2013

# A Likelihood-Based Approach for Computing the Operating Characteristics of the Standard Phase I Clinical Trial Design

Cody Chiuzan
*Medical University of South Carolina*

Elizabeth Garrett-Mayer

Sharon D. Yeatts
*Medical University of South Carolina*

**A likelihood-based approach for computing the operating characteristics of the standard phase I clinical trial design**

**Authors**: Cody Chiuzan[a] M.S., Elizabeth Garrett-Mayer[a,b] Ph.D., Sharon D Yeatts[a] Ph.D.

[a]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA

[b]Hollings Cancer Center, Medical University of South Carolina, Charleston, SC, USA

**Author for correspondence**:
Elizabeth Garrett-Mayer
Hollings Cancer Center
Medical University of South Carolina
86 Jonathan Lucas St., Rm. 118G
Charleston, SC 29425
Phone: 843-792-7764
Fax: 843-792-4233
Email: garrettm@musc.edu

This work was supported by a National Institute of Health (NIH) grant, P01 CA 154778-01 and part by the Biostatistics Shared Resource, Hollings Cancer Center, Medical University of South Carolina (P30 CA138313).

**Abstract**

In phase I clinical trials, the standard '3+3' design has passed the test of time and survived various sample size adjustments, or other dose-escalation dynamics. The objective of this study is to provide a probabilistic support for analyzing the heuristic performance of the '3+3' design. Our likelihood method is based on the evidential paradigm that uses the likelihood ratio to measure the strength of statistical evidence for one simple hypothesis over the other. We compute the operating characteristics and compare the behavior of the standard algorithm under different hypotheses, levels of evidence, and true (or best guessed) toxicity rates. Given observed toxicities per dose level, the likelihood-ratio is evaluated according to a certain *k* threshold (level of evidence). Under an assumed true toxicity scenario the following statistical characteristics are computed and compared: i) probability of weak evidence, ii) probability of favoring $H_1$ under $H_1$ (analogous to 1-α), iii) probability of favoring $H_2$ under $H_2$ (analogous to 1-β). This likelihood method allows consistent inferences to be made and evidence to be quantified regardless of cohort size. Moreover, this approach can be extended and used in phase I designs for identifying the highest acceptably safe dose and is akin to the sequential probability ratio test.

**Keywords**: phase I clinical trials, standard algorithm, likelihood method, evidential paradigm.

## 1. Introduction

Phase I trials in which new drugs or drug combinations are administered to human patients for the first time are conducted to select a dose to be used in subsequent trials. In oncology and other life threatening diseases, dose-finding studies most often aim to identify the maximum tolerated dose (MTD) defined as a dose whose probability of toxicity is closest to some acceptable, prespecified target, also known as the dose limiting toxicity (DLT) rate. The primary outcome for these trials is usually a binary indicator of the presence or absence of a DLT, with the underlying assumption that the probability of toxicity is a non-decreasing function of dose.

Despite considerable efforts since 1990 to encourage the use of model-based dose-finding designs, such as the Continual Reassessment Method (CRM) and its variants (see, e.g., O'Quigley, Pepe, and Fisher, 1990; Piantadosi, Fisher, and Grossman, 1998; Goodman, Zahurak, and Piantadosi, 1995; Yuan, Chappell, and Bailey, 2007; Cheung and Chappell, 2000; Moller, 1995), and Escalation with Overdose Control (EWOC) (Babb, Rogatko, and Zacks, 1998; Tighiouart, Rogatko, and Babb, 2005), the most common approach for dose-finding remains an 'Up-and-Down' algorithm. These methods assign patients sequentially based on prespecified decision rules. They are easy to implement, requiring no cumbersome calculations (before or during the trial) and no pre-specification of the underlying dose-toxicity model. Currently, there are a multitude of ad-hoc 'Up-and-Down' methods with the majority falling within the range of 'A+B' designs (Ivanova, 2006; Lin and Shih, 2001).

