be easily applied to simpler trials where only one treatment arm and one control are investigated. And it can easily handle a single biomarker as well.

In Guo et al. (2017a), desirably simulation results have been reported which show that SCUBA is able to discover the true subgroups with reasonable power and allocate patients to their optimal treatments. We refer details to the original publication.

7. Bayesian Efficacy Monitoring with Predictive Probability

7.1 Bayesian Efficacy Monitoring via Predictive Probability

This section describes the Bayesian Efficacy Monitoring via Predictive Probability (henceforth referred to as PP) proposed by (Lee and Liu, 2008). PP design possesses good operating characteristics. At the same time the design is more flexible compared with traditional two- or three-stage designs which can be difficult to follow exactly because the response has to be evaluated at prespecified fixed number(s) of patients.

7.1.1 Model

Denote p as the response rate. Assume p follows a beta prior, $p \sim Beta(a_0, b_0)$. It represents the investigator's previous knowledge or belief of the efficacy of the new regimen. The quantity $a_0/(a_0+b_0)$ reflects how informative the prior is. The quantities a_0 and b_0 can be considered as the number of response and the number of nonresponses, respectively. Thus, $a_0 + b_0$ can be considered as a measure of the amount of information contained in the prior. The larger the value of $a_0 + b_0$, the more informative the prior and the stronger the belief it contains.

Let X denote the number of responses among the current enrolled n patients, so we have X follow a binomial distribution, $X \sim Binomial(p, n)$. Consequently, the posterior distribution of response rate p follows a new beta distribution,

$$p|n, X = x \sim Beta(a_0 + x, b_0 + n - x). \tag{7.1}$$

Set a maximum accrual of patients to $N(N \geq n)$. Thus, the number of responses (Y) in the

potential m ($m = N - n$) future patients follows a beta-binomial distribution,

$$Y|n, m, X = x \sim Beta - Binomial(m, a_0 + x, b_0 + n - x). \tag{7.2}$$

When there are i responses in the remaining m patients, i.e., when $Y = i$, we can get the posterior distribution of response rate p,

$$p|X = x, Y = i \sim Beta(a_0 + x + i, b_0 + N - x - i). \tag{7.3}$$

Let p_0 denote a reference response rate, the effect of the standard treatment. Therefore, through (7.2) and (7.3), PP can be calculated as follows :

$$PP = \sum_{i=0}^{m} Pr(Y = i|X = x) I\{Pr(p > p_0|X = x, Y = i) \geq \theta\}, \tag{7.4}$$

where θ is the probability threshold for declaring efficacy at the end of the trial; and $I\{*\}$ is the indication function, which equals to 1 if the condition satisfies; otherwise, equals to 0.

For example, in a phase II trial, an investigator plans to enroll a maximum of $N = 15$ patients into the study. At a given time, $x = 2$ responses are observed in $n = 10$ patients. We use the prior $Beta(0.5, 0.5)$ and the efficacy declaration threshold $\theta = 0.7$. Therefore, So PP of declaring efficacy (say, $> p_0 = 30\%$) is 0.03, see Table 7.1 for the detail calculation process.

Table 7.1: Illustration of Calculating PP ($N = 15, n = 10, x = 2, p_0 = 0.3, \theta = 0.7$)

Y	$X+Y$	$Pr(Y=i\|X=x)$	$p\|X,Y \sim Beta(a,b)$ in (7.3)		$Pr(p>p_0\|X=x,Y=i)$	$Indicator$[1]	$Prod$[2]
			a	b			
0	2	0.338	2.5	13.5	0.071	0	0
1	3	0.338	3.5	12.5	0.203	0	0
2	4	0.206	4.5	11.5	0.404	0	0
3	5	0.088	5.5	10.5	0.624	0	0
4	6	aa0.026	6.5	9.5	0.804	1	0.026
5	7	0.004	7.5	8.5	0.917	1	0.004
Total		1					**0.03**

[1] ***Indicator*** denotes $I\{Pr(p > p_0|X = x, Y = i) \geq \theta\}$.
[2] ***Prod*** denotes $Pr(Y = i|X = x) \times I\{Pr(p > p_0|X = x, Y = i) \geq \theta\}$.

7.1.2 Decision Criteria

For efficacy monitoring using PP, the following two decision rules are introduced:

- **Early stopping for futility**: the trial will be stopped early and the treatment is declared inefficacious if $PP < P_L$, where P_L is chosen as a small positive constant. $PP < P_L$

indicates that it is unlikely the response rate will be larger than p_0 at the end of the trial given the current information. When this happens, we may as well terminate the trial.

- **Early stopping for efficacy**: the trial will be stopped early and the treatment is declared efficacious if $PP > P_U$, where P_U is chosen as a large positive constant. $PP > P_U$ indicates that it has a high probability to conclude that the treatment is efficacious at the end of the study, if the same trend as the current data continues. That is, the current collected data provides enough evidence to stop the trial early due to efficacy.

And the details about how to setup P_L, P_T and P_U see the future function Search.

7.1.3 Design

With any number of patients before the end, we can calculate a value of PP, and then decide whether to early stop and declare efficacy or futility by comparing PP with P_L and P_U. Exactly as the flexibility of PP design, there is not a fixed trial design. Any cohort size is adaptable. And even the cohort size can be one, so it allows continuous monitoring of the trial outcome. See the next subsection for details.

7.1.4 An Example

Consider a example that a study is expected to enroll 20 patients. During the trial, after 10 patients have assessed their primary endpoint, when there are new patients' outcomes, the decision of early stop for efficacy or futility will be made by comparing the boundary values obtained based on PP with the actual responses of primary endpoint.

If input parameters as shown on the left panel of Figure 7.1 and click Submit, we can get the result on the right panel of Figure 7.1. The futility and efficacy boundary values are shown in Table 7.2.

Figure 7.1: An Example: Bayesian Efficacy Monitoring by Predictive Probability

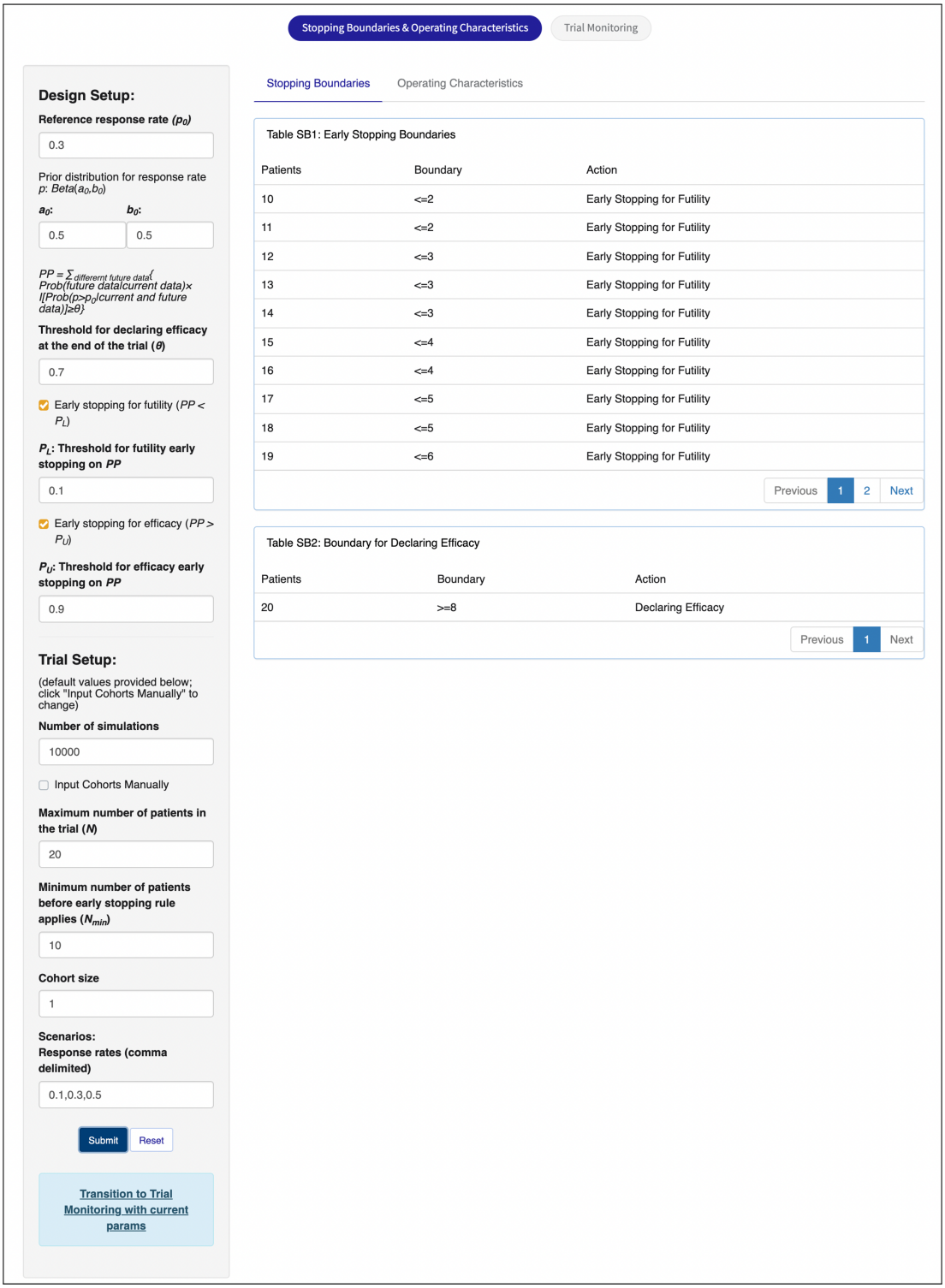


Table 7.2: Futility and Efficacy Boundary Values by Predictive Probability

Early Futility Boundary					
Number of patients (with primary endpoint assessed)	$10 \sim 11$	$12 \sim 14$	$15 \sim 16$	$17 \sim 18$	19
Early stop for futility, if number of responses	$<= 2$	$<= 3$	$<= 4$	$<= 5$	$<= 6$
Early Efficacy Boundary					
Number of patients (with primary endpoint assessed)	$10 \sim 12$	$13 \sim 15$	$16 \sim 19$		
Early stop for efficacy, if number of responses	$>= 6$	$>= 7$	$>= 8$		
Efficacy Boundary reaching the maximum sample size					
Declaring efficacy, if number of responses	$>= 8$				

Specifically, if the number of responses is less than or equal to the futility boundary, the study may be early stopped for futility (e.g., when there are 15 patients having been assessed with less than or exactly 4 responses, early stopping for futility is permitted in this trial.); if the number of responses is more than or equal to the efficacy boundary, the study may be early stopped for efficacy (e.g., when there are 16 patients having been assessed with more than or exactly 8 responses, early stopping for efficacy is permitted in this trial.). If the trial don't stop early for futility or efficacy, and more than or exactly 8 responses are observed in final 20 patients, the treatment will be considered effective, otherwise futile.

8. Bayesian Efficacy Monitoring with Posterior Probability

8.1 Bayesian Efficacy Monitoring via Posterior Probability

This section describes the Bayesian Efficacy Monitoring via Posterior Probability (henceforth referred to as PoP). PoP design possesses good operating characteristics, more flexible compared with traditional two- or three-stage designs which can be difficult to follow exactly because the response has to be evaluated at pre-specified fixed number(s) of patients, same as PP design (see Section 7.1).

8.1.1 Model

Denote θ as the response rate. Assume θ follows a prior beta distribution, $Beta(a_0, b_0)$. It represents the investigator's previous knowledge or belief of the efficacy of the new regimen. The quantity $a_0/(a_0+b_0)$ reflects how informative the prior is. The quantities a_0 and b_0 can be considered as the number of response and the number of nonresponses, respectively. Thus, a_0+b_0 can be considered as a measure of the amount of information contained in the prior. The larger the value of a_0+b_0, the more informative the prior and the stronger the belief it contains.

Let X denote the number of responses in current n patients, so we have X follow a binomial distribution, $X \sim Binomial(\theta, n)$. Consequently, the posterior distribution of response rate θ follows a new beta distribution,

$$\theta|n, X = x \sim Beta(a_0 + x, b_0 + n - x). \tag{8.1}$$

8.1.2 Decision Criteria

For efficacy monitoring using posterior probability, the following three decision rules are introduced:

- **Early stopping for futility**: let θ_{fut} be the reference response rate for futility monitoring and P_{fut} be the probability confidence threshold for futility stopping. The trial should be stopped early and the treatment is declared inefficacious if

$$Pr(\theta > \theta_{fut}|n, x) \leq P_{fut}.$$

- **Early stopping for efficacy**: let θ_{eff} be the reference response rate for efficacy monitoring and P_{eff} be the probability confidence threshold for efficacy stopping. The trial should be stopped early and the treatment is declared efficacious if

$$Pr(\theta > \theta_{eff}|n, x) \geq P_{eff}.$$

- **Criterion for declaring efficacy** at the end of the trial: let $\theta_{eff.final}$ be the reference response rate and $P_{eff.final}$ be the probability confidence threshold for declaring efficacy at the end of the trial. The treatment is declared efficacious if

$$Pr(\theta > \theta_{eff.final}|n, x) \geq P_{eff.final}.$$

For example, assume that there is a clinical trial which has enrolled **10** patients ($\boldsymbol{n = 10}$) and among these 10 patients **2** patients responds ($\boldsymbol{x = 2}$). We use the prior $a_0 = 0.5, b_0 = 0.5$. So the posterior probability of θ is as follows $\theta|n = 10, X = 2 \sim Beta(2.5, 8.5)$. If we use the $\theta_{fut} = 0.3$ as the response rate for futility, so the posterior probability of response rate being higher than 0.3 is $Pr(p > 0.3|n = 10, X = 2) = 0.25$. If we use the futility threshold $P_{fut} = 0.3$, the trial will be stopped early.

8.1.3 Design

With any number of patients before the end, we can calculate values of

$$Pr(\theta > \theta_{fut}|n, x), Pr(\theta > \theta_{eff}|n, x) \text{ and } Pr(\theta > \theta_{eff.final}|n, x),$$

and then decide whether to early stop and declare efficacy or futility by comparing them with P_{fut}, P_{eff} and $P_{eff.final}$. Exactly as the flexibility of PoP design, there is not a fixed trial design. Any cohort size is adaptable. And even the cohort size can be one, so it allows continuous monitoring of the trial outcome. See the next subsection for details.

8.1.4 An Example

Consider a example that a study is expected to enroll 20 patients. During the trial, after 10 patients have assessed their primary endpoint, when there are new patients' outcomes, the decision of early stop for efficacy or futility will be made by comparing the boundary values obtained based on PoP with the actual responses of primary endpoint.

If input parameters as shown on the left panel of Figure 8.1 and click Submit, we can get the result on the right panel of Figure 8.1. The futility and efficacy boundary values are shown in Table 8.1.

Figure 8.1: An Example: Bayesian Efficacy Monitoring by Posterior Probability

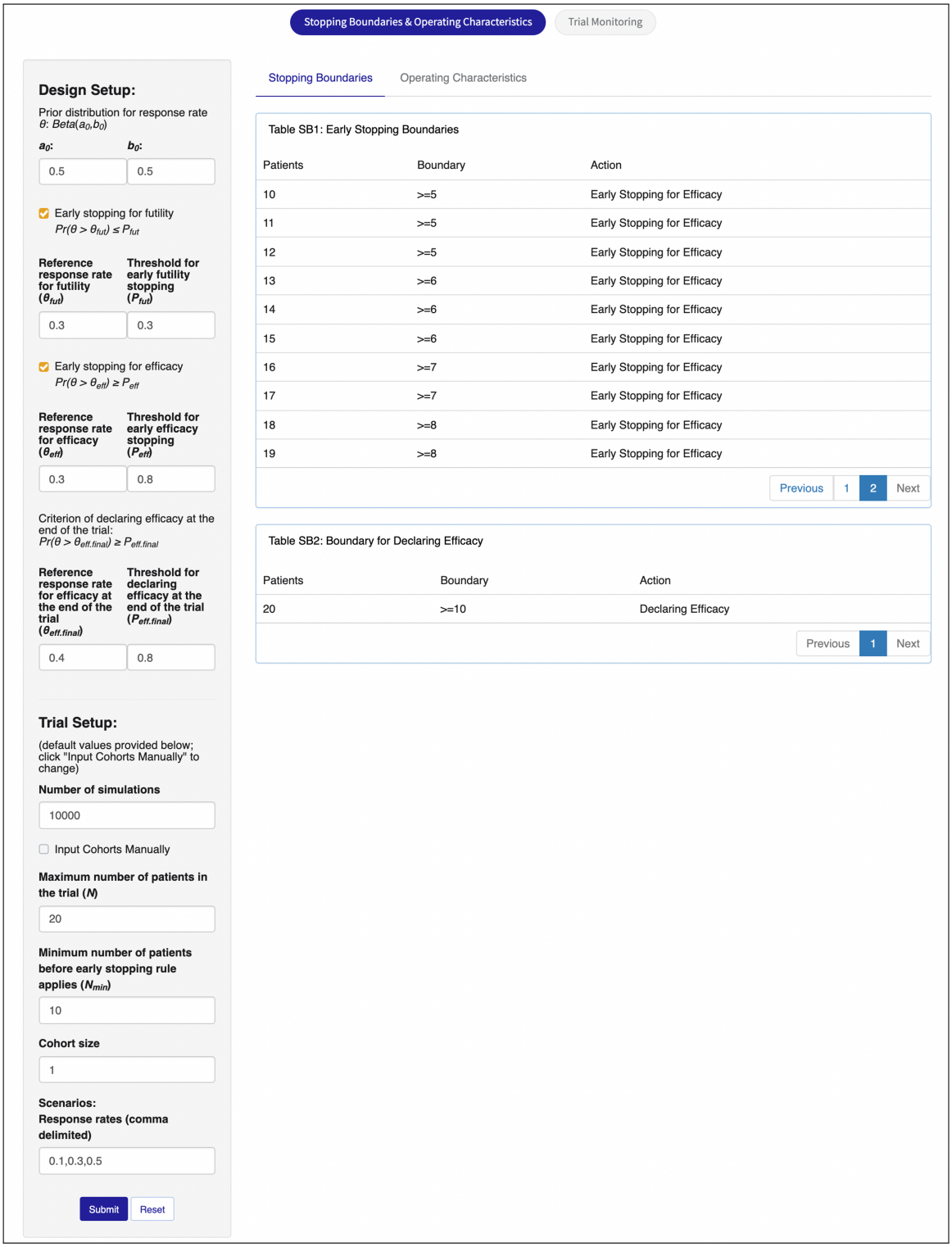


Table 8.1: Futility and Efficacy Boundary Values by Posterior Probability

Early Futility Boundary				
Number of patients (with primary endpoint assessed)	$10 \sim 13$	$14 \sim 16$	$17 \sim 19$	
Early stop for futility, if number of responses	$<= 2$	$<= 3$	$<= 4$	
Early Efficacy Boundary				
Number of patients (with primary endpoint assessed)	$10 \sim 12$	$13 \sim 15$	$16 \sim 17$	$18 \sim 19$
Early stop for efficacy, if number of responses	$>= 5$	$>= 6$	$>= 7$	$>= 8$
Efficacy Boundary reaching the maximum sample size				
Declaring efficacy, if number of responses	$>= 9$			

Specifically, if the number of responses is less than or equal to the futility boundary, the study may be early stopped for futility (e.g., when there are 17 patients having been assessed with less than or exactly 4 responses, early stopping for futility is permitted in this trial.); if the number of responses is more than or equal to the efficacy boundary, the study may be early stopped for efficacy (e.g., when there are 18 patients having been assessed with more than or exactly 8 responses, early stopping for efficacy is permitted in this trial.). If the trial don't stop early for futility or efficacy, and more than or exactly 9 responses are observed in 20 patients, the treatment will be considered effective, otherwise futile.

9. Bayesian Toxicity Monitoring

9.1 Bayesian Toxicity Monitoring via Posterior Probability

This section describes the Bayesian Toxicity Monitoring via Posterior Probability. This design is mostly the same as PoP design (see Section 8.1), the only difference being that this design is used to monitor toxicity but PoP design monitors efficacy. So this design also possesses good operating characteristics, more flexible compared with traditional two- or three-stage designs which can be difficult to follow exactly because the response has to be evaluated at pre-specified fixed number(s) of patients.

9.1.1 Model

Denote θ as the toxicity rate. Assume θ follows a prior beta distribution,

$$\theta \sim Beta(a_0, b_0).$$

It represents the investigator's previous knowledge or belief of the toxicity of the new regimen. The quantity $a_0/(a_0 + b_0)$ reflects how informative the prior is. The quantities a_0 and b_0 can be considered as the number of DLTs and the number of non-DLTs, respectively. Thus, $a_0 + b_0$ can be considered as a measure of the amount of information contained in the prior. The larger the value of $a_0 + b_0$, the more informative the prior and the stronger the belief it contains.

Let X denote the number of DLTs in current n patients,

$$X \sim Binomial(\theta, n).$$

Consequently, the toxicity distribution of toxicity rate θ follows a new beta distribution,

$$\theta | n, X = x \sim Beta(a_0 + x, b_0 + n - x).$$

For toxicity monitoring using toxicity probability, the trial should be stopped if

$$Pr(\theta > \theta_{max} | n, x) \geq \theta_T.$$

9.1.2 Design

With any number of patients before the end, we can calculate a value of $Pr(\theta > \theta_{max}|n, x)$ then decide to whether or early stop for excessive toxicity by comparing them with θ_T. Exactly as the flexibility of this design, there is not a fixed trial design. Any cohort size is adaptable. And even the cohort size can be one, so it allows continuous monitoring of the trial outcome. See the next subsection for details.

9.1.3 An Example

Consider a example that a study is expected to enroll 20 patients. During the trial, after 10 patients have assessed their primary endpoint, when there are new patients' outcomes, the decision of early stop for efficacy or futility will be made by comparing the boundary values obtained based on PoP of toxicity with the actual DLTs of primary endpoint.

If input parameters as shown on the left panel of Figure 9.1 and click Submit, we can get the result on the right panel of Figure 9.1. The futility and efficacy boundary values are shown in Table 9.1.

Figure 9.1: An Example: Bayesian Toxicity Monitoring by Posterior Probability

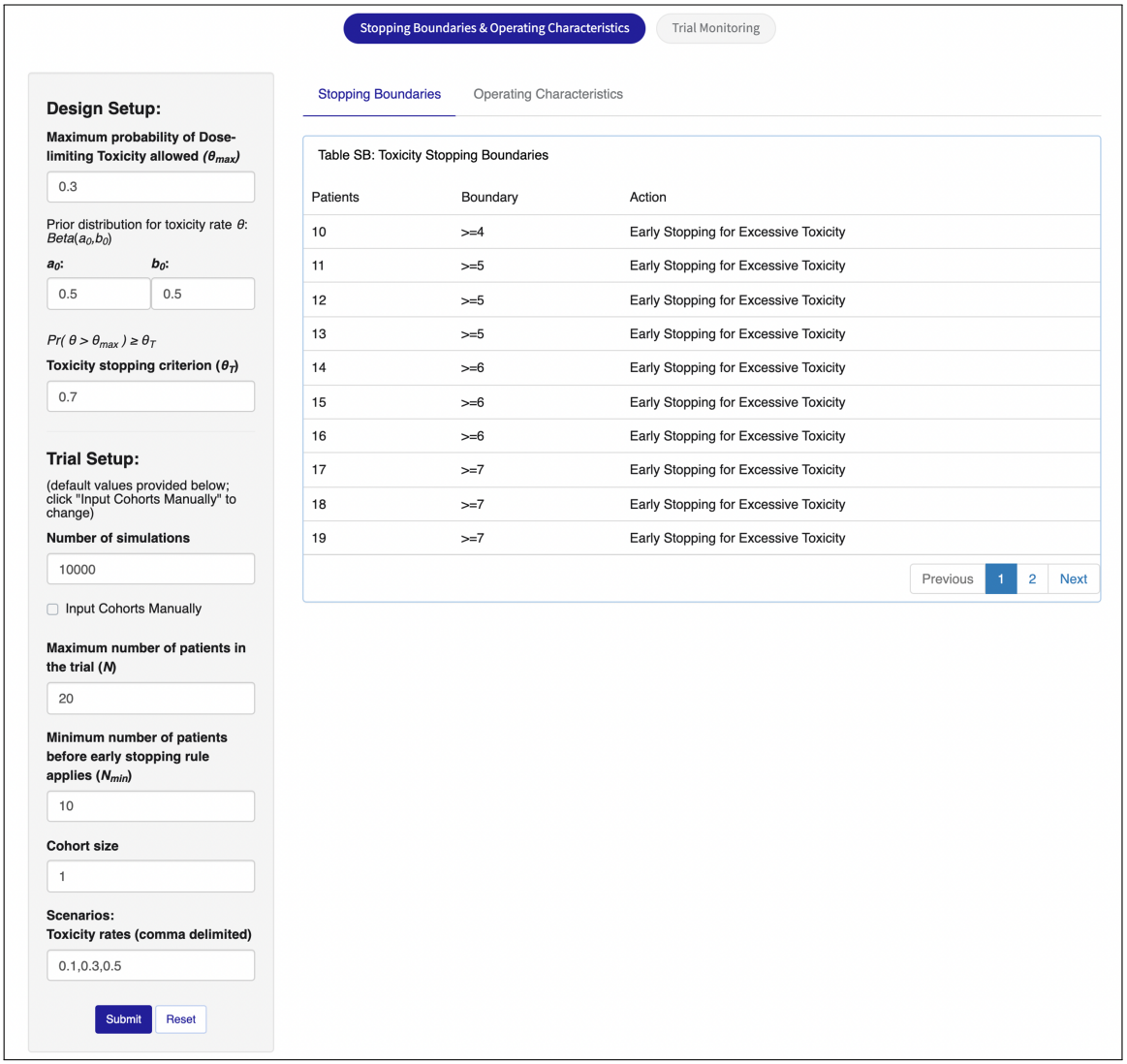


Table 9.1: Futility and Efficacy Boundary Values by Posterior Probability

Early Toxicity Boundary				
Number of patients (with primary endpoint assessed)	10	$11 \sim 13$	$14 \sim 16$	$17 \sim 19$
Early stop for excessive toxicity, if number of DLTs	$>= 4$	$>= 5$	$>= 6$	$>= 7$
Toxicity Boundary reaching the maximum sample size				
Declaring excessive toxicity, if number of DLTs	$>= 8$			

Specifically, if the number of DLTs is more than or equal to the toxicity boundary, the study may be early stopped for excessive toxicity (e.g., when there are 14 patients having been assessed with more than or exactly 6 DLTs, early stopping for excessive toxicity is permitted in this trial.).

10. Bayesian Optimal Design with Simple and Complex Endpoints (BOP2)

10.1 Introduction

This module briefly describes the Bayesian Optimal Design for phase II clinical trials (BOP2) with simple and complex endpoints (Zhou et al., 2017).

The objective of a phase II clinical trial is to evaluate the preliminary efficacy of a new treatment and to determine whether an efficacious treatment warrants investigation in a large-scale randomized phase III trial. A fundamental design feature of phase II clinical trials is the early stopping rule to prevent the exposure of an excessive number of patients to a possibly futile treatment. Numerous designs have been developed for phase II clinical trials. Among frequentist designs, the most well known one is the Simons two-stage design (Simon, 1989), which minimizes the expected sample size or the maximum sample size under the null hypothesis that the treatment is not effective while controlling the type I and type II error rates at desirable levels. Other related work includes Flemings multiple-stage test (Fleming, 1982), Ensigns optimal three-stage design (Ensign et al., 1994), and Chens optimal three-stage design (Chen, 1997), among others.

A number of Bayesian designs has been proposed for phase II trials as well. Thall and Simon (1994) propose using posterior probability to monitor phase II trials and terminate a trial if the interim data indicate that the response rate for the treatment has high posterior probability of being smaller than a prespecified threshold. Heitjan (1997) advocates the use of a persuasion probability to determine whether or not a drug is promising. Tan and Machin (2002) propose two Bayesian two-stage designs that mimic frequentist multistage designs. Lee and Liu (2008) propose a Bayesian phase II design based on posterior predictive probability, and Cai et al. (2014b) introduce a Bayesian phase II trial design that can handle delayed efficacy outcomes through the use of multiple imputation.

Traditionally, phase II clinical oncology trials focus on binary efficacy endpoints, e.g., tumor response. However, more complicated endpoints start to be adopted with the advent of novel molecular targeted agents and immunotherapies. The endpoints for such treatments may be ordinal or multivariate, and the investigators are often interested in simultaneously monitoring multiple types of events in the trial. In this module of **Bayesian Optimal Design with Simple and Complex Endpoints (BOP2)**, East Bayes uses an Rshiny app and performs trial simulation to examine the operating characteristics of the BOP2 design (Zhou et al., 2017). §10.2 introduces the Rshiny user interface and tutorial of launching trial simulations and examining results. A statistical review of the BOP2 design is provided in §10.3.

10.2 User Interface and Tutorial

10.2.1 Overview

Entering the **Bayesian Optimal Design with Simple and Complex Endpoints (BOP2)** page, users will see four main tabs: **Binary**, **Co-primary**, **EffTox** and **Ordinal** as showned in Figure 10.1. They represent the four potential endpoints of the trial.

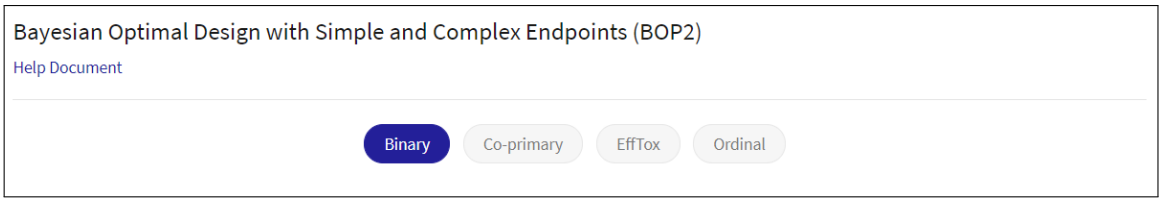


Figure 10.1: The four tabs (endpoints) in the **BOP2** module.

After clicking one of the tabs, an interface appears with **Design Setup** and **Trial Setup** on the left and the **Stopping Boundaries** and **Operating Characteristics** on the right. An example is shown in Figure 10.2 for the Binary tab.

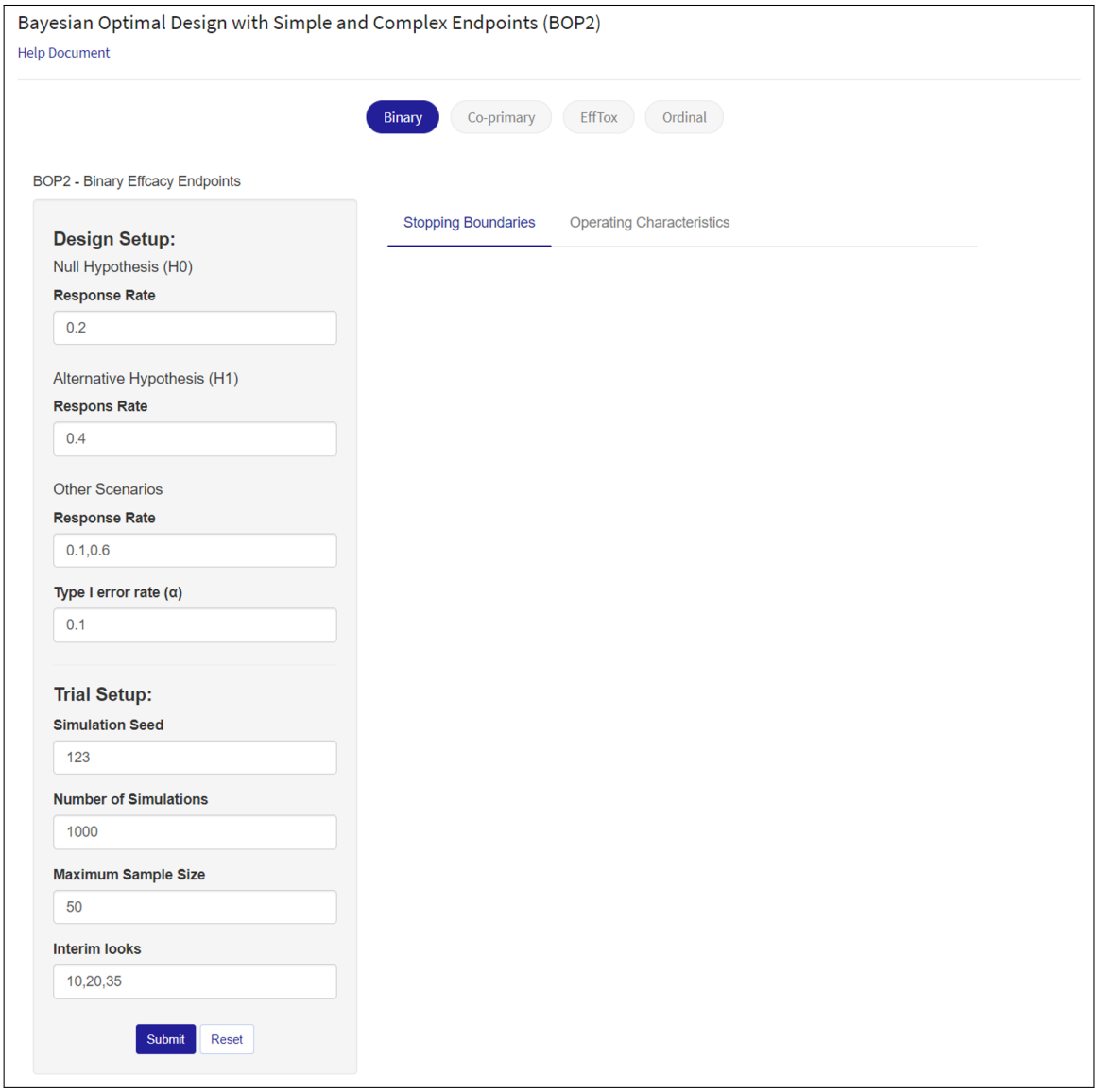


Figure 10.2: An interface for the **Binary endpoints** in **BOP2**.

10.2.2 Simulation Setup

The left box of Figure 10.2 includes two parts, **Design Setup** and **Trial Setup**. In the **Design Setup**, input arguments are required for Null Hypothesis, Alternative Hypothesis, Other Scenarios, and Type I Error Rate (Table 10.1-10.4). In the **Trial Setup**, there are four input arguments, Simulation Seed, Number of Simulations, Maximum Sample Size, and Interim Looks (Table 10.5). Users need to provide these arguments to set up BOP2 simulations.

Upon completing **Design Setup** and **Trial Setup**, users click the "Submit" button at the bottom of the page to launch simulations. Users may also click the "Reset" button next to "Submit" to clear all settings. After the simulations are launched, the results of simulations will be displayed in the **Stopping Boundaries** and **Operating Characteristics** tab on the right (Figure 10.3). Detailed steps of are elaborated next in §10.2.2.1-§10.2.2.2.

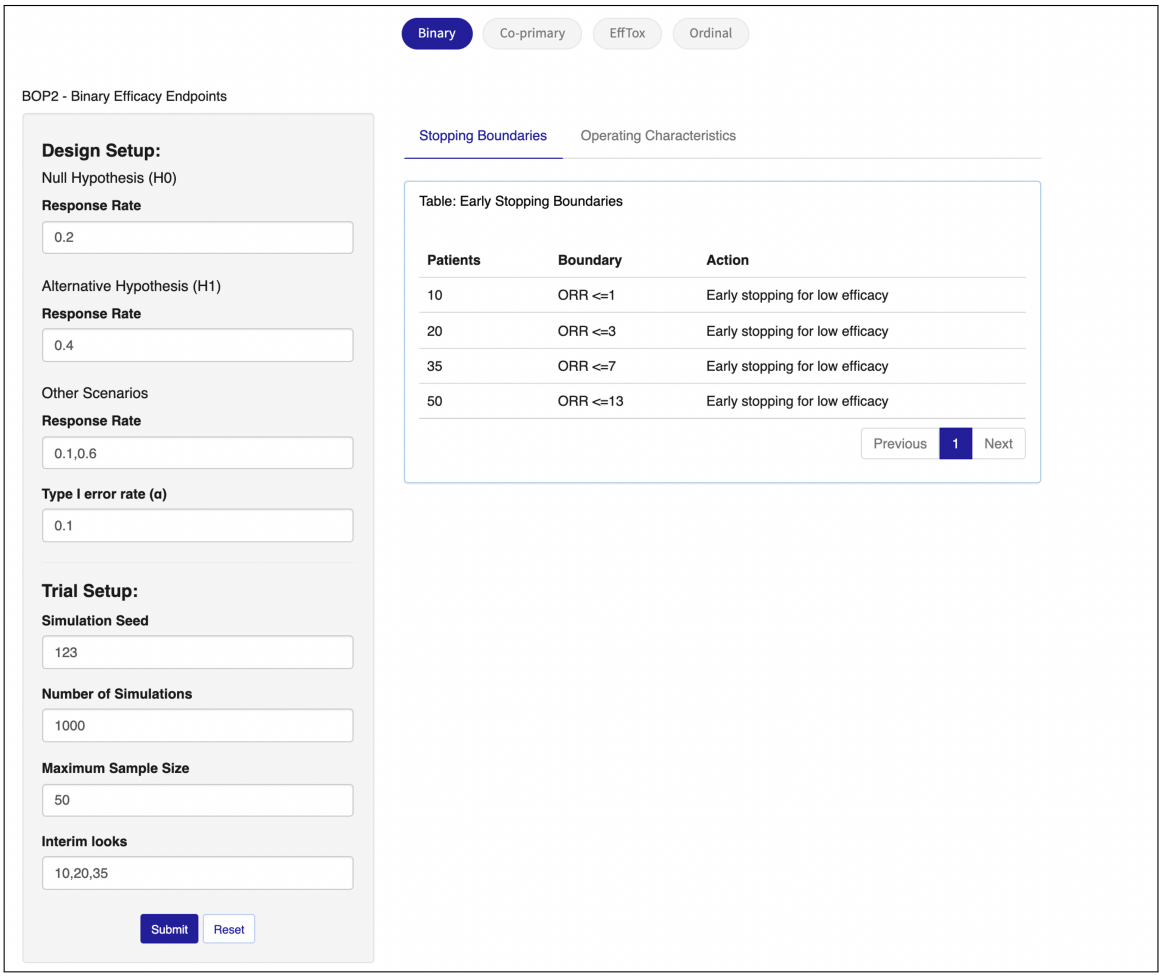


Figure 10.3: An interface with simulation results for the **Binary endpoints** in **BOP2**.

10.2.2.1 Step 1: Design Setup

In Design Setup, first specify the Response Rates under the "Null Hypothesis", the "Alternative Hypothesis", and "Other Scenarios". The detailed explanation of these input arguments for the four endpoints, Binary, Co-primary, EffTox, and Ordinal is provided in Tables 10.1, 10.2, 10.3, and 10.4, respectively.

Table 10.1: Design Setup in the **BOP2: Binary Endpoint**.

Notation	Parameters	Description
	Response Rate – Null Hypothesis	The probability of binary efficacy endpoint under the null hypothesis. The range is (0,1). The default value is 0.2.
	Response Rate – Alternative Hypothesis	The probability of binary efficacy endpoint under the alternative hypothesis. The range is less than 1, and it must be larger than the Response Rate under "Null Hypothesis". The default value is 0.4.
	Response Rate – Other Scenarios	The probabilities of binary efficacy endpoint for other scenarios. Input should be separated by commas. Each input value denotes one new scenario. The range of each input value is (0,1). The default values are "0.1,0.6", representing two scenarios.
α	Type I Error Rate	The probability of rejecting the null hypothesis when it is true. The default value is 0.1.

Table 10.2: Design Setup in the **BOP2: Co-primary Endpoint**.

Notation	Parameters	Description
$Pr(Eff1)$	Response rate of efficacy endpoint 1 – Null Hypothesis and Alternative Hypothesis	The response rate of efficacy endpoint 1 under the null hypothesis or the alternative hypothesis. The range of $Pr(Eff1)$ under the null hypothesis is (0,1) and the range of $Pr(Eff1)$ under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.1 under the null hypothesis and 0.3 under the alternative Hypothesis.
$Pr(Eff2)$	Response rate of efficacy endpoint 2 – Null Hypothesis and Alternative Hypothesis	The response rate of efficacy endpoint 2 under the null hypothesis or the alternative hypothesis. The range of $Pr(Eff2)$ under the null hypothesis is (0,1) and the range of $Pr(Eff2)$ under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.2 under the null hypothesis and 0.35 under the alternative Hypothesis.
$Pr(Eff1\&Eff2)$	Response rate of both efficacy endpoint 1 and 2 – Null Hypothesis and Alternative Hypothesis	The joint response rate of efficacy endpoints 1 and 2 under the null hypothesis or the alternative hypothesis. The ranges of them are both (0,1). The default value is 0.05 under the null hypothesis and 0.15 under the alternative hypothesis. For example, efficacy endpoints 1 and 2 denote the objective response rate (ORR) and EFS6, respectively. Here, EFS6 is a binary endpoint representing whether event-free survival at 6 months is true.
$Pr(Eff1)$	Response rate of efficacy endpoint 1 – Other Scenarios	The response rates of efficacy endpoint 1 for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.2,0.45,0.7".
$Pr(Eff2)$	Response rate of efficacy endpoint 2 – Other Scenarios	The response rates of efficacy endpoint 2 for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.2,0.45,0.6".
$Pr(Eff1\&Eff2)$	Response rate of both efficacy endpoint 1 and 2 – Other Scenarios	The joint response rates of efficacy endpoints 1 and 2 for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.2,0.4". Each combination of $Pr(Eff1)$, $Pr(Eff2)$ and $Pr(Eff1\&Eff2)$ represents one new scenario. The default values represent three scenarios.
α	Type I Error Rate	The probability of rejecting the null hypothesis when it is true. The default value is 0.1.

Table 10.3: Design Setup in the **BOP2: EffTox Endpoint**.

Notation	Parameters	Description
$Pr(Eff\&Tox)$	Probability of efficacy and toxicity – Null Hypothesis and Alternative Hypothesis	The probability of binary efficacy and toxicity endpoints under the null hypothesis or the alternative hypothesis. The range of $Pr(Eff\&Tox)$ under the null hypothesis is (0,1) and the range of $Pr(Eff\&Tox)$ under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.15 under the null hypothesis and 0.18 under the alternative Hypothesis.
$Pr(Eff) - Pr(Eff\&Tox)$	Probability of efficacy and no toxicity – Null Hypothesis and Alternative Hypothesis	The probability of efficacy and no toxicity under the null hypothesis or the alternative hypothesis. The range under the null hypothesis is (0,1) and the range under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.3 under the null hypothesis and 0.42 under the alternative Hypothesis.
$Pr(Tox) - Pr(Eff\&Tox)$	Probability of no efficacy and toxicity – Null Hypothesis and Alternative Hypothesis	The probability of no efficacy and toxicity under the null hypothesis or the alternative hypothesis. The range under the null hypothesis is (0,1) and the range under the alternative hypothesis is larger than 1 and less than that of the null hypothesis. The default value is 0.15 under the null hypothesis and 0.02 under the alternative Hypothesis.
$Pr(Eff\&Tox)$	Probability of efficacy and toxicity – Other Scenarios	The probabilities of efficacy and toxicity for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.2".
$Pr(Eff) - Pr(Eff\&Tox)$	Probability of efficacy and no toxicity – Other Scenarios	The probabilities of efficacy and no toxicity for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.25".
$Pr(Tox) - Pr(Eff\&Tox)$	Probability of no efficacy and toxicity – Other Scenarios	The probabilities of no efficacy and toxicity for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.25". Each combination of $Pr(Eff\&Tox)$, $Pr(Eff) - Pr(Eff\&Tox)$ and $Pr(Tox) - Pr(Eff\&Tox)$ represents one new scenario. The default values represent two scenarios.
α	Type I Error Rate	The probability of rejecting the null hypothesis when it is true. The default value is 0.1.

Table 10.4: Design Setup in the **BOP2: Ordinal Endpoint**

Notation	Parameters	Description
$Pr(CR)$	Probability of complete remission – Null Hypothesis and Alternative Hypothesis	The probability of complete remission under the null hypothesis or the alternative hypothesis. The range of $Pr(CR)$ under the null hypothesis is (0,1) and the range of $Pr(CR)$ under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.05 under the null hypothesis and 0.15 under the alternative hypothesis.
$Pr(PR)$	Probability of partial remission – Null Hypothesis and Alternative Hypothesis	The probability of partial remission under the null hypothesis or the alternative hypothesis. The range of $Pr(PR)$ under the null hypothesis is (0,1) and the range of $Pr(PR)$ under the alternative hypothesis is less than 1 and larger than that of the null hypothesis. The default value is 0.05 under the null hypothesis and 0.15 under the alternative hypothesis.
$Pr(CR)$	Probability of complete remission – Other Scenarios	The probabilities of complete remission for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.2,0.4".
$Pr(PR)$	Probability of partial remission – Other Scenarios	The probabilities of partial remission for other scenarios. Input should be separated by commas. The range of each input value is (0,1). The default values are "0.1,0.2,0.4". Each combination of $Pr(CR)$ and $Pr(PR)$ represents one new scenario. The default values represent three scenarios.
α	Type I Error Rate	The probability of rejecting the null hypothesis when it is true. The default value is 0.1.

10.2.2.2 Step 2: Trial Setup

Table 10.5: Trial Setup in the **BOP2** module.

Parameters	Description
Simulation Seed	The seed for random number generation. The default value is 123.
Number of Simulations	The number of simulated trials. The range is [10,10000]. The default value is 1000.
Maximum Sample Size	The maximum patient number to be enrolled in the trial. The range is $[1, +\infty]$. The default value is 50.
Interim Looks	The numbers of enrolled patients for interim analysis. Input should be integers separated by commas. Each numeral denotes one number of patients for interim analysis. The range of each single numeral is larger than 1 and less than Maximum Sample Size. The default value is "10,20,35".

10.2.2.3 Launch Simulation

Once the simulation setup is completed, users can calculate the stopping boundaries and conduct simulated clinical trials to examine the operating characteristics of the BOP2 design using the generated scenarios, by clicking the "Submit" button at the bottom. Results will be displayed on the right panel after a few seconds (Figure 10.3).

10.2.3 Simulation Results

10.2.3.1 Stopping Boundaries

Once the simulations are completed, two tabs on the right panel, **Stopping Boundaries** and **Operating Characteristics**, will appear. Figure 10.4 presents **Stopping Boundaries** under the Binary efficacy endpoints, which means the trial may be early stopped for futility if the number of responses is less than or equal to the stopping boundary. For example, when 20 patients have been assessed for efficacy and less than or equal to 3 patients responded, the trial is stopped early due to futility.

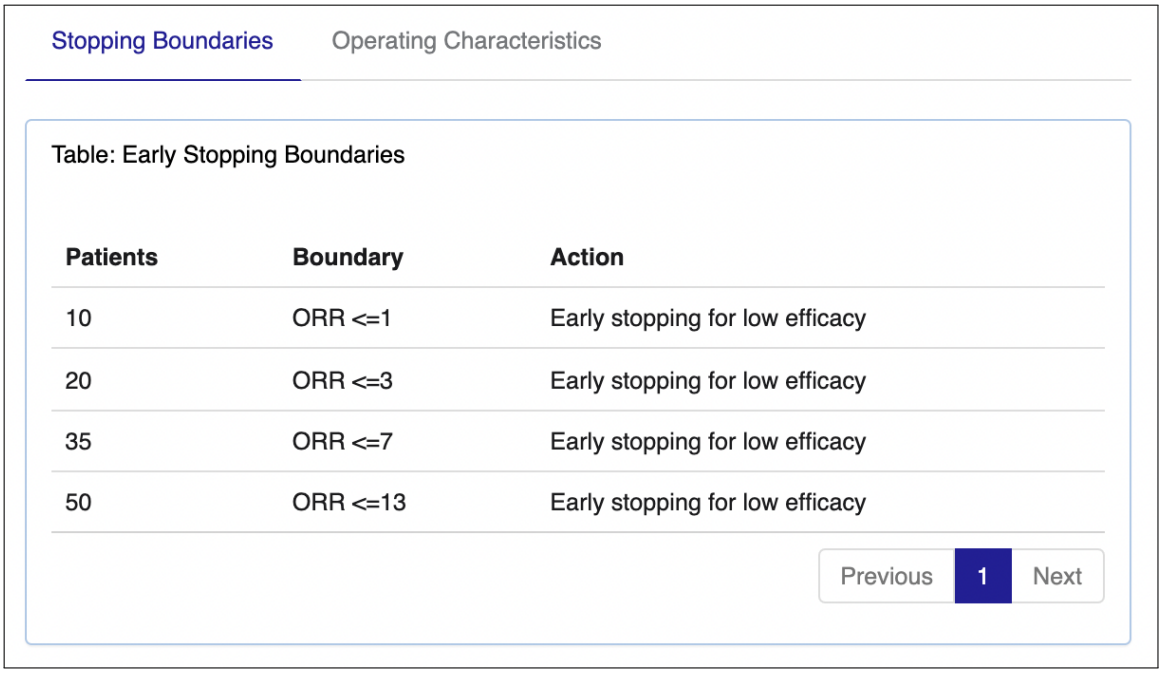


Figure 10.4: Stopping Boundaries in the **BOP2 with Binary endpoint**.

10.2.3.2 Operating Characteristics

There are three sections in **Operating Characteristics**.

A. Table: Operating Characteristics (Figure 10.5).

B. Table: Frequency of Early Stopping (Figure 10.6).

C. Frequency of Early Stopping at Interim Looks (Figure 10.7).

A. Table: Operating Characteristics. Figure 10.5 shows an example of the table.

- **Positive Trial** represents the frequency of simulated trials in which the treatment is deemed efficacy.
- **Average Sample Size** represents that the average number of patients enrolled across all the simulated trials.
- **Early Stop** represents the frequency of simulated trials that stop early due to futility or high toxicity.

Note that the sum of **Positive Trial** and **Early Stop** is not equal to 1, since there exist some simulated trials which did not stop early but the treatment is not deemed efficacy either.

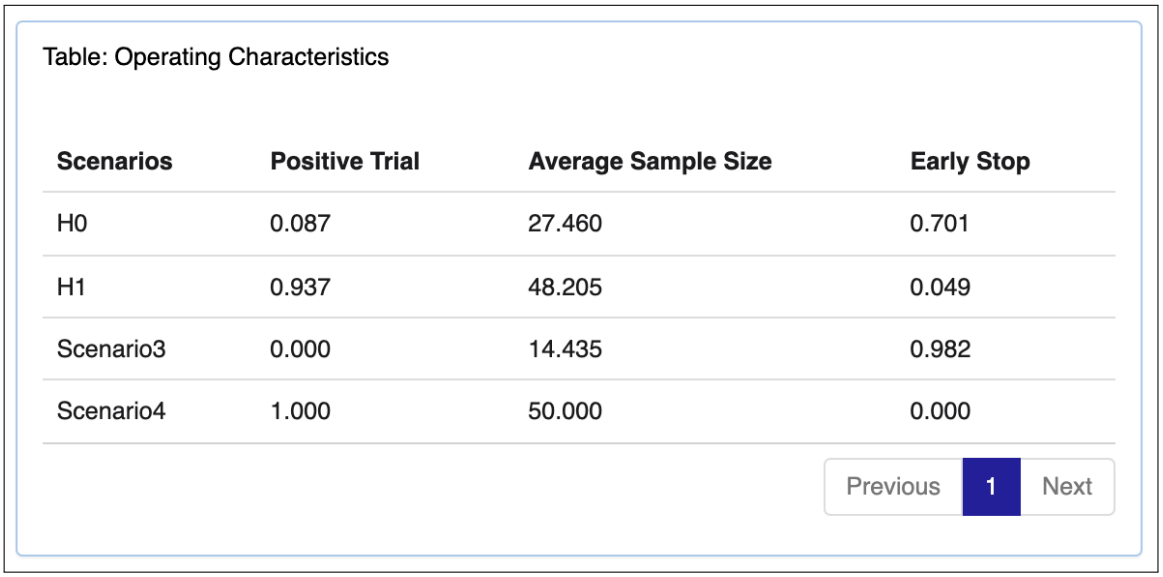


Table: Operating Characteristics

Scenarios	Positive Trial	Average Sample Size	Early Stop
H0	0.087	27.460	0.701
H1	0.937	48.205	0.049
Scenario3	0.000	14.435	0.982
Scenario4	1.000	50.000	0.000

Previous 1 Next

Figure 10.5: Operating Characteristics in the **BOP2 with Binary endpoint**.

B. Table: Frequency of Early Stopping.

In this table (Figure 10.6), **Early Stop** represents the frequency of early stopping at each interim look and **Cumulative Early Stop** the cumulative frequency of early stopping at each interim look.

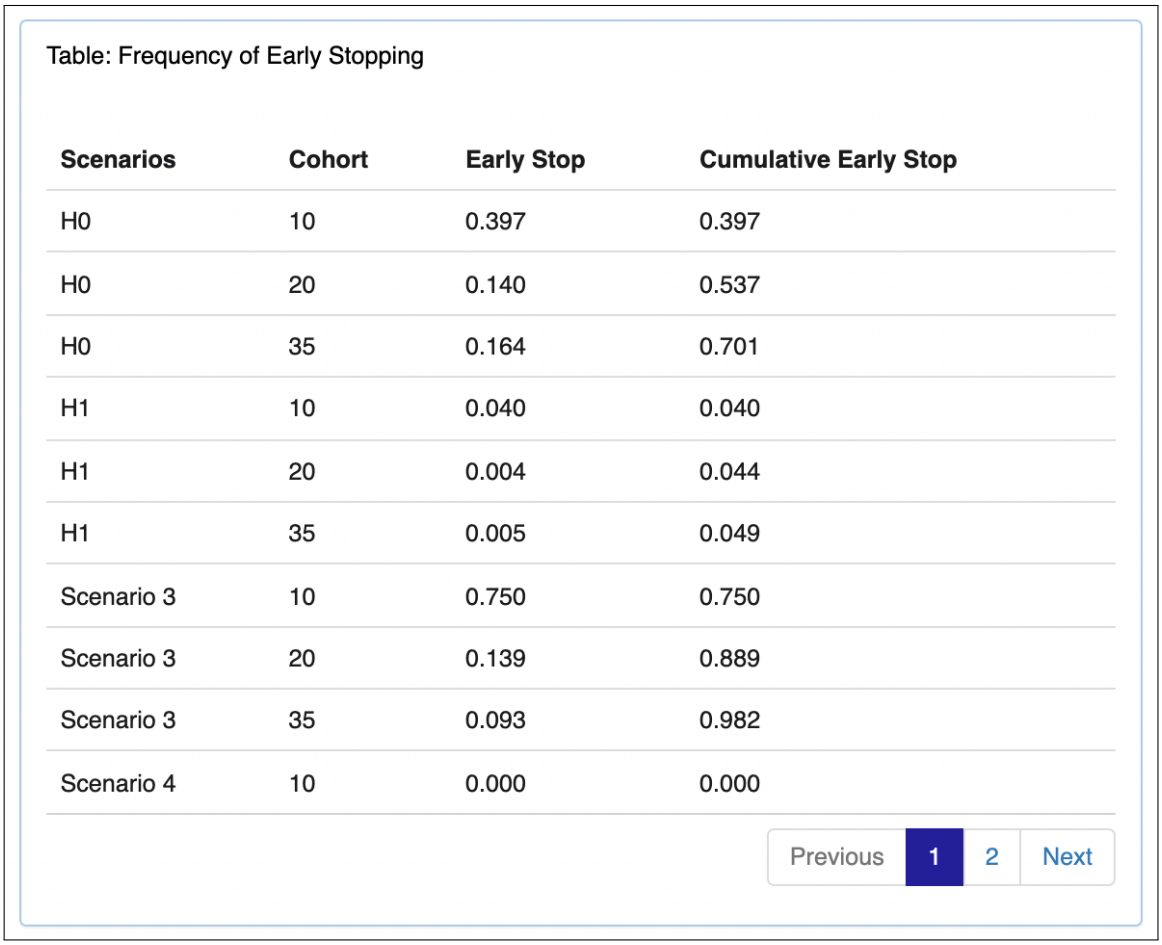


Table: Frequency of Early Stopping

Scenarios	Cohort	Early Stop	Cumulative Early Stop
H0	10	0.397	0.397
H0	20	0.140	0.537
H0	35	0.164	0.701
H1	10	0.040	0.040
H1	20	0.004	0.044
H1	35	0.005	0.049
Scenario 3	10	0.750	0.750
Scenario 3	20	0.139	0.889
Scenario 3	35	0.093	0.982
Scenario 4	10	0.000	0.000

Previous 1 2 Next

Figure 10.6: Frequency of Early Stopping in the **BOP2 with Binary endpoint**.

C. Frequency of Early Stopping at Interim Looks.

The bottom part under **Operating Characteristics** is a bar plot (Figure 10.7) of the frequency of early stopping at the interim analysis. Different colors indicate different scenarios and the sum of the numbers above same color's columns should be equal to 1.

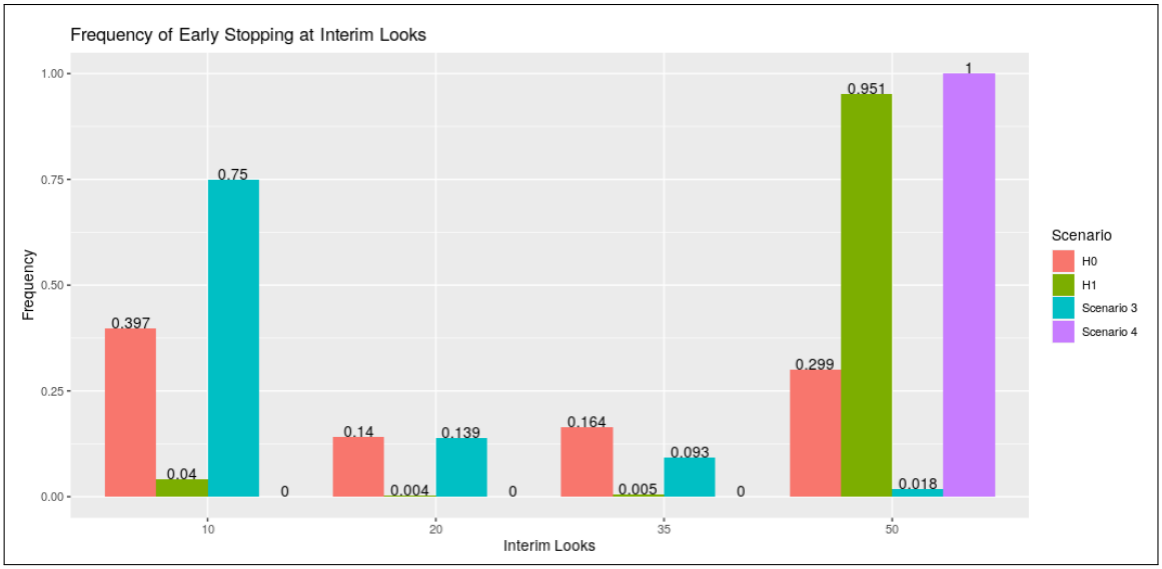


Figure 10.7: Frequency of Early Stopping at Interim Looks in the **BOP2 with Binary endpoint**.

10.3 Statistical Methods Review

10.3.1 Probability Model

Although the endpoints of the aforementioned trials take different forms, they can be unified and represented by a random variable Y that follows a multinomial distribution,

$$Y \sim Multinomial(\theta_1, \theta_2, \cdots, \theta_k), \tag{10.1}$$

where $\theta_k = Pr(Y = k)$ is the probability that Y belongs to the kth category, $k = 1, \cdots, K$. The K categories can be the actual levels of a single endpoint or the combinational levels of multiple categorical endpoints. For example, for co-primary efficacy endpoints (§10.3.4.2), Y is a multinomial variable with four categories where 1 = (OR, EFS6), 2 = (OR, no EFS6), 3 = (no OR, EFS6), and 4 = (no OR, no EFS6). For efftox endpoints (§10.3.4.3), Y is a multinomial variable with four categories: 1 = (toxicity, OR), 2 = (no toxicity, OR), 3 = (toxicity, no OR), and 4 = (no toxicity, no OR). Similarly, for ordinal efficacy endpoints (§10.3.4.4), Y is the ordinal outcome, with $Y = 1, 2, 3$, and 4 denoting CR, PR, SD, and PD, respectively.

Suppose that at an interim look, a total of n patients has been enrolled into the trial and their endpoints have been fully evaluated. Let $D_n = (x_1, \cdots, x_K)$ denote the interim data and x_k denote the number of patients with response $Y = k$, where $\sum_{k=1}^{K} x_k = n$. Assuming that $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ follows a Dirichlet prior,

$$(\theta_1, \theta_2, \cdots, \theta_K) \sim Dir(a_1, \cdots, a_K), \tag{10.2}$$

where $a_1, \cdots, a_K$ are positive hyperparameters. The posterior distribution of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta} \mid D_n \sim Dir(a_1 + x_1, \cdots, a_K + x_K). \tag{10.3}$$

We set $\sum_{k=1}^{K} a_k = 1$ such that the prior is vague and equivalent to a prior effective sample size of 1. In the special case that Y is a binary outcome (§10.3.4.1), the Dirichlet-multinomial model becomes a standard beta-binomial model.

10.3.2 BOP2 Trial Design

Let N denote the maximum sample size of the trial. The proposed BOP2 design consists of R interim looks, which occur when the number of enrolled patients reaches $n_1, \cdots, n_R$, and a final look when all N patients are enrolled. At each of these looks, the go/no-go decision is made on

the basis of the accumulating data, as described in the succeeding texts. In other words, patients are enrolled in $R+1$ cohorts of size $n_1, n_2-n_1, \cdots, n_R-n_{R-1}$ and $N-n_R$, respectively, and the go/no-go decision is made after each cohort is enrolled and their endpoints observed . When $R = N-1$, we obtain a full sequential design in which the go/no-go decision is continuously assessed after each patient. For notational brevity, we suppress the subscript of the interim sample size when this does not cause confusion.

Let $C(n)$ denote a probability cutoff, which is a function of the interim sample size n. Under the proposed design, the go/no-go decision at each interim is made on the basis of the posterior probability of the events of interest. Specifically, for the four endpoints, the interim stopping rule is described as follows. At an interim look, terminate the trial if
(Binary Endpoints, §10.3.4.1):

$$Pr(\theta_1 \leq 0.2 \mid D_n) > C(n); \tag{10.4}$$

(Co-primary Endpoints, §10.3.4.2):

$$Pr(\theta_1+\theta_2 \leq 0.1 \mid D_n) > C(n) \quad \text{and} \quad Pr(\theta_1+\theta_3 \leq 0.2 \mid D_n) > C(n); \tag{10.5}$$

(EffTox Endpoints, §10.3.4.3):

$$Pr(\theta_1+\theta_2 \leq 0.45 \mid D_n) > C(n) \quad \text{or} \quad Pr(\theta_1+\theta_3 > 0.3 \mid D_n) > C(n); \tag{10.6}$$

(Ordinal Endpoints, §10.3.4.4):

$$Pr(\theta_1 \leq 0.15 \mid D_n) > C(n) \quad \text{and} \quad Pr(\theta_1+\theta_2 \leq 0.3 \mid D_n) > C(n); \tag{10.7}$$

Unlike some existing Bayesian designs (Thall and Simon, 1994; Thall et al., 1995; Thall and Sung, 1998), which assume a constant cutoff, here we allow the cutoff $C(n)$ to be a function of the interim sample size n. Although these stopping rules have different clinical interpretations, the go/no-go decisions are all based on the evaluation of a set of the posterior probabilities of the linear combination of the model parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)^T$, for example,

$$Pr(\boldsymbol{b\theta} \leq \phi \mid D_n) > C(n), \tag{10.8}$$

where $\boldsymbol{b}$ is a design vector with elements of 0 and 1 and ϕ is a prespecified threshold.

Given $\boldsymbol{\theta} \sim Dir(a_1+x_1, \cdots, a_K+x_K)$ and a design vector $b = (b_1, \cdots, b_K)$ with elements of 0 and 1, $\boldsymbol{b\theta}$ follows a beta distribution Beta $(\sum_{k=1}^{K} b_k(a_k+x_k), \sum_{k=1}^{K}(1-b_k)(a_k+x_k))$. As a result, $Pr(\boldsymbol{b\theta} \leq \phi \mid D_n)$ can be easily evaluated as

$$Pr(\boldsymbol{b\theta} \leq \phi \mid D_n) = B(\phi; \sum_{k=1}^{K} b_k(a_k + x_k), \sum_{k=1}^{K}(1 - b_k)(a_k + x_k)), \tag{10.9}$$

where $B(\phi; \alpha, \beta)$ is the cumulative distribution function of a beta distribution with parameters α and β, evaluated at value ϕ. This property of $Pr(\boldsymbol{b\theta} \leq \phi \mid D_n)$ leads to the following result.

10.3.3 Optimizing Parameters

Suppose that appropriate null hypothesis H_0 and alternative hypothesis H_1 have been chosen to reflect clinical interests, where H_0 specifies the value of $\boldsymbol{\theta}$, under which the treatment is deemed as futile, and H_1 specifies the value of $\boldsymbol{\theta}$, under which the treatment is deemed as promising. For example, for ordinal efficacy endpoints, $H_0 : \theta_1 = 0.15$ and $\theta_1 + \theta_2 = 0.3$, and a reasonable alternative hypothesis is H_1: $\theta_1 = 0.25$ and $\theta_1 + \theta_2 = 0.5$. With complicated endpoints (e.g., two co-primary efficacy endpoints), the specification of H_1 is less straightforward and should be determined through consultation with clinicians to reflect a desirable outcome that is feasible in practice. We reject H_0 and claim that the treatment is promising if the stopping boundaries are never crossed throughout the trial (including at the end of the trial). The type I error rate and statistical power are defined as the probability of rejecting H_0 under H_0 and H_1, respectively.

The operating characteristics of the BOP2 design rely on the specification of the probability cutoff $C(n)$. Although any reasonably flexible monotonically decreasing function may be used, one particular function of $C(n)$ that is simple and yields good operating characteristics is the following two-parameter power function:

$$C(n) = 1 - \lambda(n/N)^{\gamma}, \tag{10.10}$$

where λ and γ are tuning parameters. We require that $\gamma > 0$ such that $C(n)$ is monotonically decreasing with n/N, the fraction of the accumulated information. The rationale is that at the beginning of the trial, data are sparse and a more relaxed stopping rule with a larger value of $C(n)$ may be preferred to avoid terminating the trial accidentally. When the trial proceeds and information accumulates, we have less uncertainty regarding the endpoint of interest, and thus, it is desirable to have a more stringent stopping rule with a smaller value of $C(n)$ to terminate the trial for an inefficacious treatment.

For choosing the tuning parameters λ and γ, to maximize the power of the BOP2 design while controlling the type I error rate at a certain prespecified level. This can be carried out as follows:

- Step 1: Elicit from clinicians H_0 and H_1 and the desirable type I error rate.

- Step 2: Find the values of (λ,γ) that yield the desirable type I error rate, which can be carried out through a grid search.
- Step 3: Among the set of (λ,γ) identified in step 2, select the one that yields the maximum statistical power as the optimal design parameters.

10.3.4 Examples of Four Different Endpoints

The BOP2 design is capable of handling several types of trials below in a unified framework, which use four different endpoints.

