Design and Analysis of Phase I Clinical Trials
Author(s): Barry E. Storer
Source: *Biometrics*, Vol. 45, No. 3 (Sep., 1989), pp. 925-937
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/2531693
Accessed: 20/01/2011 06:50

Design and Analysis of Phase I Clinical Trials

Barry E. Storer

Departments of Statistics and Human Oncology, University of Wisconsin-Madison,
420 North Charter St., Madison, Wisconsin 53706, U.S.A.

SUMMARY

The Phase I clinical trial is a study intended to estimate the so-called maximum tolerable dose (MTD) of a new drug. Although there exists more or less a standard type of design for such trials, its development has been largely ad hoc. As usually implemented, the trial design has no intrinsic property that provides a generally satisfactory basis for estimation of the MTD. In this paper, the standard design and several simple alternatives are compared with regard to the conservativeness of the design and with regard to point and interval estimation of an MTD (33rd percentile) with small sample sizes. Using a Markov chain representation, we found several designs to be nearly as conservative as the standard design in terms of the proportion of patients entered at higher dose levels. In Monte Carlo simulations, two two-stage designs are found to provide reduced bias in maximum likelihood estimation of the MTD in less than ideal dose–response settings. Of the three methods considered for determining confidence intervals—the delta method, a method based on Fieller's theorem, and a likelihood ratio method—none was able to provide both usefully narrow intervals and coverage probabilities close to nominal.

1. Introduction

Although the number of Phase I and Phase II clinical trials involving new cancer treatments far surpasses the number of Phase III trials, relatively little attention has been given to their design and analysis in the statistical literature. This is especially true of Phase I trials, which are trials designed to estimate the so-called maximum tolerable dose (MTD) of new therapeutic agents.

Some aspects of Phase I design have been described in a review article by Geller (1984). Typically, small groups of patients are treated at gradually escalating doses of the drug in question. Escalation continues until the number of patients experiencing a given degree of toxicity meets some set criterion, at which point the stopping dose or the next lower dose is taken as the MTD. With rare exception (Brown and Hu, 1980), there is no further analysis of the data. Even when other designs have been used (Schneiderman, 1965; Carbone et al., 1965), they have not involved objective estimation of MTD using a statistical model or consideration of sampling error. Although Anbar (1984) has considered the possible use of stochastic approximation methods in the Phase I setting, these methods are generally more appropriate for continuous, rather than binary response criteria, particularly at small sample sizes. However, some of the designs discussed here do, in a loose sense, involve elements of stochastic approximation.

A strict quantitative definition of the MTD is rarely acknowledged in clinical protocols, though it should be taken to mean some percentile of a tolerance distribution with respect to some objective definition of clinical toxicity. Because it seems closest to what is implicitly intended in the protocols with which we are familiar, the 33rd percentile is used to define

---

MTD throughout this paper. So defined, the estimation of MTD is not a novel statistical problem and is considerd at length in many standard texts (e.g., Finney, 1978). The Phase I trial, however, has features that are not adequately addressed in the bioassay literature. These include, for example, the relatively small number of experimental subjects available for treatment, the relatively long time it takes to evaluate each subject, the ethical requirement to approach the MTD conservatively, the relative subjectiveness of the response (toxic/nontoxic instead of alive/dead), the heterogeneity of the subject population, and the possible difficulty in classifying response due to early dropout for reasons unrelated to toxicity.

Although the traditional design, as described in Section 2, conforms to reasonable common sense, it does not have any intrinsic property that makes it stop close to the 33rd (or any other) percentile. Consequently, several simple alternatives are considered and described in Section 2. In Section 3 a Markov chain representation is used to evaluate the conservativeness of the designs. In Section 4, Monte Carlo simulations are used to evaluate the performance of maximum likelihood estimates (MLEs) of the 33rd percentile when these designs are used in the small-sample setting. Similarly, the small-sample performance of three standard methods of interval estimation is evaluated in Section 5.

2. Designs Considered

From a clinician's perspective, an optimal design would be one in which the MTD is defined by the dose at which the trial stops, as in current practice. Our view, however, is to consider the design a prescription for sampling, preferably a very simple one. If used correctly, the prescription should be expected to generate sample points in a reasonably efficient but conservative manner; actual estimation of MTD then follows from analysis of the observed response data. Such a combination of design and analysis should be more robust to the vagaries of patient treatment in a clinical setting, wherein the clinical protocol (design), may not be followed exactly.

Four single-stage designs are defined here:

Design A (traditional)　Groups of three patients are treated. Escalation occurs if no toxicity is observed in all three; otherwise, an additional three patients are treated at the same dose level. If only one of six has toxicity, escalation again continues; otherwise, the trial stops.

Design B　Single patients are treated. The next patient is treated at the next lower dose level if a toxic response is observed, otherwise at the next higher dose level.

Design C　Similar to design B, except that two consecutive nontoxic responses must be obtained before escalation occurs, whereas de-escalation occurs whenever a toxic response is seen.

Design D　Groups of three patients are treated. Escalation occurs if no toxicity is seen and de-escalation if more than one patient has toxicity. If a single patient has toxicity, then the next group of three is treated at the same dose level.