The most common 'A+B' design is the '3+3' algorithm, planned to sample around the $33^{rd}$ percentile (Storer, 1989). The popularity of the design is due mainly to its practical simplicity. For a limited number of dose levels ($\leq 5$), the '3+3' showed comparable properties to the CRM in terms of number of patients treated to reach the MTD (Iasonos et al., 2008). A more

recent study argued the utility of the '3+3' in practice by demonstrating (via simulations) that the standard algorithm was a better method when none of the investigational dose levels was close to the true MTD (Ji and Wang, 2013). An important limitation is its short memory (i.e., the decision rules are based on the outcomes from the most recent cohort of patients). Another drawback consists of a slow dose escalation, leading to treatment of an excessive number of patients at dose levels less likely to be efficacious (O'Quigley et al., 1990; Goodman et al., 1995; Storer, 1989). After evaluating the '3+3' operating characteristics, Lin and Shih (2001) concluded that the design does not have a fixed DLT at the MTD, such as 33%, and that it targets doses with a DLT rate between 16%-27% (Ivanova, 2006). Reiner, Paoletti, and O'Quigley (1999) concluded that this design has high error rates and frequently leads to incorrect decisions, e.g. the probability of recommending the correct MTD at the end of the trial never exceeds 44% and is actually closer to 30%. Also, estimators of the MTD are based on information collected from only six patients, and these tend to be biased or inconsistent (Storer, 1989; Brownlee, Hodges, and Rosenblatt, 1953; O'Quigley, 2006). Due to the empirical nature, the '3+3' has limited capabilities of describing and accounting for uncertainties in the observed data. However, despite the limited power of generality and the rigid design, 98% of the dose-finding cancer trials conducted between 1991 and 2006 implemented variations of the standard 'Up-and-Down' method (Rogatko et al., 2007).

Given the intuitive and transparent implementation, the '3+3' design continues to be clinicians' most popular choice for phase I trials. The objective of this study is not to argue the use of the standard '3+3' design, but to provide a method for describing the statistical properties of its heuristic performance. Our likelihood-based method can be used to compute the operating characteristics and compare the behavior of the standard design under different hypotheses,

levels of evidence, and true (or best guessed) toxicity rates. The method is based on the evidential paradigm that uses observed data to compute the likelihood ratio (LR), and then classify the level of evidence as: 1) weak evidence or 2) strong evidence in favor of one of the proposed hypotheses (denoted here as $H_1$ and $H_2$). This approach uses only the observed data as evidence for one hypothesis versus the other and provides an objective measure of the strength of this evidence. First, we present the evidential paradigm. Next, we apply our evidential approach to the '3+3' design and compute the operating characteristics under a wide range of true toxicities, four sets of simple hypotheses, and three levels of evidence providing guidance for the performance of the '3+3' based on acceptable and unacceptable DLT rates. Last, we present results from a simulation study, draw conclusions, and give some suggestions for extending the method.

## **2**. **The evidential paradigm**

Forster (2006) stated in one of his articles that "contemporary statistics is divided into three camps: classical Neyman-Pearson statistics, Bayesianism, and third, but not last, Likelihoodism." All three approaches are structured upon the likelihood ratio (LR) and the specification of a set of hypotheses, where usually the alternative represents the minimum clinically important difference. Likelihoodism is another school of thought of evidential statistics that uses data-based evidence to quantify the relative support for one model versus the other. The concept was first introduced by Hacking (1965) by stating the formal expression of the Law of Likelihood and using the likelihood ratio (*LR*) for comparing two simple hypotheses, such as $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$ for a parameter $\theta$, under the assumption that a background model is true. The evidential paradigm provides the *LR* of the two hypotheses, $LR = L(\theta_2; x) / L(\theta_1; x)$ as an

objective measure of the strength of evidence. Strong evidence supporting $\theta_2$ over $\theta_1$ exists if for a large $k$, $LR \geq k$. Similarly, strong evidence supporting $\theta_1$ over $\theta_2$ exists if $LR \leq 1/k$. Weak evidence occurs when $1/k < LR < k$, with no strong support for either one of the hypotheses. Royall (1997) and Blume (2002) established a correspondence between the values of $k$ and type I and II errors. They proposed benchmarks of 8 and 32, representing "weak" ($1 < LR < 8$), "fairly strong" ($8 < LR < 32$), and "strong" ($LR > 32$) levels of evidence. Initially, these benchmarks were derived to provide levels of evidence similar to error thresholds of $\alpha = 0.05$ and $\beta = 0.20$, in the context of a phase III trial. In phase I trials, controlling these error rates is not as stringent. However, one should be concerned with the small number of subjects enrolled (usually $< 25$). The limited amount of accumulated information (evidence) seldom generates likelihood ratios less than 8. Therefore, a value of $k = 8$ may be considered an unrealistic threshold for quantifying evidence in phase I studies.

