Published in final edited form as:

Biometrics. 2014 June ; 70(2): 389–397. doi:10.1111/biom.12158.

Simple Benchmark for Complex Dose Finding Studies

Ying Kuen Cheung

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A

Ying Kuen Cheung: yc632@columbia.edu

Summary

While a general goal of early phase clinical studies is to identify an acceptable dose for further investigation, modern dose finding studies and designs are highly specific to individual clinical settings. In addition, as outcome-adaptive dose finding methods often involve complex algorithms, it is crucial to have diagnostic tools to evaluate the plausibility of a method's simulated performance and the adequacy of the algorithm. In this article, we propose a simple technique that provides an upper limit, or a benchmark, of accuracy for dose finding methods for a given design objective. The proposed benchmark is nonparametric optimal in the sense of O'Quigley, Paoletti, and Maccario (2002), and is demonstrated by examples to be a practical accuracy upper bound for model-based dose finding methods. We illustrate the implementation of the technique in the context of phase I trials that consider multiple toxicities and phase I/II trials where dosing decisions are based on both toxicity and efficacy, and apply the benchmark to several clinical examples considered in the literature. By comparing the operating characteristics of a dose finding method to that of the benchmark, we can form quick initial assessments of whether the method is adequately calibrated and evaluate its sensitivity to the dose-outcome relationships.

Keywords

Combination therapy; Efficacy-toxicity tradeoff; Multiple toxicities; Phase I trials; Phase I/II trials; Utility scores

1. Introduction

The general goal of an early phase clinical trial is the determination of an acceptable dose, often defined in terms of toxicity and clinical response, for further investigation. In a standard phase I trial setting, the objective is to identify the maximum tolerated dose, defined as a dose associated with a pre-specified toxicity probability. There are numerous dose finding approaches for the standard setting, including model-based methods (O'Quigley, Pepe, and Fisher, 1990; Liu, Rosenberger, and Haines, 2006), random walk designs (Durham, Flournoy, and Rosenberger, 1997), stepwise procedures (Lin and Shih, 2001; Cheung, 2007) and the stochastic approximation (Cheung, 2010). Meanwhile, the

Supplementary Materials

The R code that implements Algorithm 1 for the examples in Sections 3–5 is available with this paper at the *Biometrics* website on Wiley Online Library.

complexity of outcome-adaptive dose finding designs has growth rapidly to accommodate the “nonstandard” phase I trial settings and phase I/II trials. A short list in the fast-growing literature includes the use of ordinal or multiple toxicities (Bekele and Thall, 2004; Paul, Rosenberger, and Flournoy, 2004; Yuan, Chappell, and Bailey, 2007; Lee, Cheng, and Cheung, 2011), bivariate efficacy and toxicity outcomes (O’Quigley, Hughes, and Fenton, 2001; Braun, 2002; Thall and Cook, 2004; Dragalin and Fedorov, 2006; Yin, Li, and Ji, 2006), and utility scores in combination therapy (Houede et al., 2010). These dose finding designs are typically computer intensive and operate on very elaborate dose-outcome relationships. Even as such complexity is warranted in order to make informed dose decisions in these nonstandard settings, it also complicates the evaluation of designs. While computer simulation is a routine to evaluate the operating characteristics of a dose finding method, it would be desirable to ascertain the statistical efficiency of the method in relation to some optimal benchmark. O’Quigley, Paoletti, and Maccario (OPM, 2002) introduce a nonparametric optimal design for the estimation of the maximum tolerated dose in a standard phase I trial setting, and apply the design as a benchmark for phase I dose finding methods. Cheung (2013) further exploits the OPM design to derive the sample size formulae for the continual reassessment method. In this article, we propose a general approach to generate nonparametric optimal benchmark for dose finding designs for a variety of early phase trials.

The proposed approach is a simulation-based method, and is described in Section 2. In Section 3, we discuss benchmark for outcome-adaptive designs incorporating multiple toxicities, and illustrate its application in the context of a bortezomib trial in lymphoma patients (Lee et al., 2012). Sections 4 and 5 consider benchmark designs for phase I/II studies that use both toxicity and efficacy endpoints for dosing decisions: In Section 4, we develop benchmark for the method of Thall and Cook (2004) who aim to maximize desirability in a stroke trial of thrombolytic agent; in Section 5, we illustrate the benchmark for maximizing the expected value of a utility score in combination therapy for bladder cancer patients (Houede et al., 2010). Section 6 ends this article with a discussion on the utility of benchmark designs. Technical details are provided in an Appendix.

2. The Basic Approach

In a typical dose finding study, each patient is treated at a dose chosen from a discrete set of levels $\{1, \dots, K\}$. Let $Y = Y(k)$ denote the patient’s multinomial outcome at dose k with distribution characterized by $\pi(k)$. Precisely, suppose that Y takes on $L + 1$ possible values $\omega_0 < \omega_1 < \dots < \omega_L$. We define

$$\pi_l(k) = \Pr\{Y(k) \geq \omega_l\} \text{ for } l=1, \dots, L \quad (1)$$

so that $\pi(k) = \{\pi_1(k), \dots, \pi_L(k)\}$ completely specifies the outcome distribution of dose k . Note that $\Pr\{Y(k) = \omega_l\} = \pi_l(k) - \pi_{l+1}(k)$ with the convention $\pi_0(k) \equiv 1$ and $\pi_{L+1}(k) \equiv 0$. Further let $\pi = \{\pi(1), \dots, \pi(K)\}$ denote the entire collection of the dose-outcome probabilities. Generally, a dose finding trial aims to estimate a dose $d(\pi) \in \{1, \dots, K\}$, where $d(\cdot)$ is a known design objective function determined in the planning stage of a trial.

For example, in a standard phase I trial, we deal with binary toxicity outcome with $L = 1$ and $Y \in \{0, 1\}$, so that $\pi_1(k) = \Pr\{Y(k) = 1\} = \Pr\{Y(k) = 1\}$ is the toxicity probability at dose k , and the design objective is

$$d_1(\pi) = \arg \min_k |\pi_1(k) - p| \quad (2)$$

for some pre-specified p .