Designs B through D are variations on "up and down" schemes described by Wetherill (1963) and Wetherill and Levitt (1965); they are implemented here with fixed sample sizes. Design B is not evaluated as a single-stage design by itself; rather it is included to define the two-stage designs described below. It is not a conservative design and will tend to sample around the 50th, instead of the 33rd, percentile. Design D could be considered a discretized version of the Robbins–Monro procedure (Robbins and Monro, 1951) with

equally spaced dose levels. During escalation, the dose X_j to be used at step j is given by $X_j = X_{j-1} + \Delta \cdot \text{sign}(P - \hat{P}_{j-1})$, where $\hat{P}_{j-1}$ is the observed fraction of toxic responses in the previous group of patients, P is the target fraction, and Δ is the space between doses.

None of the designs above could be expected to perform well in an arbitrary dose–response setting when implemented with fixed sample sizes. For this reason, we propose two two-stage designs, denoted BC and BD, which combine single-stage designs. The first stage follows design B until the first toxic response occurs. From the point at which the next patient is entered at the next lower dose level, the second stage design is implemented, again with fixed sample size.

All of the designs, both single-stage and two-stage, are implemented with equally spaced (presumably on a logarithmic scale) dose levels fixed in advance. In practice, dose levels may follow other schemes, the most popular being the modified Fibonacci sequence described by Schneiderman (1967), where dose levels are incremented by the diminishing multiples 2, 1.67, 1.5, 1.33, 1.33, . . . ; however, these dose increments are also equally spaced on a log scale after the first four dose levels.

The designs are compared in dose–response settings determined by three logistic curves, wherein the probability of toxic response at dose X is given by

$$P(X) = \Pr[Y = 1 \mid X] = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)},$$

where Y is the dichotomous random variable associated with response. These curves are illustrated in Figure 1. The first curve is intended to represent a relatively ideal setting, i.e., a steep dose–response and a starting dose close to the target percentile. The second curve, which probably more closely represents the typical situation, is shallower and the starting dose is several dose levels below the target. These two curves are examined with a fixed starting dose and sample sizes ranging from 12(6)36. The third curve represents the most difficult dose–response setting—a shallow curve with (relatively) closely spaced dose levels and a nonnegligible probability of a toxic response at all dose levels. This curve is examined using a fixed sample size but with the starting dose level varied from 1(1)8.
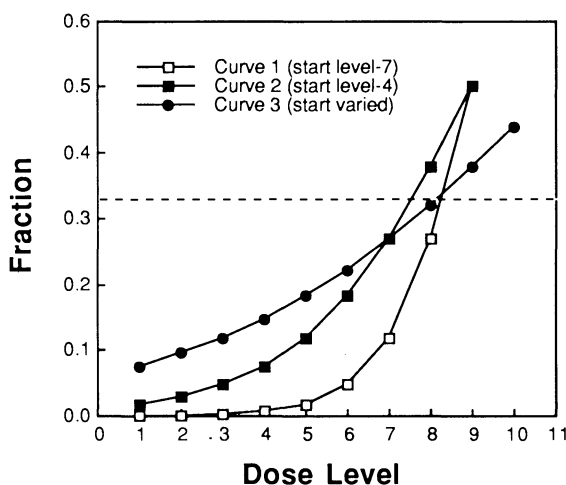

Figure 1. The three logistic dose–response curves used to compare designs. Symbols indicate the actual dose levels, which extend with equal spacing up to level 16.

3. Markov Chain Representation

3.1 *Single-Stage Designs*

The probabilistic properties of each of the designs considered here can be explicitly evaluated numerically through a representation as a discrete-time Markov chain with stationary transition probabilities. In designs B and D, each state in the Markov chain corresponds directly to the treatment of a patient or group of patients at a particular dose level. In designs A and C, two states are required for each dose level in order to preserve the property of stationarity: One state corresponds to the initial patient or group of patients evaluated at a dose level; another state corresponds to the second patient or group that is evaluated if no toxicity occurs in the first. Additionally, design A requires an absorbing state corresponding to stopping the trial.

Under a specific dose–response assumption, the transition probabilities for any of the designs are simple binomial probabilities that are easily evaluated and summarized in the transition probability matrix, **P**. Although actual clinical trials could be conducted with an essentially open-ended set of dose levels, for purposes of computation it is necessary to specify a restricted number of states in the process. This does not compromise the open-ended design, since the range of states is extended such that the probability that any of the extreme states is reached is negligibly small.

The (i, j)th element of **P**, P_{ij}, gives the probability that a trial in state i will move to state j in the next transition. Based on standard properties of this matrix, the probability that a trial in state i will move to stage j in k transitions is given by the (i, j)th element of $\mathbf{P}^k$, where $\mathbf{P}^k = \mathbf{P} \cdot \mathbf{P}^{k-1}$, $k = 2, 3, \ldots$. For a trial that starts in state i, the expected number of patients treated in state j after k transitions is given by $N_{ij}^k = m(\text{Ind}[j = i] + \sum_{h=1}^{k} P_{ij}^h)$, where m is the number of patients treated at each step, P_{ij}^k is the (i, j)th element of $\mathbf{P}^k$, and Ind is the indicator function. Depending on the design used, the expected number of patients treated at a particular dose level will be the sum of one or two corresponding N_{ij}^k.