*Error probabilities in the evidential paradigm*

A key aspect of evaluating study designs is computing the operating characteristics, such as the frequency with which the study will produce misleading or weak evidence. Ideally, the probabilities measuring how often evidence of a particular type will be observed should not affect the strength of statistical evidence quantified by the likelihood-ratio. For example, in the evidential paradigm, the probability of observing misleading evidence is a function of the fixed $k$ (strength of evidence) and sample size $n$. On the contrary, the analogous type I error from hypothesis testing is fixed at $\alpha$ and the strength of evidence at which the test rejects, $k_{\alpha,n}$, depends on $\alpha$ and $n$. In this situation, it is often possible that two tests that reject at the same $\alpha$ level could have different strengths of evidence.

The evidential paradigm defines misleading evidence as strong evidence in favor of the incorrect hypothesis, calculated under a true hypothesis. Given two simple hypotheses $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$, for $\mathbf{x} = (x_1, x_2, ..., x_n)$ i.i.d. observations, the probabilities of observing misleading evidence can be calculated as follows:

$$M_1(n,k) = P_1\left( \frac{L(\theta_2;\mathbf{x})}{L(\theta_1;\mathbf{x})} \geq k \mid H_1 \text{ true} \right)$$

$$M_2(n,k) = P_2\left( \frac{L(\theta_2;\mathbf{x})}{L(\theta_1;\mathbf{x})} \leq \frac{1}{k} \mid H_2 \text{ true} \right), \quad k \text{ fixed over } n.$$

Following the definition above, $M_1(n,k)$ represents the evidential analog to a type I error. It has been shown that for any fixed sample size *n* and any pair of probability distributions, the probability of misleading evidence under the true hypothesis satisfies the *universal bound* (Royall, 1997; Royall, 2000). That is, the probability that accumulated observations will represent strong evidence supporting the false hypothesis over the true hypothesis cannot exceed $1/k$ (with a similar bound under $H_2$):

$$M_1(n,k) = P_1\left( \frac{L(\theta_2;\mathbf{x})}{L(\theta_1;\mathbf{x})} \geq k \mid H_1 \text{ true} \right) \leq \frac{1}{k}.$$

This feature is extremely useful in sequential trials, where multiple looks at the data produce an inflation of the type I error. In the evidential approach, the probability of observing misleading evidence increases with each look, but it still remains bounded (Robbins, 1970).

The second error probability - probability of observing weak evidence - is defined as the probability that an experiment will not produce strong evidence for either hypothesis relative to the other, calculated under each true hypothesis:

$$W_1(n,k) = P_1\left(\frac{1}{k} < \frac{L(\theta_2;\mathbf{x})}{L(\theta_1;\mathbf{x})} < k \mid H_1 \text{ true}\right)$$

$$W_2(n,k) = P_2\left(\frac{1}{k} < \frac{L(\theta_2;\mathbf{x})}{L(\theta_1;\mathbf{x})} < k \mid H_2 \text{ true}\right), \quad k \text{ fixed over } n.$$

Similarly, $W_2(n,k)$ represents the evidential analog to the type II error. In his tutorial, Blume (2002) emphasizes the difference in definition and behavior between the probabilities of misleading and weak evidence and the classical type I and type II errors. Furthermore, in the context of experimental design, he shows that both probabilities of misleading and weak evidence converge to zero as the sample size increases.

**3. Methods**

Consider the general setting of the '3+3' design that uses cohorts of 3 patients with the ultimate goal of finding the MTD. The algorithm begins with 3 patients treated at the first (lowest) dose level. If 0 out of 3 patients experiences a DLT, the dose will be escalated. If 2 out of 3 patients have a DLT, the dose will be de-escalated. If 1 out of 3 patients has a DLT, 3 more patients will be enrolled at the same dose level, and if no additional patients experience a DLT, then the dose will be escalated. Otherwise (i.e., 2 or more DLTs at a dose level) de-escalation will occur. The algorithm continues until at least two patients among a cohort of 3 to 6 patients experience a DLT or the maximum dose level specified in the trial is reached. The MTD is defined as the dose level just below the toxic dose level which will have either 0 or 1 DLTs among 3 or 6 patients.

