Safety Concerns of the 3+3 Design: A Comparison to the mTPI Design

Yuan Ji and Sue-Jane Wang

Abstract The 3+3 design is the most common choice by clinicians for phase I dose-escalation oncology trials. In recent reviews, more than 90 % of phase I trials are based on the 3+3 design (Rogatko et al., Journal of Clinical Oncology 25:4982–4986, 2007). The simplicity and transparency of 3+3 allows clinicians to conduct dose escalations in practice with virtually no logistic cost, and trial protocols based on 3+3 pass IRB and biostatistics reviews briskly. However, the performance of 3+3 has never been compared to model-based designs under simulation studies with matched sample sizes. In the vast majority of statistical literature, 3+3 has been shown to be inferior in identifying the true MTD although the sample size required by 3+3 is often magnitude smaller than model-based designs. In this paper, through comparative simulation studies with matched sample sizes, we demonstrate that the 3+3 design has higher risks of exposing patients to toxic doses above the MTD than the mTPI design (Ji et al., Clinical Trials 7:653–663, 2010), a newly developed adaptive method. In addition, compared to mTPI, 3+3 does not provide higher probabilities in identifying the correct MTD even when the sample size is matched. Given the fact that the mTPI design is equally transparent, simple and costless to implement with free software, and more flexible in practical situations, we highly encourage more adoptions of the mTPI design in early dose-escalation studies whenever the 3+3 design is also considered. We provide a free software to allow direct comparisons of the 3+3 design to other model-based designs in simulation studies with matched sample sizes.

Y. Ji (✉)
Center for Clinical and Research Informatics, NorthShore, University HealthSystem, 1001 University Place, Evanston, IL 60201, USA
e-mail: yji@northshore.org

S.-J. Wang
Office of Biostatistics/Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA
e-mail: suejane.wang@fda.hhs.gov

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_11,

1 Introduction

Phase I oncology trials aim to find the maximum tolerated dose (MTD), the highest dose with toxicity rate close to a pre-specified target level, p_T. The 3+3 design [3, 4] is the leading method for phase I dose-escalation trials in oncology, as over 90 % of published phase I trials have been based on 3+3 for the past two decades [1, 5, 6]. Such popularity of 3+3 is striking since numerous model-based dose-escalation methods have been developed by biostatisticians during the same time period and almost all the new methods seemed to exhibit better performance than 3+3 [7–10].

The main reason for the popularity of the 3+3 design is due to its simplicity, transparency, and the costless implementation in practice. In contrast, it often requires a considerable amount of logistic support and complexity to implement most model-based designs. Even if the practical burden could be overcome, protocols based on model-based designs are often subject to more thorough reviews by IRB or among biostatisticians, as operating characteristics of these new designs are required. To the contrary, if the protocol is based on the 3+3 design, such requirement disappears since 3+3 has been widely used. As a result, despite the acceleration in the research development of adaptive model-based designs, the lower standard in the review process and cost-free implementation in practice makes 3+3 an increasingly popular design to physicians. Setting aside the logistic issues, we ask exactly how much better the model-based designs are than 3+3. In reviewing the statistical literature on phase I adaptive designs, we found that when comparing to 3+3, most works did not match the sample size across the designs. For example Ji et al. (2010) [2] showed that 3+3 exhibits a smaller average sample size in the computer simulations than model-based designs, and consequently 3+3 also yields a smaller percentage in identifying the true MTD in these simulations. Since the sample size is not matched in the comparison, it is difficult to assess the reason for the reduced percentage under 3+3. More importantly, since phase I trials focus on patient safety, comparisons without matching sample size cannot provide accurate assessment on the safety characteristics of designs. In fact, usually designs resulting in larger sample sizes should be safer since patients enrolled in the later stage of the trial with a larger sample size will be better protected due to more precise statistical inference.