In computer simulation, we may draw an outcome Y from a distribution π via latent variables. Precisely, for patient i , we first randomly draw a tolerance profile

$$U_{i1}, U_{i2}, \dots, U_{iL} \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1) \quad (3)$$

Then we generate, for a given k ,

$$Y_i(k) = \omega_l \quad (4)$$

if

$$U_{i,l+1} > \frac{\pi_{l+1}(k)}{\pi_l(k)} \text{ and } U_{ij} \leq \frac{\pi_j(k)}{\pi_{j-1}(k)} \text{ for all } j=1, \dots, l.$$

Because $\pi_{L+1}(k) \equiv 0$, we can arbitrarily set $U_{i,L+1} = 1$ without loss of generality. The variate $Y_i(k)$ generated by (3) and (4) will represent a simulated outcome for a patient treated at dose k in a dose finding study. In a real dose finding study, since each patient is given a dose, we only observe the outcome at that dose. In other words, we make dosing decisions and inference based on the patients' partial outcome profiles in a real study. However, in a computer simulation experiment where we know the latent variables (3), it is possible to "observe" the outcomes of the same patient at all K dose levels, i.e., a complete outcome profile $\{Y_i(1), \dots, Y_i(K)\}$, by applying (4) to each $k = 1, \dots, K$.

Proposition 1 (Monotonicity)—Suppose that, for any given $k'' > k'$,

$$\frac{\pi_{l+1}(k'')}{\pi_l(k'')} \geq \frac{\pi_{l+1}(k')}{\pi_l(k')} \text{ for } l=0, \dots, L. \quad (5)$$

Then $Y_i(k'') \geq Y_i(k')$ for any outcomes generated by (3) and (4).

In the special case of binary toxicity outcome with $L = 1$, assumption (5) reduces to monotonicity of the dose-toxicity curve, that is, $\pi_1(k'') \geq \pi_1(k')$ for $k'' > k'$. Proposition 1 then states that a toxicity at a lower dose implies toxicity at a higher dose. This assumption is the same as that in O'Quigley et al. (2002). For ordinal outcome with $L > 1$, there is no unique way to represent an increasing dose-toxicity relationship. Specifically, assumption

(5) corresponds to monotone conditional log odds; see Fleiss, Levin, and Paik (2003) who also prescribe alternative ways to model ordinal multinomial response. Importantly, monotonicity is not crucial to the validity of our proposed procedure described below; and for practical purposes, it may not be reasonable to assume monotonicity in many situations such as when dealing with bivariate outcomes (Section 4) and combination therapy (Section 5).

Proposition 2—The outcome $Y_i(k)$ generated by (3) and (4) distributes according to tail distribution π described by (1), for any given k and l .

As a consequence of Proposition 2, we can estimate π with the corresponding sample proportion based on the complete outcome profiles of n simulated patients, where n is the sample size of a study: for each k , calculate

$$\hat{\pi}_l(k) = \frac{1}{n} \sum_{i=1}^n I\{Y_i(k) \geq w_l\} = \frac{1}{n} \sum_{i=1}^n I \left[\bigcap_{j=1}^l \left\{ U_{ij} \leq \frac{\pi_j(k)}{\pi_{j-1}(k)} \right\} \right] \quad (6)$$

for $l = 1, \dots, L$, where $I[A]$ is the indicator function of the event A .

Proposition 3— $\hat{\pi}_l(k)$ is an unbiased estimator for $\pi_l(k)$ with variance achieving the Cramer-Rao lower bound without additional parametric assumptions, that is,

$$\text{var}\{\hat{\pi}_l(k)\} = \frac{\pi_l(k)\{1-\pi_l(k)\}}{n} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (7)$$

Proposition 3 implies $\hat{d}(\pi)$ is a consistent estimator for $d(\pi)$. In addition, because $\hat{d}(\pi)$ estimates $d(\pi)$ using complete outcome profiles of the simulated patients, it is expected that its performance will be superior to any dose finding designs that uses only partial outcome profiles. In the standard phase I trial setting where Y is binary, the benchmark $\hat{d}(\pi)$ reduces to the nonparametric optimal OPM design under the same monotonicity assumption as in Proposition 1.

We note that $\hat{d}(\pi)$ cannot be implemented in practice because complete outcome profiles are not observable in real studies. However, since $\hat{d}(\pi)$ can be implemented in simulation, its operating characteristics can serve as a yardstick for the performance of other dose finding designs.

Algorithm 1 (Template to compute a benchmark)

1. Specify the complete outcome distribution π for all dose levels.
2. Define objective $d(\pi)$.
3. Generate the complete outcome profile for a simulated patient according to (3) and (4).
4. Obtain $\hat{\pi}$ according to (6) based on n simulated patients.

5. Evaluate benchmark $d(\pi)$.

As Steps 1 and 2 are necessary components in writing simulation code for a dose finding method, Steps 3–5 are the additional steps needed for computing a benchmark design. It is easy and quick to implement Steps 3 and 4 in computer simulation, and Step 5 is straightforward once the design objective $d(\pi)$ is defined in Step 2. Therefore, the operating characteristics of $d(\pi)$ can be obtained by repeated simulations with only minimal additional computing effort and time. In the following sections, we will illustrate the implementation of the steps in Algorithm 1, and demonstrate the application of benchmark in a number of specific clinical contexts with different design objectives.

3. Multiple Toxicities in Phase I Trials

3.1 Design Objectives Based on Toxicity Burden

In order to account for the different types and grades of toxicities a patient can experience, Lee et al. (2011, 2012) propose to summarize the toxicity profile by a single numeric index, called toxicity burden score (TBS), and impose multiple toxicity constraints in the dose and types of toxicities. To be more specific, Table 1 gives the severity weights used in calculating the TBS based on neuropathy and low platelet count in a bortezomib trial in lymphoma patients (Lee et al., 2012): for example, a patient with a grade 1 neuropathy and a grade 3 low platelet will have $TBS = 0.19 + 0.40 = 0.59$. Let $Y(k)$ denote the TBS of a patient at dose k . In a trial with J toxicity constraints, define $v_j = \arg \min_k |\Pr\{Y(k) \leq t_j\} - p_j|$ for $j = 1, \dots, J$ where $t_1 < \dots < t_J$ are pre-specified toxicity thresholds, and $p_1 > \dots > p_J$ are the respective target rates. Lee et al. (2011) introduce a model-based extension of the continual reassessment method for multiple constraints (called CRMMC) that aims to estimate

$$d_{MC}(\pi) = \min_j v_j. \quad (8)$$

As the objective (8) reduces to (2) in the case of single constraint, i.e., $J = 1$, the CRMMC will reduce to the regular continual reassessment method accordingly.