With a maximum of 6 patients per dose, the precision with which the true DLT rate can be estimated at each dose is very poor. With such a high level of uncertainty in the true DLT rate at the defined MTD, the safety profile cannot sufficiently be established. In order to improve the MTD estimation, the *Accelerated Titration* design fits a logistic model to all data after the trial

completion and generates a point estimate (with confidence interval) for the MTD (Simon et al., 1997). Even though the median MTD is similar to that derived from traditional phase I studies, the precision of the MTD estimate is still limited because of the small sample size.

For those who choose to use the '3+3' design, we offer statistically derived properties for describing its behavior under different scenarios and for quantifying the levels of evidence. Our likelihood method can be used either in the preparatory phase of the trial for determining whether or not the '3+3' can provide a dose with an acceptable DLT rate, or after completion, for evaluating the operating characteristics at each dose level. As discussed later, it can also provide guiding principles for ascertaining toxicity of a dose when the cohort size is beyond the standard number of 3 or 6 patients used by the '3+3' algorithm.

In our proposed method, all the statistical properties are calculated per each dose level, based on observed toxicities. Let $d_j, j = 1, 2, ..., K,$ be the set of ordered dose levels, and $y^j$ be the corresponding number of observed toxicities at the $j^{\text{th}}$ dose. Let $n_j$ be the number of patients per dose with a maximum of 6. For each dose, let $P(DLT \mid dose = d_j) = p_j$, the true probability of observing a DLT at the $j^{\text{th}}$ dose. Consider the following where $P(DLT \mid dose = d_j) = p_1$ and $P(DLT \mid dose = d_j) = p_2$ are the two hypothesized DLT rates at dose $d_j$:

$$H_1^j : p_j = p_1 \text{ (unsafe dose)}$$
$$H_2^j : p_j = p_2 \text{ (acceptable dose)}$$

For a choice of $p_1$ and $p_2$ established *a priori*, we calculate the likelihood-ratio (*LR*) for each dose:

$$LR_j = \left(\frac{p_2}{p_1}\right)^{y^j} \left(\frac{1-p_2}{1-p_1}\right)^{n_j - y^j}, j = 1, 2, ..., K$$

Using the estimated likelihood-ratio and a certain benchmark *k*, we interpret the strength of evidence as follows:

i. Weak evidence (supporting neither hypothesis), if $\frac{1}{k} < LR_j < k$

ii. Evidence in favor of $H_2$, if $LR_j \geq k$

iii. Evidence in favor of $H_1$, if $LR_j \leq \frac{1}{k}$

As mentioned previously, benchmark values of $k = 8$ and 32 have been proposed to distinguish between weak, moderate, and strong evidence (Royall, 1997; Blume, 2002). In the case of '3+3', we know that a maximum sample size of 6 patients can produce only relatively modest likelihood ratios (except when there is a very large difference between $p_1$ and $p_2$, such as $p_1 = 0.60$, $p_2 = 0.10$). Thus, selecting a $k$ greater than 8 is not feasible for such small cohort sizes. After toxicity responses at a certain dose level have been observed, the strength of evidence for one hypothesis over the other is quantified and a decision is made of "too toxic", "safe" or "weak evidence". The last category can be subject to interpretability. We regard the weak evidence as not having enough information to conclude toxicity, and choose to combine it with the evidence of concluding that the dose is safe. In other words, a dose is considered safe until there is sufficient evidence to conclude that it is too toxic.