10.3.4.1 Binary Efficacy Endpoint

The aim of a phase II trial is to evaluate the efficacy of pem-brolizumab in patients with advanced small bowel adenocarcinomas. The primary endpoint is the objective response rate (ORR), defined using Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1. The treatment is regarded as futile if the ORR is $\leq 20\%$ and promising if the ORR is $\geq 40\%$. This example is used to illustrate the standard case with a binary efficacy endpoint in this module of East Bayes.

10.3.4.2 Co-primary Efficacy Endpoints

The primary objective of a phase II trial is to evaluate the efficacy of trebananib administered at 15 mg/kg IV per week in patients with persistent or recurrent carcinoma of the endometrium. The trial has two co-primary efficacy endpoints: the ORR and the event-free survival at 6 months (EFS6). The objective response (OR) is defined using RECIST version 1.1. The event-free survival (EFS) is defined as the length of time from the initiation of the treatment to disease progression, death, or beginning a subsequent therapy. The null hypothesis is that the ORR is $\leq 10\%$ and EFS6 is $\leq 20\%$. In other words, the treatment is regarded as futile if the ORR is $\leq 10\%$ and EFS6 is $\leq 20\%$. Clinically significant improvements are defined as a 20% increase in EFS6, or a 15% increase in ORR.

10.3.4.3 Joint Efficacy and Toxicity Endpoints

In a phase II clinical trial, patients with recurrent indolent non-follicular lymphoma are treated with lenalidomide in combination with rituximab. Lenalidomide is administered at 20 mg/day for days 121, and rituximab is administered at 375 mg/m^2 once on day 14 of every 28 days. The primary efficacy endpoint is the response as defined using the 1999 Cheson criteria. Because of large uncertainty regarding the safety of the combination treatment, the trial also monitors dose-limiting toxicity, defined according to the National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) . The lowest acceptable response rate is 45%, and the highest acceptable

toxicity rate is 30%.

10.3.4.4 Ordinal Efficacy Endpoints

The aim of a phase II clinical trial is to assess the efficacy of nivolumab in patients with Hodgkins lymphoma who have not experienced a successful outcome following an autologous stem cell transplant. The revised International Working Group Criteria for Malignant Lymphoma is used to define the efficacy of treatments for lymphoma, categorized as one of four levels of decreasing desirability: complete remission (CR), defined as the disappearance of all evidence of disease; partial remission (PR), defined as the regression of measurable disease and no new sites; stable disease (SD), defined as failure to attain CR, PR or progressive disease (PD); and PD, defined as evidence of any new lesion or an increase in lesion volume $> 50\%$ from the nadir of previously involved sites. In this trial, although both CR and PR are regarded as favorable responses, CR is substantially more desirable. The treatment is regarded as promising if (i) the probability of achieving CR or PR is $> 30\%$ or (ii) the probability of achieving CR is $> 15\%$, where the endpoint of the second condition is a part of the endpoints of the first condition.

11. Dose Ranging Designs

11.1 Introduction

A dose-ranging design is a clinical trial where different doses of a drug are tested against each other to establish efficacy and safety of the drug. Dose-ranging design is usually an early phase II clinical trial which includes a placebo group of subjects, and a few groups that receive different doses of the drug to be tested. One of the major goals of a phase II dose-ranging design study is to identify a correct dose before moving forward to a phase III confirmatory trial. A four-parameter sigmoid Emax (Dragalin et al., 2007) is sufficient to represent all of the observed dose response curves, except one which had a non-monotone shape. When a dose-response relationship is monotonic, the Emax model has been shown to effective and efficient for designing and analyzing dose-response data across a wide range of pharmaceutical studies.

In order to estimate Emax parameters, we adapt Bayesian methodology in this module. Using Markov Chain Monte Carlo method samples are obtained from posterior distribution. This method has the usual advantages of Bayesian methodology in particular along with the point estimates we also have variance for those estimates.

In this module, East Bayes uses an Rshiny app and performs trial simulation to examine operating characteristics of the Emax design (Dragalin et al., 2007). §11.2 introduces the Rshiny user interface and tutorial of launching trial simulations and examining results. A statistical method overview is given in §11.3.

11.2 User Interface and Tutorial

11.2.1 Overview

Entering the **Dose Ranging Designs – Emax** page, users will see two main tabs: **Inputs** and **Outputs** (Figure 11.1). In the **Inputs** tab, there are four steps: 1) **Design Parameters**, 2) **Enrollment Parameters**, 3) **Response Parameters**, and 4) **Simulation Parameters**. Users need to complete the steps 1-4 to set up simulations. Upon completing steps 1-4, users click the "Launch Simulation" button at the bottom of the page. After the simulation is launched, the results of simulations will be displayed in the **Outputs** tab. The simulation process can be monitored in real time at the top of the **Outputs** tab. Detailed steps of using this module are elaborated next in §11.2.2-§11.2.3.

Figure 11.1: The two tabs of the **Emax** module.

11.2.2 Simulation Setup

East Bayes requires users to provide input parameter values for the Emax Bayes design in four steps. After clicking on the question mark icons, a description of parameters used in the section is displayed. If there are parameters you would like to change which are not currently accessible, or designs you would like to see added to this module, please contact us by emailing support@cytel.com.

11.2.2.1 Step 1: Design Parameters

First specify the target response difference from the placebo and the prior parameters of the Emax model. See Figure 11.2. A detailed explanation of these input arguments will be provided in Table 11.1.

Click the "Apply" button (Figure 11.2) to confirm the input design parameters. The "Apply" button changes to "Edit" and can be clicked again to change design parameters as needed.

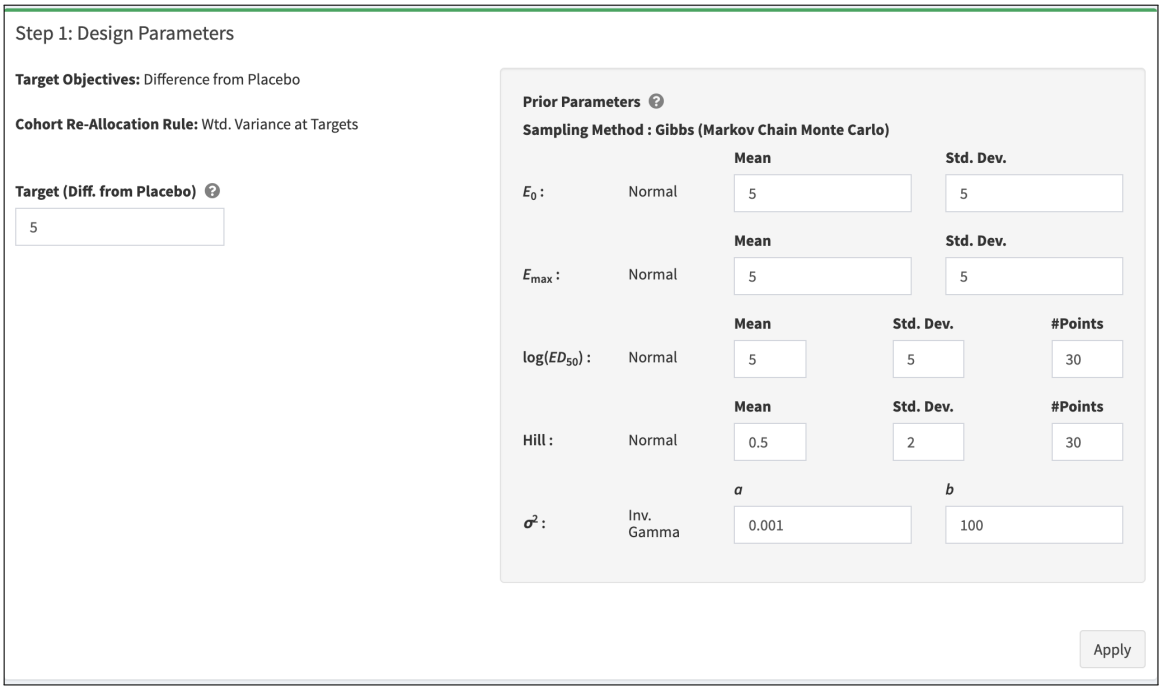


Figure 11.2: Design Parameters in the **Dose Ranging Designs – Emax** module.

Table 11.1: Design Parameters in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Target (Diff. from Placebo)	The target value is defined as the relative difference from the placebo.	$[0, 10^6]$
E_0: Mean	The mean for the prior normal distribution of the minimum value of response (E_0)	$[-10^6, 10^6]$
E_0: Std. Dev.	The standard deviation for the prior normal distribution of the minimum value of response (E_0)	$[10^{-6}, 10^6]$
E_{max}: Mean	The mean for the prior normal distribution for the difference between the maximum and minimum response (E_{max})	$[-10^6, -10^{-6}] \cup [10^{-6}, 10^6]$
E_{max}: Std. Dev.	The standard deviation for the prior normal distribution for the difference between the maximum and minimum response (E_{max})	$[10^{-6}, 10^6]$
log(ED_{50}): Mean	The mean for the prior log normal distribution of the logarithm of the value of the dose with the median response (ED_{50})	$[-13, 13]$
log(ED_{50}): Std. Dev.	The standard deviation for the prior log normal distribution of the logarithm of the value of the dose with the median response (ED_{50})	$[10^{-6}, 10^6]$
Hill: Mean	The mean for the prior truncated normal distribution of Hill (truncated to the left of 0). Hill is the slope factor that controls the rate at which response increases as a function of dose levels.	$[10^{-6}, 10^6]$
Hill: Std. Dev.	The standard deviation for the prior truncated normal distribution of Hill (truncated to the left of 0).	$[10^{-6}, 10^6]$
σ^2: a	The shape parameter a for the prior inverse gamma distribution of the variance of observation σ^2.	$[0, 10^6]$
σ^2: b	The scale parameter b for the prior inverse gamma distribution of the variance of observation σ^2.	$[0, 10^6]$

11.2.2.2 Step 2: Enrollment Parameters

First specify the rate of accural for subjects and the delay time to observe the response in the same time unit. Then specify the allocation ratio of the sample size in placebo and drug groups. The sample size for each group will be calculated according to the allocation ratio after the input of cohort size. See Figure 11.3. A detailed explanation of these input arguments will be provided in Table 11.2.

Click the "Apply" button (Figure 11.3) to confirm the input enrollment parameters. The "Apply" button changes to "Edit" and can be clicked again to change enrollment parameters as needed.

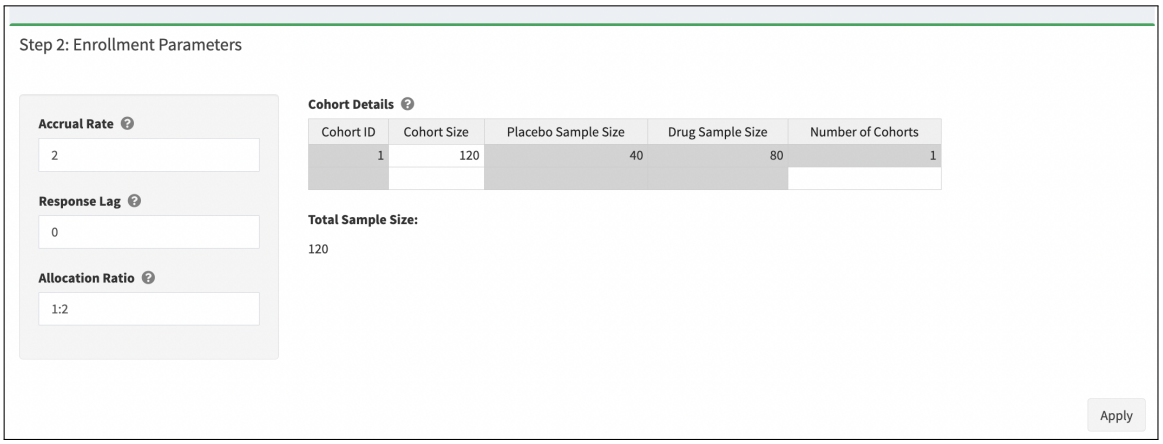


Figure 11.3: Enrollment Parameters in the **Dose Ranging Designs – Emax** module.

Table 11.2: Enrollment Parameters in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Accrual Rate	The number of patients entering the study per unit of time.	$[10^{-6}, 10^{6}]$
Response Lag	The duration between the time of the allocation of subjects to the time when their response is observed. It has the same time unit as Accrual Rate by default.	$[0, 10^{6}]$
Allocation Ratio	The proportion in which allocating patients to placebo and drug respectively in a cohort.	x:y. x and y must be positive integers less than 100
Cohort Details		
Cohort Size	The number of subjects allocated in a particular cohort.	$[1, 10^{6}]$
Number of Cohorts	The number of cohorts in one trial.	$[1, 10]$
Total Sample Size	The total number of subjects in one trial.	$[1, 10^{6}]$

11.2.2.3 Step 3: Response Parameters

In this step, users need to specify the true response value of each dose. First specify the total number of doses and the common value of the standard deviation for the responses of all doses. Then choose a certain type of the curve by clicking "Curve Family" and input particular parameters according to the curve type. The mean response of each dose is generated from the function of the selected curve type and the corresponding dose level. See Figure 11.4. A detailed explanation of these input arguments will be provided in Table 11.3.

Click the "Apply" button (Figure 11.4) to confirm the input response parameters. The "Apply" button changes to "Edit" and can be clicked again to change response parameters as needed.

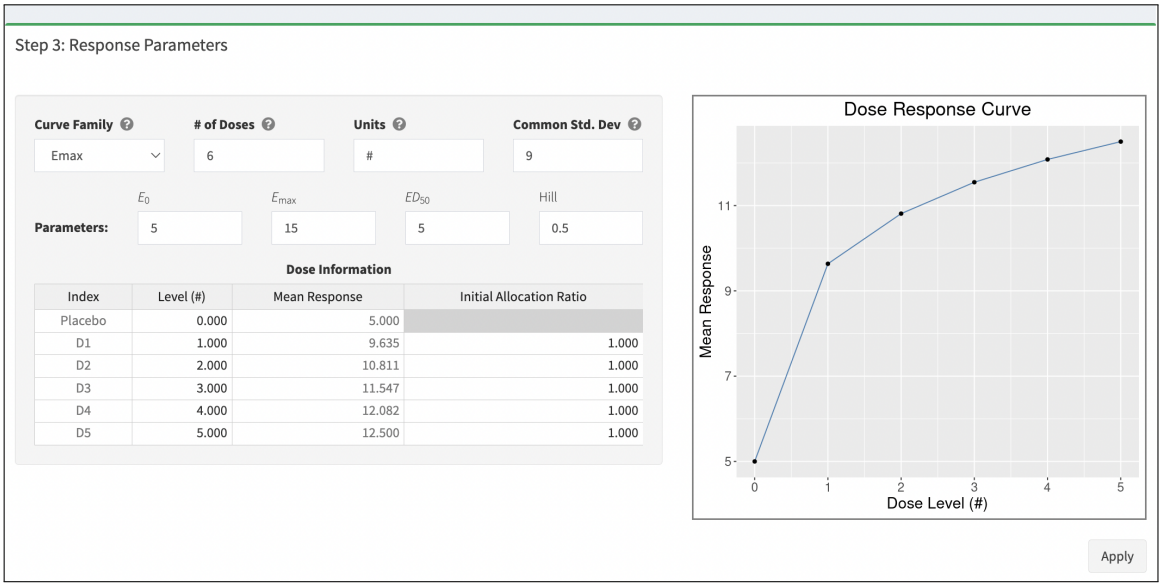


Figure 11.4: Response Parameters in the **Dose Ranging Designs – Emax** module.

When "Curve Family" is selected as "Emax", the mean response of each dose Y is generated from the formula given by,

$$E(Y \mid D) = E_0 + \frac{E_{\max} D^S}{ED_{50}^S + D^S},$$

where D is the dose level and S is a slope factor (i.e., Hill parameter). The detailed explanation of these input arguments is provided in Table 11.4.

Table 11.3: Response Parameters in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Curve Family	A mean dose-response curve from which data will be generated for this simulation study, including Emax, Four Parameter Logistic, Linear, Quadratic and Manually Input to be selected.	
# of Doses	The total number of available doses including placebo.	[2,10]
Units	The unit of the measurement for drug doses.	
Common Std. Dev.	A common value of the standard deviation for the response at each dose.	$[10^{-6}, 10^{6}]$

Table 11.4: Input arguments when "Emax" is selected in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
E_0	The y-intercept of the Emax model, i.e, the mean response for placebo.	$[-10^{6}, 10^{6}]$
E_{max}	The range of difference between the maximum and minimum response. This means that $(E_0 + E_{\text{max}})$ is the upper asymptote.	$[-10^{6}, -10^{-6}] \cup [10^{-6}, 10^{6}]$
ED_{50}	The value of the dose that gives the median response of $E_0 + \frac{1}{2}E_{\text{max}}$. This means that ED_{50} is the dose with an expected response midway between minimum and maximum responses.	$[10^{-6}, 10^{6}]$
Hill	A slope factor that controls the rate at which response increases as a function of D. For fixed E_{max} and ED_{50}, the derivative is proportional to the Hill parameter	$[10^{-6}, 10^{6}]$
Level (#)	The actual dose levels. Dose Levels must be unique values in strictly increasing order.	$[0, 10^{6}]$
Initial Allocation Ratio	The allocation ratio for the first cohort of patients.	[0,100]

When "Curve Family" is selected as "Four Parameter Logistic", the mean response of each dose Y is generated from the formula given by,

$$E(Y \mid D) = \beta + \frac{\delta}{1 + \exp\left(\frac{\theta - D}{\tau}\right)},$$

where D is the dose level. The detailed explanation of these input arguments is provided in Table 11.5.

Table 11.5: Input arguments when "Four Parameter Logistic" is selected in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
β	The minimum or maximum response value depending on whether δ is positive or negative, respectively.	$[-10^6, 10^6]$
δ	The absolute range of expected values of the response.	$[-10^6, -10^{-6}] \cup [10^{-6}, 10^6]$
θ	The value of dose that gives an expected response that is midway between the minimum and maximum values.	$[-10^6, 10^6]$
τ	The parameter is inversely proportional to the slope of the dose response curve at θ.	$[10^{-6}, 10^6]$
Level (#)	The actual dose levels. Dose Levels must be unique values in strictly increasing order.	$[0, 10^6]$
Initial Allocation Ratio	The allocation ratio for the first cohort of patients.	[0,100]

When "Curve Family" is selected as "Linear", the mean response of each dose Y is generated from the formula given by,

$$E(Y \mid D) = E_0 + d * D,$$

where D is the dose level, E_0 represents the y-intercept, i.e., the mean response for placebo, and d represents the slope factor. The detailed explanation of these input arguments is provided in Table 11.6.

Table 11.6: Input arguments when "Linear" is selected in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Intercept	The y-intercept of the Emax model, i.e, the mean response for placebo.	$[-10^6, 10^6]$
Slope	The slope factor.	$[-10^6, -10^{-6}] \cup [10^{-6}, 10^6]$
Level (#)	The actual dose levels. Dose Levels must be unique values in strictly increasing order.	$[0, 10^6]$
Initial Allocation Ratio	The allocation ratio for the first cohort of patients.	[0,100]

When "Curve Family" is selected as "Quadratic", the mean response of each dose Y is generated from the formula given by,

$$E(Y \mid D) = E_0 + B_1 * D + B_2 * D^2,$$

where D is the dose level, E_0 represents the y-intercept, i.e., the mean response for placebo, and B_1 and B_2 represent the linear coefficient and the quadratic coefficient, respectively. The detailed explanation of these input arguments is provided in Table 11.7.

Table 11.7: Input arguments when "Quadratic" is selected in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Intercept	The y-intercept of the Emax model, i.e, the mean response for placebo.	$[-10^6, 10^6]$
Linear Coeff	The linear Coefficient	$[-10^6, 10^6]$
Quadratic Coeff	The quadratic Coefficient	$[-10^6, -10^{-6}] \cup [10^{-6}, 10^6]$
Level (#)	The actual dose levels. Dose Levels must be unique values in strictly increasing order.	$[0, 10^6]$
Initial Allocation Ratio	The allocation ratio for the first cohort of patients.	[0,100]

When "Curve Family" is selected as "Manually Input", the mean response of each dose is input manually.

Table 11.8: Input arguments when "Manually Input" is selected in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
Level (#)	The actual dose levels. Dose Levels must be unique values in strictly increasing order.	$[0, 10^6]$
Mean Response	Mean or average value of the response	$[-10^6, 10^6]$
Initial Allocation Ratio	The allocation ratio for the first cohort of patients.	[0,100]

11.2.2.4 Step 4: Simulation Parameters

In this step, users need to specify the the number of simulations, one-sided Type I error , random seed. Numbers of samples in the burn-in period and steady state samples, and starting values of $\log(ED_{50})$ and Hill need to be specified for the MCMC sampling. See Figure 11.5. A detailed explanation of these input arguments will be provided in Table 11.9.

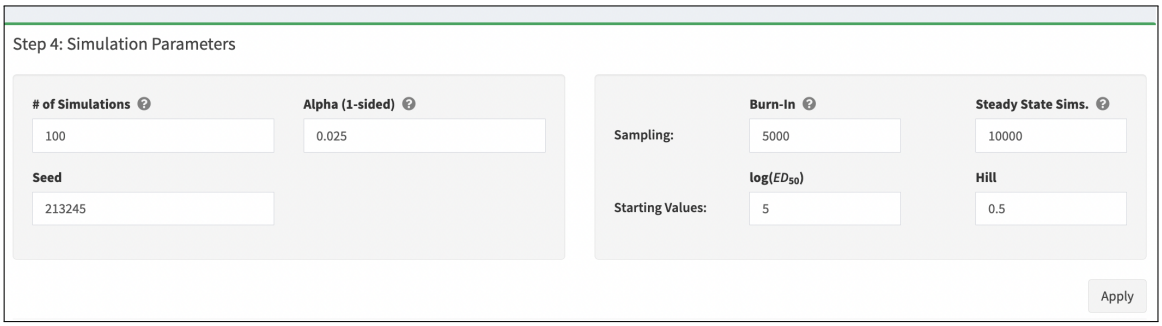


Figure 11.5: Simulation Parameters in the **Dose Ranging Designs – Emax** module.

Click the "Apply" button (Figure 11.5) to confirm the input simulation parameters. The "Apply" button changes to "Edit" and can be clicked again to change design parameters as needed.

Table 11.9: Simulation Parameters in the **Dose Ranging Designs – Emax** module.

Parameters	**Description**	**Range**
# of Simulations	The total number of simulations to be executed.	[1,5000]
Alpha (1-sided)	One-sided Type I error rate	[0.001,0.999]
Seed	Random seed number	$[0, 10^6]$
MCMC Parameters		
Sampling: Burn-In	The number of the initial MCMC iterations that are removed from the final analysis.	[0,10000]
Sampling: Steady State Samples	The number of samples which are collected from MCMC chains of posterior distributions of the parameters in order to calculate the Bayesian estimates. The chains are assumed to have reached the stationary distribution after the burn-in period.	[1000,10000]
Starting Values: log(ED_{50})	The initial value for the logarithm of ED_{50} in the MCMC sampling.	[-13,13]
Starting Values: Hill	The initial value for parameter Hill in the MCMC sampling.	$[10^{-6}, 10^6]$

11.2.3 Simulation Results

The **Outputs** tab is primarily used for viewing the simulation jobs and simulation results, and for downloading simulation results. Simulation results (figures and tables) can be downloaded in CSV format. Hereinafter, we use simulation results and operating characteristics interchangeably.

In the **Outputs** tab, the **History** panel exhibits the progress of all simulations users launched (Figure 11.6). The simulations are displayed in ascending order by the submit time. Once an ongoing simulation is selected, click the "Delete" button to delete the corresponding simulation.

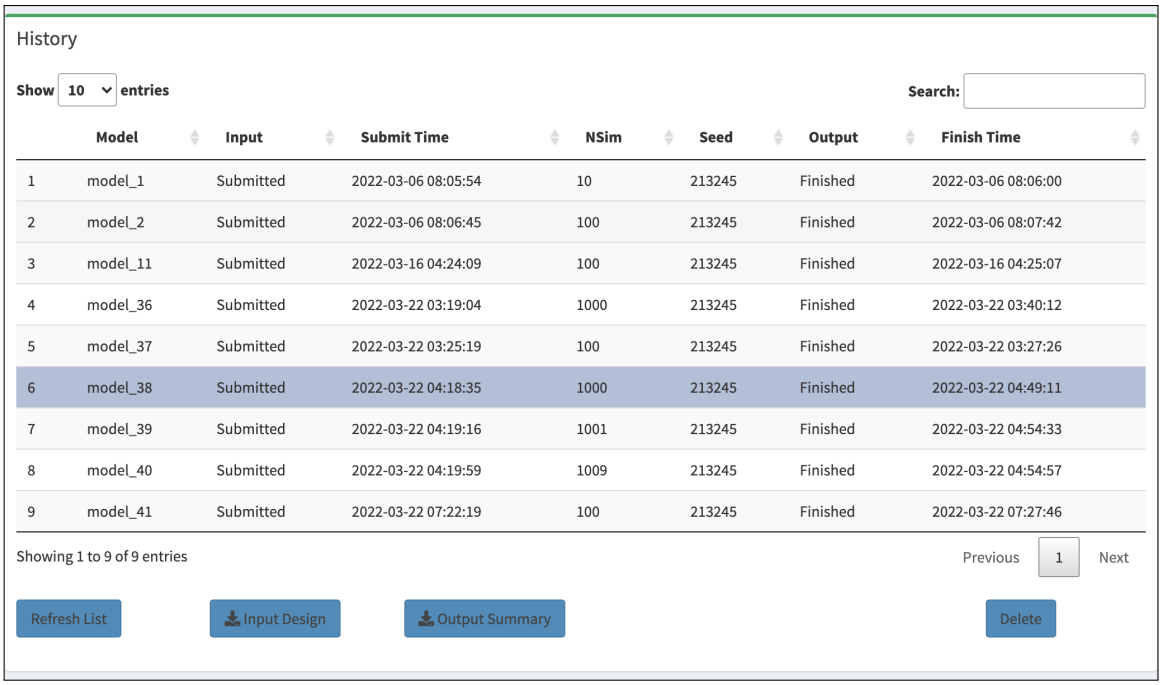


Figure 11.6: Simulation progress in the **Dose Ranging Designs – Emax** module.

Select a finished simulation to show the simulation results (Figure 11.7). The design settings are firstly displayed at the **View Input** panel. Click the **View Output** panel to view the results of simulation. Once a finished simulation is selected, click "Input Design" and "Output Summary" to download a CSV file including simulation settings or simulation results separately. The simulation results are divided into three parts, i.e, **Summary**, **Estimates** and **Target Analysis**.

11.2.3.1 Summary

There are three tables in the **Summary** section (Figure 11.8):

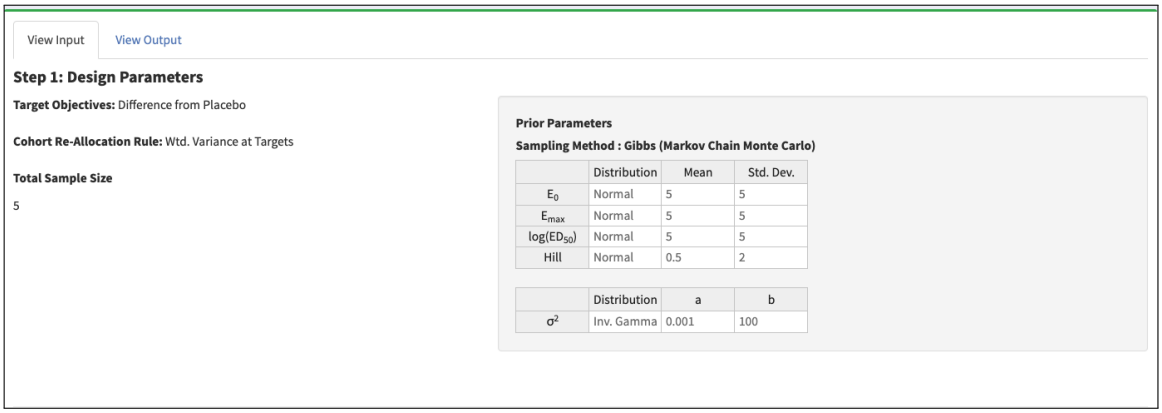


Figure 11.7: View Inputs and Outputs in the **Dose Ranging Designs – Emax** module.

- **Enrollment Specifications**: This table shows the average number of patients assigned to the placebo and the drug in the simulated trials, the average accrual duration and the average duration of the study.
  - **Pbo. Sample Size**: The average number of patients treated at the placebo in the simulated trials.
  - **Drug Sample Size**: The average number of patients treated at the drug in the simulated trials.
  - **Total Sample Size**: The average total number of patients treated at both the placebo and the drug in the simulated trials.
  - **Accr. Dur.**: The average duration of the patient accrual in the simulated trials.
  - **Study Dur.**: The average study duration when the responses of all patients are observed in the simulated trials.
- **Average Sample Size**: This table shows the average number of patients treated at the placebo and all treatment arms of the drug in the simulated trials.
- **Test Statistics**: This table provides the observed value for the associated test Statistics, the estimated power for these tests, and the pooled standard deviation.
  - **tnmax** & **tnmax Power**: t test comparing the mean response of the placebo with that of the dose group which has the most allocation. This test statistic for continuous response

Figure 11.8: Summary in the outputs of the **Dose Ranging Designs – Emax** module.

is defined as follows.

$$t_{n_{\max}} = \frac{\frac{1}{n_{j^*}} \sum_{i=1} y_{ij^*} - \frac{1}{n_0} \sum_{i=1} y_{i0}}{s\left(\sqrt{\frac{1}{n_{j^*}} + \frac{1}{n_0}}\right)},$$

where y_{ij} denotes the response of the ith subject $(i = 1, \cdots, n_j)$ observed at dose j, $j = 0$ denotes the placebo, j^* is the dose index that has the maximum subject allocation, and s denotes the pooled standard deviation defined as,

$$s = \sqrt{\frac{\sum_j (n_j - 1) s_j^2}{\sum_j n_j - J - 1}}. \tag{11.1}$$

Here, $s_j^2 = \frac{1}{n_j - 1} \sum_j (y_{ij} - \bar{y}_j)^2$ and $\bar{y}_j$ represents the average response of subjects at dose j.

For each simulation, significance is determined by comparing $|t_{n_{\max}}|$ with the $1 - \alpha/2$ percentile of the t-distribution with $N - J - 1$ degrees of freedom, and N denotes the total number of subjects in the trial.

– **tslope** & **tslope Power**: test of trend in the case of continuous endpoints. We assume the model is $y_j = \alpha + \beta d_i + \epsilon_i$, where y_i is the response of subject i and d_i is the dose

assigned to subject i, $i = 1, \cdots, N$. N denotes the total number of subjects in the trial. The t-test statistic for the slope to measure the dose response effect is

$$t_{\text{slope}} = \frac{\hat{\beta}}{\text{se}\left(\hat{\beta}\right)},$$

where

$$\hat{\beta} = \frac{\sum_i \left(y_i - \bar{y}\right)\left(d_i - \bar{d}\right)}{\sum_i \left(d_i - \bar{d}\right)^2},$$

$$\text{se}\left(\hat{\beta}\right) = \sqrt{\frac{s^2}{\sum_i \left(d_i - \bar{d}\right)^2}}.$$

Here, $\bar{y}$ is the mean response of total N subjects, $\bar{d} = \frac{\sum_i d_i}{N}$, $s^2 = \frac{1}{N-2}\sum_i \left(y_i - \hat{\alpha} - \hat{\beta} d_i\right)^2$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{d}$.

- **Pooled Standard Deviation**: See (11.1).

11.2.3.2 Estimates

This section displays two tables and one line plot as shown in Figure 11.9.

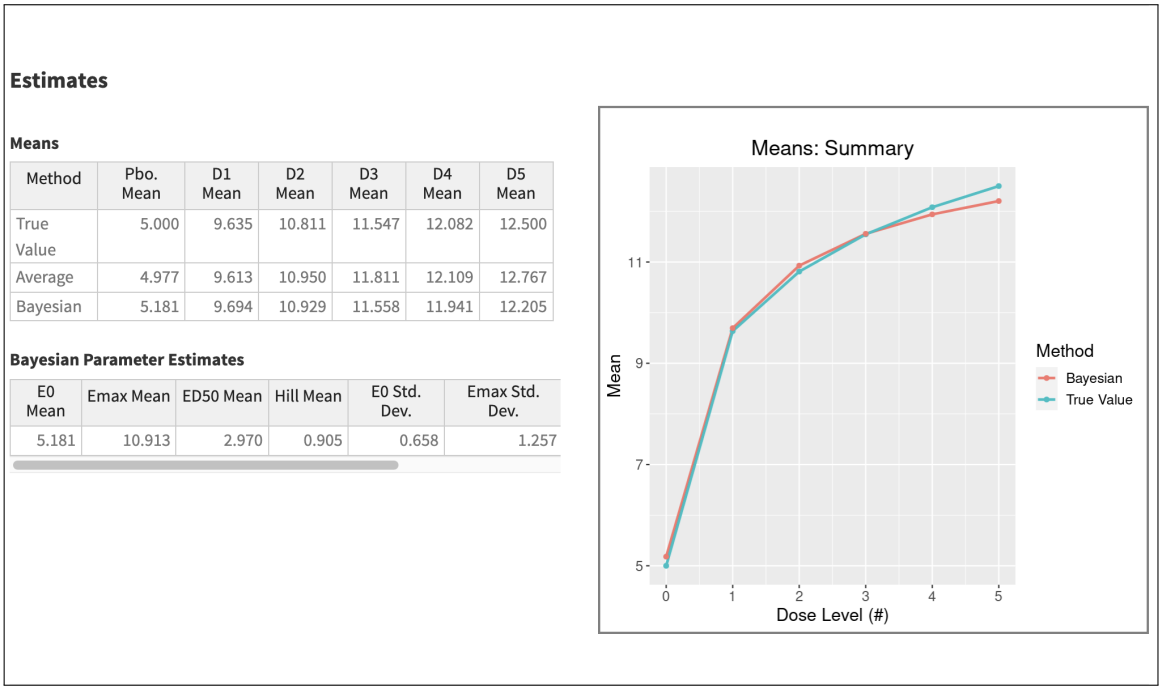
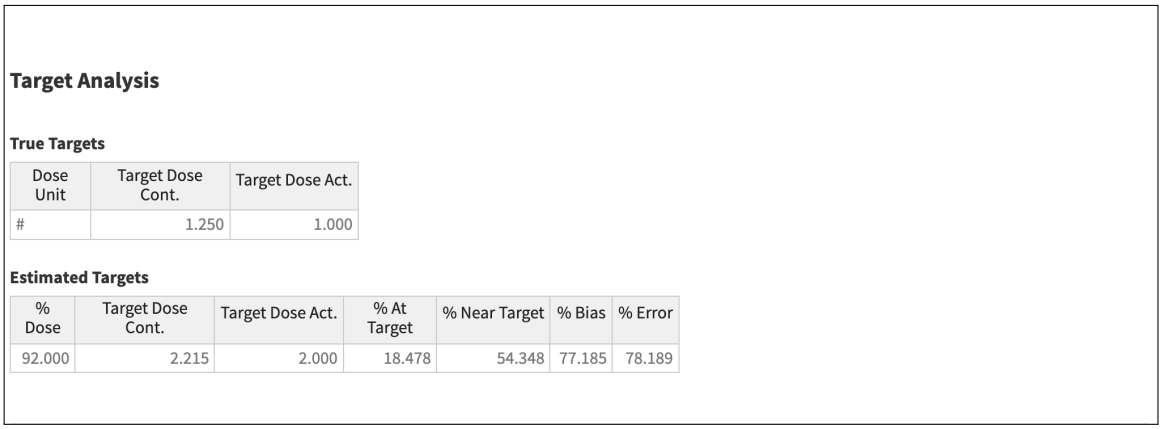


Figure 11.9: Estimates in the outputs of the **Dose Ranging Designs – Emax** module.

- **Means**: This table displays the estimated mean response of each dose group by Bayesian methods.
- **Bayesian Parameter Estimates**: This table displays the estimates of the parameters by Bayesian methods.
- **Mean: Summary**: This line plot displays the estimates of the mean response and the true response value of each dose group.

11.2.3.3 Target Analysis

In the **Target Analysis** section, there are two tables summarizing the information on the true target dose and the estimated target dose from simulation (Figure 11.10).



Target Analysis

True Targets

Dose Unit	Target Dose Cont.	Target Dose Act.
#	1.250	1.000

Estimated Targets

% Dose	Target Dose Cont.	Target Dose Act.	% At Target	% Near Target	% Bias	% Error
92.000	2.215	2.000	18.478	54.348	77.185	78.189

Figure 11.10: Target Analysis in the outputs of the **Dose Ranging Designs – Emax** module.

- **True Targets**: This table contains the information on the true target dose on the continuous scale as well as from the actual doses available. East Bayes calculates the true target dose on the continuous scale based on the dose response curve (§11.2.2.3) and the target objective specified in the design section (§11.2.2.1). If the calculated dose is in the range of studied dose levels, it will be rounded to the nearest dose value. If the target dose level on the continuous scale is out of the range of the studied doses, the cells corresponding to the true target dose on the continuous scale and the actual target dose are left empty, which indicates the target dose is not achievable in the range of studied doses.
  - **Target Dose Cont.**: The true target dose on the continuous scale.

  - **Target Dose Act.**: The true target dose within the studied dose range, obtained by rounding the true target dose on the continuous scale to a dose within the range of studied doses.

- **Estimated Targets**: This table summarizes the information on estimated target doses. If the true target dose on the continuous scale is out of the range of the studied doses, the table is left empty. Explanations of the entries in this table are given below:

  - **% Dose**: The percentage of successfully finding the target dose.
  - **Target Dose Cont.**: The estimated target dose on the continuous scale averaged over those simulations which successfully identify a target dose.
  - **Target Dose Act.**: The estimated target dose within the studied dose range, obtained by rounding the estimated target dose on the continuous scale to a dose within the range of studied doses.
  - **% At Target**: The percentage of times that the true target dose is selected as a target dose.
  - **% Near Target**: The percentage of times that the estimated target dose is adjacent to the true target dose.
  - **% Bias**: The percentage of bias in estimating the target dose.
  - **% Error**: The percentage of mean square error in estimating the target dose. This is a measure of how well the estimated mean response fits the true one.

11.3 Statistical Methods Review

11.3.1 Emax Bayesian Design

11.3.1.1 Probability Model

Suppose there are $(J+1)$ doses including placebo denoted by $d_0, d_1, d_2, \cdots, d_J$ (d_0 denotes placebo, i.e., $d_0 = 0$). The mean response observed at dose D, $D \in \{d_0, d_1, d_2, \cdots, d_J\}$ is given by

$$E(Y \mid D) = E_0 + \frac{E_{\max} D^S}{ED_{50}^S + D^S}, \qquad \text{with } S > 0, ED_{50} > 0. \tag{11.2}$$

Let y_{ij} denote the response of the ith subject $(i = 1, \cdots, n_j)$ observed at dose j. And we assume y_{ij} follows a normal distribution with independent error $\epsilon_{ij} \sim N(0, \sigma^2)$,

$$y_{ij} = E(Y \mid d_j) + \epsilon_{ij}, \qquad \text{for } i = 1, \cdots, n_j; j = 0, 1, \cdots, J.$$

$E(Y \mid D)$ is a monotonically increasing function of D. The minimum of $E(Y \mid D)$ occurs at $D = d_0$, i.e., $D = 0$, where $E(Y \mid D) = E_0$. The upper asymptote is $E_0 + E_{\max}$, so that E_0 is the baseline (minimal) response and $E_{\max}$ is the range of $E(Y \mid D)$ values.

$$E(Y \mid D = ED_{50}) = E_0 + \frac{E_{\max}}{2},$$

so that ED_{50} is the dose with an expected response midway between minimum and maximum responses.

The derivative of $E(Y \mid D)$ at $D = ED_{50}$ is

$$E(Y \mid D = ED_{50}) = E_0 + \frac{E_{\max}}{2}.$$

For fixed $E_{\max}$ and ED_{50}, the derivative is proportional to S. S is often called the **Hill** parameter.

Likelihood Function:

$$L(\boldsymbol{y} \mid E_0, E_{\max}, ED_{50}, S, \sigma) \propto \sigma^{-N} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=0}^{J} n_j \left[\bar{y}_j - \left(E_0 + \frac{E_{\max} d_j^S}{ED_{50}^S + d_j^S} \right) \right]^2 \right\}$$

where N denotes the total number of subjects in the trial, and $\bar{y}_j$ average response of subjects at dose j.

Prior Models: The joint Bayesian model can be written as

$$L(\boldsymbol{y} \mid E_0, E_{\max}, ED_{50}, S, \sigma) p(E_0) p(E_{\max}) p(ED_{50}) p(S) p(\sigma) \tag{11.3}$$

The following independent prior distributions are used,

$$E_0 \sim N(m_{E_0}, s^2_{E_0})$$

$$E_{\max} \sim N(m_{E_{\max}}, s^2_{E_{\max}})$$

$$\log(ED_{50}) \sim N(m_{ED_{50}}, s^2_{ED_{50}})$$

$$S \sim N_+(m_S, s^2_S)$$

$$\sigma^2 \sim InverseGamma(a, b)$$

where $N_+()$ denotes a truncated normal distribution with the left truncation of 0.

Posterior Inference: The model is flexible enough to adequately approximate many different families of parametric monotone dose-response curves. There is no available closed form representation of the joint posterior distribution of the parameters, so an MCMC sampling algorithm is used.

Based on the joint model (11.3), posterior samples for the parameters are obtained using MCMC simulations. Posterior inference will be based on the sampled values from $B + 1$ to T, where is a user-specified burn-in period. The default values are $B = 5000$, $T - B = 10000$.

11.3.1.2 Trial Design

The Emax design allocate the first cohort of patients with user-specified initial allocation ratios (§11.2.2.3). The allocation rule for subsequent cohorts is defined in §11.3.1.3.

11.3.1.3 Target and Allocation Ratio

In East Bayes, users define the actual target value of response as difference from the placebo. Let q index the target dose and $g(d_j)$ denote the posterior mean of the expected response at dose d_j. We use an expected utility u_j of assigning dose d_j to a single future subject,

$$u_j = \text{var}\left(g(d_j)\right) \Pr(d_j = d_q),$$

where the quantities var $(g(d_j))$ and $\Pr(d_j = d_q)$ are estimated using MCMC chains. Allocation is performed on a cohort-by-cohort basis, where all subjects in the current cohort are allocated in a block randomization fashion using rounded values computed from randomization ratios

$$r_j = \frac{u_j}{\sum_{k=1}^{J} u_k}.$$

The placebo does is intentionally left out of this calculation as it is assumed that a pre-specified number of subjects will be allocated to the placebo for each cohort.

Part IV

Group Sequential Methodologies

12. Bayesian Group Sequential Designs

12.1 Introduction

This module provides Bayesian approaches to the monitoring of group sequential designs (GSD). Bayesian approaches offer more flexibility in terms of defining success and futility criteria at interim analysis while also allowing for the inclusion of prior information on the treatment effect. The implementation is based on Gerber et al. (2016). The R package **gsbDesign** in Gerber et al. (2016) is used to evaluate the operating characteristics of the Bayesian group sequential designs. In the module **Bayesian Group Sequential Designs**, we consider clinical trials with interim analyses and provide options to include multiple success and/or futility criteria at each interim and final analysis. Simulations are used to generate operating characteristics for the Bayesian group sequential designs.

In this module, we compare a treatment with a control where the following different endpoints are currently available:

- **Normal**: The effect size is the difference of two means,
- **Binomial**: The effect size is the logarithm of odds ratio,
- **Time-to-event**: The effect size is the logarithm of hazard ratio.

Although the original implementation only supports two-arm Bayesian designs with normal endpoints, and known standard deviations of the effects in the treatment and control arms, we extended the framework to support both binomial and time-to-event endpoints using some transformation and large sample approximation theory.

Group sequential designs are adaptive designs that have one or more interim analyses, where decisions are made so that we could to continue the trial until the end or stop early because of success or futility. One of the main advantages is that if the treatment is not effective, the futile trials can be stopped early. For a sponsor, this could save a lot of time and money. On the other

hand, trials can be also stopped early for success, which may result in faster access to the new treatment. As the major aspect of a group sequential design lies on the decision at each interim analysis on whether to stop or continue the trial, Bayesian methodologies are used in the monitoring of group sequential clinical trials as they are well suited for decision-making process. Bayesian framework also helps to incorporate external information by bringing in informative priors. Gerber et al. (2016) considers Bayesian group sequential designs that incorporate decision making based on the posterior distribution of the difference between the treatment and the control arms.

In this module several stopping criteria based on this posterior distribution could be combined. This combination of multiple criteria goes beyond the scope of the significance testing framework in classical group sequential designs (Gerber et al., 2016). The package **gsbDesign** not only incorporates prior information but also allows the user to specify multiple decision criteria that are provided using two thresholds given below:

- Threshold on the effect size,
- Threshold on the posterior probability.

Evaluating the operating characteristics of the Bayesian group sequential design is essential once stopping criteria have been defined to correspond with clinical decision-making. In order to do that, some true effects of the control and the treatment arms are assumed and the probability of stopping for success or futility, as well as the expected sample size (ESS), are calculated (Gerber et al., 2016).

In this module, currently both non-informative as well as user defined informative priors can be used. In a future release we are planning to use **Meta Analytic Predictive (MAP) Priors** [now as a separate module in East Bayes] in designing group sequential studies.

12.2 User Interface and Tutorial

12.2.1 Normal Endpoints

Upon entering the **Bayesian Group Sequential Designs – Normal Endpoint** page, two main tabs are presented: **Inputs** and **Outputs**. The first tab allows users to conduct simulations, and the second tab allows users to visualize/download simulation results. In the **Inputs** tab, there are two steps (Figure 12.1): 1) Design Settings, and 2) Simulation Settings. Users need to complete **Step 1**, and click the "Apply" button to edit **Step 2**. Upon completion of both two steps, users click the "Simulate" button at the bottom of the page.

After the simulation is launched, the results of simulations will be displayed in the **Outputs** tab. In the **Outputs** tab, users may also click the "Download All" button to download a Rds file including inputs and outputs of this simulation job, or click the "Back to Input" button to reset these settings. Detailed steps of using this module are described in §12.2.1.1–§12.2.1.2.

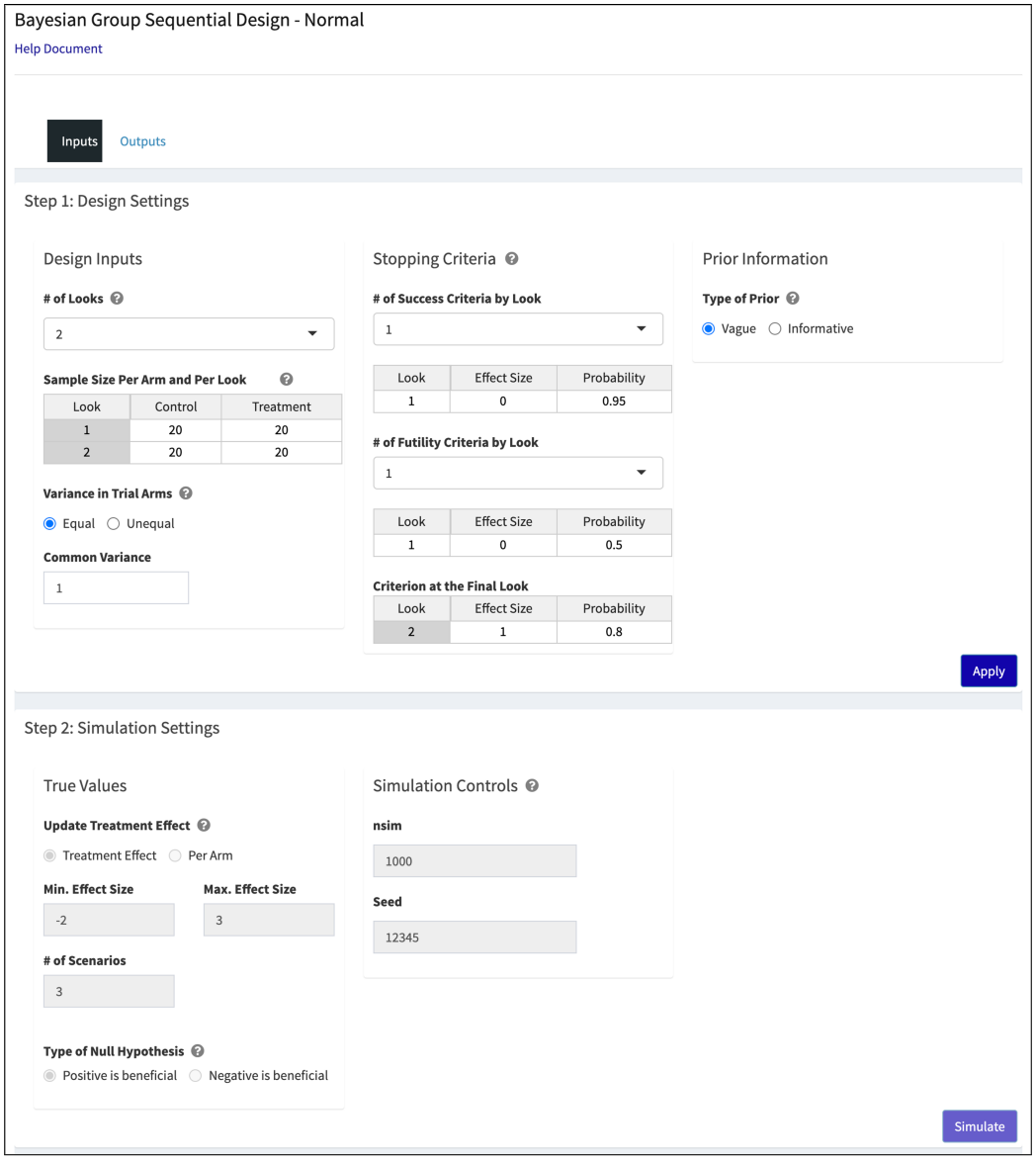


Figure 12.1: Inputs in the **Bayesian Group Sequential Designs – Normal Endpoint**.

12.2.1.1 Inputs

Step 1: Design Settings

In Design Settings, three parts, Design Inputs, Stopping Criteria and Prior Information, need to specify. The detailed explanation of these input arguments is provided in Tables 12.1, 12.2, and 12.3.

Table 12.1: Design Inputs in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Parameters	**Description**
# of Looks	The number of looks. The range is $[1, 5]$. The default value is 2.
Sample Size Per Arm and Per Look	The sample size allocated to the control and treatment arms at each look. The range is $[1, 10000]$. The default values are both $\{20, 20\}$ for the control and treatment arms.
When "Equal" of **Variance in Trial Arms** is selected,	
Common Variance	The common variance of the control and treatment arms. The range is $(0, +\infty)$. The default value is 1.
When "Unequal" of **Variance in Trial Arms** is selected,	
Control	The variance of the control arm. The range is $(0, +\infty)$. The default value is 1.
Treamtent	The variance of the treatment arm. The range is $(0, +\infty)$. The default value is 2.

Click the "Apply" button (Figure 12.2) to confirm the input design settings. The "Apply" button changes to "Edit" and can be clicked again to change design settings as needed.

Table 12.2: Stopping Criteria in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Parameters	**Description**
Success Criteria	
# of Success Criteria by Look	The number of success criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding success criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding success criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in corresponding the success criterion. The range is $(0, 1)$. The default value is 0.95.
Futility Criteria	
# of Futility Criteria by Look	The number of futility criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding futility criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding futility criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in the corresponding futility criterion. The range is $(0, 1)$. The default value is 0.5.
Criterion at the Final Look	
Effect Size	Effect threshold in the criterion at the final look. The range is $(-\infty, +\infty)$. The default value is 1.
Probability	Probability threshold in the criterion at the final look. The range is $(0, 1)$. The default value is 0.8.

Table 12.3: Prior information in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Parameters	**Description**
When "Informative" of **Type of Prior** and **Prior on** "Effect Size" are selected,	
Mean	Prior treatment effect mean. The range is $(-\infty, +\infty)$. The default value is 3.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.
When "Informative" of **Type of Prior** and **Prior on** "Arm-wise" are selected,	
Control Mean	Prior effect mean of the control arm. The range is $(-\infty, +\infty)$. The default value is 3.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Treamtent Mean	Prior effect mean of the treamtent arm. The range is $(-\infty, +\infty)$. The default value is 3.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.

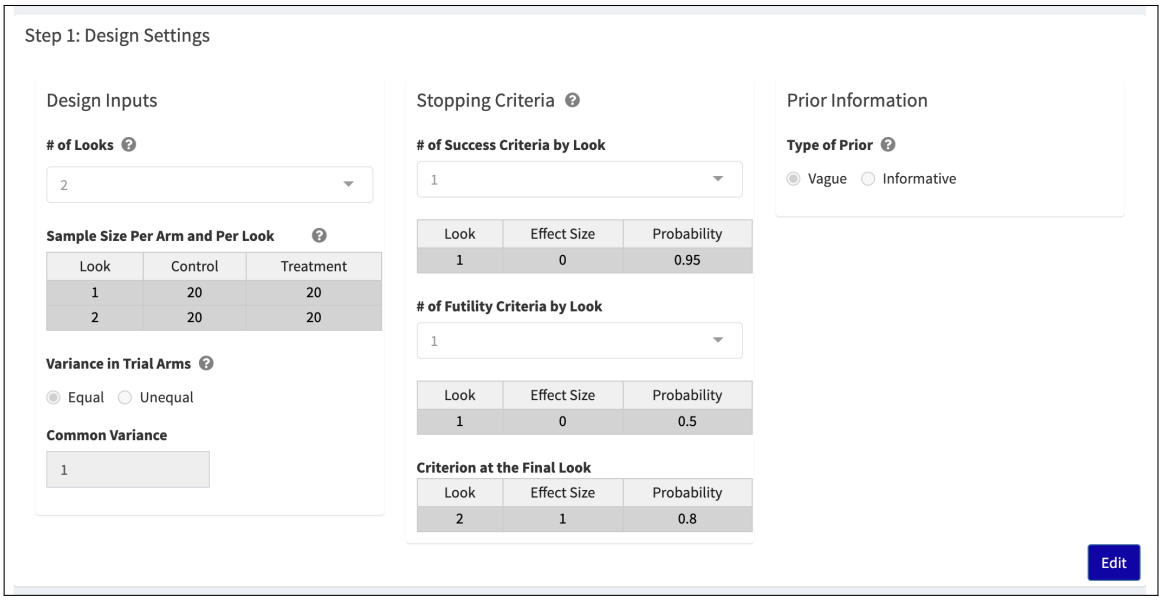


Figure 12.2: Apply design settings in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Step 2: Simulation Settings

In Simulation Settings, two parts, True Values and Simulation Controls, need to specify. For True Values, there are two ways to specify scenarios, setting effect size ("Treatment Effect" is selected, see Figure 12.3) or effect per arm ("Per Arm" is selected, see Figure 12.4).

- "Treatment Effect": Set the minimum effect size (Min. Effect Size), the maximum effect size (Max. Effect Size), and the number of scenarios ($1 \leq$ # of Scenarios ≤ 20). These generated scenarios would be a sequnce with the length of **# of Scenarios** and the identical increments from **Min. Effect Size** to **Max. Effect Size**.

- "Per Arm": First set the number of scenarios, and then the effect per arm need to specify manually.

And there are two sets of radio button, "Positive is beneficial" and "Negative is beneficial". The options are the directions of the hypothese. In the simulation, the success and futility criteria are fixed as specified in Design Settings and we change the direction of scenarios accordingly.

The detailed explanation of these input arguments in Simulation Controls is provided in Table 12.4.

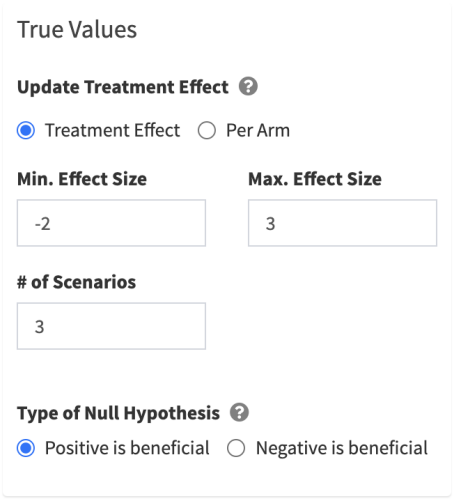


Figure 12.3: True Values when "Treatment Effect" is selected in the **Bayesian Group Sequential Designs – Normal Endpoint**.

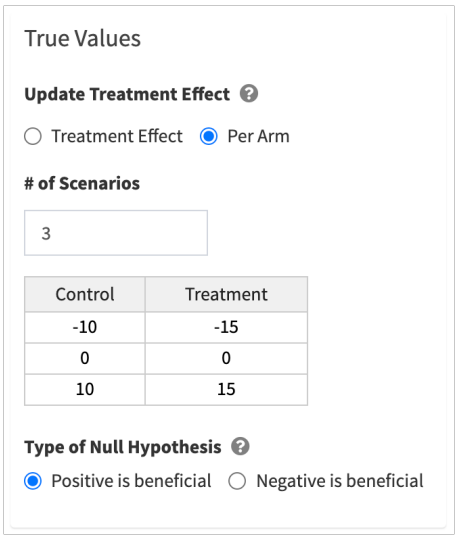


Figure 12.4: True Values when "Per Arm" is selected in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Table 12.4: Simulation Controls in the **Bayesian Group Sequential Designs – Normal Endpoint**.

Parameters	Description
nsim	The number of simulated trials. The maximum number of simulated trials allowed is 10,000. The default value is 1000.
Seed	The random seed of simulation. The default value is 12345.

12.2.1.2 Outputs

In the Outputs tab, users can view the simulation results, and download simulation results with rds format.

Details of the Simulation Results

The simulation results are divided into two parts, A. Expected Sample Size, and B. Simulation Outputs.

A. Expected Sample Size

The table and the figure show the expected sample size for each scenario.

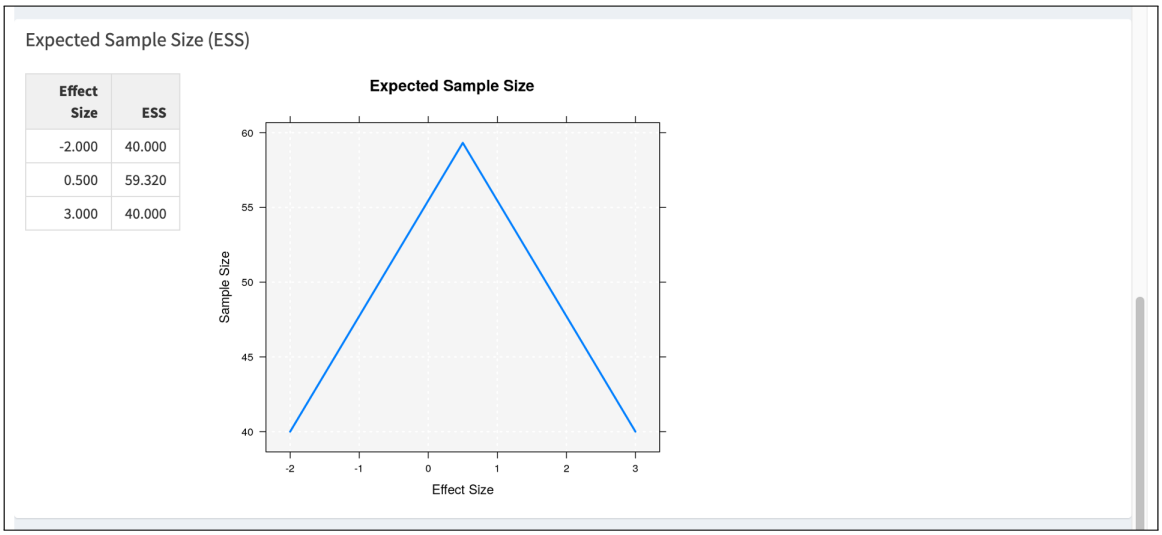


Figure 12.5: Expected Sample Size in the **Bayesian Group Sequential Designs – Normal Endpoint**.

B. Simulation Outputs

These tables and figures show the operating characteristics under each scenario. For **Operating**

Characteristics,

- Prob. Success: probability of early declaring efficacy at each look.
- Prob. Futility: probability of early declaring futility at each look.

For **Cumulative Operating Characteristics**,

- Prob. Success: cumulative probability of early declaring efficacy at each look.
- Prob. Futility: cumulative probability of early declaring futility at each look.
- Prob. Indeterminate: cumulative probability of indeterminate decision at each look.

Download Simulation Results

There is a "Download All" button found at the left bottom in the **Output** tab. Click it to download a Rds file, which includes all inputs and outputs of the launched simulation job. Users may also load it using **readRDS** function and with **lattice** package loaded in R.

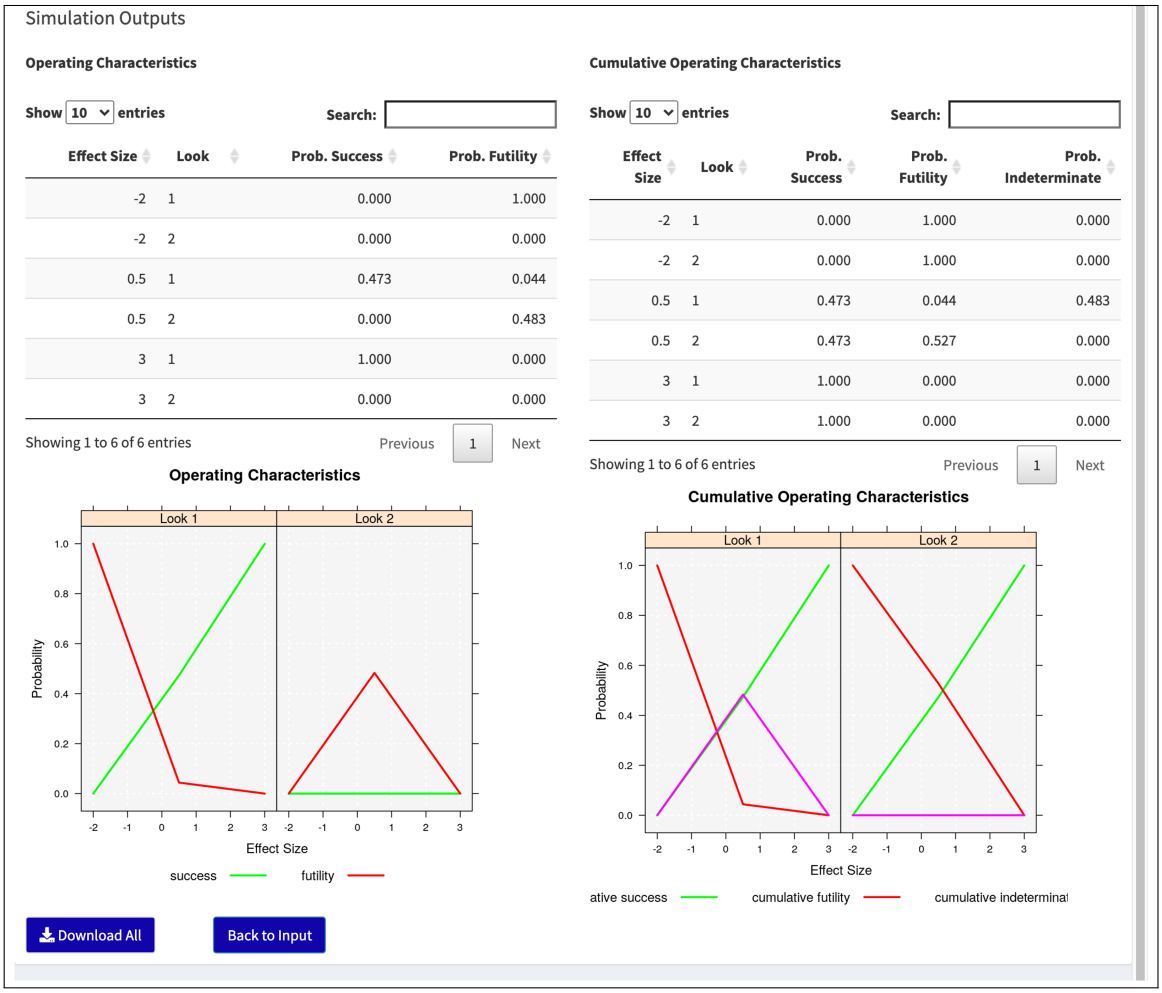


Figure 12.6: Simulation Outputs in the **Bayesian Group Sequential Designs – Normal Endpoint**.

12.2.2 Binomial Endpoints

Upon entering the **Bayesian Group Sequential Designs – Binomial Endpoint** page, two main tabs are presented: **Inputs** and **Outputs**. The first tab allows users to conduct simulations, and the second tab allows users to visualize/download simulation results. In the **Inputs** tab, there are two steps (Figure 12.7): 1) Design Settings, and 2) Simulation Settings. Users need to complete **Step 1**, and click the "Apply" button to edit **Step 2**. Upon completion of both two steps, users click the "Simulate" button at the bottom of the page.

After the simulation is launched, the results of simulations will be displayed in the **Outputs** tab. In the **Outputs** tab, users may also click the "Download All" button to download a Rds file including inputs and outputs of this simulation job, or click the "Back to Input" button to reset these settings. Detailed steps of using this module are described in §12.2.2.1–§12.2.2.2.

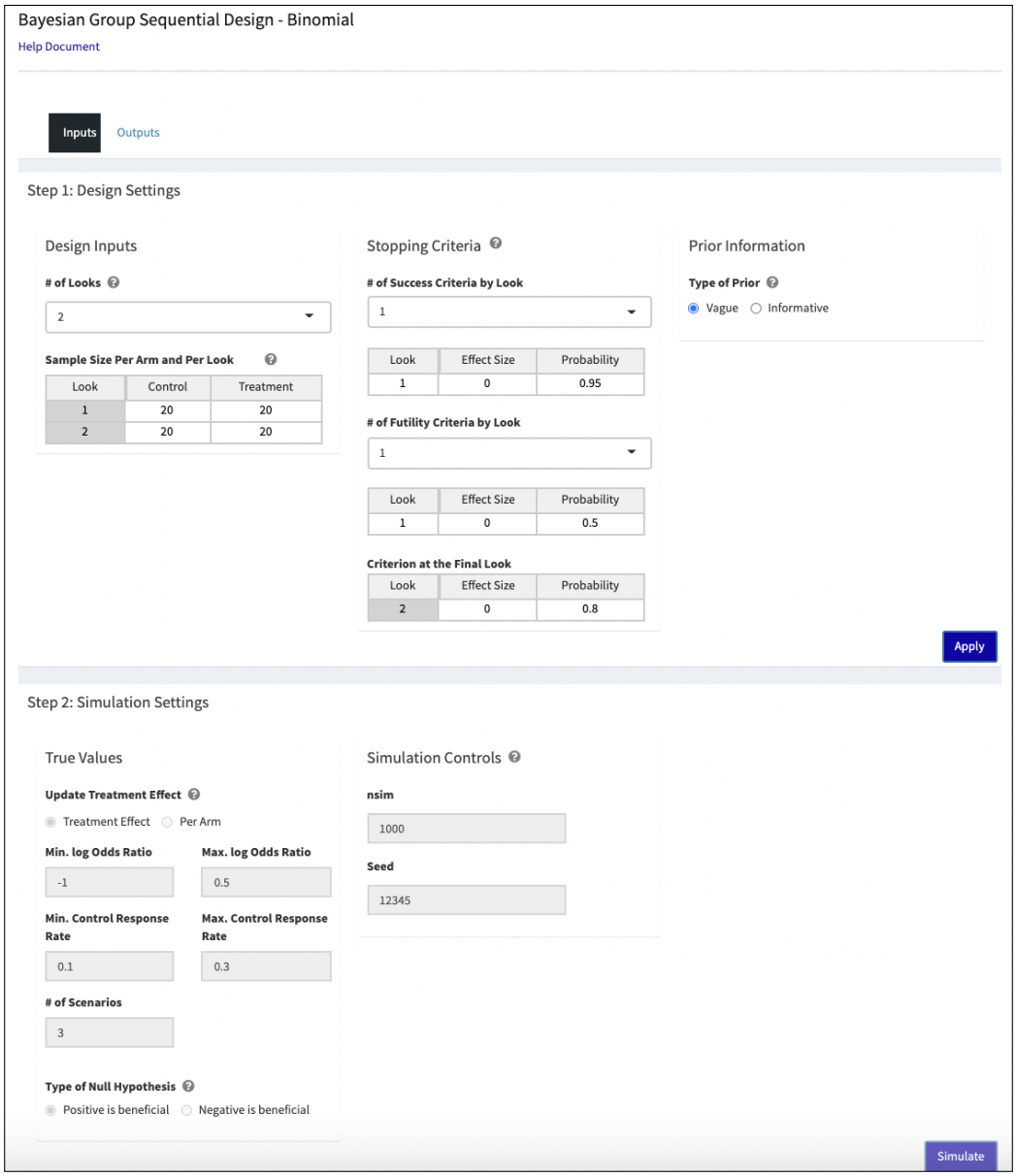


Figure 12.7: Inputs in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

12.2.2.1 Inputs

Step 1: Design Settings

In Design Settings, three parts, Design Inputs, Stopping Criteria and Prior Information, need to specify. The detailed explanation of these input arguments is provided in Tables 12.5, 12.6, and 12.7.