Another approach that accounts for multiple toxicities is proposed by Bekele and Thall (BT, 2004), who aim at a dose with a target expected total toxicity burden (TTB). A TTB is a sum of severity weights and is in essence the same as TBS; the only difference is the elicitation process of the weights assigned to the different grades and types of toxicities. If the same weights are used, the computations of TBS and TTB will be identical. We shall use the term TBS from now on while referring to both of these concepts. The BT method aims to estimate

$$d_{BT}(\pi) = \arg \min_k |E_{\pi}\{Y(k)\} - \tau|, \quad (9)$$

where τ is a pre-specified of mean toxicity burden, and

$$E_{\pi}\{Y(k)\} = \sum_{l=0}^L \{\pi_l(k) - \pi_{l+1}(k)\} w_l. \text{ We note that since the TBS is a function of the}$$

grades and types of toxicities (cf. Table 1), it takes on a finite number of values. Therefore, the distribution of $Y(k)$ can be completely specified by $\{\pi_1(k), \dots, \pi_L(k)\}$ for some finite L .

3.2 Numerical Illustrations

As we have explicitly worked out the form of the design objectives (8) and (9) per Step 2 of Algorithm 1, we next illustrate how to generate the complete outcome profile of a patient per Step 3 of the algorithm. Table 2 enumerates the distribution π under simulation Scenario 6 in Lee et al. (2011); the last column in the table gives a numerical example of the tolerance profile of patient i in a simulated trial, from which we can derive the complete outcome profile according to (4): For example, we can verify that $Y_i(4) = \omega_{15} = 2.53$ because $U_{i,16} = 0.77 > \pi_{16}(4)/\pi_{15}(4) = 0.71$ with $U_{ij} \leq \pi_j(4)/\pi_{j-1}(4)$ for all $j \leq 15$. Likewise, we can obtain $\{Y_i(1), Y_i(2), Y_i(3), Y_i(4), Y_i(5)\} = \{0, 0, 0, 2.53, 3.38\}$ for this patient, who would have had no toxicity should she or he receive dose level 3 or lower, but a severe toxicity (i.e., TBS 1.5) for dose levels 4 and 5. In the same manner, we can generate the complete outcome profiles for all n patients in a trial, estimate π using the complete outcomes via (6) per Step 4 of Algorithm 1, and obtain a realization of the benchmark design by plugging $\hat{\pi}$ into the design objective functions (8) and (9) per Step 5 of the algorithm.

3.3 Method Comparison

Lee et al. (2011) attempted to compare the CRMMC and the BT method in the context of the above mentioned bortezomib study. The sample size was 18; and there were $K = 5$ dose levels. The TBS outcomes were computed based on Table 1. For the CRMMC, the authors considered $J = 2$ constraints with toxicity thresholds $t_1 = 1$ and $t_2 = 1.5$, and target rates $p_1 = 0.25$ and $p_2 = 0.10$. For the BT method, Lee et al. (2011) set the target mean toxicity burden $\tau = 0.72$, so that $d_{MC}(\pi) = d_{BT}(\pi)$ under most of their simulation scenarios.

As the CRMMC and the BT method are tailored for different objectives, respectively $d_{MC}(\pi)$ and $d_{BT}(\pi)$, a direct comparison between them may not always be appropriate. Table 3 shows the operating characteristics of the CRMMC and the BT method against their respective benchmarks $d_{MC}(\hat{\pi})$ and $d_{BT}(\hat{\pi})$ in the scenarios in Lee et al. (2011): The results of the two methods are extracted from Table 3 in their paper, and that of the benchmark designs are evaluated based on 10,000 simulation replicates.

Under Scenarios 1–4, where the two design objectives (8) and (9) are identical, the two methods appear to have comparable accuracy. For example, in Scenario 4, both methods correctly select dose level 5 with a 0.57 probability. However, the ratio of the probability of correct selection of the CRMMC relative to its benchmark $d_{MC}(\hat{\pi})$ is $0.57 \div 0.65 \times 100\% = 88\%$. In this sense, the efficiency is quite high. By contrast, the correct selection ratio of the BT method against $d_{BT}(\hat{\pi})$ is only 76% under this scenario.

Under Scenario 5, where $d_{MC}(\pi) = d_{BT}(\pi) = 3$, the CRMMC appears to be superior by selecting dose 3 with a 0.64 probability, compared to 0.48 for the BT method. By observing that the benchmark $d_{BT}(\hat{\pi})$ makes a correct selection with a 0.54 probability, we note that this is a “difficult” scenario for the design objective (9). In fact, the BT method is slightly more efficient than the CRMMC in this scenario, with respective correct probability ratios

$89\% = 0.48 \div 0.54 \times 100\%$ and $83\% = 0.64 \div 0.77 \times 100\%$. The benchmark designs thus provide a basis for direct comparison between the two methods.

As expected, the CRMMC selects the correct dose uniformly less often than its benchmark $d_{MC}(\pi)$ in all scenarios; but the efficiency is generally quite high: The ratio of the probability of correctly selecting $d_{MC}(\pi)$ ranges from 79% to 88%. The performance of the BT method by contrast varies across scenarios: the efficiency against $d_{BT}(\pi)$ ranges from 64% to 88%. From a broader perspective, this comparison between the two methods is not entirely fair, because the prior of the BT method has not been properly chosen for the simulation setup in Lee et al. (2011), whereas the CRMMC has been carefully calibrated. In this specific context, the comparison to the benchmark design $d_{BT}(\pi)$ indicates that the BT method may be improved by a careful prior elicitation process.

4. Efficacy-Toxicity Trade-off in Phase I/II Trials

4.1 Design Objective Based on Desirability

Thall and Cook (2004) develop an outcome-adaptive, Bayesian model-based dose finding procedure for a study of thrombolytic agents in acute ischemic stroke patients. The main therapeutic endpoint of a thrombolytic agent is to re-establish blood flow (reperfusion). By the same therapeutic mechanism, however, it could also cause intracranial hemorrhage. A third possible outcome for a patient is to experience neither. The general study goal is to find a dose with “good” reperfusion rate without excessive hemorrhage. To be precise, let $F_T(k)$ and $F_E(k)$ respectively denote the probability of toxicity (hemorrhage) and the probability of response (reperfusion without hemorrhage) at dose k . Thall and Cook (2004) propose to express the desirability δ_k of dose k as a function of the response and toxicity rates of the dose; that is, $\delta_k = \delta\{F_E(k), F_T(k)\}$, and aim to find a dose that has the largest desirability.