3.2 *Two-Stage Designs*

The two-stage designs BC and BD are evaluated similarly, except that the distribution of stopping states from the first stage yields a distribution of initial states for the second-stage Markov chain, leading to a slightly modified computation for the expected numbers of patients treated in a given state after k transitions as $N_{Ij}^k = \sum_i \Pr[I = i] \cdot N_{ij}^k$, where the subscript I denotes the random distribution of initial states and N_{ij}^k is as above. The expected number of patients treated at a given dose level is then the sum of the appropriate N_{ij}^k from the first stage and the one (design BD) or two (design BC) appropriate N_{Ij}^k from the second stage.

3.3 *Results*

Figure 2 presents the expected fractions of patients that would be treated at dose levels above the 50th percentile of the tolerance distribution for the three dose–response situations. The choice of dose levels above the 50th percentile is arbitrary and merely provides a point of reference. Although the two-stage designs are less conservative than either the standard design or their single-stage counterparts, the absolute level of conservativeness does not appear unreasonable in light of the increased sampling in the vicinity of the target percentile.
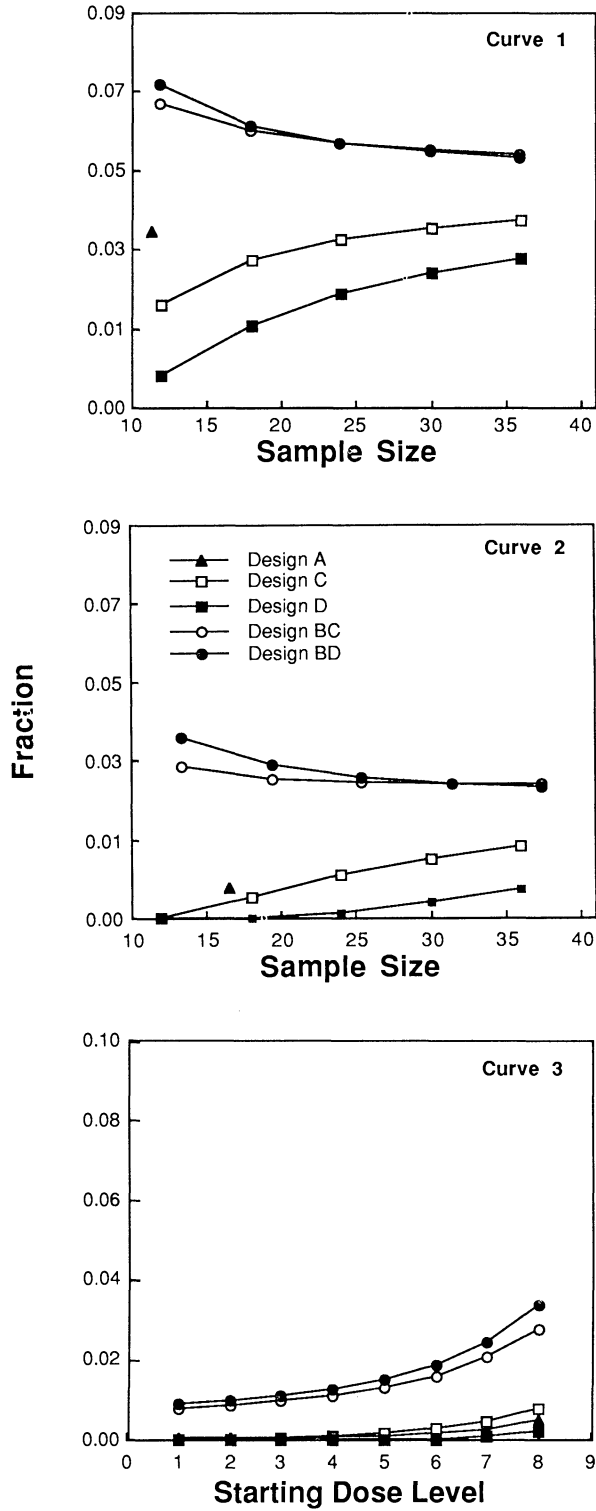
Figure 2. Expected fraction of patients treated above the 50th percentile for the various designs. Results for A, BC, and BD are plotted as a function of expected sample size.

4. Estimation of MTD

In the standard design, the actual stopping dose or the next lower dose is usually taken as the estimate of MTD, perhaps depending on the actual amount and degree of toxicity observed in the last group of evaluated patients. Of course, no dose level need necessarily be close to the true MTD, except fortuitously. In a comparison of the different designs, the actual stopping dose level is taken to be the estimate of MTD for design A. The exact distribution of stopping dose levels is readily computed in a given dose–response setting using the Markov property.

