An Integrated Dose-Finding Tool for Phase I Trials in Oncology

Shengjie Yang,

Program of Computational Genomics & Medicine, Northshore University HealthSystem

Sue-Jane Wang,

Center for Drug Evaluation and Research

US Food and Drug Administration

Yuan Ji [*]

Program of Computational Genomics & Medicine, Northshore University HealthSystem

Department of Public Health Sciences, The University of Chicago, Chicago, USA

September 22, 2015

Abstract

In the past 25 years, the 3+3 design has been the most popular approach for planning phase I dose-finding trials in oncology. During the same time period, major development of more efficient model-based designs has been made by statistical researchers

[*] Correspondence: 1001 University Place, Evanston, IL 60201. E-mail: koaeraser@gmail.com

aiming to improve the clinical practice of dose finding in oncology. Despite the effort, 3+3 is still the most frequently used designs in practice. Part of the reason is due to the lack of software tools that allow comparison of different designs, including 3+3 and other model-based methods, in a head-to-head and easy-to-use fashion. To this end, we introduce NextGen-DF, a next-generation tool for designing oncology dose-finding trials that allows for construction, comparison, and calibration of multiple designs via internet, in real time, and independent of computer operating systems. Through NextGen-DF, we present massive and user-generated comparison results based on over 4 million simulated trials, which clearly indicate the inferiority of 3+3. To our knowledge, the reported crowd-sourcing results are the largest and most objective comparison across major dose-finding methods to date. NextGene-DF is expected to improve patient care and drug development by providing safer and more efficient designs for phase I oncology trials. NextGen-DF is available at www.compgenome.org/NGDF.

Keywords: 3+3; Bayesian design; CRM; mTPI; Next-generation dose finding; Webtool.

1 Introduction

In oncology, phase I dose-finding trials are the initial first-in-human experiments to test the safety of a new treatment. Ascending doses of the treatment are given to patients sequentially and dose-limiting toxicity (DLT) events are monitored with a goal to find the maximum tolerated dose (MTD), the highest dose with a probability of toxicity of DLT less than a target rate, usually denoted by p_T. For more than two decades, the majority of phase I oncology trials is based on the 3+3 design [1]. In recent in-depth reviews [2, 3], it has been shown that in practice more than 90% of the trials use the 3+3 design. These statistics illustrate the popularity of the 3+3 design. It would be of interest to learn whether 3+3 leads to a high success rate in phase I oncology trials.

Despite the popularity of 3+3, the performance of Phase I trials in oncology has been miserable. The largest and most recent survey [4] of 835 drug developers, including biotech and large pharmaceutical firms, showed alarming failure rates of oncology trials compared with those in other disease indications, with the largest discrepancy seen in phase I trials. In particular, the survey pointed out that non-oncology phase I trials were twice likely to produce an FDA approval than oncology phase I trials. Such failure is likely due to the wrong doses recommended from the phase I studies. Without selecting the right dose, a promising treatment could fail later in confirmatory studies due to either high toxicity caused by over dosing or lack of efficacy caused by under dosing.

The standard 3+3 design has been shown by numerous statistical publications to be an inferior, over-simplified and algorithm-driven method when compared to model-based trial designs [5, 6, 7, 8, 9, 10, 11]. For example in [9] we demonstrated that the mTPI design [8], when compared to 3+3 based on matched sample size, was safer in protecting patients from being exposed to overly toxic doses, and still possessed sufficient power in identifying the true MTD.

The question is then, why is there such a discrepancy between the popularity of 3+3 in practice and its unpopularity in the scientific literature? Also, why did 3+3 remain to be the most frequently used design when phase I trials in oncology failed miserably? Simplicity of the 3+3 design makes it the most popular choice in practice. In particular, this was the case in the late 1980s and early 1990s when computing power was limited. However, given humans rapid gain in PC computing over the last two decades, more sophisticated model-based designs with better safety protection and higher chance of finding the MTD should outweigh methods that are inferior but require less computation. There is a practical need to upgrade designs and conduct for phase I oncology trials. Currently, there are few computational tools that allow direct comparison of multiple designs through a user-friendly computing environment. Also, most efficient adaptive designs cannot provide a transparent interface and require advanced IT infrastructure to implement in real-world settings. We believe that this is the main reason why better designs are still unable to penetrate the practical barrier.

Aiming at providing a user-friendly next-generation tool for the community, we introduce NextGen-DF (Next-Generation Dose Finding) for phase I oncology trials. It is a next-generation tool for three reasons. First, NextGen-DF is a web tool allowing investigators to obtain, compare, and record statistical designs online in a web browser. This spares users from needing to download, organize, and maintain software packages for various designs. Second, NextGen-DF does not require a companion and dependable operating system or computing software. It runs on an internet browser, such as Firefox, Safari, or Chrome, and can be accessed from a PC, Mac, Tablet, or even a smart phone anywhere in the world via internet. Third, NextGen-DF is based on optimized computer programs that achieve ultra fast online computing and allows the simulation results to be obtained and visualized within seconds, except for one function that includes comparison to a third-party program.

The remainder of the paper is organized as follows. Section 2 introduces the dose-finding designs in the NextGen-DF tool, including the 3+3 design, CRM, mTPI, and a new practical design derived from mTPI based on user preference. Section 3 describes the detail of NextGen-DF as a web tool. Section 4 presents the crowd-sourcing comparison results of the designs in user-generated simulations. We end the paper with a brief discussion in Section 5.

2 Methods in NextGen-DF

2.1 The mTPI Design

Phase I oncology trials aim to identify the maximum tolerated dose (MTD) of an investigational drug, the highest dose with a probability of dose-limiting toxicity (DLT) lower than (or close to, in some cases) a targeted probability, denoted by p_T. Patients are enrolled and treated sequentially at ascending dose levels of the investigational drug, and escalate or de-escalate the dose level depending on the observed DLTs. Proper designs provide rules of escalation and de-escalation based on the observed DLT outcomes, aiming to locate the MTD quickly without exposing patients to excessive toxicity.

NextGen-DF implements and compares three main-stream phase I dose-finding designs, the standard 3+3 design [1], the continual reassessment method (CRM) [6], and the modified toxicity probability interval design known as mTPI [8, 9]. These three designs receive more attention in practice. The CRM design has been thoroughly studied and discussed by the research community since 1990, with numerous extensions and modifications. The mTPI design is relatively new but has already shown great potential in practice [12, 13, 14], recognized and adopted quickly by practitioners. NextGen-DF implements the standard 3+3 design, the CRM design provided in the R package "dfCRM", and the mTPI design. Additionally, customized designs based on mTPI are also included to incorporate personalized decisions. We plan to discuss these four designs next, starting from mTPI.

The modified toxicity probability interval (mTPI) design is a model-based method for phase I clinical trials [8, 9]. It is an extension from the toxicity probability interval (TPI) method [7], a founding method that established the use of toxicity probability intervals for dose finding. The mTPI design employs a simple hierarchical model based on independent beta priors. This setting has been shown to be effective in standard phase I dose finding studies with single agents [15] as well as in recent two-dimensional dose finding development [16]. Suppose a total of d doses is prespecified for a dose finding trial. Let q_i denote the true but unknown toxicity probability for dose i, $i = 1, \dots, d$. After launching the trial, at any moment the observed data include n_i patients treated at dose i and the corresponding $x_i \in [0, n_i]$ experiencing DLTs. The likelihood function for data $\{(x_i, n_i), i = 1, \dots, d\}$ is a product of binomial probabilities,

$$l(\boldsymbol{q}) \propto \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{n_i - x_i} \tag{1}$$

where $\boldsymbol{q} = (q_1, \dots, q_d)$ is the vector containing all the q_i's. The mTPI method calculates the unit probability mass (UPM) of three probability intervals corresponding to underdosing, proper dosing, and overdosing, and assigns a dose for future patients based on a Bayesian decision rule. Specifically, the underdosing interval is defined as $(0, p_T - \varepsilon_1)$ which corresponds to a dose escalation (E), the overdosing interval $(p_T + \varepsilon_2, 1)$ corresponds to a dose de-escalation (D), and the proper dosing interval $(p_T - \varepsilon_1, p_T + \varepsilon_2)$ corresponds to staying at the current dose (S). 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. Define the UPMs for three intervals as

$$\begin{aligned} UPM(D, i) &= \frac{P(q_i - p_T > \varepsilon_2 | data)}{1 - p_T - \varepsilon_2}, \\ UPM(S, i) &= \frac{P(-\varepsilon_1 \leq q_i - p_T \leq \varepsilon_2 | data)}{\varepsilon_1 + \varepsilon_2}, \\ UPM(E, i) &= \frac{P(q_i - p_T < -\varepsilon_1 | data)}{p_T - \varepsilon_1}, \end{aligned}$$

each of which is the ratio between the posterior probability and length of a dosing interval.