In this paper, we construct a comprehensive simulation study to evaluate the operating characteristics of 3+3 and a newly developed adaptive design known as the modified toxicity probability interval (mTPI) method [2, 6]. In doing so we match the sample size between the two designs. The main intent of choosing the mTPI design for comparison is because mTPI is equally simple, transparent, and costless to implement. In other words, the logistic burden of mTPI and 3+3 is comparable, which allows us to focus on the simulation performance. Albeit being recently introduced to the society, mTPI has already received attention from both research and industry entities [11, 12]. For example, through personal communication we are informed that almost all phase I oncology trials conducted

at Merck Co., Inc. in the past 2 years have been based on the mTPI design or its variations. Recently, phase I trials based on the mTPI design has been published [13, 14]. Considering the short time period since the publication of the mTPI design, this popularity is encouraging.

In a nutshell, the 3+3 design consists of a set of deterministic rules that dictate dose-escalation decisions based on observed patient outcomes. For example, if out of three treated patients 0, 1, or more than 1 toxicities are observed, 3+3 will recommend escalating dose level, continuing at the same dose level, or de-escalating dose level, respectively (see, e.g., [15, 16]).

The mTPI design uses a Bayesian statistics framework and a beta/binomial hierarchical model to compute the posterior probability of three intervals that reflect the relative distance between the toxicity rate of each dose level to the target rate p_T. Let p_d denote the probability of toxicity for dose d, $d = 1, \ldots, D$, where D is the total number of candidate doses. Using the posterior samples for p_d, mTPI computes the unit probability mass, defined as

$$\text{UPM}_{(a,b)}(d) = \frac{Pr\{p_d \in (a,b) \mid data\}}{b-a}, \tag{1}$$

for three intervals corresponding to *under*-, *proper*-, and *over*-dosing, in reference to whether a dose is lower, close to, or higher than the MTD, respectively. Specifically, the under-dosing interval is defined as $(0, p_T - \epsilon_1)$ and implies that the dose level is lower than the MTD, the over-dosing interval $(p_T + \epsilon_2, 1)$ implies that the dose level is higher than the MTD, and the proper-dosing interval $(p_T - \epsilon_1, p_T + \epsilon_2)$ suggests that the dose level is close to the MTD. Here ϵ_1 and ϵ_2 are small fractions, say 0.05. Inference is robust with respect to the choice of ϵ, as shown in [2]. Large UPM values for each interval imply large per-unit posterior probability mass for that interval, therefore implying the corresponding decision: if $\text{UPM}(d)$ is the largest for under-, proper-, or over-dosing interval, the decision should be to escalate (E), stay (S) at dose d, or de-escalate (D), respectively. Therefore, assuming that dose d is currently used to treat patients, the mTPI design assigns the next cohort of patients based on the decision rule $\mathbf{B}_d$, given by

$$\mathbf{B}_d = \arg \max_{m \in \{D,S,E\}} \text{UPM}(m,d), \tag{2}$$

where $\text{UPM}(m,d)$ is the value of UPM for the dosing interval associated with decision m. Decisions D, S, or E warrant the use of dose $(d-1)$, d, or $(d+1)$ for the next cohort of patients, respectively. Ji et al. [2] proved that the decision rule $\mathbf{B}_d$ is consistent and optimal in that it minimizes the posterior expected loss, in which the loss function is determined to achieve equal prior expected loss for the three decisions, D, S, and E. More importantly, all the dose-escalation decisions for a given trial can be pre-calculated under the mTPI design and presented in a two-way table (Fig. 1). Once the trial starts, clinicians can easily monitor the trial and