The MLE of MTD defined by any percentile (X_p, where $\Pr[Y = 1 \mid X_p] = p$) is given by $\hat{X}_p = (k_p - \hat{\alpha})/\hat{\beta}$, where $k_p = \mathrm{logit}(p)$, and $\hat{\alpha}$ and $\hat{\beta}$ are MLEs of the logistic parameters obtained via a standard Newton–Raphson maximization scheme. Although this method is usually well behaved, in small samples it is possible that $\hat{\beta} = 0$ (which can be detected analytically prior to maximization, simply by observing whether the score equation for β is satisfied by 0) or that $\hat{\beta} < 0$, which results in nonsensical estimates of the MTD. It is also possible that strict convergence cannot be obtained with respect to $\hat{\alpha}$ and $\hat{\beta}$ (which can also be checked prior to maximization). This situation occurs when the log-likelihood can be maximized by putting unit mass at a single point, above which all responses (if any) are toxic and below which all responses (if any) are nontoxic. In this case, $\hat{X}_p$ is not unique and is the same for any p.

Although MLEs of the logistic parameters, and hence of the MTD, have desirable large-sample properties, one would by no means expect this to be true in the small-sample setting encountered here, especially when combined with the stochastic nature of the sampling schemes. Hence, a series of Monte Carlo experiments using the previously described dose-response curves was carried out to examine the performance of MLEs of the MTD in combination with the various trial designs.

In simulated clinical trials, the outcome for each patient was determined by generating a pseudorandom U(0, 1) deviate using a standard IMSL subroutine. Response was taken to be toxic if the deviate generated was less than or equal to the underlying $P(X)$ determined by the curve under consideration at the current dose level. Sample sizes for the first two curves varied from 12(6)36, and for the two-stage designs applied to the size of the second stage. For the third curve the sample size was fixed at 24 and the starting dose level was varied from 1(1)8 (see Figure 1). Each combination of curve, design, and sample size or starting dose was simulated 5,000 times, using initial seeds chosen independently from a table of random digits.

For purposes of evaluating the designs, attention will be restricted to successful simulations; i.e., those where $0 < \hat{\beta} < \infty$. The probability of obtaining a successful result thus becomes one basis for comparing the designs. Table 1 gives the fraction of 5,000 replications that were successful in each of the simulations. Except for the single-stage designs in the very smallest sample sizes, all of the alternative designs yield useable estimates a large fraction of the time. Unfortunately, the application of the logistic model to the sample data generated by design A does not give convergent estimates a usefully large fraction of the time and will not be considered further.

Within the group of successful simulations, MLEs of MTD from designs C, D, BC, and BD and empirical estimates from design A are compared on the basis of median (Figure 3) and lower and upper quartiles (Table 2) of estimates. These are used rather than mean or mean squared error because a small fraction of estimates may be extremely large (due to division by $\hat{\beta} \approx 0$) and the usual measures are sensitive to these outliers. Since the dose scale itself is meaningless, results here and elsewhere are expressed in terms of the corresponding percentile of the underlying tolerance distribution.

Results using curve 1 were very similar for all of the alternative designs. At sample sizes of 18 or above all showed very little bias and similar interquartile ranges. For design A,

Table 1
Fraction of successful trials for various designs in three dose–response settings

Curve 1 (starting dose level = 7)

Sample size[a]	12	18	24	30	36
Design					
A	.310				
C	.603	.849	.944	.978	.992
D	.399	.737	.884	.947	.972
BC	.793	.916	.963	.989	.996
BD	.652	.837	.935	.960	.984

[a] Expected sample size for design A is 11.4. Expected additional sample size from B stage for designs BC and BD is 2.9.

Curve 2 (starting dose level = 4)

Sample size[b]	12	18	24	30	36
Design					
A	.465				
C	.469	.843	.959	.987	.996
D	.172	.603	.850	.938	.974
BC	.854	.950	.988	.996	.999
BD	.689	.875	.953	.978	.993

[b] Expected sample size for design A is 16.5. Expected additional sample size from B stage for designs BC and BD is 4.4.

Curve 3 ($n = 24$[c])

Starting dose level	1	2	3	4	5	6	7	8
Design								
A	.623	.578	.555	.540	.538	.307	.247	.176
C	.871	.905	.923	.938	.957	.965	.975	.980
D	.725	.768	.807	.835	.860	.888	.898	.912
BC	.961	.962	.972	.977	.980	.984	.981	.986
BD	.899	.906	.915	.926	.927	.939	.936	.941

[c] Expected sample size for design A ranges from 21.5 to 8.7. Expected additional sample size from B stage for designs BC and BD ranges from 5.8 to 2.9.

38.5% of trials stop at the dose level below the true MTD [$P(X) = .269$] and 40.4% at the dose level above [$P(X) = .5$]. Results from curves 2 and 3, on the other hand, reveal that the designs can perform very differently in less than ideal settings. It is also clear that the two-stage designs have less bias than single-stage designs for a given sample size, i.e., the additional patients treated in the first stage contribute more to reducing bias than the same number added to the single-stage version of the same design, at least on average. Design BD in particular has negligible bias even with sample sizes as small as 18 in the second stage. Although the performance of even the two-stage designs falls off somewhat at the extreme starting doses used in curve 3, it does so to a much lesser extent that the others. A minor improvement in the variability of the estimates is also obtained with the two-stage designs, though only at the larger sample sizes.