In order to assess the '3+3' behavior in the context of likelihood inference, we also calculate the probabilities of escalation (dose is safe) and non-escalation (dose is unsafe) based on the algorithmic rules. For any dose $d_j$, let $y_1^j$ be the count of DLTs in the first cohort of 3 patients, and let $y_2^j$ be the count of DLTs in the second cohort of 3 patients. Then, $y_1^j$ and $y_2^j$ are

two independent binomial random variables, with $n_j = 3$ and $P(DLT \mid dose = d_j) = p_j$ for $j = 1, 2, ..., K$. Thus, for any given $p_j$, the following probabilities hold true:

$$\begin{aligned} P(escalation \mid d_j) &= P\left((y_1^j = 0) \cup (y_1^j = 1, y_2^j = 0)\right) \\ &= P(y_1^j = 0) + P(y_1^j = 1) \cdot P(y_2^j = 0) \\ &= (1 - p_j)^3 + 3p_j(1 - p_j)^5 \\ P(non - escalation \mid d_j) &= 1 - P(escalation \mid d_j) \end{aligned}$$

*Simulation set-up*

We calculated the operating characteristics of the '3+3'design using the likelihood method. For each dose, $H_1(p_1)$ is defined as the unsafe DLT rate, and $H_2(p_2)$ is considered an acceptable DLT rate. Data were simulated under a wide range of true toxicity rates (from 0.05 to 0.70). The following performance characteristics were considered:

- Probability of weak evidence
- Probability of favoring $H_1$ under $H_1$, $P(favors\ H_1 \mid H_1)$, analogous to $1 - \alpha$
- Probability of favoring $H_2$ under $H_2$, $P(favors\ H_2 \mid H_2)$, analogous to $1 - \beta$

The statistical properties were further compared using several cutoffs for the likelihood-ratio. We highlight the results by presenting the benchmarks with the most distinctive behavior as far as classifying the levels of evidence, i.e., $k = 1, 2, \text{and } 8$. All simulation scenarios were conducted with 10,000 trials each using the statistical software R (2009).

Numerous hypotheses $(p_1, p_2)$ were tested and compared in terms of operating characteristics. In this paper, we discuss four of the most interesting and relevant sets of hypotheses. The values for the first two scenarios: (A) $(p_1 = 0.40, p_2 = 0.15)$ and

(B) ($p_1 = 0.50, p_2 = 0.10$) were selected to be consistent with what is assumed regarding the '3+3' design: it targets a DLT rate around 20-30%. These cases characterize situations when the absolute difference between the hypotheses is greater than or equal to 30% with a midpoint close to 30%. The other two scenarios: (C) ($p_1 = 0.15, p_2 = 0.05$) and (D) ($p_1 = 0.50, p_2 = 0.30$) were selected to demonstrate the poor behavior of the '3+3' design in selecting the MTD when both hypotheses are either below or above a DLT rate of 0.20. These scenarios can be very well encountered in practice. For agents where lethal or life threatening toxicities are expected (e.g., cytotoxic agents), investigators might be only willing to accept relatively low DLT rates. Contrarily, the target DLT probability can be set high when toxicities are transient and nonfatal. This may be the case for biologic agents used in immunotherapy, such as vaccines or adoptive cell therapy, where a higher DLT rate (greater than 25-30%) might be considered tolerable.

## 4. Results

*Levels of evidence*

Tables 1 and 2 show the estimated likelihood ratios (*LR*) and the decisions regarding dose safety. For each dose one of the following decisions can be made: "acceptable dose", "toxic dose" or "weak evidence", i.e., not having enough evidence to conclude neither. Several benchmark values of $k$ were considered. However, we present the results for $k$ = 1, 2, and 8 that mark a significant change in the level of evidence. The scale of statistical evidence is not discrete, i.e., evidence does not suddenly move from one category to another. For example, for scenario (A) benchmarks of 3, 4, and 5 generate the same conclusions as $k$ = 2 regarding the strength of evidence. The same stands for values of 7 and 8.

Table 1 illustrates scenarios (A) $(p_1 = 0.40, p_2 = 0.15)$ and (B) $(p_1 = 0.50, p_2 = 0.10)$. For $k = 1$, the likelihood approach generates the same inferences as the '3+3" algorithm, by favoring one of the two hypotheses (Table 1). For a threshold of 1, the weak evidence category is eliminated. For $k = 2$, the only weak evidence category is 2 DLTs out of 6 for scenario (A) $(p_1 = 0.40, p_2 = 0.15)$. Weak levels of evidence are common under $k = 8$ (Table 1). A cutoff of 8 denotes fairly strong evidence and with cohorts of only 3 or 6 patients at a dose level we expect weak evidence, especially when the difference between the hypotheses values is not very large.

