UNIVERSITY OF WISCONSIN
CLINICAL CANCER CENTER

BIOSTATISTICS

Technical Report # 54

November 1989

SMALL SAMPLE CONFIDENCE SETS FOR THE MTD IN A PHASE I CLINICAL TRIAL

Barry Storer

Department of Human Oncology
and
Department of Statistics

UNIVERSITY OF WISCONSIN-MADISON
Small Sample Confidence Sets for the MTD in a Phase I Clinical Trial

Barry E. Storer

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

Summary

Properties of procedures for interval estimation of the MTD in a Phase I clinical trial using a two-stage stochastic sampling scheme (Storer, 1989) are examined. Although the likelihood function for the data arising from such a scheme is identical to one arising under a fixed binomial sampling assumption, intervals based on the exact distribution of the usual large-sample statistics under this assumption do not offer improvement over unadjusted intervals. However, consideration of the distribution of these statistics based on the true stochastic sampling scheme can lead to the construction of intervals with correct coverage probabilities which do not depend on the true values of the model parameters. Membership in the confidence set can be evaluated by Monte Carlo simulation at a restricted maximum likelihood estimate of the nuisance slope parameter, despite strong upward bias from the up-and-down sampling scheme in its estimation. The lack of information in the small sample setting is reflected in a large proportion of confidence intervals which include infinite values for the MTD, especially when the dose response curve is shallow. Intervals based on a likelihood ratio criterion perform best in this regard.

Key words: Exact distributions; Interval estimation; Stochastic sampling; Up-and-down design.

1. Introduction and Background

Up-and-down designs have been employed for the purpose of estimating particular percentiles of a tolerance distribution. In Storer (1989) we described the possibility of their use in the setting of the Phase I clinical trial in cancer, where the percentile of interest is referred to as the MTD (Maximum Tolerable Dose) and is defined as the dose where a specified fraction of patients, $q_0$, experience dose-limiting toxicity. Assuming a logistic model for the underlying dose-response relationship, several stochastic sampling schemes in conjunction with maximum likelihood were used to estimate the MTD. A Monte Carlo evaluation of the
small-sample performance of these methods indicated that two two-stage designs provided a
reasonable basis for point estimation, but also showed that large-sample methods of interval
estimation were not generally suitable, due either to the small sample size, the nature of
the sampling scheme, or both. The purpose of this report is to examine the feasibility and
performance of exact inference procedures in the small-sample setting. In the present section
we review the basic setup and derive the likelihood for the model. In Section 2 we consider
the use of exact small-sample procedures for inference. Section 3 presents simulation results
which validate the performance of these procedures.

The two-stage designs proposed by Storer (1989), denoted ‘BC’ and ‘BD’, are imple-
mented using a set of fixed dose levels \( \ldots, a_{-1}, a_0, a_1, \ldots \), which in most applications are
assumed to represent an approximately equally spaced logarithmic transformation of the
original dose. We assume that the underlying probability of a toxic event for a patient treated
at a transformed dose \( x \) taking on one of the values \( a_i \) is given by \( \logit[p(x, \theta)] = \mu + \beta x \),
where \( \theta = (\mu, \beta)' \), and define the MTD as \( \gamma = (\pi - \mu)/\beta \), where \( \pi = \logit(q_0) \). In this report
we assume \( q_0 = 1/3 \). Also define the triplet \((y_i, x_i, n_i), i = 1, \ldots, m\), where, for the \( i \)th pa-
tient or group of patients, \( n_i \) is the number of patients treated, \( x_i \) is the transformed dose at
which they are treated, and \( y_i \) is the number of patients observed to have a toxic response. A
random number \( m_1 \) of single individuals will be entered in the first stage and a fixed number
\( m_2 \) of individuals or groups will be entered in the second stage, with \( m = m_1 + m_2 \). The
stages of the design ‘BD’, on which this report will focus, are described below:

‘B’ stage—Single patients are entered starting at a pre-specified dose level \( x_1 = a_0 \). If the current dose level is \( x_i = a_k \) then: if \( y_i = 0 \) and \( y_{i-1} = 0 \) (or \( i = 1 \))
then \( x_{i+1} = a_{k+1} \); if \( y_i = 1 \) and \( y_{i-1} = 1 \) (or \( i = 1 \)) then \( x_{i+1} = a_{k-1} \); otherwise
the first stage terminates and \( m_1 = i \). If \( x_{m_1} = a_k \) then the starting dose level
for the second stage is \( x_{m_1+1} = a_{k-1} \) or \( a_k \), depending on whether \( y_{m_1} = 1 \) or \( 0 \),
respectively.