5. Confidence Intervals for the MTD

One drawback of the usual procedures for estimating MTD is that they make no allowance for sampling error in the estimate. In the setting of maximum likelihood estimation, several methods are available for computing confidence intervals, any of which might prove

Table 2
True percentile associated with lower and upper quartiles of estimates of MTD *for various designs in three dose–response settings**

Curve 1 (starting dose level = 7)

Sample size[a]	18		24		30		36	
Design								
A	.269	.500						
C	.235	.402	.255	.404	.264	.396	.271	.393
D	.260	.419	.269	.404	.278	.401	.283	.398
BC	.253	.409	.267	.400	.274	.398	.274	.389
BD	.273	.421	.279	.406	.280	.402	.285	.397

[a] Expected sample size for design A is 11.4. Expected additional sample size from B stage for designs BC and BD is 2.9.

Curve 2 (starting dose level = 4)

Sample size[b]	18		24		30		36	
Design								
A	.182	.378						
C	.204	.385	.227	.386	.244	.380	.255	.380
D	.200	.348	.229	.386	.251	.391	.261	.387
BC	.241	.395	.255	.389	.265	.386	.271	.383
BD	.248	.415	.267	.404	.279	.400	.284	.397

[b] Expected sample size for design A is 16.5. Expected additional sample size from B stage for designs BC and BD is 4.4.

Curve 3 (*n* = 24[c])

Starting dose level	1		3		5		7	
Design								
A	.148	.269	.148	.269	.182	.321	.269	.378
C	.191	.339	.208	.333	.227	.344	.253	.359
D	.172	.302	.202	.317	.235	.334	.275	.360
BC	.231	.370	.237	.369	.246	.366	.263	.370
BD	.235	.374	.244	.372	.258	.378	.283	.389

[c] Expected sample size for design A ranges from 21.5 to 8.7. Expected additional sample size from B stage for designs BC and BD ranges from 5.8 to 2.9.

* Results for design A are based on the actual stopping dose from 5,000 simulations. Results for other designs are based on maximum likelihood estimates in successful fraction of 5,000 simulations.

satisfactory in large samples, but not necessarily in the small sample sizes under consideration here. For this reason we examined the small-sample properties of three types of confidence intervals: (1) intervals derived using the delta method, (2) intervals based on Fieller's theorem, and (3) intervals based on a likelihood ratio criterion. Since the properties of the intervals varied considerably by method, we focus on these differences rather than on a comparison among designs of the results of a specific method of interval computation.

5.1 *Methods Compared*

The simplest interval is given by a straightforward application of the multivariate delta method. Letting V_α and V_β denote the usual asymptotic variance estimates for $\hat{\alpha}$ and $\hat{\beta}$, respectively, and $V_{\alpha\beta}$ the asymptotic covariance estimate, then a large-sample estimate of the variance of $\hat{X}_p$ is given by $V_d = [V_\alpha + 2\hat{X}_p V_{\alpha\beta} + \hat{X}_p^2 V_\beta]/\hat{\beta}^2$. Therefore, an approximate $100(1 - \alpha)\%$ confidence interval for X_p is given by

$$\{X_p: \hat{X}_p - Z_{\alpha/2} V_d^{1/2} \leq X_p \leq \hat{X}_p + Z_{\alpha/2} V_d^{1/2}\},$$
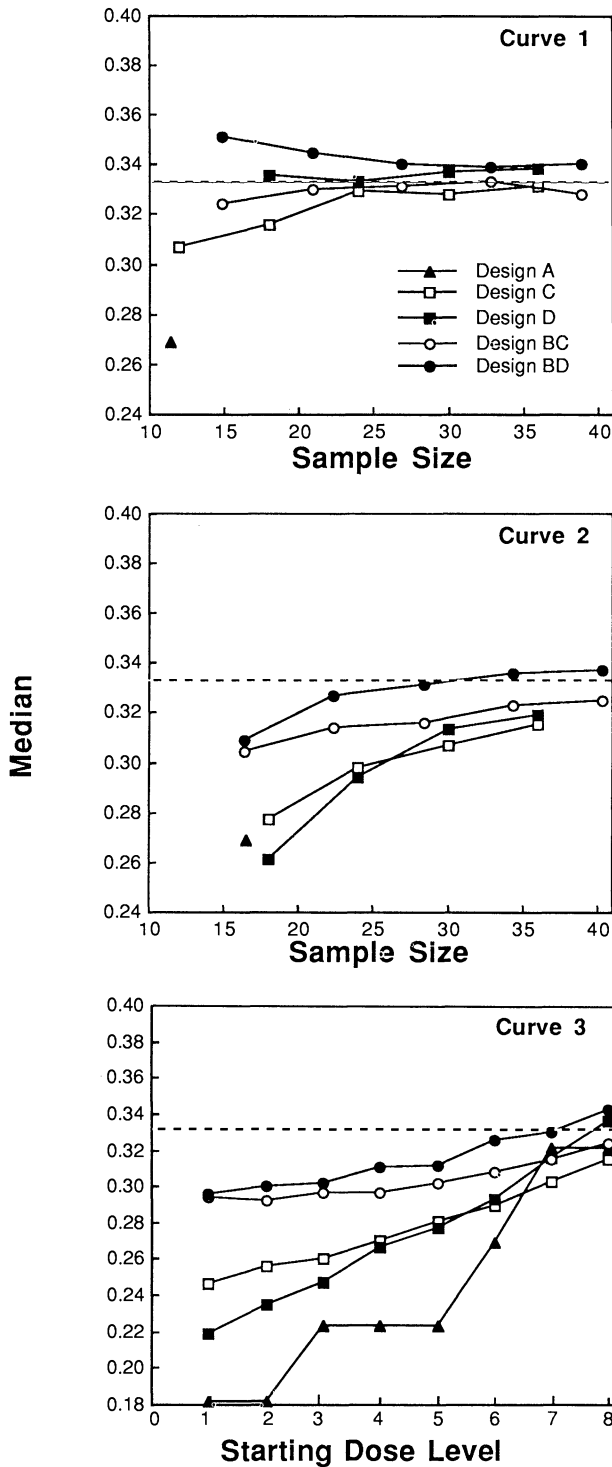
Figure 3. Median of estimates of MTD (33rd percentile) for the various designs, expressed as the true percentile of the underlying dose–response curve. Results for A, BC, and BD are plotted as a function of expected sample size. Results for C, D, BC, and BD are plotted where the fraction of successful simulations (out of 5,000) was greater than .85.