Table 2 displays scenarios (C) $(p_1 = 0.15, p_2 = 0.05)$ and (D) $(p_1 = 0.50, p_2 = 0.30)$. For scenario (C), the likelihood method disagrees with the '3+3' algorithm only for 1 out of 6 DLTs and $k = 1$. For $k = 2$ and $k = 8$, 1 out of 6 and 0 out of 3 DLTs are classified as weak evidence. In the high toxicity scenario (D), 2 out of 6 DLTs are considered acceptable ($k = 1$). For $k = 2$, only 3 out of 6 DLTs are regarded as weak evidence, the rest being in good agreement with the '3+3' rules. For $k = 8$ all categories are classified as weak evidence.

*Operating characteristics*

The operating characteristics of the likelihood method were further evaluated by computing the following probabilities over a range of true DLT rates: *P(weak evidence supporting neither hypothesis)*, $P(\textit{favors } H_1 \mid H_1), P(\textit{favors } H_2 \mid H_2)$, and comparing them to the operating characteristics of the '3+3' design. In figures 1 – 5, the solid lines mark the probabilities of favoring $H_2$ - dose is acceptable (black) and favoring $H_1$ - dose is unsafe (gray) generated by the likelihood method. Using the same coloristic, the dashed lines represent the analogous probabilities derived from the standard '3+3' algorithm: probability of escalation - dose is

acceptable (black) and probability of non-escalation – dose is unsafe (gray). The gray dotted line marks the level of weak evidence.

With a zero probability of weak evidence for $k = 1$, scenarios (A) $(p_1 = 0.40, p_2 = 0.15)$ and (B) $(p_1 = 0.50, p_2 = 0.10)$ are in perfect agreement with the '3+3' design (Figure 1A and 1B). For true DLT rates of 0.15 and 0.10, the probability of correctly declaring the dose acceptable is of 81% and 88%, respectively, as shown by the height of the black solid and dashed lines where they intersect the vertical line $H_2$ at 0.15 and 0.10 (Figure 1A and 1B). When the true DLT rates are 0.40 and 0.50, the chances of correctly declaring the dose unsafe are 69% and 82%, respectively, as shown by the height of the gray solid and dashed lines where they intersect the vertical line $H_1$ at 0.40 and 0.50 (Figure 1A and 1B).

For scenario (C) $(p_1 = 0.15, p_2 = 0.05)$ and $k = 1$, the algorithm has a higher probability of correctly favoring $H_2$ (dose is acceptable) than the likelihood method (97% vs. 86%) at a true DLT of 0.05, but a lower probability of correctly favoring $H_1$ (dose is unsafe) (19% vs. 40% at true DLT of 0.15) (Figure 1C). For the high toxicity scenario (D) $(p_1 = 0.50, p_2 = 0.30)$, the likelihood method performs better in declaring a dose acceptable (69% vs. 49% at true DLT of 0.30), but has almost 15% lower chances in identifying true toxicity when the DLT rate is 0.50 compared to the '3+3' model (Figure 1D).

For $k = 2$, the probability of weak evidence steps away from zero for all scenarios but (B) $(p_1 = 0.50, p_2 = 0.10)$, which displays the same behavior as for $k = 1$ (Figure 2B). For scenario (A) $(p_1 = 0.40, p_2 = 0.15)$, the probability of weak evidence peaks at 19% between the two hypotheses, represented in the figure by a dotted dark gray line. In this case, the likelihood method and the '3+3' algorithm match in correctly favoring $H_2$ (dose is acceptable) (Figure 2A).

Hypotheses (C) $(p_1 = 0.15, p_2 = 0.05)$ produce the highest probabilities of weak evidence for *k* = 2, i.e., 98% for a true DLT of 0.05 and 80% for a true DLT of 0.15 (Figure 2C). Surprisingly, this probability is almost the same as the probability of escalation from the '3+3' design. The likelihood method and the algorithm correctly declare a dose unsafe with very similar frequencies. For scenario (D) $(p_1 = 0.50, p_2 = 0.30)$, the most notable difference in statistical properties regards the probability of correctly favoring $H_1$. Hence, the likelihood method has only an 18% chance of declaring the dose unsafe for a true DLT of 0.50, compared to 83% from the '3+3' algorithm (Figure 2D).