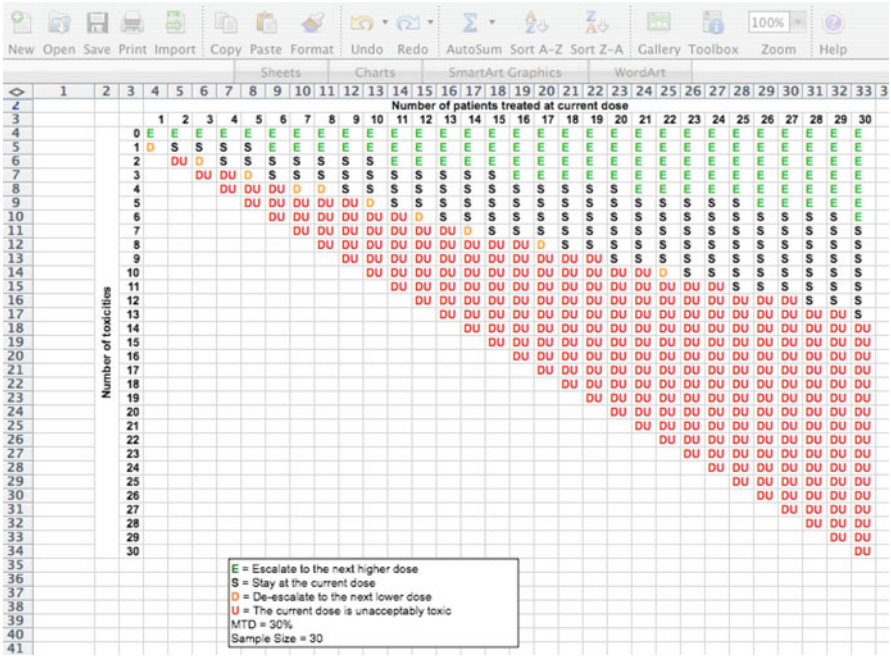

Fig. 1 Dose-finding spreadsheet of the mTPI method. The spreadsheet is generated based on a Beta/Binomial model and pre-calculated before a trial starts. The letters in different colors are computed based on the decision rules under the mTPI method and represent different dose-finding actions. In addition to actions *D*, *S*, and *E*, the table includes action *U*, which is defined as the execution of the *dose exclusion rule* in mTPI.

select the appropriate doses following the pre-calculated table. The simplicity and transparency of mTPI makes it a strong candidate as a model-based counterpart of the 3+3 design in practice. A software in Excel is provided at https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=72 We will show surprising and important findings and make a recommendation to use mTPI in future phase I trials based on these findings.

2 Comparison of 3+3 and mTPI

2.1 Simulation Setup

We perform computer simulation of phase I trials based on the 3+3 and mTPI designs and compare their operating characteristics summarized over thousands of simulated trials.

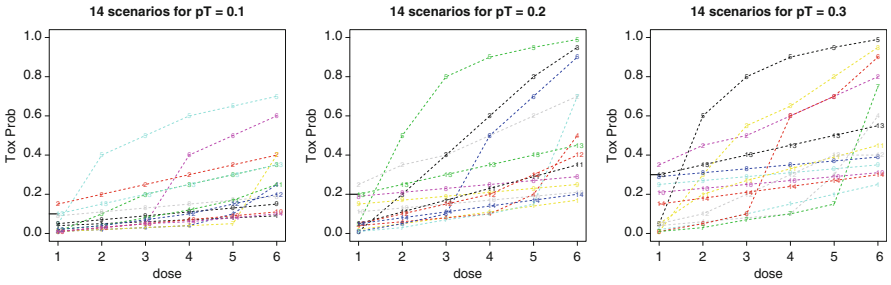

Fig. 2 Dose-response patterns for the 42 clinical scenarios in the simulation. For each of the $p_T = 0.1, 0.2, 0.3$ values, 14 scenarios are constructed.

2.1.1 Clinical Scenarios

We consider 6 doses in the simulated trials. We construct 14 scenarios for each of the three target p_T values, resulting in a total of 42 scenarios. In each scenario, true toxicity probabilities are specified for the 6 doses. These scenarios are set up to capture a wide range of dose–response shapes in practice, as shown in Fig. 2 (see also a discussion in Ji et al., 2012 [17]). Specifically, Scenario 1 represents a case where all doses are safe and low; Scenario 2 represents a case where all doses are high; in Scenarios 3–4 doses cover a wide range of toxicity probabilities and the toxicity probability of one dose equals p_T; Scenarios 5–7 also cover a wide range of toxicity probabilities but the MTD is bracketed by two adjacent doses; In Scenario 8–10, dose toxicity probabilities do not vary much and center around the target p_T; Scenarios 11–12 are similar to Scenarios 8–10, except doses have a wider range of toxicity; lastly, Scenarios 13–14 represent two rare cases in which the MTD is the lowest and highest dose, respectively.

2.1.2 Values of p_T

In practice, the target p_T values are rarely larger than 30 % as it implies unnecessary exposure of patients to doses with high toxicity. Below, we make three choices of p_T: 0.1, 0.2, and 0.3, i.e., the target toxicity rates of the MTD in our simulated trials are 10 %, 20 %, or 30 %. For each p_T and each scenario, we simulate 2,000 trials.