where $Z_{\alpha/2}$ is the upper $\alpha/2$ critical value of the standard normal distribution. This interval has a simple form familiar to most medical investigators and is easily computed with output available from standard programs that fit a logistic curve. It has finite endpoints whenever $\hat{X}_p$ is finite and unique.

A slightly more complicated computation is based on the quantity $\hat{\alpha} - k_p + X_p\hat{\beta}$, which is asymptotically normally distributed with mean 0 and variance estimated by $V_f = V_d\hat{\beta}^2$. An approximate $100(1 - \alpha)\%$ confidence interval for X_p is given by

$$\{X_p: |\hat{\alpha} - k_p + X_p\hat{\beta}|^2 \leq X_\alpha^2 V_f\},$$

where X_α^2 is the upper α percentile of the chi-squared distribution with 1 degree of freedom. We will refer to this as Fieller's interval, which is not necessarily symmetric about $\hat{X}_p$. The endpoints of a closed interval are obtained as a solution of a quadratic equation in X_p, provided that $Z_{\alpha/2}|\hat{\beta}|/V_\beta^{1/2} < 1$. When this is not the case, which corresponds to failing to reject a level α test of the null hypothesis $\beta = 0$, based on $\hat{\beta}/V_\beta^{1/2}$, then the range of values of X_p satisfying the inequality will either be infinite (the entire real line) or will comprise two disjoint half-lines, the area between which is exclusive. Though the latter outcome is often referred to as Fieller's paradox, it is always the case that one of the half-intervals will include X_p, and this half-interval is the natural choice for the confidence interval. The infinite ends of the half-intervals could be thought of as connected at a point corresponding to $\beta = 0$. The discarded half-interval corresponds to values such that $\beta < 0$, which in the present context are nonsensical.

A final confidence interval is based on a likelihood ratio criterion and comprises

$$\{X_p = (k_p - \alpha)/\beta: 2[L(\hat{\alpha}, \hat{\beta}) - L(\alpha, \beta)] \leq X_\alpha^2\},$$

where L is the log-likelihood function for the logistic model. The confidence interval is determined by a direct search procedure essentially equivalent to that of Williams (1986), except that the search is restricted to $\beta > 0$.
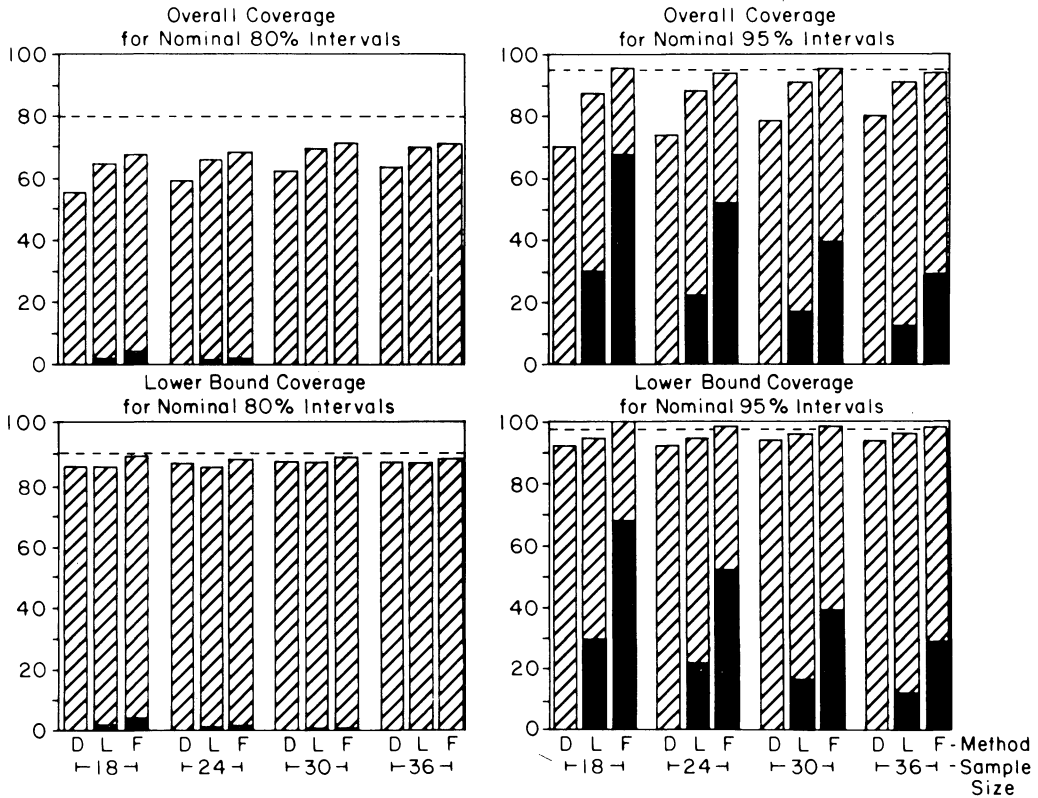
5.2 *Results*