Table 12.5: Design Inputs in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Parameters	Description
# of Looks	The number of looks. The range is $[1, 5]$. The default value is 2.
Sample Size Per Arm and Per Look	The sample size allocated to the control and treatment arms at each look. The range is $[1, 10000]$. The default values are both $\{20, 20\}$ for the control and treatment arms.

Click the "Apply" button (Figure 12.8) to confirm the input design settings. The "Apply" button changes to "Edit" and can be clicked again to change design settings as needed.

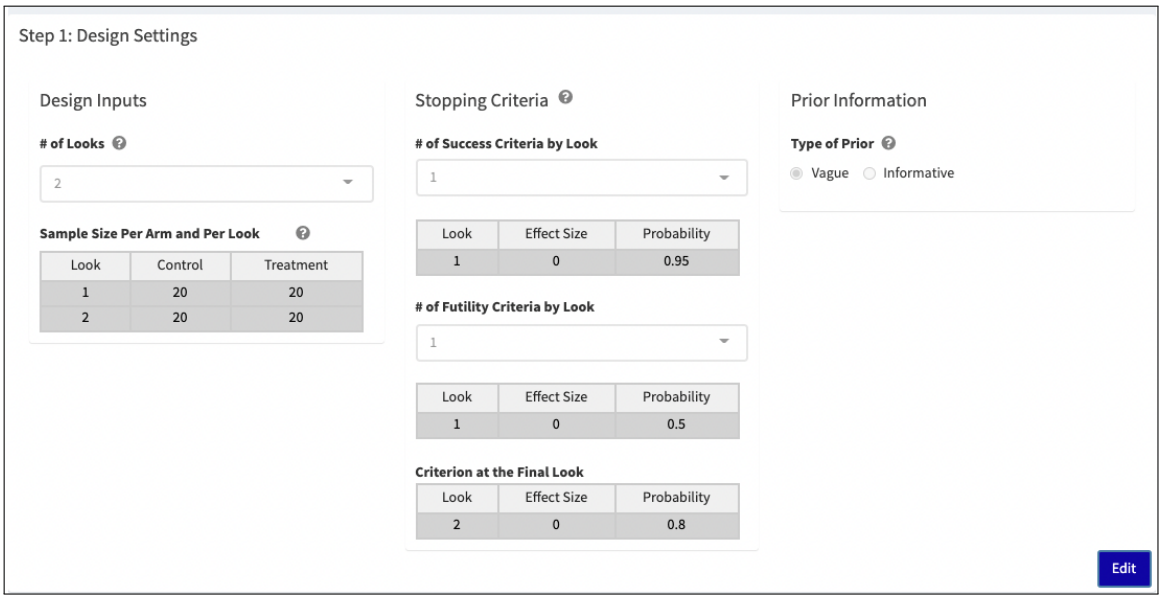


Figure 12.8: Apply design settings in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Table 12.6: Stopping Criteria in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Parameters	**Description**
Success Criteria	
# of Success Criteria by Look	The number of success criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding success criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding success criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in corresponding the success criterion. The range is $(0, 1)$. The default value is 0.95.
Futility Criteria	
# of Futility Criteria by Look	The number of futility criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding futility criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding futility criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in the corresponding futility criterion. The range is $(0, 1)$. The default value is 0.5.
Criterion at the Final Look	
Effect Size	Effect threshold in the criterion at the final look. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in the criterion at the final look. The range is $(0, 1)$. The default value is 0.8.

Table 12.7: Prior information in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Parameters	**Description**
When "Informative" of **Type of Prior** and **Prior on** "Effect Size" are selected,	
Log Odds Ratio Mean	Prior log odds ratio mean. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.
When "Informative" of **Type of Prior** and **Prior on** "Arm-wise" are selected,	
Control Log Odds Mean	Prior log odds mean of the control arm. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Treatment Log Odds Mean	Prior log odds mean of the treamtent arm. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.

Step 2: Simulation Settings

In Simulation Settings, two parts, True Values and Simulation Controls, need to specify. For True Values, there are two ways to specify scenarios, setting effect size ("Treatment Effect" is selected, see Figure 12.9) or effect per arm ("Per Arm" is selected, see Figure 12.10).

- "Treatment Effect": Set the minimum log odds ratio (Min. log Odds Ratio), the maximum log odds ratio (Max. log Odds Ratio), and the number of scenarios ($1 \leq$ # of Scenarios ≤ 20). These generated scenarios would be a sequnce with the length of **# of Scenarios** and the identical increments from **Min. log Odds Ratio** to **Max. log Odds Ratio**.

- "Per Arm": First set the number of scenarios, and then the effect per arm need to specify manually.

And there are two sets of radio button, "Positive is beneficial" and "Negative is beneficial". The options are the directions of the hypothese. In the simulation, the success and futility criteria are fixed as specified in Design Settings and we change the direction of scenarios accordingly.

The detailed explanation of these input arguments in Simulation Controls is provided in Table 12.8.

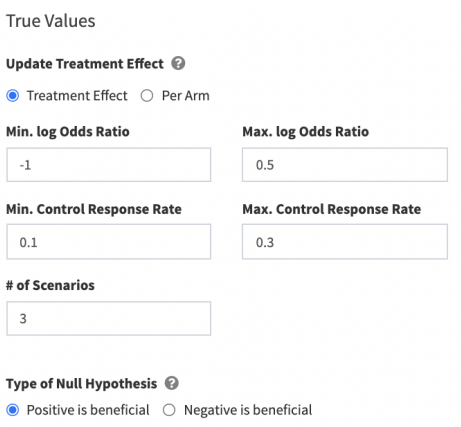


Figure 12.9: True Values when "Treatment Effect" is selected in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

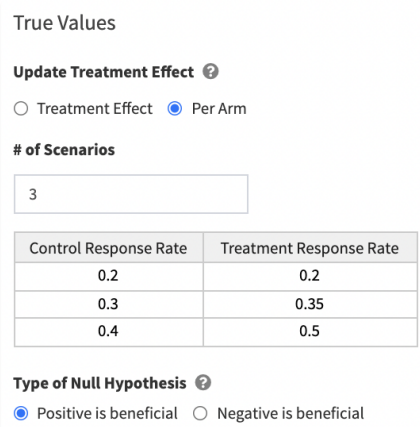


Figure 12.10: True Values when "Per Arm" is selected in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Table 12.8: Simulation Controls in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Parameters	Description
nsim	The number of simulated trials. The maximum number of simulated trials allowed is 10,000. The default value is 1000.
Seed	The random seed of simulation. The default value is 12345.

12.2.2.2 Outputs

In the Outputs tab, users can view the simulation results, and download simulation results with rds format.

Details of the Simulation Results

The simulation results are divided into two parts, A. Expected Sample Size, and B. Simulation Outputs.

A. Expected Sample Size

The table and the figure show the expected sample size for each scenario.

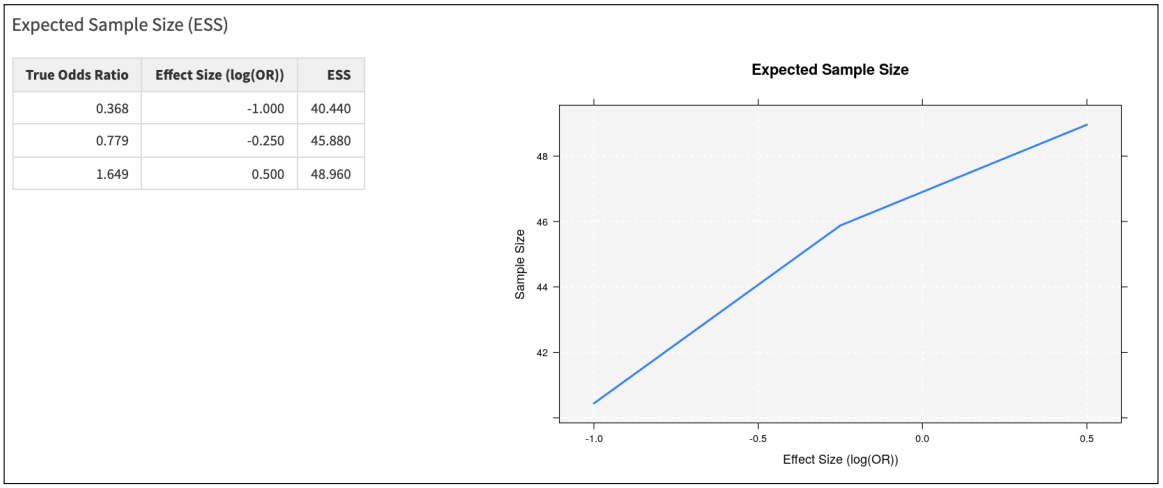


Figure 12.11: Expected Sample Size in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

B. Simulation Outputs

These tables and figures show the operating characteristics under each scenario. For **Operating**

Characteristics,

- Prob. Success: probability of early declaring efficacy at each look.
- Prob. Futility: probability of early declaring futility at each look.

For **Cumulative Operating Characteristics**,

- Prob. Success: cumulative probability of early declaring efficacy at each look.
- Prob. Futility: cumulative probability of early declaring futility at each look.
- Prob. Indeterminate: cumulative probability of indeterminate decision at each look.

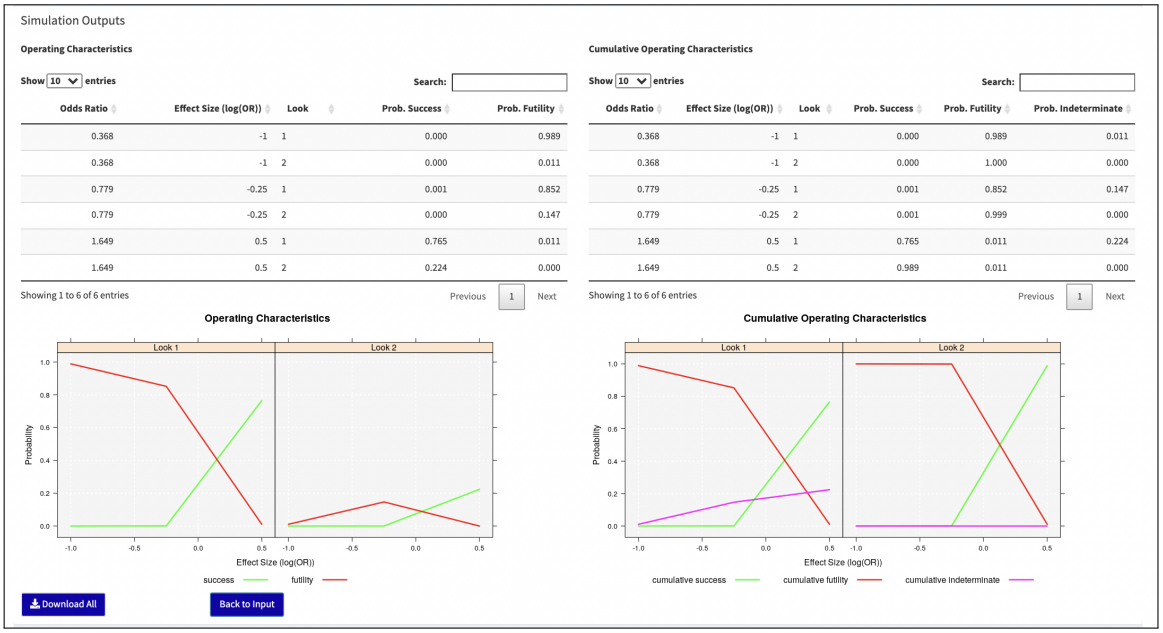


Figure 12.12: Simulation Outputs in the **Bayesian Group Sequential Designs – Binomial Endpoint**.

Download Simulation Results

There is a "Download All" button found at the left bottom in the **Output** tab. Click it to download a Rds file, which includes all inputs and outputs of the launched simulation job. Users may also load it using **readRDS** function and with **lattice** package loaded in R.

12.2.3 Time-to-Event Endpoints

Upon entering the **Bayesian Group Sequential Designs – Time to Event Endpoint** page, two main tabs are presented: **Inputs** and **Outputs**. The first tab allows users to conduct simulations, and the second tab allows users to visualize/download simulation results. In the **Inputs** tab, there are two steps (Figure 12.13): 1) Design Settings, and 2) Simulation Settings. Users need to complete **Step 1**, and click the "Apply" button to edit **Step 2**. Upon completion of both two steps, users click the "Simulate" button at the bottom of the page.

After the simulation is launched, the results of simulations will be displayed in the **Outputs** tab. In the **Outputs** tab, users may also click the "Download All" button to download a Rds file including inputs and outputs of this simulation job, or click the "Back to Input" button to reset these settings. Detailed steps of using this module are described in §12.2.3.1–§12.2.3.2.

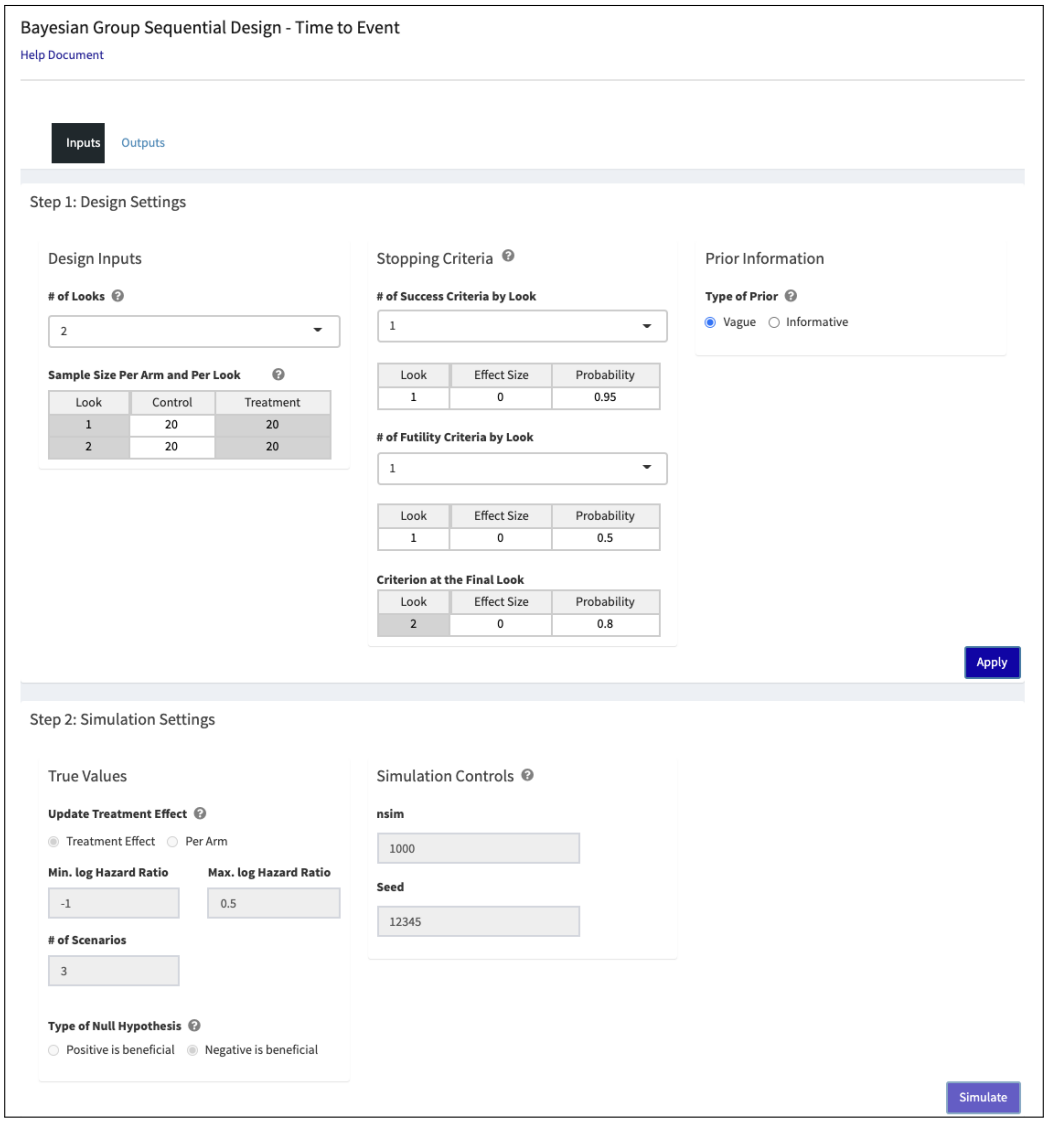


Figure 12.13: Inputs in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

12.2.3.1 Inputs

Step 1: Design Settings

In Design Settings, three parts, Design Inputs, Stopping Criteria and Prior Information, need to specify. The detailed explanation of these input arguments is provided in Tables 12.9, 12.10, and 12.11.

Table 12.9: Design Inputs in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Parameters	Description
# of Looks	The number of looks. The range is $[1, 5]$. The default value is 2.
Sample Size Per Arm and Per Look (Editable only for the control arm)	The sample size allocated to the control arm at each look. By default, the sample sizes allocated to the control and treatment arms are the same at each look. The range is $[1, 10000]$. The default values are $\{20, 20\}$ for the control arm.

Click the "Apply" button (Figure 12.14) to confirm the input design settings. The "Apply" button changes to "Edit" and can be clicked again to change design settings as needed.

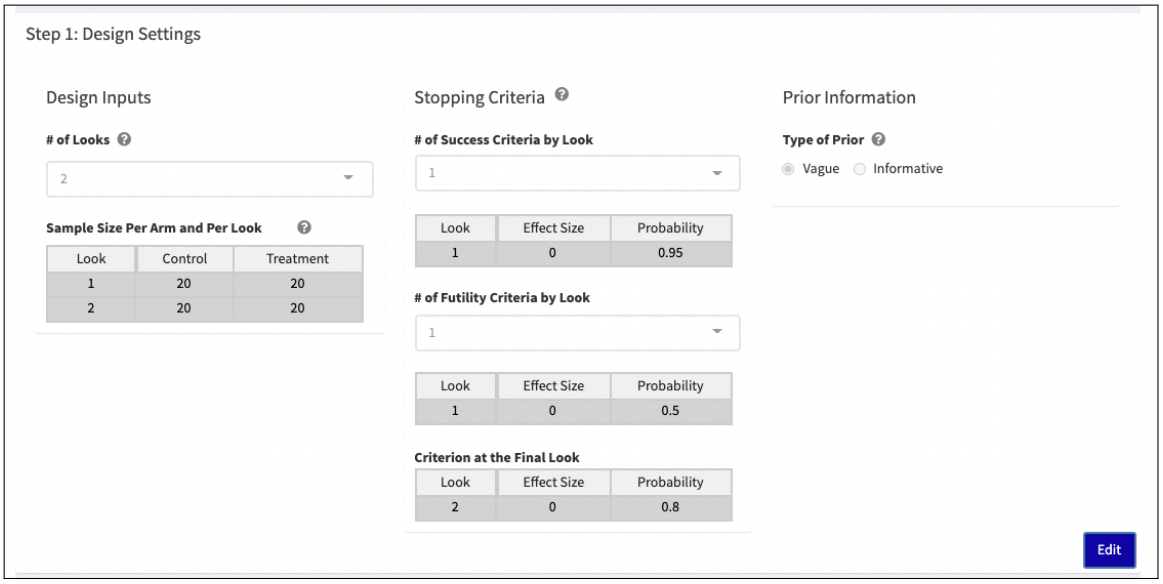


Figure 12.14: Apply design settings in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Table 12.10: Stopping Criteria in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Parameters	**Description**
Success Criteria	
# of Success Criteria by Look	The number of success criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding success criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding success criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in corresponding the success criterion. The range is $(0, 1)$. The default value is 0.95.
Futility Criteria	
# of Futility Criteria by Look	The number of futility criteria at interim looks. The range is [0, # of Looks - 1]. The default value is 1.
Look	The interim look number of the corresponding futility criterion. The range is [1, # of Looks - 1]. The default value is 1.
Effect Size	Effect threshold in the corresponding futility criterion. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in the corresponding futility criterion. The range is $(0, 1)$. The default value is 0.5.
Criterion at the Final Look	
Effect Size	Effect threshold in the criterion at the final look. The range is $(-\infty, +\infty)$. The default value is 0.
Probability	Probability threshold in the criterion at the final look. The range is $(0, 1)$. The default value is 0.8.

Table 12.11: Prior information in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Parameters	**Description**
When "Informative" of **Type of Prior** and **Prior on** "Effect Size" are selected,	
Log Hazard Ratio Mean	Prior log hazard ratio mean. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.
When "Informative" of **Type of Prior** and **Prior on** "Arm-wise" are selected,	
Control Log Hazard Rate Mean	Prior log hazard rate mean of the control arm. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Control	Effective sample size on the control arm. The range is $(0, +\infty)$. The default value is 2.
Treatment Log Hazard Rate Mean	Prior log hazard rate mean of the treamtent arm. The range is $(-\infty, +\infty)$. The default value is 1.
Effective Sample Size on Treamtent	Effective sample size on the treatment arm. The range is $(0, +\infty)$. The default value is 1.

Step 2: Simulation Settings

In Simulation Settings, two parts, True Values and Simulation Controls, need to specify. For True Values, there are two ways to specify scenarios, setting effect size ("Treatment Effect" is selected, see Figure 12.15) or effect per arm ("Per Arm" is selected, see Figure 12.16).

- "Treatment Effect": Set the minimum log hazard ratio (Min. log Hazard Ratio), the maximum log hazard ratio (Max. log Hazard Ratio), and the number of scenarios ($1 \leq$ # of Scenarios ≤ 20). These generated scenarios would be a sequnce with the length of **# of Scenarios** and the identical increments from **Min. log Hazard Ratio** to **Max. log Hazard Ratio**.

- "Per Arm": First set the number of scenarios, and then the effect per arm need to specify manually.

And there are two sets of radio button, "Positive is beneficial" and "Negative is beneficial". The options are the directions of the hypothese. In the simulation, the success and futility criteria are fixed as specified in Design Settings and we change the direction of scenarios accordingly.

The detailed explanation of these input arguments in Simulation Controls is provided in Table 12.12.

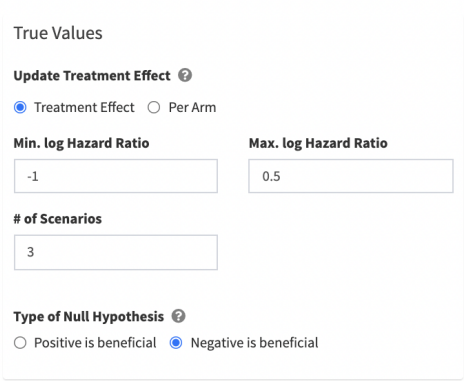


Figure 12.15: True Values when "Treatment Effect" is selected in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

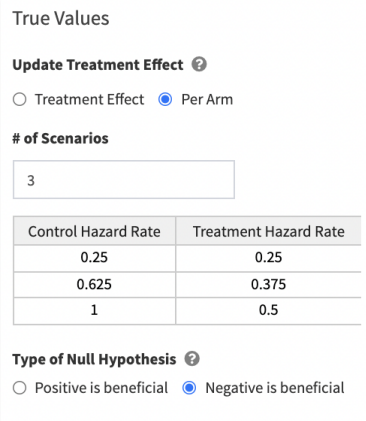


Figure 12.16: True Values when "Per Arm" is selected in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Table 12.12: Simulation Controls in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Parameters	Description
nsim	The number of simulated trials. The maximum number of simulated trials allowed is 10,000. The default value is 1000.
Seed	The random seed of simulation. The default value is 12345.

12.2.3.2 Outputs

In the Outputs tab, users can view the simulation results, and download simulation results with rds format.

Details of the Simulation Results

The simulation results are divided into two parts, A. Expected Sample Size, and B. Simulation Outputs.

A. Expected Sample Size

The table and the figure show the expected sample size for each scenario.

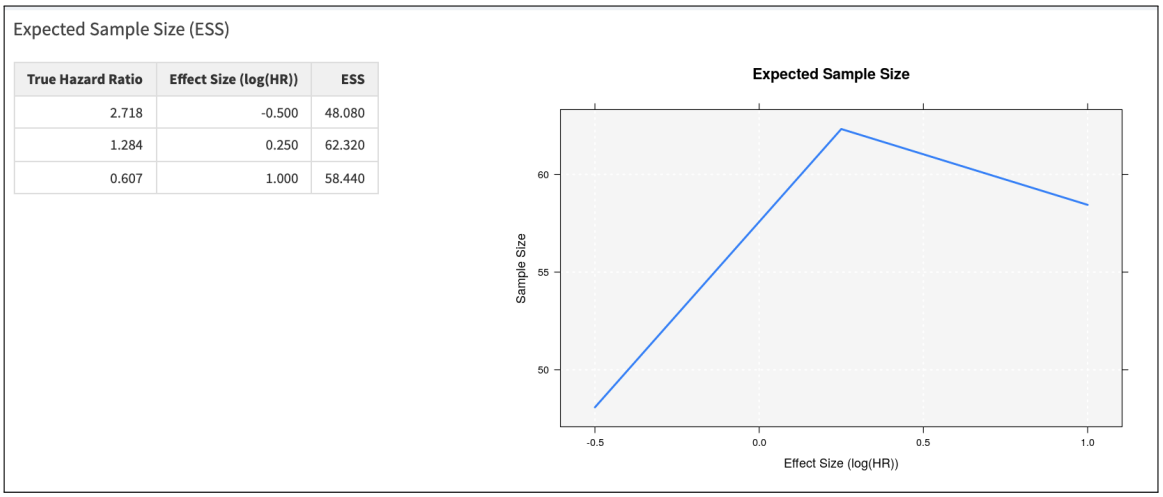


Figure 12.17: Expected Sample Size in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

B. Simulation Outputs

These tables and figures show the operating characteristics under each scenario. For **Operating**

Characteristics,

- Prob. Success: probability of early declaring efficacy at each look.
- Prob. Futility: probability of early declaring futility at each look.

For **Cumulative Operating Characteristics**,

- Prob. Success: cumulative probability of early declaring efficacy at each look.
- Prob. Futility: cumulative probability of early declaring futility at each look.
- Prob. Indeterminate: cumulative probability of indeterminate decision at each look.

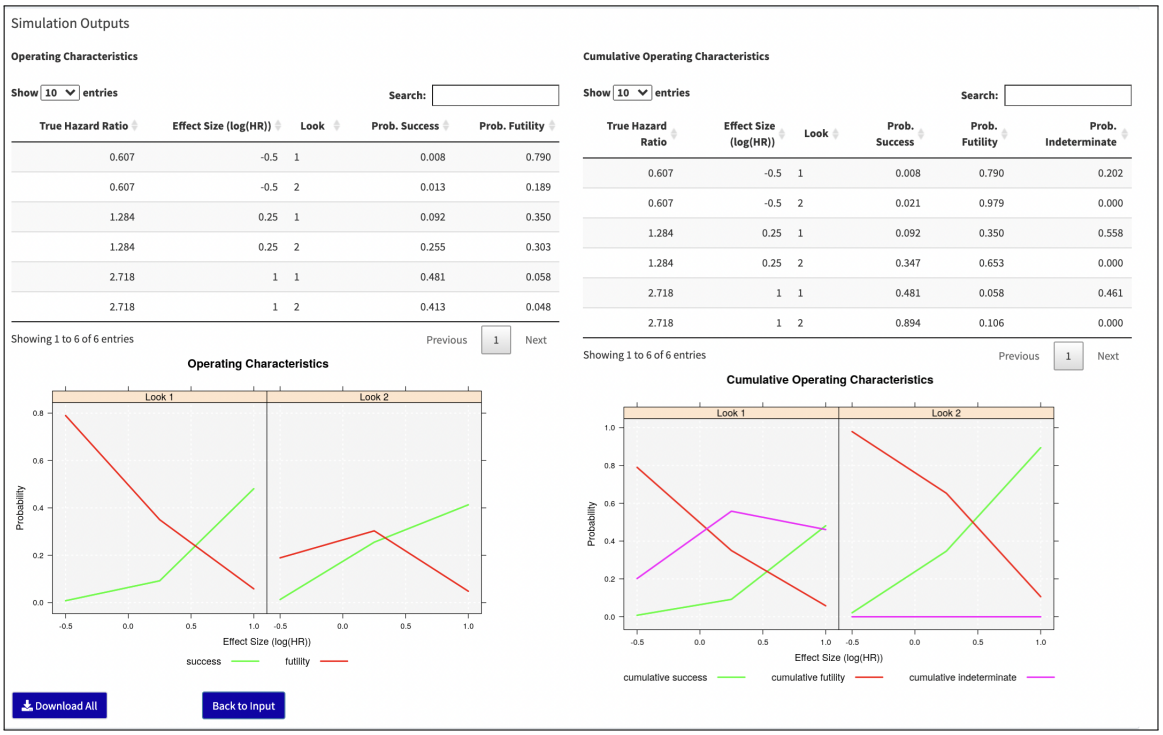


Figure 12.18: Simulation Outputs in the **Bayesian Group Sequential Designs – Time to Event Endpoint**.

Download Simulation Results

There is a "Download All" button found at the left bottom in the **Output** tab. Click it to download a Rds file, which includes all inputs and outputs of the launched simulation job. Users may also load it using **readRDS** function and with **lattice** package loaded in R.

12.3 Statistical Methods Review

12.3.1 Normal Endpoints

Here first we discuss a general setup of the Bayesian design for adaptive two-arm clinical trials with zero, one, or more interim analyses. At each analysis, the success and futility criteria are evaluated to decide if the trial should be stopped. The model for normal endpoints assumes continuous outcome data with error terms that are also normally distributed. We use $N(\cdot,\cdot)$ and $N_P(\cdot,\cdot)$ to denote normal distribution parametrized by variance and precision, respectively. Because sometimes parametrizing the normal distributions in terms of precision as opposed to variance makes the analytical expressions simpler.

Based on the posterior distribution of the treatment effect the sopping criteria are constructed. This treatment effect denotes the improvement of the treatment over the control and denoted by δ. Although in principle an arbitrary number of success and futility criteria could be specified at each analysis, we choose to restrict the number of maximum criteria from a practical implementation point of view.

We follow the same formulation as given in Gerber et al. (2016) to specify the success and the futility criteria. The criteria is given by:

$$P(\delta > s|\text{data}) \geq p \tag{12.1}$$

$$P(\delta < f|\text{data}) \geq q, \tag{12.2}$$

respectively. Note that, s and f are user-specified thresholds for δ. Also, p and q are user specified probability thresholds for success and futility, respectively.

Prior information in terms of prior distribution could be put either on treatment effect (δ) or on the effect in the control arm (μ_1) and the treatment arm (μ_2) individually. Currently, we only work with the prior that are distributed normally. In order to denote the variance for the control arm and the treatment we use $\frac{\sigma_1^2}{n_{10}}$ and $\frac{\sigma_2^2}{n_{20}}$ where

$\sigma_j, j = 1, 2:$ standard deviation for the control arm $(j = 1)$and the treatment arm $(j = 2)$

$n_{j0}, j = 1, 2:$ quantification of prior information per arm.

Other parameters of the design can be specified as

$I:$	the number of interim analyses including the final analysis
$n_{ji}, i = 1, \cdots, I:$	the number of patients per arm and per interim analysis. Hence, the total number of patients in arm j at interim i is $N_{ji} = \sum_{k=1}^{i} n_{jk}$
s_{ir} and $p_{ir}, i = 1, \cdots, I, r = 1, \cdots, M:$	effect and probability thresholds for each success criterion at each interim analysis, respectively
f_{ir} and $q_{ir}, i = 1, \cdots, I, r = 1, \cdots, M:$	effect and probability thresholds for each futility criterion at each interim analysis, respectively.

Note that, for our implementation we decide to have $I = 5$ and $M = I - 1 = 4$. All criteria have to be fulfilled to stop for futility or success at an interim or at the final analysis. If the trial does not stop for success or for futility, it continues until the end.

Operating Characteristics (OC): Simulation of any clinical trial model can be broken into a scenario and a design. Different true value of δ gives rise to different scenarios and a set of parameters - sample sizes, stopping criteria, prior specification create the design. The important operating characteristics are the probabilities of success and futility at each interim analysis, and the expected sample size. In this module, we report those specific characteristics as the primary output. First, we simulate a large number of trials given some true treatment effects of interest. Accuracy depends on the number of trials. At each interim analysis, we compute the posterior distribution of the treatment effect given the data and evaluate the stopping criteria based on the trials those are not stopped at the previous interim analysis. Note that, while simulating, the prior could be specified in two ways - (1) specified on treatment effect and (2) specified on both arms individually (Gerber et al., 2016).

12.3.1.1 Prior on treatment effect δ

Let us denote $Y_{ijk} \sim N(\mu_j, \sigma_j^2)$ for the observations for treatment $j = 1, 2$ at interim $i = 1, ..., I$ for subject $k = 1, ..., n_{ji}$.

The combined treatment effect at interim i is $D_i = \bar{Y}_{2i} - \bar{Y}_{1i}$ with $\bar{Y}_{ji} = (\sum_{l=1}^{i} \sum_{k=1}^{n_l} Y_{jlk})/N_{ji}$ and $N_{ji} = n_{j1} + \cdots + n_{ji}$. Thus,

$$D_i \sim N(\delta, \sigma_1^2/N_{1i} + \sigma_2^2/N_{2i}) \text{ where } \delta = \mu_2 - \mu_1.$$

Let us also assume that the prior information is available for the treatment effect δ as

$$\sim N(\alpha_0, \sigma_1^2/n_{10} + \sigma_2^2/n_{20}).$$

This specification of prior reflects the information on the treatment effect as if n_{10} and n_{20} patients had been treated with the control and the test treatment, respectively.

The prior precision is denoted by $\beta_0 = n_{10}n_{20}/(n_{10}\sigma_2^2 + n_{20}\sigma_1^2)$ and the precision of the observed treatment effect at interim i is denoted by $B_i = N_{1i}N_{2i}/(N_{1i}\sigma_2^2 + N_{2i}\sigma_1^2)$.

Using Bayes' theorem, the posterior is proportional to the likelihood times the prior. Here, the likelihood and the prior are $D_i|\delta \sim N_P(\delta, B_i)$ and $\delta \sim N_P(\alpha_0, \beta_0)$, respectively. Because of conjugacy, we get a normally distributed posterior here. In other words the posterior expectation is a weighted average of the prior mean and the sample mean, and the posterior precision is the sum of the prior and sample precisions. Thus, a sequential update yields the normal posterior distribution at interim i with expectation $\alpha_i = w_i\alpha_0 + (1 - w_i)D_i$ with $w_i = \beta_0/\beta_i$ and precision $\beta_i = \beta_0 + B_i$. To characterize the distribution of D_i, we use the fact that the sequence $Z_1 = D_1\sqrt{B_1}, \cdots, Z_I = D_I\sqrt{B_I}$ is multivariate normal distribution with $E[Z_i] = \delta\sqrt{B_i}$, for $i = 1, \cdots, I$ and $\text{COV}[Z_i, Z_j] = \sqrt{B_i/B_j}$, $1 \leq i \leq j \leq I$ (Gerber et al., 2016).

Simulation When evaluating the operating characteristics of a design, a range of true treatment effect that constitutes the scenarios, denoted by $\delta_u, u = 1, \cdots, U$ is considered. A complete set of interim treatment effects, D_i for $i = 1, \cdots, I$, is generated for a large number of trials, denoted by T_0 and each of the scenarios. We use the canonical joint distribution for δ in order to simulate the D_i. At each interim analysis, the posterior distribution is updated and the decision criteria are applied.

12.3.1.2 Prior information on control and treatment arms

We consider the combined arm-wise treatment response at interim i and it is given by

$$\bar{Y}_{ji} = \sum_{l=1}^{i}\sum_{k=1}^{n_l} Y_{jlk}/N_{ji} \text{ and } N_{ji} = n_{j1} + \cdots + n_{ji}.$$

Let the prior information be available for both the control and treatment arms: $\mu_j \sim N_P(\eta_{j0}, \gamma_{j0})$ with $\gamma_{j0} = n_{j0}/\sigma_j^2$. In this case, update for posterior parameter is done per arm: $\mu_j|\bar{Y}_{ji} \sim N_P(\eta_{ji}, \gamma_{ji})$ where $\eta_{ji} = w_{ji}\eta_{j0} + (1 - w_{ji})\bar{Y}_{ji}$ and $\gamma_{ji} = \gamma_{j0} = N_{ji}/\sigma_j^2$. The posterior distribution for the treatment effect is given by $\delta|\bar{Y}_{2i}, \bar{Y}_{1i} \sim N_P(\tilde{\alpha}_i, \tilde{\beta}_i)$, where $\tilde{\alpha}_i = \eta_{2i} - \eta_{1i}$ and $\tilde{\beta}_i = (1/\gamma_{1i} + 1/\gamma_{2i})^{-1}$. We generate the observed look-wise average treatment response for a

large number of trials T_0. They are denoted by $\tilde{Y}_{ji} = \sum_{k=1}^{n_{ji}} Y_{jik}/n_{ji}$ under a different true average control and treatment responses μ_{10} and μ_{20}. The combined j-th arm treatment response is then $(n_{ji}\tilde{Y}_{ji} + N_{j,i-1}\bar{Y}_{j,i-1})/(n_{ji} + N_{j,i-1})$.

At each interim analysis the posterior distribution is updated arm-wise and converted to the treatment effect. The decision criteria are then applied to the posterior distribution of the treatment effect.

In both the cases, the OC are then derived by computing the proportion of trials for which the success and/or futility criteria are fulfilled. It is important to note that the denominator for the computation of the proportion is not the same at each interim. Because, at interim $i + 1$, we only have to consider the trials that continued from the previous analysis i and those two could be different. Therefore, T_0 must be large enough to ensure that enough simulated trials are continued to the final analysis. The simulation is summarized in pseudo-algorithms 2 and 3, respectively as shown in Gerber et al. (2016).

12.3.1.3 Expected sample size

The expected sample size (ESS) in a group sequential design is an important OC. It is computed as $\sum_{i=1}^{I} \pi_i \left(n_{1i} + n_{2i}\right)$ where π_i denotes the probability of stopping at the i-th interim. Once the probabilities of stopping for futility and stopping for success are available, the expected sample size is fairly straightforward to calculate (Gerber et al., 2016).

Algorithm 2 Pseudo-algorithm for simulation when prior is on treatment effect

for a large T_0 and each δ_u **do**

for $i = 1, 2, \cdots, I$ **do**

- Simulate $D_i^{(t)}, t = 1, \cdots, T_{i-1}$ with T_{i-1} the number of trials not stopped at $i-1$-th interim.
- Compute the Bayesian update of the posterior distribution recursively:

$$\beta_i = \beta_0 + B_i; \qquad \alpha_i^{(t)} = w_i \alpha_0 + (1 - w_i) D_i^{(t)}$$

- Compute T_i^S the number of trials fulfilling all success criteria at i-th interim.
- Compute probability of success at look i as T_i^S / T_{i-1}.
- Compute T_i^F the number of trials fulfilling all futility criteria at i-th interim.
- Compute probability of futility at look i as T_i^F / T_{i-1}.
- Set $T_i = T_{i-1} - T_i^S - T_i^F$.

end forloop for i

end forloop for δ_u

Algorithm 3 Pseudo-algorithm for simulation when prior is on both treatment arms

for a large T_0 and each plausible μ_{10} and μ_{20} do **do**

for $i = 1, 2, \cdots, I$ **do**

- Simulate $\tilde{Y}_{ji}^{(t)}$, $t = 1, \cdots, T_{i-1}$ and $j = 1, 2$, with T_{i-1} the number of trials not stopped at $i-1$-th interim.
- Compute $\bar{Y}_{j,i} = (n_{ji}\tilde{Y}_{ji} + N_{j,i-1}\bar{Y}_{j,i-1})/(n_{ji} + N_{j,i-1})$.
- Compute the Bayesian update for the posterior distribution per arm recursively:

$$\gamma_{ji} = \gamma_{j0} = N_{ji}/\sigma_j^2; \qquad \eta_{ji}^{(t)} = w_{ji}\eta_{j0} + (1 - w_{ji})\bar{Y}_{ji}^{(t)}.$$

- Convert arm-wise posterior distributions to posterior distribution of treatment effect:

$$\tilde{\alpha}_i^{(t)} = \eta_{2i}^{(t)} - \eta_{1i}^{(t)}; \qquad \tilde{\beta}_i^{(t)} = (1/\gamma_{1i} + 1/\gamma_{2i})^{-1}.$$

- Compute T_i^S the number of trials fulfilling all success criteria at i-th interim.
- Compute probability of success at look i as T_i^S/T_{i-1}.
- Compute T_i^F the number of trials fulfilling all futility criteria at i-th interim.
- Compute probability of futility at look i as T_i^F/T_{i-1}.
- Set $T_i = T_{i-1} - T_i^S - T_i^F$.

end forloop for i

end forloop for μ_{10} and μ_{20}

12.3.2 Binomial Endpoints

For binomial endpoints or proportion data, we use large-sample approximation theory so that we can use normal approximations for binary data. Following Spiegelhalter et al. (2004) and Agresti (2003), for binary data, we form an appropriate approximate normalized likelihood that can then be used with the setup shown in section 12.3.1. Examples of clinical trials with binary endpoints using R package **gsbDesign** can also be found in Gsponer et al. (2014).

Suppose the data comprise a series of observations in which an event has occurred or not, and we wish to compare the probability of such events under two different treatments. For two events with probabilities p_T and p_C from treatment and control arm, respectively, the odds ratio (OR) is defined as

$$\text{OR} = \frac{p_T}{(1-p_T)} \Big/ \frac{p_C}{(1-p_C)}$$

which is a standard way of describing the changes in the chances of events due to a treatment, on a scale between 0 and ∞. In order to make the assumption of a normal likelihood more plausible, it is convenient to work with the natural logarithm of the odds ratio so that it takes values on the whole range between $-\infty$ and $+\infty$ (Spiegelhalter et al., 2004). Thus we have

$$\log(\text{OR}) = \theta = \log\left(\frac{p_T}{1-p_T}\right) - \log\left(\frac{p_C}{1-p_C}\right)$$

and so the treatments are compared through their difference on the logit scale.

For normal endpoints we have known difference of means and variance i.e. σ_1^2 and σ_2^2 (see section 12.3.1). For binomial endpoints, we have the logarithm of the odds ratio as mean. For binary data, the estimated variance is a function of sample sizes and estimated response rates. Therefore, we need to calculate the variance for approximated normal likelihood along with the mean.

Now, when we want to estimate $log(\text{OR})$ from the data, we first need estimated response rate for control arms (p_C) and treatment arm (p^T). Using $\hat{p}_C$ and $\hat{p}_T$ the approximate variance of $\log(\hat{\text{OR}})$ is

$$Var[\log(\hat{\text{OR}})] = \frac{1}{n_T\hat{p_T}} + \frac{1}{n_T(1-\hat{p_T})} + \frac{1}{n_C\hat{p_C}} + \frac{1}{n_C(1-\hat{p_C})} \tag{12.3}$$

Please see the appendix below (section 12.3.2.1) to see the full derivation.

Now, in this module, in order to generate scenarios we take input for true response rate for control and treatment arms. Hence we compute the variance using the true response rates instead of using the estimates as shown in equation 12.3. We use that as the variance for the approximated normal likelihood.

For the prior specification, note that for normal we specify prior on treatment effect or both treatment arms separately (see sections 12.3.1.1 and 12.3.1.2). Here for binomial endpoints we use logit transformation to make the endpoints normally distributed. Therefore, those two options are equivalent to specifying prior on $\log(\text{OR})$ or on $\log(\text{Odds})$ for both the arms separately.

The rest of the statistical theory for binomial endpoints is very similar to that for normal endpoints as we are transforming the likelihood in this case to an approximated normal distribution to get a normally distributed posterior distribution using conjugacy.

12.3.2.1 Appendix: Derivation of the variance

For binomial endpoints *effect size* is the logarithm of the odds ratio of the response rate of treatment arm to that of control arm. The corresponding variance can be calculated as follows:

$$
\begin{aligned}
Var[log(\hat{OR})] &= Var\left[log\left(\frac{\frac{\hat{p_T}}{1-\hat{p_T}}}{\frac{\hat{p_C}}{1-\hat{p_C}}}\right)\right] \\
&= Var\left[log\left(\frac{\hat{p_T}}{1-\hat{p_T}}\right) - log\left(\frac{\hat{p_C}}{1-\hat{p_C}}\right)\right] \\
&= Var\left[log\left(\frac{\hat{p_T}}{1-\hat{p_T}}\right)\right] + Var\left[log\left(\frac{\hat{p_C}}{1-\hat{p_C}}\right)\right] \\
&\approx \left(\frac{1}{\hat{p_T}(1-\hat{p_T})}\right)^2 Var(\hat{p_T}) + \left(\frac{1}{\hat{p_C}(1-\hat{p_C})}\right)^2 Var(\hat{p_C}) \quad \text{(using delta method)} \\
&= \left(\frac{1}{\hat{p_T}(1-\hat{p_T})}\right)^2 \frac{\hat{p_T}(1-\hat{p_T})}{n_T} + \left(\frac{1}{\hat{p_C}(1-\hat{p_C})}\right)^2 \frac{\hat{p_C}(1-\hat{p_C})}{n_C} \\
&= \frac{1}{n_T}\left(\frac{(1-\hat{p_T})+\hat{p_T}}{\hat{p_T}(1-\hat{p_T})}\right) + \frac{1}{n_C}\left(\frac{(1-\hat{p_C})+\hat{p_C}}{\hat{p_C}(1-\hat{p_C})}\right) \\
&= \frac{1}{n_T\hat{p_T}} + \frac{1}{n_T(1-\hat{p_T})} + \frac{1}{n_C\hat{p_C}} + \frac{1}{n_C(1-\hat{p_C})}
\end{aligned}
$$

where $\hat{p_C}$ and $\hat{p_T}$ are the estimates of the response rates of the control and the treatment arm, respectively and n_C and n_T are the sample sizes for control and treatment arm, respectively. Note that this approximation would work better when both p_C and p_T are smaller.

12.3.3 Time-to-Event Endpoints

For time-to-event endpoints, we have a set of measurements of time to some event often referred to as survival data. This event is assumed to occur with hazard rate $h(t)$, which is the chance of an event in a short interval of time following t. Survival under two different interventions with hazard rates $h_2(t)$ and $h_1(t)$ may be compared by their hazard ratio, HR $= h_2(t)/h_1(t)$: the common "proportional hazards" assumption assumes HR is constant with time (Spiegelhalter et al., 2004).

The hazard ratio varies between 0 and ∞, and once again similar to the binomial endpoints, in order to make the assumption of a normal likelihood more plausible, it is convenient to work with the natural logarithm of the hazard ratio so that it takes values on the whole range between $-\infty$ and $+\infty$ (Spiegelhalter et al., 2004).

$$\log(HR) = \log\left[\frac{h_2(t)}{h_1(t)}\right]$$

Suppose that we have two-arm trial with the treatment arm (T), and the control arm (C). For time-to-event or survival data, following the large-sample approximation in the particular case of **equal allocation** and **same follow-up** as given in pages 28-29 of Spiegelhalter et al. (2004), we take standard deviation (σ) = 2 and adopt a normal likelihood.

For the prior specification, again note that for normal endpoints we specify prior on treatment effect or both treatment arms separately (see sections 12.3.1.1 and 12.3.1.2). Here for time-to-event endpoints, we use logarithm of hazard ratios to make the endpoints normally distributed. Therefore, those two options are equivalent to specifying prior on $\log(\text{HR})$ or on $\log(\text{Hazard Rate})$ for the control and the treatment arm separately.

The rest of the statistical theory for survival or time-to-event endpoints is very similar to that for normal endpoints as we are transforming the likelihood in this case to an approximated normal distribution to get a normally distributed posterior distribution using conjugacy.

13. Phase II/III Seamless Designs with Binary Endpoint

On East Bayes, we extended the simulation scheme introduced in Thall and Simon (1994) for phase 2-3 seamless design, including three arms in the phase 2 stage, two doses (high and low) and placebo. At the end of the phase 2 stage, the design will make a go/no-go decision based on **Bayesian go/no-go criteria** first and then select one of high and low doses as the treatment arm in phase 3 based on **Bayesian selection criteria**.

13.1 Binary Outcome

With binary outcome, a two-sided hypothesis z-test will be performed based on data of the selected treatment dose integrated from both phase 2 and 3 and data of control arm only from phase 3 at the end of phase 3 stage,

$$H_0 : p_T = p_C \quad \text{vs} \quad H_1 : p_T \neq p_C,$$

where p_T and p_C represent response probabilities of the selected treatment dose and placebo respectively.

13.1.1 Model

Let N_{H2}, N_{L2} and N_{C2} denote sample sizes, Y_{H2}, Y_{L2} and Y_{C2} numbers of patients with response and p_H, p_L and p_C response probabilities for three arms in phase 2 (H, L, C represent high dose, low dose and control). The sampling models are:

$$Y_{H2} \sim Bin(N_{H2}, p_H),$$

$$Y_{L2} \sim Bin(N_{L2}, p_L),$$

$$Y_{C2} \sim Bin(N_{C2}, p_C);$$

Priors in simulation:

$$p_H \sim Beta(\alpha_{H0}, \beta_{H0}),$$

$$p_L \sim Beta(\alpha_{L0}, \beta_{L0}),$$

$$p_C \sim Beta(\alpha_{C0}, \beta_{C0}).$$

The conditional posterior distribution of p is (here we suppress the subscript H, L and C):

$$p \mid N_2, Y_2 \sim Beta(\alpha_0 + Y_2, \beta_0 + N_2 - Y_2),$$

13.1.2 Decision Criteria

Bayesian go/no-go and selection decision with binary outcome from phase 2 on East Bayes are as below:

- **Bayesian go/no-go criteria** at the end of phase 2 stage based on two indicators, $h.go$ and $l.go$. If $h.go = 1$ or $l.go = 1$, go to phase 3. Otherwise, not go.

  Let $h = Pr(p_H > p_C + \delta_0 | data)$ and $l = Pr(p_L > p_C + \delta_0 | data)$,

$$h.go = \begin{cases} 1, \text{ if} & h \geq \eta_1 \\ \sim Bin(1, h.go.p), \text{ if} & \eta_2 < h < \eta_1 \\ 0, \text{ if} & h \leq \eta_2 \end{cases}$$

$$l.go = \begin{cases} 1, \text{ if} & l \geq \eta_1 \\ \sim Bin(1, l.go.p), \text{ if} & \eta_2 < l < \eta_1 \\ 0, \text{ if} & l \leq \eta_2 \end{cases}$$

  where $h.go.p = \frac{h-\eta_2}{\eta_1-\eta_2}$, $l.go.p = \frac{l-\eta_2}{\eta_1-\eta_2}$ and δ_0 denotes the expected difference between the probabilities of treatment dose and placebo.

- **Bayesian selection criteria** after making go decision based on one indicator, $h.select$. If $h.select = 1$, select the high dose ($T = H$). Otherwise, select the low dose ($T = L$).

$$h.select = \begin{cases} 1, \text{ if} & Pr(p_H > p_L | data) > \xi \\ 0, \text{ if} & Pr(p_H > p_L | data) \leq \xi \end{cases}$$

- (Criteria of z.test) Let N_{T3}, N_{C3} denote sample sizes and Y_{T3}, Y_{C3} numbers of patients with response for two arms in phase 3 (T, C represent selected treatment dose and control). The sampling models are the same,

$$Y_{T3} \sim Bin(N_{T3}, p_T),$$

$$Y_{C3} \sim Bin(N_{C3}, p_C).$$

And the estimated probabilities of two arms are,

$$\hat{p}_T = \frac{Y_{T2} + Y_{T3}}{N_{T2} + N_{T3}},$$

$$\hat{p}_C = \frac{Y_{C3}}{N_{C3}}.$$

If $1 - \Phi(Z) < z.test.\alpha/2$, we will think the selected treatment dose and placebo are significantly different, where $\Phi(*)$ denotes the standard normal distribution function and

$$Z = \frac{|\hat{p}_T - \hat{p}_C|}{\sqrt{\hat{p}_T(1-\hat{p}_T)/(N_{T2}+N_{T3}) + \hat{p}_C(1-\hat{p}_C)/N_{C3}}}.$$

13.1.3 Program Input and Output

13.1.3.1 Input

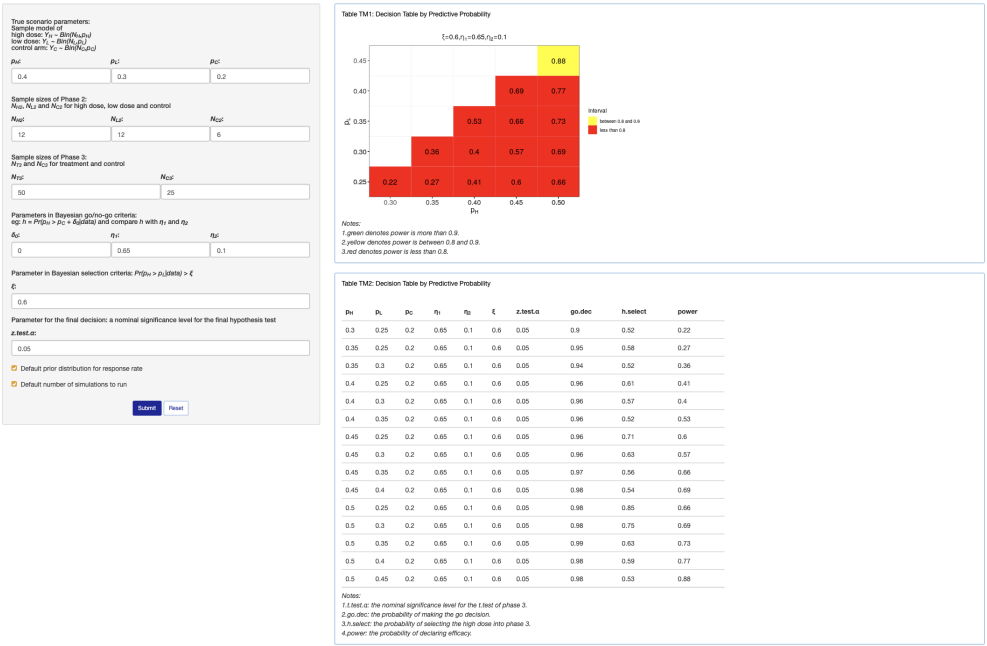
- p_H, p_L, p_C: true scenario parameters for three arms.
- N_{H2}, N_{L2}, N_{C2}: sample sizes of three arms in phase 2.
- N_{T3}, N_{C3}: sample sizes of treatment and control arms phase 3.
- δ_0, η_1, η_2, ξ: parameters in Go/No-Go and Selection decisions.
- $z.test.\alpha$: parameter for the final decision, a nominal significance level (or say the corresponding critical value) for the final hypothesis test in phase 3.
- α_{H0}, β_{H0}, α_{L0}, β_{L0}, α_{C0}, β_{C0}: parameters of prior distributions of the response rate
- Number of simulated trials

13.1.3.2 Output

- Decision table
- Probability of Go decision, probability of high-dose selection and Power

13.1.3.3 An Example (Figure 13.1)

Figure 13.1: An Example: Phase II/III Seamless Design with Binary Outcome



14. Phase II/III Seamless Designs with Continuous Endpoint

14.1 Introduction

The main objective in clinical trials is to find effective treatments for patients. Traditionally, phase II trials start to establish initial efficacy of a new treatment and phase III trials confirm the treatment's effectiveness. Modern clinical trials consider seamless phase II/III designs in which phase II compares multiple treatment arms and phase III selects one arm for testing against a control. Bischoff and Miller (2009) proposed a new test procedure for a new seamless phase II/III trial design. After a provisional sample size calculation in the planning stage, a portion of the planned sample is recruited at the first stage (phase II), the best treatment is estimated, and the sample size is recalculated on the basis of the observed variability. In the second stage (phase III), patients are randomized to the control arm and the estimated best treatment arm.

Here, we describe a module in East Bayes, **Phase II/III Seamless Design with Continuous Endpoint**, which performs trial simulation to examine the operating characteristics of the seamless design and searches the optimal parameter for sample size re-estimation using the method of Bischoff and Miller (2009). In this module, we extend the simulation scheme introduced in Thall and Simon (1994) and consider including three arms in phase II of drug development, two treatment arms (like two doses of one new agent) and one control arm. At the end of the first stage, a Bayesian Go/No-Go decision will be made first and if Go the design will select one of two treatments as the only arm for phase III based on a Bayesian rule. In the second stage, patients are randomized to the control and the selected treatment arm with a sample size re-estimation.

The remainder of the manual is organized as follows. §14.2 introduces the user interface and a tutorial on launching trial simulations and visualizing results. Statistical details of the seamless design are provided in §14.3. In particular, §14.3.1 introduces the simulation scheme with the

Bayesian Go/No-Go decision and selection rule, and §14.3.2 describes the method of sample size re-estimation in Bischoff and Miller (2009).

14.2 User Interface and Tutorial

14.2.1 Overview

The **Phase II/III Seamless Design — Continuous Endpoint** page of East Bayes has three main tabs: **Simulation Setup**, **Simulation Results**, and **SSR Calculator**. The first tab, **Simulation Setup**, allows users to conduct simulations; the second tab, **Simulation Results**, to visualize/download simulation results; and the third tab, **SSR Calculator**, to re-estimate sample size at the interim analysis. In the **Simulation Setup** tab, there are three steps (Figure 14.1). Step 1: **Input Simulation Parameters**, Step 2: **Input Design Parameters**, and Step 3: **Generate Scenarios**. Upon completing Steps 1-3, click the "Launch Simulation" button at the bottom of the page to begin the simulation using the current parameters, or click the "Reset" button to clear all settings and enter new parameters. After the simulation completes, the results will be displayed in the **Simulation Results** tab. Step-by-step instructions are shown in §14.2.2–§14.2.4.

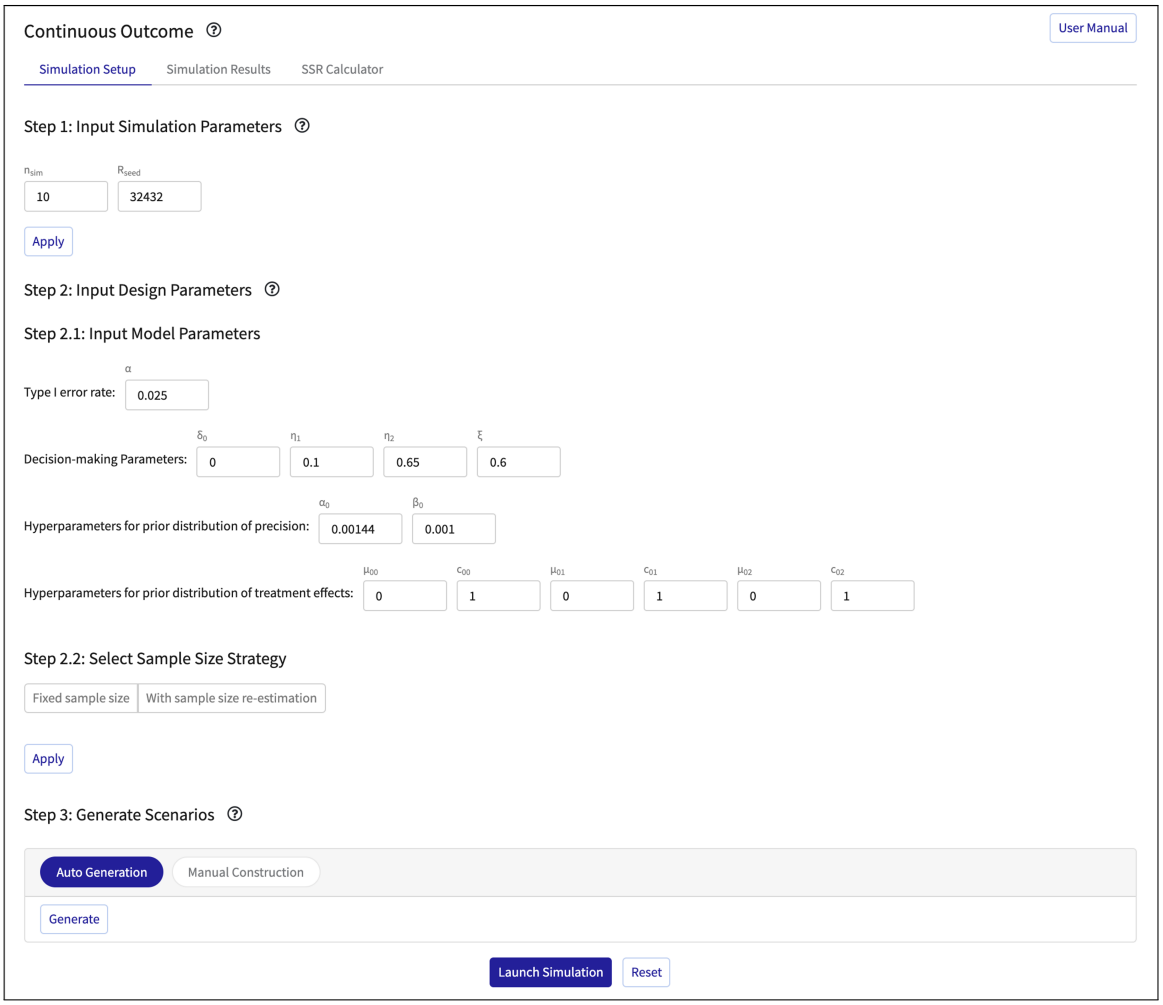


Figure 14.1: Simulation Setup in the **Phase II/III Seamless Design — Continuous Endpoint** module.

14.2.2 Simulation Setup

East Bayes requires users to provide input parameter values for the seamless design in three steps. When hovering over the question mark icons, a description of parameters used in the section is displayed. If there are parameters you would like to change which are not currently accessible, or designs you would like to see added to this module, please contact us by emailing support@cytel.com.

14.2.2.1 Step 1: Input Simulation Parameters

First specify the number of simulations (n_{sim}) and the simulation seed value (R_{seed}). See Figure 14.2. A detailed explanation of these input arguments will be provided in Table 14.1.

Click the "Apply" button (Figure 14.2) to confirm the input simulation parameters. The "Apply" button changes to "Edit" and can be clicked again to change trial parameters as needed.

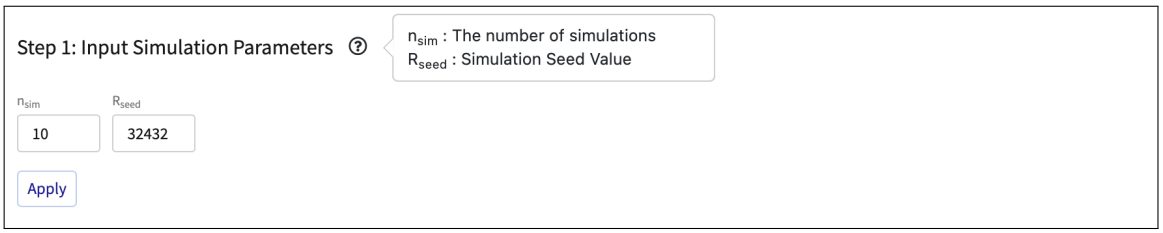


Figure 14.2: Input Simulation Parameters in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Table 14.1: Simulation parameters in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
n_{sim}	Number of simulations	The maximum number of simulations allowed is 10,000. The default value is 10.
R_{seed}	Simulation seed value	A number used to initialize a pseudo random number generator in the simulation. The default value is 32432.

14.2.2.2 Step 2: Input Design Parameters

First enter the desired model parameters in their respective entry fields, and then click one of the "Fix sample size" and "With sample size re-estimation" buttons to select a sample size strategy. Different strategies require different design parameters. For a detailed parameter description list, see Table 14.2–14.5 next.

Figure 14.3: Input Design Parameters in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Table 14.2: Model parameters in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
α	Type I error rate	The probability of wrongly rejecting the null hypothesis described in §14.3.2.1. The range is $(0, 1)$ and the default value is 0.025.
δ_0	Meaningful effect difference between the treatment and control arms	When the treatment arm exhibits a better response by a margin of δ_0 over the control arm, it is regarded promising. The range is $[0, +\infty)$ and the default value is 0. This is used for Bayesian Go/No-Go decision in Stage 1.
η_1	Parameters in the Bayesian Go/No-Go criteria	The lower probability threshold when making the Go/No-Go decision. The range is $(0, \eta_2)$ and the default value is 0.1.
η_2		The upper probability threshold when making the Go/No-Go decision. The range is $(\eta_1, 1)$ and the default value is 0.65. See details of Bayesian Go/No-Go criteria in §14.3.1.1
ξ	Parameter in the Bayesian selection rule	The probability threshold when selecting a better treatment arm. The range is $(0, 1)$ and the default value is 0.6. See details of Bayesian selection rule in §14.3.1.2.