2.1.3 Matching Sample Size

A unique feature in our comparison is that we attempt to match the average sample size of the 3+3 and mTPI designs for each of the clinical scenarios used in the simulation study. To achieve this, for each scenario we first apply the 3+3 design to 2,000 simulated trials and obtain the mean of the 2,000 sample sizes. We then apply

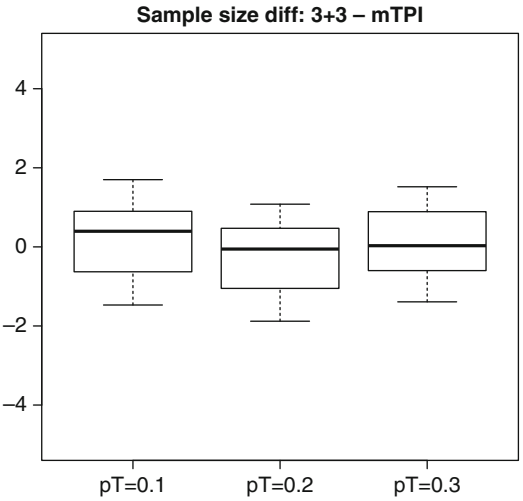

Fig. 3 Difference in the average sample size per trial between 3+3 and mTPI. Each boxplot summarizes the differences for 14 scenarios for a given target toxicity p_T value.

the mTPI design, in which we need to specify the maximum sample size. The mTPI design stops the trial when the total number of patients enrolled is equal or larger than the maximum sample size. We calibrate the maximum sample sizes of mTPI for each p_T value and each scenario, so that the average sample sizes over simulated trials under both designs are similar across all the scenarios. Figure 3 shows the differences of the average sample sizes (over 2,000 simulated trials) between 3+3 and mTPI. The two designs exhibit comparable sample sizes overall. Our calibration of mTPI only involves varying the maximum sample size, while keeping all the other design features unchanged.

2.1.4 Variations of the 3+3

To account for different target p_T values, we use one of the two 3+3 variations (3+3^L and 3++3^H). See Fig. 4. Briefly, the two designs only differ when 6 patients have been treated at a dose, and 1 or 2 of them experience the toxicity. In one variation, 3+3^L, we would stop the trial and declare that the MTD has been exceeded if 2 out of 6 patients experienced toxicity at the dose; in the other variation, called 3+3^H, we would stop the trial and declare that the MTD is that dose. Likewise, 3+3^H would escalate if 1 toxicity is observed from 6 patients, while 3+3^L would stop and declare the dose to be the MTD. Here, ^{L} or ^{H} means that the target toxicity rate p_T of the MTD is low or high. We use 3+3^L for trials with $p_T = 0.1$ or $p_T = 0.2$, and the 3+3^H for trials with $p_T = 0.3$.