None of the intervals proved completely satisfactory with respect to the criteria of coverage probability and width. Though performance differed among designs, these differences were in general not large. The results presented here are those obtained with design BC and curve 2.

Figure 4 presents the actual coverage probabilities for the three methods for nominal 80% and 95% confidence intervals. These are given both for the overall coverage and for the coverage of the lower bound only, which should be 90% and 97.5%, respectively. Although Fieller's interval clearly performs the best overall, the absolute level of performance is good only for the 95% intervals. Coverage for delta method intervals is markedly anticonservative, even at the largest sample sizes. Likelihood ratio intervals are intermediate in performance, but closer to Fieller's intervals than to delta method intervals.

It is interesting that the lower-bound coverage probabilities are closer to the nominal level than are the overall coverage probabilities, and that there are much smaller differences among the methods for these probabilities. This is true even after taking into account the halving of the absolute amount of miscoverage that would occur if the miscoverage were symmetric. Since the point estimate of MTD has negligible bias, particularly at the larger sample sizes, it is apparent that the estimated variability of $\hat{X}_p$ must tend to be smaller when it is below X_p than when it is above.

Unfortunately, many of the covering intervals for the two most accurate methods were infinite, especially for the 95% intervals. This fraction is indicated by the solid portion of the bars in Figure 4. For this coverage level, the majority of Fieller's intervals were infinite

Figure 4. Coverage probabilities for nominal 80% and 95% confidence intervals for the MTD (33rd percentile) using three methods (D: Delta; L: Likelihood ratio; F: Fieller's). Results are for design BC and curve 2. Solid portions of the bars indicate the fraction of infinite intervals or bounds.

Table 3
True percentile associated with median lower and upper bounds of confidence intervals for the MTD *computed by three methods[a]*

Sample size[b]	18		24		30		36	
80% confidence interval								
Delta method	.218	.402	.229	.406	.239	.407	.246	.408
Likelihood ratio	.227	.471	.234	.463	.242	.454	.248	.446
Fieller's	.213	.478	.225	.466	.236	.456	.242	.447
95% confidence interval								
Delta method	.188	.452	.196	.452	.205	.453	.211	.451
Likelihood ratio	.155	.885	.172	.787	.182	.709	.193	.661
Fieller's	.000	1.00	.000	1.00	.106	.903	.146	.729

[a] Based on successful fraction of 5,000 replications using design BC and curve 2.
[b] Expected additional sample size from B stage is 4.4.

except at the largest sample size; the proportion of infinite likelihood ratio intervals, though relatively smaller, was still undesirably high. Although the results considered here are restricted to simulations giving a finite and determinate estimate of MTD, it should be noted that in situations where the estimate of MTD is indeterminate, the likelihood ratio method can yield finite intervals, whereas Fieller's intervals (and delta method intervals) are always infinite.

The median lower and upper bounds of confidence intervals for the three methods are presented in Table 3. As would be expected from the above, delta method intervals are consistently narrowest and Fieller's intervals consistently widest. The differences are substantial for the 95% intervals.

6. Discussion

As was remarked earlier, the literature treating statistical aspects of Phase I clinical trials is quite small. One can find occasional discussions concerning the choice of starting dose or the spacing of dose levels (e.g., Schneiderman, 1965; Goldsmith, Slavik, and Carter, 1975; Penta et al., 1979), but none of which we are aware that consider the design of the trial and estimation of MTD as a whole. Although these results are not directly relevant to some Phase I settings (for example, where MTD is defined as a much different percentile), they do suggest that simple types of designs can be combined with relatively standard analytic techniques to provide a more quantitatively rational basis for estimation than that usually employed.

It is not surprising that the two-stage designs perform better than the single-stage designs. Although they are still not entirely independent of the unknown underlying dose–response curve, they are relatively less sensitive to an unlucky choice of dose levels and perhaps in practice would prove satisfactory most of the time. Nevertheless, incorporation of a systematic basis for sequentially modifying the dose levels, as in Wu (1985), might be desirable. Although his methods are not directly applicable to the clinical trial setting, some modifications might be useful in determining, for example, the sample size and dose levels for a third stage of the trial, if a two-stage design did not give satisfactory results with the chosen number of patients.