Generally, the desirability function $\delta(q_E, q_T)$ can be specified from functions that are increasing in q_E and decreasing in q_T . However, the exact form of the function should be obtained through an elicitation process. For illustration purposes, we consider the function Thall and Cook (2004) use for the stroke trial:

$$\delta(q_E, q_T) = 1 - \frac{\sqrt{(1-q_E)^2 + q_T^2}}{\sqrt{(1-p_E)^2 + p_T^2}} \quad (10)$$

where (p_E, p_T) satisfies

$$p_T = -0.045 + \frac{0.0347}{p_E} - \frac{0.147}{p_E^2} \text{ and } q_T(q_E - p_E) = (q_E - 1)(q_T - p_T).$$

The desirability function (10) reaches its maximum at 1 under the best possible scenario for a dose with $q_E = 1$ and $q_T = 0$, equals 0 for (q_E, q_T) falling on the “target contour”, and < 0 for (q_E, q_T) outside the contour; see Figure 1 in Thall and Cook (2004) for a geometrical interpretation of (10). As such, the study objective is to estimate

$$d_{ET}(\pi) = \begin{cases} \arg \max_k \delta_k & \text{if } \delta_k \geq 0 \text{ for some } k \\ 0 & \text{if } \delta_k < 0 \text{ for all } k. \end{cases} \quad (11)$$

Using the notation developed in Section 2, the trinary outcome in this trial corresponds to $L = 2$ with $w_l = l$ for $l = 0, 1, 2$. Specifically, we define the outcome $Y = 0$ for no reperfusion or toxicity, $Y = 1$ for response, and $Y = 2$ for toxicity, so that $F_E(k) = \pi_1(k) - \pi_2(k)$ and $F_T(k) = \pi_2(k)$. Consequently, we can define the benchmark design $d_{ET}(\pi)$ by replacing δ_k with $\hat{\delta}_k$ in (11), where $\hat{\delta}_k$ is calculated according to (10) by setting $q_E = \pi_1(k) - \pi_2(k)$ and $q_T = \pi_2(k)$. The benchmark design can be generated using Algorithm 1; and the steps are similar to that outlined in Section 3.2 and are omitted here.

4.2 Method Comparison

The Thall-Cook method was applied to a stroke trial with $K = 5$ dose levels and a sample size of 72 subjects, with two different dose-outcome models: the proportional odds (PO) model that was used in an earlier work by Thall and Russell (1998) and the continuation ratio (CR) model that is more flexible than the PO model in terms of the number of regression parameters.

Table 4 shows the operating characteristics of the Thall-Cook procedures against the benchmark $d_{ET}(\pi)$. The results of the Thall-Cook procedures are extracted from Table 1 in Thall and Cook (2004), and that of $d_{ET}(\pi)$ are based on 10,000 simulation replicates. Overall, the Thall-Cook procedure using the CR model performs well with over 86% correct selection in the “easy” scenarios where the benchmark $d_{ET}(\pi)$ makes over 90% correct selection (i.e., Scenarios 1, 4, 5, 6). The CR model performs not as well in Scenario 3 and is correct only about half the time; under the scenario, the accuracy of the benchmark dips below 90%, indicating this as a relatively difficult scenario. Therefore, the method with the CR model is behaving in a reasonable manner by performing well under easy scenarios and not so well under difficult scenarios.

By contrast, the procedure with the PO model performs well under Scenario 3 but relatively poorly under the easy scenarios (e.g. Scenario 4). Thall and Cook (2004) give an in-depth look into the superior performance of the PO model in Scenario 3, and found that the PO model happened to fit this scenario much more closely than the CR model. By showing the method’s inferior performance in the easy scenarios, our simulation results further confirm that the PO model is not sufficiently flexible. In this specific context, the benchmark design serves as a yardstick to “rank” the difficulty of the simulation scenarios, so that we can detect when a dose finding method is not behaving reasonably.

5. Utility-based Optimization of Combination Therapy

5.1 Design Objective Based on Expected Utility

The desirability objective attempts to summarize the marginal toxicity and response rates into a single parameter value for each dose. An alternative approach to incorporate toxicity and clinical response is to directly assign a utility score to each possible joint outcome of response and toxicity, and aim to find a dose with maximum expected utility. Houede et al.

(2010) take this approach and apply Bayesian model-based design in a trial of combination therapy of a biological agent and a chemotherapy in patients with untreated, advanced bladder cancer. In the trial, there are $K_1 = 4$ levels of the biological agent and $K_2 = 3$ levels of gemcitabine. Thus, there are a total of $K = 12$ levels in this study, and each combination level is indexed with a pair $k = (k_1, k_2)$. Let Y denote the utility outcome of a patient. The authors consider three levels for toxicity and three for response, and hence Y takes on 9 possible values for each patient, i.e., $L = 8$. The study aims to identify a combination that maximizes the expected utility, that is, having design objective

$$d_{util}(\pi) = \arg \max_{(k_1, k_2)} E_{\pi}\{Y(k_1, k_2)\} = \arg \max_{(k_1, k_2)} \sum_{l=0}^L \{\pi_l(k_1, k_2) - \pi_{l+1}(k_1, k_2)\} w_l. \quad (12)$$

5.2 Method Comparison

Like the TBS, the utility score for each toxicity-response outcome is to be pre-defined. Table 5 presents the three sets of utility scores elicited for the bladder cancer trial: (A) the first set of utility was obtained using Delphi method; (B) the second set was chosen to reflect one oncologist's opinion; and (C) the third set was hypothetical scores that put greater value on response. Table 5 also gives the distribution π of the utility scores for each of the twelve combinations under Scenario 2 in Houede et al. (2010); the distribution can be obtained by information given in Supplementary Table 2 in the Web-based supplementary materials in their paper. In this scenario, under all three utility definitions, the target dose $d_{util}(\pi) = (2, 2)$, whereas (4,1) is the worst combination.