The dose-assignment rule B_i, chooses the decision among D, S, or E, that produces the largest value of UPM, that is,

$$B_i = \begin{cases} D, & \text{if } UPM(D,i) > UPM(S,i) \vee UPM(E,i), \\ S, & \text{if } UPM(S,i) > UPM(D,i) \vee UPM(E,i), \\ E, & \text{if } UPM(E,i) > UPM(D,i) \vee UPM(S,i), \end{cases} \tag{2}$$

Ji et al. (2010) showed that decision B_i is optimal in that it minimizes the posterior expected loss, with the loss determined to achieve equal prior expected loss for the three decisions, D, S, and E. The mTPI design uses a set of independent and vague priors $q_i \sim Beta(1,1)$. Combined with the likelihood in (1), the posterior distribution of q_i follows independent $Beta(1+x_i, 1+n_i-x_i)$, for $i = 1, \ldots, d$. When strong prior information on the toxicity of the candidate doses is available, informative beta priors can replace the vague priors. The choice of the independent prior models might seem counter-intuitive here since the toxicity probabilities are in general believed to be monotone increasing over dose levels. Therefore, theoretically it is desirable to introduce dependent models for q_i's. However, we believe that for phase I trials with small sample sizes, especially when only a small number of patients are treated at a given dose, may not benefit from dependent models alone. Designs using dependent prior models need be complemented by practical rules [17, 18] to ensure the safety and desirability of the designs.

In addition to the Bayes rule in (2), we recommend two additional rules for additional protection of patient safety.

- **Safety rule 1 (early termination):** Suppose that dose 1 has been used to treat patients. If $P(q_1 > p_T|data) > \xi$ for a ξ close to 1 (say, $\xi = 0.95$), then terminate the trial due to excessive toxicity. Otherwise, terminate the trial when the maximum sample size is reached.

- **Safety rule 2 (dose exclusion):** Suppose that the decision is E, to escalate from dose i to $(i+1)$. If $P(q_{i+1} > p_T|data) > \xi$, for a ξ close to 1 (say, $\xi = 0.95$), then treat the next cohort of patients at dose i instead of $(i+1)$ and exclude doses $(i+1)$ and higher from the trial, i.e., these doses will never be used again in the trial.

At the end of the trial when the toxicity outcomes of all the enrolled patients are observed, a dose will be selected as the estimated MTD for subsequent studies. We view this as an estimation problem that is separated from the design for dose finding. Following Ji et al. (2010), we propose to select the MTD by performing an isotonic regression that borrows strength cross doses. We first compute the posterior mean $\hat{q}_i$ under the beta posterior distribution and then apply the pooled adjacent violators algorithm (PAVA) [19] on $\hat{q}_i$ so that the resulting transformed values $\hat{q}_i^*$ increase with the dose levels. That is, $\hat{q}_j^* \geq \hat{q}_i^*$ for $j \geq i$. The recommended MTD is the dose with a toxicity probability closest to p_T, i.e.,

$$\text{Estimated MTD } = \arg\min_i |\hat{q}_i^* - p_T|.$$

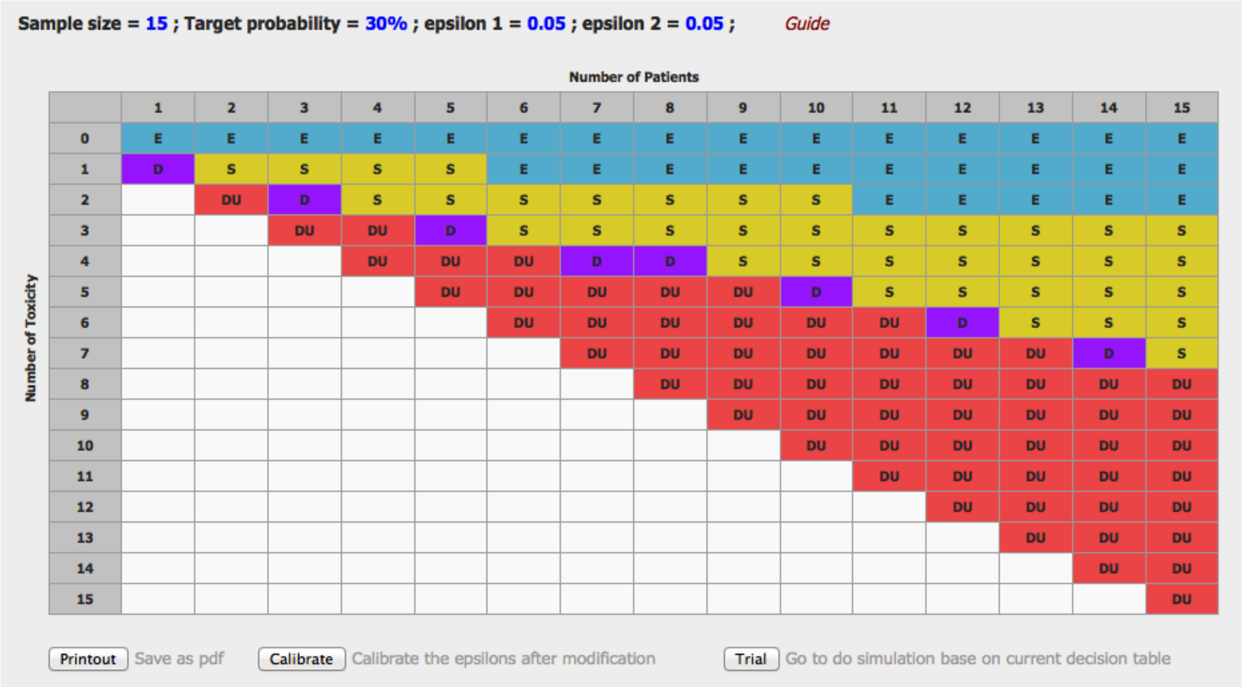
2.2 The Customized Designs

The second design in NextGen-DF is new and derived from mTPI based on user preference. We call it the *customized design*. Customized designs are rule-based and derived directly from modifications of the decision table provided by mTPI. To obtain a customized design, one simply starts by generating a decision table under the mTPI design for a trial. This only requires providing the maximum sample size and the target MTD toxicity rate p_T. A decision table from mTPI can be generated in NextGen-DF with these two trial parameters, such as the one in Figure 1. Then in NextGen-DF a customized design can be constructed by changing any decision in the table to a different one when desired. NextGen-DF allows the new customized designs to be compared head-to-head with mTPI in simulated clinical trials, thus allowing users to see the impact of the changes on the operating characteristics of the designs. For example, when p_T is set at 0.3, mTPI will perform S, i.e., staying at the current dose, if 3 DLTs are observed from 6 treated patients (row 3, column 6 in the decision table in Figure 1). A user might wonder how well the design would perform if the decision for 3 (DLTs) out of 6 (patients) were D, to de-escalate. NextGen-DF allows users to change S to D, de-escalate (see Quick Start Guide at `www.compgenome.org/NGDF`). Once the table is changed, a customized design is constructed that uses a different decision table from mTPI. NextGen-DF then allows users to compare the customized design with existing designs via simulations.