For *k* = 8, the probability of weak evidence is above 90% for both $H_1$ and $H_2$ in scenarios (A) $(p_1 = 0.40, p_2 = 0.15)$ and (D) $(p_1 = 0.50, p_2 = 0.30)$ (Figure 3A and 3D). The likelihood method never favors $H_2$ (dose is acceptable) for all four sets of hypotheses and it has a zero probability of declaring a dose unsafe for $(p_1 = 0.50, p_2 = 0.30)$ (Figure 3D). This dramatic behavior of either always declaring a dose safe or unsafe when it is not the case can have serious implications in a dose-finding trial. We caution against using a *k* threshold of 8, and underline the small level of evidence that can be reached with only 6 patients per dose.

Since the category of weak evidence does not offer any clear guidance in making a decision, it can easily be subject to interpretability. One option is to consider it as not having enough information to conclude toxicity and to combine it with the evidence of favoring $H_2$ (safe dose). This is akin to "innocent until proven guilty": we assume a dose is acceptably safe until there is sufficient evidence to declare it unsafe. These results are displayed for all four sets of hypotheses, *k = 2* (Figure 4) and *k = 8* (Figure 5). With this action, for *k* = 2, the likelihood method and algorithm reach agreement for scenario (C) $(p_1 = 0.15, p_2 = 0.05)$, but the probability

of identifying an unsafe dose remains low at 18% (Figure 4C). Moreover, for the high toxicity hypotheses (D) $(p_1 = 0.50, p_2 = 0.30)$, the combined probabilities of weak and correctly declaring a dose safe increase to 92% for 0.15 true null DLT rate compared to 81% for $k = 1$ (Figure 4A). For scenario (D), the probability of declaring a dose safe reaches almost 96% for the likelihood method, but the performance of identifying toxicity is still poor (Figure 4D).

**5. Discussion**

In this paper we have explored the operating characteristics of the standard '3+3' design in a likelihood framework based on the evidential paradigm. We offer a probabilistic support for analyzing its behavior for different sets of hypotheses, levels of evidence, and true (acceptable and unacceptable) toxicity rates. Our likelihood approach has been developed to accommodate any combination of those three, with R functions available upon request. This way, any investigator (statistician or non-statistician) that intends to implement the standard algorithm has the opportunity of testing its properties under different conditions and selecting the hypotheses consistent with good performance characteristics at each dose level.

The evidential paradigm is an ideal setting for monitoring clinical trials with likelihood inference. Compared to the Neyman-Pearson methodology, in this framework the strength of evidence is quantified solely by the likelihood-ratio and it is amenable to sequential evaluation of the data. Moreover, the strength of evidence is dissociated from the probability of observing misleading evidence. In our simulation study we presented three likelihood-ratio thresholds: $k = 1$, 2, and 8. For the four sets of hypotheses considered, $k = 8$ is too stringent given the amount of evidence generated from a maximum of 6 patients per dose. In this case, the probability of weak evidence was over 80% and if combined with the probability of favoring $H_2$ it almost always

declared a dose safe even when it was truly toxic. The choice of $k = 2$ appears to be the most reasonable one, generating operating characteristics in good agreement with the '3+3' design. This benchmark also produces high probabilities of correctly favoring $H_1$ (dose is unsafe) and correctly favoring $H_2$ (dose is acceptable). These high probabilities stand especially for sets of hypotheses for which the unsafe DLT rate is greater than 0.30 and the acceptable DLT rate is less than 0.20, with a midpoint around 30%. For more extreme scenarios: low $(p_1 = 0.15, p_2 = 0.05)$ or high toxicity $(p_1 = 0.50, p_2 = 0.30)$, both likelihood method and the '3+3' design have less than 20% probability of identifying a toxic dose. This means that for extreme scenarios the '3+3' algorithm has no probabilistic support and it should not be used.