Prior distributions for precision and treatment effects of three arms

Table 14.3: Prior distributions in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
α_0	Hyperparameters for prior distribution for precision	Hyperparameters of the gamma prior distribution for the precision. The ranges are both $(0, +\infty)$ and the default values are 0.00144 and 0.001 for α_0 and β_0, respectively. See details of the priors in §14.3.1
β_0		
μ_{00}	Hyperparameters for prior distribution for the treatment effect of the control arm	Hyperparameters of the normal prior distribution for the mean response of the control arm. The ranges are $(-\infty, +\infty)$ and $(0, +\infty)$ and the default values are 0 and 1 for μ_{00} and c_{00}, respectively.
c_{00}		
μ_{01}	Hyperparameters for prior distribution for the treatment effect of treatment arm 1	Hyperparameters of the normal prior distribution for the mean response of treatment arm 1. The ranges are $(-\infty, +\infty)$ and $(0, +\infty)$ and the default values are 0 and 1 for μ_{01} and c_{01}, respectively.
c_{01}		
μ_{02}	Hyperparameters for prior distribution for the treatment effect of treatment arm 2	Hyperparameters of the normal prior distribution for the mean response of treatment arm 2. The ranges are $(-\infty, +\infty)$ and $(0, +\infty)$ and the default values are 0 and 1 for μ_{02} and c_{02}, respectively.
c_{02}		

Design parameters when selecting "Fixed sample size"

Table 14.4: Design parameters when selecting "Fixed sample size" in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
n_{10}	Sample size of the control arm in stage 1	The number of patients treated at the control arm in stage 1. The range is $[10, 10000]$ and the default value is 50.
n_{11}	Sample size of treatment arm 1 in stage 1	The number of patients treated at treatment arm 1 in stage 1. The range is $[10, 10000]$ and the default value is 50.
n_{12}	Sample size of treatment arm 2 in stage 1	The number of patients treated at treatment arm 2 in stage 1. The range is $[10, 10000]$ and the default value is 50.
n_{20}	Sample size of the control arm in stage 2	The number of patients treated at the control arm in stage 2. The range is $[10, 10000]$ and the default value is 100.
n_{2t}	Sample size of the selected treatment arm in stage 2	The number of patients treated at the selected treatment arm in stage 2. The range is $[10, 10000]$ and the default value is 100.

Trial parameters when selecting "With sample size re-estimation"

Table 14.5: Trial parameters when selecting "With sample size re-estimation" in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
Δ	Treatment effect	Under "With sample size re-estimation", the sample sizes of the trial is estimated with type I error rate less than or equal to α and the power larger than or equal to $(1-\beta)$ when at least one treatment arm exhibits a better response by a margin of of Δ over the control arm. The range is $(0, +\infty)$ and the default value is 4. This is used in Stage 2.
β	Type II error rate	The power, $(1-\beta)$, is the probability of correctly rejecting the null hypothesis. The range is $(0, 1)$ and the default value is 0.2.
γ	Error selection rate	The sample size for each arm at stage 1 should be set to guarantee that the probability of selecting the inferior treatment arm for stage 2 is smaller than or equal to γ. The range is $(0, 1)$ and the default value is 0.1. See details in §14.3.2.3.
For "Fixed σ^2"		
σ^2	Variance of treatment effect	The variance of treatment effect. The range is $(0, +\infty)$ and the default value is 144.
For "Point mass prior of σ^2"		
n_v	Point mass prior for variance of treatment effect	The variance of treatment effect is treated as a discrete random variable taking a set of n_v distinctive values with probabilities p. The ranges of the possible values are $(0, +\infty)$ and the ranges of the probabilities are $(0, 1)$, and the sum of all the probabilities is equal to 1.
σ^2		
p		

14.2.2.3 Step 3: Generate scenarios

There are two ways to generate scenarios, automatically (see Figure 14.4) or manually (see Figure 14.5). In East Bayes, we assume the true treatment effects of patients in each arm are independent and follow a Gaussian distribution. For each scenario, the means and variances of three arms need to be specified.

Auto Generation (Figure 14.4)
Upon clicking the "Generate" button, three or six scenarios will be created automatically, each of which contains the true means and variances of three arms. If users select "With sample size re-estimation" and "Point mass prior of σ^2", six scenarios will be created. Otherwise, three will be created.

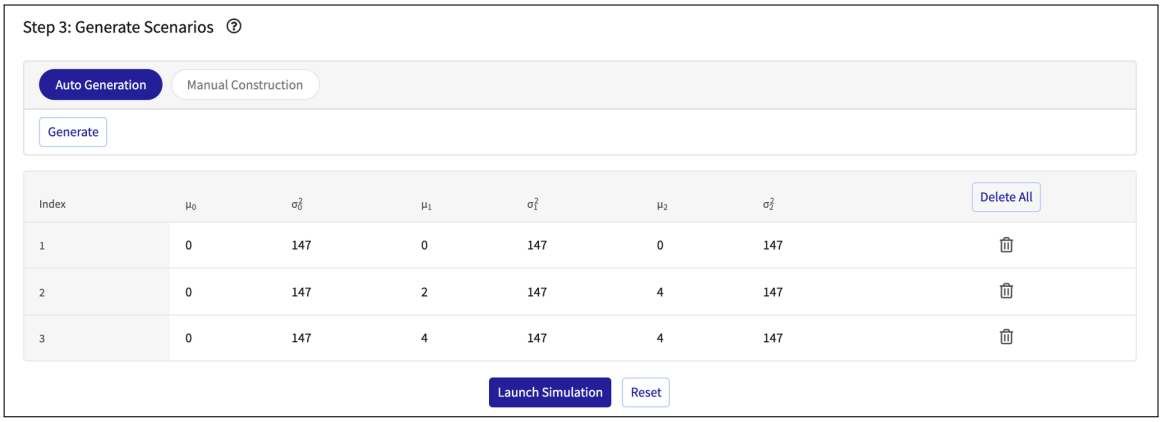


Figure 14.4: Automatically generate scenarios in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Manual Construction (Figure 14.5)

Manually input mean and variance for each arm, then click the "Add" button to create a new scenario.

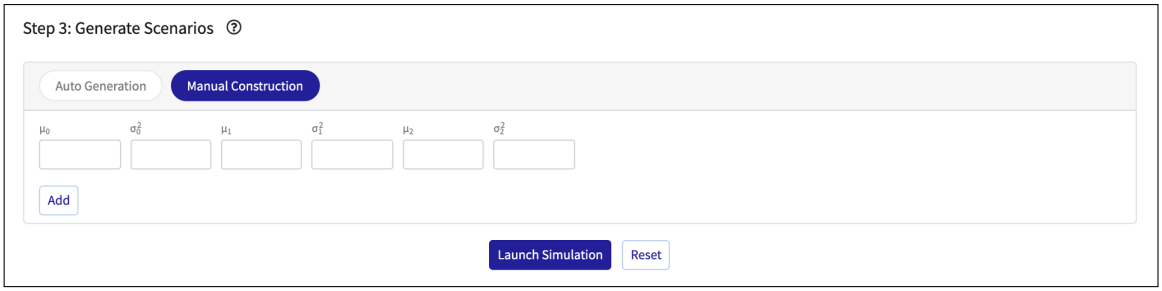


Figure 14.5: Manually generate scenarios in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Once scenarios are generated, click the "Launch Simulation" button at the bottom of the page to run n_{sim} (set in Step 1) simulations for each scenario and selected sample size strategy (set in Step 2) combination.

14.2.2.4 Launch Simulation

Once Steps 1-3 are completed, click the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 14.4 and 14.5) to submit the job. A "**Success**" message will be displayed (Figure 14.6) to indicate the simulation has been successfully launched. Users may click the "OK" button in the pop-up box to proceed to **Simulation Results** tab and track the simulation processing status and visualize simulation results.

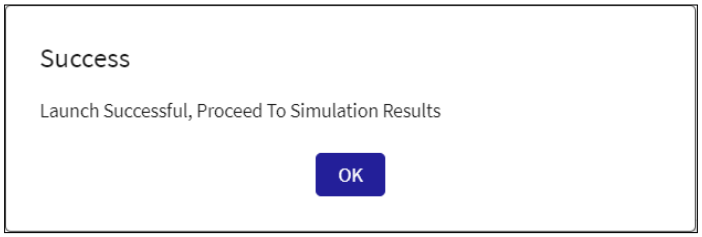


Figure 14.6: The "**Success**" message after launching simulation in the **Phase II/III Seamless Design — Continuous Endpoint** module.

14.2.3 Simulation Results

The **Simulation Results** tab is primarily used for viewing the simulation jobs and simulation results (§14.2.3.1), for restoring simulation settings to reproduce the simulation results or make change in the simulation set as needed (§14.2.3.2), and for downloading simulation reports (§14.2.3.3). Simulation results (figures and tables) can be downloaded in Word format, with accompanying statistical sections in a trial protocol. Hereinafter, we use simulation results and operating characteristics interchangeably.

14.2.3.1 View simulation results

Once simulations are completed, a message appears in the **Running Simulations** panel, and the simulation results are automatically loaded into the **Simulation History** panel (Figure 14.7), a mail icon is used to indicate new results that have not been viewed. The duration displayed depends on the availability of computing resources, and includes the waiting time after submitting the simulation.

Simulation results for other modules can be viewed by using the "Select a Design Category" drop-down box (Figure 14.7).

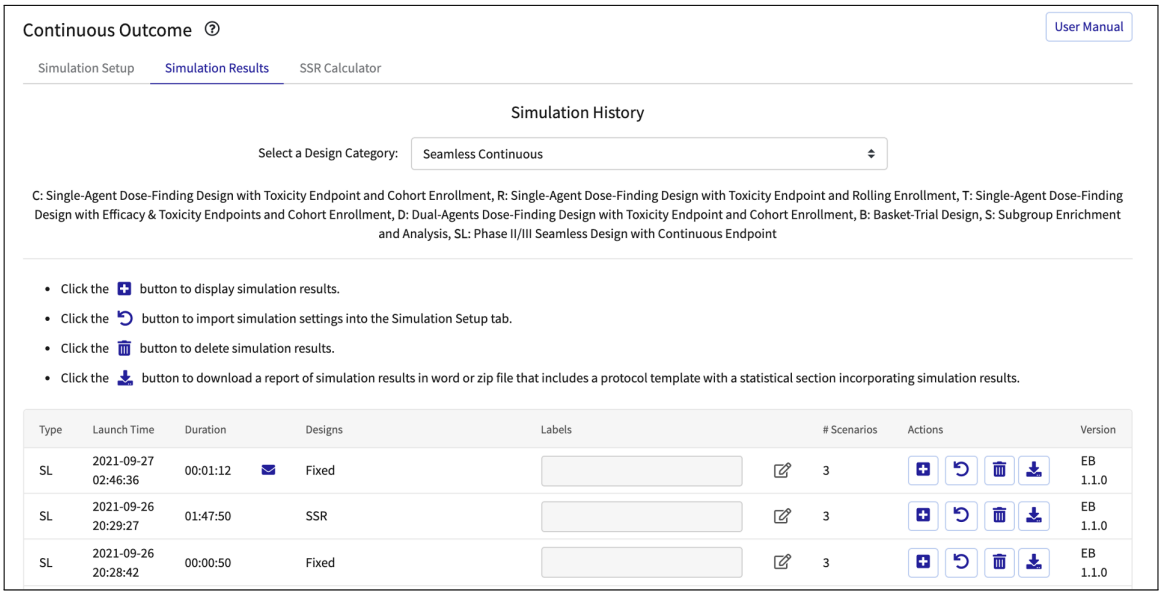
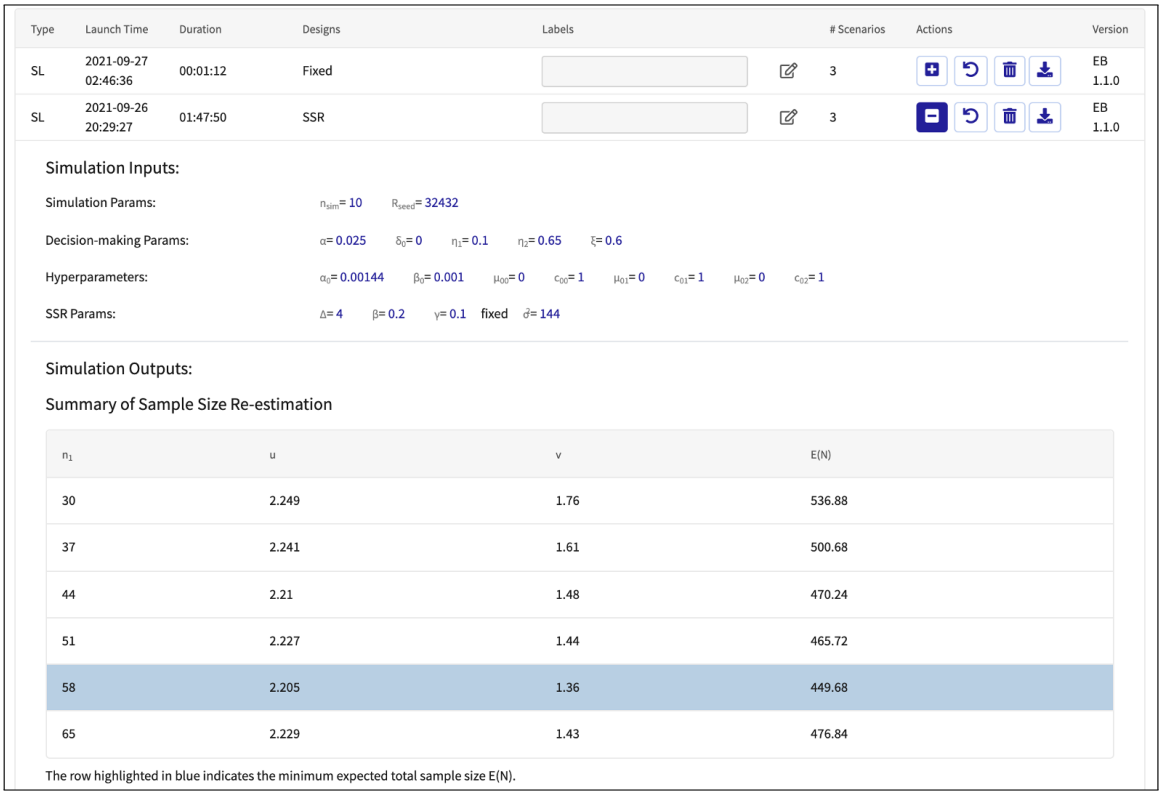


Figure 14.7: Simulation Results in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Click the button ⊞ to expand the panel and view simulation results (Figure 14.8). The simulation and trial parameters are displayed at the top of each simulation job (Figure 14.8) followed by the results in both tabular and graphical form.

If a set of simulation results is no longer needed, click the 🗑 button to delete the selected simulation results. There is no un-delete option.



Type	Launch Time	Duration	Designs	Labels	# Scenarios	Actions	Version
SL	2021-09-27 02:46:36	00:01:12	Fixed		3		EB 1.1.0
SL	2021-09-26 20:29:27	01:47:50	SSR		3		EB 1.1.0

Simulation Inputs:

Simulation Params: n_{sim}= 10 R_{seed}= 32432

Decision-making Params: α= 0.025 δ_0= 0 η_1= 0.1 η_2= 0.65 ξ= 0.6

Hyperparameters: α_0= 0.00144 β_0= 0.001 μ_{00}= 0 c_{00}= 1 μ_{01}= 0 c_{01}= 1 μ_{02}= 0 c_{02}= 1

SSR Params: Δ= 4 β= 0.2 γ= 0.1 fixed ∂= 144

Simulation Outputs:

Summary of Sample Size Re-estimation

n_1	u	v	E(N)
30	2.249	1.76	536.88
37	2.241	1.61	500.68
44	2.21	1.48	470.24
51	2.227	1.44	465.72
58	2.205	1.36	449.68
65	2.229	1.43	476.84

The row highlighted in blue indicates the minimum expected total sample size E(N).

Figure 14.8: View the simulation results in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Details of the Simulation Results

Simulation results are first summaried across scenarios and then elaborated by each scenario. There are three sections of simulation results:

A. Summary of sample size re-estimation. (Figure 14.9, only available upon selecting the sample size strategy "With sample size re-estimation" in the simulation setting).
B. Summary of performance. (Figure 14.10).
C. Detailed results by scenarios. (Figure 14.11).

A. Summary of sample size re-estimation.

Figure 14.9 shows the expected total sample size for the trial, $E(N)$, when the sample size for each arm in stage 1 is n_1. The summary is only available when "With sample size re-estimation" is selected in the simulation setting.

Summary of Sample Size Re-estimation

n_1	u	v	E(N)
14	2.295	1.81	169.298
19	2.272	1.59	155.423
24	2.259	1.46	149.31
29	2.249	1.38	147.899

The row highlighted in blue indicates the minimum expected total sample size E(N).

Figure 14.9: Summary of sample size re-estimation in the **Phase II/III Seamless Design — Continuous Endpoint** module.

B. Summary of performance.

Figure 14.10 shows scenario-specific summary statistics. They are explained in full detail next.

- **Freq. of Go**: The frequency of making the Go decision at the end of stage 1 across all simulated trials. Here, "Go" means a treatment arm from phase II will be selected for testing in phase III.
- **Freq. of Selecting Treatment 1**: The frequency of selecting treatment 1 to enter stage 2 at the end of stage 1 across all simulated trials.
- **Freq. of Selecting Treatment 2**: The frequency of selecting treatment 2 to enter stage 2 at the end of stage 1 across all simulated trials.
- **Power 1**: The frequency of declaring one of treatment arms to be better than the control arm at the end of the trial using a one-sided superiority t-test across all simulated trials.
- **Power 2**: (Only available upon selecting the sample size strategy "With sample size re-estimation" in the simulation setting) The frequency of declaring one of treatment arms to be better than the control arm at the end of the trial using the statistic proposed by Bischoff and Miller (2009) and described in §14.3.1.3 across all simulated trials.
- **E(N) (s.d.)**: The average total number of patients treated in three arms in the simulated trials and its standard deviation.

Summary of Performance

Scenario	(μ_0, σ_0^2), (μ_1, σ_1^2), (μ_2, σ_2^2)	Freq. of Go	Freq. of Selecting Treatment 1	Freq. of Selecting Treatment 2	Power 1	Power 2	E(N) (s.d.)
1	(0,22), (0,22), (0,22)	0.9	0.6	0.3	0	0	89.2 (3.327)
2	(0,22), (2,22), (4,22)	1	0	1	0.9	0.9	88.2 (3.155)
3	(0,22), (4,22), (4,22)	1	0.6	0.4	1	0.8	89.6 (2.836)
4	(0,66), (0,66), (0,66)	0.6	0.2	0.4	0	0	158 (65.498)
5	(0,66), (2,66), (4,66)	1	0.4	0.6	0.9	0.8	213.6 (38.552)
6	(0,66), (4,66), (4,66)	1	0.6	0.4	0.9	0.9	212.4 (32.729)

Power 2: Only for the sample size strategy of with sample size re-estimation

Figure 14.10: Summary of performance in the **Phase II/III Seamless Design — Continuous Endpoint** module.

C. Detailed results by scenarios.

The detailed simulation results are presented and arranged by scenarios. There are three bar plots for **Freq. of Selection**, **Power 1**, and **Power 2**, and one box plot for **Treatment Effect Difference**.

- **Freq. of Selection**: These three bars denote the frequencies of three selection decisions at the end of stage 1 among all simulated trials, separately. The three selection decisions are,
  - **No Selection**: No treatment is seleted as promising at the end of stage 1.
  - **Treatment 1**: Selecting treatment 1 to enter stage 2.
  - **Treatment 2**: Selecting treatment 2 to enter stage 2.
- **Power 1**: The frequencies of the following three decisions at the end of the trial using a one-sided superiority t-test across all simulated trials.
- **Power 2**: (Only available upon selecting the sample size strategy "With sample size re-estimation" in the simulation setting) The frequencies of the following three decisions at the end of the trial using the statistic proposed by Bischoff and Miller (2009) and described in §14.3.1.3 across all simulated trials.

  For **Power 1** and **Power 2**, the three trial decisions are,
  - **No Promising**: No treatment arms are selected at the end of stage 2.
  - **Treatment 1**: Treatment 1 is promising and selected at the end of stage 2, i.e., better than the control arm.
  - **Treatment 2**: Treatment 2 is promising and selected at the end of stage 2, i.e., better than the control arm.
- **Treatment Effect Difference**: The difference in the treatment effect between the treatment and control arms among the simulated trials.
  - **NS1**: The treatment effect difference between treatment 1 and the control arm in these simulated trials that stop at the end of stage 1 with "No Selection".
  - **NS2**: The treatment effect difference between treatment 2 and the control arm in these simulated trials that stop at the end of stage 1 with "No Selection".

- **NP11 (NP12)** : The treatment effect difference between treatment 1 (treatment 2) and the control arm in these simulated trials that enter stage 2 with one of treatment arms, but do not declare the treatment arm to be better than the control arm at the end of stage 2 using a one-sided superiority t-test.
- **P11 (P12)**: The treatment effect difference between treatment 1 (treatment 2) and the control arm in these simulated trials that declare one of treatment arms to be better than the control arm at the end of stage 2 using a one-sided superiority t-test.
- **NP21 (NP22)** : The treatment effect difference between treatment 1 (treatment 2) and the control arm in these simulated trials that enter stage 2 with one of treatment arms, but do not declare the treatment arm to be better than the control arm at the end of stage 2 using the statistic proposed by Bischoff and Miller (2009) and described in §14.3.1.3.
- **P21 (P22)**: The treatment effect difference between treatment 1 (treatment 2) and the control arm in these simulated trials that declare one of treatment arms to be better than the control arm at the end of stage 2 using the statistic proposed by Bischoff and Miller (2009) and described in §14.3.1.3.

Figure 14.11: Detailed results by scenarios in the **Phase II/III Seamless Design — Continuous Endpoint** module.

14.2.3.2 Restore simulation setup

Users can "restore" the simulation input settings from the simulation results by clicking the button (yellow arrow in Figure 14.12). When clicked, this button navigates to the **Simulation Setup** page and recreates the original simulation input.

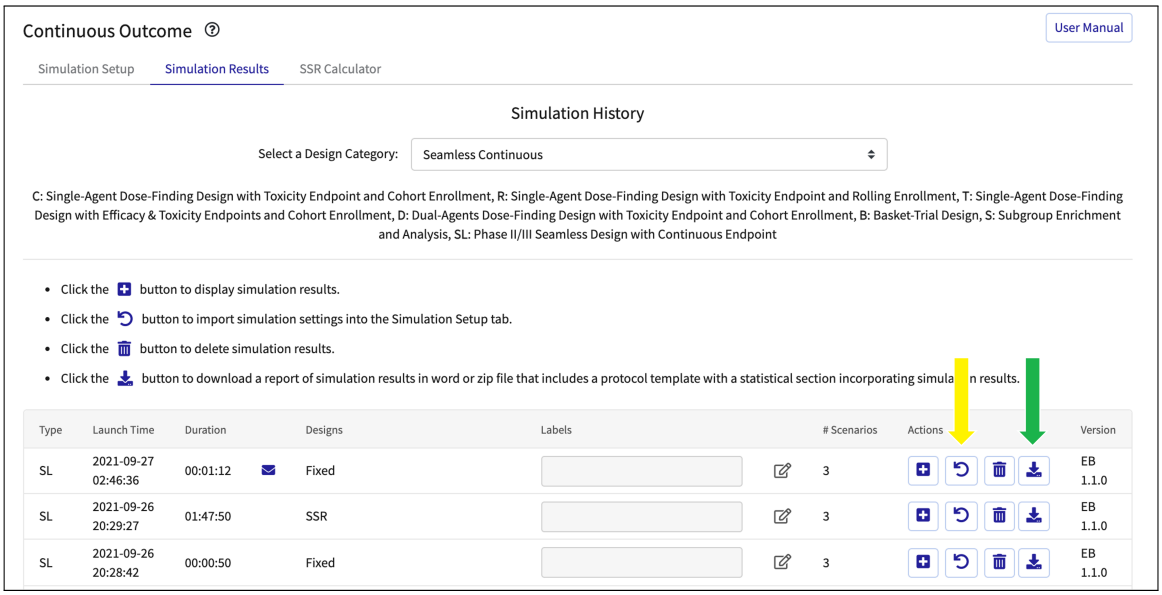


Figure 14.12: Restore simulation setup and download simulation results in the **Phase II/III Seamless Design — Continuous Endpoint** module.

14.2.3.3 Download simulation results

The download button (green arrow in Figure 14.12) creates and downloads a Word document, which includes three parts:

- Part A: Complete simulation results for the method and scenarios users selected,
- Part B: Detailed technical descriptions of the designs,
- Part C: References.

14.2.4 SSR Calculator

In the **SSR Calculator** tab, users can calculate the sample size for each arm at stage 2 using the method described in §14.3.2 when stage 1 of the trial is completed and data collected.

Specify the tuning parameter for power, v, the within-group variance at stage 1, S_1^2, and the sample size of each arm at stage 1, n_1; and click the "Estimate" button to calculate the sample size for each arm at stage 2 as shown in Figure 14.13. See detailed parameter descriptions in Table 14.6.

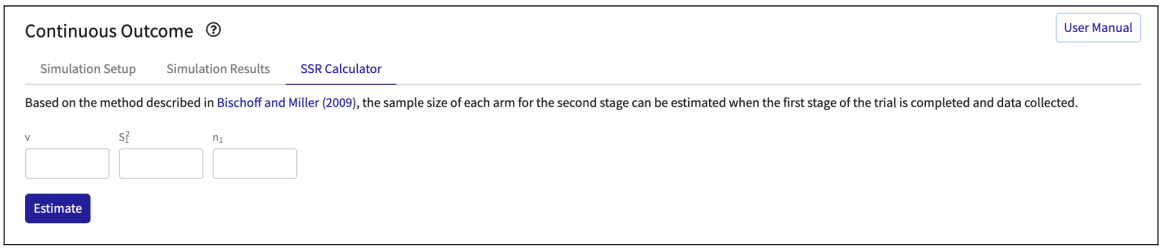


Figure 14.13: SSR Calculator in the **Phase II/III Seamless Design — Continuous Endpoint** module.

Table 14.6: Input parameters in the **SSR Calculator** tab of the **Phase II/III Seamless Design — Continuous Endpoint** module.

Notation	Parameters	Description
v	The tuning parameter for power	v is a parameter that ensure the power is at least $(1-\beta)$, and can be calculated and shown for sample size re-estimation. See the example in Figure 14.9. The range is $(0, 10]$.
S_1^2	The within-group variance at stage 1	S_1^2 is the sample variance calculated at the end of stage 1 using observed data. See how to calculate S_1^2 in (14.4). The range is $(0, +\infty)$.
n_1	The sample size of each arm at stage 1	The sample sizes n_1 of three arms at stage 1 are the same in the method described in §14.3.2. The range is $[10, 10000]$.

14.3 Statistical Methods

This section describes the method of sample size re-estimation in Bischoff and Miller (2009). Consider a seamless clinical trial with two stages, phase II and phase III, and three arms, two treatment arms (e.g., two doses of a new drug) and one control arm. All three arms will be evaluated at the end of Stage 1. If both treatment arms are not promising, stop the trial. If at least one of treatment arms is promising, select only one treatment arm and proceed to Stage 2. At the end of Stage 2, assess whether the selected treatment arm is better than the control.

Let y_{ijk} denote the response of patient i in arm j, $j \in \{0, 1, 2\}$, in stage k, $k \in \{1, 2\}$. Arm $j = 0$ corresponds to the control arm, whereas arms $j = 1, 2$ correspond to two treatment arms. For Stage 1 of the trial, randomly allocate $\sum_{j=0}^{2} n_{1j}$ patients to the three arms with n_{1j} patients to arm j. At the end of Stage 1, whether or not proceed to Stage 2 is decided based on Bayesian Go/No-Go decision criteria (§14.3.1.1) and the better of the two treatment arms, denoted as $\hat{t}$, is chosen for testing in Stage 2 based on a Bayesian selection rule (§14.3.1.2). In Stage 2, randomly allocate $(n_{20} + n_{2\hat{t}})$ patients to arms 0 and $\hat{t}$, i.e., n_{20} to arm 0 and $n_{2\hat{t}}$ to arm $\hat{t}$.

There are two sample size strategies in East Bayes. One is "Fixed sample size", which means that users need to specify sample sizes of treatment and control arms at both stages, n_{1j}, $j = 0, 1, 2$, n_{20} and $n_{2\hat{t}}$. The other is "With sample size re-estimation", which means that the sample sizes are calculated using the method in §14.3.2. By default, $n_{10} = n_{11} = n_{12} = n_1$ and $n_{20} = n_{2\hat{t}} = N_2$.

14.3.1 Probability Model

Bischoff and Miller (2009) assumes y_{ijk}'s are independent and follow a Gaussian distribution,

$$y_{ijk} \sim N(\mu_j, \frac{1}{\tau_j}).$$

Let $Y_{jk} = \{y_{ijk} \mid i = 1, \cdots, n_{kj}\}$ denote the set of response for arm j in stage k.

For simplicity, suppress the subscript of stage k when introducing the probability model. Given the priors of μ_j and τ_j,

$$\begin{aligned} \mu_j|\tau_j &\sim N(\mu_{0j}, \frac{1}{c_{0j}\tau_j}), \\ \tau_j &\sim Gamma(\alpha_0, \beta_0), \end{aligned} \tag{14.1}$$

the joint posterior distribution of μ_j and τ_j is,

$$\begin{aligned} P(\mu_j, \tau_j | Y_j) \propto & P(\tau_j) p(\mu_j | \tau_j) P(Y_j | \mu_j, \tau_j) \\ \propto & \tau_j^{\alpha_0 - 1} e^{-\beta_0 \tau_j} \tau_j^{1/2} exp\left(-\frac{c_0 \tau_j}{2}(\mu_j - \mu_{0j})^2\right) \tau_j^{n/2} exp\left(-\frac{\tau_j}{2}\sum_i (y_{ij} - \mu_j)^2\right). \end{aligned}$$

After integrating out μ_j, a Gamma marginal posterior for τ_j can be derived as

$$\tau_j | Y_j \sim Gamma\left(\alpha_0 + \frac{n_j}{2}, \beta_0 + \frac{1}{2}\sum_i (y_{ij} - \overline{y}_j)^2 + \frac{n_j c_0}{2(n_j + c_0)}(\overline{y}_j - \mu_{0j})^2\right), \tag{14.2}$$

where $Y_j = \{y_{ij} \mid i = 1, \cdots, n_j\}$, $\overline{y}_j = \sum_{i=1}^{n_j} y_{ij}/n_j$, and n_j denotes the number of patients included in Y_j for arm j. The Bayesian Go/No-Go decision criteria and selection rule are based on the marginal posterior distributions of μ_j, $p(\mu_j \mid Y_j)$, which is given by,

$$p(\mu_j \mid Y_j) = p(\mu_j \mid \tau_j, Y_j) p(\tau_j \mid Y_j). \tag{14.3}$$

With the marginal posterior distributions of τ_j and μ_j, (14.2) and (14.3), it is easy to draw samples for $p(\mu_j \mid Y_j)$ by the next processes,

1. sample one τ_j^* from the marginal posterior distributions of τ_j, $p(\tau_j \mid Y_j)$.
2. then sample one μ_j^* from the conditional posterior distribution of μ_j, $p(\mu_j \mid \tau_j^*, Y_j)$, with one fixed τ_j^* sampled from step 1,

where the conditional posterior distribution of μ_j is

$$\mu_j | Y_j, \tau_j \sim N\left(\frac{n_j}{n_j + c_{0j}}\overline{y}_j + \frac{c_{0j}}{n_j + c_{0j}}\mu_{0j}, \frac{1}{n_j \tau_j + c_{0j}\tau_j}\right).$$

14.3.1.1 Bayesian Go/No-Go Criteria

The Bayesian Go/No-Go criteria at the end of Stage 1 are based on two posterior probabilities, $p_{g1} = Pr(\mu_1 > \mu_0 + \delta_0 \mid Y_{01}, Y_{11})$ and $p_{g2} = Pr(\mu_2 > \mu_0 + \delta_0 \mid Y_{01}, Y_{21})$, where δ_0 is a clinical treatment effect difference between the treatment and control arms specified by users. And the two probabilities assess the chance that the treatment effect of the treatment arm j is clinically better than that of the control arm 0, $j = 1, 2$ separately. Define

$$g_j = \begin{cases} 1, \text{ if } p_{gj} > \eta_2, \\ \sim Bin(1, \tilde{p}_{gj}), \text{ if } \eta_1 < p_{gj} \leq \eta_2, \\ 0, \text{ if } p_{gj} \leq \eta_1, \end{cases}$$

where $\tilde{p}_{gj} = \frac{p_{gj}-\eta_1}{\eta_2-\eta_1}$ and $Bin(n,p)$ denotes the binomial distribution with n independent experiment(s) and the success probability of p. If $g_1 = 1$ or $g_2 = 1$, select one of treatment arms and proceed to Stage 2 (Go). Otherwise, stop the trial (No Go).

Following the criteria, when p_{g1} and p_{g2} are both small, i.e., less than or equal to a small fraction η_1, such as 0.1, it is unlikely that the two treatment arms are more efficacious than the control, and stopping the trial early (No Go) can prevent patients from being exposed to the ineffective investigational treatments.

14.3.1.2 Bayesian Selection Rule

If the "Go" decision is made, one of treatment arms is selected to tested in Stage 2. At the end of Stage 1, define

$$s_2 = \begin{cases} 1, \text{ if } Pr(\mu_2 > \mu_1 \mid Y_{21}, Y_{11}) > \xi \\ 0, \text{ if } Pr(\mu_2 > \mu_1 \mid Y_{21}, Y_{11}) \leq \xi \end{cases}$$

If $s_2 = 1$, select arm 2 ($\hat{t} = 2$). Otherwise, select arm 1 ($\hat{t} = 1$).

If $\xi > 0.5$, this rule is friendly to arm 1, and vice versa. For example, assuming two doses of an investigational agent are tested as the two treatment arms, arm 1 represents the lower dose of the agent, and arm 2 the higher dose, one may prefer the lower dose due to the safety if it has similar treatment effect as the higher dose.

14.3.1.3 Final Inference

In East Bayes, we provide two methods to decide whether one treatment arm is better than the control arm at the end of Stage 2.

- A one-sided t-test including data of the control and selected arms at both stages, $Y_{j\cdot} = \{y_{ijk} \mid i = 1, \cdots, n_{kj}, \quad j = 0, \hat{t}, \quad k = 1, 2\}$. We call the power calculated by this method "Power 1" in East Bayes.

- The test statistic recommended by Bischoff and Miller (2009),

$$\xi = (\overline{y}_{\cdot\hat{t}\cdot} - \overline{y}_{\cdot 0\cdot})\sqrt{\frac{N}{2S_1^2}},$$

  where $\overline{y}_{\cdot j\cdot}, j \in \{0, \hat{t}\}$ denotes the average treatment effect of arm j at Stage 2, $N = n_1 + N_2$,

and

$$S_1^2 = \frac{1}{3(n_1 - 1)} \sum_{j=0}^{2} \sum_{i=1}^{n_1} \left(y_{ij1} - \overline{y}_{\cdot j1} \right)^2 . \tag{14.4}$$

Here we follow the default settings of Bischoff and Miller (2009), and set $n_{10} = n_{11} = n_{12} = n_1$ and $n_{20} = n_{2\hat{t}} = N_2$. Then the trial ends with declaring the treatment arm promising only when $\xi > u$, where u is calibrated. See §14.3.2. We call the power calculated by this method "Power 2" in East Bayes.

14.3.2 Sample Size Re-estimation

For Stage 1 of the trial, randomly allocate n_1 patients to each of the three arms, i.e., $n_{10} = n_{11} = n_{12} = n_1$. At the end of Stage 1, if one treatment arm $\hat{t}$ is more promising and selected for testing in the next stage, compute the sample size for each arm at Stage 2, $n_{20} = n_{2\hat{t}} = N_2$, as,

$$N_2 = N_2(S_1^2) = \lceil \max(vS_1^2 - n_1, n_{2,\min}) \rceil,$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x, S_1^2 is calculated by (14.4) and $n_{2,\min}$ is a minimal required number of patients per arm for Stage 2. In East Bayes, we set $n_{2,\min} = 0$ by default. And v is chosen to guarantee the power larger than $(1 - \beta)$ given the hypotheses in §14.3.2.1, where β is the desired type II error rate specified by users.

14.3.2.1 Hypothesis

With the unknown effects of three arms, μ_j, $j = 0, 1, 2$, the classical multiple testing problem,

$$H_{0j} : \mu_j \leq \mu_0 \quad \text{vs} \quad H_{1j} : \mu_j > \mu_0, \quad j = 1, 2,$$

is usually handled by controlling the family-wise type I error rate,

$$P\left[\cup_{j \in I}(\text{rejection of } H_{0j})\right],$$

where $I \subseteq J = \{1, 2\}$ is the subset of true H_{0j}, that is, $I = \{j \mid \mu_j \leq \mu_0, j = 1, 2\}$.

However, one does not test all hypotheses in the final analysis. Once a treatment arm is selected in the interim analysis at the end of Stage 1, of interested is the selected treatment arm, $\hat{t}$, and then one may test the following hypothesis at the end of Stage 2,

$$H_{0\hat{t}} : \mu_{\hat{t}} \leq \mu_0 \quad \text{vs} \quad H_{1\hat{t}} : \mu_{\hat{t}} > \mu_0.$$

Hence, the type I error rate is the probability of wrongly rejecting the null hypothesis, which means,

$$\sum_{j \in I} P(\text{rejection of } H_{0\hat{t}} \mid \hat{t} = j) P(\hat{t} = j)$$

and the power is the probability of correctly rejecting the null hypothesis, which means,

$$\sum_{j \in J \setminus I} P(\text{rejection of } H_{0\hat{t}} \mid \hat{t} = j) P(\hat{t} = j).$$

After the second stage, estimate the treatment effects of the control arm and selected treatment arm based on all data of two stages, i.e.,

$$\hat{\mu}_j = \bar{y}_{\cdot j\cdot} = \frac{1}{n_1 + N_2}\left(n_1\bar{y}_{\cdot j1} + N_2\bar{y}_{\cdot j2}\right), \quad j \in \{0, \hat{t}\}.$$

Use ξ as the test statistic,

$$\xi = (\hat{\mu}_{\hat{t}} - \hat{\mu}_0)\sqrt{\frac{N}{2S_1^2}} = (\overline{y}_{\cdot\hat{t}\cdot} - \overline{y}_{\cdot 0\cdot})\sqrt{\frac{N}{2S_1^2}},$$

where $N = n_1 + N_2$ and the variance is based on data from the first stage only. This approach was first proposed by Stein (1945). In this case one can change u to control type I error rate according to the rule,

$$\text{reject } H_0 \Leftrightarrow \xi > u.$$

14.3.2.2 Optimal u and v

With pre-specified type I and II error rates, α and β, the optimal procedure in Bischoff and Miller (2009) is to find the smallest u to control the type I error rate at α, and then find the smallest v to maintain the power to be at least $(1-\beta)$. An algorithm to determine u and v, and an optimal stage 1 sample size is given as follows.

Optimal u

Given $n_1 \in \mathbb{Z}^+$, $n_{2,\min} \in \mathbb{N}$, and $\forall\sigma^2 > 0$, the optimal critical value u is defined as the smallest one with the type I error rate less than or equal to α, that is,

$$\sum_{j\in I} P(\xi > u \mid \hat{t} = j)P(\hat{t} = j) \le \alpha,$$

where $I \subseteq J = \{1, 2\}$ is the subset of true H_{0j}. Then the optimal u is the solution of the equation

$$\int_0^\infty \int_{-\infty}^\infty \left(1 - \Phi\left(a\sqrt{\frac{n_1 + 2n_{2,\min}}{n_1}} + u\sqrt{\frac{2w(n_1 + n_{2,\min})}{(3n_1 - 3)n_1}}\right)^2\right) \times \phi(a) f_{3n_1-3}(w)dadw = \alpha, \tag{14.5}$$

where $\phi(x)$ and $f_{3n_1-3}(x)$ are the probability density functions of $N(0,1)$ and $\chi^2_{3n_1-3}$ distributions, separately, and $\Phi(x)$ is the cumulative distribution function of $N(0,1)$.

Considering that the type I error rate decreases with increasing u, the optimal u is approximated by the bisection method using (14.5) with a range of in [0,10] in East Bayes. If no optimal u

can be identified in [0,10], error messages will be reported in the simulation results.

Optimal v

Given $n_1 \in \mathbb{Z}^+$, $n_{2,\min} \in \mathbb{N}$, a fixed σ^2, and $\max\{\mu_1 - \mu_0, \mu_2 - \mu_0\} \geq \Delta$, the optimal v is defined as the smallest one with the power larger than $(1-\beta)$, that is

$$\sum_{j \in J \setminus I} P(\xi > u \mid \hat{t} = j) P(\hat{t} = j) \geq 1 - \beta,$$

which means that the optimal v will guarantee a power of at least $(1-\beta)$. If the true effects of two treatment arms are both better than that of the control arm, i.e., $\mu_1 > \mu_0$ and $\mu_2 > \mu_0$, then $I = \emptyset$ and the power is $\sum_{j \in 1,2} P(\xi > u \mid \hat{t} = j) P(\hat{t} = j)$.

Let $S_1^2 = \frac{\sigma^2}{3n_1 - 3} w$ and then

$$N_2(S_1^2) = m_2(w, \sigma^2) = \lceil \max\left(v \frac{\sigma^2}{3n_1 - 3} w - n_1, n_{2,\min} \right) \rceil.$$

Here, σ^2 denotes the unknown true variance of treatment effects of three arms and w follows the $\chi^2_{3n_1-3}$ distribution. When treatment arm 1 is selected to be tested in Stage 2, i.e., $\hat{t} = 1$, for fixed σ^2, we have

$$\begin{aligned} &P(\xi > u, \hat{t} = 1) = P(\xi > u \mid \hat{t} = 1) P(\hat{t} = 1) \\ &= \int_0^\infty \int_{-\infty}^\infty \Phi\left((\mu_1 - \mu_0) \frac{n_1 + m_2(w, \sigma^2)}{\sqrt{\sigma^2 (n_1 + 2 m_2(w, \sigma^2))}} + a \sqrt{\frac{n_1}{n_1 + 2 m_2(w, \sigma^2)}} \right. \\ &\left. - u \sqrt{\frac{2w(n_1 + m_2(w, \sigma^2))}{(3n_1 - 3)(n_1 + 2 m_2(w, \sigma^2))}} \right) \Phi\left(a + (\mu_1 - \mu_2) \sqrt{\frac{n_1}{\sigma^2}} \right) \phi(a) f_{3n_1-3}(w) da dw \end{aligned} \tag{14.6}$$

where $f_{3n_1-3}(x)$ is the probability density function of the $\chi^2_{3n_1-3}$ distribution. The probability $P(\xi > u, \hat{t} = 2)$ is given by interchanging μ_1 and μ_2 in (14.6).

Considering the effect of at least Δ over the control arm, without loss of generality, let $\mu_0 = 0$, $\mu_1 \in [0, \Delta]$, and $\mu_2 = \Delta$. It is easily checked that for $\mu_0 = 0$ and $\mu_2 = \Delta$ there exist a different $\mu_1 \in [0, \Delta]$ with equal power. To guarantee the power for all $\mu_1 \in [0, \Delta]$ larger than or equal to $(1-\beta)$, we compute the power for μ_1 in a finite and discrete subset of $[0, \Delta]$ by numerical integration using (14.6) and determine the minimal power over all μ_1. Then for fixed n_1, $n_{2,\min}$ and σ^2, we determine the smallest v such that the minimal power is larger than or equal to $(1-\beta)$.

Since a larger v leads to a larger sample size at Stage 2 and a higher power, the optimal u is approximated by the bisection method in a default range in East Bayes. If the optimal v can not be found in the default range, error messages will be reported in the simulation results. For the point mass prior of σ^2 specified by users, we guarantee the power only for the maximal value of the prior.

14.3.2.3 Optimal n_1

For a fixed σ^2 the expected number of patients for the whole trial is

$$r_{\sigma^2}(n_1) = E_{\sigma^2}(3n_1 + 2N_2) = 3n_1 + 2\int_0^\infty m_2(w, \sigma^2) f_{3n_1-3}(w) dw.$$

Given a prior π for the unknown parameter σ^2, the expected number of patients for the whole trial is

$$E\left(r_{\sigma^2}(n_1)\right) = 3n_1 + 2\int\int_0^\infty m_2(w, \sigma^2) f_{3n_1-3}(w) dw \pi(d\sigma^2).$$

In East Bayes, one may set a point mass prior for σ^2 with up to 10 possible values. Only the optimal n_1 with the minimal expected total sample size will be used for simulation.

Estimation Error of Choosing an Inferior Treatment

For the true better treatment arm t^* and a fixed σ^2, if $\mu_{t^*} - \mu_j \geq \Delta$, $j \neq 0, t^*$, the probability of selecting the inferior treatment arm j to enter the second stage, $P(\hat{t} = j)$, is smaller than or equal to γ if and only if $n_1 \geq n_{1,\text{min}}$, where

$$n_{1,\text{min}} = \lceil 2\frac{max\{\sigma^2\}}{\Delta^2}\left(\Phi^{-1}(1-\gamma)\right)^2 \rceil.$$

Part V

Master Protocols

15. Basket Trial Designs

15.1 Introduction

Basket trials are a type of master protocol in which a treatment is evaluated in more than one indications (baskets). For example, a BRAF inhibitor can be tested simultaneously in multiple cancer types all harboring BRAF mutations (Hyman et al., 2015) in a single trial (NCT01524978), as opposed to multiple trials each of which focusing on a single cancer type. Empowered by breakthroughs in genomics, complex diseases like cancer are further subdivided by biomarkers in addition to the histology, paving the foundation for complex studies like basket trials. In essence, a basket trial is a multi-arm phase 2 or phase 3 study investigating a treatment for multiple diseases or sub-diseases, and basket trials are usually without randomized control. Here and hereinafter, we use the terminology "basket" or "arm" to represent a group of patients with the same disease type or subtype that are treated by the same drug or drug combination in a multi-arm intervention trial.

Usually, each arm in a basket trial is compared with a historical control. Patients enrolled in a basket trial are often composed of a heterogeneous group across multiple indications, such as different cancer types. Therefore, it is difficult to evaluate time-to-event endpoints (e.g., progression-free survival (PFS) or overall survival (OS)), and the primary endpoints in a basket trial is often response rates (e.g., objective response rate (ORR) or pathological complete response (pCR)), which are less sensitive to the effects of population heterogeneity.

In screening new treatments, there might be a scientific rationale to expect some degree of similarity in treatment effect across arms. There exists two common approaches as to whether or not borrow information in the design and analysis of trial trial data: pooled analysis and independent analysis. If the treatment effect is assumed homogeneous across different baskets, a pooled analysis may be preferred, in which the data across all the arms are combined. However, the homogeneity assumption often fails in practice. For example, in BRAF V600 study, while BRAF V600E-mutant melanoma and hairy cell leukemia are responsive to BRAF inhibition, BRAF-mutant colon cancer

is not (Flaherty et al., 2010; Tiacci et al., 2011; Prahallad et al., 2012). When the homogeneity assumption is not valid, a separate stand-alone analysis for each arm is a simple alternative. However, conducting an independent evaluation in each arm is time- and resource-consuming. Also, the trial sample size may be inflated under independent arms when compared to designs that borrow information. Recently, adaptive designs that borrow information via model-based inference have been proposed, such as works in (Thall et al., 2003; Berry et al., 2013; Neuenschwander et al., 2016; Simon et al., 2016; Cunanan et al., 2017; Liu et al., 2017; Chu and Yuan, 2018a,b; Hobbs and Landin, 2018; Psioda et al., 2019). Using the observed data, these methods borrow information by prior distributions that shrink the arm-specific estimates to a centered value.

In East Bayes, we implement a module of **Basket Trial Designs** and use simulation-based power calculation to evaluate four Bayesian approaches, including the Bayesian hierarchical model (BBHM) proposed by Berry et al. (2013), the calibrated Bayesian hierarchical model (CBHM) by Chu and Yuan (2018a), the exchangeabilitynonexchangeability (EXNEX) method in Neuenschwander et al. (2016) and a novel multiple cohort expansion (MUCE) method in Lyu et al. (2020). Users may choose a desirable designs based on provided software in this module.

15.2 User Interface and Tutorial

15.2.1 Overview

Entering the **Basket Trial Designs** page, users will see two main tabs: **Simulation Setup** and **Simulation Results**. These two tabs allow users to conduct simulations and visualize/download simulation results. In the **Simulation Setup** tab, there are three steps (Figure 15.1): 1) **Set trial parameters**, 2) **Select designs**, and 3) **Generate scenarios**. Users need to complete the steps 1-3 to set up simulations for a single design or multiple designs. Upon completing steps 1-3, users click the "Launch Simulation" button at the bottom of the page. Users may also click the "Reset" button next to **Launch Simulation** to clear all the settings. After the simulations are launched, the results of simulations will be displayed in the **Simulation Results** tab. The simulation process can be monitored in real time at the top of the **Simulation Results** tab. Detailed steps of using this module are elaborated next in §15.2.2-§15.2.3.

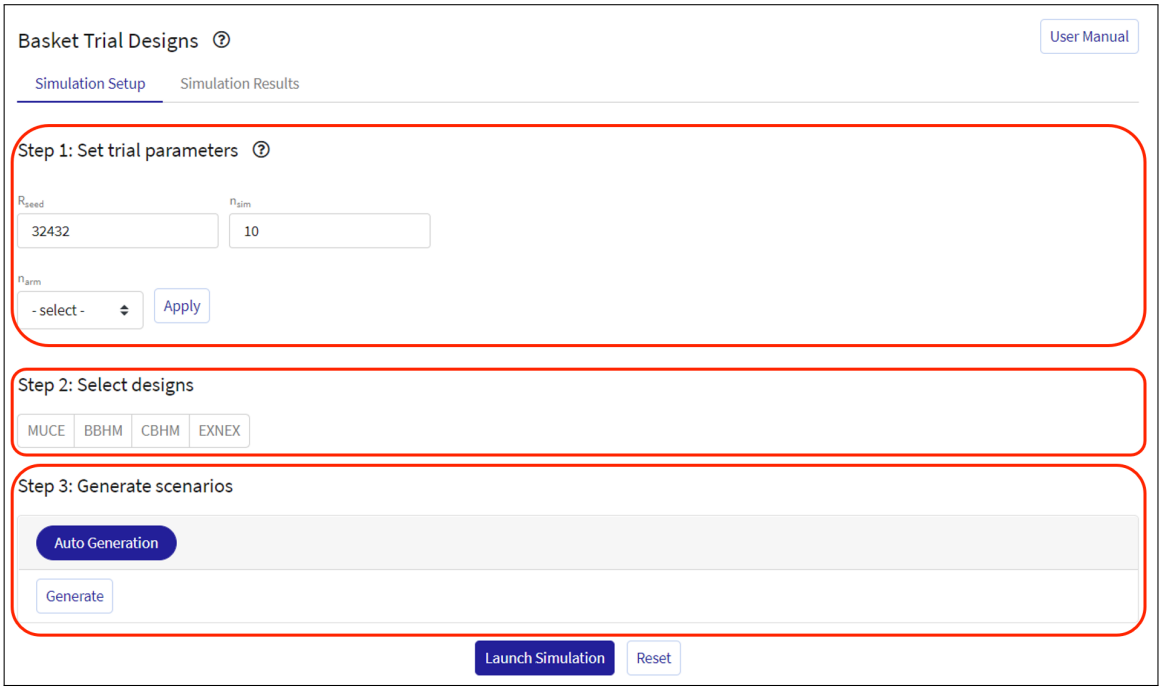


Figure 15.1: Simulation Setup in the **Basket Trial Designs** module.

15.2.2 Simulation Setup

In the **Basket Trial Designs** module, East Bayes provides four designs, BBHM, CBHM, EXNEX, and MUCE, for simulation. Users can choose up to four design configurations for simultaneous comparison in the **Simulation Setup** tab each time. A design configuration means a design such as MUCE, along with the designs settings, such as sample size. Request to allow more than four design configurations by emailing support@cytel.com.

15.2.2.1 Step 1: Set trial parameters

Specify the number of simulated trials (n_{sim}) and the random seed of simulation (R_{seed}). Then select a number of arms ($n_{arm}, 2 \leq n_{arm} \leq 10$) from the dropdown box. Upon selection, manually type in the reference response rate (R_{ref}), the target response rate (R_{target}), and the type I error rate (α) for each arm. See Figure 15.2.

Click the "Reset" button to clear all the settings. Users may click the  icon (right after the cell of Arm 1) to copy and paste the value of Arm 1 into other arms.

Hover mouse over the question mark icon, and a description will be displayed explaining the meaning of the parameters. The detailed description of the above six input arguments is provided in Table 15.1.

Click the "Apply" button in Figure 15.2 to confirm and submit the trial parameters. And click the "Edit" button to enable the edits.

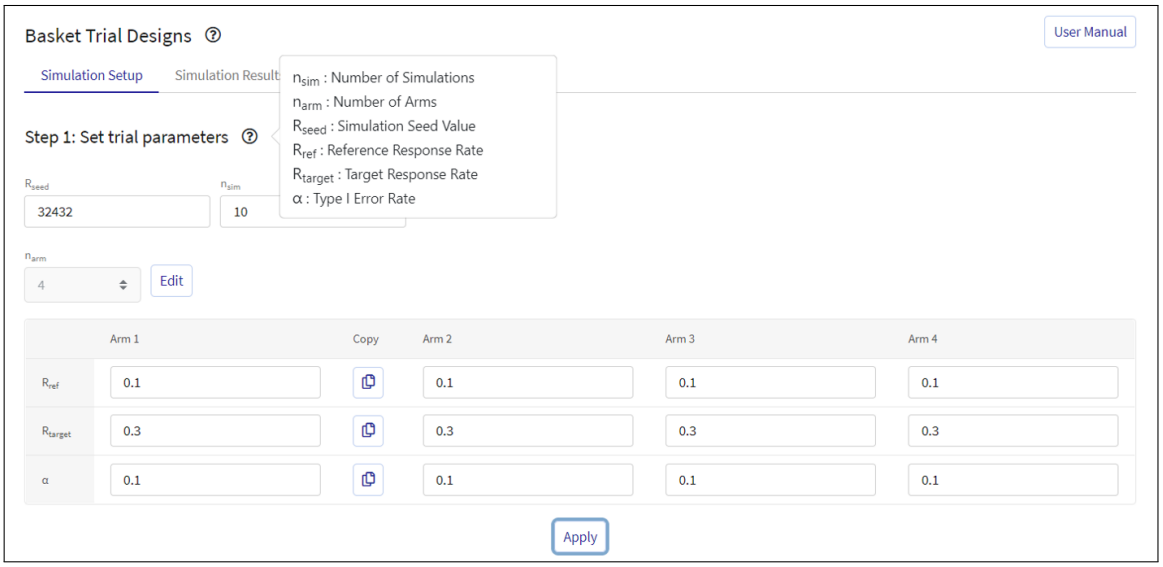


Figure 15.2: Set trial parameters in the **Basket Trial Designs** module.

Table 15.1: Input parameters for trials in the **Basket Trial Designs** module.

Notation	Parameters	Description
n_{sim}	Number of simulated trials	The number of simulated trials to be conducted for each scenario. The maximum number allowed is 10,000. Default value is 1,000.
R_{seed}	Random seed of simulation	A number used to initialize a pseudorandom number generator in the simulation. Default value is 32432.
n_{arm}	Number of arms	The number of arms in the trial. The range is $[2, 10]$.
R_{ref}	Reference response rate	The reference response rate (also called the historical control rate) is the largest rate considered to be not promising. Default value is 0.1.
R_{target}	Target response rate ($R_{target} > R_{ref}$)	The target response rate is the smallest rate considered to be promising. Default value is 0.3.
α	Type I error rate	The probability of rejecting null when the null hypothesis is true. Default value is 0.1.

15.2.2.2 Step 2: Select designs

To select a design, click the button with the design's name on it. Up to four design configurations may be selected for comparison. Upon selection of a design, specify the maximum sample size for each arm (n), interim analysis parameters, and when needed, advanced design parameters. See Figure 15.3.

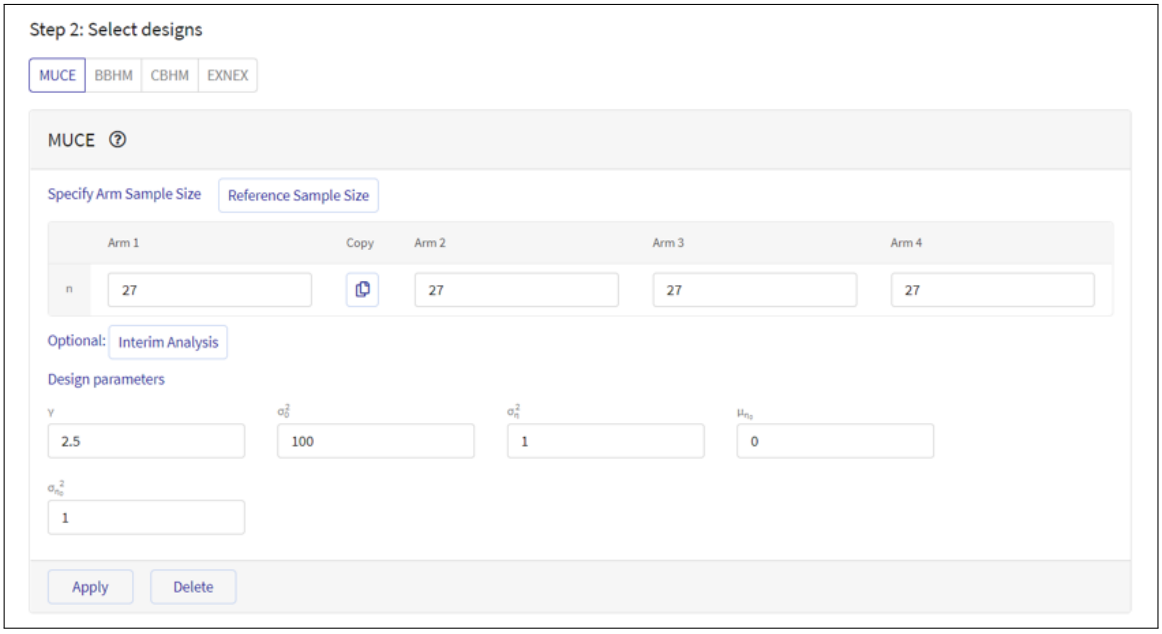


Figure 15.3: Select designs in the **Basket Trial Designs** module.

Specify arm sample size

East Bayes provides a function to facilitate sample size specification. It generates "reference sample size" as candidate for simulations. Users can first try the reference sample size, generate simulation results, calibrate the sample size based on the results, and finally decide an appropriate sample size. Click the "Show Reference Sample Size" button in Figure 15.3 to expand the reference sample size section (Figure 15.4). East Bayes provides three sets of sample sizes under power $(1-\beta)$ of 70%, 80% and 90%, respectively, which are calculated by the one-sided equality Z-test with the standard deviation based on the target rate for one-sample proportion, $n = \frac{(Z_\alpha + Z_\beta)^2 R_{target}(1-R_{target})}{(R_{target}-R_{ref})^2}$. Users can also manually type in a different power value and click the icon button to obtain a new reference sample size. These numbers can be used to help users to provide the maximum sample size for each arm. By clicking the icon (at the end of each row), the sample sizes in the corresponding row will be loaded as the required maximum sample size.

Click the "Hide the reference sample size" button to hide the reference sample size section. Similar in Step 1, users may click the icon right after the cell of Arm 1 to copy and paste the sample size of Arm 1 into other arms.

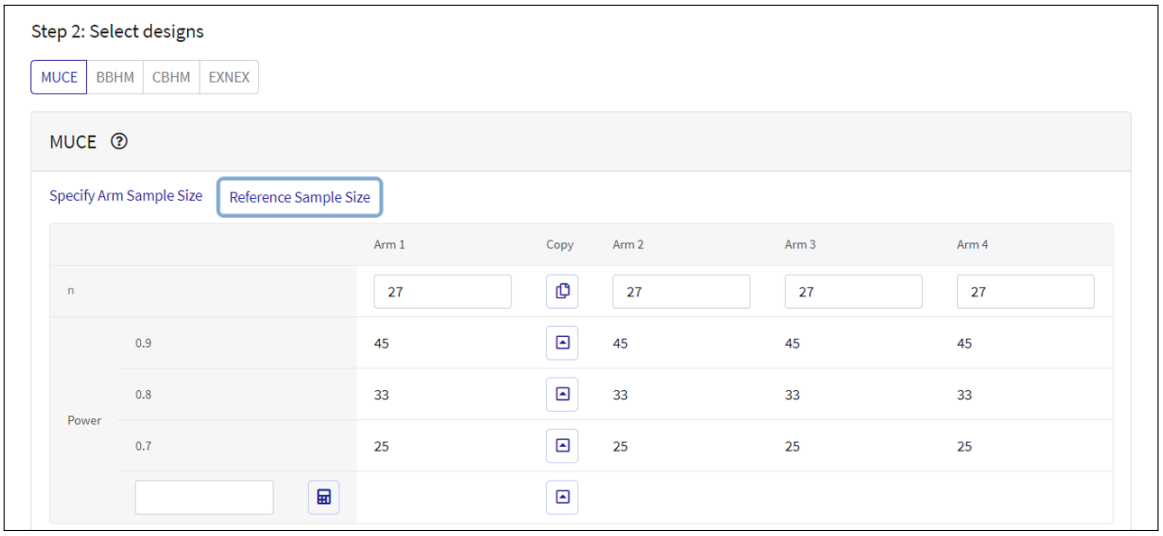


Figure 15.4: Display the reference sample size in **Step 2: Select designs** in the **Basket Trial Designs** module.

Interim analysis (optional)

Check the box behind the **Optional: Interim Analysis** in Figure 15.3 to expand the section of interim analysis parameters specification. Using the enrollment speed (S_{enroll}) of Arm 1 as a benchmark, users can manually type in the enrollment speeds for other arms that are relative to Arm 1. A value greater or less than 1 means a faster or slower patients accrual than Arm 1, respectively. And users can specify the probability threshold of futility stopping ($P_{futility}$) for interim analysis.

When checked, two interim analyses will be applied by default. There are two possibilities. First, if all the arms are assumed to take the same amount of time to enroll the total number of patients (arm sample size) and the speed of enrollment is constant, the first interim analysis is performed when each arm enrolls half (50%) of the sample size of the arm, and the second time is when each arm enrolls 75% of the total sample size. Otherwise, the first interim is conducted when the fastest arm enrolls half of the sample size of the arm, and the second interim is conducted when the slowest arm enrolls half of the sample size of the arm. For example, for a three-arms basket trial with the maximum sample size set at $(40, 80, 20)$ for three arms, if the enrollment speed is $S_{enroll} = (1, 2, 0.5)$, the enrollment time of all three arms are the same. Assuming a constant

enrollment speed, the two interim analyses will be performed when three arms enroll $(40 \times 0.5, 80 \times 0.5, 20 \times 0.5) = (20, 40, 10)$ patients and $(40 \times 0.75, 80 \times 0.75, 20 \times 0.75) = (30, 60, 15)$ patients, respectively; if the enrollment speed is $S_{enroll} = (1, 4, 0.75)$, two interim analyses will be performed when the fastest arm enrolls half patients (Arm 2) and the slowest arm enrolls half patients (Arm 1), which result in sample sizes $(10, 40, 3)$ for interim 1 and $(20, 80, 15)$ for interim 2. Request to allow other interim analysis options by emailing support@cytel.com.

Design parameters

The default values of advanced design parameters are recommended. See detailed explanation of each parameter in §15.3 next.

Click the "Apply" button in Figure 15.3 to confirm and submit the trial parameters. Click the "Edit" button to enable the edit mode and all design parameters can be modified. Click the "Delete" button to remove the selected designs.

Hover mouse over the question mark icon next to the design name, and a description will be displayed explaining the meaning of the parameters of this design. The detailed description of the above input arguments is provided in Table 15.2 below.

Table 15.2: Input parameters for designs in the **Basket Trial Designs** module.

Notation	Parameters	Description
n	Maximum sample size	The maximum number of patients to be treated in the trial for each arm. The value is an integer between $(0, 1000]$.
S_{enroll}	Relative enrollment speed	The enrollment speed relative to Arm 1. The range is $(0, \infty)$. Default value is 1 for all arms, which means all arms have the same enrollment speed. A value of 0.5 means the arm enrolls half of the speed of Arm 1, whatever it is.
$P_{futility}$	Futility stopping threshold	The probability threshold of futility stopping at an interim analysis. See stopping criteria in §15.3. Default value is 0.1.

15.2.2.3 Step 3: Generate scenarios

There are two ways to generate scenarios, automatically (in below **Auto Generation** tab, see Figure 15.5) or through manual construction, see Figure 15.6.

Auto Generation (Figure 15.5)

Click the "Generate" button to automatically create three to six scenarios, each of which contains the true response rates for n_{arm} arms. Scenario 1 is a global null scenario in which all arms are not promising with the response rate set at the reference response rate R_{ref}. Scenario 2 is a global alternative scenario in which all arms are promising with the response rate set at the target response rate R_{target}. Other scenario(s) are mixed scenarios with some but not all arms promising.

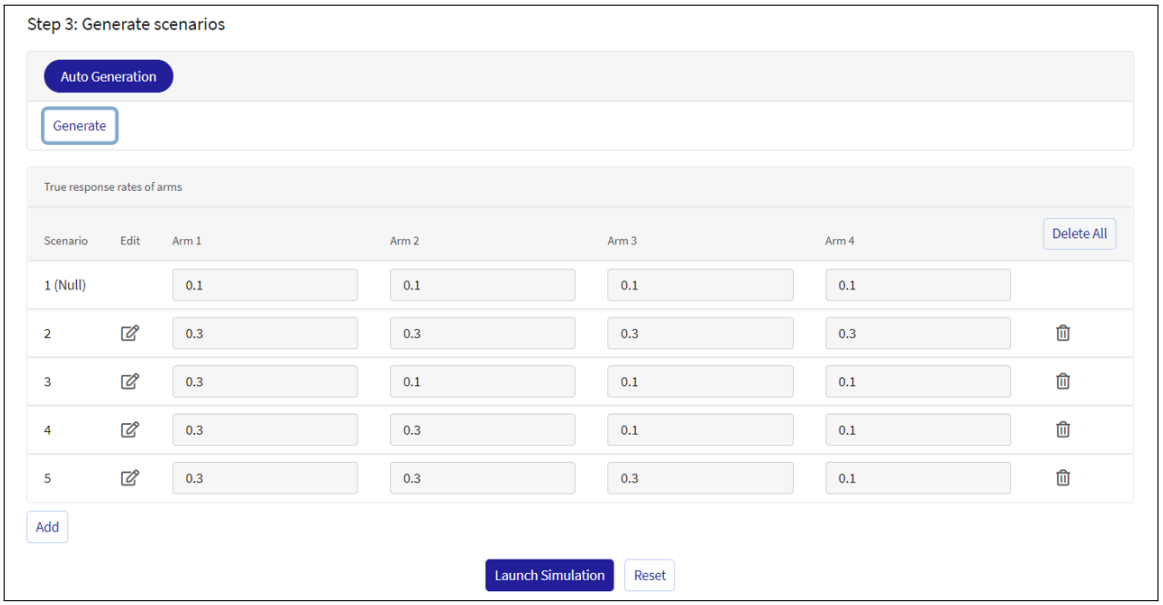


Figure 15.5: Automatically generate scenarios in the **Basket Trial Designs** module.

Manual Construction (Figure 15.6)

Click the "Add" button to create a new scenario. The format of input must be numeral between 0 and 1, each representing the true response rate of each arm. After completing the input, click the icon button to confirm it.