2.3 The CRM Design

We use the "dfCRM" R package described in [20] for the implementation of CRM. NextGen-DF directly calls the R package and use the "getprior()" function in the package to specify the initial guesses of toxicity probabilities associated with the doses. We also use the default dose-response model specified by the R package, which is an "empirical" model. Briefly, the CRM model assumes a parametric dose-response curve, in the form of

$$q_i(q_{i0}, \beta) = q_{i0}^{\beta},$$



Sample size = 15 ; Target probability = 30% ; epsilon 1 = 0.05 ; epsilon 2 = 0.05 ;

Number of Patients

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
1	D	S	S	S	S	E	E	E	E	E	E	E	E	E	E
2		DU	D	S	S	S	S	S	S	S	E	E	E	E	E
3			DU	DU	D	S	S	S	S	S	S	S	S	S	S
4				DU	DU	DU	D	D	S	S	S	S	S	S	S
5					DU	DU	DU	DU	DU	D	S	S	S	S	S
6						DU	DU	DU	DU	DU	DU	D	S	S	S
7							DU	DU	DU	DU	DU	DU	DU	D	S
8								DU	DU	DU	DU	DU	DU	DU	DU
9									DU	DU	DU	DU	DU	DU	DU
10										DU	DU	DU	DU	DU	DU
11											DU	DU	DU	DU	DU
12												DU	DU	DU	DU
13													DU	DU	DU
14														DU	DU
15															DU

Guide
Number of Toxicity
Printout
Save as pdf
Calibrate
Calibrate the epsilons after modification
Trial
Go to do simulation base on current decision table

Figure 1: A decision table generated by mTPI for a trial of 15 patients and MTD target toxicity probability $p_T = 0.3$. The letters “E”, “S”, and “D”represent decisions of dose escalation, dose stay (unchanged), dose de-escalation, respectively. Letter “U” means the dose is too toxic and unacceptable, and should be excluded from the trial. The decisions in the Table are precalculated based on a Bayesian hierarchical model in Ji et al. (2010). NextGen-DF allows users to change any decisions in this table to different decisions, which would result in new customized designs.

where q_{i0} is an initial guess of the toxicity probability for dose i and β is a power parameter. Given a set of initial guesses $\hat{\mathbf{q}}_0 = \{\hat{q}_{10}, \ldots, \hat{q}_{d0}\}$, the unknown parameter β is estimated based on a Bayesian procedure in which a prior $p(\beta)$ is used to update the binomial likelihood (1). The posterior $p(\beta \mid data)$ is obtained, and the recommended dose for the next cohort of patients is given by

$$\hat{i} = \arg\min_i |q_i(\hat{q}_{i0}, \hat{\beta}) - p_T| \tag{3}$$

where $\hat{\beta}$ is the posterior mean $E(\beta \mid data)$.

Lee and Cheung (2009) [21] proposed a calibration procedure that returns a set of prior guesses $\{q_{10}, \ldots, q_{d0}\}$ by imposing constraints on the dose response curves. Specifically, they considered two sets of constraints as

$$\begin{aligned} q_{i-1}(q_{i-1,0}, b_i) + q_i(q_{i0}, b_i) = 2p_T \quad &\text{and} \quad q_i(q_{i,0}, b_{i+1}) + q_{i+1}(q_{i+1,0}, b_{i+1}) = 2p_T \\ q_{i-1}(q_{i-1,0}, b_i) = p_T - \delta \quad &\text{and} \quad q_{i+1}(q_{i+1,0}, b_{i+1}) = p_T + \delta \end{aligned} \tag{4}$$

where $(p_T - \delta, p_T + \delta)$ is called an "indifference" interval. Similar to the "proper dosing" interval in the mTPI design, the indifference interval contains the doses whose toxicity probabilities are close approximations of p_T, i.e., doses that can be approximated as the MTD. Solving above two equations iteratively for all doses i, one gets an estimated guess $\hat{\mathbf{q}}_0$ for the entire trial, and assigns $\hat{i}$ in (3) to the next cohort of patients. The trial proceeds until the DLT outcomes from the next cohort of patients are observed, at which point $\hat{i}$ is re-estimated. The trial stops when the maximum sample size is reached.

2.4 The 3+3 Design

Based on [1] we implement the 3+3 design described in Figure 2. To our knowledge, this is the standard 3+3 design used by the pharmaceutical industry and many research institutes. Below is a detailed algorithm for practical implementation.

The 3+3 Design:

1). Start trial by treating three patients at the initial dose (usually the lowest dose).

2). Denote the dose level being used to treat patients as the current dose level. Accrue and treat three patients at the current dose level.

2a. If the maximum number of patients has been accrued, stop the trial. The MTD is inconclusive.

3). Check the number of patients at the current dose level.

3a. If there are three patients, go to 4).

3b. If there are six patients, go to 5).

4). Check the number of toxicities (among three patients) at the current dose level.

4a. If there are zero toxicities, escalate and go to 7).

4b. If there is one toxicity, stay at the current dose and go to 2).

4c. If there are two or three toxicities, declare that the MTD has been exceeded and go to 6).

5). Check the number of toxicities (among six patients) at the current dose level.

5a. If there are zero toxicities, stop the trial and declare that the MTD is the current dose.

5b. If there is one toxicity, and the MTD has been exceeded, stop the trial and declare that the MTD is the current dose; otherwise, go to 7).

5c. If there are two or more than two toxicities, declare that the MTD has been exceeded and go to 6).

6). The MTD has been exceeded.

6a. If the current dose is the lowest dose, stop the trial and declare that the MTD is lower than the lowest dose level.

6b. If the next-lower dose level has six patients, stop the trial and declare that the MTD is the next lower dose level; otherwise, the next-lower dose level has three patients; set the current dose level to be the next-lower dose level and go to 2).

7). Escalate if possible.

7a. If the current dose level is the highest dose level, stop the trial and declare that the MTD is the highest dose level.

7b. Otherwise, escalate to the next-higher dose level and go to 2).

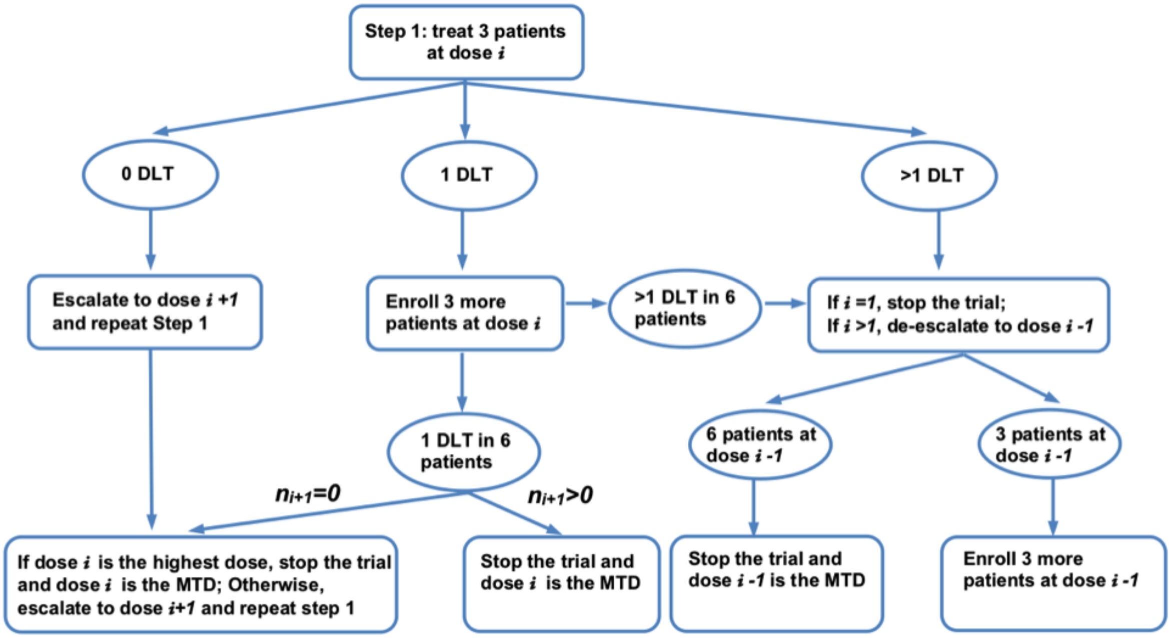

Step 1: treat 3 patients at dose i
0 DLT
1 DLT
>1 DLT
Escalate to dose i +1 and repeat Step 1
Enroll 3 more patients at dose i
>1 DLT in 6 patients
If i =1, stop the trial; If i >1, de-escalate to dose i -1
1 DLT in 6 patients
6 patients at dose i -1
3 patients at dose i -1
ni+1=0
ni+1>0
If dose i is the highest dose, stop the trial and dose i is the MTD; Otherwise, escalate to dose i+1 and repeat step 1
Stop the trial and dose i is the MTD
Stop the trial and dose i -1 is the MTD
Enroll 3 more patients at dose i -1