Using the information given in Table 5, we ran the benchmark design $d_{util}(\hat{\pi})$ and compared it to the model-based method by Houede et al. (2010) for the combination therapy trial with $n = 48$. The benchmark was computed by replacing π with $\hat{\pi}$ in (12). As the simulation outputs are very elaborate, for illustration purposes, we present here the comparison using the "R statistic" considered in their paper. In particular, for the scenario in Table 5, we summarize the benchmark accuracy by

$$R = \frac{\sum_{k_1, k_2} E_{\pi}\{Y(k_1, k_2)\} \times \Pr\{d_{util}(\hat{\pi}) = (k_1, k_2)\} - E_{\pi}\{Y(4, 1)\}}{E_{\pi}\{Y(2, 2)\} - E_{\pi}\{Y(4, 1)\}} \quad (13)$$

The expression (13) reaches its maximum at 1 if the benchmark selects the correct dose (2,2) with probability one, and minimum at 0 if it always selects the worst dose (4,1). This R statistic corresponds to Equation (14) in Houede et al. (2010) for their method, which yields 0.81, 0.68, and 0.45 respectively under the three utility definitions (Table 4 in their paper). For $d_{util}(\hat{\pi})$, we obtained $R = 0.99$ under all three sets of utility scores. This suggests that Scenario 2 has comparable level of difficulty in terms of identifying the correct combination regardless of the choice of the utility definition, and that the model-based, utility-based method is sensitive to the numerical utilities. From a practical viewpoint, this underlines the importance of using a relevant set of utilities and calibrating the model correspondingly.

6. Discussion

As demonstrated in the examples in the previous sections, modern dose finding studies are highly specific to individual clinical settings in terms of trial objective $d(\pi)$ and outcome Y . The challenge for the study statistician then is to adequately evaluate a dose finding method tailored to a study using computer simulation. The evaluation process is complex for two reasons. First, most dose finding procedures are outcome-adaptive and involve computer intensive algorithms such as Markov Chain Monte Carlo. It poses the practical burdens on the programmers to check for coding errors and plausibility of the simulation results, in order to truly assess the validity of the method. In this article, we propose a general approach to compute a benchmark that can serve as a yardstick for method evaluation. The benchmark design $d(\hat{\pi})$ sets an upper limit of the accuracy that we can expect for a given dose finding objective $d(\pi)$. By comparing a dose finding method against its benchmark, we can assess the plausibility of the simulation results of a method's performance. We can also gain insights whether there is room for improvement—if a method has uneven performance across scenarios relative to the benchmark, it may indicate that the model and prior distribution should be further calibrated. Importantly, the proposed approach is easy to implement. The only additional computational effort involves Steps 3 and 4 in Algorithm 1, which can be easily incorporated in the simulation programs for a dose finding method. Also, the proposed approach is versatile and is applicable to the most common clinical scenarios such as bivariate toxicity-efficacy outcomes and combination therapies.

Second, method evaluation becomes especially difficult in nonstandard phase I trials and phase I/II trials where the study outcome is more than a single binary value and the dose-outcome relationship gets intricate. As the complexity of simulation scenarios grow in these situations, it can be hard to gauge how a reasonable dose finding method should behave under different scenarios. In the context of hypothesis testing, a reasonable statistical test procedure is expected to yield higher power under a greater effect size. In the context of these complex dose finding studies, however, because the simulation scenarios involve a large number of parameters and elaborate models, the concept of “effect size” can be elusive. The benchmark accuracy (e.g. R statistic in Section 5.2) summarizes a complex dose-outcome relationship into a numerical index that can serve as an effect size of the given scenario. By examining the performance of the benchmark design under a wide range of simulation scenarios, we can rank the degrees of difficulty of the scenarios in terms of the benchmark accuracy. As the choice of simulation scenarios can influence the perceived performance of a dose finding method, a benchmark facilitates the standardization of scenarios.

Our intention is not to comment on the relative efficiencies of competing dose finding methods, although we have demonstrated that benchmarking provides a basis of comparison of methods that address similar but not identical trial objectives (cf. Section 3). Rather, our focus is on the diagnosis of the inadequacy of model specification of a particular method. We note that some accuracy index of a dose finding method under a given scenario should not be the only measure of success and should not be interpreted in absolute terms. In conducting a dose finding trial, there are additional ethical constraints such as coherence (Cheung, 2005) that may prevent an outcome-adaptive method from achieving near the

optimal benchmark in terms of accuracy. Therefore, these ethical factors should be considered when selecting from a competing set of dose finding methods. Having said this, in our experience so far, we find it useful to run the benchmark design under a large number of scenarios so as to form a “yardstick” for the performance of dose finding methods: Instead of using the absolute value of an accuracy index under a specific scenario, we can examine whether the relative performance of the dose finding method across scenarios is in agreement with that of the benchmark. If a dose finding method performs well in the difficult scenarios but relatively poorly in the easy ones (e.g. PO model in Section 4), it may be an indication of inadequacy. In addition to using an accuracy index, it may be useful to compare the variance in the estimation of the dose-outcome distribution π . While the variance estimation for a dose finding method is specific to the model and outcome, the variance of $\hat{\pi}$ for the benchmark design can be easily computed using (7). Given the typically enormous computing effort on coding an outcome-adaptive method and the minimal additional work required by the benchmark, such a diagnostic tool should be proven viable and valuable in most dose finding trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thanks the reviewers and editors for their comments. This work was supported by NIH grant R01 NS055809.

References

- Bekele BN, Thall PF. Dose finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association*. 2004; 99:26–35.
- Braun T. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials*. 2002; 23:240–256. [PubMed: 12057877]
- Cheung YK. Coherence principles in dose finding studies. *Biometrika*. 2005; 92:863–873.
- Cheung YK. Sequential implementation of stepwise procedures identifying the maximum tolerated dose. *Journal of the American Statistical Association*. 2007; 102:1448–1461.
- Cheung YK. Stochastic approximation and modern model-based designs for dose finding clinical trials. *Statistical Science*. 2010; 25:191–201. [PubMed: 21197369]
- Cheung YK. Sample size formulae for the Bayesian continual reassessment method. *Clinical Trials*. 2013 in press. Online 21 August 2013. 10.1177/1740774513497294
- Dragalin V, Fedorov VV. Adaptive designs for dose-finding based on efficacy-toxicity response. *Journal of Statistical Planning and Inference*. 2006; 136:1800–1823.
- Durham SD, Flournoy N, Rosenberger WF. A random walk rule for phase I clinical trials. *Biometrics*. 1997; 53:745–760. [PubMed: 9192462]
- Fleiss, J.; Levin, B.; Paik, MC. *Statistical Methods for Rates and Proportions*. New York: Wiley-Interscience; 2003.
- Houede J, Thall PF, Nguyen H, Paoletti X, Kramar A. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Biometrics*. 2010; 66:532–540. [PubMed: 19673865]
- Lee SM, Cheng B, Cheung YK. Continual reassessment method with multiple toxicity constraints. *Biostatistics*. 2011; 12:386–398. [PubMed: 20876664]
- Lee SM, Horseman DL, Martin P, Leonard JP, Cheung YK. Toxicity burden score: a novel approach to summarize multiple toxic effects. *Annals of Oncology*. 2012; 23:537–541. [PubMed: 21536663]