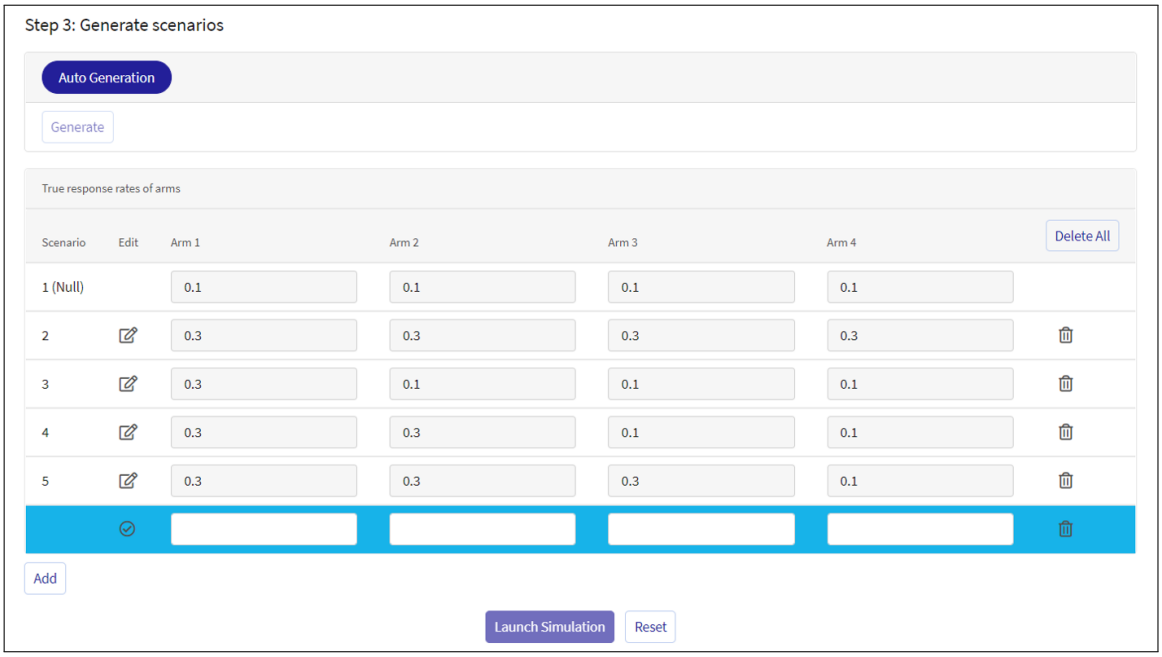


Figure 15.6: Manually generate scenarios in the **Basket Trial Designs** module.

The generated scenarios are displayed as a list (Figures 15.5 and 15.6) which appears below the generation section. Click the icon to edit the corresponding scenario.

Click the icon (at the end of each row) to delete the corresponding scenario. The first (Null) scenario is always included in order to benchmark designs. Click the "Delete All" button to delete all scenarios (including the Null scenario).

15.2.2.4 Launch Simulation

Once the steps 1-3 are completed, users can conduct simulated clinical trials to examine the operating characteristics of the selected designs using the selected scenarios. Click the "Launch Simulation" button at the bottom of **Simulation Setup** tab (Figures 15.5 and 15.6). A "**Success**" message will be displayed on the screen (Figure 15.7) to indicate that the simulations have been successfully launched. Users may click the "OK" button in the pop-up box to track the simulation processing status and simulation results.

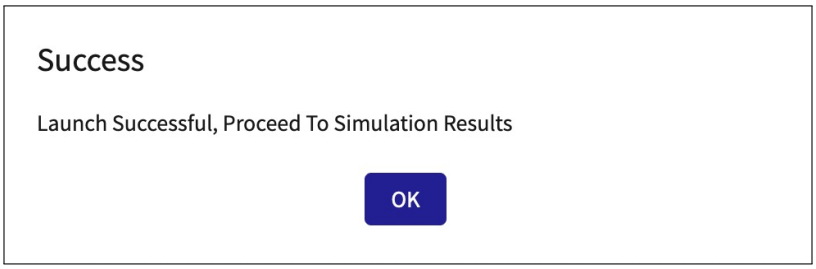


Figure 15.7: "**Success**" message after launching simulation in the **Basket Trial Designs** module.

15.2.3 Simulation Results

In the **Simulation Results** tab, users can view the simulation progress and simulation results (§15.2.3.1), restore the simulation settings if needed (§15.2.3.2), and download East Bayes's proprietary intelligent simulation reports (§15.2.3.3). Specifically, all the simulation results (figures and tables) can be downloaded in Word format, accompanying the statistical sections in a trial protocol. Hereinafter, we use simulation results and operating characteristics interchangeably.

15.2.3.1 View simulation results

In the **Simulation Results** tab, the **Running Simulations** panel exhibits the progress of ongoing simulation (Figure 15.8). The ongoing simulations are displayed in ascending order by the launch time. Click the icon "×" to delete the corresponding simulation.

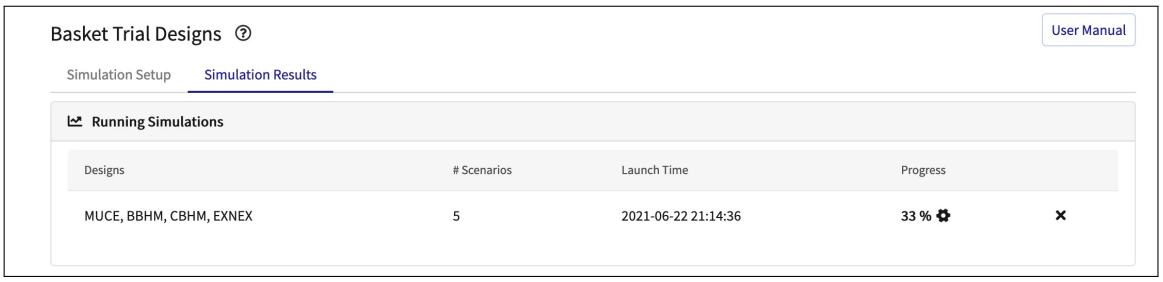


Figure 15.8: Simulation progress in the **Basket Trial Designs** module.

Once the simulations are completed, the **Running Simulations** panel in Figure 15.8 will disappear, green "*simulation result created*" massages will appear instead and stay at the same place of the **Running Simulations** panel unless explicitly dismissed by clicking the icon "×" at the end of the corresponding row, and the simulation results will be automatically loaded into the **Simulation History** panel (Figure 15.9), with the blue mail icon shown to indicate new results. All the previously completed simulations are also listed in the **Simulation History** panel. Simulation results for other modules can also be viewed under the **Simulation History** by dropping down the "Select a Design Category" button (Figure 15.9). Click the button to delete the selected simulation results.

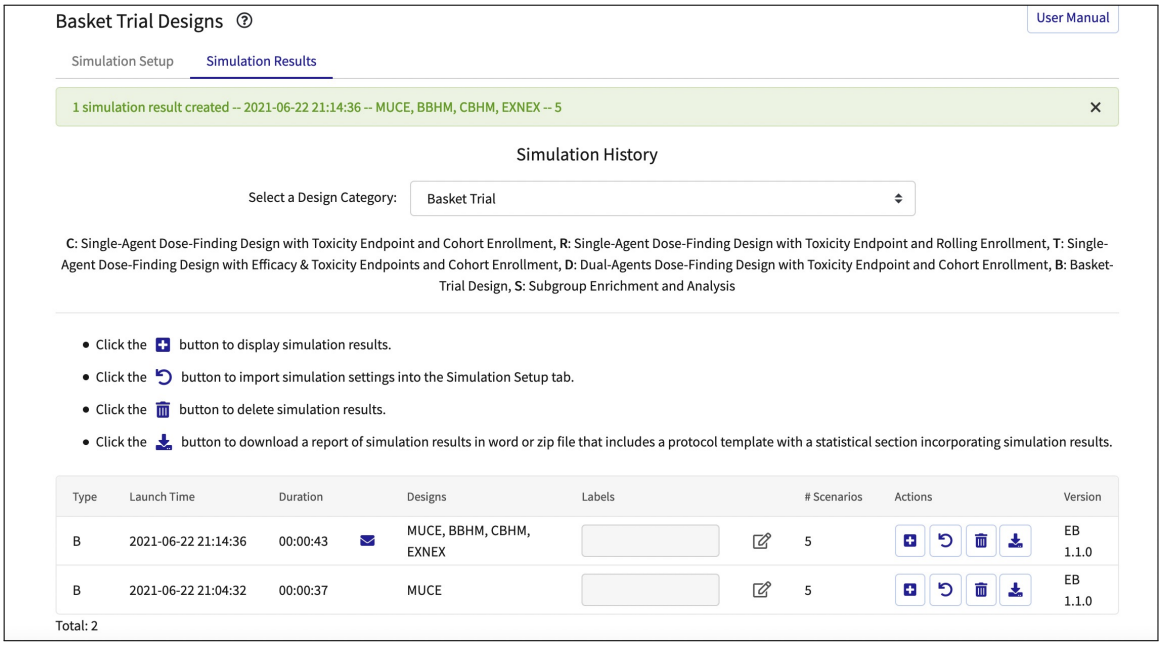


Figure 15.9: Simulation Results in the **Basket Trial Designs** module.

Click the ⊞ button to unfold the simulation results (Figure 15.10). The design settings are firstly displayed at the top of each simulation study. Then the results of simulation are shown in two ways: figures and tables. See next.

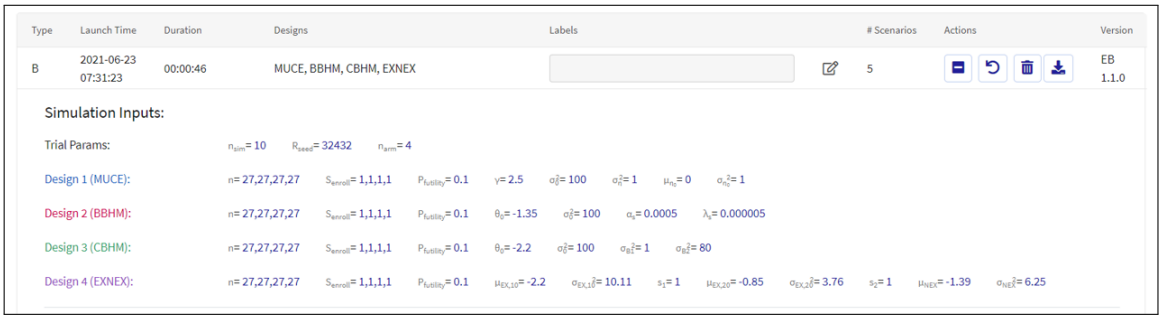


Figure 15.10: View the simulation results in the **Basket Trial Designs** module.

Details of the Simulation Results

The simulation results are divided into two parts, i.e, Simulation Result Summary and Tabulated Results by Scenarios. Each part can be viewed or hidden by clicking the button for that part (Figure 15.11).

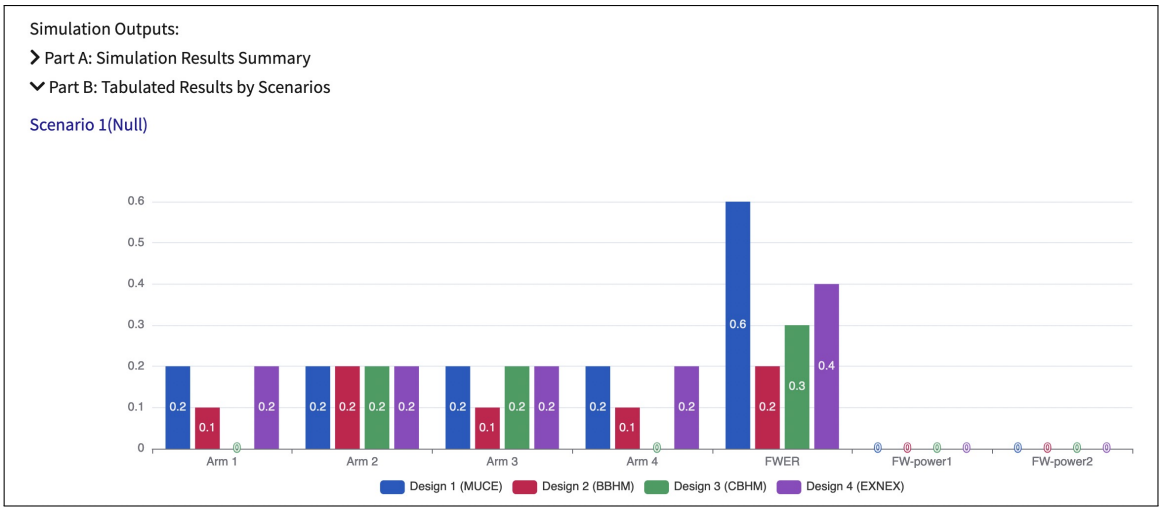


Figure 15.11: View each part of the simulation results in the **Single-Agent Dose-Finding Designs with Toxicity Endpoint and Cohort Enrollment** module.

Part A: Simulation Results Summary (Figures 15.12 and 15.13)

There are two sections in the Simulation Results Summary.

1. Line plots showing three frequentist summary statistics of the simulation results for all the designs from two aspects: **Family-wise Type I Error Rate** and **Family-wise Power** (Figure 15.12).
    - The three frequentist summary statistics are explained in full detail next.
        - **Family-wise Type I Error Rate:** The proportion of simulated trials in which any true null is rejected, i.e., any false discovery is made. In other words, it is the proportion of simulated trials in which any arm is wrongly declared to be more efficacious than historical controls.
        - **Family-wise Power:** Two subtypes of powers are considered.
            * **Family-wise Power 1:** The proportion of simulated trials in which only true efficacious arms are correctly declared to be more efficacious than the histor-

ical controls, and no true inefficacious arms are wrongly declared to be more efficacious than the historical controls.

  - **Family-wise Power 2:** The proportion of simulated trials in which all true efficacious arms are correctly declared to be more efficacious than the historical controls, and no true inefficacious arms are declared to be more efficacious than the historical controls.

- For each line plot, the x-axis is the index of scenario and the y-axis is the value of summary statistics. Lines with different colors represent different designs.
- The plots are interactive for better visualization.
  - Hover the mouse on a dot and a box will display the value of each design at the corresponding scenario (e.g. top left plot in Figure 15.12: Family-wise Type I Error Rate).
  - Hover the mouse on the design label to highlight the corresponding line and fade the others (e.g. top right plot in Figure 15.12: Family-wise Power 1).
  - Click the design label to hide the corresponding line and click again to change it back (e.g. bottom left plot in Figure 15.12: Family-wise Power 2).

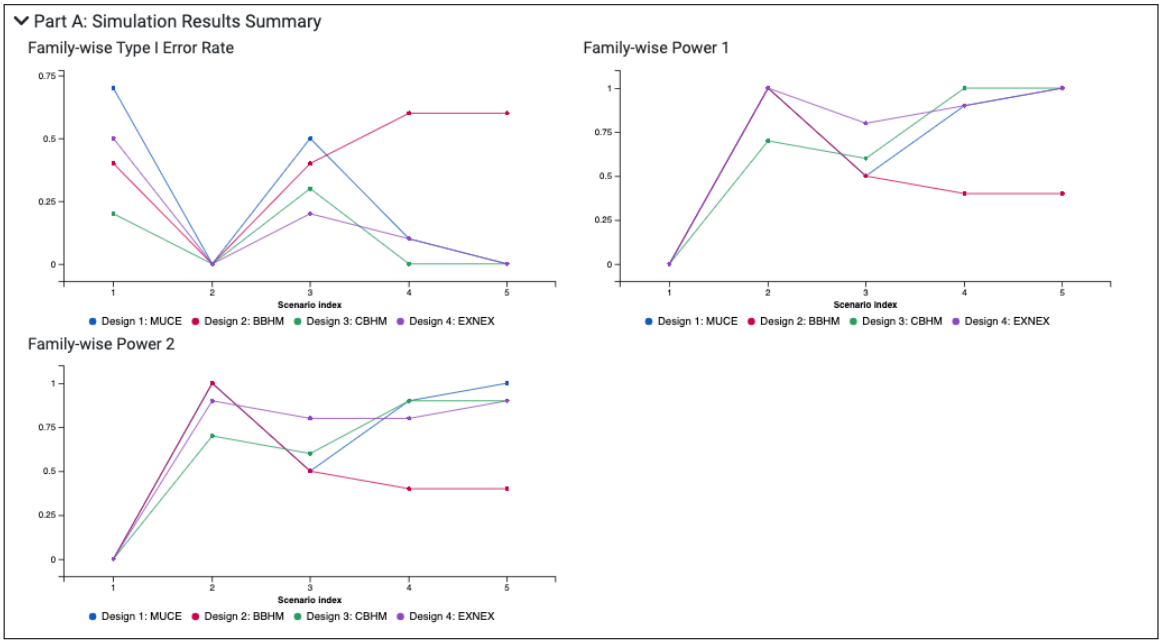


Figure 15.12: Simulation result plots in the **Basket Trial Designs** module.

2. A table showing trial settings and probability thresholds used in the final analysis for all designs (Figure 15.13).

    - The table shows the trial parameters specified in step 1 (§15.2.2) and the probability thresholds for the rejection of null in the final analysis for all the selected designs. The trial parameters displayed include the reference response rate (R_{ref}), the target response rate (R_{target}), and the type I error rate (α), for each arm.

Trial Settings and Probability Thresholds for Final Analysis

Arm	R_{ref}	R_{target}	α	Probability thresholds for rejection of Null at the final analysis			
				Design 1 (MUCE)	Design 2 (BBHM)	Design 3 (CBHM)	Design 4 (EXNEX)
1	0.1	0.3	0.1	0.878	0.967	1	0.795
2	0.2	0.4	0.1	0.736	0.754	0.849	0.87
3	0.2	0.3	0.1	0.866	0.323	0.734	0.838
4	0.1	0.4	0.1	0.896	0.998	1	0.796

Figure 15.13: Trial settings and probability thresholds for the final analysis in the **Basket Trial Designs** module.

Part B: Tabulated Results by Scenarios (Figure 15.14)

Full simulation results are presented in bar plots and tables arranged by scenario (Figure 15.14). For each scenario, the simulation results are summarized from the following three frequentist aspects.

1. **Type I error rate / Power:** A bar plot showing the arm-wise type I error rate & power and family-wise type I error rate & power (FWER & FW-power).
   - Bars with different colors represent different designs.
   - The first n_{arm} clusters of bars report the arm-wise type I error rate & power, and the last three clusters report the FWER and two family-wise powers.
   - Four statistics are explained in detail next.
     - **Arm-wise type I error rate & power:** The proportion of simulated trials in which the null hypothesis for an arm is rejected, i.e., the proportion of simulated trials in which the arm is declared to be more efficacious than the historical control. This is the arm-wise type I error rate if the arm is actually not more efficacious than the historical control in this arm, and is the arm-wise power otherwise.
     - **Family-wise type I error rate & power (FWER & FW-power)**
       - **Family-wise type I error rate (FWER):** The proportion of simulated trials in which at least one arm is wrongly declared to be more efficacious than historical controls in any arm.
       - **Family-wise power 1 (FW-power1):** The proportion of simulated trials in which only true efficacious arms are correctly declared to be more efficacious than the historical controls, and no true inefficacious arms are wrongly declared to be more efficacious than the historical controls.
       - **Family-wise power 2 (FW-power2):** The proportion of simulated trials in which all true efficacious arms are correctly declared to be more efficacious than the historical controls, and no true inefficacious arms are declared to be more efficacious than the historical controls.

     For detailed descriptions, please refer to **Simulation Results Summary** above.
2. **Response Rate Estimation:** A table is provided (Figure 15.14) reporting the accuracy and the precision of the estimates of response rates. The first two columns summarize the scenario settings, with the index and its true response rate of each arm; the subsequent columns report the average bias of response rate estimates and their standard deviation. The bias is defined as the difference between the posterior mean of response rate and the true response rate. The average is taken across all the simulated trials.

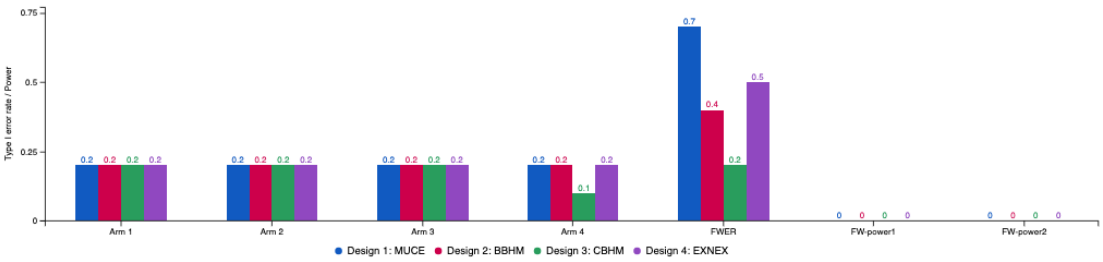
3. **Interim Analysis:** A table is provided (Figure 15.14) summarizing the statistics of interim analysis, if any.
    - **Average sample size (s.d.):** The average number of patients treated in a simulated trial and its standard deviation, averaging across all the simulated trials.
    - **Current** # of patients treated: The numbers of patients treated for each arm when the 1st and the 2nd interim analyses are performed, respectively.
    - **Probability of futility stopping:** The proportion of simulated trials in which an arm is stopped early due to futility at the 1st or the 2nd interim analysis.

When calculating the standard deviation, we use n_{sim} as the denominator instead of $(n_{sim} - 1)$ in East Bayes.

Part B: Tabulated Results by Scenarios

Scenario 1(NULL)

Type I error rate / Power



Response Rate Estimation

Arm	True response rate	Average bias (s.d.) of response rate estimates			
		Design 1 MUCE	Design 2 BBHM	Design 3 CBHM	Design 4 EXNEX
1	0.1	0 (0.06)	0.01 (0.03)	0.01 (0.05)	0.02 (0.06)
2	0.1	-0.01 (0.05)	0.01 (0.03)	0.01 (0.03)	0.03 (0.08)
3	0.1	0.01 (0.05)	0.01 (0.03)	0.01 (0.04)	-0.01 (0.06)
4	0.1	0.04 (0.05)	0.01 (0.03)	0.02 (0.05)	0.01 (0.06)

Interim Analysis

Arm	Average sample size (s.d.)				At the 1st interim								At the 2n			
					Current # of patients treated				Probability of futility stopping				Current # of patients treated			
	Design 1 MUCE	Design 2 BBHM	Design 3 CBHM	Design 4 EXNEX	Design 1 MUCE	Design 2 BBHM	Design 3 CBHM	Design 4 EXNEX	Design 1 MUCE	Design 2 BBHM	Design 3 CBHM	Design 4 EXNEX	Design 1 MUCE	Design 2 BBHM	Design 3 CBHM	Design 4 EXNEX
1	24.4 (5.2)	20.6 (5.817)	18.6 (5.869)	24.5 (4.225)	14	14	14	14	0.2	0.4	0.6	0.1	21	21	21	21
2	24.4 (5.2)	20 (5.422)	20 (5.422)	23.8 (5.212)	14	14	14	14	0.2	0.4	0.4	0.2	21	21	21	21
3	25.7 (3.9)	18 (5.254)	19.9 (6.139)	20.5 (6.5)	14	14	14	14	0.1	0.6	0.5	0.5	21	21	21	21
4	27 (0)	20.5 (6.5)	18.6 (5.869)	23.8 (5.212)	14	14	14	14	0	0.5	0.6	0.2	21	21	21	21

Figure 15.14: Simulation results by scenario in the **Basket Trial Designs** module.

15.2.3.2 Restore simulation setup

Users can restore the simulation settings from the simulation results by clicking the button at the upper right corner of each simulation results panel (yellow arrow in Figure 15.15) and the display will switch to the **Simulation Setup** page with the same simulation settings restored. This is useful to restore the old simulation settings for reproducible results.

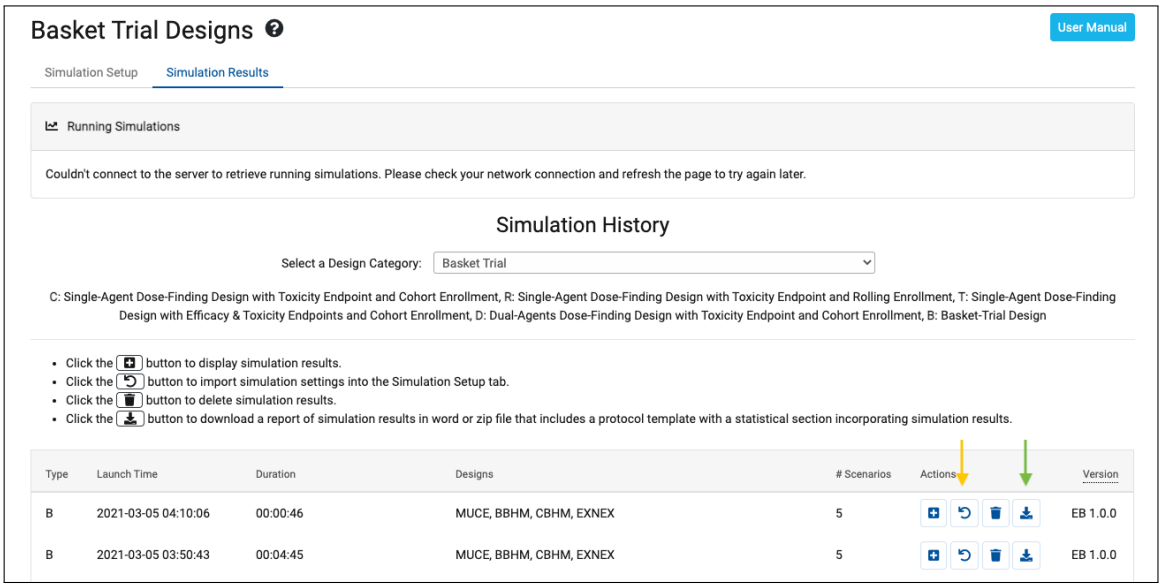


Figure 15.15: Restore simulation setup and download simulation results in the **Basket Trial Designs** module.

15.2.3.3 Download simulation results

A button is placed at the upper right corner of each simulation results panel (green arrow in Figure 15.15). Click it to download East Bayes's proprietary word file with complete simulation results under the designs and scenarios users specified in the simulation settings tab. Users could update the simulation settings and results tailored for their trials. Contact us via email (`support@cytel.com`) for consulting services.

15.3 Statistical Methods Review

15.3.1 Bayesian Hierarchical Model (BBHM)

Berry et al. (2013) apply a Bayesian hierarchical model to phase II basket trial designs that borrows information across arms.

15.3.1.1 Probability Model

Consider a phase II basket trial that evaluates the efficacy of a new treatment in K different arms (indications). Let n_k and y_k denote the number of patients and responders in arm k, respectively. Denote by p_k the true and unknown response rate for arm k. The objective of the trial is to test the null hypothesis that the response rate, p_k, of the arm is less than a reference response rate, π_{k0},

$$H_{0k} : p_k \leq \pi_{k0}$$

versus the alternative hypothesis that the response rate is at least as high as a target rate, π_{k1},

$$H_{1k} : p_k \geq \pi_{k1},$$

for each arm k, $k = 1, 2, \ldots, K$.

BBHM models the log-odds of response rate for each arm k, including an adjustment for the targeted π_{k1} rates, defined as

$$\theta_k = \log\left(\frac{p_k}{1-p_k}\right) - \log\left(\frac{\pi_{k1}}{1-\pi_{k1}}\right).$$

Assume θ_k follow a normal prior distribution with unknown mean θ and variance σ^2

$$\theta_k \mid \theta \overset{iid}{\sim} N(\theta, \sigma^2).$$

The hyperparameters θ and σ^2 are given conjugate hyperpriors,

$$\theta \sim N(\theta_0, \sigma_0^2), \quad \sigma^2 \sim \text{Inv-Gamma}(\alpha_s, \lambda_s),$$

where α_s and λ_s are the shape and scale parameters of the inverse gamma distribution, respectively. This prior construction assumes that the arm-specific treatment effect θ_k's across different arms are exchangeable and shrinks to a shared mean θ, thus enabling information borrowing across arms. The degree of shrinkage or information borrowing is determined by the value of σ^2. The smaller the σ^2, the stronger the borrowing. In the extreme cases, $\sigma^2 = 0$ means all θ_k's equal θ which is

the pooled analysis, and $\sigma^2 = \infty$ is equivalent to the independent approach, where θ_k are assumed independent and distinct.

In short, the hierarchical models are:

$$\begin{aligned}
&\text{Likelihood:} && y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k) \\
&\text{Transformation:} && \theta_k = \log\left(\frac{p_k}{1-p_k}\right) - \log\left(\frac{\pi_{k1}}{1-\pi_{k1}}\right) \\
&\text{Prior for } \theta_k: && \theta_k \mid \theta, \sigma^2 \sim N(\theta, \sigma^2) \\
&\text{Hyperpriors:} && \theta \sim N(\theta_0, \sigma_0^2) \\
& && \sigma^2 \sim \text{Inv-Gamma}(\alpha_s, \lambda_s)
\end{aligned}$$

Following Berry et al. (2013), by default, East Bayes assigns a non-informative inverse gamma prior Inv-Gamma$(0.0005, 0.000005)$ for σ^2, and uses the average of θ_k under the null rates $\theta_0 = \frac{1}{K}\sum_{k=1}^{K}\left(\log\left(\frac{\pi_{k0}}{1-\pi_{k0}}\right) - \log\left(\frac{\pi_{k1}}{1-\pi_{k1}}\right)\right)$ and a large variance $\sigma_0^2 = 10^2$ for the prior of θ, creating a nearly non-informative prior. The inverse gamma prior gives a $E(\sigma^2) = 10^2$ and $Var(\sigma^2) = 2 \times 10^7$.

15.3.1.2 Trial Design

Suppose $L(\geq 0)$ interim looks are planned, and the l-th interim analysis is conducted after n_k^l patients have been enrolled in arm k. Let $\mathcal{D}^l \equiv \{(n_k^l, y_k^l) : k = 1, 2, \ldots, K\}$ denote the observed data at interim analysis l, where y_k^l is the number of responders among the n_k^l patients. Denote $\mathcal{D}^{L+1} \equiv \{(n_k^{L+1}, y_k^{L+1}) : k = 1, 2, \ldots, K\}$ the observed data at the end of the trial, where n_k^{L+1} is the prespecified maximum sample size for arm k and y_k^{L+1} is the total number of responders. The proposed BBHM basket trial design with L interim looks is describe as follows:

1. Enroll n_k^1 patients in k-th arm, $k = 1, 2, \ldots, K$.
2. Given the data $\mathcal{D}^l$ at the l-th interim look, $l = 1, 2, \ldots, L$,
   (a) *[Futility stopping]* If the posterior probability that the response rate of arm k, p_k, is greater than $(\pi_{k0} + \pi_{k1})/2$ is small, i.e.,

$$Pr\{p_k > \frac{\pi_{k0} + \pi_{k1}}{2} \mid \mathcal{D}^l\} < P_{\textit{futility}},$$

   stop the accrual to the k-th arm for futility;
   (b) Otherwise, continue to enroll patients until reaching the next interim analysis.

3. Once the maximum sample size is reached or all the arms have stopped, evaluate the efficacy for each arm based on all the observed data. If the posterior probability that the response rate, p_k, is greater than π_{k0} is large, i.e.,

$$Pr\{p_k > \pi_{k0} \mid \mathcal{D}^{L+1}\} > \phi_k,$$

arm k is declared efficacious and promising; otherwise, it is considered not promising.

Step 2 is optional, since the BBHM design does not require an interim look. However, it is useful to allow interim in practice for early stopping. The probability thresholds for the interim analysis $P_{\textit{futility}}$ and for the final analysis $\{\phi_k : k = 1, 2, \ldots, K\}$, are calibrated through simulations to achieve a prespecified type I error rate for each arm under the global null scenario. In brief, assume n_{sim} trials are simulated under the Null scenario. For arm k, suppose T_k out of n_{sim} trials are early stopped due to futility. From the remaining $(n_{sim} - T_k)$ trials, we can obtain $(n_{sim} - T_k)$ posterior probabilities $p(p_k > \pi_{k0} \mid H_{k0})$. Denote them as $\{P_i = Pr\{p_k > \pi_{k0} \mid \mathcal{D}_i^{L+1}\}, i = 1, \ldots, n_{sim} - T_k\}$, where $\mathcal{D}_i^{L+1}$ is the observed data at the end of i-th trial under the null scenario. Then sort the samples $\{P_i\}$ to obtain a set of order statistics $\{P_{(i)}, i = 1, \ldots, n_{sim} - T_k\}$, where $P_{(i)} \leq P_{(j)}$, for $i < j$. Finally, $\phi_k = P_{(n_{sim} - T_k - n_{sim} \times \alpha_k)}$ so that $n_{sim} \times \alpha_k$ out of n_{sim} trials are rejected under the Null scenario, i.e., the type I error rate is α_k.

15.3.2 Calibrated Bayesian Hierarchical Model (CBHM)

Chu and Yuan (2018a) proposed a calibrated Bayesian hierarchical model (CBHM) as an extension of BBHM, which estimates σ^2 from the observed data instead of using a prior.

15.3.2.1 Probability Model

Consider a phase II basket trial that evaluates the efficacy of a new treatment in K different arms (indications). Let p_k denote the true and unknown response rate for arm k. The objective of the trial is to test whether the new treatment is effective in each of the arms

$$H_{0k} : p_k \leq \pi_{k0} \quad \text{versus} \quad H_{1k} : p_k \geq \pi_{k1}, \quad \text{for } k = 1, 2, \ldots, K,$$

where π_{k0} is the reference response rate (also called the historical response rate), and π_{k1} is the target response rate under which the treatment is regarded as promising.

Suppose at a certain moment, n_k patients from arm k have been enrolled, among which y_k patients respond favorably to the treatment. CBHM assumes that y_k follows a hierarchical model

$$\begin{aligned}
&\text{Likelihood:} && y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k) \\
&\text{Transformation:} && \theta_k = \log\left(\frac{p_k}{1-p_k}\right) \\
&\text{Prior for } \theta_k : && \theta_k \mid \theta, \sigma^2 \sim N(\theta, \sigma^2) \\
&\text{Hyperpriors:} && \theta \sim N(\theta_0, \sigma_0^2)
\end{aligned} \tag{15.1}$$

The same as Berry et al. (2013), the above prior construction assumes that the arm-specific treatment effect θ_k's across different arms are exchangeable and shrinks to a shared mean θ, thereby enabling information borrowing across arms. The degree of shrinkage or information borrowing is determined by the value of σ^2. Following Chu and Yuan (2018a), by default, East Bayes uses the average of θ_k under the null rates $\theta_0 = \frac{1}{K}\sum_{k=1}^{K} \log\left(\frac{\pi_{k0}}{1-\pi_{k0}}\right)$ and a large variance $\sigma_0^2 = 10^2$ for the prior of θ, creating a vague prior.

15.3.2.2 Calibration of shrinkage parameter σ^2

Unlike the BBHM approach (Berry et al., 2013) in §15.3.1, which assigns a prior to σ^2 and estimates it from the data, CBHM defines σ^2 in (15.1) as a function of the measure of homogeneity among the arms. The idea is that the function is prespecified and calibrated in a way such that when the treatment effects across arms are homogeneous, small σ^2 is induced so that strong information borrowing occurs and thus improves power, and when the treatment effects across arms

are heterogeneous, large σ^2 is induced so that little or no borrowing across groups occur, thereby controlling the type I error rate. In what follows, Chu and Yuan (2018a) use a homogeneity measure to determine and calibrate the estimation of parameter σ^2.

Specifically, CBHM adopts the chi-squired test statistic to measure homogeneity, given by

$$T = \sum_{k=1}^{K} \frac{(O_{0k} - E_{0k})^2}{E_{0k}} + \sum_{k=1}^{K} \frac{(O_{1k} - E_{1k})^2}{E_{1k}}$$

where O_{0k} and O_{1k} denote the observed counts of non-responses and responses for arm k (i.e. $n_k - y_k$ and y_k), and E_{0k} and E_{1k} are the "expected" counts of non-responses and responses, given by

$$E_{0k} = n_k \frac{\sum_k n_k - \sum_k y_k}{\sum_k n_k} \quad \text{and} \quad E_{1k} = n_k \frac{\sum_k y_k}{\sum_k n_k}$$

A smaller value of T indicates higher homogeneity in the treatment effect across arms.

Then CBHM links the shrinkage parameter σ^2 with T through the following two-parameter exponential model

$$\sigma^2 = g(T) = \exp\{a + b \times \log(T)\}, \tag{15.2}$$

where a and b are tuning parameters that characterize the relationship between σ^2 and T. Also $b > 0$ is required so that greater homogeneity (i.e. a small value of T) leads to stronger shrinkage (i.e. a small value of σ^2). The values of a and b in (15.2) are calibrated using the following three-step simulation-based procedure:

1. Simulate the case in which the treatment is effective for all arms. Specifically, R replicates of data are generated by simulating $\boldsymbol{y} = (y_1, \ldots, y_K)$ from *Binomial*$(\boldsymbol{n}, \boldsymbol{\pi_1})$, where $\boldsymbol{n} = (n_1, \ldots, n_K)$ and $\boldsymbol{\pi_1} = (\pi_{11}, \ldots, \pi_{K1})$ and then calculate T for each simulated dataset. Let H_{B1} denote the median of T from R simulated datasets.
2. Simulate the cases in which the treatment effect is heterogeneous across arms. Let $\boldsymbol{\pi(k)} = (\pi_{11}, \ldots, \pi_{k1}, \pi_{(k+1)0}, \ldots, \pi_{K0})$ denote scenario in which the treatment is effective for the first k arms with the target response rate of π_{k1}, but not effective for arms $(k+1)$ to K with the reference response rate of π_{k0}. Given a value of k, we generate R replicates of data by simulating $\boldsymbol{y}$ from Binomial$(\boldsymbol{n}, \boldsymbol{\pi(k)})$, calculate T for each simulated dataset and then obtain its median H_{B2k}. Repeat this for $k = 1, 2, \ldots, K-1$ and define
$$H_{B2} = \min_k(H_{B2k}).$$
3. Let σ^2_{B1} denote a prespecified small value (the default value is 1 in East Bayes) for shrinkage parameter σ^2 under which strong shrinkage or information borrowing occurs under the hierarchical model (equation (15.1)), and let σ^2_{B2} denote a prespecifed large value (the default value

is 80 in East Bayes) of shrinkage parameter σ^2, under which little shrinkage or information borrowing occurs. Solve a and b in equation (15.2) based on the following two equations

$$\begin{cases} \sigma^2_{B1} = g(H_{B1}; a, b) = \exp\{a + b \times \log(H_{B1})\} \\ \sigma^2_{B2} = g(H_{B2}; a, b) = \exp\{a + b \times \log(H_{B2})\} \end{cases} \tag{15.3}$$

which enforces strong and weak shrinkage respectively. The solution of the equations (15.3) is given by

$$\begin{aligned} a &= \log(\sigma^2_{B1}) - \frac{\log(\sigma^2_{B2}) - \log(\sigma^2_{B1})}{\log(H_{B2}) - \log(H_{B1})} \log(H_{B1}) \\ b &= \frac{\log(\sigma^2_{B2}) - \log(\sigma^2_{B1})}{\log(H_{B2}) - \log(H_{B1})} \end{aligned}$$

East Bayes's take: While we report the procedure from Chu and Yuan (2018a), we leave the users to assess the procedure in §15.3.2.2. We would probably take a formal empirical Bayes approach instead, such as the procedure in Carlin and Louis (2010).

15.3.2.3 Trial Design

CBHM applies the same trial design as that in BBHM (§15.3.1).

15.3.3 ExchangeabilityNonexchangeability (EXNEX) Method

Neuenschwander et al. (2016) proposed the exchangeabilitynonexchangeability (EXNEX) approach that allows each arm-specific parameter to be exchangeable with other similar arm parameters or nonexchangeable with any of them.

15.3.3.1 Probability Model

Consider a phase II basket trial that evaluates the efficacy of a new treatment in K different arms (indications). Let n_k and y_k denote the number of patients and responders in arm k, respectively. Denote by p_k the true and unknown response rate for arm k. A natural sampling model for y_k given n_k and p_k is binomial model, $y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k)$.

The objective of the trial is to test whether the new treatment is effective in each of the arms

$$H_{0k} : p_k \leq \pi_{k0} \quad \text{versus} \quad H_{1k} : p_k \geq \pi_{k1},$$

for $k = 1, 2, \ldots, K$, where π_{k0} and π_{k1} are the reference and target response rates for arm k, respectively. Let $\theta_k = \log\left(\frac{p_k}{1-p_k}\right)$ denote the log-odds of the response rate. EXNEX models the θ_k's with a mixture distribution,

$$\theta_k \mid \boldsymbol{w}_k, \boldsymbol{\theta}_{\text{EX}}, \boldsymbol{\sigma}^2_{\text{EX}}, \boldsymbol{\theta}_{\text{NEX}}, \boldsymbol{\sigma}^2_{\text{NEX}} \sim \sum_{c=1}^{C} w_{kc} N(\theta_{\text{EX},c}, \sigma^2_{\text{EX},c}) + w_{k0} N(\theta_{\text{NEX},k}, \sigma^2_{\text{NEX},k}). \tag{15.4}$$

In other words, with probability w_{kc}, θ_k belongs to an exchangeability (EX) component c, and with probability w_{k0}, θ_k belongs to a nonexchangeability (NEX) component. Here, $\sum_{c=0}^{C} w_{kc} = 1$. The parameters of the EX components, $\theta_{\text{EX},c}$ and $\sigma^2_{\text{EX},c}$ are shared across arms within component c. In contrast, the parameter of the NEX components, $\theta_{\text{NEX},k}$ and $\sigma^2_{\text{NEX},k}$ are arm-specific. The number of EX components C and the weights of the components $\boldsymbol{w}_k = (w_{k1}, \ldots, w_{kC}, w_{k0})$ are prespecified by the investigator. By default, the same NEX components and mixture weights are specified for all arms, $\theta_{\text{NEX},1} = \ldots = \theta_{\text{NEX},K} = \theta_{\text{NEX}}$, $\sigma^2_{\text{NEX},1} = \ldots = \sigma^2_{\text{NEX},K} = \sigma^2_{\text{NEX}}$, and $\boldsymbol{w}_1 = \ldots = \boldsymbol{w}_K = \boldsymbol{w}$. For the prior specification, in each EX component c, a normal prior is assigned to $\theta_{\text{EX},c}$, and a half-normal (HN) prior with scale parameter s_c is assigned to $\sigma_{\text{EX},c}$,

$$\theta_{\text{EX},c} \sim N(\mu_{\text{EX},c0}, \sigma^2_{\text{EX},c0}), \quad \sigma_{\text{EX},c} \sim \text{HN}(s_c).$$

In East Bayes, the default settings Neuenschwander et al. (2016) is used for EXNEX: A mixture of two $(C = 2)$ EX distributions and one NEX distribution with weights $\boldsymbol{w} = (0.25, 0.25, 0.5)$ is

chosen by default. Therefore, in brief, East Bayes applies the following hierarchical model:

$$
\begin{aligned}
&\text{Likelihood:} && y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k) \\
&\text{Transformation:} && \theta_k = \log\left(\frac{p_k}{1-p_k}\right) \\
&\text{Prior for } \theta_k: && \theta_k \mid \boldsymbol{w}, \boldsymbol{\theta}_{\text{EX}}, \boldsymbol{\sigma}^2_{\text{EX}}, \theta_{\text{NEX}}, \sigma^2_{\text{NEX}} \sim \\
& && 0.25N(\theta_{\text{EX},1}, \sigma^2_{\text{EX},1}) + 0.25N(\theta_{\text{EX},2}, \sigma^2_{\text{EX},2}) + 0.5N(\theta_{\text{NEX}}, \sigma^2_{\text{NEX}}) \\
&\text{Hyperpriors:} && \theta_{\text{EX},1} \sim N(\mu_{\text{EX},10}, \sigma^2_{\text{EX},10}),\ \sigma_{\text{EX},1} \sim \text{HN}(s_1) \\
& && \theta_{\text{EX},2} \sim N(\mu_{\text{EX},20}, \sigma^2_{\text{EX},20}),\ \sigma_{\text{EX},2} \sim \text{HN}(s_2)
\end{aligned}
\tag{15.5}
$$

Following Neuenschwander et al. (2016), weakly-informative priors are used in East Bayes by default. Specifically, for the priors of the NEX parameters, we fix the mean θ_{NEX} at the log-odds of a plausible guess for the response probability (e.g. the mean of the middle of reference and target response rates across arms, $p_w = \frac{1}{K}\sum_{k=1}^{K}\frac{\pi_{k1}+\pi_{k0}}{2}$), and the variance σ^2_{NEX} at a value that corresponds to approximately one observation, $\sigma^2_{\text{NEX}} = 1/p_w + 1/(1-p_w)$, for all arms. For EX components, we place $N\left(\log\left(\frac{\pi_0}{1-\pi_0}\right), 1/\pi_0 + 1/(1-\pi_0) - 1\right)$ and $N\left(\log\left(\frac{\pi_1}{1-\pi_1}\right), 1/\pi_1 + 1/(1-\pi_1) - 1\right)$ prior on $\theta_{\text{EX},1}$ and $\theta_{\text{EX},2}$, respectively, where $\pi_0 = \frac{1}{K}\sum_{k=1}^{K}\pi_{k0}$ and $\pi_1 = \frac{1}{K}\sum_{k=1}^{K}\pi_{k1}$ are the average reference and target response rate across arms; and half-normal priors with scale parameter $s_1 = s_2 = 1$ on $\sigma_{\text{EX},1}$ and $\sigma_{\text{EX},2}$.

15.3.3.2 Trial Design

The original EXNEX design does not have a futility or efficacy stopping rule, but for fair comparison, the same rules as those in BBHM (§15.3.1) are available in East Bayes.

15.3.4 Multiple Cohort Expansion (MUCE) Method

The multiple cohort expansion (MUCE) design is originally proposed by Lyu et al. (2020), for trials with multiple arms, include basket trials. The MUCE is based on a class of Bayesian hierarchical models including a latent probit prior that allows for different degrees of borrowing across arms. Furthermore, instead of using the posterior interval of the estimated response rate to declare futility or efficacy, as in BBHM (§15.3.1), CBHM (§15.3.2) and EXNEX (§15.3.3), MUCE applies a formal Bayesian hypothesis test to make statistical inference.

15.3.4.1 Probability Model

Consider a phase II basket trial that evaluates the efficacy of a new treatment in K different arms (indications). Suppose n_k patients have been treated in arm k, and y_k of them respond. Let p_k denote the true and unknown response rate for the arm k. We assume y_k follows a binomial distribution conditional on n_k and p_k, $y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k)$. Whether arm k is effective can be examined by the following hypothesis test:

$$H_{0k} : p_k \leq \pi_{k0} \quad \text{versus} \quad H_{1k} : p_k > \pi_{k0}, \tag{15.6}$$

where π_{k0} is the reference response rate for arm k.

MUCE constructs a formal Bayesian testing framework for (15.6). Let λ_k be a binary indicator of the hypothesis, such that $\lambda_k = 0$ (or 1) represents that hypothesis H_{0k} (or H_{1k}) is true. Firstly, a prior model for p_k is built under each hypothesis. Let $\theta_k = \log\left(\frac{p_k}{1-p_k}\right)$ denote the log-odds of the response rate. The null hypothesis $p_k \leq \pi_{k0}$ is equivalent to $\theta_k \leq \theta_{k0}$, and the alternative hypothesis is equivalent to $\theta_k > \theta_{k0}$, where $\theta_{k0} = \log\left(\frac{\pi_{k0}}{1-\pi_{k0}}\right)$. Conditional on λ_k, MUCE assumes

$$\theta_k \mid \lambda_k = 0 \sim \text{Trunc-Cauchy}(\theta_{k0}, \gamma; (-\infty, \theta_{k0}]),$$
$$\theta_k \mid \lambda_k = 1 \sim \text{Trunc-Cauchy}(\theta_{k0}, \gamma; (\theta_{k0}, \infty)),$$

where Trunc-Cauchy$(\theta, \gamma; A)$ denotes a Cauchy distribution with location θ and scale γ truncated to interval A.

Secondly, prior models for the probabilities of the hypotheses (i.e. priors for the probabilities of $\{\lambda_k = 1\}$) are constructed. MUCE uses a probit model as the prior model for λ_k. Let Z_k be a latent Gaussian random variable, and $\lambda_k = I(Z_k < 0)$, where $I(\cdot)$ is an indicator function. Z_k is assumed to follow a normal distribution,

$$Z_k \sim N(\eta_k, \sigma_0^2).$$

Here, $\mathrm{E}(Z_k) = \eta_k$, in which η_k characterizes the effect of arm k. The arm-specific effects are then separately modeled by common priors,

$$\eta_k \mid \eta_0, \sigma_\eta \overset{iid}{\sim} N(\eta_0, \sigma_\eta^2).$$

Lastly, give η_0 a hyperprior, $\eta_0 \sim N(\mu_{\eta_0}, \sigma_{\eta_0}^2)$.

In brief, the entire hierarchical models are summarized in the following display:

$$\begin{aligned}
&\text{Likelihood:} && y_k \mid n_k, p_k \sim \text{Binomial}(n_k, p_k);\\
&\text{Transformation:} && \theta_k = \log\left(\frac{p_k}{1-p_k}\right), \theta_{k0} = \log\left(\frac{\pi_{k0}}{1-\pi_{k0}}\right);\\
&\text{Prior for } (\theta_k \mid \lambda_k): && \theta_k \mid \lambda_k = 0 \sim \text{Trunc-Cauchy}(\theta_{k0}, \gamma; (-\infty, \theta_{k0}]),\\
& && \theta_k \mid \lambda_k = 1 \sim \text{Trunc-Cauchy}(\theta_{k0}, \gamma; (\theta_{k0}, \infty));\\
&\text{Prior for } \lambda_k: && \lambda_k = \begin{cases} 0, & \text{if } Z_k < 0,\\ 1, & \text{if } Z_k \geq 0; \end{cases}\\
&\text{Latent probit regression:} && Z_k \mid \eta_k, \sigma_0^2 \sim N(\eta_k, \sigma_0^2);\\
&\text{Arm-specific effects:} && \eta_k \mid \eta_0, \sigma_\eta^2 \sim N(\eta_0, \sigma_\eta^2);\\
&\text{Hyperprior:} && \eta_0 \mid \mu_{\eta_0}, \sigma_{\eta_0}^2 \sim N(\mu_{\eta_0}, \sigma_{\eta_0}^2),
\end{aligned} \tag{15.7}$$

In East Bayes, the values of the hyperparameters $\gamma = 2.5$, $\mu_{\eta_0} = 0$, $\sigma_0^2 = 100$, $\sigma_\eta^2 = 1$ and $\sigma_{\eta_0}^2 = 1$ are used by default.

15.3.4.2 Trial Design

Suppose $L(\geq 0)$ interim looks are planned, and the l-th interim analysis is conducted after n_k^l patients have been enrolled in arm k. Let $\mathcal{D}^l \equiv \{(n_k^l, y_k^l) : k = 1, 2, \ldots, K\}$ denote the observed data at interim analysis l, where y_k^l is the number of responders among the n_k^l patients. Denote $\mathcal{D}^{L+1} \equiv \{(n_k^{L+1}, y_k^{L+1}) : k = 1, 2, \ldots, K\}$ the observed data at the end of the trial, where n_k^{L+1} is the prespecified maximum sample size for arm k and y_k^{L+1} is the total number of responders. The proposed phase II basket trial design with L interim looks is describe as follows:

1. Enroll n_k^1 patients in k-th arm, $k = 1, 2, \ldots, K$.
2. Given the data $\mathcal{D}^l$ at the l-th interim look, $l = 1, 2, \ldots, L$,
   (a) *[Futility stopping]* If the posterior probability that the hypothesis of arm k, H_{1k}, is true (i.e., $\lambda_k = 1$) is small, i.e.,

$$Pr\{\lambda_k = 1 \mid \mathcal{D}^l\} < P_{\textit{futility}},$$

stop the accrual to the k-th arm for futility;

(b) Otherwise, continue to enroll patients until reaching the next interim analysis.

3. Once the maximum sample size is reached or all the arms have stopped, evaluate the efficacy for each arm based on all the observed data. If the posterior probability that that the hypothesis of arm k, H_{1k}, is true (i.e., $\lambda_k = 1$) is large, i.e.,

$$Pr\{\lambda_k = 1 \mid \mathcal{D}^{L+1}\} > \phi_k,$$

arm k is declared efficacious and promising; otherwise, it is considered not promising.

Similar in BBHM (§15.3.1), Step 2 is optional. In East Bayes, the probability threshold for futility interim analysis, $P_{\textit{futility}}$, and for the final analysis, $\{\phi_k : k = 1, 2, \ldots, K\}$, are calibrated through simulations to achieve a prespecified type I error rate for each arm, under the null scenario. See the detailed calibration process in §15.3.1.

15.3.4.3 Discussion

MUCE is also used as a design for cohort expansion clinical trials. Finally, MUCE is a sophisticated method, the detail of which is in Lyu et al. (2020).

Part VI

Real-World Evidence

16. Meta-Analytic-Predictive (MAP) Priors

16.1 Introduction

This module briefly describes the design of a Bayesian Meta-Analytic-Predictive (MAP) priors from historical data of the past clinical trials. Along with the functionality of creating a new MAP prior, this interface also provides an easier way to visualize and compare different MAP priors.

Historical information has been always useful when designing clinical trials, but it could also be incorporated in the analysis. Although, the formal use of historical information in the analysis is controversial, but when incorporated, historical data allow us to reduce the number of subjects, which brings down the cost and the trial duration, facilitates recruitments and may be more ethical (Schmidli et al., 2014).

Techniques for incorporating historical information are well developed in the earlier phases of drug development, occasionally in phase II studies, special areas such as medical devices and pediatric studies. Also, clinical trials where control arm is entirely replaced by historical information are popular in phase II oncology trial but could lead to biases. Regardless of whether the information on control is to be used in design or analysis, there is a need to provide a quantitative summary of the available historical data. One direct way to consider the most appropriate summary is the predictive distribution of the control parameter in the new trial and in Bayesian paradigm the predictive distribution can then be used as a prior distribution to be used into the final analysis (Schmidli et al., 2014; Neuenschwander et al., 2010).

Use of historical data in analysis needs a more careful look because overly optimistic use of historical data may be inappropriate due to prior-data conflict. This approach is similar to a meta-analytic-combined analysis of historical and new data assuming the exchangeability of the parameters across the trials.

In the meta-analytic approach to incorporate historical data, the generated MAP prior is further robustified (Schmidli et al., 2014). The robust prior is a mixture prior with two components. The first one which is derived from historical data is a MAP prior and the second one is an additional weakly informative or non-informative component that robustify against the prior-data conflict. The weight on that second component is basically the prior probability that the current trial differs systematically from the historical data. The choice of mixing weights determines how quickly historical information is discounted with increasing prior-data conflict. The important thing to note here is that MAP prior is not available in analytical form. A kernel-density estimate from the MCMC sample can be used to describe the MAP prior. In order to do a tractable posterior analysis, MAP prior is approximated by a mixture of conjugate priors where Kullback-Leibler (KL) divergence is used a a measure of discrepancy. When the control data and the robust prior are in clear conflict, the prior information will be essentially discarded in the posterior analysis. Adaptive design of trials could minimize this particular risk though (Schmidli et al., 2014; Neuenschwander, 2011).

When we design a clinical trial, we need to specify the number of subjects allocated to the control and the treatment arm. If the historical data is used, it is very important to know the prior effective sample size (ESS) which is the equivalent number of subjects corresponding to the prior information.

In this module, using R Bayesian evidence synthesis Tools (**RBesT**) package we provide an interface to borrow strength from historical information in clinical trials. Once relevant historical information has been identified, RBesT supports the derivation of informative priors via the Meta-Analytic-Predictive (MAP) approach.

In §16.2 of this document we introduce the R-shiny interface for creating and visualizing new MAP prior from historical data and comparing those. Statistical method overview is given in §16.3.

16.2 User Interface and Tutorial

On entering the Meta-Analytic-Predictive (MAP) prior UI interface, the users will see main two tabs - **Create New Prior** and **Prior Comparison and Visualization** as Figure 16.1. The first tab gives the option to create a new MAP prior from historical data and the second one provides to visualize those and compare different MAP priors (maximum 5 now).

Figure 16.1: Two tabs in the **MAP** module.

New priors can be designed under **Create New Prior** tab and generated priors can be visualized and/or compared under **Prior Comparison and Visualization** tab.

16.2.1 Creating New Prior

After clicking this tab, another two-tabbed window containing **Inputs** and **Outputs** tabs appear. In the **Input** tab, there are three steps to follow to design a new prior as shown in Figure 16.2 below.

16.2.1.1 Inputs

Step 1: Type of Priors

There are two sets of radio button - **Prior Derivation** and **Effect Parameter**. The options for effect parameter depends on the selected prior derivation. Prior Derivation has two options - **Control** and **Effect Size**. MAP prior is derived for the control group baseline effect using historical data exclusively on the control group when **Control** is selected. On the other hand, MAP prior is derived for the treatment effect based on historical data from two-arm studies when **Effect Size** is selected.

- When prior derivation is **Control** then **Effect Parameter** could be either **Mean** or **Proportion** which denote normal or binomial endpoint, respectively (see Figure 16.3 below).
- When prior derivation is **Effect Size** then **Effect Parameter** could be **Log Odds Ratio** or **Log Hazard Ratio** or **Difference of Means** which denotes treatment effect size for binomial or survival or normal endpoint, respectively (see Figure 16.4 below).

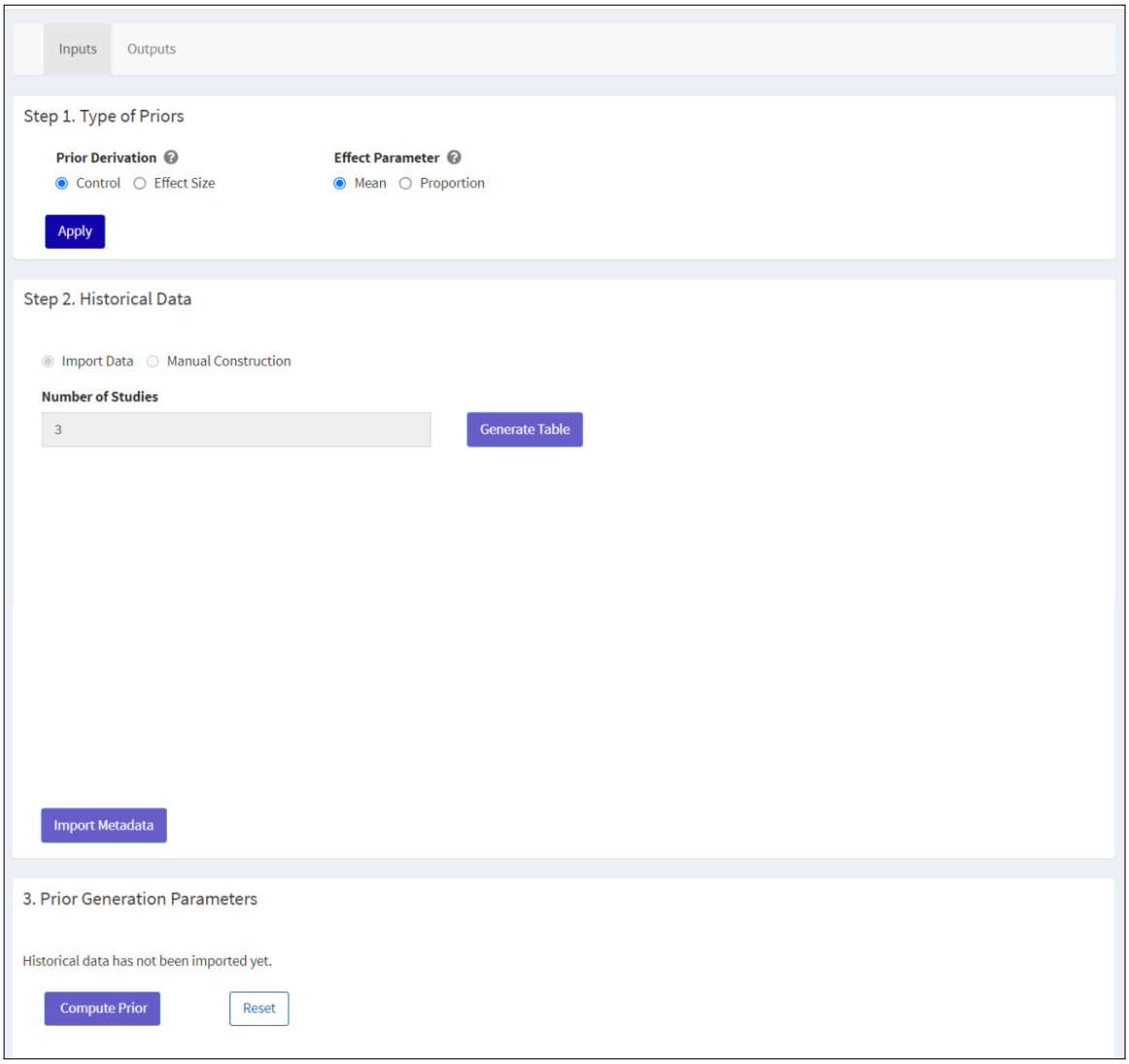


Figure 16.2: Tabs under create new prior.

Choices for **Step 2** and **Step 3** is updated based on the selection from **Step 1** - **Prior Derivaton** and **Effect Parameter**. In order to move to **Step 2**, the "Apply" button in **Step 1** needs to be clicked and it changes to "Edit" for updating the selected options and subsequently resets all the input parameters.

Step 2: Historical Data

In this step we appropriately select the source of the historical data. Currently, there are two ways to input data - Import Data and Manual Construction. If Import Data is selected then a file browser interface to upload a *.csv* **or** *.xlsx* **file (with header) is shown (see Fig-**

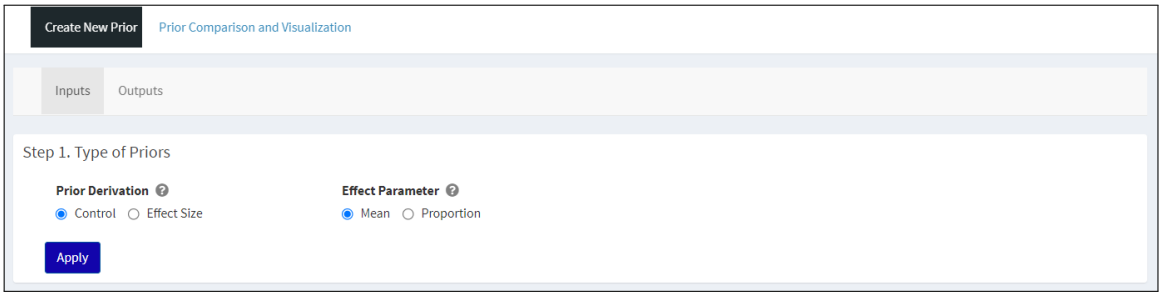


Figure 16.3: Effect parameters when prior dervation is **Control**

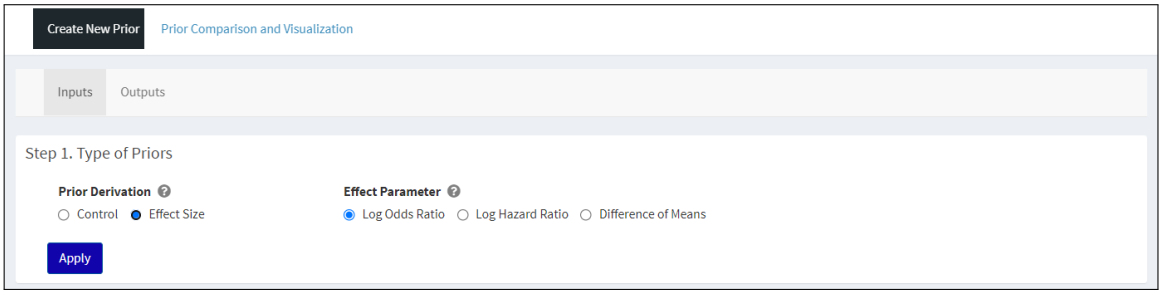


Figure 16.4: Effect parameters when prior dervation is **Effect Size**

ure 16.5). In case a user wants to know about the appropriate data format, there is a button named "Download Data Template" and a sample file starts downloading once clicked.

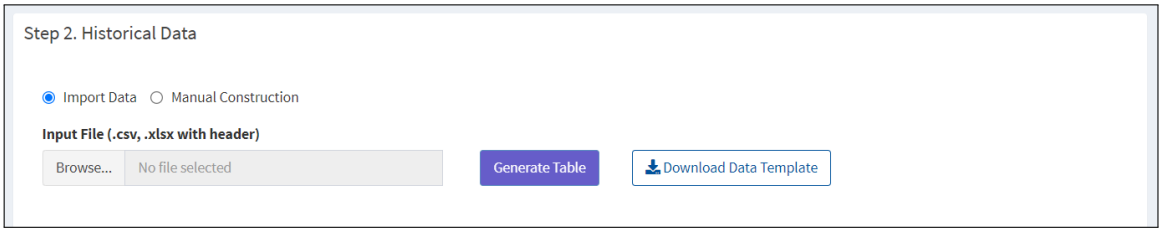


Figure 16.5: Data import from file

In order to enter data manually, "Generate Table" button can be clicked after giving the appropriate Number of Studies (maximum 10) and an appropriate editable empty table is shown there where user can put the values in each cell. One example is given below in Figure 16.6.

On the other hand when the Import Data is selected and appropriate file with header is uploaded, a column a selection panel is shown on the screen. Appropriate column needs to be selected from the dropdown list before clicking Generate Table (see and example in

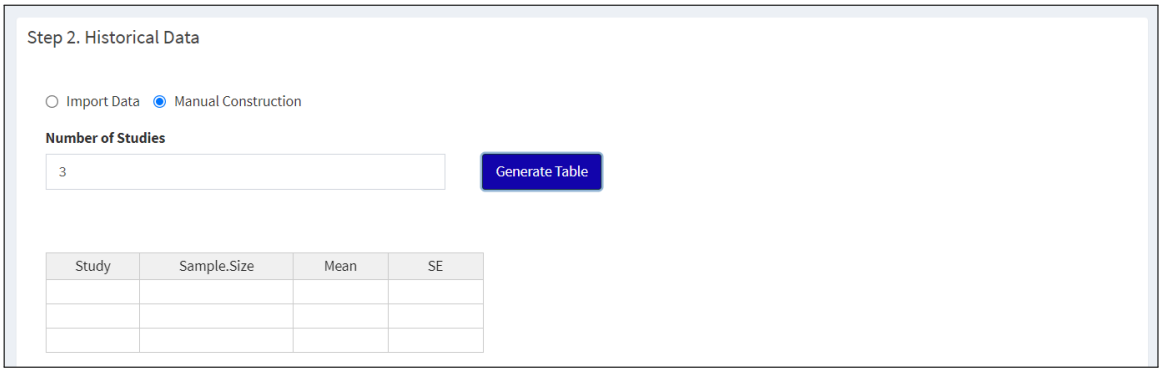


Figure 16.6: Manual entry for historical data

Figure 16.7).