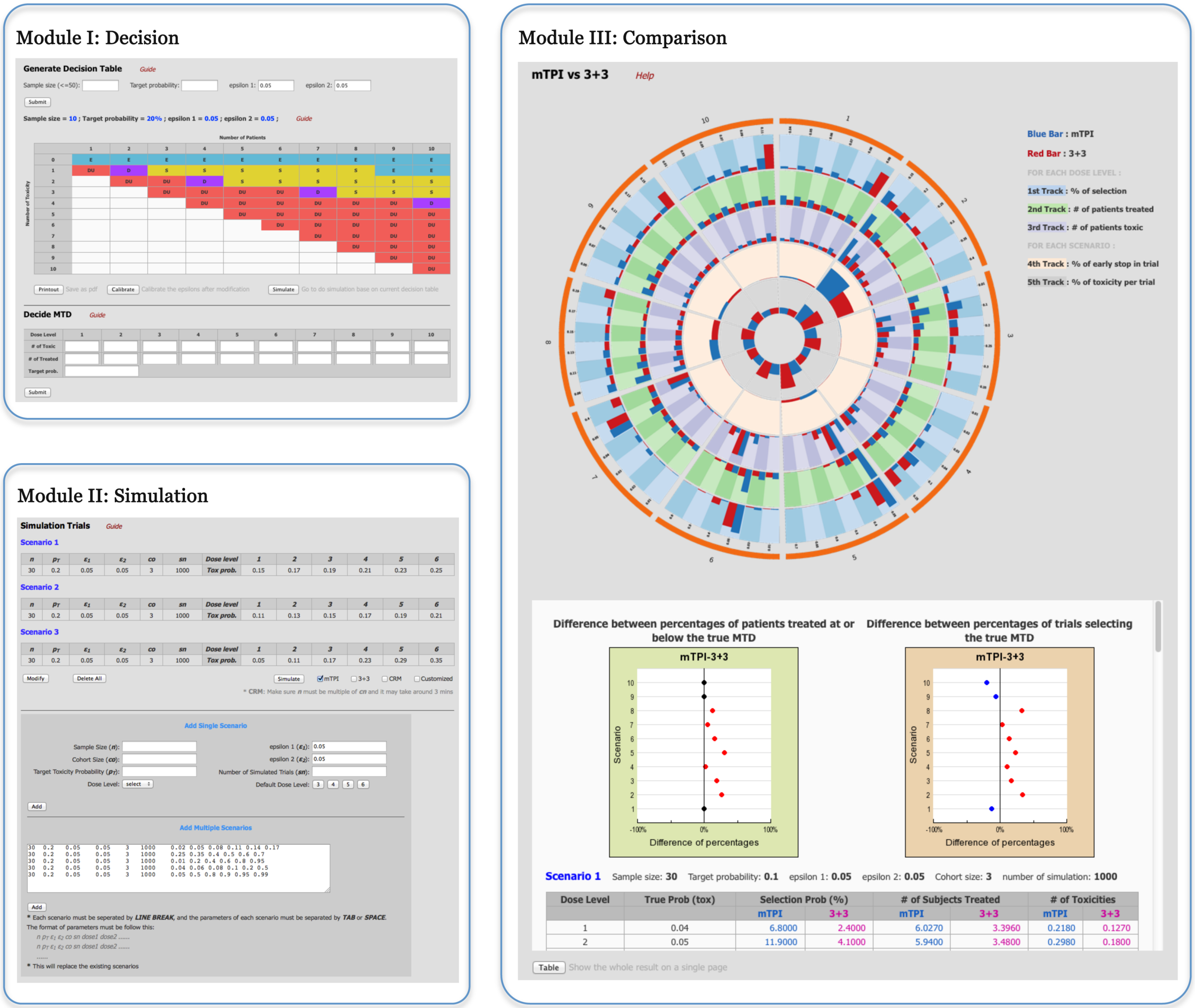
Figure 2: Schema of the standard 3+3 design. Assuming dose i is the dose currently being used in the trial, the schema illustrates all the possible decisions under 3+3. Notation n_{i+1} refers to the number of patients that have been treated at dose level $i+1$, the next higher dose.

3 NextGen-DF Web Tool

NextGen-DF contains three modules (Figure 3), each providing a different utility.

Module I, *Decision*, generates decision tables under the mTPI design. These decisions are constructed based on the Bayesian hierarchical models and statistical inference under the mTPI design. The use of independent priors in mTPI makes the implementation of mTPI effortless and transparent, greatly enhancing its feasibility in practice. In particular, a decision table (Figure 3, Module I) that contains all the possible decisions can be precalculated using the posterior beta distributions. The decision table will then be used repeatedly to guide dose finding decisions. For detail, see Ji et al. (2010) and Ji and Wang (2013). The Decision module also provides a function based on the isotonic transformation [9] that provides an estimated MTD given the data from a completed trial.

Innovatively, we allow the users to freely modify any decision entries in the decision table, which could result in a new and customized decision table. Each customized table corresponds to a new customized design. As an option, NextGen-DF can also back calibrate ε's in the mTPI design and see if there are any values of ε's that would give the same table users changed to. That is, NextGen-DF can attempt to calibrate the mTPI design ε's trying to match the new decision table users specified. In particular, NextGen-DF sets up a search program in the background which will traverse all the combinations of epsilons in the domain of $(0 < \varepsilon_1 < p_T,\ p_T < \varepsilon_2 < 1)$ with a step size of 0.001. Each combination will generate an mTPI decision table, and the numbers of matched and mismatched decisions between each mTPI table and the customized table will be enumerated. NextGen-DF reports an mTPI table that matches the customized table when possible. If none of the mTPI tables matches the customized table, NextGen-DF reports one that has the smallest mismatches. See Quick Start Guide at `www.compgenome.org/NGDF` for an example. Users can opt to using



Module I: Decision

Generate Decision Table *Guide*

Sample size (<=50): ____ Target probability: ____ epsilon 1: 0.05 epsilon 2: 0.05

Submit

Sample size = 10 ; Target probability = 20% ; epsilon 1 = 0.05 ; epsilon 2 = 0.05 ; *Guide*

Number of Patients

Number of Toxicity	1	2	3	4	5	6	7	8	9	10
0	E	E	E	E	E	E	E	E	E	E
1	DU	D	S	S	S	S	S	S	E	E
2		DU	DU	D	S	S	S	S	S	S
3			DU	DU	DU	DU	D	S	S	S
4				DU	DU	DU	DU	DU	DU	D
5					DU	DU	DU	DU	DU	DU
6						DU	DU	DU	DU	DU
7							DU	DU	DU	DU
8								DU	DU	DU
9									DU	DU
10										DU

Printout Save as pdf Calibrate Calibrate the epsilons after modification Simulate Go to do simulation base on current decision table

Decide MTD *Guide*

Dose Level	1	2	3	4	5	6	7	8	9	10
# of Toxic										
# of Treated										
Target prob.										