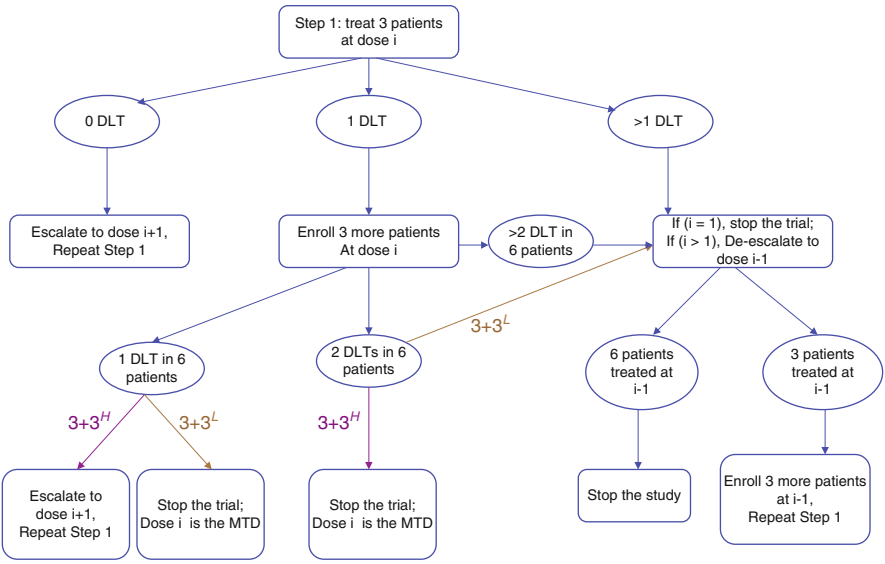

Fig. 4 Schema of the enhanced 3+3 design. The two versions of $3+3^L$ and $3+3^H$ represent the cases where the MTD is defined as the highest dose on which no more than 1 and 2 dose-limiting toxicities (DLT) are observed from 6 patients, respectively.

2.2 *Performance Evaluation*

Summarizing results from 42 scenarios over three different p_T values for three designs can be subjective depending on the criterion used in the comparison. Since the average sample sizes between the two methods are roughly matched, we focus our comparison on two summary statistics simultaneously,

$$n_{>MTD} = \text{ the number of patients treated above the true MTD}$$
$$\%Sel_{MTD} = \text{ the percentage of selecting the true MTD.}$$

$n_{>MTD}$ directly evaluates the safety of each design since under matched sample size; a smaller $n_{>MTD}$ value implies fewer toxicities. To calculate $\%Sel_{MTD}$, we need to decide which doses will be considered as the MTD for each scenario.

2.3 *Main Results*

Figure 5 summarizes the comparison between the 3+3 and mTPI designs, regarding the differences in $n_{>MTD}$ and $\%Sel_{MTD}$. We present the comparison results of $n_{>MTD}$ in the left panel. Comparing to the mTPI design, the 3+3 design has lower

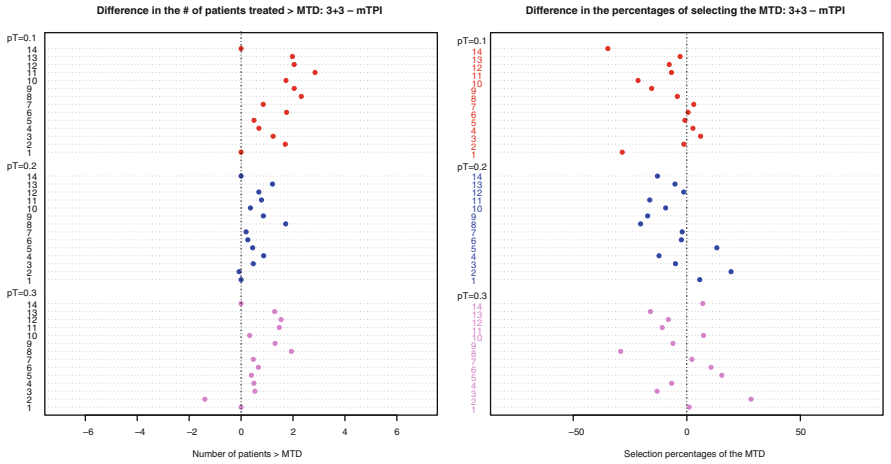

Fig. 5 Comparison between 3+3 and mTPI based on matched sample sizes. The left panel presents the differences in the numbers of patients treated at doses above the MTD ($n_{>MTD}$), i.e., values of ($n_{>MTD3+3}$ - $n_{>MTDmTPI}$) for all 42 scenarios. The right panel presents the differences in the selection percentages of the true MTD ($\%Sel_{MTD}$), i.e., values of ($\%Sel_{MTD3+3}$ - $\%Sel_{MTDmTPI}$) for all 42 scenarios. The three colors in the plots represent the results corresponding to the three different p_T values.

$n_{>MTD}$ values for two scenarios, higher $n_{>MTD}$ for 34 scenarios, and the same $n_{>MTD}$ for six scenarios. In words, 40 out of 42 times, mTPI treats fewer or the same number of patients at doses higher than the MTD than 3+3. In addition, Fig. 6 examines the *overall toxicity percentage*, defined as

$$\frac{\text{the total number of toxicities over all simulated trials}}{\text{the total number of patients treated over all simulated trials}} \times 100\%.$$

Only in one out of 42 scenarios, the 3+3 design exhibits a lower overall toxicity percentage than the mTPI design.