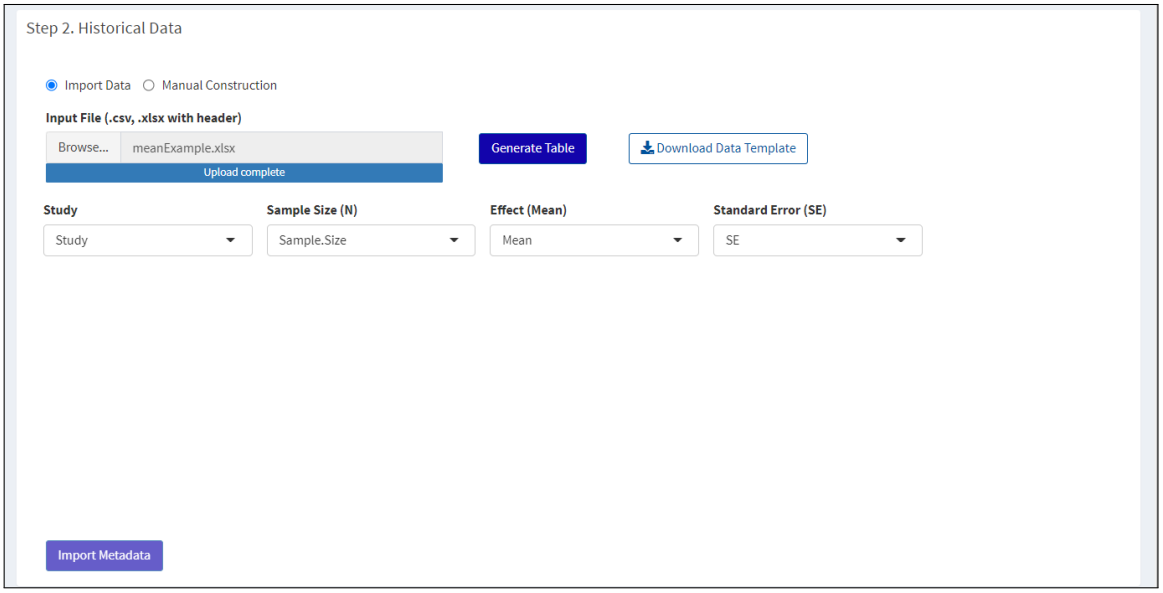


Figure 16.7: Appropriate column header selection before generating the table from an input file

Once the Generate Table is clicked the imported file is shown as a table. Below the imported table there is a button names "Import Metadata". Once clicked the data frame in the table is checked for the appropriate format. If it is successful, the historical data is imported for generating the MAP prior (see Figure 16.8). There is also an "Edit" button, which helps to edit data in the table that one enters.

Please refer Table 16.1 to Table 16.5 for the ranges of the entries in the table for different effect parameters.

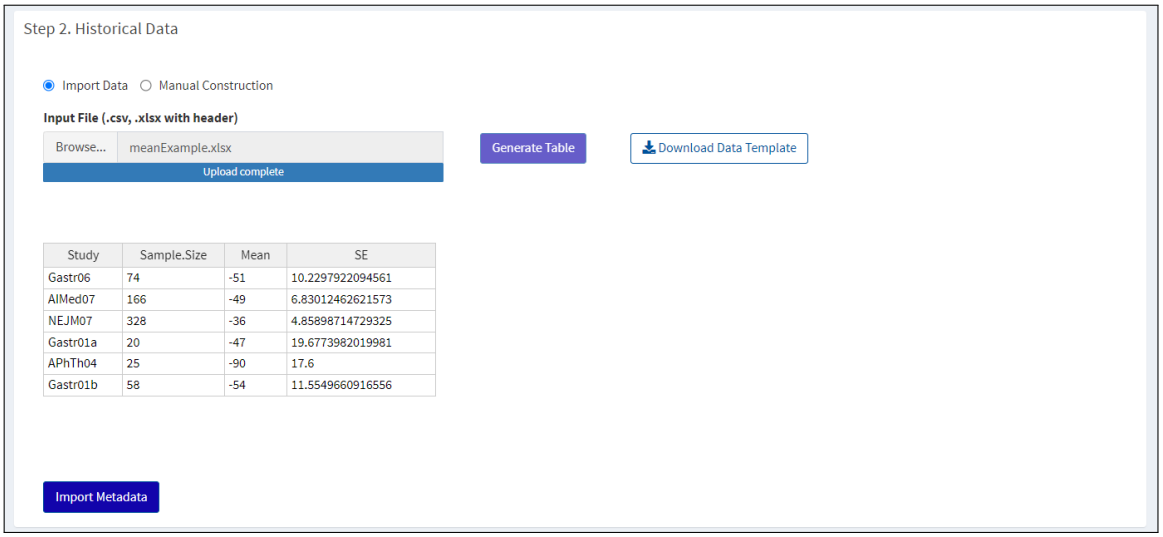


Figure 16.8: Imported historical data from a file

Table 16.1: Data structure for historical data when prior derivation is **Control** and effect parameter is **Mean**

Input Parameter	**Meaning**	**Data Type**	**Range**
Study	Name of the sudy	Alphanumeric string	Upto length 25
Sample.Size	Size of the sample	Integer	$[1, 10^5]$
Mean	Mean of data	Real	$(-\infty, +\infty)$
SE	Standard error of data	Real	$(0, +\infty)$

Once historical data is successfully entered or imported, we move to Step 3, where prior generation parameters needs to be entered.

Step 3: Prior Generation Parameters

In this step all the parameters related to prior generation including heterogeneity, effect prior, robustness, number of mixture components and MCMC computational parameters are specified (see Figure 16.9 and 16.10). Also, when **Effect Size** is chosen as prior derivation in Step 1, then prior generation parameters has two additional inputs namely **Effect Prior** and **Min. Effect Size** as shown in Figure 16.10.

Heterogeneity (τ) parameter has currently three options - **Low**, **High** and **Known** τ. Please refer §16.3 for details. For **Known** τ choice, the constant value can be entered as a positive real number. Choice for **Effect Prior** and **Min. Effect Size** appear only when the prior derivation is

Table 16.2: Data structure for historical data when prior derivation is **Control** and effect parameter is **Proportion**

Input Parameter	Meaning	Data Type	Range
Study	Name of the sudy	Alphanumeric string	Upto length 25
Sample.Size	Size of the sample	Integer	$[1, 10^5]$
Frequency	Numer of patients with response	Integer	[0, Sample.Size]

Table 16.3: Data structure for historical data when prior derivation is **Effect Size** and effect parameter is **Log Odds Ratio**

Input Parameter	Meaning	Data Type	Range
Study	Name of the sudy	Alphanumeric string	Upto length 25
Freq.cnt	Number of responders in control arm	Integer	[0, Sample.Size.cnt]
Sample.Size.cnt	Sample size of the control arm	Integer	$[1, 10^5]$
Freq.trt	Number of responders in treatment arm	Integer	[0, Sample.Size.trt]
Sample.Size.trt	Sample size of the treatment arm	Integer	$[1, 10^5]$

Effect Size (refer to last two rows in Table 16.6).

There is an option to robustify the MAP prior and that can be selected by checking the box next to **Add Robust Component** and typing the mixture weight for the component in the box labelled as **Component Weight**. In order to use the MAP prior conveniently, the kernel density estimate from the MCMC samples are approximated by a mixture of conjugate priors. Users have an option to find the **Number of Mixture Components** automatically or set it manually. Finally, there are three computational parameters - first one is **Random Seed**, which is useful to make the results exactly reproducible, second one is **No. of MCMC Runs** and the last one is **ESS Computational Method** or methods by which effective sample size (ESS) is calculated.

The main advantage of using historical information is the possibility to reduce the number of

Table 16.4: Data structure for historical data when prior derivation is **Effect Size** and effect parameter is **Log Hazard Ratio**

Input Parameter	Meaning	Data Type	Range
Study	Name of the sudy	Alphanumeric string	Upto length 25
NEvents.cnt	Number of events in the control arm	Integer	$[1, 10^5]$
NEvents.trt	Number of events in the treatment arm	Integer	$[1, 10^5]$

control patients, as the informative prior is effectively equivalent to a certain number of control patients. This is called the effective sample size (ESS). Note that the moment matching approach leads to conservative (small) ESS estimates while the Morita (Morita et al., 2008) method tends to estimates liberal (large) ESS estimates when used with mixtures. Also, number of MCMC runs includes 2000 burn-in iterations.

Once all the parameters are successfully entered there is a button named "Compute Prior" and once that is clicked the computation for MAP prior begins.

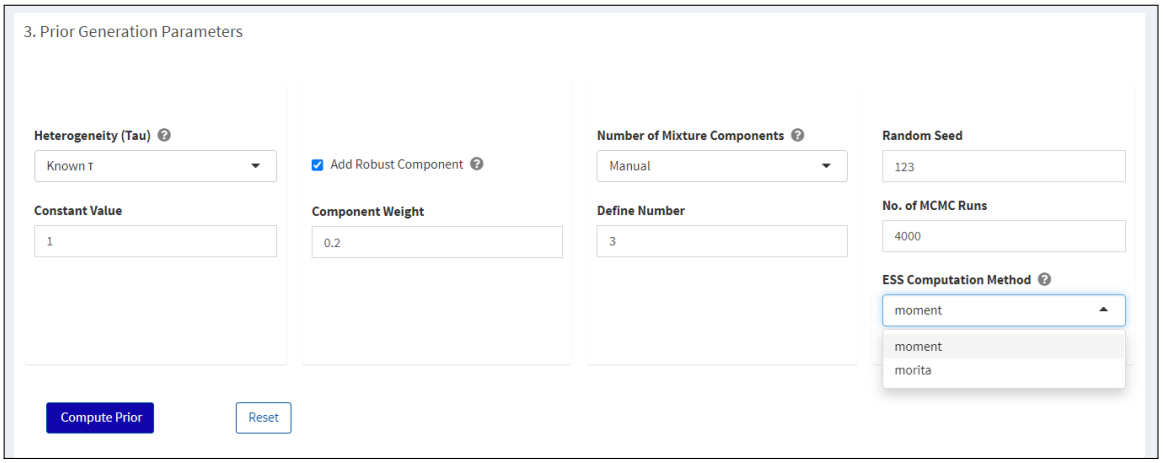


Figure 16.9: Prior generation parameter when historical data is borrowed from control arm only

All the choices for input prior generation parameters can be seen in Figures 16.9 and 16.10. The details of the range and default values are given in Table 16.6.

There is also a "Reset" button next to the "Compute Prior". All the input parameters and output

Table 16.5: Data structure for historical data when prior derivation is **Effect Size** and effect parameter is **Difference of Means**

Input Parameter	Meaning	Data Type	Range
MeanDiff	Mean effect difference between the treatment and control arms	Real	$(-\infty, +\infty)$
Sample.Size.cnt	Sample size of the control arm	Integer	$[1, 10^5]$
Sample.Size.trt	Sample size of the treatment arm	Integer	$[1, 10^5]$
SE.cnt	Standard error of the control arm	Real	$(0, +\infty)$
SE.trt	Standard error of the treatment arm	Real	$(0, +\infty)$

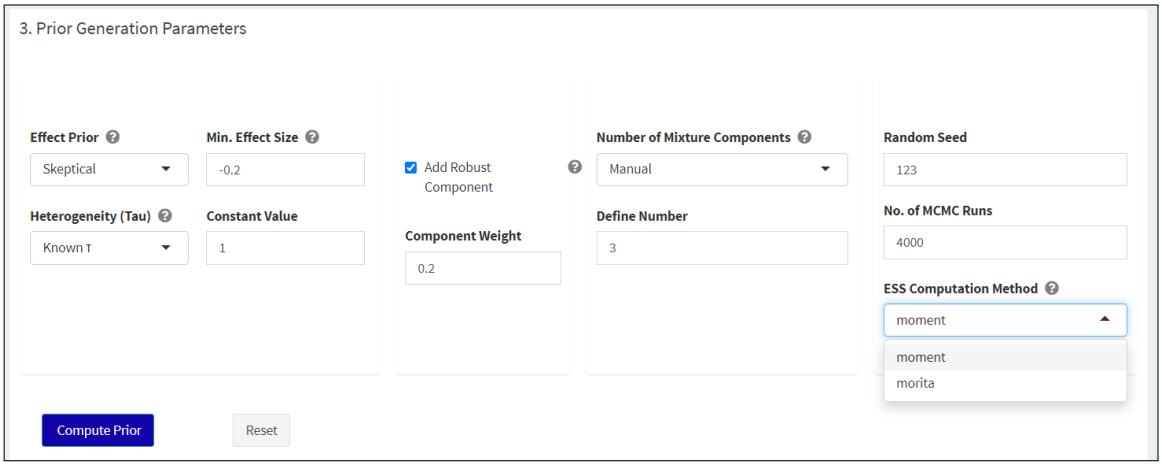


Figure 16.10: Prior generation parameter when historical data is borrowed from two arm studies

results get reset once "Reset" button is clicked.

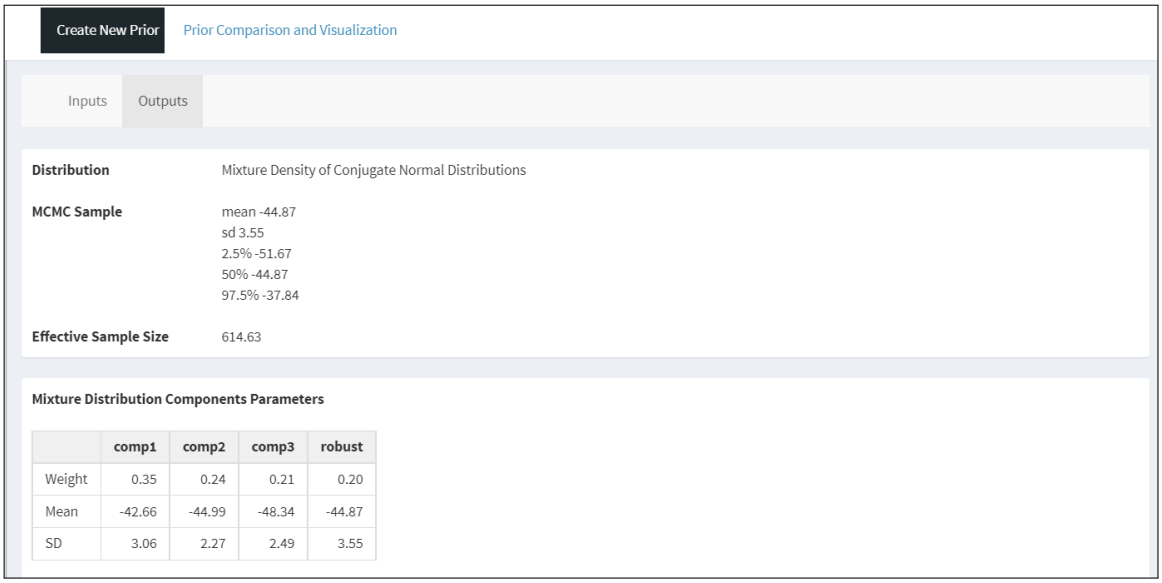
16.2.1.2 Outputs

Upon a successful computation of MAP prior, the summary from MCMC samples and density plots and other important information are shown on **Outputs** tab. Under this tab we have the information

Table 16.6: Different prior generation parameters and their choices

Input Parameter	**Meaning**	**Range**	**Default**
Heterogeneity (Tau)	Prior for the between-trial heterogeneity of the random effects meta analytic model.	{High, Low, Known τ}	High
Constant Value	Between trial standard deviation	$(0, +\infty)$	1
Component weight	Weight of robust component	$(0, 1)$	0.2
Number of Mixture Components	Component to fit a mixture model	$[1, 50]$	3
Random Seed	Random seed to make the results reproducible	$[1, 10^7]$	123
No. of MCMC Runs	Total MCMC iterations including 2000 burn-in iterations	$[4000, 15000]$	4000
ESS Computation Method	Possible ways to calculate effective sample size	{moment, mortia}	moment
Effect Prior	Type of effect prior	Skeptical, Enthusiastic	Skeptical
Min. Effect Size	Minimum Effect size	$(-\infty, +\infty)$	-0.2

regarding the generated MAP prior. The name of the distribution is **Mixture Density of Conjugate Normal Distributions** or **Mixture Density of Conjugate Beta Distributions** depending on the endpoint whether it is normal or binary. Next, we show the summary statistics for the MCMC samples and effective sample size (ESS). In the next section the parameters and mixture weights for all the components in the mixture distribution are shown. In Figure 16.11, for example,we show all the information related to MAP prior for an example when **Prior Derivation** and **Effect Parameter** are selected as **Control** and **Mean**, respectively. Below the summary information, we have three plots - first one is the forest plot of the estimated mean and standard deviation (SD), next one is the kernel density plot for MAP prior from the MCMC samples and the final one is the density plots for the components of the MAP prior along with the robust component if that is present (see



	comp1	comp2	comp3	robust
Weight	0.35	0.24	0.21	0.20
Mean	-42.66	-44.99	-48.34	-44.87
SD	3.06	2.27	2.49	3.55

Figure 16.11: Information regarding the generated MAP prior

Figure 16.12 as an example)

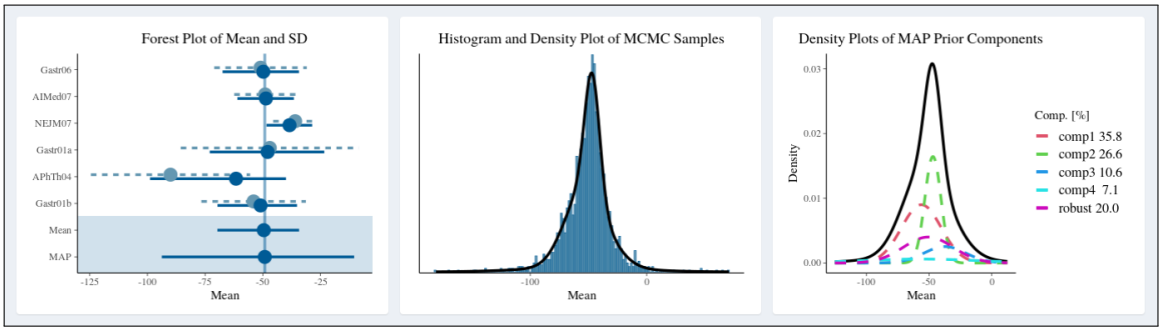


Figure 16.12: Information regarding the generated MAP prior

At the end of the **Output** page, users have options to save the MAP prior result in a file in the cloud, load the results of a MAP prior from a file in the cloud. Apart from "Save MAP Prior" and "Load MAP prior", user can also download the *.rds* file related to a MAP prior to the local machine by selecting the name of the MAP prior and clicking the "Download MAP Prior". There is also a delete option where user can delete a selected MAP prior from the cloud storage by clicking the button "Delete MAP Prior". Note that, while saving the file the name of the file cannot contain any special character. These options can be found in Figure 16.13.

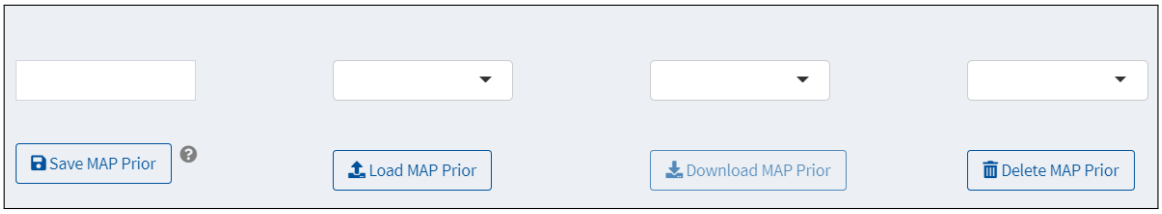
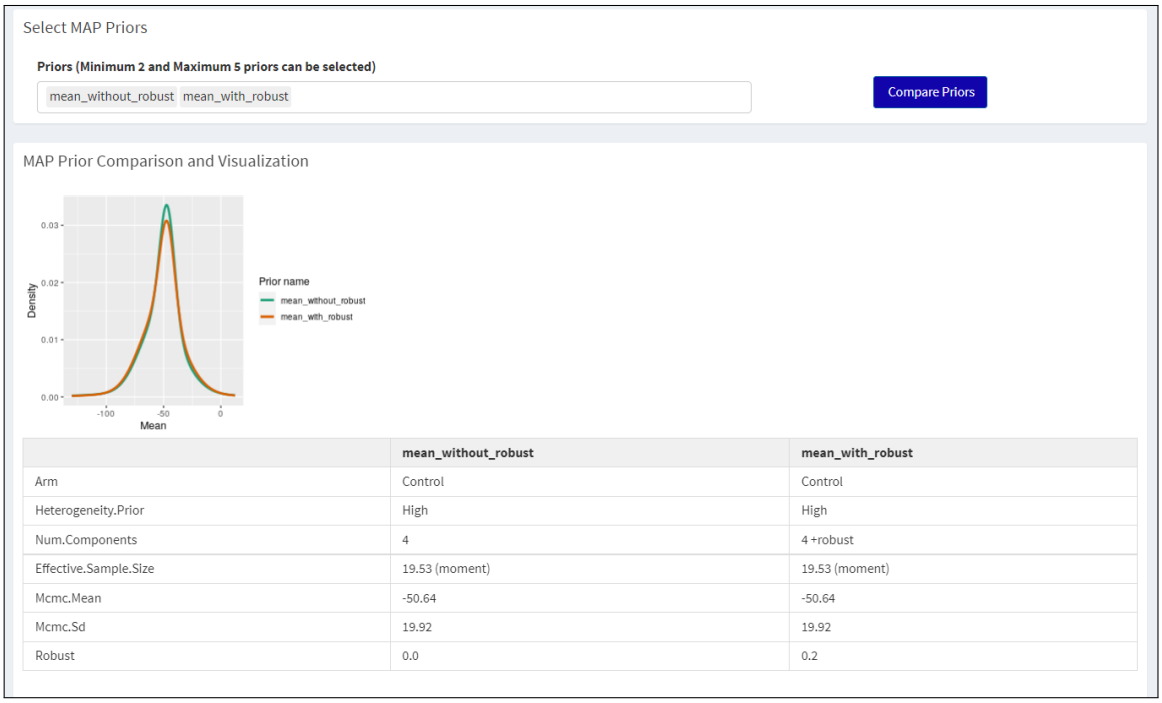


Figure 16.13: Save, Load, Download or Delete a MAP prior

16.2.2 Prior Comparison and Visualization

In the **Prior Comparison and Visualization** tab, there is a selection box where minimum two and maximum five MAP prior could be selected and compared side by side. Once the appropriate MAP priors are selected and "Compare Priors" is clicked, the densities of selected MAP priors are shown together overlapping each other. Below this plot, corresponding **Prior Generation Parameters** are shown side by side. User can quickly compare those parameters at a glance (see Figure 16.14 as an example). Next, there is a set of radio buttons to choose a prior that is in this set. Once particular prior is selected, under the **Description of Given Priors** historical data, summary of MCMC samples, parameters of the MAP prior components and a set of three plots similar to the **Outputs** tab are shown for that particular selected MAP prior (see Figure 16.15 as an example). This part is very similar to the results shown under **Outputs** tab, so please refer §16.2.1.2 for details.



	mean_without_robust	mean_with_robust
Arm	Control	Control
Heterogeneity.Prior	High	High
Num.Components	4	4 +robust
Effective.Sample.Size	19.53 (moment)	19.53 (moment)
Mcmc.Mean	-50.64	-50.64
Mcmc.Sd	19.92	19.92
Robust	0.0	0.2

Figure 16.14: MAP Prior comparison and visualization

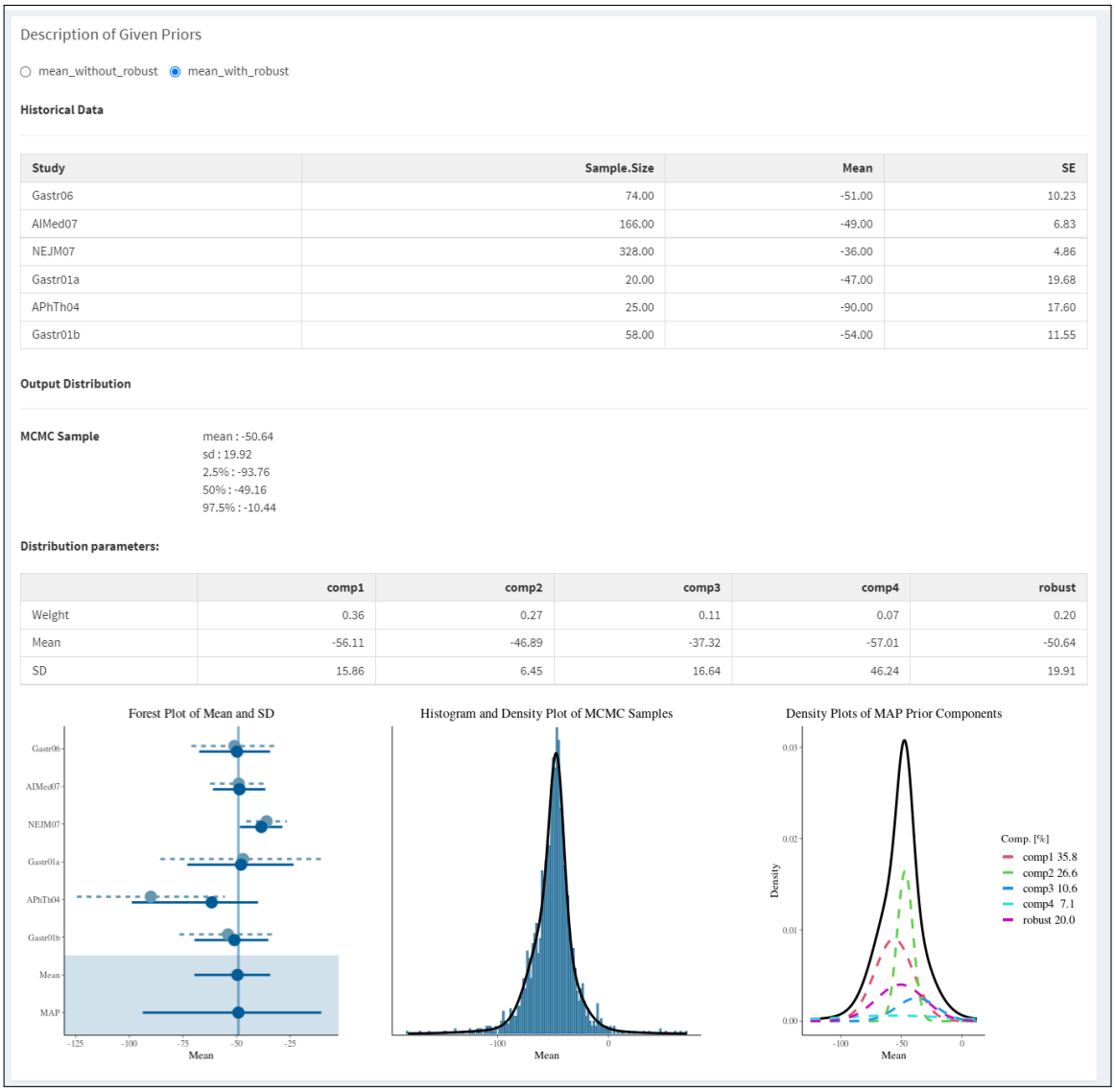


Figure 16.15: Description of selected prior

16.3 Statistical Methods Review

16.3.1 Meta-Analytic-Predictive (MAP) prior generation

Table 16.7: Prior Derivation, Effect Parameters and Endpoints

Prior Derivation	Effect Parameters	Description	Endpoints
Control	Mean	Treatment effect for control arm	Normal
Control	Proportion	Response rate for control arm	Binomial
Effect Size	Difference of Means	Difference of treatment effects in a two-arm study	Normal
Effect Size	Log Odds Ratio	Log of odds ratio for treatment rates in a two-arm study	Binomial
Effect Size	Log Hazard Ratio	Log of hazard rates for a two-arm study	Survival

Based on the RBesT package supports the generation of Meta-Analytic-Predictive (MAP) prior using historical meta data. The tool generates a prior from user-imported historical metadata either for a single control arm or for the effect size for Normal (Δ Means), Binomial (log-OR) and Survival (log-HR) endpoints. Several MAP priors can also be compared visually using this tool. In Table 16.7 we see the endpoints and the descriptions based on the selection of **Prior Derivation** and **Effect Parameter**. In the following sections, we assume the historical data are on effect sizes but the same theory is applied when the historical data are from single control arm as well.

16.3.2 Historical Data: Observed Effect Sizes

Let us denote the historical effect size and parameters of the H historical trials by $Y_{\mathcal{H}} = \{Y_1, Y_2, \cdots, Y_H\}$ and $\theta_{\mathcal{H}} = \{\theta_1, \theta_2, \cdots, \theta_H\}$ where $\mathcal{H} = \{1, 2, \cdots, H\}$ and let us also denote the data and the parameters in the new trial by Y_* and θ_*. We could write the structure as a hierarchical model

$$Y_h|\theta_h \sim F(\theta_h; n_h), \quad \theta_h|\eta \sim G(\eta), \quad \eta \sim P$$

where $\mathcal{H} = \{1, 2, \cdots, H, *\}$ and $n_1, n_2, \cdots, n_H, n_*$ are the sample sizes of the trials. These sample sizes (events) are needed to compute standard errors s_h, where $h \in \mathcal{H}$. Also, F, G, P are sampling, exchangeability (random-effects), and hyper-prior distribution, respectively. Inference for control parameter θ_* in the new trial is based on both Y_* and $Y_{\mathcal{H}}$. A MAP prior is denoted by $p(\theta_*|Y_{\mathcal{H}})$ and derived from the historical data at the design stage. Finally at the end of the trial the current data Y_* is combined with the MAP prior using Bayes' rule i.e $p(\theta_*|Y_1, Y_2, \cdots, Y_H, Y_*) \propto p(Y_*|\theta_*)p(\theta_*|Y_1, Y_2, \cdots, Y_H)$.

16.3.3 MAP approach

Each Y_h is assumed to be available estimates for trial specific parameters $\theta_1, \ldots, \theta_H$. In the MAP approach the historical data is used to predict the effect size estimate to be observed in the actual trial (θ^*).

16.3.3.1 Likelihood and prior

The historical data Y_h from n_h patients in the h-th trial are distributed as

$$Y_h|\theta_h \sim N(\theta_h, s_h^2), \quad h = 1, 2, \cdots, H. \tag{16.1}$$

The similarity of new and historical trials is expressed by the following prior

$$\theta_1, \theta_2, \ldots, \theta_H, \theta^*|\mu, \tau^2 \sim N(\mu, \tau^2) \tag{16.2}$$

Further locally uniform prior is assumed for μ which for known between trial variance (τ^2) results in the predictive distribution of interest -

$$\theta^*|Y_1, Y_2, \ldots Y_H, \tau \sim N\left(\frac{\sum w_h Y_h}{\sum w_h}, \frac{1}{\sum w_h} + \tau^2\right), \tag{16.3}$$

where $w_h = \frac{1}{s_h^2 + \tau^2}$. This shows how the heterogeneity parameter (τ) controls the information on θ^* borrowed from the historical trials data.

In a random-effects model, the prior distribution for the heterogeneity parameter (τ) is taken to control the degree of prior belief on the relevance of the historical data. For example on the log-OR scale, using a Half-Normal with standard deviation 1 puts around 5% probability to $\tau > 2$ which correspond to 5% chance that the historical data carries no relevance about θ^* in the new trial. The tool implements such prior for τ with two options:

- **High heterogeneity**: 5% probability that historical data has no relevance about θ^*

- **Low heterogeneity**: 5% probability that historical data has no relevance about θ^*
- **Known** τ: A user defined known between trial standard deviation can also be used.

The tool implements hierarchical model and offers two options for the prior for μ:

- **Skeptical**: This puts around 5% chance that the effect size exceeds the minimal clinically relevant effect size.
- **Enthusiastic**: This ensures around 5% chance that the effect size is less or equal to zero.

Based on the above equations 16.1 and 16.2 and hyper-prior, MAP prior distribution $p_H(\theta_*) \equiv p(\theta_*|Y_1, Y_2, \cdots, Y_H)$ for the new trial can be derived. Markov Chain Monte Carlo (MCMC) samples can be generated as $\theta_*^{(1)}, \theta_*^{(2)}, \cdots, \theta_*^{(M)}$, where M is the number of samples.

16.3.3.2 Approximation of MAP Prior

A kernel-density estimate from the MCMC samples can be used to describe the MAP prior. But there are practical disadvantages of working with such density estimate due to a large number of parameters. An approximated, compact and tractable representation is a mixture of conjugate priors (Schmidli et al., 2014), e.g., for normal endpoint, the mixture prior can be written as

$$\hat{p}_H(\theta_*) \equiv \hat{p}(\theta_*|Y_1, Y_2, \cdots, Y_H) = \sum_{k=1}^{K} w_k \, N(\theta_*|\mu_k, \sigma_k^2),$$

such that $\sum_{k=1}^{K} w_k = 1$. According to Diaconis (1985), any prior can be closely approximated in this way. The number of components, K, the weights of the mixture components $\{w_k\}_{k=1}^{K}$ and the corresponding hyperparameters need to be specified in order to derive this closed-form representation. The Kullback-Leibler (KL) divergence is used to compute the distance between the exact MAP prior $p_H(\theta_*)$ and the approximated MAP prior $\hat{p}_H(\theta_*)$. The KL divergence is written as

$$KL(p_H(\theta_*), \hat{p}_H(\theta_*)) = \int log\{p_H(\theta_*)\} \, p_H(\theta_*) \, d\theta_* - \int log\{\hat{p}_H(\theta_*)\} \, p_H(\theta_*) \, d\theta_*. \qquad (16.4)$$

The best approximation in terms of mixing weights $\{w_k\}_{k=1}^{K}$ and hyperparameters of the conjugate prior can be obtained by maximizing the second term of the equation 16.4. Note that a Monte-Carlo estimate of the integral is given by $\frac{1}{M} \sum_{i=1}^{M} log\{\hat{p}_H(\theta_*^{(i)})\}$ where $\theta_*^{(i)}$ is a sample from the posterior distribution. This term is identical to the log-likelihood of the MCMC sample with mixture model $\hat{p}_H(\theta_*)$ and hence those mixing weights and hyperparameters can also be obtained as

maximum-likelihood estimates (Schmidli et al., 2014). Currently the choice of the number of components, K can be done numerically. The entire approximation can also be carried out as a non-parametric Bayesian density estimation process using mixture of Dirichlet process or mixture of Polya trees (Hjort et al., 2010; Müller et al., 2015).

16.3.3.3 Robustification of the MAP prior

Use of historical data in a new trial requires careful selection of the historical trials because exchangeability of the parameters is an important assumption here. In order to address the possible prior-data conflict, a robust version of MAP prior is often preferred, where the MAP prior is added with a vague non-informative or weakly informative conjugate prior (Schmidli et al., 2014) and expressed as

$$\hat{p}_{HR}(\theta_*) = (1 - w_R)\hat{p}_H(\theta_*) + w_R p_R(\theta_*) \tag{16.5}$$

where $\hat{p}_H(\theta_*)$ is the approximated MAP prior and $p_R(\theta_*)$ is the weakly informative or non-informative conjugate prior. w_R is the prior probability that new trial systematically differs from historical trials. The choice of w_R in equation 16.5 determined how quickly the effect of historical data goes away with the increase of prior-data conflict. When historical data and new data are in clear conflict, the prior is discarded if the MAP prior is robust. Also note that, if this vague prior is proper then the mixing weight can be interpreted as a probability, but for an improper flat prior, it won't be the case. From the above equation 16.5 we can see that the robust MAP prior is again a mixture of conjugate priors, therefore the posterior is also a mixture of conjugate posteriors with updated mixture weights.

16.3.3.4 Effective sample size(ESS) of the robust MAP prior

While borrowing strength from historical trial information, it is useful to quantify the prior effective sample size (ESS). For conjugate priors, the ESS is relatively easy to obtain for the exponential family of distributions. For example, for binary endpoints ESS = $a + b$ for the prior $Beta(a, b)$. For non-conjugate prior, normal approximations can be used (Morita et al., 2008). The ESS is the sample size such that the expected information of the posterior under a non-informative prior is the same as the information of the informative prior $p(\theta_*)$ where the information is evaluated at the mode of the informative prior (Schmidli et al., 2014).

17. Adaptive design with MAP Prior (Binary Outcome)

17.1 Introduction

This module briefly describes a Bayesian adaptive design with Meta-Analytic-Predictive (MAP) prior (Binary Outcome) (Schmidli et al., 2014), which utilize historical data of the past clinical trials. This interface provides an easier way to perform trial simulation and examine the operating characteristics of the adaptive design with MAP prior (Binary Outcome).

Historical information is useful for clinical trial design as it can save time and effort, and reduce the number of subjects. This data must be used in an appropriate way as it might create a prior data conflict, i.e., bias.

One way to introduce data is to include it from previous studies, which might provide useful information about potential treatment effects and/or variability for the control group in a new study. This can be used for sample size and power calculations. The historical control group information also can be incorporated into analyses of treatment group effects.

Sometimes past information might not be relevant for the new trial. Hence there is a need to introduce few subjects or the data from the current phase to derive correct results. This can be overcome by deriving a Bayesian meta-analytic-predictive prior from historical data, which is then combined with the new data.

For more information on how to derive the Bayesian-meta-analytic prior, see **Meta-Analytic-Predictive (MAP) prior generation** section in **Meta-Analytic-Predictive (MAP) Prior** help.

This generated MAP prior is further robustified, with a mixture of two components. The first component is derived from historical data and the second component robustifies against the prior-data conflict.

The **Adaptive Design with MAP Prior (Binary Outcome)** - feature can be used extensively

in phase II clinical trials. You can use this feature to check the efficacy of the investigated drug in the treatment arm against a marketed drug in the control arm. In this type of trial, patients are enrolled and randomized to the two arms by a fixed ratio.

The efficacy observation is completed by concluding whether the investigated drug (treatment arm) is better than the marketed drug (control arm) based on the data collected in the trial. Adaptive design with Meta Analytic Prior uses information collected from historical trials of the marketed drug in the control arm.

In §17.2 of this document we introduce the interface for creating and launching new adaptive designs with MAP prior from historical data and comparing those. Statistical method overview is given in §17.3.

17.2 User Interface and Tutorial

Adaptive Design with MAP Prior (Binary Outcome) has three tabs, **Setup**, **Results**, and **Result Details**.

17.2.1 Setup

17.2.1.1 Design

You can add a maximum of four designs using this section. In the **Design** section enter the parameters in **Planned Sample Size**, **Control Arm Prior**, and **Treatment Arm Prior**.

Planned Sample Size

Enter the parameters **Interim Look** and **Final** for **Control Arm** and **Treatment Arm** in the **Planned Sample Size** section.

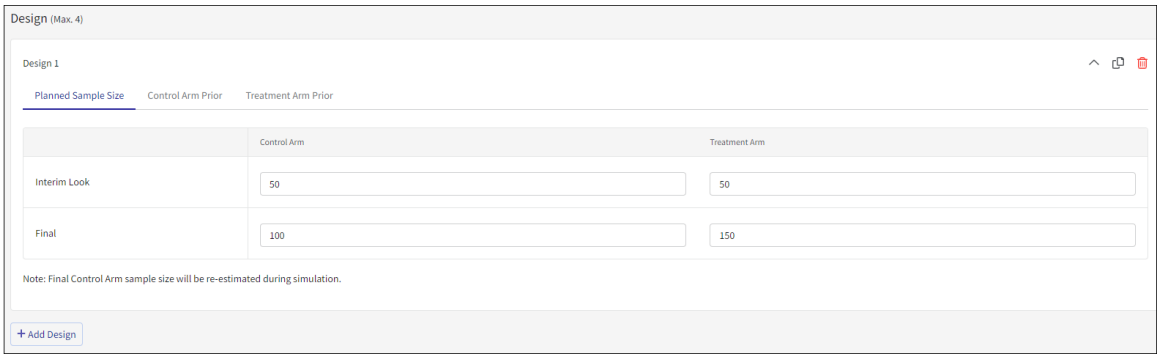


Figure 17.1: Designing Simulation – Planned Sample Size

Table 17.1: Designing Simulation – Planned Sample Size

Input Parameter	Range	Data Type
Interim Look Control Arm sample size	[1,10000]	Integer
Interim Look Treatment Arm sample size	[1,10000]	Integer
Final Control Arm sample size	[1,10000]	Integer
Final Treatment Arm sample size	[1,10000]	Integer

Final Control Arm sample size is re-estimated during simulation.

After entering details in the section, you can move to the **Control Arm Prior**.

Control Arm Prior

Enter the details in the **Control Arm Prior** section in the following order:

In **Prior Generation Parameters** section (Table 17.2),

1. Select Heterogeneity, Add Robust Component and enter Component Weight (if Robust Component is selected).
2. Select Mixture Components, enter No. of MCMC Runs.
3. Select ESS Computational methods.

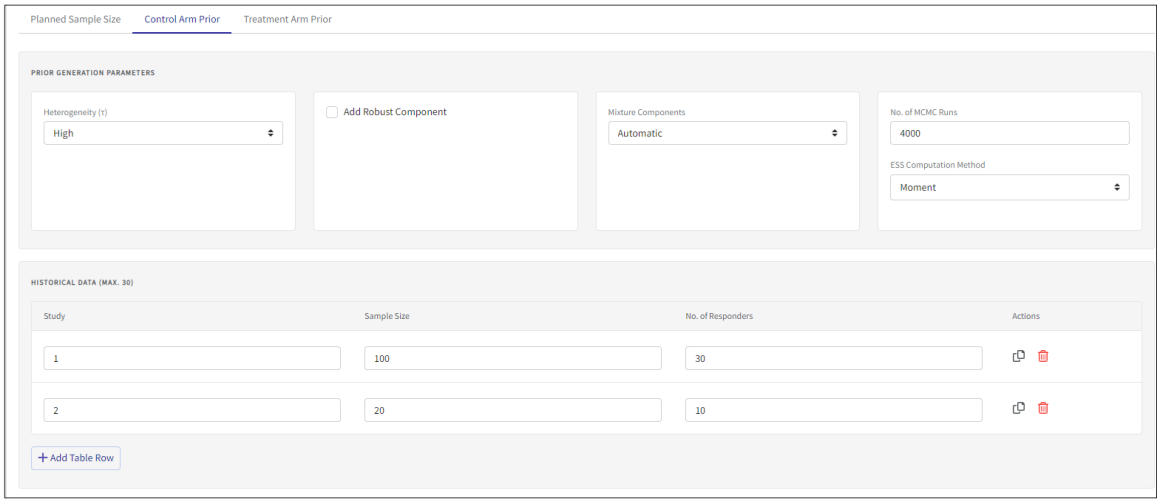


Figure 17.2: Designing Simulation – Control Arm Prior

For more information on **Heterogeneity**, **Add Robust Component**, **Number of Mixture Components**, **No. of MCMC Runs**, and **ESS Computation Method**, see **Meta-Analytic-Predictive (MAP) prior generation** section in **Meta-Analytic-Predictive (MAP) Prior** help.

In **Historical Data** section, enter the parameters for Historical Data (Table 17.3).

Table 17.2: Designing Simulation – Control Arm Prior

Input Parameter	**Description**	**Range**	**Data Type**
Heterogeneity (τ)	Prior for the between-trial heterogeneity (τ) of the random effects meta analytic model.	{High, Low, Known}	Categorical
If **Heterogeneity (τ)** is selected as **Known** τ, enter τ as follows:			
Heterogeneity (τ)	Between trial standard deviation	[1e-6, 1e+6]	Numeric
Component weight	Weight of robust component	(0,1)	Numeric
Number of Mixture Components		{Automatic, Manual}	Categorical
If **Number of Mixture Components** is selected as **Manual**, enter the **Number of Mixture Components** as follows:			
Number of Mixture Components	Components to fit a mixture model	[1, 10]	Integer
No. of MCMC Runs	MCMC iterations	[4000, 15000]	Integer
ESS Computation Method	Options to calculate effective sample size	{Moment, Morita}	Categorical

Table 17.3: Designing Simulation – Control Arm Prior – Historical Data

Input Parameter	**Description**	**Range**	**Data Type**
Study	Study Name	30 Characters	String
Sample Size	Sample size in each historical trial	[1, 10000]	Integer
No. of Responses	Number of patients with responses in each historical trial	[0, Sample Size]	Integer

Treatment Arm Prior

In the Treatment Arm Prior section,

1. Enter the response rate for treatment arm as α_t and β_t.

Figure 17.3: Designing Simulation – Treatment Arm Prior

Table 17.4: Designing Simulation – Treatment Arm Prior

Input Parameter	Description	Range	Data Type
α_t	Shape parameter for beta prior distribution of response rate in the treatment arm	[1e-6, 1e+6]	Numeric
β_t	Shape parameter for beta prior distribution of response rate in the treatment arm	[1e-6, 1e+6]	Numeric

Once this step is completed, move to the **Scenarios** section.

17.2.1.2 Scenarios

You can add scenarios using the following modes:

- **Add Auto**: Values for true response rate are added by default. By default, true response rate of control arm p_c is 0.2.
- **Add Manual**: Enter the values for true response rate of treatment arm p_t manually for each scenario. In the first scenario, the value you enter for true response rate of treatment arm p_t is also added in control arm p_c field as it must be same in scenario 1. Here, value for control arm p_c is carried over for the next scenarios as well.

You can add a maximum of four scenarios. Copy or delete a scenario using the "Copy" and "Delete" icons under **Actions**.

Figure 17.4: Adding Scenarios

Table 17.5: Adding Scenarios

Input Parameter	Description	Range	Data Type
p_c	True response rate of control arm	(0,1)	Numeric
p_t	True response rate of treatment arm	(0,1)	Numeric

17.2.1.3 Simulation Parameters

Enter the **Simulation Parameters** as follows:

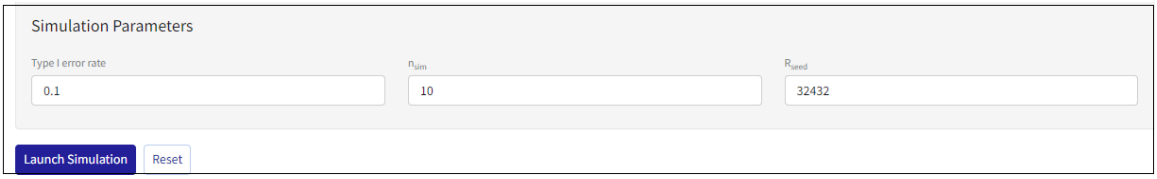


Figure 17.5: Simulation Parameters

Table 17.6: Simulation Parameters

Input Parameter	Description	Range	Data Type
Type I error rate	Type I error rate	(0, 1)	Numeric
n_{sim}	Number of simulations	[1,10000]	Integer
R_{seed}	Simulation seed value	[1,1e+6]	Integer

17.2.2 Results

After you launch the simulation, you can check the status in the **Results** tab.

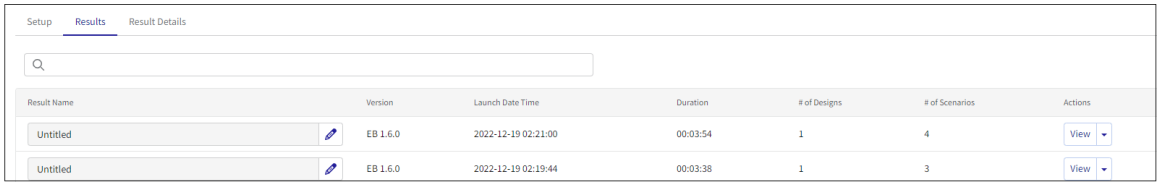


Figure 17.6: Results

You can provide a specific result name for each simulation. In the above screenshot, all the details related to the simulation are displayed; like **Launch Date Time**, **Duration**, **# of Designs**, **# of Duration**, **Actions** (**Download Report** and **Delete Simulation Result**).

Click **View** under the **Actions** column in the **Results** tab to view Simulation Output. The detailed output is available in the **Result Details** tab.

17.2.3 Result Details

The following output parameters are available for the **Result Details** tab:

- **Summary of Performance**
- **Intermediate Output: MAP Priors**

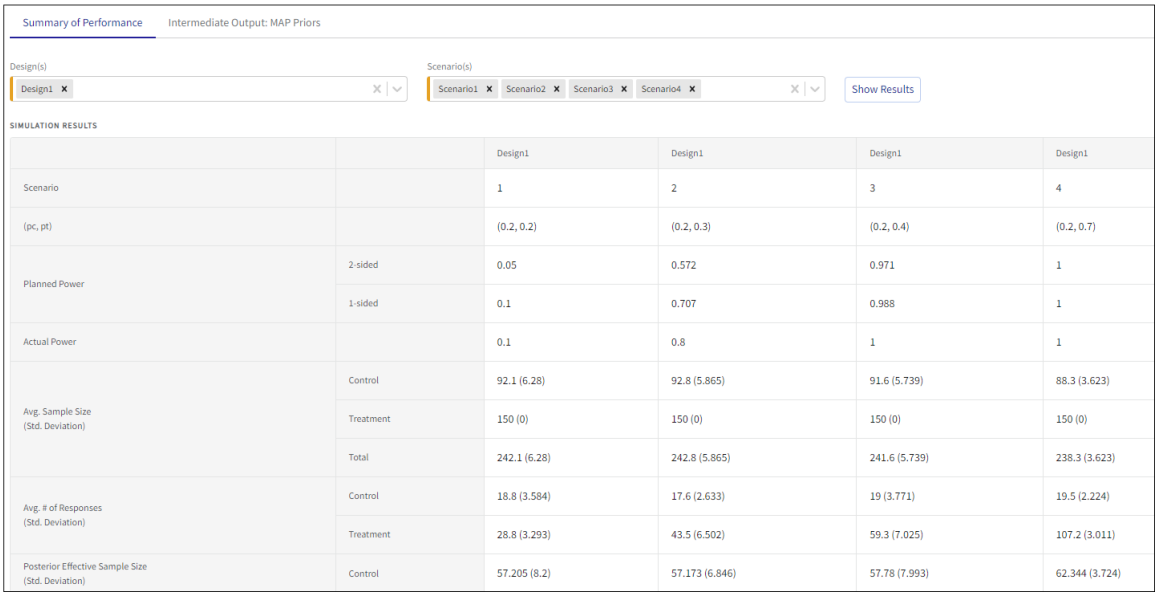
Summary of Performance

- **Simulation Results**

Select the design and scenarios in the following drop-down box to generate the tables for respective design and scenario.

Table 17.7: Simulation Results

Output Parameter	Description	Data Type
Planned Power (2-sided)	The 2-sided power using two-sample z-test given planned sample size, response rates of treatment and control arms, and type I error rate	Numeric
Planned Power (1-sided)	The 1-sided power using two-sample z-test given planned sample size, response rates of treatment and control arms, and type I error rate	Numeric
Actual Power	The proportion of simulated trials in which the treatment arm produces more desired result than the control arm control.	Numeric
Avg. Sample Size (Std. Deviation) – Control	The average number of patients treated at the control arm in a simulated trial and its standard deviation, averaging across all the simulated trials.	Numeric
Avg. Sample Size (Std. Deviation) – Treatment	The average number of patients treated at the treatment arm in a simulated trial and its standard deviation, averaging across all the simulated trials.	Numeric
Avg. Sample Size (Std. Deviation) – Total	The average number of patients treated in a simulated trial and its standard deviation, averaging across all the simulated trials.	Numeric
Avg. # of Responses (Std. Deviation) – Control	The average number of patients which experience efficacy outcome at the control arm in a simulated trial and its standard deviation, averaging across all the simulated trials.	Numeric
Avg. # of Responses (Std. Deviation) – Total	The average number of patients, which experience efficacy outcome at the treatment arm in a simulated trial and its standard deviation, averaging across all the simulated trials.	Numeric
Posterior Effiective Sample Size (Std. Deviation) – Control	The average posterior effective sample size of the control arm in a simulated trial using the computation method (Moment/Morita) which is specified by users in input page after the interim look is completed.	Numeric



Summary of Performance | Intermediate Output: MAP Priors

Design(s): Design1 ×

Scenario(s): Scenario1 × Scenario2 × Scenario3 × Scenario4 ×

Show Results

SIMULATION RESULTS

		Design1	Design1	Design1	Design1
Scenario		1	2	3	4
(pc, pt)		(0.2, 0.2)	(0.2, 0.3)	(0.2, 0.4)	(0.2, 0.7)
Planned Power	2-sided	0.05	0.572	0.971	1
	1-sided	0.1	0.707	0.988	1
Actual Power		0.1	0.8	1	1
Avg. Sample Size (Std. Deviation)	Control	92.1 (6.28)	92.8 (5.865)	91.6 (5.739)	88.3 (3.623)
	Treatment	150 (0)	150 (0)	150 (0)	150 (0)
	Total	242.1 (6.28)	242.8 (5.865)	241.6 (5.739)	238.3 (3.623)
Avg. # of Responses (Std. Deviation)	Control	18.8 (3.584)	17.6 (2.633)	19 (3.771)	19.5 (2.224)
	Treatment	28.8 (3.293)	43.5 (6.502)	59.3 (7.025)	107.2 (3.011)
Posterior Effective Sample Size (Std. Deviation)	Control	57.205 (8.2)	57.173 (6.846)	57.78 (7.993)	62.344 (3.724)

Figure 17.7: Results Details – Summary of Performance

- **Power Comparison**

Power Comparison graph is displayed based on **Design** as well as **Scenarios**.

Graph based on **Designs** is displayed on **Result Details** tab in **Simulation Output** section as follows: (Figure 17.8)

Graph based on **Scenarios** is displayed on **Result Details** tab in **Simulation Output** section as follows: (Figure 17.9)

Interpretation: Actual Power denotes the proportion of simulated trials in which the treatment arm is doing better than control. For scenario 1, planned power (1-sided) must be type I error rate that you enter (planned power (2-sided) must be half of type I error rate that you enter). Note that this is a null scenario. And the result of scenario 1 is used to calibrate the threshold of declaring promising treatment arm. For other scenarios, if actual power is larger than planned power, that shows the design has its strength, and it does utilize historical data.

Figure 17.8: Power Comparison - Designs

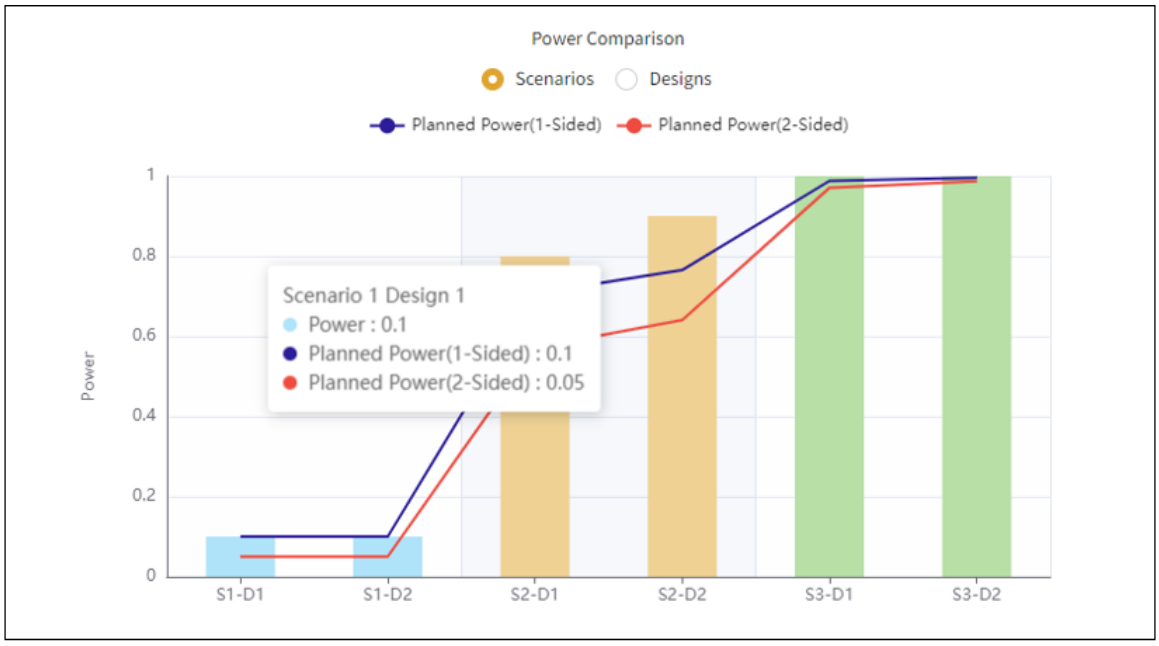


Figure 17.9: Power Comparison - Scenarios

- **Sample Size Comparison (Control Arm)**

Sample Size Comparison (Control Arm) graph is displayed based on **Design** as well as **Scenarios** (Same as **Power Comparison**).

Graph based on **Designs** is displayed on **Result Details** tab in **Simulation Output** section as follows: (Figure 17.10)

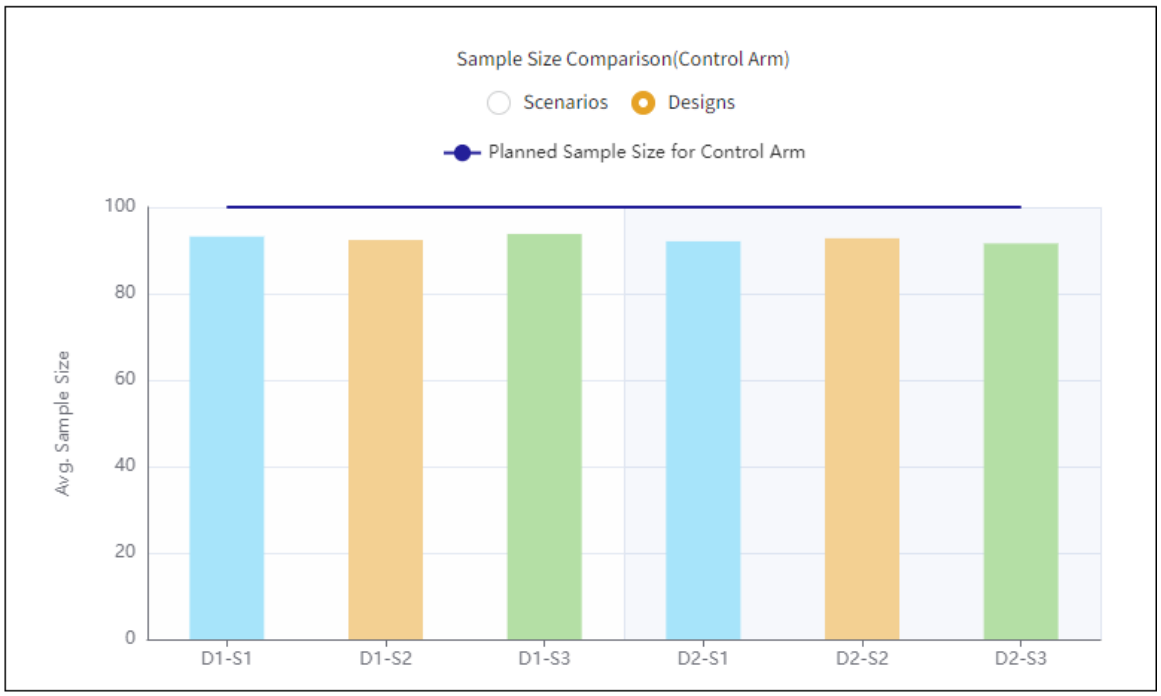


Figure 17.10: Sample Size Comparison (Control Arm) - Designs

Graph based on **Scenarios** is displayed on **Result Details** tab in **Simulation Output** section as follows: (Figure 17.11)

Interpretation: When the actual power is larger than or equal to planned power in a scenario, the lower the actual sample size than planned sample size, the better the design.

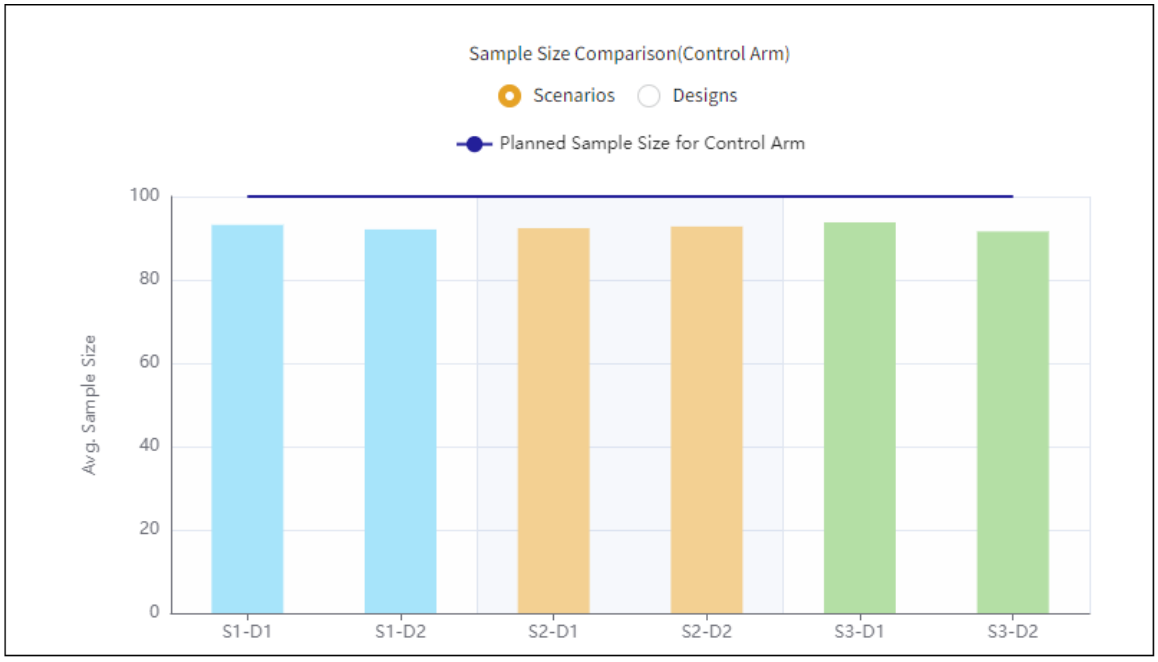


Figure 17.11: Sample Size Comparison (Control Arm) - Scenarios

Intermediate Output: MAP Priors

- **MAP Prior Visualization**

Density plot of the mixture of beta distributions for each design (Figure 17.12).

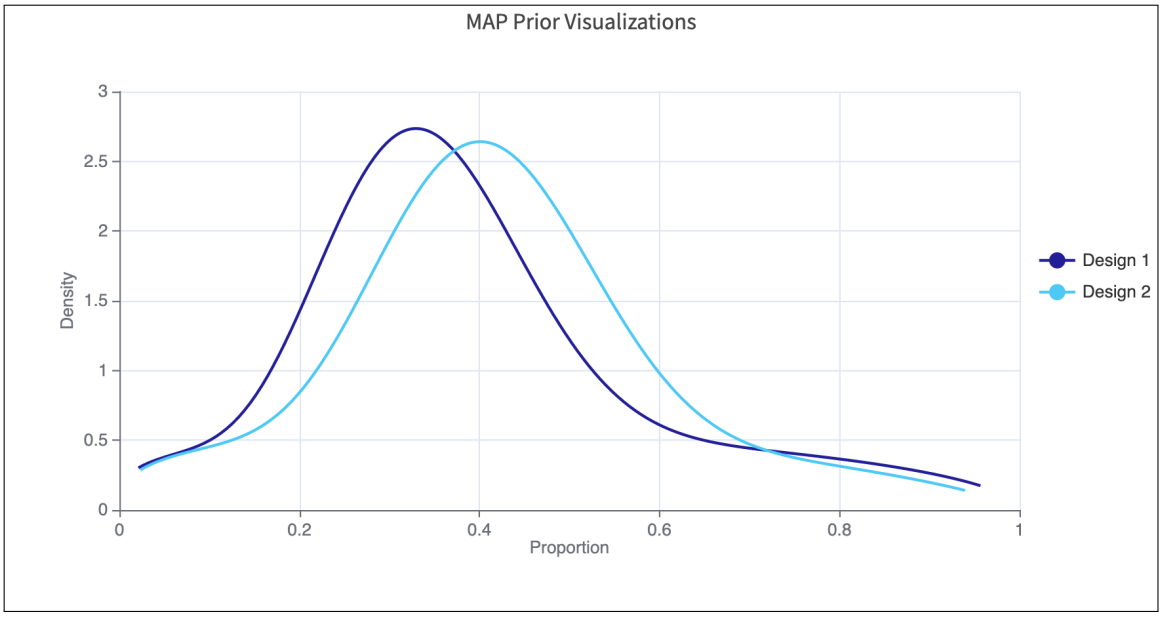


Figure 17.12: MAP Prior Visualization

- **MAP Prior Comparison**

The output is available in **Intermediate Output: MAP Priors** on **Result Details** tab in **Simulation Output** section as follows:

Table 17.8: MAP Prior Comparison

Output Parameter	**Description**	**Data Type**
Heterogeneity Prior	Prior for the between-trial heterogeneity (τ) of the random effects meta analytic model.	Numeric
No. of Components	The number of components to fit a mixture model. (Specify in input page if **Manual** selected; otherwise, its computed automatically)	Integer
Robust	Weight of robust component. (If selecting to add robust component and specify the weight of robust component in input page, its numeric; otherwise, NA)	Numeric or NA
Effective Sample Size	Computed effective sample size of the map prior on the control arm using the computation method (Moment/Morita).	String

- **Computed RMAP Prior Parameters**

Components in computed MAP prior for each design are displayed as follows:

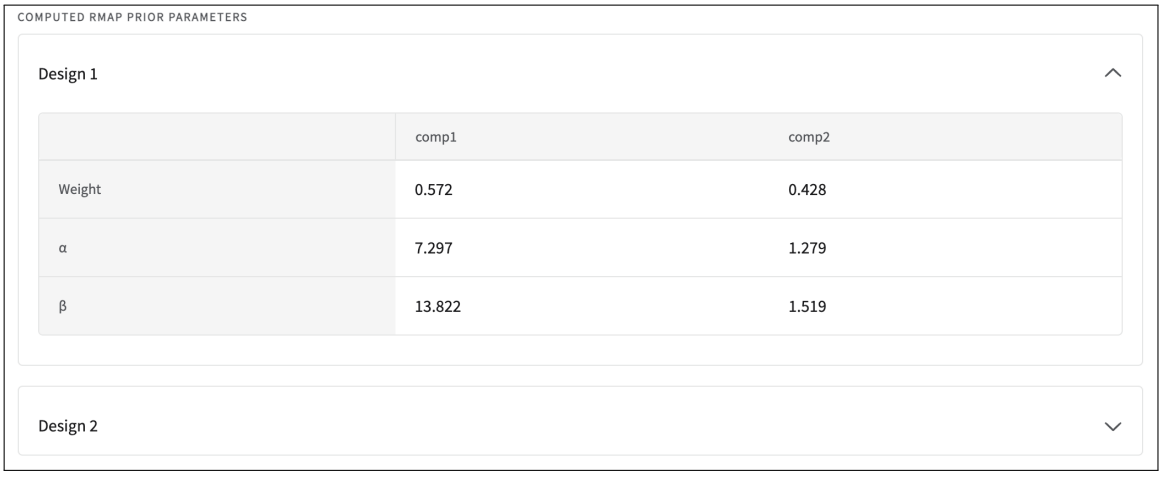


Figure 17.13: Computed RMAP Prior Parameters

Table 17.9: Computed RMAP Prior Parameters

Output Parameter	**Description**	**Data Type**
Weight	Weight of each beta distribution in the mixture for each design	Numeric
α	Shape parameter of each beta distribution in the mixture for each design	Numeric
β	Shape parameter of each beta distribution in the mixture for each design	Numeric

17.3 Statistical Method Review

17.3.1 Adaptive design with MAP prior (Binary Outcome)

Consider two arms, treatment and control, in the innovative design. Number of responses in the control and treatment arms, y_c and y_t, are generated through binomial distributions:

$$y_c \sim B(n_c, p_c) \quad \text{and} \quad y_t \sim B(n_t, p_t).$$

Specify the number of patients in control and treatment arms; denoted as n_c and n_t, and true response rates as p_c and p_t respectively.

Towards the end of the trial, posterior probability is calculated to check whether treatment arm is better than the control arm with different priors imposed on response rates of treatment and control.