Submit

Module II: Simulation

Simulation Trials *Guide*

Scenario 1

n	p_T	ε_1	ε_2	co	sn	Dose level	1	2	3	4	5	6
30	0.2	0.05	0.05	3	1000	Tox prob.	0.15	0.17	0.19	0.21	0.23	0.25

Scenario 2

n	p_T	ε_1	ε_2	co	sn	Dose level	1	2	3	4	5	6
30	0.2	0.05	0.05	3	1000	Tox prob.	0.11	0.13	0.15	0.17	0.19	0.21

Scenario 3

n	p_T	ε_1	ε_2	co	sn	Dose level	1	2	3	4	5	6
30	0.2	0.05	0.05	3	1000	Tox prob.	0.05	0.11	0.17	0.23	0.29	0.35

Modify Delete All Simulate ☑ mTPI ☐ 3+3 ☐ CRM ☐ Customized

* **CRM**: Make sure n must be multiple of cn and it may take around 3 mins

Add Single Scenario

Sample Size (n): ____ epsilon 1 (ε_1): 0.05

Cohort Size (co): ____ epsilon 2 (ε_2): 0.05

Target Toxicity Probability (p_T): ____ Number of Simulated Trials (sn): ____

Dose Level: select Default Dose Level: 3 4 5 6

Add

Add Multiple Scenarios

```
30  0.2   0.05    0.05    3   1000    0.02 0.05 0.08 0.11 0.14 0.17
30  0.2   0.05    0.05    3   1000    0.25 0.35 0.4 0.5 0.6 0.7
30  0.2   0.05    0.05    3   1000    0.01 0.2 0.4 0.6 0.8 0.95
30  0.2   0.05    0.05    3   1000    0.04 0.06 0.08 0.1 0.2 0.5
30  0.2   0.05    0.05    3   1000    0.05 0.5 0.8 0.9 0.95 0.99
```

Add

* Each scenario must be seperated by ***LINE BREAK***, and the parameters of each scenario must be separated by ***TAB*** or ***SPACE***.
The format of parameters must be follow this:

n p_T ε_1 ε_2 co sn dose1 dose2 ......
n p_T ε_1 ε_2 co sn dose1 dose2 ......
......

* This will replace the existing scenarios

Module III: Comparison

mTPI vs 3+3 *Help*

Blue Bar : mTPI
Red Bar : 3+3
FOR EACH DOSE LEVEL :
1st Track : % of selection
2nd Track : # of patients treated
3rd Track : # of patients toxic
FOR EACH SCENARIO :
4th Track : % of early stop in trial
5th Track : % of toxicity per trial

1
2
3
4
5
6
7
8
9
10
Difference between percentages of patients treated at or below the true MTD
mTPI-3+3
Scenario
10
9
8
7
6
5
4
3
2
1
-100%
0%
100%
Difference of percentages
Difference between percentages of trials selecting the true MTD
mTPI-3+3
Scenario
-100%
0%
100%
Difference of percentages

Scenario 1 Sample size: **30** Target probability: **0.1** epsilon 1: **0.05** epsilon 2: **0.05** Cohort size: **3** number of simulation: **1000**

Dose Level	True Prob (tox)	Selection Prob (%)		# of Subjects Treated		# of Toxicities	
		mTPI	3+3	mTPI	3+3	mTPI	3+3
1	0.04	6.8000	2.4000	6.0270	3.3960	0.2180	0.1270
2	0.05	11.9000	4.1000	5.9400	3.4800	0.2980	0.1800

Table Show the whole result on a single page

Figure 3: Workflow of NextGen-DF. Module I Decision includes four functionalities: 1) generate decision table, 2) calibrate design parameters based on mTPI, 3) personalize decision table to generate new rule-based designs, and 4) estimate the MTD after the trial is completed. Module II Simulation includes two functionalities: 1) add or modify simulation scenarios, and 2) simulate trials using mTPI, 3+3, CRM, or the new rule-based designs from Module I. Module III Comparison includes one functionality: compare the safety and reliability of the designs through graphical and tabular displays.

the customized design nonetheless and comparing the designs in Modules II and III.

Module II, *Simulation*, sets up computer simulation studies for comparing 3+3, CRM, mTPI and potentially the customized designs generated from the Decision module. In this module, users need to provide simulation scenarios for the clinical trial, such as the number of doses, maximum sample size, target MTD probability p_T, and the true toxicity probabilities for the doses. NextGen-DF allows users to provide multiple scenarios to test different dose-response shapes. NextGen-DF provides multiple means of generating simulation scenarios, by manually inputting one scenario at a time, batch inclusion of multiple scenarios in a user-prepared file, or inclusion of recommended build-in scenarios in the software. Investigators will be able to change any aspects of the scenarios at any time, even after the scenarios have been selected and the simulation studies have been conducted. This feature is useful in practice as investigators are likely to modify scenarios multiple times after reviewing the simulation results.

Module III, *Comparison*, displays the simulation results of different designs in graphical and tabular forms (Figure 3). A circos plot [22] is generated summarizing results into a single figure, greatly facilitating the visualization of the simulation outcomes. Aside from the usual summary statistics such as the average numbers of patients treated and experienced toxicity at each dose, or the average sample size and overall percentage of toxicity, we introduce two additional summary statistics [9] to evaluate and compare the safety and reliability of any two designs out of mTPI, 3+3, CRM, and the customized design. The two summary statistics are

- $n_{<=MTD}\%$: percentage of the patients treated at or below the true MTD.
- $Sel_{MTD}\%$: percentage of simulated trials selecting the true MTD.

The value $n_{<=MTD}\%$ directly evaluates the safety of a design in terms of the patients treated at doses lower than or equal to the true MTD. The value of $Sel_{MTD}\%$ measures the reliability of the design in selecting the true MTD. To calculate $n_{<=MTD}\%$ and $Sel_{MTD}\%$, we need to decide which dose will be considered as the MTD for each scenario. Note that in many scenarios, none of the doses have toxicity probabilities exactly equal to p_T. However, some dose-toxicity probabilities are close enough to be declared the MTDs. This is consistent with the real-world belief that the exact value p_T is almost never achieved for a dose. To this end, we assume that any doses with true toxicity probabilities in the proper dosing interval $(p_T - \varepsilon_1, p_T + \varepsilon_2)$ would be considered as an estimated MTD. For example, if users set $\varepsilon_1 = \varepsilon_2 = 0.05$, any dose in the interval (0.05, 0.15) when $p_T = 0.1$, (0.15, 0.25) when $p_T = 0.2$, and (0.25, 0.35) when $p_T = 0.3$ will be considered as an estimated MTD. If no dose is in the proper dosing interval, the highest dose with true toxicity probability less than p_T is considered the MTD. If the MTD could not be identified (e.g., if all the doses have toxicity probabilities higher than p_T), the correct decision is not to select any dose. In that case, the percentage of ”none” selection (i.e., no dose selected as the MTD) is used to compare among the designs.

The computational speed of the simulation runs is ultra fast, costing a few seconds to complete the simulations for all the designs as long as one does not include the CRM in the comparison. The CRM is based on the R package “dfCRM” which takes longer time to compute. Nevertheless, even if CRM is included, NextGen-DF usually take minutes to complete thousands of simulated trials, which should be acceptable for most users. Exclusion of CRM would shorten the computation to seconds, allowing for real-time calibration. For convenience, we provide a time estimate for completing the simulation studies when CRM is included in the comparison. All the results can be saved locally as files for future reference.

4 Results

4.1 Overall Results

Since the launch of NextGene-DF website at http://www.compgenome.org/NGDF, within a few months there have been over 100 registered users from 25 different countries and regions spanning America, Europe, Asia, and Oceania. The affiliations of these users are mostly academic institutions and industry giants (see website above). Below we report user-generated simulation results (through using NextGene-DF) comparing different designs, which provide unbiased and crowd-sourcing evaluation of these designs. As will be seen next, 3+3 is the inferior design, which addresses some recent questions raised in the literature [23].

We recorded simulated trials and results conducted by NextGene-DF users and incorporated the results into a database. The database consists of results for 3+3, CRM and mTPI based on **1,275** different scenarios and **4,327,385** simulated trials. To our knowledge, this is the largest study evaluating the three designs to date. In each scenario, users specified a number of doses and a set of true toxicity probabilities for all the doses, generated toxicity response data based on the true toxicity probabilities of the doses, and conducted trials on computer using each of the three designs. At the end of simulation, users summarized the operating characteristics of the three designs and compared their performance. We summarize the pairwise comparisons of mTPI, 3+3 and CRM in terms of reliability (Sel_{MTD}) and safety ($n_{\leq MTD}$) based on the results database, see Figure 4. The first three boxplots compare the reliability of the three designs. We calculate the average reliability of each design across all the scenarios, and present the difference of the average reliability between any two designs. The right three boxplots show the differences in the average safety. In all boxplots, a higher than zero value indicates that the first design of the pair has a better performance. For example, the most left boxplot implies that mTPI is more reliable in finding the true MTD than 3+3. Examining all six boxplots, it is clear that mTPI is the most desirable design with higher reliability and safety and 3+3 is the least desirable design.