We direct attention to the right panel of Fig. 5 which compares $\%Sel_{MTD}$ between the two designs. In 10 out of 42 scenarios, 3+3 has a higher selection percentage of the true MTD than mTPI. Among these scenarios, the 3+3 design selects the MTD up to about 25 % more often than the mTPI design (Scenario 2 for $p_T = 0.3$). In the remaining 32 scenarios, mTPI selects the MTD more often than 3+3, up to more than 40 % (Scenario 14 for $p_T = 0.1$). A closer examination reveals that 3+3 has higher $\%Sel_{MTD}$ values in scenarios when none of the doses has a toxicity probability close to p_T or when the MTD is at the lower or higher end of the dosing set. We performed additional simulations and confirmed this finding. We found that when the MTD is out of the range of the dosing set, 3+3 usually has a higher selection percentage than mTPI. In other words, 3+3 is a better method when none of the investigational doses is close to the true MTD. This advantage seems to be of limited utility in practice since usually doses are chosen based on scientific and historical data, anticipating some of them are close to the MTD, not the opposite.

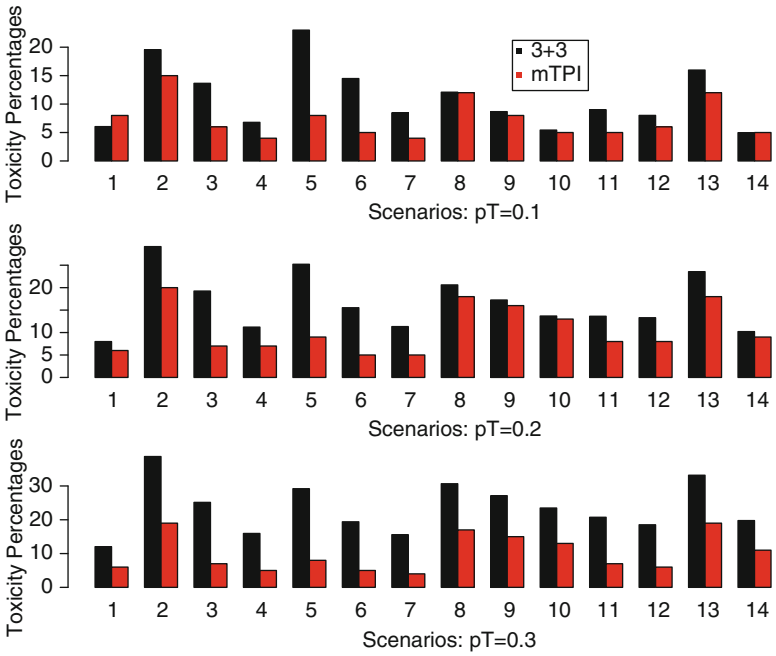

Fig. 6 Overall toxicity percentages for the 3+3 and mTPI designs across all the simulated trials.

Summarizing the two plots in Fig. 5 and considering that (1) the overall sample sizes between the two designs are roughly matched for all the scenarios and (2) the 42 scenarios are constructed to cover a wide range of practical dose–response shapes, *we conclude that the 3+3 design is more likely to treat patients at toxic doses above the MTD and less likely to identify the true MTD than the mTPI design.*

3 Conclusion and Discussion

The mTPI has all the attractive properties 3+3 enjoys for practical considerations and implementations. In addition, compared to the 3+3 design, the mTPI design is safer in treating fewer patients at doses above the MTD, and in general yielding higher probabilities in identifying the true MTD.

In practice, a single value n must be provided as the maximum sample size for the mTPI design in any dose escalation study. In implementing the mTPI design, we recommend a sample size of $n = k \times (d + 1)$ to ensure that the design will reach the highest dose if needed and still has one more cohort to use. Here k is the cohort size and d the number of doses.

It is commonly accepted that phase I trials are of small sizes. This mythology is poorly addressed in the literature. Small phase I trials often provide wrong

recommended doses for phase II, resulting in either low efficacy or high toxicity if the recommended doses are too low or too high, respectively. More discussion and investigation on the proper sample sizes of phase I trials are needed. For example, a streamlined and seamless phase I/II design may result in higher power in the identification of safe and effective doses [18] due to increased sample sizes from the seamless features.