‘D’ stage—Groups of three patients are entered starting at \( x_{m_1+1} \) as determined
by the first stage. If the current dose level is \( x_i = a_k \) then: if \( y_i = 0 \) then
\( x_{i+1} = a_{k+1} \); if \( y_i = 1 \) then \( x_{i+1} = a_k \); and if \( y_i > 1 \) then \( x_{i+1} = a_{k-1} \). That is,
one escalates, stays at the same level, or de-escalates as the observed fraction of
toxic responses is less than, equal to, or greater than \( 1/3 \).

The ‘B’ stage as described above is a slight modification of the original description (Storer,
1989), so that de-escalation will occur until the first non-toxic response. The reasons for this will be discussed in Section 3.

Define the vectors \( y = (y_1, \ldots, y_m)' \), \( x = (x_1, \ldots, x_m)' \), and \( n = (n_1, \ldots, n_m)' \) and let \( Y, X, \) and \( N \) denote the associated vectors of random variables. Under the sampling schemes described the distribution of \( Y_i \) depends conditionally only on the associated values of \( x_i \) and \( n_i \), and the latter are completely determined by the previous history of the sample. Thus, for \( j = 2, \ldots, m \) we have

\[
\Pr_\theta[Y_j = y_j | x_j, n_j, (y_i, x_i, n_i), i = 1, \ldots, j - 1] = \Pr_\theta[Y_j = y_j | x_j, n_j]
\] (1)

and

\[
\Pr_\theta[X_j = x_j, N_j = n_j | (y_i, x_i, n_i), i = 1, \ldots, j - 1] = 1.
\] (2)

By recursively conditioning on the previous history of the sample using (1) and (2) and assuming that \((x_1, n_1)\) are given, we have that

\[
\Pr_\theta[Y = y, X = x, N = n] \propto \prod_{i=1}^{m} [p(x_i, \theta)]^{y_i} [1 - p(x_i, \theta)]^{n_i - y_i}.
\] (3)

Thus, the unconditional probability of a given realization of \((Y, X, N)\) under the sampling scheme 'BD' is equal to the conditional probability of the same realization of \( Y \) under fixed binomial sampling from the logistic model, given the realization of \((X, N)\). Hence, maximum likelihood estimation of \( \theta \) can proceed in the usual manner. A similar result was noted by O'Quigley, Pepe and Fisher (1990) for their Continual Reassessment Method, and in fact holds for any sequential sampling scheme in which the conditional history of the sample path completely determines the next dose level to be visited and the number of patients to be evaluated.

Figure 1 presents some results regarding point estimation of \( \gamma \) and \( \beta \) from a series of Monte Carlo simulations described in detail in Section 3. Though point estimation is not the major focus of this paper, these results confirm the very reasonable performance of \( \hat{\gamma} \) over a much wider range of settings than previously examined. They also reveal a rather striking upward bias in \( \hat{\beta} \), particularly when the true value of \( \beta \) is relatively small. The impact of this bias on interval estimation will be considered in the next section.

2. Interval Estimation

We described in Storer (1989) a Markov chain representation of various stochastic sampling schemes that might be employed in the Phase I trial setting. For the designs mentioned
here, the states of the Markov chain representing the second ('D') stage comprise a single class which is recurrent and aperiodic. Hence, this process will have a limiting stationary state distribution which is independent of the distribution of starting dose levels (from the 'B' stage). It is tempting to speculate that the large-sample behavior of maximum likelihood estimates derived from data generated under the stochastic sampling model should be similar to that of estimates derived from data generated via fixed sampling at the various dose levels in the same proportions as that of the equilibrium distribution. Nevertheless, our results indicated that even if this is true, the sample sizes typically employed in Phase I trials fall far short of what must be required.