4.2 Analysis of Simulation Results

We perform a careful subgroup analysis of the simulation results based on 593 scenarios in which both CRM and mTPI were used by the NextGene-DF users. In the other scenaiors, only one of the two methods was used. The goal is to identify factors that characterize these scenarios and associate them with the performance of different dose-finding methods. We focus on the comparison of CRM and mTPI, excluding 3+3, as each method has been shown to be superior than 3+3 in the literature.

Setup. For each scenario, a set of true toxicity probabilities $\mathbf{p} = (p_1, \ldots, p_d)$ are specified by NextGene-DF users. A large number of clinical trials (mostly between 1,000 to 10,000 trials) is simulated for each scenario based on CRM and mTPI designs, and two main summary statistics are calculated as part of the operating characteristics of the designs. They are $n_{<\leq MTD}$ and Sel_{MTD}, as defined in §3 of the main text. For each scenario and each design, we obtain a pair of values for $n_{<\leq MTD}$ and Sel_{MTD}. We then take the difference in each value between the two designs (mTPI - CRM), and represent the differences as diff$-n$ and diff$-sel$. A positive/negative value of diff$-n$ or diff$-sel$ means that mTPI is safer/less safe, or more/less likely to locate the true MTD than CRM. We use two diff's as our response variables for subgroup analysis. For predictive variables, or covariates, we define six variables based on the characteristics of the scenarios. They are provided in Table 1. Specifically, for

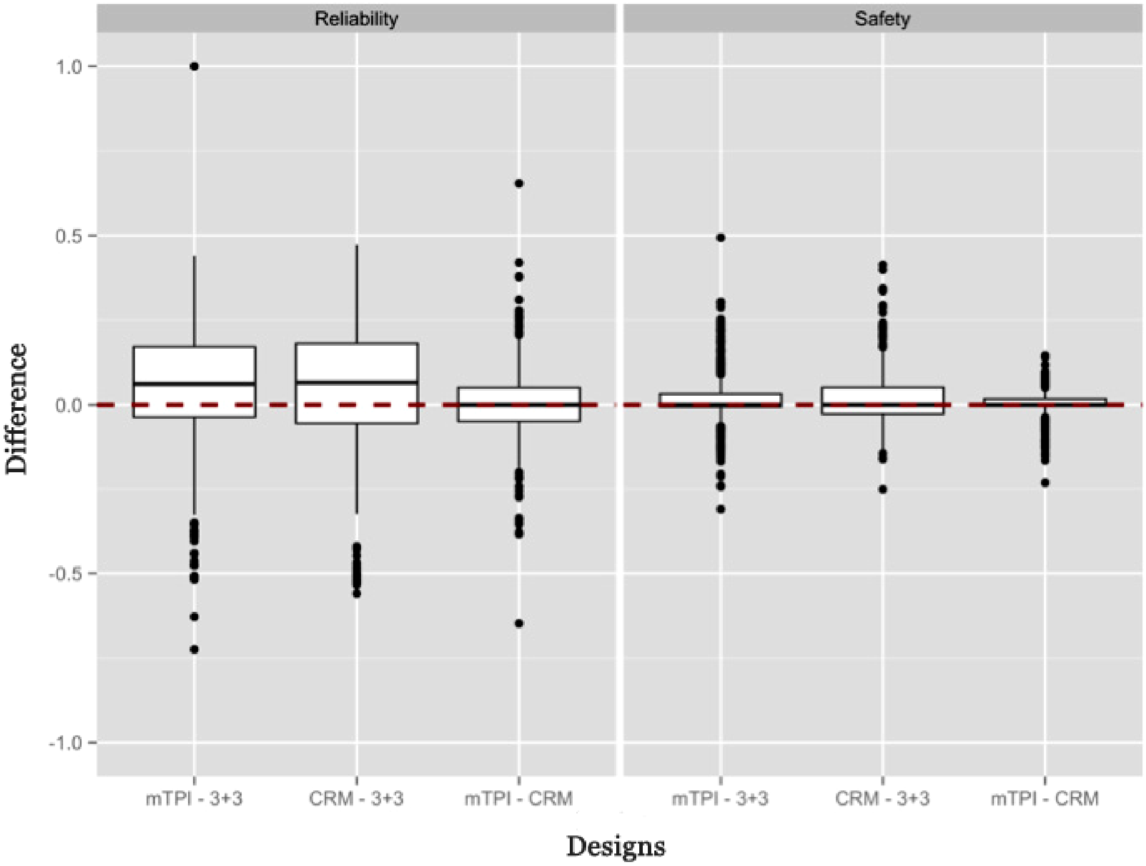


Figure 4: Performance comparison. Boxplots of pair-wise differences in reliability and safety for the three designs, 3+3, CRM, and mTPI. The reliability is calculated as the average percentage of selecting the true MTD across simulated trials for a given scenario. The safety is calculated as the average percentage of patients treated at or below the MTD across simulated trials for a given scenario. Each barplot summarizes the differences of reliability or safety between two designs, in the form of design A minus design B. Therefore, if the value is above zero, the first design of the pair has a better performance. For example, the most left boxplot implies that mTPI is more reliable in finding the true MTD than 3+3.

Table 1: Six covariates summarized for each scenario.

Covariate ID	Covariate Name	Mean	Stand. Dev.
x_1	p_T	0.259	0.059
x_2	dose-sd	0.155	0.075
x_3	min-tox	0.111	0.120
x_4	num-dose	5.086	1.180
x_5	MTD-rank	55.9	33.6
x_6	linearity	22.6	23.0

each scenario we define 1) "p_T" as the true toxicity probability of the true MTD, 2) "dose-sd" as the standard deviation of the vector $\mathbf{p}$, 3) "min-tox" as the minimum, $\min \mathbf{p}$, 4) "num-dose" as d, the number of doses, 5) "MTD-rank" as the position of the true MTD relative to the dose index $\{1, 2, \ldots, d\}$ that takes values 1, 2, 3, and 4, if the true MTD is lower than dose 1, lies in the first half of the doses, lies in the second half of the doses, and higher than dose d, respectively, and 6) "linearity", which we describe below.:

"linearity" measures how linear the dose response curve is when treating $\mathbf{p}$ as the true response rates and dose index $\{1, 2, \ldots, d\}$ as the actual dose level. Since in the NextGene-DF simulation, there is no need to input the actual dose level, we use the dose index as a surrogate. We then fit a linear model

$$logit(p_i) = \mu_p + \beta_p * i + \epsilon,$$

where $logit(p) = \log(p) - \log(1 - p)$ is the logit transformation. We then take the residual sum of square of the regression above and use that as the value for "linearity". The larger the value, the worse the linear fit, and hence the less linear (at the logit scale) the dose response curve.