For the treatment arm, you can assume p_t, following a simple beta distribution $Beta(\alpha_t, \beta_t)$ and for the control arm, p_c to follow a mixture of beta distributions computed through a MAP prior method.

You can consider one interim analysis in the design. In this case, n_c and n_t are split into n_{1c}, n_{2c} and n_{1t}, n_{2t}, i.e.,

$$n_c = n_{1c} + n_{2c} \quad \text{and} \quad n_t = n_{1t} + n_{2t}.$$

Enter the n_{1c} and n_{1t} as planned sample sizes at the interim analysis and n_c and n_t as planned sample sizes at the final of the trial.

At the interim analysis, n_{2c} is re-estimated using the posterior effective sample size computed by the **Moment** or **Morita** (Morita et al., 2008) methods.

Let ESS_I denotes the posterior effective sample size based on data of the control arm after the interim analysis with a MAP prior. The re-estimated sample size for the control arm after the interim analysis is calculated as

$$\hat{n}_{2c} = \max(n_c - ESS_I, n_{\min})$$

Default value of $n_{\min}$) is fixed as 5.

In most cases, ESS_I is larger than n_{1c} as the computation of ESS_I is based on information from both n_{1c} patients in the control arm and the MAP prior. However, in some special cases, the MAP prior may differ a lot from the actual data of the control arm, which results in $ESS_I < n_{1c}$ and also $\hat{n}_{2c} > n_{2c}$.

In that case, the MAP prior and $\hat{n}_{2c}$ cannot be used in the rest of the trial. In the engine, you can enroll n_{2c} patients for the control arm and make a final decision with a beta prior $Beta(\alpha_c, \beta_c)$ imposed on p_c, where $\alpha_c = \beta_c = 0.5$ by default in East Bayes.

Part VII

Sample Size Calculation

18. Sample Size Calculation for Binary Outcome

In this Module, we implement the sample size calculation for binary endpoint, which include the following functions shown in Table 18.1.

Table 18.1: Function implementation in sample size calculation for binary endpoint.

Number of Arms	Test Objectives	One- or/and Two-sided	Contents	Section
One	Equality	One-sided & Two-sided	Z-test	Section 18.1.1
	Equivalence	-	Z-test	Section 18.1.2
	Non-inferiority	-	Z-test	Section 18.1.3
	Superiority	-	Z-test	Section 18.1.3
	Agreement	One-sided & Two-sided	Cohen's Kappa	Section 18.1.4
Two (independent)	Equality	One-sided & Two-sided	Z-test	Section 18.2.1
	Equivalence	-	Z-test	Section 18.2.2
	Non-inferiority	-	Z-test	Section 18.2.3
	Superiority	-	Z-test	Section 18.2.3
Two (paired)	McNemar's test	One-sided & Two-sided		Section 18.3

18.1 Single arm

Let $x_i, i = 1, \cdots, n$ be the binary response observed from ith subject. In clinical research, x_i could be the indicator for the response of tumor in cancer trials, i.e., $x_i = 1$ for responder or $x_i = 0$ for non-responder. It is assumed that x_i's are i.i.d. with $P(x_i = 1) = p$, where p is the true response

rate. Since p is unknown, it is usually estimated by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Also, let $\epsilon = p - p_0$ be the difference between the true response rate of a test drug (p) and a reference value (p_0). In practice, it is of interest to test for equality (i.e., $p = p_0$), non-inferiority (i.e., $p - p_0$ is greater than or equal to a pre-determined non-inferiority margin), superiority (i.e., $p - p_0$ is greater than a pre-determined superiority margin), and equivalence (i.e., the absolute difference between p and p_0 is within a difference of clinical importance). The following are details of sample size calculation with single arm.

18.1.1 Test Objective: Equality

18.1.1.1 Methods

- **Hypothesis:** To test whether there is a difference between the true response rate of the test drug and the reference value, the following hypotheses are usually considered,

  $(\boldsymbol{Two-sided})$

  $$H_0 : \epsilon = 0 \quad versus \quad H_1 : \epsilon \neq 0$$

  $(\boldsymbol{One-sided})$

  $$H_0 : \epsilon \leq 0 \quad versus \quad H_1 : \epsilon > 0$$

- **Formula:** Using the value of p to compute the standard deviation in z-test statistic, we can get sample size n from,

  $(\boldsymbol{Two-sided})$

  $$n = \frac{(z_{\alpha/2} + z_\beta)^2 p(1-p)}{\epsilon^2}$$

  $(\boldsymbol{One-sided})$

  $$n = \frac{(z_{\alpha} + z_\beta)^2 p(1-p)}{\epsilon^2}$$

  where z_α is the upper αth quantile of the standard normal distribution.

18.1.1.2 Input and Output

- **Input:**

  1. p_0: a reference value(response rate for the historical control)

2. p: true response rate of the test drug
3. α: type I error rate
4. β: type II error rate (Power: $1-\beta$)

- **Output:** sample size n

18.1.1.3 An Example (Single-arm Equality Two-sided Test)

Suppose that the response rate of the patient population under study after treatment by a test drug is expected to be around 50% (i.e., $p = 0.50$). At $\alpha = 0.05$, the required sample size for having an 80% power (i.e., $1-\beta = 0.8$) for correctly detecting a difference between the post-treatment response rate and the reference value of 30% (i.e., $p_0 = 0.30$) can be obtained by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: One**, **Test Objective: Equality** and **1 or 2 Sided Test: 2-Sided**.
- Input $\boldsymbol{p_0, p, \alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

Then the computed sample size is 50 using Z-test in this situation, shown in Figure 18.1.

18.1.2 Test Objective: Equivalence

18.1.2.1 Methods

- **Hypothesis:** To establish equivalence, the following hypotheses are usually considered,

$$H_0 : |p - p_0| \geq \delta \quad versus \quad H_1 : |p - p_0| < \delta,$$

or

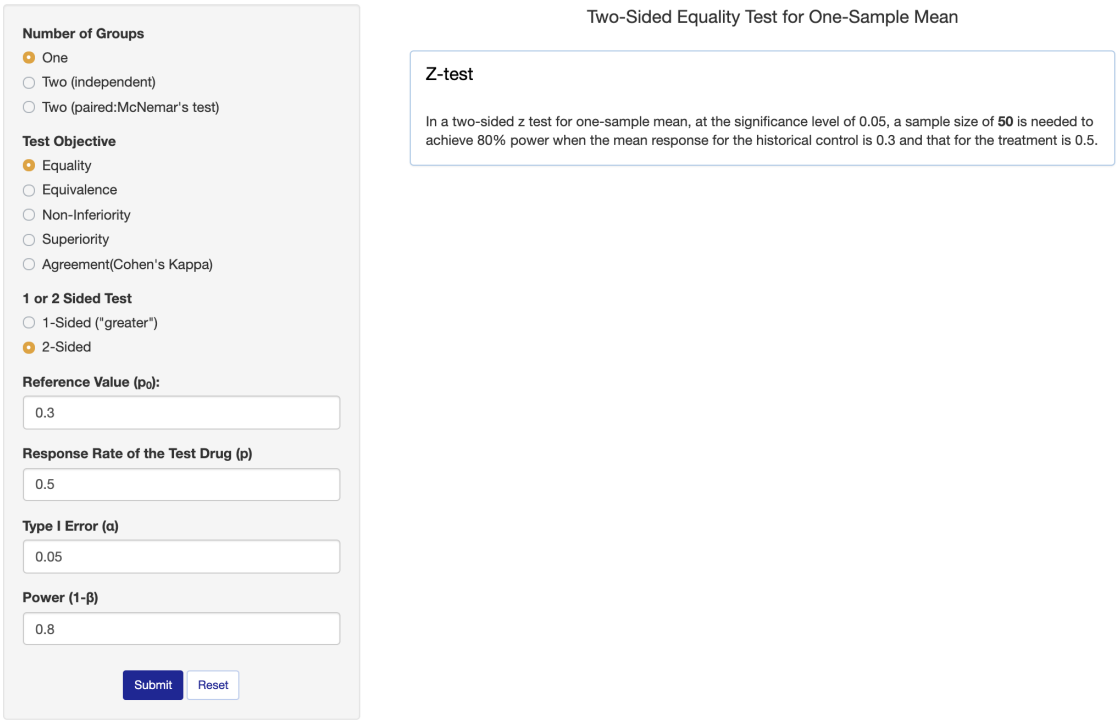
$$H_0 : |\epsilon| \geq \delta \quad versus \quad H_1 : |\epsilon| < \delta,$$

The proportion of the responses is concluded to be equivalent to the reference value of p_0 if the null hypothesis is rejected at a given significance level.

- **Formula:** Using the value of p to compute the standard deviation in z-test statistic, we can get sample size n from,

$$n = \frac{(z_\alpha + z_{\beta/2})^2 p(1-p)}{(\delta - |\epsilon|)^2}$$

Figure 18.1: An Example (Single-arm Equality Two-sided Test)



18.1.2.2 Input and Output

- **Input:**
  1. δ ($\delta > 0$): equivalence margin
  2. p_0: a reference value
  3. p: true response rate of the test drug
  4. α: type I error rate
  5. β: type II error rate (Power: $1 - \beta$)
- **Output:** sample size n

18.1.2.3 An Example (Single-arm Equivalence Test)

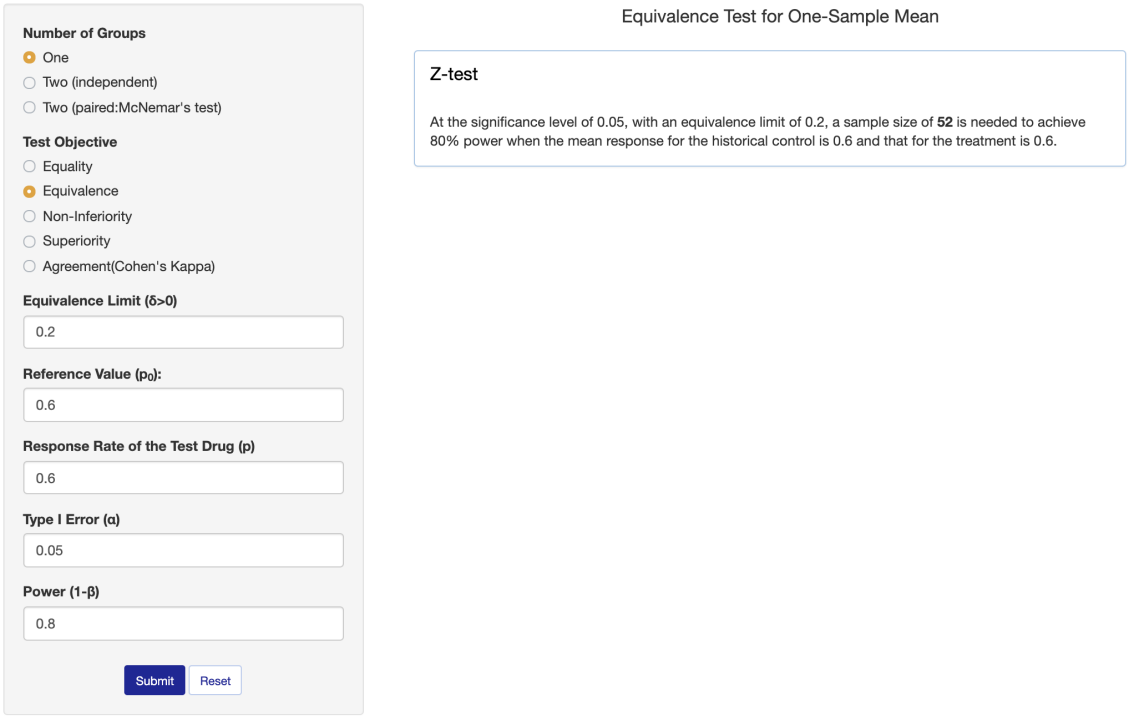
Assume that one brand name drug for a certain disease on the market has a responder rate of 60% (i.e., $p_0 = 0.60$). It is believed that a 20% difference in responder rate is of no clinical significance (i.e., $\delta = 0.2$). Hence, the investigator wants to show the study drug is equivalent to the market drug

in terms of responder rate. At $\alpha = 0.05$, assuming that the true response rate is 60% (i.e., $p = 0.60$), the sample size required for achieving an 80% power can be obtained by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: One** and **Test Objective: Equivalence**.
- Input δ, p_0, p, α and $1 - \beta$.
- Click **Submit**.

Then the computed sample size is 52 using Z-test in this situation, shown in Figure 18.2.

Figure 18.2: An Example (Single-arm Equivalence Test)



18.1.3 Test Objective: Non-Inferiority/Superiority

18.1.3.1 Methods

- **Hypothesis:** The problem of testing non-inferiority and superiority can be translated into the following hypotheses,

$(\boldsymbol{Noninferiority})$

$$H_0 : \epsilon \leq -\delta \quad versus \quad H_1 : \epsilon > -\delta$$

$(\boldsymbol{Superiority})$

$$H_0 : \epsilon \leq \delta \quad versus \quad H_1 : \epsilon > \delta$$

where δ ($\delta > 0$) is the superiority or non-inferiority margin.

- **Formula:** Using the value of p to compute the standard deviation in z-test statistic, we can get sample size n from,
  $(\boldsymbol{Noninferiority})$
  $$n = \frac{(z_\alpha + z_\beta)^2 p(1-p)}{(\epsilon + \delta)^2}.$$
  $(\boldsymbol{Superiority})$
  $$n = \frac{(z_\alpha + z_\beta)^2 p(1-p)}{(\epsilon - \delta)^2}.$$

18.1.3.2 Input and Output

- **Input:**
  1. δ ($\delta > 0$): non-inferiority or superiority margin
  2. p_0: a reference value
  3. p: true response rate of the test drug
  4. α: type I error rate
  5. β: type II error rate (Power: $1 - \beta$)
- **Output:** sample size n

18.1.3.3 An Example (Single-arm Non-Inferiority Test)

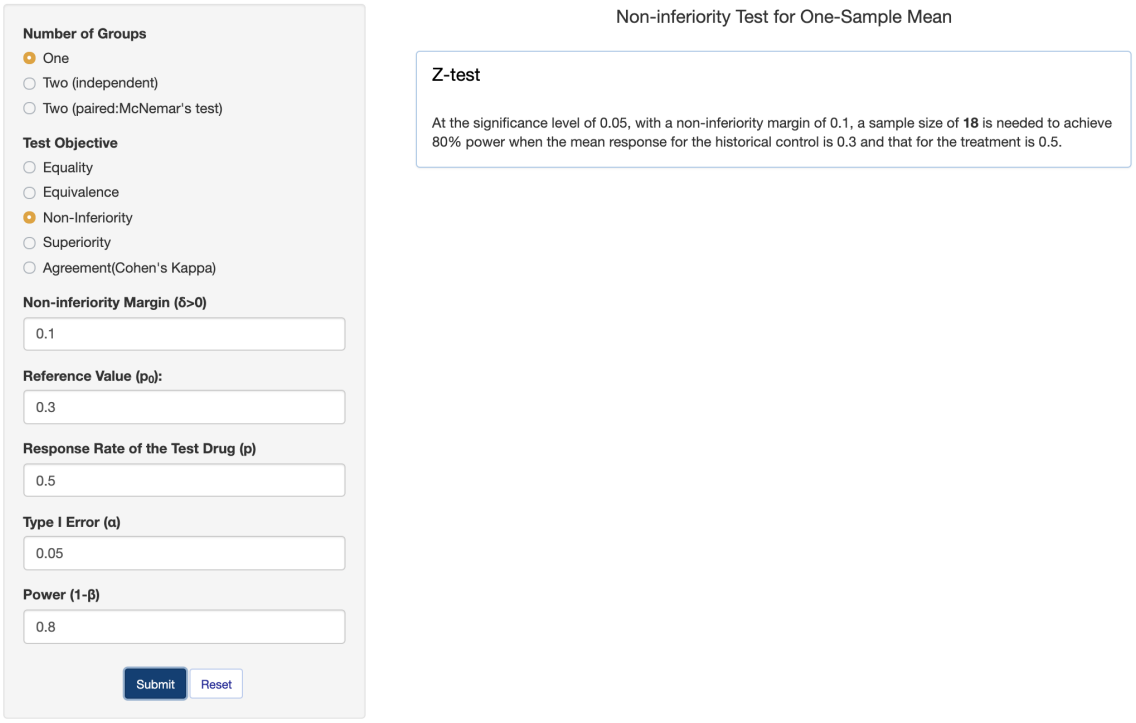
For a certain disease, we wish to show that the majority of patients whose change after treatment by a test drug is at least as good as the reference value (30%) ($p_0 = 0.3$). Also assume that a difference of 10% in responder rate is considered of no clinical significance ($\delta = 0.1$). Assume the true response rate is 50% ($p = 0.5$). At $\alpha = 0.05$, the required sample size for having an 80% power (i.e., $1 - \beta = 0.8$) can be obtained by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.

- Select **Number of Groups: One** and **Test Objective: Non-Inferiority**.
- Input δ, p_0, p, α and $1-\beta$.
- Click **Submit**.

Then the computed sample size is 18 using Z-test in this situation, shown in Figure 18.3.

Figure 18.3: An Example (Single-arm Non-Inferiority Test)



18.1.4 Cohen's Kappa

In some clinical trials, to check inter-rater reliability, independent sets of measurements are taken by more than one rater and the responses are checked for agreement. For a binary response, Cohens Kappa test can be used to check inter-rater reliability. Conventionally, the kappa coefficient is used to express the degree of agreement between two raters when the same two raters rate each of a sample of n subjects independently. A simple example is given in the Table 18.2, where p_{ij} denotes the true proportion of the corresponding evaluations by $Rater$ 1 and $Rater$ 2 (e.g., p_{10} denotes that $Rater$ 1 thinks it's positive but $Rater$ 2 thinks it's negative), $p_{i\cdot} = p_{i1} + p_{i0}$ and $p_{\cdot j} = p_{1j} + p_{0j}$.

Table 18.2: Proportional Distribution by Two Rater

		$Rater\ 2$		
		positive	negative	
$Rater\ 1$	positive	p_{11}	p_{10}	$p_{1\cdot}$
	negative	p_{01}	p_{00}	$p_{0\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 0}$	

Kappa coefficient κ takes the form,

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o ($p_o = p_{11} + p_{00}$) is the proportion of rater pairs exhibiting agreement and p_e ($p_e = p_{1\cdot}p_{\cdot 1} + p_{0\cdot}p_{\cdot 0}$) is the proportion expected to exhibit agreement by chance alone. Thus "perfect agreement" would be indicated by $\kappa = 1$, and no agreement (other than that expected by chance) means that $\kappa = 0$.

18.1.4.1 Methods

- **Hypothesis:** The hypotheses of interest are

  $(\boldsymbol{Two-sided})$

  $$H_0 : \kappa = k_0 \quad versus \quad H_1 : \kappa \neq k_1$$

  $(\boldsymbol{One-sided})$

  $$H_0 : \kappa = k_0 \quad versus \quad H_1 : \kappa > k_1$$

- **Formula:** We can get sample size n from,

  $(\boldsymbol{Two-sided})$

  $$n = [\frac{z_{\alpha/2}\sqrt{Q_0} + z_{\beta}\sqrt{Q_1}}{k_1 - k_0}]^2$$

  $(\boldsymbol{One-sided})$

  $$n = [\frac{z_{\alpha}\sqrt{Q_0} + z_{\beta}\sqrt{Q_1}}{k_1 - k_0}]^2$$

  where Q_0 (Q_1) can be caculated by using k_0 (k_1) with

  $$Q_0(Q_1) = (1-p_e)^{-4}\left\{\sum_i p_{ii}[(1-p_e) - (p_{\cdot i} + p_{i\cdot})(1-p_o)]^2 + (1-p_o)^2 \sum_i \sum_{j \neq i} p_{ij}(p_{\cdot i} + p_{j\cdot})^2 - (p_o p_e - 2p_e + p_o)^2\right\}$$

Note that all of the values needed are uniquely determined by $p_{1\cdot}$, $p_{\cdot 1}$, k_0 and k_1. Specifically,

$$p_{0\cdot} = 1 - p_{1\cdot}$$

$$p_{\cdot 0} = 1 - p_{\cdot 1}$$

$$p_e = p_{1\cdot}p_{\cdot 1} + p_{0\cdot}p_{\cdot 0}$$

$$p_o = \begin{cases} k_0(1-p_e) + p_e & \text{for } Q_0 \\ k_1(1-p_e) + p_e & \text{for } Q_1 \end{cases}$$

$$p_{00} = (p_o - p_{1\cdot} + p_{\cdot 0})/2$$

$$p_{11} = p_o - p_{00}$$

$$p_{10} = p_{1\cdot} - p_{11}$$

$$p_{01} = p_{\cdot 1} - p_{11}$$

18.1.4.2 Input and Output

- **Input:**

  1. $p_{1\cdot}$: proportion that $Rater$ 1 gives positive evaluation
  2. $p_{\cdot 1}$: proportion that $Rater$ 2 gives positive evaluation
  3. k_0: reference value of Kappa coefficient
  4. k_1: expected value of Kappa coefficient
  5. α: type I error rate
  6. β: type II error rate (Power: $1-\beta$)

- **Output:** sample size n

18.1.4.3 An Example (Single-arm Cohen's Kappa Test)

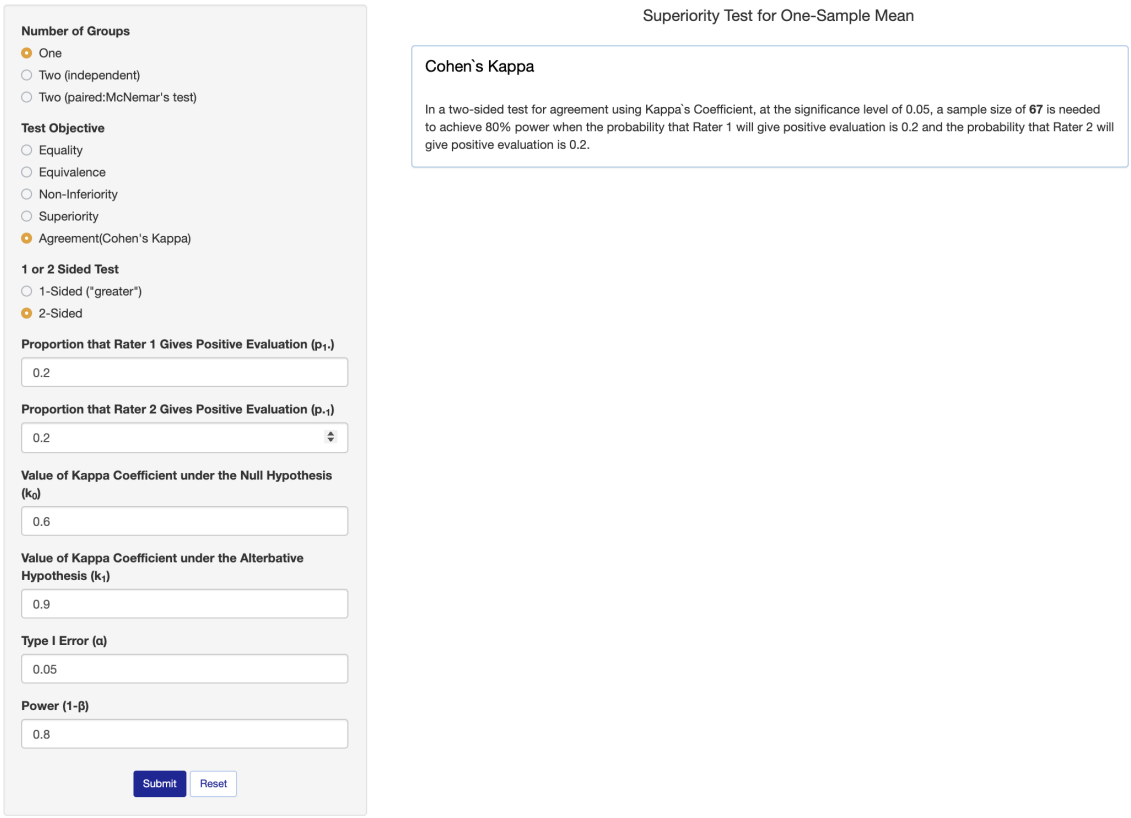
As an example, suppose two evaluation methods are asked to rate a group of cancer patients and to decide whether or not the status of each exhibits positive. We expect each method to identify 20% of patients to be positive ($p_{1\cdot} = p_{\cdot 1} = 0.20$). Let κ denote the level of agreement. The null hypothesis is $H_0 : \kappa = 0.6$, but we expect Kappa coefficient is 0.9. At $\alpha = 0.05$, the required sample size for having an 80% power (i.e., $1 - \beta = 0.8$) can be obtained by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.

- Select **Number of Groups: One**, **Test Objective: Agreement(Cohen's Kappa)** and **1 or 2 Sided Test: 2-Sided**.
- Input $p_1., p_{.1}, k_0 = 0.6, k_1 = 0.9, \alpha$ and $1 - \beta$.
- Click **Submit**.

Then the computed sample size is 67 in this situation, shown in Figure 18.4.

Figure 18.4: An Example (Single-arm Cohen's Kappa Test)



18.2 Two arms (independent)

Let x_{ij} be a binary response from the jth subject in the ith treatment group, $j = 1, \cdots, n_i, i = 1, 2$. For a fixed i, it is assumed that x_{ij}'s are i.i.d. with $P(x_{ij} = 1) = p_i$. In practice, p_i is usually

estimated by the observed proportion in the ith treatment group,

$$\hat{p}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}.$$

Let $\epsilon = p_t - p_c$ be the difference between the true mean response rates of a test drug (p_t) and a control (p_c). It is of interest to test for equality (i.e., $p_t = p_c$), non-inferiority (i.e., $p_t - p_c$ is greater than or equal to a pre-determined non-inferiority margin), superiority (i.e., $p_t - p_c$ is greater than a pre-determined superiority margin), and equivalence (i.e., the absolute difference between p_t and p_c is within a difference of clinical importance). The following are details of sample size calculation with two arms.

18.2.1 Test Objective: Equality

18.2.1.1 Methods

- **Hypothesis:** To test whether there is a difference between the mean response rates of the test drug and the reference drug, the following hypotheses are usually considered,

  $(\boldsymbol{Two-sided})$

  $$H_0 : \epsilon = 0 \quad versus \quad H_1 : \epsilon \neq 0$$

  $(\boldsymbol{One-sided})$

  $$H_0 : \epsilon \leq 0 \quad versus \quad H_1 : \epsilon > 0$$

- **Formula:** We can get sample sizes n_t and n_c from

  $(\boldsymbol{Two-sided})$

  $$n_c = \frac{(z_{\alpha/2} + z_\beta)^2}{\epsilon^2}[\frac{p_t(1-p_t)}{k} + p_c(1-p_c)]$$

  $(\boldsymbol{One-sided})$

  $$n_c = \frac{(z_{\alpha} + z_\beta)^2}{\epsilon^2}[\frac{p_t(1-p_t)}{k} + p_c(1-p_c)]$$

  and $n_t = kn_c$

18.2.1.2 Input and Output

- **Input:**

  1. p_c: true response rate of control treatment
  2. p_t: true response rate of the test drug

3. k ($k = n_t/n_c$): subject ratio of test control versus treatment
4. α: type I error rate
5. β: type II error rate (Power: $1 - \beta$)

- Output: sample sizes n_t and n_c

18.2.1.3 An Example (Two-arms (independent) Equality Two-sided Test)

In this example, suppose that a difference of $\epsilon = 20\%$ in clinical response of cure is considered of clinically meaningful difference between the two agents for a certain disease. Assuming that the true cure rate for control treatment and the test drug are 65% ($p_c = 0.65$ and $p_t = p_c + \epsilon = 0.85$), respectively, at $\alpha = 0.05$, the sample sizes for having an 80% power (i.e., $1 - \beta = 0.8$) with $k = 1$ (equal allocation) can be determined by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: Two**, **Test Objective: Equality** and **1 or 2 Sided Test: 2-Sided**.
- Input $\boldsymbol{p_c, p_t, k, \alpha}$ and $\boldsymbol{1 - \beta}$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 18.5.

18.2.2 Test Objective: Equivalence

18.2.2.1 Methods

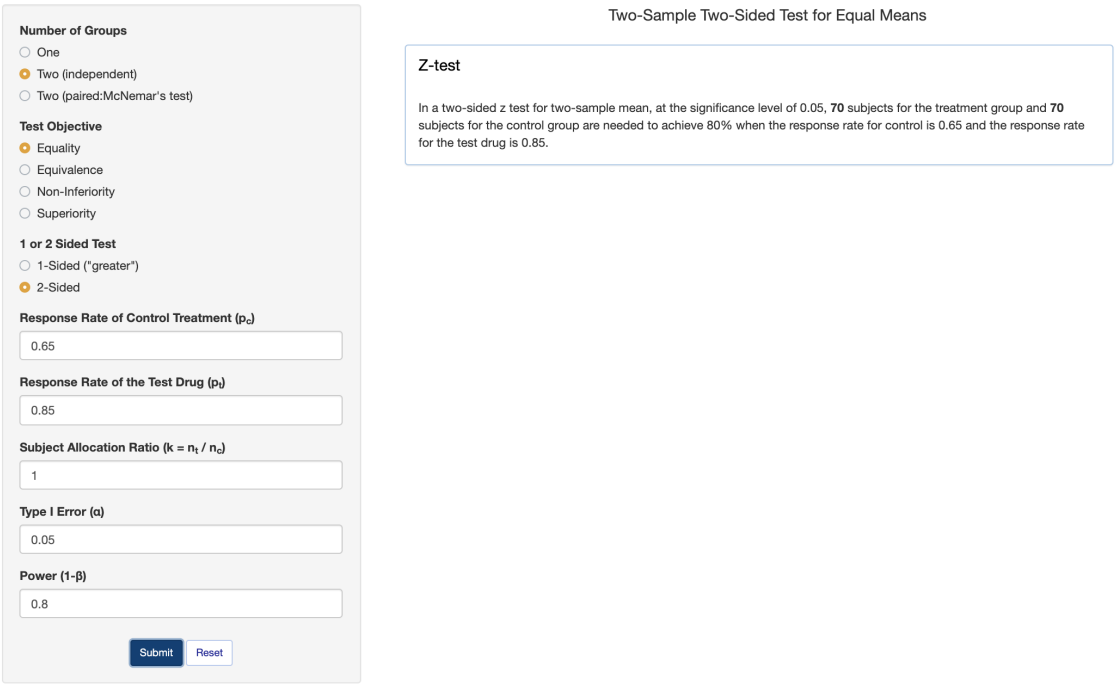
- **Hypothesis:** To establish equivalence, the following hypothesis is usually considered,

$$H_0 : |\epsilon| \geq \delta \quad versus \quad H_1 : |\epsilon| < \delta$$

- **Formula:** We can get sample sizes n_t and n_c from

$$n_c = \frac{(z_\alpha + z_{\beta/2})^2}{(\delta - |\epsilon|)^2}[\frac{p_t(1-p_t)}{k} + p_c(1-p_c)] \quad \text{and} \quad n_t = kn_c.$$

Figure 18.5: An Example (Two-arms (independent) Equality Two-sided Test)



18.2.2.2 Input and Output

- **Input:**
  1. δ ($\delta > 0$): equivalence margin
  2. p_c: true response rate of control treatment
  3. p_t: true response rate of the test drug
  4. k ($k = n_t/n_c$): subject ratio of test control versus treatment
  5. α: type I error rate
  6. β: type II error rate (Power: $1 - \beta$)
- **Output:** n_t and n_c

18.2.2.3 An Example (Two-arms (independent) Equivalence Test)

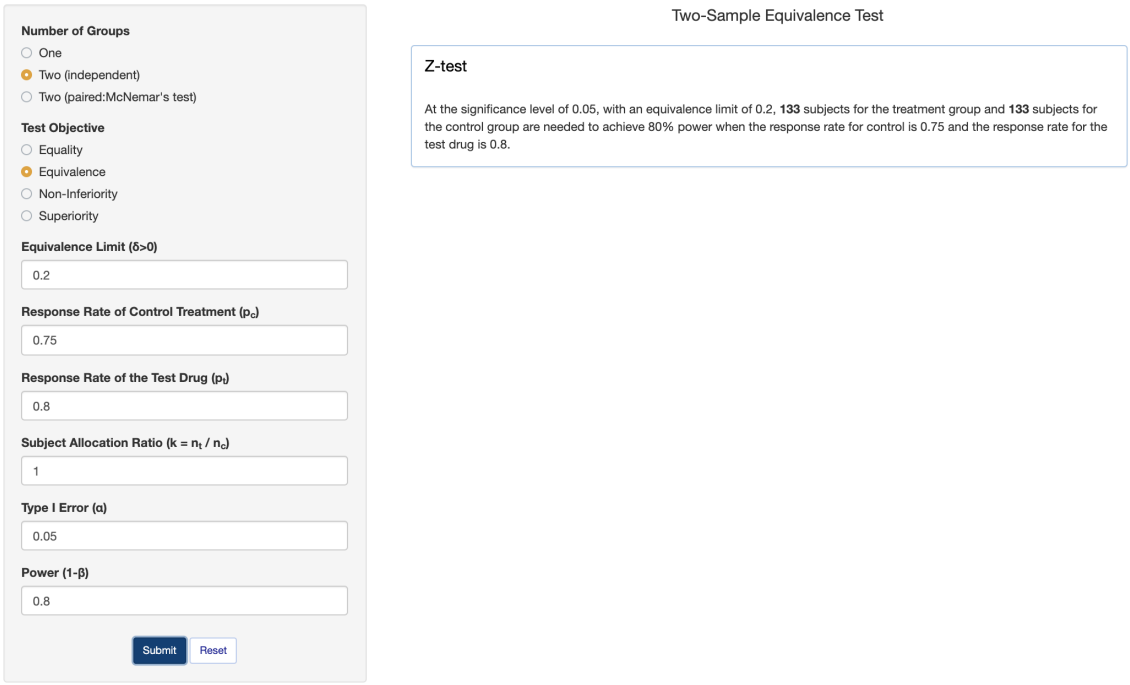
For establishment of equivalence, suppose the true cure rate for the two agents are 75% ($p_c = 0.75$) and 80% ($p_t = 0.80$) and the equivalence limit is 20% (i.e., $\delta = 0.20$). At $\alpha = 0.05$, the sample sizes for having an 80% power (i.e., $1 - \beta = 0.8$) with $k = 1$ (equal allocation) can be determined

by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: Two** and **Test Objective: Equivalence**.
- Input $\delta, p_c, p_t, k, \alpha$ and $1-\beta$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 18.6.

Figure 18.6: An Example (Two-arms (independent) Equivalence Test)



18.2.3 Test Objective: Non-Inferiority/Superiority

18.2.3.1 Methods

- **Hypothesis:** The problem of testing non-inferiority and superiority can be translated into the following hypotheses,
  $(Noninferiority)$

$$H_0 : \epsilon \leq -\delta \quad versus \quad H_1 : \epsilon > -\delta$$

$(\boldsymbol{Superiority})$

$$H_0 : \epsilon \leq \delta \quad versus \quad H_1 : \epsilon > \delta$$

where δ ($\delta > 0$) is the superiority or non-inferiority margin.

- **Formula:** We can get sample sizes n_t and n_c from

$(\boldsymbol{Noninferiority})$

$$n_c = \frac{(z_\alpha + z_\beta)^2}{(\epsilon + \delta)^2}[\frac{p_t(1-p_t)}{k} + p_c(1-p_c)]$$

$(\boldsymbol{Superiority})$

$$n_c = \frac{(z_\alpha + z_\beta)^2}{(\epsilon - \delta)^2}[\frac{p_t(1-p_t)}{k} + p_c(1-p_c)]$$

18.2.3.2 Input and Output

- **Input:**

  1. δ ($\delta > 0$): non-inferiority or superiority margin
  2. p_0: a reference value
  3. p: true response rate of the test drug
  4. α: type I error rate
  5. β: type II error rate (Power: $1 - \beta$)

- **Output:** sample sizes n_t and n_c

18.2.3.3 An Example (Two-arms (independent) Non-Inferiority Test)

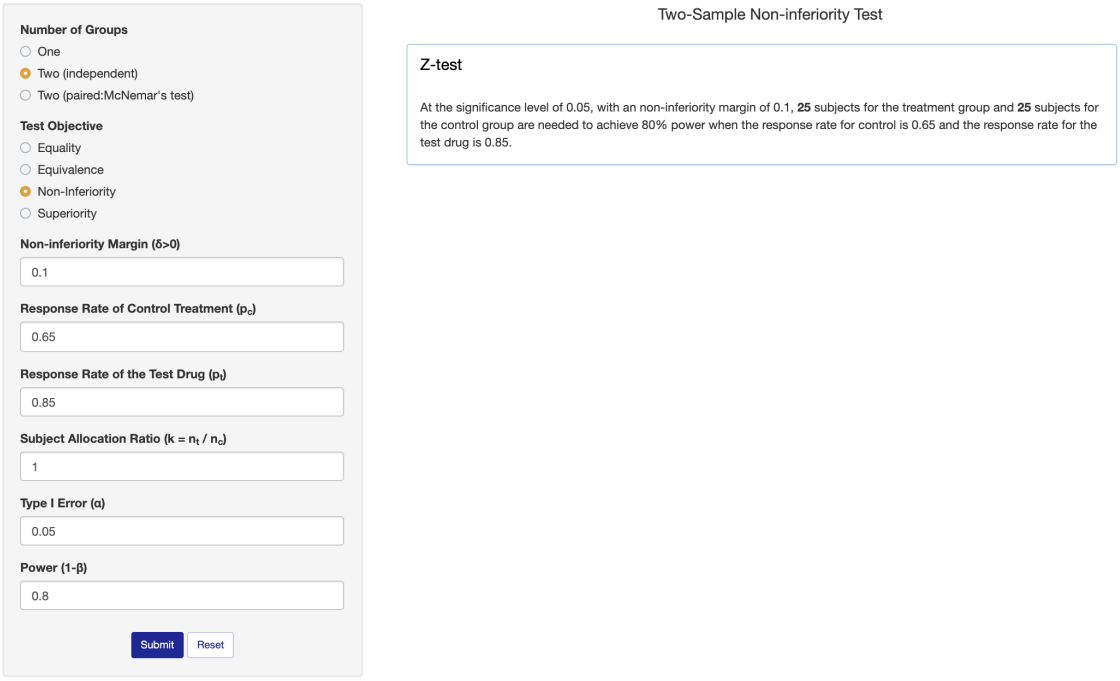
Now, suppose it is of interest to establish non-inferiority of the test drug as compared to the active control agent. Similarly, we consider the difference less than 10% is of no clinical importance. Thus, the non-inferiority margin is chosen to be 10% (i.e., $\delta = 0.10$). Also, suppose the true mean cure rates of the treatment agents and the active control are 85% and 65% (i.e., $p_t = 0.85$ and $p_c = 0.65$), respectively. Then, at $\alpha = 0.05$, the sample size for having an 80% power (i.e., $1 - \beta = 0.8$) with $k = 1$ (equal allocation) can be determined by the following steps,

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: One** and **Test Objective: Non-Inferiority**.
- Input $\boldsymbol{\delta, p_c, p_t, k, \alpha}$ and $\boldsymbol{1 - \beta}$.

- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 18.7.

Figure 18.7: An Example (Two-arms (independent) Non-Inferiority Test)



18.3 Two arms (paired): McNemar's Test

For a given laboratory test, test results are usually summarized as either normal or abnormal. Let x_{ij} denote the binary response from the ith ($i = 1, 2, \ldots, n$) subject in the jth treatment where $j = 1$ denotes pre-treatment and $j = 2$ post-treatment, and $x_{ij} = 1$ denotes that the response is normal and $x_{ij} = 0$ abnormal. The test results can be summarized in Table 18.3, where $n_{ij}, i, j = 1, 0$ are

Table 18.3: Test Results of Two Arms Paired

		Post-treatment	
		normal	abnormal
Pre-treatment	normal	n_{11}	n_{10}
	abnormal	n_{01}	n_{00}

defined by as follows,

$$n_{11} = \sum_{i=1}^{n} x_{i1}x_{i2}$$

$$n_{10} = \sum_{i=1}^{n} x_{i1}(1 - x_{i2})$$

$$n_{01} = \sum_{i=1}^{n} (1 - x_{i1})x_{i2}$$

$$n_{00} = \sum_{i=1}^{n} (1 - x_{i1})(1 - x_{i2})$$

Define,

$$p_{11} = n_{11}/n$$

$$p_{10} = n_{10}/n$$

$$p_{01} = n_{01}/n$$

$$p_{00} = n_{00}/n$$

$$p_{1\cdot} = p_{11} + p_{10}$$

$$p_{\cdot 1} = p_{11} + p_{01}$$

18.3.1 Methods

- **Hypothesis:** It is of interest to test whether there is a categorical shift after treatment. A categorical shift is defined as either a shift from 0 (abnormal) in pre-treatment to 1 (normal) in post-treatment or a shift from 1 (normal) in pre-treatment to 0 (abnormal) in post-treatment. Thus, the hypothesis of interest is

  $(\boldsymbol{Two-sided})$

  $$H_0 : p_{1\cdot} = p_{\cdot 1} \quad versus \quad H_1 : p_{1\cdot} \neq p_{\cdot 1}$$

  $(\boldsymbol{One-sided})$

  $$H_0 : p_{1\cdot} = p_{\cdot 1} \quad versus \quad H_1 : p_{1\cdot} > p_{\cdot 1}$$

  which is equivalent to

  $(\boldsymbol{Two-sided})$

  $$H_0 : p_{10} = p_{01} \quad versus \quad H_1 : p_{10} \neq p_{01}$$

  $(\boldsymbol{One-sided})$

  $$H_0 : p_{10} = p_{01} \quad versus \quad H_1 : p_{10} > p_{01}$$

- **Formula:** We can get sample size n from

  $(\boldsymbol{Two-sided})$

$$n = \frac{[z_{\alpha/2}\sqrt{p_{10}+p_{01}} + z_{\beta}\sqrt{p_{10}+p_{01}-(p_{10}-p_{01})^2}]^2}{(p_{10}-p_{01})^2}$$

  $(\boldsymbol{One-sided})$

$$n = \frac{[z_{\alpha}\sqrt{p_{10}+p_{01}} + z_{\beta}\sqrt{p_{10}+p_{01}-(p_{10}-p_{01})^2}]^2}{(p_{10}-p_{01})^2}$$

18.3.2 Input and Output

- **Input:**

  1. p_{10}: probability of shifting from normal to abnormal
  2. p_{01}: probability of shifting from abnormal to normal
  3. α: type I error rate
  4. β: type II error rate (Power: $1-\beta$)

- **Output:** sample size n

18.3.3 An Example (Two-arms (paired) McNemar's Test)

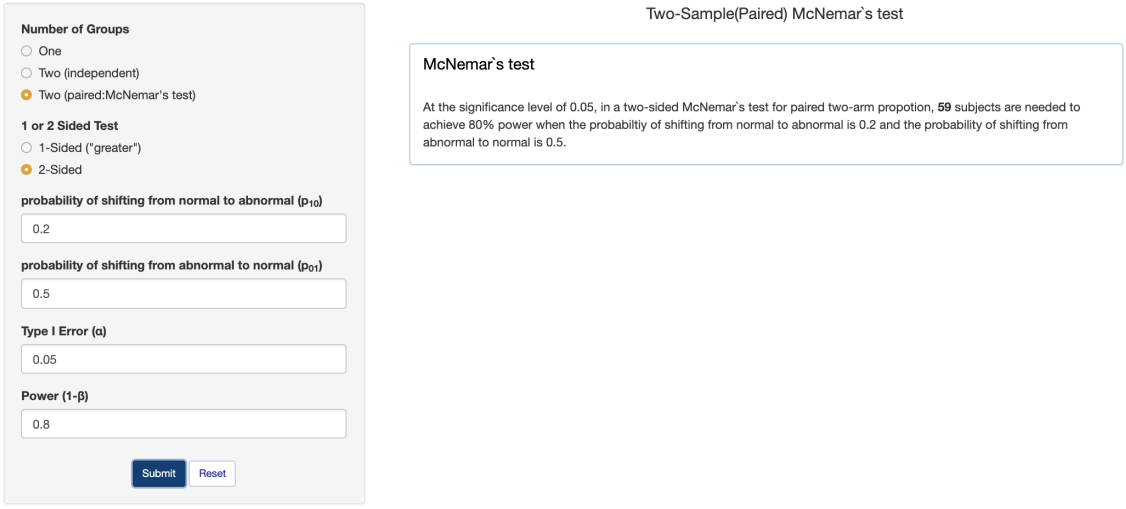
Consider a study, it is expected that about 50% ($p_{10} = 0.50$) of patients will shift from 1 (abnormal pre-treatment) to 0 (normal post-treatment) and 20% ($p_{01} = 0.20$) of patients will shift from 0 (normal pre-treatment) to 1 (abnormal post-treatment).

The investigator would like to select a sample size such that there is an 80% ($1-\beta = 0.80$) power for detecting such a difference if it truly exists at the 5% ($\alpha = 0.05$) level of significance. The required sample size can be obtained as follows:

- Select **SAMPLE SIZE: Binary Outcome**.
- Select **Number of Groups: One** and **Test Objective: Non-Inferiority**.
- Input $\boldsymbol{\delta, p_c, p_t, k, \alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

Then the computed sample size in this situation is shown in Figure 18.8.

Figure 18.8: An Example (Two-arms (paired) McNemar's Test)



19. Sample Size Calculation for Continuous Outcome

In this module, we implement the sample size calculation for continuous endpoint, which include the following functions shown in Table 19.1.

Table 19.1: Function implementation in sample size calculation for binary endpoint.

Number of Arms	Test Objectives	One- or/and Two-sided	Contents	Section
One	Equality	One-sided & Two-sided	Z-test & T-test	Section 19.1.1
	Equivalence	-	Z-test & T-test	Section 19.1.3
	Non-inferiority	-	Z-test & T-test	Section 19.1.2
	Superiority	-	Z-test & T-test	Section 19.1.2
	Correlation	One-sided & Two-sided	Z-test & T-test	Section 19.1.4
Two (independent)	Equality	One-sided & Two-sided	Z-test & T-test	Section 19.2.1
	Equivalence	-	Z-test & T-test	Section 19.2.2
	Non-inferiority	-	Z-test & T-test	Section 19.2.3
	Superiority	-	Z-test & T-test	Section 19.2.3
Two (paired)	Paired	One-sided & Two-sided	T-test	Section 19.3
Multipe	ANOVA	-	F-test	Section 19.4

19.1 Single arm

To compare a new drug to a placebo control, one single-sample study will be conducted. This single sample will consist of observations from a single treatment using the new drug when the mean is

to be compared to a specified constant, the reference response. Let $\epsilon = \mu_t - \mu_c$ be the difference between the expected mean response (μ_t) of the new drug and a reference response value (μ_c) from the control. The main reference for this section is Chow et al. (2017).

19.1.1 Test Objective: Equality

19.1.1.1 Methods

To test whether there is a difference between the mean response of the test drug and the reference value, the following hypotheses and calculation formulas are usually considered,

- **Hypothesis:**
  - $(\boldsymbol{Two-sided})$if there is a difference between μ_t and μ_c,

$$H_0 : \epsilon = 0 \quad versus \quad H_1 : \epsilon \neq 0$$

  - $(\boldsymbol{One-sided})$if there is a positive difference between μ_t and μ_c, that is $\mu_t > \mu_c$,

$$H_0 : \epsilon \leq 0 \quad versus \quad H_1 : \epsilon > 0$$

- **Formula:**
  - for ***T-test***, we search for sample size n that satisfies the following conditions,
    $(\boldsymbol{Two-sided})$

$$T_{n-1}\left\{t_{\alpha/2,n-1}|\frac{\sqrt{n\epsilon^2}}{\sigma}\right\} - T_{n-1}\left\{-t_{\alpha/2,n-1}|\frac{\sqrt{n\epsilon^2}}{\sigma}\right\} = \beta$$

    $(\boldsymbol{One-sided})$

$$T_{n-1}\left\{t_{\alpha,n-1}|\frac{\sqrt{n\epsilon^2}}{\sigma}\right\} = \beta$$

  - for ***Z-test***, we can get sample size n from,
    $(\boldsymbol{Two-sided})$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\epsilon^2}$$

    $(\boldsymbol{One-sided})$

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\epsilon^2}$$

19.1.1.2 Input and Output

- **Input:**
  1. ϵ: difference between the true mean response of a test drug (μ_t) and a reference value (μ_c)
  2. α: type I error rate
  3. β: type II error rate (Power: $1-\beta$)
  4. σ: standard deviation (we assume standard deviation is known when z-test and unknown when t-test) and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_n^{i=1}(x_i - \overline{x})^2$
- **Output:** sample size n

19.1.1.3 An Example (Single-arm Equality Two-sided Test)

Consider an example concerning a study of osteoporosis (or decreased bone mass). Usually, the measure of bone density is SD.

Suppose that the mean bone density before the treatment is 1.5 SD ($\mu_c = 1.5$ SD) and after treatment is expected to be 2.0 SD ($\mu_t = 2$ SD) with the standard deviation ($\sigma = 1$). At $\alpha = 0.05$, the required sample size for having an 80% power ($1-\beta = 0.8$) for correctly detecting a difference of $\epsilon = 0.5$ SD change from pre-treatment to post-treatment can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: One**, **Test Objective: Equality** and **1 or 2 Sided Test: 2-Sided**.
- Input $\boldsymbol{\mu_t, \mu_c, \sigma, \alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

Then the computed sample sizes are 34 using T-test and 32 using Z-test in this situation, shown in Figure 19.1.

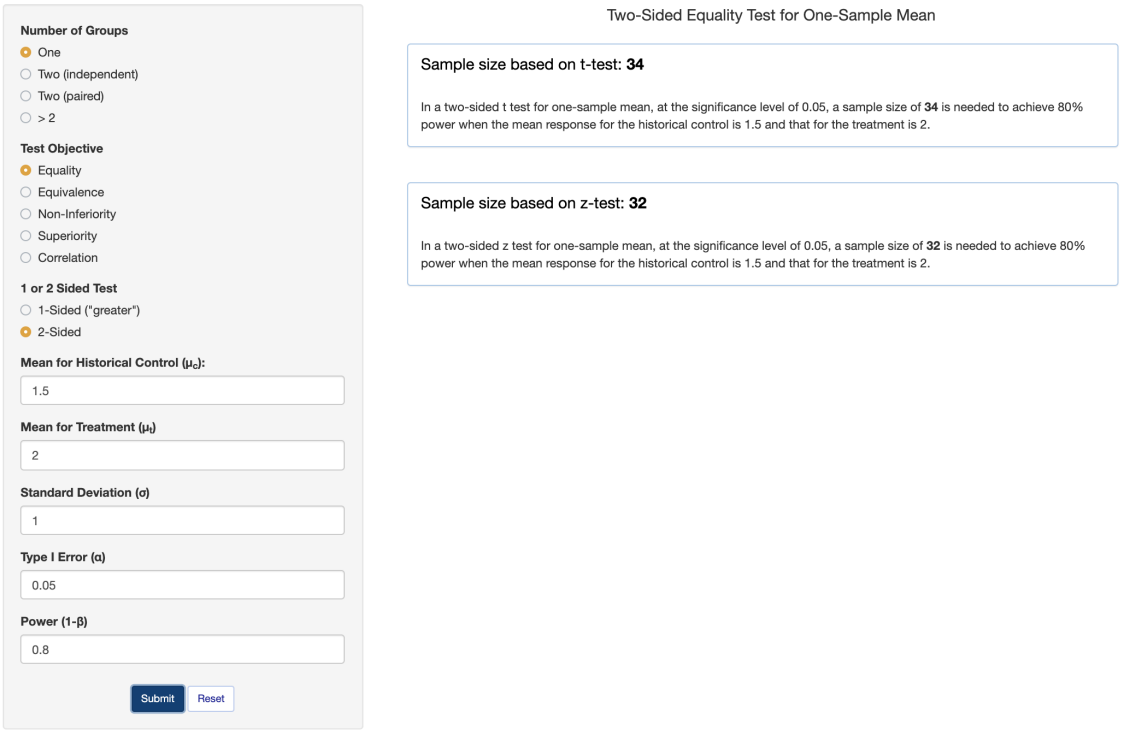
19.1.2 Test Objective: Non-Inferiority/Superiority

19.1.2.1 Methods

The problem of testing non-inferiority and superiority can be explained by the following hypotheses,

- **Hypothesis:**

Figure 19.1: An Example (Single-arm Equality Two-sided Test)



  - $(\boldsymbol{Non-inferiority})$ if the new drug μ_t is not much worse than the placebo control μ_c. In other words, $\epsilon = \mu_t - \mu_c$ is not too small,

$$H_0 : \epsilon \leq -\delta \quad versus \quad H_1 : \epsilon > -\delta$$

  - $(\boldsymbol{Superiority})$ if the new drug μ_t is much better than the placebo control μ_c. In other words, $\epsilon = \mu_t - \mu_c$ is big enough,

$$H_0 : \epsilon \leq \delta \quad versus \quad H_1 : \epsilon > \delta$$

    where δ ($\delta > 0$) is the non-inferiority or superiority margin.

- **Formula:**

  - for ***T-test***, we search for a n that satisfies $(\boldsymbol{Noninferiority})$

$$T_{n-1}\left\{t_{\alpha,n-1}\Big|\frac{\sqrt{n}(\epsilon+\delta)}{\sigma}\right\} = \beta$$

$(\boldsymbol{Superiority})$

$$T_{n-1}\left\{t_{\alpha,n-1}\Big|\frac{\sqrt{n}(\epsilon-\delta)}{\sigma}\right\}=\beta$$

– for ***Z-test***, we can get sample size n from

$(\boldsymbol{Noninferiority})$

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\epsilon+\delta)^2}$$

$(\boldsymbol{Superiority})$

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\epsilon-\delta)^2}$$

19.1.2.2 Input and Output

- **Input:**

  1. δ: superiority or non-inferiority margin
  2. ϵ: difference between the true mean response of a test drug (μ_t) and a reference value (μ_c)
  3. α: type I error rate
  4. β: type II error rate (Power: $1-\beta$)
  5. σ: standard deviation (we assume standard deviation is known when z-test and unknown when t-test) and $\hat{\sigma}^2=\frac{1}{n-1}\sum_n^{i=1}(x_i-\overline{x})^2$

- **Output:** sample size n

19.1.2.3 An Example (Single-arm Non-inferiority Test)

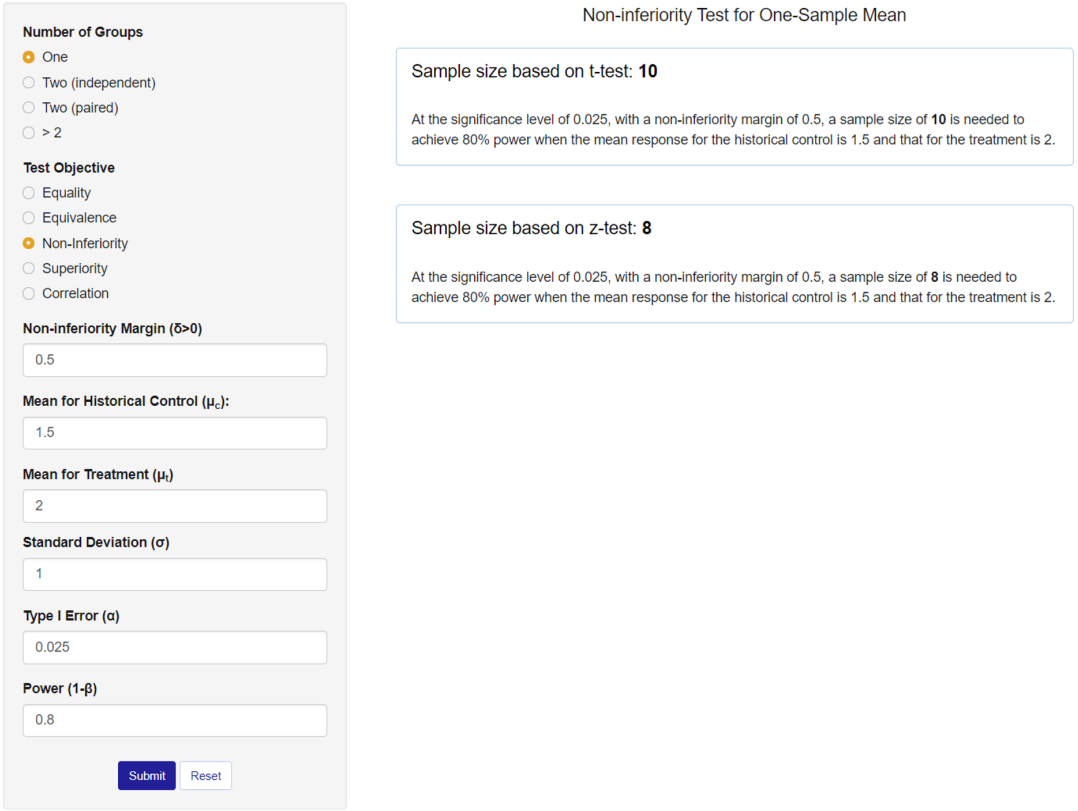
In the study of osteoporosis, we wish to show that the mean bone density post-treatment is no less than pre-treatment by a clinically meaningful difference $\delta=0.5$ SD. We know mean bone density pre-treatment is 1.5 ($\mu_c=1.5$). Suppose the expected mean bone density post-treatment is 2.0 ($\mu_t=2.0$) with standard deviation of 1 ($\sigma=1$). At $\alpha=0.025$, the required sample size for having an 80% power ($1-\beta=0.8$) can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: One** and **Test Objective: Non-inferiority**.
- Input $\boldsymbol{\delta, \mu_c, \mu_t, \sigma, \alpha}$ and $\boldsymbol{1-\beta}$.

- Click **Submit**.

Then the computed sample sizes are 10 using T-test and 8 using Z-test in this situation, shown in Figure 19.2.

Figure 19.2: An Example (Single-arm Non-inferiority Test)



19.1.3 Test Objective: Equivalence

19.1.3.1 Methods

The objective is to test how close the treatment effect of the test drug is to a gold standard on average. The following hypothesis will be considered,

$$H_0 : |\epsilon| \geq \delta \quad versus \quad H_1 : |\epsilon| < \delta.$$

- For ***T-test***, we search for sample size n that satisfies

$$T_{n-1}\left\{t_{\alpha,n-1}|\frac{\sqrt{n}(\delta - |\epsilon|)}{\sigma}\right\} = \frac{\beta}{2}$$

- For ***Z-test***, we can get sample size n from,

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\delta - |\epsilon|)^2}$$

19.1.3.2 Input and Output

- **Input:**
  1. δ: equivalence margin; $\delta > 0$
  2. ϵ: difference between the true mean response of a test drug (μ_t) and a reference value (μ_c)
  3. σ: standard deviation (we assume standard deviation is known when z-test and unknown when t-test) and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_n^{i=1}(x_1 - \overline{x})^2$
  4. α: type I error rate
  5. $1 - \beta$: power (β is type II error rate)
- **Output:** sample size n

19.1.3.3 An Example (Single-arm Equivalence Test)

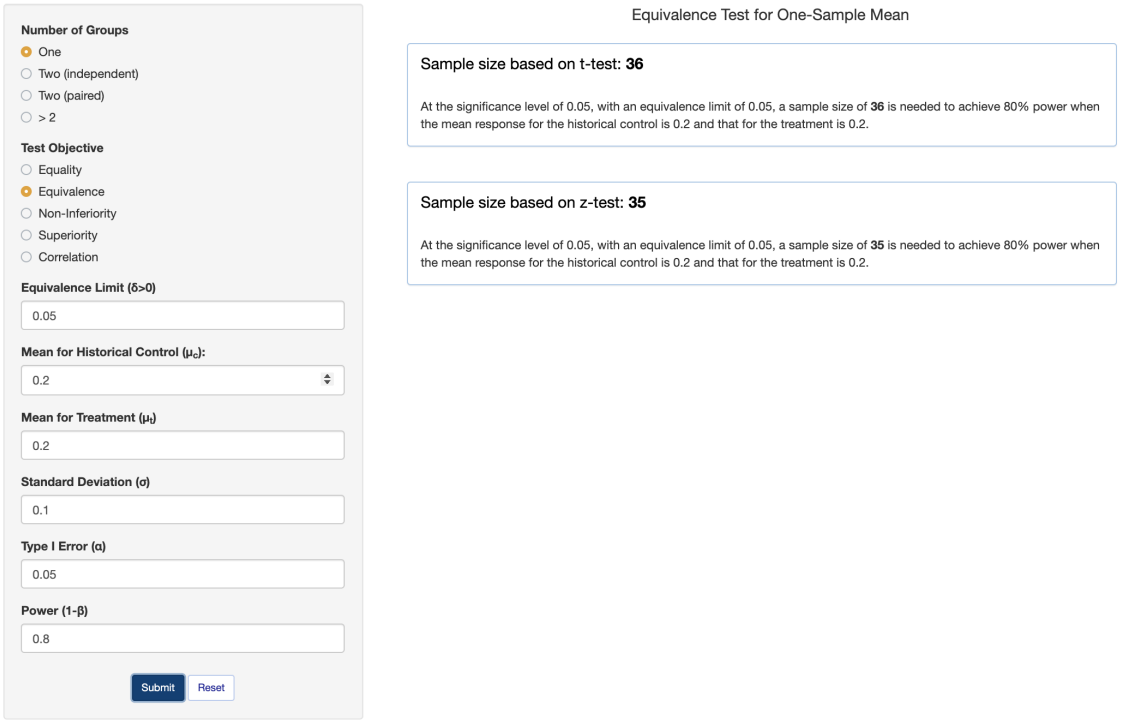
Consider an example concerning the effect of a test drug on body weight change in terms of body mass index (BMI) before and after the treatment.

Suppose clinicians consider that a less than 5% change in BMI from baseline (pre-treatment) to endpoint (post-treatment) is not a safety concern for the indication of the disease under study. Thus, we consider $\delta = 0.05$ as the equivalence margin. The objective is then to demonstrate safety by testing equivalence in mean BMI between pre-treatment and post-treatment of the test drug. Assume the true BMI before and after the treatment are both 0.2 ($\mu_c = \mu_t = 0.2$) and the difference of them is 0 ($\epsilon = 0$) and the standard deviation is 10% ($\sigma = 0.1$), with $\alpha = 0.05$, the sample size required for achieving an 80% power $(1 - \beta = 0.8)$ can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: One** and **Test Objective: Equivalence**.
- Input $\boldsymbol{\delta}, \boldsymbol{\mu_c}, \boldsymbol{\mu_t}, \boldsymbol{\sigma}, \boldsymbol{\alpha}$ and $\mathbf{1 - \beta}$.
- Click **Submit**.

Then the computed sample sizes are 36 using T-test and 35 using Z-test in this situation, shown in Figure 19.3.

Figure 19.3: An Example (Single-arm Equivalence Test)



19.1.4 Test Objective: Correlation

This subsection introduces the single-arm correlation test. The correlation coefficient ρ is calculated as

$$\rho = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}},$$

indicating that the relationship of only two variables is being examined, e.g. the relationship of patient's age (X) and treatment effect of a certain drug (Y). The main reference for this subsection is Zar (2010).

19.1.4.1 Methods

To test whether there is a correlation between two variables, the following hypotheses and calculation formulas are usually considered,

- **Hypothesis:**

- $(\boldsymbol{Two-sided})$ if there is a correlation between the two variables,

$$H_0 : \rho = 0 \quad versus \quad H_1 : \rho = r$$

where $r \neq 0$.

- $(\boldsymbol{One-sided})$ if there is a positive correlation between the two variables,

$$H_0 : \rho = 0 \quad versus \quad H_1 : \rho > r$$

where $r > 0$.

- **Formula:** We can use both t-test and z-test to calculate the sample size for the hypothesis. Both tests use Fishers Transformation, denoted as $C(r) = 0.5log(\frac{1+r}{1-r})$.

  - For ***Z-test***: Given a sample correlation r based on n observations that are from a population with true correlation parameter ρ, $C(r)$ follows a normal distribution with mean $C(\rho)$ and variance $1/\sqrt{n-3}$.

$$C(r) \sim N(C(\rho), 1/\sqrt{n-3})$$

Thus, under H_0, $\sqrt{n-3}C(r) \sim N(0,1)$ since $C(\rho) = 0.5log(1) = 0$. The sample sizes required to achieve the power $1-\beta$ and control type I error rate at α are as follows:

$(\boldsymbol{Two-sided})$

$$n = (\frac{z_{\alpha/2} + z_\beta}{C(r)})^2 + 3$$

$(\boldsymbol{One-sided})$

$$n = (\frac{z_\alpha + z_\beta}{C(r)})^2 + 3$$

  - For ***T-test***: The t-test for significance of r is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

If we find the critical t value, denoted as t_c, above which we will reject H_0, then we can get r_c.

$$r_c = \sqrt{\frac{t_c^2}{t_c^2 + n - 2}}$$

The sample size calculation involves the transformation proposed by Pearson and Hartley (1996):

$$C_r = C(r) + \frac{r}{2(n-1)},$$

$$C_{r_c} = C(r_c).$$

The sample size required can be obtained by solving the following equations iteratively.

$(\boldsymbol{Two-sided})$

$$1-\beta = \Phi\{(C_r - C_{r_c})\sqrt{n-3}\} + \Phi\{(-C_r - C_{r_c})\sqrt{n-3}\}$$

$(\boldsymbol{One-sided})$

$$1-\beta = \Phi\{(C_r - C_{r_c})\sqrt{n-3}\}$$

19.1.4.2 Input and Output

- **Input:**
  1. r: correlation coefficient under alternative hypothesis, or expected correlation coefficient
  2. α: type I error rate
  3. β: type II error rate (Power: $1-\beta$)
- **Output:** sample size n

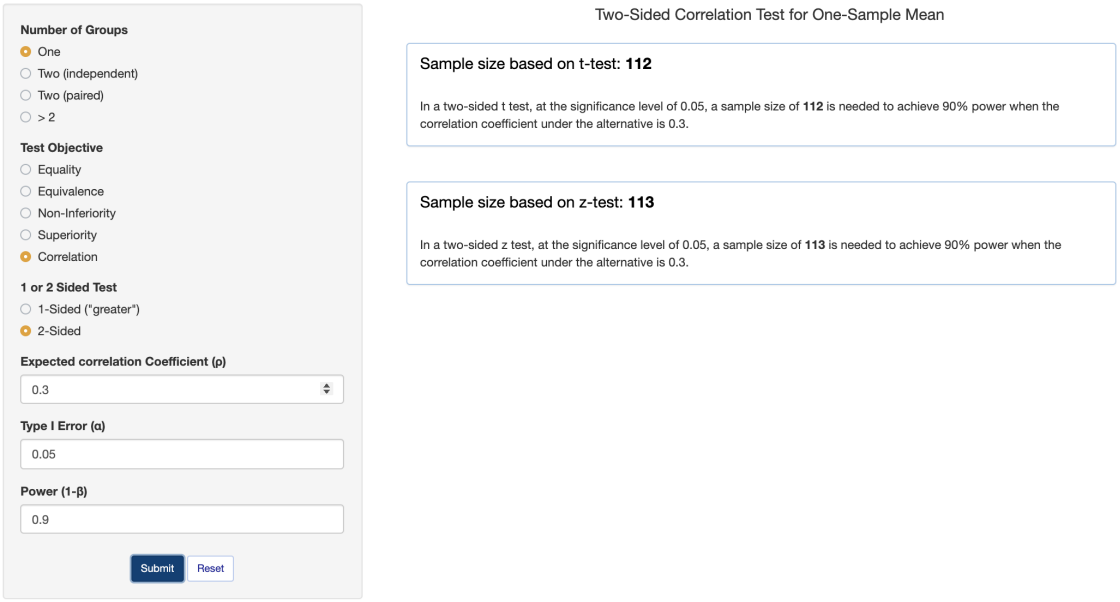
19.1.4.3 An Example (Single-arm Correlation Test)

Consider a situation where we want to test whether the treatment effect of a certain new drug is associated with the patient age. It's hoped that the correlation coefficient between the treatment effect of this new drug and the patient age is 0.3 ($r = 0.3$).And we want the design with type I error rate of 0.05 ($\alpha = 0.05$) and power of 90% ($1-\beta = 0.9$). The sample sizes can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: One**, **Test Objective: Correlation** and **1 or 2 Sided Test: 2-Sided**.
- Input $\boldsymbol{r}$, $\boldsymbol{\alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

This will calculate the sample sizes for this design and the output is shown in the right panel. The computed sample sizes are 112 using T-test and 113 using Z-test in this situation, shown in Figure 19.4.