Our likelihood approach can be extended and implemented for cohorts of sizes other than 3 and 6. It is not uncommon for other cohort sizes to arise, either due to design or happenstance. In those cases, there are not commonly used rules for declaring dose safe or unsafe. The likelihood-ratio method with a fixed *k* and pre-stated hypotheses for the DLT rates allows consistent inferences to be made and evidence to be quantified regardless of cohort size. Table 3 illustrates one of these situations with a cohort of 5 patients at a dose level. For hypotheses $(p_1 = 0.40, p_2 = 0.15)$ and $k = 2$ as likelihood-ratio threshold, the method declares toxicity for 2 or more DLTs. In the same manner, one can experiment with different cohort sizes and *k* values and have a decision rule ready for any unexpected situation. In conclusion, our approach offers great potential in both phase I designs for identifying the highest acceptably safe dose and is akin to the sequential probability ratio test (Wald 1945).

**Acknowledgements**

The authors would like to thank Dr.Viswanathan Ramakrishnan (Department of Public Health Sciences, Medical University of South Carolina) for the insightful comments during the preparation of this manuscript. We also thank Dr. Michael I. Nishimura (Department of Surgery, Loyola University Chicago) for the funding grant and clinical expertise.

## References


1. Babb J., Rogatko A., and Zacks S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* **17**, 1103-1120.
2. Berger J. O., and Wolpert R. L. (1988). *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics.
3. Blume J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine* **21**(17), 2563-2599.
4. Brownlee K. , Hodges J., and Rosenblatt M. (1953) The up-and-down method with small samples. *Journal of American Statistical Association* **48**, 262-277.
5. Cheung Y. K., and Chappell R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56**(4), 1177-1182.
6. Forster M. R. (2006). Counterexamples to a likelihood theory evidence. *Mind Mach* **16**, 319-338.
7. Goodman S. N., Zahurak M. L., and Piantadosi S. (1995) Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* **14**(11), 1149-1161.
8. Hacking I. (1965). *Logic of statistical inference*. London: Cambridge University Press.
9. Iasonos A., Wilton A. S., Riedel E. R., Seshan V. E., and Spriggs D. R. (2008) A comprehensive comparison of the continual reassessment method to the standard 3+3. *Clinical Trials* **5**(5), 465-477.
10. Ivanova A. (2006). Escalation, group and A + B designs for dose-finding trials. *Statistics in Medicine* **25**(21), 3668-3678.
11. Ji Y., and Wang S. J. (2013). Modified Toxicity Probability Interval Design: A Safer and More Reliable Method. *Journal of Clinical Oncology* **31**(14), 1785-91.
12. Lin Y., and Shih W. J. (2001). Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics* **2**(2), 203-215.
13. Moller S. (1995). An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in Medicine* **14**(9-10), 911-922.
14. O'Quigley J., Pepe M., and Fisher L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* **46**(1), 33-48.
15. O'Quigley J. (2006). Theoretical study of the continual reassessment method. *Journal of Statistical Planning and Inference* **136**, 1765-1780.
16. Piantadosi S., Fisher J. D., and Grossman S. (1998). Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemotherapy and Pharmacology Journal* **41**(6), 429-436.
17. R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for Statistical Computing.
18. Reiner E., Paoletti X., and O'Quigley J. (1999) Operating characteristics of the standard phase I clinical trial design. *Computational Statistics and Data Analysis* **30**(3), 303-315.
19. Robbins H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *Annals of Mathematical Statistics* **41**(5), 1397-1409.
20. Rogatko A., Schoeneck D., Jonas W. et al. (2007). Translation of innovative designs into phase I trials. *Journal of Clinical Oncology* **25**(31), 4982-4986.
21. Royall R. (1997). *Statistical Evidence: A likelihood paradigm*. London: Chapman & Hall/CRC.
22. Royall R. (2000). On the probability of observing misleading statistical evidence. *Journal of American Statistical Association* **95**(451), 760-768.
23. Simon R., Freidlin B., Rubinstein L., Arbuck S. G. et al. (1997) Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute* **89**(15), 1138-1147.
24. Storer B. E. (1989). Design and Analysis of Phase I Clinical Trials. *Biometrics* **45**(3): 925-937.
25. Tighiouart M., Rogatko A., and Babb J. S. (2005). Flexible Bayesian methods for cancer phase I

clinical trials. Dose escalation with overdose control. *Statistics in Medicine* **24**(14), 2183-2196.
26. Wald A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**(2): 117-186.
27. Yuan Z., Chappell R., and Bailey H. (2007). The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics* **63**(1): 173-179.