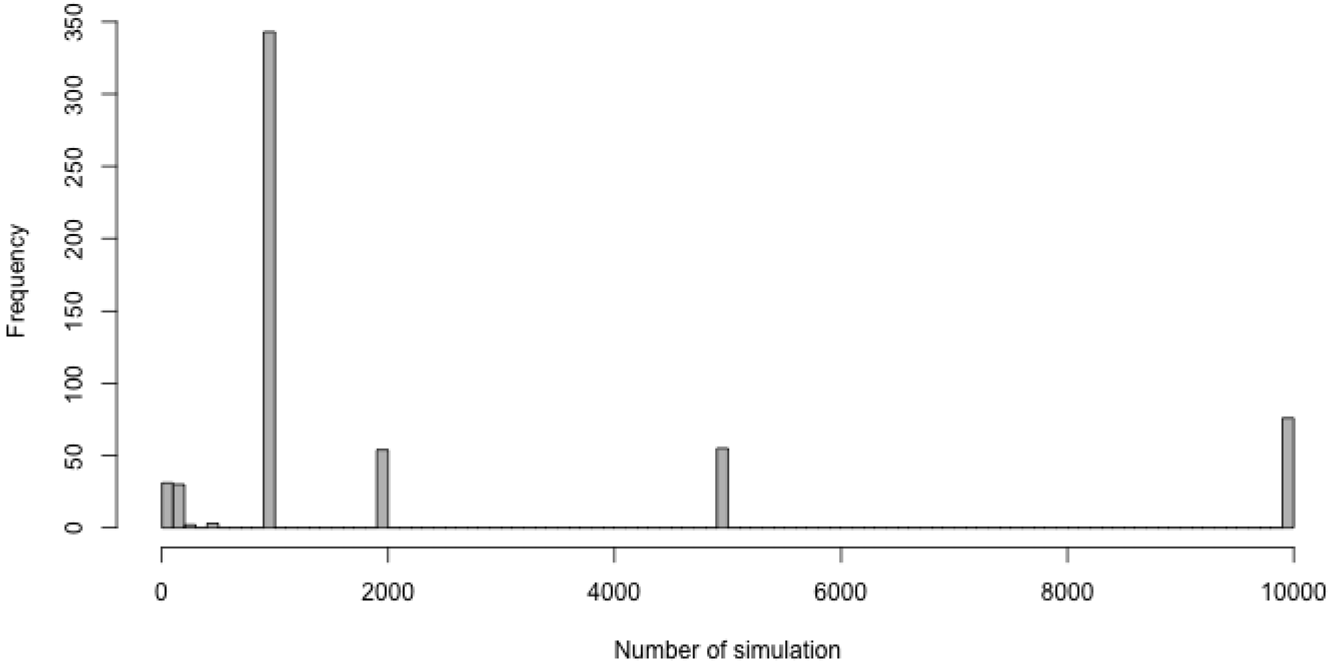


Figure 5: Histogram showing the frequencies of the numbers of simulated trials for all 593 scenarios.

Subgroup Analysis of Scenarios We first perform a clustering analysis for the matrix $[p_{si}]$ where $s = 1, \ldots, 593$ index the scenarios and $i = 1, \ldots, 6$ index the covariates. Figure 6 provides a summary of the subgroup analysis. The rows are scenarios, which are clustered based on the Euclidean distance. The columns are the six covariates and the two response variables, differences diff$-sel$ in reliability and diff$-n$ in safety. Figure 6 shows that there are many row clusters, implying the user-generated scenarios are different. Covariate values are standardized and transformed into $(0, 1)$. Looking at the covariate columns, most covariates display a diverse range. "Linearity" is mostly blue, meaning that most scenarios reflect a linear dose-response curve. "Min-tox" ($\min \mathbf{p}$) is largely blue as well, suggesting that the lowest dose has a small toxicity probability in most scenarios. In the last two columns, we display the two differences between CRM and mTPI in realiability and safety; color red suggests mTPI is better, yellow CRM, and white suggests two are equivalent. Comparing the positions of yellow and red bars in the two columns along with the color display of the six covariates, we suspect that covariates "dose-sd" and "MTD-rank" might be associated with the two outcomes. We then perform a formal regression analysis.

Regression analysis. We perform a linear regression between each response variable, among diff$-n$ and diff$-sel$, and the six covariates. We fit a multivariate model with only main effects,

$$E(y_s) = a + \sum_{j=1}^{7} b_j x_{is},$$

where y_s is the value of either diff$-n$ or diff$-sel$ in scenario s and x_{is} is the value of covariate x_i in scenario s. Varible "MTD-rank" is treated as a categorical variable. We obtain two fitted regression models using the least square estimates. We find that the most significant variable is x_2, "dose-sd", which measures the standard deviation of the vector

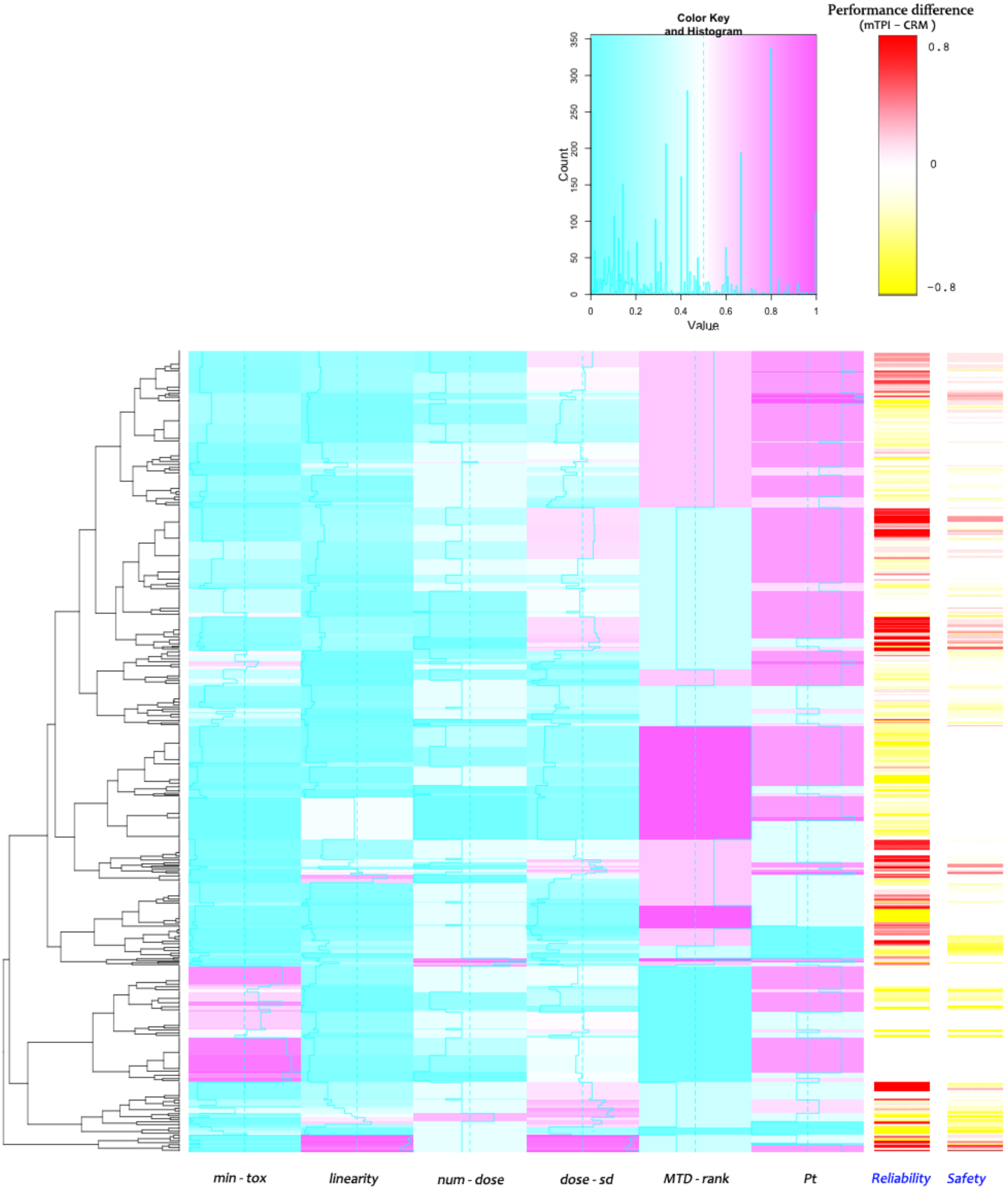


Figure 6: Subgroup Analysis. Differences in reliability and safety between CRM and mTPI (right two columns) are displayed according to the clusters of six covariates in Table 1 for 593 scenarios specified by NextGene-DF users.

$\mathbf{p} = \{p_1, \ldots, p_d\}$. The larger the standard deviation, the safer and more likely to find the true MTD mTPI is compared with CRM. The actually value of p_T, variable x_1 is also significant in both models. A highter p_T value favors CRM over mTPI in reliability (finding the true MTD), and favors mTPI over CRM in safety (treating more patients at doses lower than MTD). "MTD-rank" is significant for safety analysis, favoring mTPI. Linearity, variable x_6 is not significant in both analysis, implying that neither a linear nor nonlinear dose response curve favors mTPI or CRM. The analysis output is given in the Appendix.