Figure 19.4: An Example (Single-arm Correlation Test)



19.2 Two arms (independent)

To compare a new drug to a standard treatment, one two-samples study will be conducted. These two samples will consist of observations from the treatment using this new drug and this standard treatment. Let $\epsilon = \mu_t - \mu_c$ be the difference between the expected mean response of this new drug (μ_t) and this standard treatment (μ_c). In practice, it may be desirable to have an unequal treatment allocation, i.e., $n_c/n_t = k$ for some k, where n_t and n_c denote sample sizes for treatment and control respectively. Note that $k = 1/2$ indicates a 2 to 1 test-control allocation, whereas $k = 2$ indicates a 1 to 2 test-control allocation.

19.2.1 Test Objective: Equality

19.2.1.1 Methods

To test whether there is a difference between the mean response of the test drug and the reference value, the following hypotheses and calculation formulas are usually considered,

- **Hypothesis:**

– $(\boldsymbol{Two-sided})$ if there is a difference between μ_t and μ_c,

$$H_0 : \epsilon = 0 \quad versus \quad H_1 : \epsilon \neq 0$$

– $(\boldsymbol{One-sided})$ if there is a positive difference between μ_t and μ_c, that is $\mu_t > \mu_c$, or $\epsilon > 0$,

$$H_0 : \epsilon \leq 0 \quad versus \quad H_1 : \epsilon > 0$$

- **Formula:**

– for ***T-test***, we search for n_t that satisfies

$(\boldsymbol{Two-sided})$

$$T_{(1+k)n_t-2}\left\{t_{\alpha/2,(1+k)n_t-2}|\frac{\sqrt{n_t\epsilon^2}}{\sigma\sqrt{1+1/k}}\right\} - T_{(1+k)n_t-2}\left\{-t_{\alpha/2,(1+k)n_t-2}|\frac{\sqrt{n_t\epsilon^2}}{\sigma\sqrt{1+1/k}}\right\} = \beta$$

$(\boldsymbol{One-sided})$

$$T_{(1+k)n_t-2}\left\{t_{\alpha,(1+k)n_t-2}|\frac{\sqrt{n_t\epsilon^2}}{\sigma\sqrt{1+1/k}}\right\} = \beta$$

and $n_c = kn_t$.

– for ***Z-test***, we can get sample sizes n_t and n_c from,

$(\boldsymbol{Two-sided})$

$$n_t = \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2(1+1/k)}{\epsilon^2}$$

$(\boldsymbol{One-sided})$

$$n_t = \frac{(z_{\alpha} + z_\beta)^2\sigma^2(1+1/k)}{\epsilon^2}$$

and $n_c = kn_t$.

19.2.1.2 Input and Output

- **Input:**

1. $\epsilon = \mu_t - \mu_c$: the expected mean difference between a test drug (μ_t) and a standard treatment (μ_c)
2. $k = n_c/n_t$: treatment allocation ratio
3. α: type I error rate
4. β: type II error rate (Power: $1 - \beta$)

5. σ: variance. Assume that variance is known when z-test and unknown t-test, we often use the pooled variance to estimate it.

$$\hat{\sigma}^2 = \frac{1}{n_c + n_t - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})$$

- **Output:**

1. n_t: sample size of treatment group
2. n_c: sample size of control group

19.2.1.3 An Example (Two-arms (Independent) Equality Two-sided Test)

Consider a pharmaceutical company that is interested in conducting a clinical trial to compare two cholesterol lowering agents through a parallel design. The primary efficacy parameter is the low density lipoprotein (LDL), because most of the cholesterol is bound to LDLs. In what follows, we will consider the situation where the intended trial is for testing equality of mean responses in LDL.

In this example, suppose a difference of 5% ($\epsilon = \mu_t - \mu_c = 0.05$) in percent change of LDL is considered of clinically meaningful difference. Assuming that the standard deviation is 10% ($\sigma = 10\%$), with $\alpha = 0.05$, the sample sizes required for achieving an 80% power $(1 - \beta = 0.8)$ can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: Two (independent)**, **Test Objective: Equality** and **1 or 2 Sided Test: 2-Sided**.
- Input $\boldsymbol{\epsilon = \mu_t - \mu_c, \sigma, k, \alpha}$ and $\boldsymbol{1 - \beta}$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 19.5.

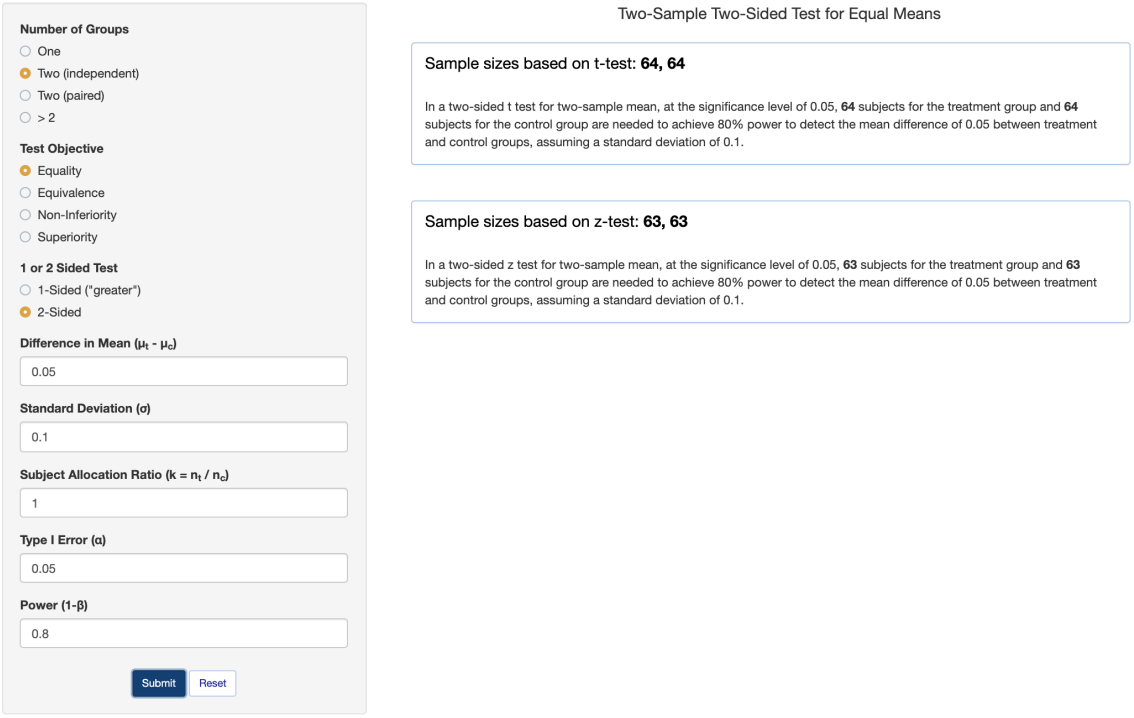
19.2.2 Test Objective: Equivalence

19.2.2.1 Methods

The objective is to test how close the treatment effect of the test drug and the standard treatment are. The following hypothesis will be considered,

$$H_0 : |\epsilon| \geq \delta \quad versus \quad H_1 : |\epsilon| < \delta$$

Figure 19.5: An Example (Two-arms (Independent) Equality Two-sided Test)



- For ***T-test***, we search for n_t that satisfies

$$T_{(1+k)n_t-2}\left\{t_{\alpha,(1+k)n_t-2}|\frac{\sqrt{n_t}(\delta-|\epsilon|)}{\sigma\sqrt{1+1/k}}\right\}=\frac{\beta}{2}$$

- For ***Z-test***, we can get sample sizes n_t and n_c from

$$n_t=\frac{(z_\alpha+z_{\beta/2})^2\sigma^2(1+1/k)}{(\delta-|\epsilon|)^2}\quad\text{and}\quad n_c=kn_t$$

19.2.2.2 Input and Output

- **Input:**

  1. δ: equivalence margin
  2. $\epsilon=\mu_t-\mu_c$: the true mean difference between a test drug (μ_t) and a standard treatment (μ_c)
  3. $k=n_c/n_t$: treatment allocation ratio

4. σ: variance (we assume variance is known when z-test and unknown t-test)
5. α: type I error rate
6. β: type II error rate (Power: $1-\beta$)

- **Output:**

1. n_t: sample size of treatment group
2. n_c: sample size of control group

19.2.2.3 An Example (Two-arms (Independent) Equivalence Test)

Consider a pharmaceutical company that is interested in conducting a clinical trial to compare two cholesterol lowering agents through a parallel design. The primary efficacy parameter is the low density lipoprotein (LDL), because most of the cholesterol is bound to LDLs. In what follows, we will consider the situation where the intended trial is testing for therapeutic equivalence.

For establishment of equivalence, suppose the true mean difference is 1% ($\epsilon = 0.01$) and the equivalence limit is 5% ($\delta = 0.05$). Assuming that the standard deviation is 10% ($\sigma = 10\%$), with $\alpha = 0.05$, the sample sizes required for achieving an 90% power $(1-\beta = 0.9)$ can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: Two (independent)** and **Test Objective: Equivalence**.
- Input $\boldsymbol{\delta, \epsilon = \mu_t - \mu_c, \sigma, k, \alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 19.6.

19.2.3 Test Objective: Non-Inferiority/Superiority

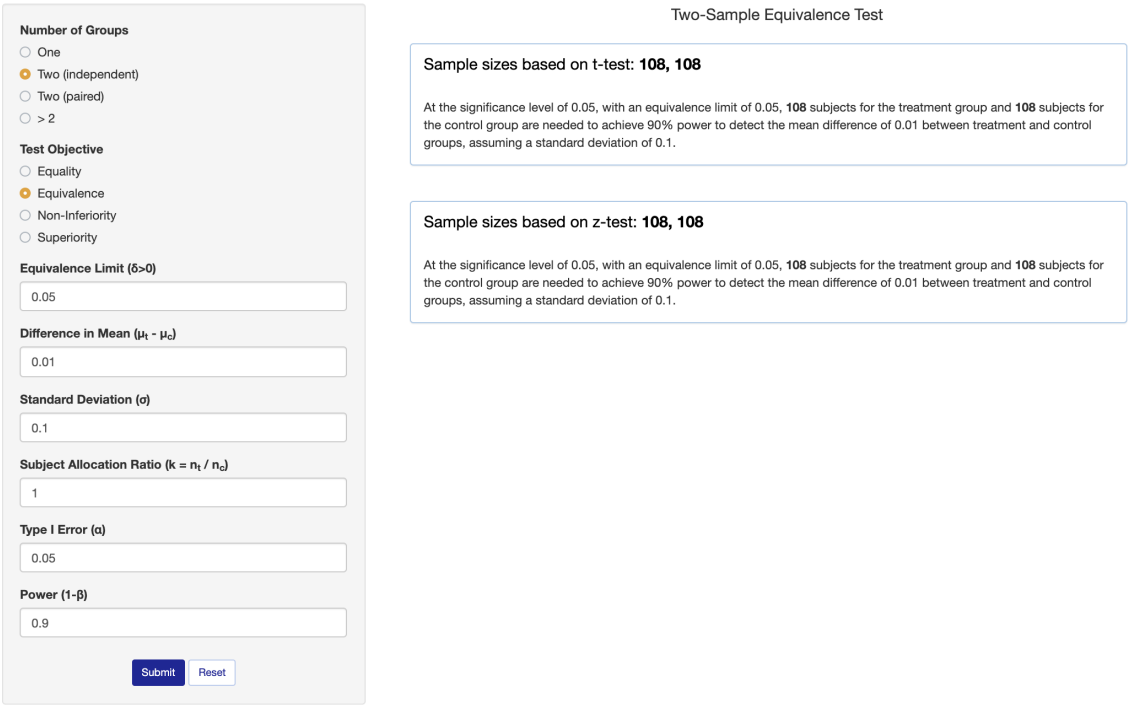
19.2.3.1 Methods

The problem of testing non-inferiority and superiority can be explained by the following hypotheses,

- **Hypothesis:**
  - $(\boldsymbol{Non-inferiority})$ The objective is to confirm that the new drug μ_t is not much worse than the standard treatment μ_c. In other words, $\epsilon = \mu_t - \mu_c$ is not too small,

$$H_0 : \epsilon \leq -\delta \quad versus \quad H_1 : \epsilon > -\delta$$

Figure 19.6: An Example (Two-arms (Independent) Equivalence Test)



  - $(\boldsymbol{Superiority})$ The objective is to confirm that the new drug μ_t is much better than the standard treatment μ_c. In other words, $\epsilon = \mu_t - \mu_c$ is big enough,

$$H_0 : \epsilon \leq \delta \quad versus \quad H_1 : \epsilon > \delta$$

    where δ ($\delta > 0$) is the superiority or non-inferiority margin.

- **Formula:**

  - For ***T-test***, we search for n_t that satisfies

    $(\boldsymbol{Non-inferiority})$

$$T_{(1+k)n_t-2}\left\{t_{\alpha,(1+k)n_t-2}|\frac{\sqrt{n_t}(\epsilon+\delta)}{\sigma\sqrt{1+1/k}}\right\} = \beta$$

    $(\boldsymbol{Superiority})$

$$T_{(1+k)n_t-2}\left\{t_{\alpha,(1+k)n_t-2}|\frac{\sqrt{n_t}(\epsilon-\delta)}{\sigma\sqrt{1+1/k}}\right\} = \beta$$

    and $n_c = kn_t$.

– For ***Z-test***, we can get sample sizes n_t and n_c from

$(\boldsymbol{Non-inferiority})$

$$n_t = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1 + 1/k)}{(\epsilon + \delta)^2}$$

$(\boldsymbol{Superiority})$

$$n_t = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1 + 1/k)}{(\epsilon - \delta)^2}$$

and $n_c = k n_t$.

19.2.3.2 Input and Output

- **Input:**

  1. δ: superiority or non-inferiority margin
  2. $\epsilon = \mu_t - \mu_c$: the true mean difference between a test drug (μ_t) and a standard treatment (μ_c)
  3. $k = n_c/n_t$: treatment allocation ratio
  4. σ: variance (we assume variance is known when z-test and unknown t-test)
  5. α: type I error rate
  6. β: type II error rate (Power: $1 - \beta$)

- **Output:**

  1. n_t: sample size of treatment group
  2. n_c: sample size of control group

19.2.3.3 An Example (Two-arms (Independent) Non-inferiority Test)

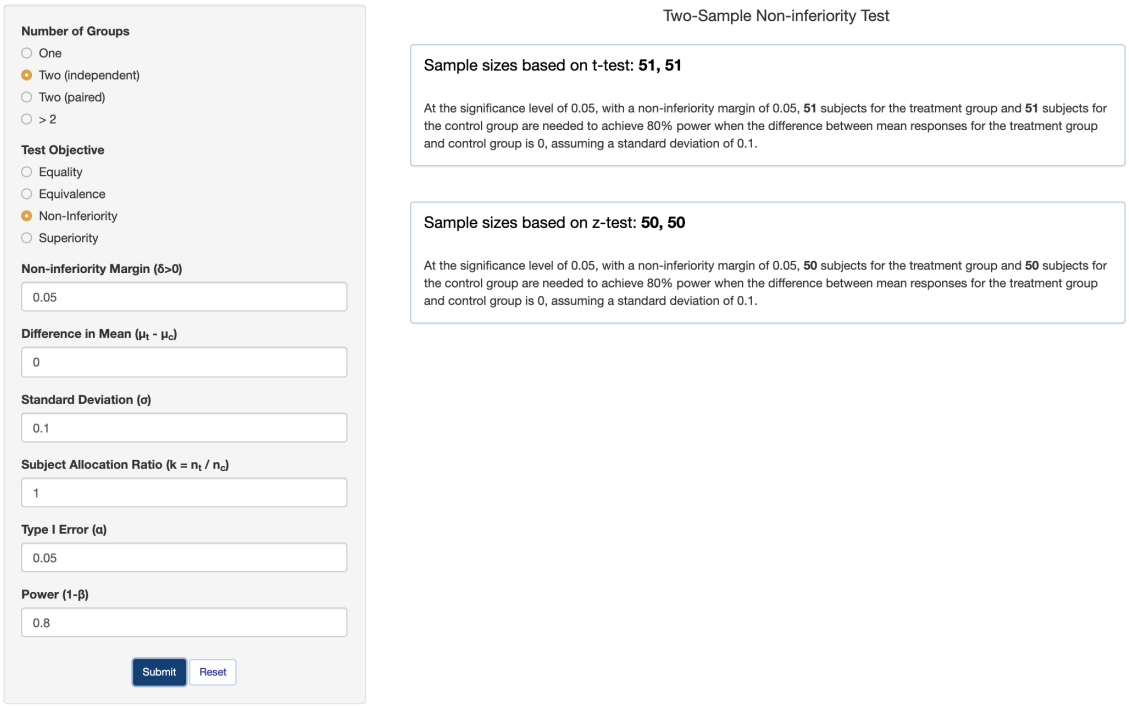
Suppose that the pharmaceutical company is interested in establishing non-inferiority of the test drugas compared to the active control agent. Similarly, we assume that the non-inferiority margin is chosen to be 5% ($\delta = 0.05$). Also, suppose the true difference in mean LDL between treatment groups is 0% ($\epsilon = \mu_t - \mu_C = 0$). Assuming that the standard deviation is 10% ($\sigma = 10\%$), with $\alpha = 0.05$, the sample sizes required for achieving an 80% power $(1 - \beta = 0.8)$ can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: Two (independent)** and **Test Objective: Non-inferiority**.

- Input $\delta, \epsilon = \mu_t - \mu_c, \sigma, k, \alpha$ and $1 - \beta$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 19.7.

Figure 19.7: An Example (Two-arms (Independent) Non-inferiority Test)



19.3 Two arms (paired)

19.3.1 Methods

Let $\epsilon_d = \mu_1 - \mu_2$ be the difference between the true mean response of two paired groups (μ_1 and μ_2). Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) an indication of improvement (worsening) of the test drug as compared to the reference value.

- **Hypothesis:** The hypothesis of interest is
  $(\boldsymbol{Two-sided})$

$$H_0 : \epsilon_d = 0 \quad versus \quad H_1 : \epsilon_d \neq 0$$

$(\boldsymbol{One-sided})$

$$H_0 : \epsilon_d \leq 0 \quad versus \quad H_1 : \epsilon_d > 0$$

- **Formula:** Denote $\triangle_d = \epsilon_d/\sigma_d$ be the effect size. And we use the ***T-test*** here to calculate n that satisfies

  $(\boldsymbol{Two-sided})$

  $$T_{n-1}\left\{t_{\alpha/2,n-1}\middle|\sqrt{n}\triangle_d\right\} - T_{n-1}\left\{-t_{\alpha/2,n-1}\middle|\sqrt{n}\triangle_d\right\} = \beta$$

  $(\boldsymbol{One-sided})$

  $$T_{n-1}\left\{t_{\alpha,n-1}\middle|\sqrt{n}\triangle_d\right\} = \beta$$

19.3.2 Input and Output

- **Input:**
  - if we "Enter the effect size directly",
    1. $\triangle_d$: the effect size, could be calculated by $\triangle_d = (\mu_1 - \mu_2)/\sigma_d$, where μ_1 and μ_2 are mean response of two groups, and σ_d is the standard deviation of pre-post difference
    2. α: type I error rate
    3. β: type II error rate (Power: $1 - \beta$)
  - if "Calculate the effect size" is needed,
    1. μ_1: mean response of group 1
    2. μ_2: mean response of group 2
    3. σ_d: standard deviation of pre-post difference
    4. α: type I error rate
    5. β: type II error rate (Power: $1 - \beta$)
- **Output:** n

19.3.3 An Example (Two-arms (paired) Equality Test)

Consider a standard two-period paired design for the trial whose objective is to establish therapeutic equality between a test drug and a standard therapy. The sponsor is interested in having an 80% ($1 - \beta = 0.80$) power for establishing therapeutic equality. Based on the results from previous studies, it is estimated that the variance is 20% ($\sigma_d = 0.20$). Suppose mean response of group 2

is 1.3 and mean response of group 1 is 1.2. That is, the true mean difference is 10% ($\mu_2(test) - \mu_1(reference) = 0.10$) and effect size $\Delta = 0.50$. The sample sizes can be obtained by the following steps,

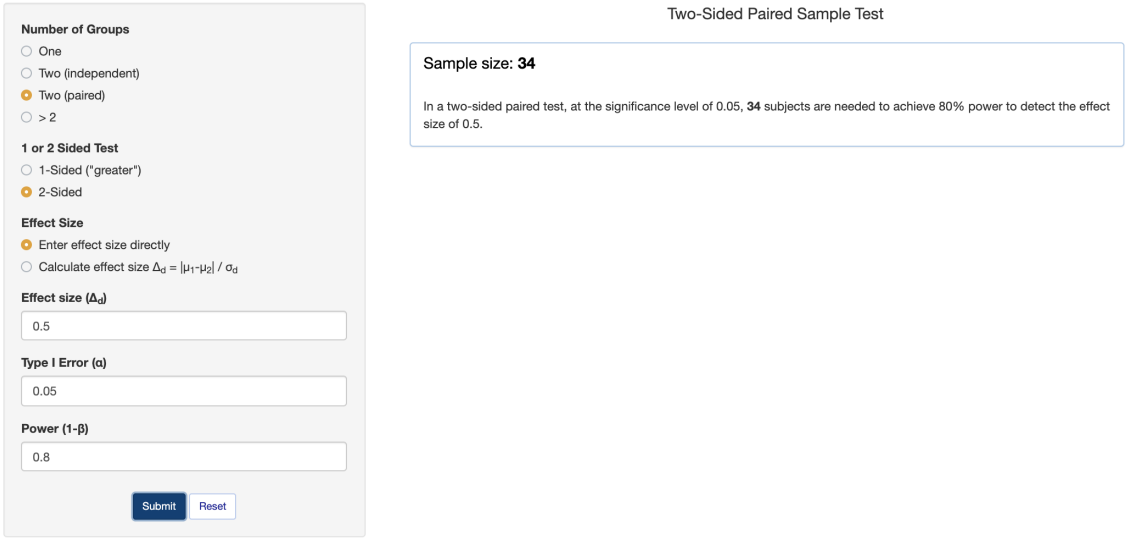
- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups: Two (paired)**, **1 or 2 Sided Test: 2-Sided** and **Effect Size: Enter effect size directly**.
- Input $\boldsymbol{\Delta_d}, \boldsymbol{\alpha}$ and $\mathbf{1} - \boldsymbol{\beta}$.
- Click **Submit**.

or,

- Select **Number of Groups: Two (paired)**, **1 or 2 Sided Test: 2-Sided** and **Effect Size: Calculate effect size** $\Delta_d = |\mu_1 - \mu_2|/\sigma_d$.
- Input $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\sigma_d}, \boldsymbol{\alpha}$ and $\mathbf{1} - \boldsymbol{\beta}$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 19.8.

Figure 19.8: An Example (Two-arms (paired) Equality Test)



19.4 Multiple arms

19.4.1 Methods

Let x_{ij} be the j-th subject from the $i-$th treatment group, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Consider the following one-way analysis of variance (ANOVA) model:

$$x_{ij} = \mu_i + \epsilon_{ij},$$

where μ_i is the fixed effect of the ith treatment and ϵ_{ij} is a random error in observing x_{ij}. It is assumed that ϵ_{ij} are i.i.d. normal random variables with mean 0 and variance σ^2. Let

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \mu_i)^2$$

$$SSA = \sum_{i=1}^{m}(\mu_i - \overline{\mu})^2,$$

where

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad \text{and} \quad \overline{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mu_i$$

Then σ^2 can be estimate by

$$\hat{\sigma}^2 = \frac{SSE}{m(n-1)}$$

- **Hypothesis:** The hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \qquad \text{versus} \qquad H_1 : \mu_i \neq \mu_j \quad (1 \leq i \leq j \leq m).$$

- **Formula:** Under the null hypothesis H_0, $F_A = \frac{nSSA/(m-1)}{SSE/[m(n-1)]}$ follows F-distribution. So H_0 is rejected at the α level of significance if

$$F_A = \frac{nSSA/(m-1)}{SSE/[m(n-1)]} > F_{\alpha,m-1,m(n-1)}$$

where $F_{\alpha,m-1,m(n-1)}$ is the α upper quantile of the F-distribution with $m-1$ and $m(n-1)$ degrees of freedom.

Under the alternative hypothesis H_1, the power of this test is given by

$$P(F_A > F_{\alpha,m-1,m(n-1)})$$

Hence, the sample size needed to achieve power $1-\beta$ can be obtained by $P(F_A > F_{\alpha,m-1,m(n-1)}) = 1-\beta$.

19.4.2 Input and Output

- **Input:**
  - If we "Enter Effect Size Directly",
    1. m: number of groups
    2. f: effect size
    $$f = \frac{\sigma_m}{\sigma} = \sqrt{\frac{\sigma_m^2}{\sigma^2}}$$
    where $SSA/(m-1)$ is approximately σ_m^2 and $SSE/m(n-1)$ is approximately σ^2.
    3. α: type I error rate
    4. β: type II error rate (Power: $1-\beta$)
  - If "Calculate Effect Size" is needed,
    1. m: number of groups
    2. μ_i: mean of group i ($1 \leq i \leq m$)
    3. σ: common standard deviation
    4. α: type I error rate
    5. β: type II error rate (Power: $1-\beta$)
- **Output:** n for per group

19.4.3 An Example (Multiple-arms One-Way ANOVA Test)

Suppose that we are interested in conducting a four-arm ($m = 4$) parallel group, double-blind, randomized clinical trial to compare four treatments. The comparison will be made with a significance level of $\alpha = 0.05$. Assume that the standard deviation within each group is $\sigma = 3.5$ and that the true mean responses for the four treatment groups are given by,

$$\mu_1 = 8.25, \quad \mu_2 = 9.75, \quad \mu_3 = 9.00 \quad \text{and} \quad \mu_3 = 10.00.$$

Then, $f = 0.391$. The sample sizes required for achieving an 80% power ($1-\beta = 0.8$) can be obtained by the following steps,

- Select **SAMPLE SIZE: Continuous Outcome**.
- Select **Number of Groups:** > 2 and **How to Determine Effect Size (f): Enter effect size directly**.

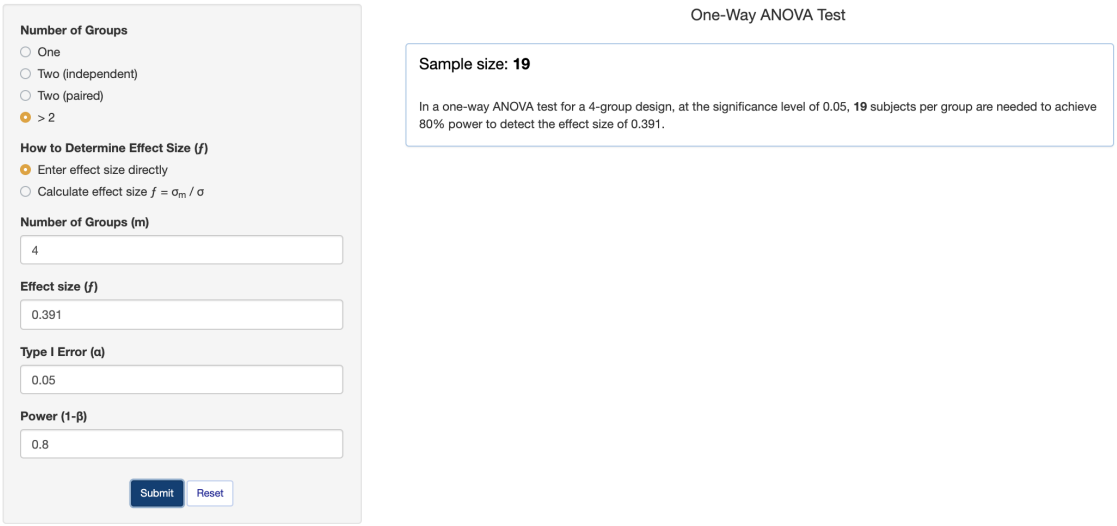
- Input m, f, α and $1-\beta$.
- Click **Submit**.

or,

- Select **Number of Groups:** >2 and **How to Determine Effect Size** (f)**: Calculate effect size** $f=\sigma_m/\sigma$.
- Input m, $\mu_i(i=1,\cdots,4)$, α and $1-\beta$.
- Click **Submit**.

Then the computed sample sizes in this situation are shown in Figure 19.9.

Figure 19.9: An Example (Multiple-arms One-Way ANOVA Test)



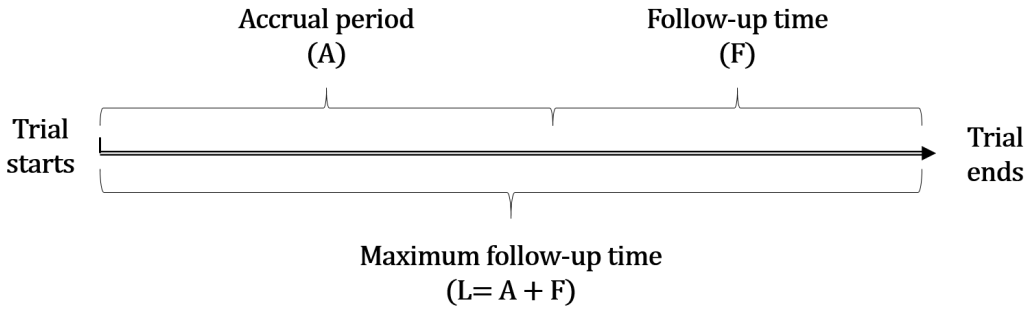
20. Sample Size Calculation for Time-to-Event Outcome

In this section, we implement the sample size calculation for time-to-event endpoint, which include the following functions shown in Table 20.1.

Table 20.1: Function implementation in sample size calculation for binary endpoint.

Number of Arms	Test Objectives	One- or/and Two-sided	Contents	Section
One	Equality	One-sided & Two-sided		Section 20.1
Two	Equality	One-sided & Two-sided	Logrank test	Section 20.2

Before the text, there is three important symbols A, F and L for time-to-event endpoint introduced as follows:



20.1 Single arm

In a study with a single arm, we assume for planning purposes that the survival times follow an exponential distribution with hazard $h(t;\lambda) = \lambda$ and survival distribution $S(t;\lambda) = e^{-\lambda t}$. After the trial is completed, we obtain a series of independent survival times $t_1, t_2, \cdots, t_n$ and indicators $\delta_1, \delta_2, \cdots, \delta_n$, with $\delta_i = 1$ for event occuring, $\delta_i = 0$ otherwise, where n is the total number of subjects in the trial. According to Moore (2016), $\hat{\lambda} = d/V$, where

$$d = \sum_{i=1}^{n} \delta_i \text{ and } V = \sum_{i=1}^{n} t_i.$$

20.1.1 Methods

- **Hypothesis:** The hypothesis of interest is

  $(\boldsymbol{Two-sided})$

  $$H_0 : \lambda_t = \lambda_c \qquad \text{versus} \qquad H_1 : \lambda_t \neq \lambda_c,$$

  $(\boldsymbol{One-sided})$

  $$H_0 : \lambda_t = \lambda_c \qquad \text{versus} \qquad H_1 : \lambda_t < \lambda_c,$$

  where λ_t and λ_c are the hazard rates for the current treatment and historical reference, respectively. The hypothesis is equivalent to

  $(\boldsymbol{Two-sided})$

  $$H_0 : m_t = m_c \qquad \text{versus} \qquad H_1 : m_t \neq m_c,$$

  $(\boldsymbol{One-sided})$

  $$H_0 : m_t = m_c \qquad \text{versus} \qquad H_1 : m_t > m_c,$$

  where m_t and m_c are median survival time for the current treatment and historical reference, respectively, or

  $(\boldsymbol{Two-sided})$

  $$H_0 : HR = 1 \qquad \text{versus} \qquad H_1 : HR \neq 1,$$

  $(\boldsymbol{One-sided})$

  $$H_0 : HR = 1 \qquad \text{versus} \qquad H_1 : HR < 1,$$

  where $HR = \lambda_t/\lambda_c = m_c/m_t$ is the hazard ratio for the current treatment and historical reference.

- **Formula:** We can get sample size n from

  $(\boldsymbol{Two-sided})$

  $$n_d = \frac{(z_{\alpha/2} + z_\beta)^2}{\Delta^2}$$

  $(\boldsymbol{One-sided})$

  $$n_d = \frac{(z_{\alpha} + z_\beta)^2}{\Delta^2}$$

  where $\Delta = log(\lambda_t/\lambda_c)$ and z_α is the upper αth quantile of the standard normal distribution, and

  $$n = \frac{n_d}{P(\delta = 1)},$$

  where n_d is the number of event required, n the total simple size required and the proportion of event occuring

  $$P(\delta = 1) = 1 - \frac{1}{A\lambda_t}(e^{-\lambda_t F} - e^{-\lambda_t(A+F)}).$$

20.1.2 Input and Output

- **Input:**

  1. m_c ($m_c = \frac{log(2)}{\lambda_c}$): median survival time for historical control
  2. m_t ($m_t = \frac{log(2)}{\lambda_t}$): median survival time for treatment, or HR ($HR = m_c/m_t$): hazard ratio
  3. A: length of accrual period
  4. L ($L = A + F$): maximum follow-up time
  5. α: type I error rate
  6. β: type II error rate (Power: $1 - \beta$)

- **Output:** number of event required n_d and total simple size n

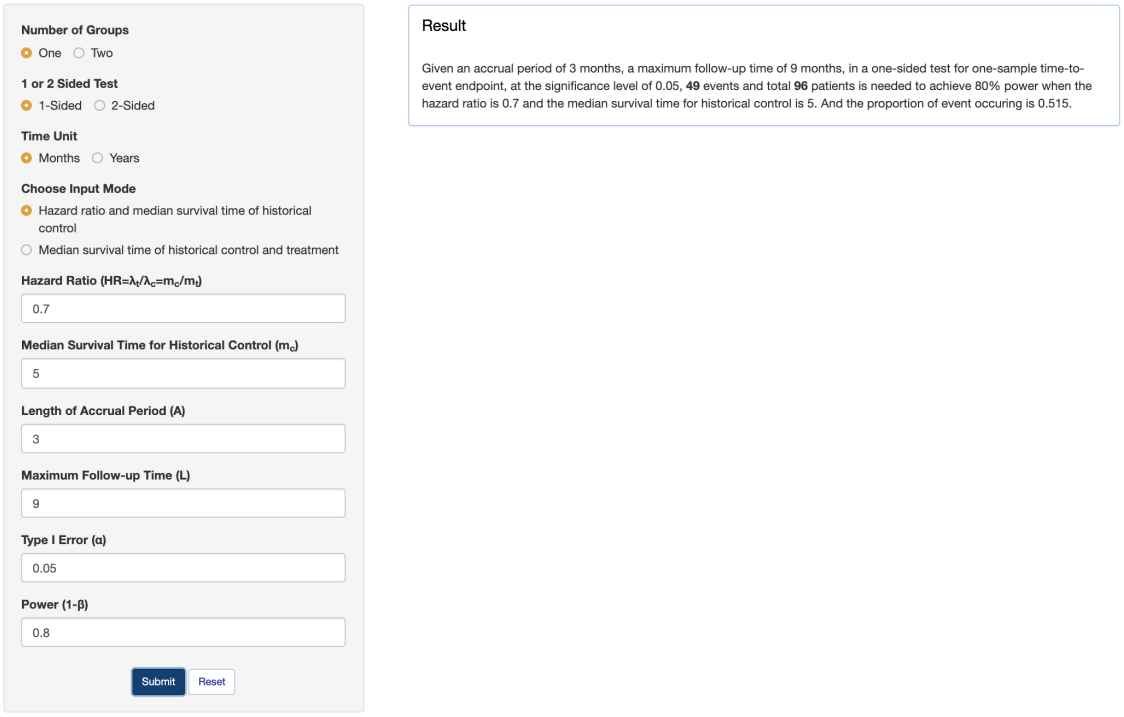
20.1.3 An Example (Single-arm One-sided Test)

Consider a example where we plan a single sample clinical trial with a 5% ($\alpha = 0.05$) significance level (one-sided) test, and we need 80% ($1 - \beta = 0.8$) power to detect a hazard ratio of 0.7 ($HR = 0.7$). Suppose that the null hypothesis rate is $m_c = 7$ months, and the alternative hypothesis hazard rate is $m_t = m_c/HR = 10$ months. We suppose now that the accrual period is $A = 3$ months and that the follow-up period is an additional $F = 6$ months (i.e., maximum follow-up time $L = 9$ months). To obtain an estimate of the number of patients, we follow these steps,

- Select **SAMPLE SIZE: Time To Event**.
- Select **Number of Groups: One**, **1 or 2 Sided Test: 1-Sided**, **Time Unit: Months** and **Choose Input Mode: Hazard ratio and median survival time of historical control**.
- Input $\boldsymbol{HR}, \boldsymbol{m_c}, \boldsymbol{A}, \boldsymbol{L}, \boldsymbol{\alpha}$ and $\boldsymbol{1-\beta}$.
- Click **Submit**.

Then the computed sample size in this situation is shown in Figure 20.1.

Figure 20.1: An Example (Single-arm One-sided Test)



20.2 Two arms

20.2.1 Methods

- **Hypothesis:** The hypothesis of interest is $(\boldsymbol{Two-sided})$

$$H_0 : \lambda_t = \lambda_c \qquad \text{versus} \qquad H_1 : \lambda_t \neq \lambda_c,$$

$(\boldsymbol{One-sided})$

$$H_0 : \lambda_t = \lambda_c \qquad \text{versus} \qquad H_1 : \lambda_t < \lambda_c,$$

where λ_t and λ_c are the hazard rates for the current treatment and historical reference, respectively. The hypothesis is equivalent to

$(\boldsymbol{Two-sided})$

$$H_0 : m_t = m_c \qquad \text{versus} \qquad H_1 : m_t \neq m_c,$$

$(\boldsymbol{One-sided})$

$$H_0 : m_t = m_c \qquad \text{versus} \qquad H_1 : m_t > m_c,$$

where m_t and m_c are median survival time for the current treatment and historical reference, respectively, or

$(\boldsymbol{Two-sided})$

$$H_0 : HR = 1 \qquad \text{versus} \qquad H_1 : HR \neq 1,$$

$(\boldsymbol{One-sided})$

$$H_0 : HR = 1 \qquad \text{versus} \qquad H_1 : HR < 1,$$

where $HR = \lambda_t/\lambda_c = m_c/m_t$ is the hazard ratio for the current treatment and historical reference.

- **Formula:** We can get sample sizes n_t and n_c from

$(\boldsymbol{Two-sided})$

$$n_d = \frac{[(1+k)(z_{\alpha/2} + z_\beta)]^2}{k\Delta^2}$$

$(\boldsymbol{One-sided})$

$$n_d = \frac{[(1+k)(z_{\alpha} + z_\beta)]^2}{k\Delta^2}$$

where

1. $\Delta = log(\lambda_t/\lambda_c)$.
2. $k = n_t/n_c$
3. z_α is the upper αth quantile of the standard normal distribution.

and

$$n = \frac{n_d}{P(\delta = 1)}, \quad n_c = \frac{n}{1+k} \quad \text{and} \quad n_t = \frac{kn}{1+k},$$

where n_d is the number of event required,n the total simple size required and $P(\delta = 1)$ is the combined probability of event occuring. According to Schoenfeld (1983), we have

$$P(\delta = 1) = \frac{P(\delta_c = 1)}{1+k} + \frac{kP(\delta_t = 1)}{1+k} = \frac{P(\delta_c = 1) + kP(\delta_t = 1)}{1+k},$$

where $P(\delta_c = 1)$ and $P(\delta_t = 1)$ are probabilities of event occuring for control and treatment, respectively, and are calculated as:

$$P(\delta_c = 1) = 1 - \frac{1}{A\lambda_c}(e^{-\lambda_c F} - e^{-\lambda_c(A+F)}),$$

$$P(\delta_t = 1) = 1 - \frac{1}{A\lambda_t}(e^{-\lambda_t F} - e^{-\lambda_t(A+F)}).$$

20.2.2 Input and Output

- **Input:**

  1. m_c ($m_c = \frac{log(2)}{\lambda_c}$): median survival time for historical control
  2. m_t ($m_t = \frac{log(2)}{\lambda_t}$): median survival time for treatment, or HR ($HR = m_c/m_t$): hazard ratio
  3. k ($k = n_t/n_c$): subject ratio of test control versus treatment
  4. A: length of accrual period
  5. L ($L = A + F$): maximum follow-up time
  6. α: type I error rate
  7. β: type II error rate (Power: $1 - \beta$)

- **Output:** number of event required n_d, total sample size n, sample size for control arm n_c and for test treatment n_t

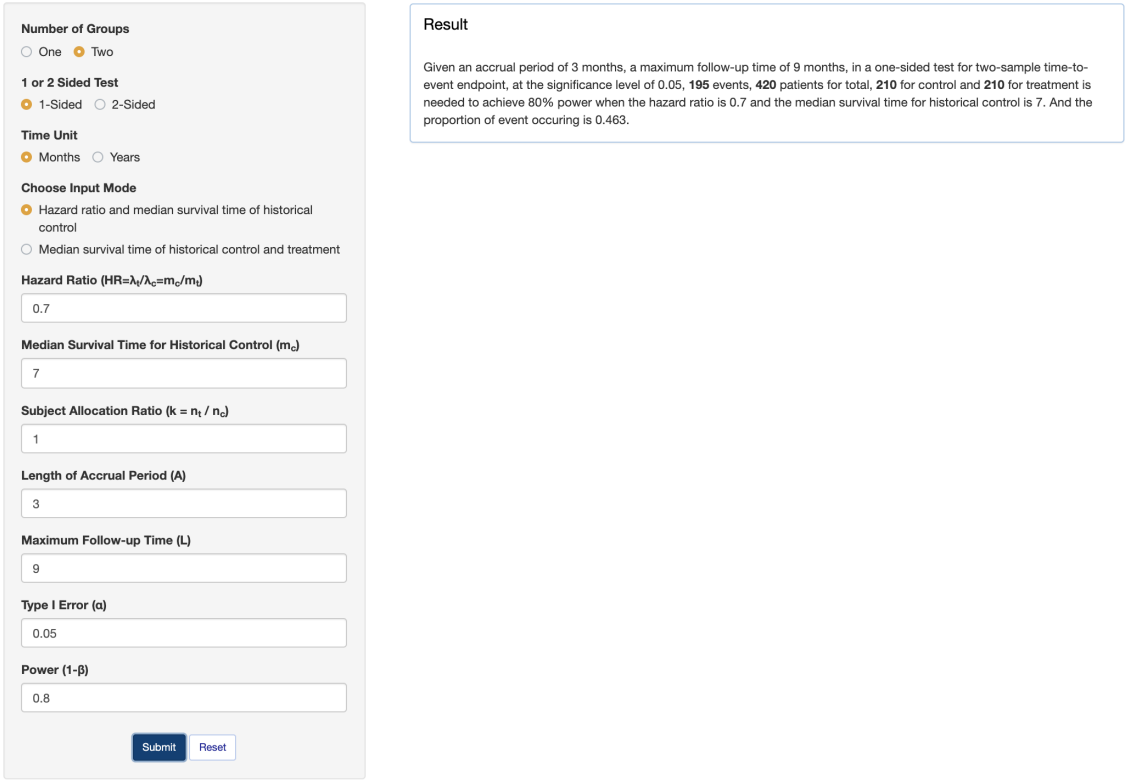
20.2.3 An Example (Two-arms One-sided Test)

Consider a example where we plan a single sample clinical trial with a 5% ($\alpha = 0.05$) significance level (one-sided) test, and we need 80% ($1 - \beta = 0.8$) power to detect a hazard ratio of 0.7 ($HR = 0.7$). Suppose that the null hypothesis rate is $m_c = 7$ months, and the alternative hypothesis hazard rate is $m_t = m_c/HR = 10$ months. We suppose now that the accrual period is $A = 3$ months and that the follow-up period is an additional $F = 6$ months (i.e., maximum follow-up time $L = 9$ months). To obtain an estimate of the number of patients with $k = 1$ (equal allocation), we follow these steps,

- Select **SAMPLE SIZE: Time To Event**.
- Select **Number of Groups: Two**, **1 or 2 Sided Test: 1-Sided**, **Time Unit: Months** and **Choose Input Mode: Hazard ratio and median survival time of historical control**.
- Input HR, m_c, k, A, L, α and $1-\beta$.
- Click **Submit**.

Then the computed sample size in this situation is shown in Figure 20.2.

Figure 20.2: An Example (Two-arms One-sided Test)



21. Simon's Two-Stage Design

This section introduces the sample size calculation for Phase Ib/II clinical trial using Simon's two-stage design (Simon, 1989).

21.1 Method

The Simon's two-stage design is a one-sided one-sample design in which the treatment is tested against a historical control in its response rate. The hypothesis of interest in this design is

$$H_0 : p \leq p_0 \quad versus \quad H_1 : p \geq p_1$$

where p_0 is uninteresting response rate, which is often the historical response rate, and p_1 is expected response rate.

The design consists of two stages. In the first stage, n_1 patients will be recruited and treated and number of responses in the first stage (x_1) is assumed that $x_1 \sim Bin(n_1, p)$. If there are r_1 or fewer responses among these n_1 patients, i.e., $x_1 \leq r_1$, the study will be early terminated and accept the null hypothesis. Otherwise, additional n_2 patients will be enrolled in the second stage and number of responses in the second stage (x_2) is assumed that $x_2 \sim Bin(n_2, p)$, resulting in a total number sample size of $n = n_1 + n_2$. If there are less than or exactly r responses among these n patients,i.e., $x = x_1 + x_2 \leq r$, we also accept the null hypothesis and claim that the treatment is not promising. The process of the design is shown in Figure 21.1.

2. Enumeration

For specified values of p_0, p_1, and type I/II error rates, α and β, we enumerate all of designs with

$$n \in [1, n_{max}],\ n_1 \in [1, n-1],\ r_1 \in [0, n_1] \text{ and } r \in [r_1, n].$$

We can get the expected sample size $EN = n_1 + (1 - PET)n_2$, where PET represents the probability of early termination after the first stage and depends on the true probability of response p (assumed as p_0):

$$PET = B(r_1; p_0, n_1) = \sum_{i=0}^{r_1} \binom{n_1}{i} p_0^i (1 - p_0)^{n_1 - i},$$

where $B(*)$ denotes the cumulative binomial distribution. Then determine that

- Optimal Two-stage Design : satisfies the error probability constraints and minimizes the expected sample size (EN) when the response probability is p_0.
- Minimax Two-stage Design : satisfies the error probability constraints and minimizes the total sample size (n).

3. Start of Enumeration

   The search over n could be ranged from a lower value of about

$$\overline{p}(1 - \overline{p})\left[\frac{z_\alpha + z_\beta}{p_1 - p_0}\right]^2,$$

where $\overline{p} = (p_0 + p_1)/2$ and z_α is the upper αth quantile of the standard normal distribution, to ensure that there are a nontrivial ($n_1, n_2 > 0$) two-stage design.

21.2 Program Input and Output

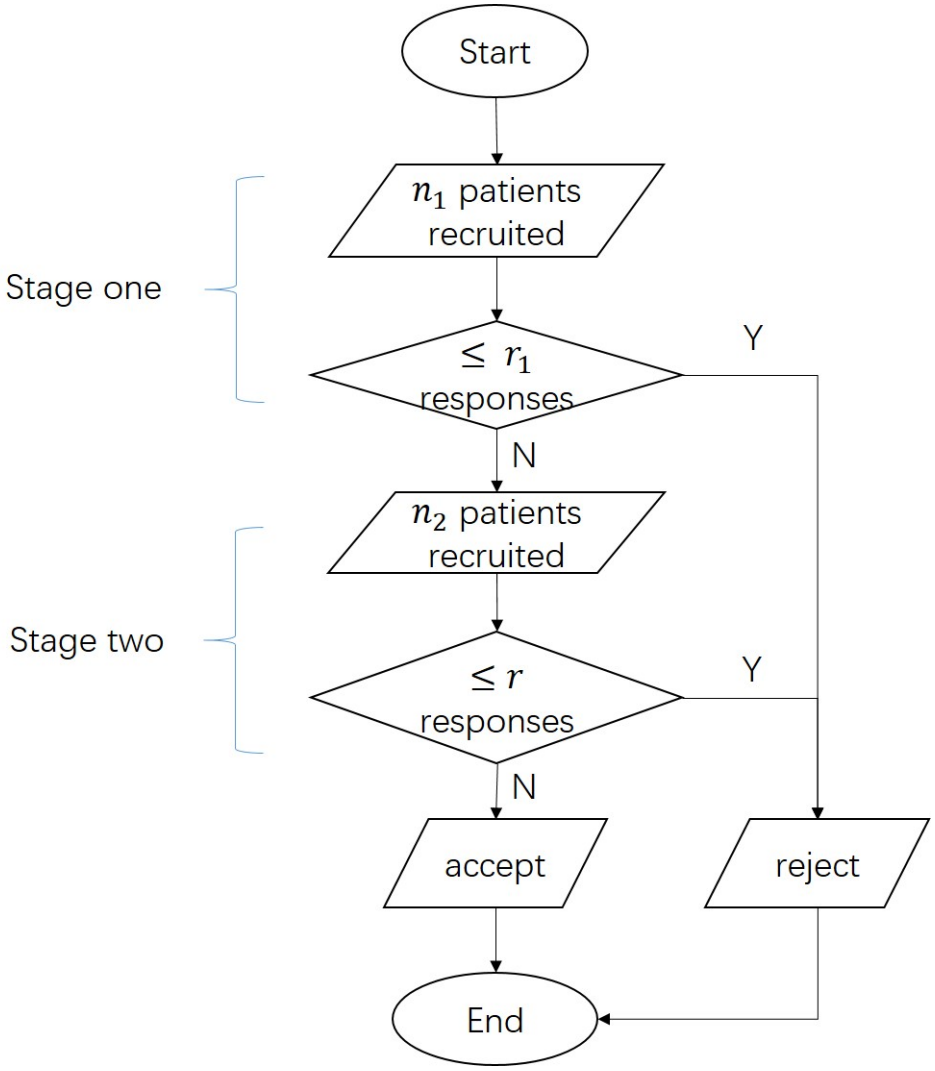
1. Input: p_0, p_1, α, β, n_{max}.
   - p_0 : uninteresting response rate or the historical response rate of the control
   - p_1 : desirable target response rate
   - α : type I error rate
   - β : type II error rate (Power: $1 - \beta$)
   - n_{max} : maximum sample size allowed when searching n
2. Output: r_1, n_1, r, n, EN and PET for the Optimal and Minimax designs.
   - r_1: the first stage threshold to stop the trial for futility, i.e., if there are r_1 or less responses, the trial will be early terminated.
   - n_1: the number of patients studied in the first stage.

- r:
- n: the total sample size.
- PET: the probability of early termination after the first stage under the null when the response probability is p_0.
- EN: the expected sample size, $EN = n_1 + (1 - PET)(n - n_1)$, under the null when the response probability is p_0.

21.3 Protocol Template

A Simons two-stage Optimal(/Minimax) design will be used to allow early stopping if the response is not sufficiently promising to warrant further development (i.e. $< p_0$). This design tests a null hypothesis that the true response rate is less than p_0 against a specific one-sided alternative hypothesis that the true response is at least p_1. The type I error rate is α (one-sided) and the type II error rate is β. Under these assumptions, a total of n patients are planned for enrollment. Based on the above design considerations, n_1 patients will be enrolled to the first stage. If $\leq r_1$ patient in the cohort achieves a response, then enrollment will be early terminated. If at least $r_1 + 1$ patients achieve a response among the first n_1 patients, then an additional $n - n_1$ patients will be enrolled to the second stage. The null hypothesis will be rejected if at least $r + 1$ responses are observed among the n patients.

Figure 21.1: Flow Chart of Simon's Two Stage.



Reference

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Ala, U., Karreth, F. A., Bosia, C., Pagnani, A., Taulli, R., Léopold, V., Tay, Y., Provero, P., Zecchina, R., and Pandolfi, P. P. (2013). Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proceedings of the National Academy of Sciences*, 110(18):7154–7159.

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17(10):1103–1120.

Berger, J. O., Wang, X., and Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics*, 24(1):110–129.

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials*, 10(5):720–734.

Bischoff, W. and Miller, F. (2009). A seamless phase II/III design with sample-size re-estimation. *Journal of biopharmaceutical statistics*, 19(4):595–609.

Braun, T. M. and Wang, S. (2010). A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics*, 66(3):805–812.

Bril, G., Dykstra, R., Pillers, C., and Robertson, T. (1984). Algorithm as 206: Isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):352–357.

Cai, C., Ji, Y., and Yuan, Y. (2014a). A Bayesian dose-finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society Series C*, 63(1):159–173.

Cai, C., Liu, S., and Yuan, Y. (2014b). A Bayesian design for phase II clinical trials with delayed responses based on multiple imputation. *Statistics in medicine*, 33(23):4017–4028.

Carlin, B. P. and Louis, T. A. (2010). *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC.

Catenacci, D. V. (2015). Next-generation clinical trials: Novel strategies to address the challenge of tumor molecular heterogeneity. *Molecular oncology*, 9(5):967–996.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7(4):875–892.

Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in medicine*, 16(23):2701–2711.

Chen, Z., Krailo, M. D., Sun, J., and Azen, S. P. (2009). Range and trend of expected toxicity level (ETL) in standard A+ B designs: a report from the Children's Oncology Group. *Contemporary Clinical Trials*, 30(2):123–128.

Cheung, Y. K. (2011). *Dose Finding by the continual reassessment method*. CRC Press.

Cheung, Y. K. and Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chow, S.-C., Shao., J., Wang, H., and Lokhnygina, Y. (2017). *Sample size calculations in clinical research*. Chapman and Hall/CRC.

Chu, Y. and Yuan, Y. (2018a). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials*, 15(2):149–158.

Chu, Y. and Yuan, Y. (2018b). BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):723–740.

Conaway, M. R., Dunbar, S., and Peddada, S. D. (2004). Designs for Single-or Multiple-Agent phase I Trials. *Biometrics*, 60(3):661–669.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B., and Gönen, M. (2017). An efficient basket trial design. *Statistics in medicine*, 36(10):1568–1579.

Diaconis, P. (1985). Quantifying prior opinion (with discussions). *Bayesian Statist.*, 2:133–156.

Doussau, A., Geoerger, B., Jiménez, I., and Paoletti, X. (2016). Innovations for phase I dose-finding designs in pediatric oncology clinical trials. *Contemporary clinical trials*, 47:217–227.

Dragalin, V., Hsuan, F., and Padmanabhan, S. K. (2007). Adaptive designs for dose-finding studies based on sigmoid $E_{\max}$ model. *Journal of Biopharmaceutical Statistics*, 17(6):1051–1070.

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in medicine*, 13(17):1727–1736.

FDA (2013). Guidance for industry: codevelopment of two or more new investigational drugs for use in combination.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Flaherty, K. T., Puzanov, I., Kim, K. B., Ribas, A., McArthur, G. A., Sosman, J. A., O'Dwyer, P. J., Lee, R. J., Grippo, J. F., Nolop, K., et al. (2010). Inhibition of mutated, activated BRAF in metastatic melanoma. *New England Journal of Medicine*, 363(9):809–819.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, pages 143–151.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Freidlin, B., McShane, L. M., and Korn, E. L. (2010). Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160.

Gerber, F., Gsponer, T., et al. (2016). gsbDesign: an R package for evaluating the operating characteristics of a group sequential Bayesian design. *Journal of Statistical Software*, 69(11):1–27.

Goodman, S. N., Zahurak, M. L., and Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Statistics in medicine*, 14(11):1149–1161.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Gsponer, T., Gerber, F., Bornkamp, B., Ohlssen, D., Vandemeulebroecke, M., and Schmidli, H. (2014). A practical guide to Bayesian group sequential designs. *Pharmaceutical statistics*, 13(1):71–80.

Guo, W., Ji, Y., and Catenacci, D. V. (2017a). A subgroup cluster-based Bayesian adaptive design for precision medicine. *Biometrics*, 73(2):367–377.

Guo, W., Ji, Y., and Li, D. (2019). Rolling Toxicity Probability Interval Design to Shorten the Duration and Maintain Safety of phase I Trials. *Journal of Biopharmaceutical Statistics*, 29(2):411424.

Guo, W., Wang, S.-J., Yang, S., Lynn, H., and Ji, Y. (2017b). A Bayesian interval dose-finding design addressing Ockham's razor: mTPI-2. *Contemporary clinical trials*, 58:23–33.

Heitjan, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in medicine*, 16(16):1791–1802.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

Hobbs, B. P. and Landin, R. (2018). Bayesian basket trial design with exchangeability monitoring. *Statistics in medicine*, 37(25):3557–3572.

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., et al. (2015). Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine*, 373(8):726–736.

Ivanova, A., Flournoy, N., and Chung, Y. (2007). Cumulative cohort design for dose-finding. *Journal of Statistical Planning and Inference*, 137(7):2316–2327.

Ivanova, Anastasia and Wang, Kai (2006). Bivariate isotonic design for dose-finding with ordered groups. *Statistics in medicine*, 25:2018–26.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Ji, Y., Li, Y., and Nebiyou Bekele, B. (2007). Dose-finding in phase I clinical trials based on toxicity probability intervals. *Clinical Trials*, 4(3):235–244.

Ji, Y., Liu, P., Li, Y., and Nebiyou Bekele, B. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):653–663.

Ji, Y. and Wang, S.-J. (2013). Modified toxicity probability interval design: a safer and more reliable method than the 3+ 3 design for practical phase I trials. *Journal of Clinical Oncology*, 31(14):1785.

Ji, Y. and Yang, S. (2017). On the interval-based dose-finding designs. *arXiv preprint arXiv:1706.03277*.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., et al. (2011). The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Lee, J. J. and Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, 5:93–106.

Lee, S. M. and Cheung, Y. K. (2011). Calibration of prior variance in the Bayesian continual reassessment method. *Statistics in Medicine*, 30(17):2081–2089.

Li, D. H., Whitmore, J. B., Guo, W., and Ji, Y. (2017). Toxicity and Efficacy Probability Interval Design for Phase I Adoptive Cell Therapy Dose-Finding Clinical Trials. *Clinical Cancer Research*, 23(1):13–20.

Lin, R. and Yin, G. (2016). Bayesian optimal interval design for dose finding in drug-combination trials. *Statistical methods in medical research*, page 0962280215594494.

Lin, X. and Ji, Y. (2020a). Probability Intervals of Toxicity and Efficacy (PRINTE) design for CAR-T Dose Finding Clinical Trials. *Statistical Methods in Medical Research*, Accepted.

Lin, X. and Ji, Y. (2020b). The joint i3+3 (Ji3+3) design for phase I/II adoptive cell therapy clinical trials. *Contemporary Clinical Trials*, In press.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621.

Liu, M., Wang, S.-J., and Ji, Y. (2020). The i3+ 3 design for phase I clinical trials. *Journal of Biopharmaceutical Statistics*, 30(2):294–304.

Liu, R., Liu, Z., Ghadessi, M., and Vonk, R. (2017). Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemporary clinical trials*, 63:67–72.

Liu, S. and Yuan, Y. (2015a). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523.

Liu, S. and Yuan, Y. (2015b). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statal Society: Series C (Applied Stats)*, 64.

Lyu, J., Ji, Y., Zhao, N., and Catenacci, D. V. (2019). AAA: triple adaptive Bayesian designs for the identification of optimal dose combinations in dual-agent dose finding trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(2):385–410.

Lyu, J., Zhou, T., Yuan, S., Guo, W., and Ji, Y. (2020). MUCE: Bayesian Hierarchical Modeling for the Design and Analysis of Phase 1b Multiple Expansion Cohort Trials. *arXiv preprint arXiv:2006.07785*.

Mair, P., Hornik, K., and de Leeuw, J. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software*, 32(5):1–24.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in medicine*, 24(3):329–339.

Mander, A. P. and Sweeting, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Statistics in medicine*, 34(8):1261–1276.

McLachlan, G. J. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Moore, D. F. (2016). *Applied Survival Analysis Using R*. Springer, Switzerland.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.

Mullard, A. (2015). NCI-MATCH trial pushes cancer umbrella trial paradigm. *Nature Reviews Drug Discovery*, 14(8):513.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Neuenschwander, B. (2011). From historical data to priors. In *Proceedings of Biopharmaceutical Section, Joint Statistical Meetings, American Statistical Association, Miami Beach, Florida*, pages 3466–3474.

Neuenschwander, B., Branson, M., and Gsponer, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in medicine*, 27(13):2420–2439.

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18.

Neuenschwander, B., Matano, A., Tang, Z., Roychoudhury, S., Wandel, S., and Bailey, S. (2015). Bayesian industry approach to phase I combination trials in oncology. *Statistical Methods in Drug Combination Studies*, 2015:95–135.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15(2):123–134.

Nie, L., Rubin, E. H., Mehrotra, N., Pinheiro, J., Fernandes, L. L., Roy, A., Bailey, S., and de Alwis, D. P. (2016). Rendering the 3+ 3 design to rest: more efficient approaches to oncology dose-finding trials in the era of targeted therapy.

Normolle, D. and Lawrence, T. (2006). Designing dose-escalation trials with late-onset toxicities using the time-to-event continual reassessment method. *Journal of Clinical Oncology*, 24(27):4426–4433.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, pages 33–48.

Peter, F, Thall, Richard, C, Herrick, Hoang, Q, Nguyen, and and, J. (2014). Effective sample size for computing prior hyperparameters in Bayesian phase I-II dose-finding. *Clinical Trials*.

Porter, D. L., Levine, B. L., Kalos, M., Bagg, A., and June, C. H. (2011). Chimeric Antigen ReceptorModified T Cells in Chronic Lymphoid Leukemia. *New England Journal of Medicine*, 365(8):725–733. PMID: 21830940.

Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R. L., Bardelli, A., and Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF (V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–103.

Psioda, M. A., Xu, J., Jiang, Q., Ke, C., Yang, Z., and Ibrahim, J. G. (2019). Bayesian adaptive basket trial design using model averaging. *Biostatistics*.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Riviere, M.-K., Yuan, Y., Dubois, F., and Zohar, S. (2014). A Bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharmaceutical statistics*, 13(4):247–257.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Robertson, T. (1988). Order restricted statistical inference. Technical report.

Ruberg, S. J., Chen, L., and Wang, Y. (2010). The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical trials*, 7(5):574–583.

Sargent, D. J., Conley, B. A., Allegra, C., and Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9):2020–2027.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

Schoenfeld, D. A. (1983). Sample-Size Formula for the Proportional-Hazards Regression Model. *Biometrics*, 39(2):499–503.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, 110(509):303–312.

Shi, Y. and Yin, G. (2013). Escalation with overdose control for phase I drug-combination trials. *Statistics in medicine*, 32(25):4400–4412.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*, 10(1):1–10.

Simon, R., Geyer, S., Subramanian, J., and Roychowdhury, S. (2016). The Bayesian basket design for genomic variant-driven phase II trials. In *Seminars in oncology*, volume 43, pages 13–18. Elsevier.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, 10(20):6759–6763.

Sivaganesan, S., Laud, P. W., and Müller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in medicine*, 30(4):312–323.

Skolnik, J. M., Barrett, J. S., Jayaraman, B., Patel, D., and Adamson, P. C. (2008). Shortening the timeline of pediatric phase I trials: the rolling six design. *Journal of Clinical Oncology*, 26(2):190–195.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, pages 925–937.

Tan, S.-B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in medicine*, 21(14):1991–2012.

Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693.

Thall, P. F. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*, pages 337–349.

Thall, P. F., Simon, R. M., and Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in medicine*, 14(4):357–379.

Thall, P. F. and Sung, H.-G. (1998). Some extensions and applications of a bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in medicine*, 17(14):1563–1580.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in medicine*, 22(5):763–780.

Tiacci, E., Trifonov, V., Schiavoni, G., Holmes, A., Kern, W., Martelli, M. P., Pucciarini, A., Bigerna, B., Pacini, R., Wells, V. A., et al. (2011). BRAF mutations in hairy-cell leukemia. *New England Journal of Medicine*, 364(24):2305–2315.

Tighiouart, M., Li, Q., and Rogatko, A. (2017). A Bayesian adaptive design for estimating the maximum tolerated dose curve using drug combinations in cancer phase I clinical trials. *Statistics in medicine*, 36(2):280–290.

Tighiouart, M., Piantadosi, S., and Rogatko, A. (2014). Dose finding with drug combinations in cancer phase I clinical trials using conditional escalation with overdose control. *Statistics in medicine*, 33(22):3815–3829.

Wages, N. A. (2017). Identifying a maximum tolerated contour in two-dimensional dose finding. *Statistics in medicine*, 36(2):242–253.

Wages, N. A. and Conaway, M. R. (2014). Phase I/II adaptive design for drug combination oncology trials. *Statistics in medicine*, 33(12):1990–2003.

Wages, N. A., Portell, C. A., Williams, M. E., Conaway, M. R., and Petroni, G. R. (2017). Implementation of a model-based design in a phase Ib study of combined targeted agents. *Clinical Cancer Research*, 23(23):7158–7164.

Xu, Y., Trippa, L., Müller, P., and Ji, Y. (2016). Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences*, 8(1):159–180.

Yan, F., Mandrekar, S. J., and Yuan, Y. (2017). Keyboard: a novel Bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, 23(15):3994–4003.

Yang, S., Wang, S.-J., and Ji, Y. (2015). An integrated dose-finding tool for phase I trials in oncology. *Contemporary clinical trials*, 45:426–434.

Yin, G. and Yuan, Y. (2009). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(2):211–224.

Yuan, S., Zhou, T., Lin, Y., and Ji, Y. (2021). The CI3+3 Design for Dual-Agent Combination Dose-Finding Clinical Trials. *arXiv preprint arXiv:2103.13693*.

Yuan, Y., Lin, R., Li, D., Nie, L., and Warren, K. E. (2018). Time-to-event Bayesian optimal interval design to accelerate phase I trials. *Clinical Cancer Research*, 24(20):4921–4930.

Zar, J. H. (2010). *Biostatistical Analysis*. Prentice-Hall, 5th edition.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539.

Zhou, H., Lee, J. J., and Yuan, Y. (2017). BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in medicine*, 36(21):3302–3314.

Zhou, T., Guo, W., and Ji, Y. (2019a). PoD-TPI: Probability-of-Decision Toxicity Probability Interval Design to Accelerate Phase I Trials. *Statistics in Biosciences*, pages 1–22.

Zhou, Y., Lee, J. J., and Yuan, Y. (2019b). A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Statistics in Medicine*, 38(28):S5299–S5316.