- Lin Y, Shih WJ. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics*. 2001; 2:203–215. [PubMed: 12933550]
- Liu G, Rosenberger WF, Haines LM. Sequential designs for logistic phase I clinical trials. *Journal of Biopharmaceutical Statistics*. 2006; 16:605–621. [PubMed: 17037261]
- O'Quigley J, Hughes, Fenton T. Dose-finding designs for HIV studies. *Biometrics*. 2001; 57:1018–1029. [PubMed: 11764240]
- O'Quigley J, Paoletti X, Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics*. 2002; 3:51–56. [PubMed: 12933623]
- O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical studies in cancer. *Biometrics*. 1990; 46:33–48. [PubMed: 2350571]
- Paul RK, Rosenberger WF, Flournoy N. Quantile estimation following non-parametric phase I clinical trials with ordinal response. *Statistics in Medicine*. 2004; 23:2483–2495. [PubMed: 15287079]
- Thall PF, Cook JD. Dose finding based on efficacy-toxicity trade-offs. *Biometrics*. 2004; 60:684–693. [PubMed: 15339291]
- Thall PF, Russell KT. A strategy for dose finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*. 1998; 54:251–264. [PubMed: 9544520]
- Yin G, Li Y, Ji Y. Bayesian dose finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics*. 2006; 62:777–784. [PubMed: 16984320]
- Yuan Z, Chappell R, Bailey H. The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics*. 2007; 63:173–179. [PubMed: 17447942]

Appendix

Proofs of propositions

Proof of Proposition 1

Suppose that $k'' > k'$ and the uniform variates are generated by (4) such that $Y_i(k') = w_l$. Then it is easy to derive from (4) that $U_{ij} = \pi_j(k')/\pi_{j-1}(k')$ for all $j = 1, \dots, l$. The assumption (5) in Proposition 1 then implies $U_{ij} = \pi_j(k'')/\pi_{j-1}(k'')$ for all $j = 1, \dots, l$. By (4), we have $Y_i(k'') = w_l$. But since this holds for all l , we have $Y_i(k'') = Y_i(k')$ as desired.

Proof of Proposition 2

For any given $l = 1, \dots, L$, we have

$$\begin{aligned} \Pr \{Y_i(k) \geq \omega_l\} &= \Pr \left\{ \bigcap_{j=1}^l U_{ij} \leq \frac{\pi_j(k)}{\pi_{j-1}(k)} \right\} \\ &= \prod_{j=1}^l \Pr \left(U_{ij} \leq \frac{\pi_j(k)}{\pi_{j-1}(k)} \right) \\ &= \frac{\pi_1(k)}{\pi_0(k)} \times \frac{\pi_2(k)}{\pi_1(k)} \times \dots \times \frac{\pi_l(k)}{\pi_{l-1}(k)} \\ &= \pi_l(k). \end{aligned}$$

The first equality is derived from (4), the second equality is by independence of U_{ij} s in (3), the third is by property of standard uniform distribution, and the last is due to $\pi_0(k) \equiv 1$.

Proof of Proposition 3

By Proposition 2, the indicator $I\{Y_i(k) \geq w_l\}$ is a Bernoulli variable with probability $\pi_l(k)$.

Thus, $E\{\hat{\pi}_l(k)\} = E\left[n^{-1} \sum_{i=1}^n I\{Y_i(k) \geq w_l\}\right] = n^{-1} \sum_{i=1}^n \pi_l(k) = \pi_l(k)$. That is, $\hat{\pi}_l(k)$ is unbiased for $\pi_l(k)$. In addition,

$$\text{var}\{\hat{\pi}_l(k)\} = \text{var}\left[\frac{1}{n} \sum_{i=1}^n I\{Y_i(k) \geq w_l\}\right] = \frac{1}{n^2} \sum_{i=1}^n \pi_l(k) \{1 - \pi_l(k)\} = \frac{\pi_l(k) \{1 - \pi_l(k)\}}{n}.$$

Table 1

Severity weights of toxicity types and grades in the bortezomib study (Lee et al., 2012)

Toxicity type	Grade	Weight
Neuropathy	0	0.00
	1	0.19
	2	0.64
	3	1.03
	4	2.53
Low platelet count	0	0.00
	1 or 2	0.17
	3	0.40
	4	0.85

Table 2

Distribution of TBS under Scenario 6 in Lee et al. (2011).

l	α_l	$\pi_l(1)$	$\pi_l(2)$	$\pi_l(3)$	$\pi_l(4)$	$\pi_l(5)$	U_{il}
0	0.00	1.00	1.00	1.00	1.00	1.00	
1	0.17	0.53	0.63	0.70	0.89	0.94	0.81
2	0.19	0.40	0.48	0.53	0.78	0.86	0.65
3	0.36	0.28	0.39	0.45	0.72	0.82	0.12
4	0.40	0.25	0.35	0.41	0.66	0.76	0.54
5	0.59	0.21	0.29	0.35	0.59	0.70	0.18
6	0.64	0.20	0.28	0.33	0.55	0.66	0.63
7	0.81	0.09	0.22	0.32	0.54	0.65	0.07
8	0.85	0.06	0.19	0.31	0.52	0.63	0.04
9	1.03	0.05	0.16	0.25	0.45	0.55	0.37
10	1.04	0.03	0.14	0.25	0.44	0.54	0.76
11	1.20	0.02	0.12	0.23	0.38	0.47	0.16
12	1.43	0.01	0.11	0.23	0.37	0.46	0.15
13	1.49	0.01	0.11	0.23	0.36	0.45	0.52
14	1.88	0.01	0.10	0.23	0.35	0.43	0.61
15	2.53	0.01	0.10	0.23	0.34	0.42	0.12
16	2.70	0.00	0.04	0.12	0.24	0.34	0.77
17	2.93	0.00	0.02	0.05	0.14	0.21	0.52
18	3.38	0.00	0.01	0.02	0.07	0.13	0.60
$E_{\pi}\{Y(k)\}$		0.25	0.51	0.81	1.28	1.6	

Table 3

Comparison of the CRMCMC and the BT method against their respective benchmarks. With 10,000 replicates, the Monte Carlo simulation errors for the benchmark designs are no greater than 0.5%.