Though we have seen no data precisely defining a tolerance distribution for human subjects with a toxic drug, the logistic model used here is easy to work with and likely to be reasonably robust with respect to the actual shape of the distribution, provided that the MTD is not chosen to be too extreme a percentile. McLeish and Tosh (1983) discuss methods of conservatively estimating very small percentiles, but using sample sizes far in excess of those realizable in a Phase I trial. Although additional flexibility could be achieved by incorporating an additional parameter in the logistic model to accommodate asymmetry (as in Wu, 1985), it seems unlikely that this would have much impact on the estimates used here (where MTD is defined as the 33rd percentile) if the design has successfully concentrated sampling in the lower half of the distribution.

The most unsatisfactory part of this work has been with respect to the provision of confidence intervals for the MTD; perhaps no method can be expected to perform well in this context with small sample sizes. Although the lower bound of simple delta method intervals always exists (when the MLE is determinate) and did not seem to perform too badly here in terms of coverage, one would not necessarily expect this result to hold in other settings. Williams (1986) reports quite reasonable performance by likelihood ratio intervals for the 50th percentile in sample sizes comparable to those considered here; however, we suspect that the dose–response situations studied were more ideal, i.e., with sampling uniformly spaced over a range symmetric about the true median.

Finally, we suggest that additional improvement in Phase I trial design could be obtained by incorporating randomization into the design. Given the heterogeneity of the typical population of patients eligible for a Phase I study, it is obvious that decisions to enter or not enter patients at particular dose levels, based on the anticipated response, could bias the outcome no matter how sophisticated the design. Since the designs considered here determine the treatment assignments sequentially, they are not readily amenable to randomization. This will be one area of future work.

RÉSUMÉ

L'essai clinique de phase I est une étude destinée à estimer la dose maximale tolérée (D.M.T.) d'une nouvelle molécule. Bien qu'il existe plus ou moins un schéma type pour de tels essais, son développement s'est fait largement de manière improvisée. Tel qu'il est habituellement appliqué, ce schéma ne garantit pas toujours une estimation satisfaisante de la D.M.T. Dans cet article, nous comparons ce schéma standard et plusieurs alternatives simples en regard (a) du conservatisme du schéma, (b) de l'estimation ponctuelle et par intervalles de la D.M.T. (percentile 33%) sur de petits échantillons. En utilisant une représentation par un processus de Markov, plusieurs schémas se révèlent presque aussi conservatifs que le schéma standard au sens de la proportion de malades inclus à des doses supérieures. Des simulations montrent que deux schémas à deux étapes permettent de réduire le biais dans l'estimation du maximum de vraisemblance de la D.M.T. Des trois méthodes étudiées pour déterminer des intervalles de confiances—la méthode delta, un méthode reposant sur le théorème de Fieller et la méthode du rapport de vraisemblance—aucune ne fournit à la fois des intervalles suffisamment étroits pour être utilisables et des probabilités de recouvrement proches du niveau nominal.

REFERENCES

Anbar, D. (1984). Stochastic approximation methods and their use in bioassay and Phase I clinical trials. *Communications in Statistics—Theory and Methods* **13**, 2451–2467.
Brown, B. and Hu, M. (1980). Setting dose levels for the treatment of testicular cancer. In *Biostatistics Casebook*, R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses (eds), 123–152. New York: Wiley.
Carbone, P. P., Krant, M. J., Miller, S. P., Hall, T. C., Shnider, B. I., Colsky, J., Horton, J., Hosley, H., Miller, J. M., Frie, E., and Schneiderman, M. (1965). The feasibility of using randomization schemes early in the clinical trials of new chemotherapeutic agents: Hydroxyurea. *Clinical Pharmacology and Therapeutics* **6**, 17–24.
Finney, D. J. (1978). *Statistical Methods in Biological Assays.* London: Griffin.
Geller, N. (1984). Design of Phase I and Phase II clinical trials in cancer: A statistician's view. *Cancer Investigation* **2**, 483–491.
Goldsmith, M. A., Slavik, M., and Carter, S. K. (1975). Quantitative prediction of drug toxicity in humans from toxicology in small and large animals. *Cancer Research* **35**, 1354–1364.
McLeish, D. and Tosh, D. (1983). The estimation of extreme quantiles in logit bioassay. *Biometrika* **70**, 625–632.
Penta, J. S., Rozencweig, M., Guarino, A. M., and Muggia, F. M. (1979). Mouse and large animal toxicology studies of twelve antitumor agents: Relevance to starting dose for Phase I clinical trials. *Cancer Chemotherapy and Pharmacology* **3**, 97–101.
Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **29**, 351–356.
Schneiderman, M. A. (1965). How can we find an optimal dose? *Toxicology and Applied Pharmacology* **7**, 44–53.
Schneiderman, M. A. (1967). Mouse to man: Statistical problems in bringing a drug to clinical trial, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **4**, 855–866. Berkeley, California: University of California Press.
Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society, Series B* **25**, 1–48.
Wetherill, G. B. and Levitt, H. (1965). Sequential estimation of points of a psychometric function. *British Journal of Mathematical and Statistical Psychology* **18**, 1–10.
Williams, D. A. (1986). Interval estimation of the median lethal dose. *Biometrics* **42**, 641–645.
Wu, C. F. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association* **80**, 974–984.