5 Discussion

The main contribution of this work is not on the methodology development as all three dose-finding designs, 3+3, CRM, and mTPI, as they have been thoroughly developed and investigated in the scientific literature. Instead, we present NextGen-DF as an integrated web tool consisting of all three designs allowing side-by-side and real time comparison on the internet. In addition, NextGen-DF provides a new customized design based on user input of decision table generated by mTPI. We allow any changes to be made in the customized design, but such a design must be examined through simulations using thousands of simulated trials in the "Simulation" and "Comparison" modules (Module II and III) in our NextGene-DF tool. Customized designs with inconsistent decisions will for sure lose in the comparison and be screened off. However, potentially comparable and more desirable designs could be generated as well. NextGen-DF's ultra-fast speed allows users to potentially modify the customized design repeatedly in practice. Through the web tool we summarize the crowdsourcing results comparing the three trial designs, which provides convincing evidence supporting the inferiority of 3+3 and desirability of mTPI. These results will guide the public in choosing the most efficient methods in practice. We will continue to publish user-generated results as an objective evaluation of the available designs. Within a few months, NextGene-DF has already attracted real-world users from major academic and industrial institutions (see `http://www.compgenome.org/NGDF/`). It is expected that NextGene-DF will generate a consensus from users on the choice of the best designs for phase I dose-finding trials in practice. This achieves a triple goal of improved user experience in practice, better protection for patients, and higher chance of finding the true MTD for drug developers.

When compared to trials for other complications, Hay et al. [4] showed miserable and nearly the worst success rate and likelihood of approval for phase I oncology studies. The authors noted

"... Unfortunately, in oncology, when all indications are considered, only around 1 in 15 drugs entering clinical development in phase 1 achieves FDA approval compared with close to 1 in 8 using the lead indication methodology."

We believe that NextGene-DF will address this deficiency and help improve the success rate of cancer drug development.

Acknowledgement

Yuan Ji's research is partly supported by NIH 2R01 CA132897-06.

Appendix

```
######## R output for the linear regression of diff-sel on x1--x6 ###########

lm(formula = diff-sel ~ x[, "pt"] + x[, "sd"] + x[, "min"] + x[, "dosenumber"] +
    factor(x[, "mtd_rank"]) + x[, "variability"])

Residuals:
     Min       1Q   Median       3Q      Max
-0.58863 -0.05015  0.00154  0.04920  0.50577

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.903e-02  4.113e-02   0.949  0.34293
x[, "pt"]                -2.265e-01  8.621e-02  -2.627  0.00885 **
x[, "sd"]                 6.418e-01  9.053e-02   7.089 3.90e-12 ***
x[, "min"]               -7.842e-02  8.923e-02  -0.879  0.37981
x[, "dosenumber"]        -1.803e-02  3.657e-03  -4.931 1.07e-06 ***
factor(x[, "mtd_rank"])2  4.796e-02  2.435e-02   1.969  0.04938 *
factor(x[, "mtd_rank"])3  4.609e-02  2.877e-02   1.602  0.10975
factor(x[, "mtd_rank"])4 -2.701e-02  3.472e-02  -0.778  0.43678
x[, "linearity"]       -4.688e-05  2.385e-04  -0.197  0.84424
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
######## R output for the linear regression of diff-n on x1--x6 ###########

lm(formula = diff-n ~ x[, "pt"] + x[, "sd"] + x[, "min"] + x[, "dosenumber"] +
    factor(x[, "mtd_rank"]) + x[, "variability"])

Residuals:
      Min        1Q    Median        3Q       Max
-0.198847 -0.013517  0.001551  0.015829  0.124232

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)               -9.009e-02  1.401e-02  -6.428 2.68e-10 ***
x[, "pt"]                  1.882e-01  2.938e-02   6.406 3.08e-10 ***
x[, "sd"]                  2.046e-01  3.085e-02   6.631 7.60e-11 ***
x[, "min"]                 5.053e-02  3.041e-02   1.662   0.0971 .
x[, "dosenumber"]         -6.790e-03  1.246e-03  -5.449 7.47e-08 ***
factor(x[, "mtd_rank"])2  3.346e-02  8.299e-03   4.032 6.26e-05 ***
factor(x[, "mtd_rank"])3  5.035e-02  9.805e-03   5.135 3.84e-07 ***
factor(x[, "mtd_rank"])4  5.608e-02  1.183e-02   4.741 2.68e-06 ***
x[, "linearity"]        3.578e-06  8.127e-05   0.044   0.9649
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

References

[1] BE Storer. Design and analysis of phase i clinical trials. *Biometrics*, 45:925–937, 1989.

[2] A Rogatko, D Schoeneck, W. Jonas, M. Tighiouart, FR. Khuri, and A. Porter. Translation of Innovative Designs Into Phase I Trials. *Journal of Clinical Oncology*, 25:4982–4986, 2007.

[3] C. Le Tourneau, JJ. Lee, and LL. Siu. Dose Escalation Methods in Phase I Cancer Clinical Trials. *Journal of National Cancer Institute*, 101:708–720, 2009.

[4] M Hay, DW Thomas, JL Craighead, C Economides, and J Rosenthal. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

[5] PF Thall and S-J Lee. Practical model-based dose-finding in phase i clinical trials:

Methods based on toxicity. *International Journal of Gynecological Cancer*, 13(3):251–261, 2003.

[6] J O'Quigley, M Pepe, and L Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48, 1990.

[7] Y Ji, Y Li, and B Bekele. Dose-finding in phase i clinical trials based on toxicity probability intervals. *Clinical Trials*, 4:235–244, 2007.

[8] Y Ji, P Liu, Y Li, and B Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):653–663, 2010.

[9] Y Ji and S Wang. 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):1785–1791, 2013.

[10] Karla V Ballman. Phase i trial improvement: A question of patient selection, trial design, or both? *Journal of Clinical Oncology*, 32(6):489–490, 2014.

[11] O Sverdlov, WK Wong, and Y Ryeznik. Adaptive clinical trial designs for phase i cancer studies. *Statistics Survey*, pages 2–44, 2014.

[12] F Didier, MM Lisa, DJP Andrew, G Lia, GD Steven, A Isabelle, I Robert, G Ryan, VS Arne, W Ruixue, and G Birgit. A phase i study of the anti-insulin like growth factor type 1 receptor (igf-1r) antibody dalotuzumab in pediatric patients with advanced solid tumors. *Journal of Clinical Oncology*, 31:suppl; abstr 10026, 2013.

[13] TA Yap, L Yan, A Patnaik, I Fearen, D Olmos, K Papadopoulos, R D Baird, L Delgado, A Taylor, L Lupinacci, et al. First-in-man clinical trial of the oral pan-akt inhibitor mk-

2206 in patients with advanced solid tumors. *Journal of Clinical Oncology*, 29(35):4688–4695, 2011.

[14] ADJ Pearson, SM Federico, I Aerts, DR Hargrave, SG DuBois, R Iannone, R Geschwindt, R Wang, TM Trippett, and B Geoerger. A phase i study of ridaforolimus (mk-8669) in pediatric patients with advanced solid tumors. *Journal of Clinical Oncology*, 31(15), 2013.

[15] SK Fan, Y Lu, and Y Wang. A simple bayesian decision-theoretic design for dose-finding trials. *Statistics In Medicine*, 31(28):3719–3730, 2012.

[16] AP Mander and MJ Sweeting. A product of independent beta probabilities dose escalation design for dual-agent phase i trials. *Statistics in medicine*, 34(8):1261–1276, 2015.

[17] SN Goodman, ML Zahurak, and S Piantadosi. Some practical improvements in the continual reassessment method for phase i studies. *Statistics in medicine*, 14(11):1149–1161, 1995.

[18] S Piantadosi, JD Fisher, and S Grossman. Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer chemotherapy and pharmacology*, 41(6):429–436, 1998.

[19] T. Robertson, F. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. Wiley, New York, 1998.

[20] K Cheung. *dfcrm: Dose-finding by the continual reassessment method*, 2013. R package version 0.2-2.

[21] Shing M Lee and Ying Kuen Cheung. Model calibration in the continual reassessment method. *Clinical Trials*, 6(3):227–238, 2009.

[22] M Krzywinski, J Schein, I Birol, J Connors, R Gascoyne, D Horsman, SJ Jones, and MA Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[23] AR Hansen, DM Graham, GR Pond, and LL. Siu. Phase 1 trial design: is 3 + 3 the best? *Cancer Control*, 21:200–208, 2014.