Design	Percent of Selecting Dose				
	1	2	3	4	5
Scenario 1 in Lee et al.					
$\Pr\{\text{TBS}(k) \mid 1.0\}$	0.05	0.25	0.40	0.45	0.55
$\Pr\{\text{TBS}(k) \mid 1.5\}$	0.01	0.10	0.21	0.29	0.41
$E\{\text{TBS}(k)\}$	0.37	0.72	1.00	1.24	1.56
CRMCMC	24	58	16	3	0
$d_{MC}(\hat{\pi})$	19	70	10	1	0
BT	18	61	20	2	0
$d_{BT}(\hat{\pi})$	8	69	20	3	0
Scenario 2 in Lee et al.					
$\Pr\{\text{TBS}(k) \mid 1.0\}$	0.05	0.05	0.25	0.45	0.55
$\Pr\{\text{TBS}(k) \mid 1.5\}$	0.01	0.01	0.10	0.24	0.35
$E\{\text{TBS}(k)\}$	0.33	0.37	0.72	1.13	1.48
CRMCMC	2	25	62	11	0
$d_{MC}(\hat{\pi})$	1	18	74	6	0
BT	0	27	60	12	1
$d_{BT}(\hat{\pi})$	0	8	80	12	0
Scenario 3 in Lee et al.					
$\Pr\{\text{TBS}(k) \mid 1.0\}$	0.05	0.05	0.08	0.25	0.45
$\Pr\{\text{TBS}(k) \mid 1.5\}$	0.01	0.01	0.02	0.10	0.24
$E\{\text{TBS}(k)\}$	0.33	0.37	0.45	0.72	1.13
CRMCMC	0	3	31	57	9
$d_{MC}(\hat{\pi})$	1	1	24	67	6
BT	0	3	36	46	14
$d_{BT}(\hat{\pi})$	0	0	16	72	13
Scenario 4 in Lee et al.					

Design		Percent of Selecting Dose				
		1	2	3	4	5
$\Pr\{\text{TBS}(k)$	1.0}	0.05	0.05	0.08	0.12	0.25
$\Pr\{\text{TBS}(k)$	1.5}	0.00	0.01	0.02	0.04	0.10
$E\{\text{TBS}(k)\}$		0.30	0.33	0.45	0.54	0.72
CRM		0	2	6	36	57
$d_{MC}(\hat{\pi})$		0	2	6	27	65
BT		0	1	14	28	57
$d_{BT}(\hat{\pi})$		0	0	2	23	75
Scenario 5 in Lee et al.						
$\Pr\{\text{TBS}(k)$	1.0}	0.05	0.05	0.25	0.45	0.55
$\Pr\{\text{TBS}(k)$	1.5}	0.00	0.01	0.05	0.10	0.20
$E\{\text{TBS}(k)\}$		0.33	0.37	0.59	0.90	1.16
CRM		1	17	64	17	1
$d_{MC}(\hat{\pi})$		1	7	77	14	1
BT		0	8	48	39	6
$d_{BT}(\hat{\pi})$		0	1	54	43	2
Scenario 6 in Lee et al.						
$\Pr\{\text{TBS}(k)$	1.0}	0.05	0.16	0.25	0.45	0.55
$\Pr\{\text{TBS}(k)$	1.5}	0.01	0.10	0.23	0.35	0.43
$E\{\text{TBS}(k)\}$		0.25	0.51	0.81	1.28	1.56
CRM		16	52	27	4	0
$d_{MC}(\hat{\pi})$		15	66	18	1	0
BT		5	52	37	5	1
$d_{BT}(\hat{\pi})$		1	35	56	8	0

Table 4

Comparison of the Thall-Cook procedure against the desirability benchmark. With 10,000 replicates, the Monte Carlo simulation errors for the benchmark design are no greater than 0.5%.

Method	Percent of Selecting Dose					
	1	2	3	4	5	None
Scenario 1 in Thall and Cook (2004)						
$(F_E, F_T) =$	(0.05,0.01)	(0.20,0.02)	(0.35,0.03)	(0.60,0.04)	(0.80,0.05)	
$\delta =$	-0.74	-0.48	-0.22	0.22	0.54	
PO	0.0	0.0	0.2	22.5	74.2	3.1
CR	0.0	0.0	0.7	5.8	92.8	0.7
$d_{ET}(\hat{\pi})$	0.0	0.0	0.0	0.1	99.8	0.1
Scenario 2 in Thall and Cook (2004)						
$(F_E, F_T) =$	(0.57,0.01)	(0.58,0.03)	(0.60,0.06)	(0.62,0.20)	(0.64,0.32)	
$\delta =$	0.21	0.20	0.18	-0.31	-1.00	
PO	0.7	10.8	72.2	15.2	0.6	0.5
CR	0.1	20.5	61.9	16.1	0.9	0.5
$d_{ET}(\hat{\pi})$	40.1	25.5	31.9	1.2	0.0	1.4
Scenario 3 in Thall and Cook (2004)						
$(F_E, F_T) =$	(0.20,0.02)	(0.40,0.03)	(0.60,0.04)	(0.68,0.06)	(0.74,0.20)	
$\delta =$	-0.48	-0.13	0.22	0.32	-0.26	
PO	0.0	0.4	19.8	71.6	6.5	1.7
CR	0.0	1.6	32.2	49.4	15.7	1.0
$d_{ET}(\hat{\pi})$	0.0	0.0	12.7	85.2	1.1	1.0
Scenario 4 in Thall and Cook (2004)						
$(F_E, F_T) =$	(0.52,0.01)	(0.62,0.015)	(0.71,0.02)	(0.79,0.025)	(0.86,0.03)	
$\delta =$	0.12	0.29	0.45	0.58	0.69	
PO	0.1	1.7	10.1	34.3	53.7	0.0
CR	0.0	0.1	1.1	4.6	94.0	0.1
$d_{ET}(\hat{\pi})$	0.0	0.0	0.4	4.5	95.1	0.0
Scenario 5 in Thall and Cook (2004)						