We note that comparison between CRM and 3+3 have been investigated by various authors [19–21] and thus is not included in this paper. A downside of CRM is the lack of easy ways for implementation in practice. We have included the CRM design in our software so that interested users can examine all three designs together, 3+3, CRM, and mTPI.

References

[1] 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.

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

[3] B.E. Storer. An evaluation of phase I clinical trials designs in the continuous dose-response setting. *Statistics in Medicine*, 48:2399–2408, 2001.

[4] B.E. Storer. Design and analysis of phase I clinical trials. *Biometrics*, 45:925–937, 1989.

[5] 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.

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

[7] J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: A practical design for phase i clinical trials in cancer. *Biometrics*, 46:33–48, 1990.

[8] SM Berry, BP Carlin, JJ Lee, and P Müller. *Bayesian Adaptive Methods for CLinical Trials*. CRC, Boca Raton, FL, 2011.

[9] YK Cheung. *Dose Finding by the Continual Reassessment Method*. CRC, Boca Raton, FL, 2011.

[10] B. Neuenschwander, M. Branson, and T. Gsponer. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27:2420–2439, 2008.

[11] M.S. Blanchard and J.A. Longmate. Toxicity equivalence range design (TEQR): A practical Phase I design. *Comtemporary Clinical Trials*, 32:114–121, 2011.

[12] G.J. Hather and H. Mackey. Some Notable Properties of the Standard Oncology Phase I Design. *Journal of Biopharmaceutical Statistics*, 19:543–555, 2009.

[13] M. Fanale, L. Fayad, B. Pro, F. Samaniego, M. Liboon, C. Nunez, S. Horowitz, P. Anderlini, U. Popat, Y. Ji, LW. Kwak, and A. Younes. Phase I study of bortezomib plus ICE (BICE) for the treatment of relapsed/refractory Hodgkin lymphoma. *British Journal of Haematology*, 154:284–286, 2011.

[14] TA Yap, L Yan, A Patnaik, I Fearen, D Olmos, K Papadopoulos, RD Baird, L Delgado, A Taylor, L Lupinacci, R Riisnaes, LL Pope, SP Heaton, G Thomas, MD Garrett, DM Sullivan, JS de Bono, and AW Tolcher. First-in-Man Clinical Trial of the Oral Pan-AKT Inhibitor MK-2206 in Patients With Advanced Solid Tumors. *Journal of Clinical Oncology*, pages 4688–4695, 2011.

[15] NK Ibrahim, N Desai, and S et al Legha. Phaes I and pharmacokinetic study of ABI-007, a Cremophor-free, protein-stablized, nanoparticle formulation of paclitaxel. *Clinical Cancer Research*, 7(5):1038–1044, 2002.

[16] D Strumberg, H Richly, and RA et al Hilger. Phase I clinical and pharmacokinetic study of the novel Raf kinase and vascular endothelial growth factor receptor inhibitor BAY 43-9006 in patients with advanced refractor solid tumors. *Journal of Clinical Oncology*, 23(5):965–972, 2005.

[17] Y Ji, L Feng, EJ Shpall, P Kebriaei, R Champlin, D Berry, and L Cooper. Bayesian Continual Reassessment Method for Dose-Finding Trials Infusing T Cells with Limited Sample Size. *Journal of Biopharmaceutical Statistics*, 22:1206–1219, 2012.

[18] F Xie, Y Ji, and L Tremmel. A Bayesian adaptive design for multi-dose,randomized,placebo-controlled phase I/II trials. *Contemporary Clinical Trials*, 33:739–748, 2012.

[19] J O'Quigley. Another Look at Two Phase I Clinical Trial Designs. *Statistics in Medicine*, 18:2683–2690, 1999.

[20] E.L. Korn, D. Midthune, T.T. Chen, L.V. Rubinstein, M.C. Christian, and R. Simon. A comparison of two phase I designs. *Statistics in Medicine*, 13:1799–1806, 1994.

[21] S.N. Goodman, M.L. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*, 14:1149–1161, 1995.