Method	Percent of Selecting Dose					
	1	2	3	4	5	None
$(F_E, F_T) =$						
	(0.05, 0.18)	(0.20, 0.22)	(0.35, 0.26)	(0.47, 0.30)	(0.58, 0.33)	
$\delta =$	-1.03	-0.90	-0.85	-0.94	-1.07	
PO	0.0	2.1	1.9	0.3	0.1	95.6
CR	0.1	0.9	1.6	1.4	0.2	97.3
$d_{ET}(\pi)$	0.0	0.0	0.0	0.0	0.1	99.9
Scenario 6 in Thall and Cook (2004)						
$(F_E, F_T) =$						
	(0.15, 0.08)	(0.38, 0.18)	(0.52, 0.25)	(0.59, 0.30)	(0.62, 0.35)	
$\delta =$	-0.66	-0.50	-0.64	-0.89	-1.18	
PO	0.2	12.7	1.9	0.0	0.0	85.2
CR	0.4	11.4	1.3	0.0	0.0	86.9
$d_{ET}(\pi)$	0.0	0.1	0.7	0.1	0.0	99.1

Table 5

Distribution $\pi(k_1, k_2)$ of the utility scores for each combination under Scenario 2 in Houede et al. (2010). For toxicity, $T = 0$ if no grade 3/4 toxicity occurs; $T = 1$ if grade 3/4 toxicity is resolved within 2 weeks; $T = 2$ if grade 3/4 toxicity is not resolved within 2 weeks. For response, $E = 0$ for progressive disease, 1 for stable disease, and 2 for tumor response.

(A) Delphi method														
I	(T, E)	ω_I	(1, 3)	(2, 3)	(3, 3)	(4, 3)	(1, 2)	(2, 2)	(3, 2)	(4, 2)	(1, 1)	(2, 1)	(3, 1)	(4, 1)
0	(2,0)	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	(1,0)	10	0.93	0.92	0.90	0.88	0.99	0.99	0.98	0.90	0.98	0.97	0.95	0.86
2	(0,0)	25	0.80	0.80	0.80	0.80	0.82	0.91	0.90	0.80	0.70	0.71	0.74	0.72
3	(2,1)	40	0.80	0.80	0.80	0.80	0.80	0.90	0.90	0.80	0.60	0.65	0.70	0.70
4	(2,2)	52	0.58	0.55	0.50	0.44	0.76	0.87	0.83	0.47	0.59	0.58	0.58	0.39
5	(1,1)	60	0.47	0.43	0.35	0.27	0.75	0.79	0.67	0.35	0.56	0.57	0.55	0.34
6	(0,1)	76	0.14	0.12	0.10	0.08	0.24	0.53	0.44	0.08	0.20	0.16	0.13	0.07
7	(1,2)	82	0.14	0.12	0.10	0.08	0.18	0.52	0.44	0.08	0.09	0.08	0.08	0.04
8	(0,2)	100	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.02	0.01	0.01	0.00
$E_{\pi}\{Y(k_1, k_2)\}$			47.2	46.2	44.2	42.2	54.6	65.8	61.8	43.4	44.6	44.9	45.1	39.1

(B) A single oncologist														
I	(T, E)	ω_I	(1, 3)	(2, 3)	(3, 3)	(4, 3)	(1, 2)	(2, 2)	(3, 2)	(4, 2)	(1, 1)	(2, 1)	(3, 1)	(4, 1)
0	(2,0)	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	(1,0)	10	0.93	0.92	0.90	0.88	0.99	0.99	0.98	0.90	0.98	0.97	0.95	0.86
2	(0,0)	40	0.80	0.80	0.80	0.80	0.82	0.91	0.90	0.80	0.70	0.71	0.74	0.72
3	(2,1)	40	0.80	0.80	0.80	0.80	0.80	0.90	0.90	0.80	0.60	0.65	0.70	0.70
4	(2,2)	40	0.58	0.55	0.50	0.44	0.76	0.86	0.83	0.47	0.57	0.58	0.58	0.39
5	(1,1)	50	0.47	0.43	0.35	0.27	0.75	0.79	0.67	0.35	0.56	0.57	0.55	0.34
6	(1,2)	80	0.14	0.12	0.10	0.08	0.24	0.53	0.44	0.08	0.20	0.16	0.13	0.07
7	(0,1)	90	0.00	0.00	0.00	0.00	0.07	0.03	0.00	0.00	0.12	0.09	0.06	0.03
8	(0,2)	100	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.02	0.01	0.01	0.00
$E_{\pi}\{Y(k_1, k_2)\}$			42.1	41.2	39.5	37.8	50.0	61.5	56.6	38.9	43.6	42.5	41.7	36.0

(C) Hypothetical

I	(T, E)	ω	(1, 3)	(2, 3)	(3, 3)	(4, 3)	(1, 2)	(2, 2)	(3, 2)	(4, 2)	(1, 1)	(2, 1)	(3, 1)	(4, 1)
0	(2, 0)	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0	(1, 0)	5	0.93	0.92	0.90	0.88	0.99	0.99	0.98	0.90	0.98	0.97	0.95	0.86
0	(0, 0)	10	0.80	0.80	0.80	0.80	0.82	0.91	0.90	0.80	0.70	0.71	0.74	0.72
0	(2, 1)	30	0.80	0.80	0.80	0.80	0.80	0.90	0.90	0.80	0.60	0.65	0.70	0.70
0	(1, 1)	40	0.58	0.55	0.50	0.44	0.76	0.87	0.83	0.47	0.57	0.58	0.58	0.39
0	(0, 1)	50	0.25	0.25	0.25	0.25	0.25	0.61	0.60	0.20	0.20	0.18	0.16	0.13
0	(2, 2)	80	0.25	0.25	0.25	0.25	0.20	0.60	0.60	0.20	0.10	0.10	0.10	0.10
0	(1, 2)	90	0.14	0.12	0.10	0.08	0.18	0.52	0.44	0.08	0.09	0.08	0.07	0.04
0	(0, 2)	100	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.02	0.01	0.01	0.00
$E_{\pi}\{Y(k_1, k_2)\}$			41.8	41.4	40.5	39.6	43.2	65.7	64.1	38.0	32.2	32.9	35.6	30.6