To summarize briefly, the simplest interval estimates of $\gamma$ are provided by the delta method and a method related to Fieller's theorem. Letting $\hat{V}_\mu$ and $\hat{V}_\beta$ denote the usual large sample variance estimates for $\hat{\mu}$ and $\hat{\beta}$ in the model (3) and $\hat{V}_{\mu \beta}$ the large sample covariance estimate, then asymptotic $100(1 - \alpha)\%$ confidence interval estimates for $\gamma$ using the two methods are given, respectively, by

$$\{\gamma : (\hat{\gamma} - \gamma)^2 / \hat{V}_\mu < \chi_\alpha^2 \}$$  \hspace{1cm} (4)

and

$$\{\gamma : (\hat{\mu} - \mu + \gamma \hat{\beta})^2 / \hat{V}_f < \chi_\alpha^2 \},$$  \hspace{1cm} (5)

where $\hat{V}_d = (\hat{V}_\mu + 2\hat{\gamma} \hat{V}_{\mu \beta} + \hat{\gamma}^2 \hat{V}_\beta) / \hat{\beta}^2$, $\hat{V}_f = \hat{V}_\mu + 2\hat{\gamma} \hat{V}_{\mu \beta} + \hat{\gamma}^2 \hat{V}_\beta$, and $\chi_\alpha^2$ is the upper $100(1 - \alpha)$ percentile of the $\chi^2$ distribution with one degree of freedom. Delta method intervals are always finite when $0 < \hat{\beta} < \infty$, but showed marked undercoverage. Fieller's intervals were frequently infinite (including cases of two disjoint infinite half-intervals), and the accuracy of actual coverage probabilities depended on the nominal coverage. At least part of the poor performance of the delta method is immediately attributable to the upward bias in $\hat{\beta}$, which will deflate the estimated variance of $\gamma$, sometimes to a considerable degree.

We also investigated a somewhat more computationally complex procedure involving inversion of a likelihood ratio test of the composite hypothesis $H_0 : \gamma = \gamma_0$, $\beta > 0$. This is based on $L(y, x, n | \gamma_0) = L(y, x, n | \hat{\gamma}) / L(y, x, n | \hat{\theta})$, where $L(\cdot)$ is the likelihood represented by (3), $\hat{\theta}$ is the unrestricted maximum likelihood estimate of $\theta$, and $\hat{\gamma}$ is the maximum likelihood estimate in the restricted parameter space defined by $H_0$. Confidence sets based on the usual large sample approximation to the distribution of $2 \log \lambda(Y, X, N | \gamma)$ are obtained as

$$\{\gamma : 2 \log \lambda(y, x, n | \gamma) < \chi_\alpha^2 \}. \hspace{1cm} (6)$$
These intervals did not markedly outperform (4) or (5). Coverage probabilities, though improved over the delta method, were still less than nominal. Infinite intervals also occurred, though with less frequency than with Fieller's intervals.

Our results regarding asymptotic likelihood ratio intervals were different from those of Williams (1986), who reported quite reasonable performance of intervals for a 50th percentile. Although his sample sizes were somewhat larger than ours, he also used fixed sample points spaced more uniformly and symmetrically about the target percentile. Hence, it remained unclear whether the undercoverage we observed was due to the small sample size, the stochastic sampling scheme, or both. Therefore, in the remainder of this section, we discuss two procedures which are exact in the small-sample setting, and one of which accounts as well for the nature of the sampling scheme.

Following the principle of test inversion, a $100(1 - \alpha)\%$ confidence set for $\gamma$ is given by

\[
\{\gamma : \sup_{\beta > 0} \Pr_\theta[T(Y, X, N|\gamma) \geq t(y, x, n|\gamma)] > \alpha\},
\]  

where $T(\cdot)$ and $t(\cdot)$ refer to any of statistics inside (4)-(6) with nominal $\chi^2$ distributions. The exact distribution of $T(Y, X, N|\gamma)$ is discrete regardless of the sampling scheme. For the binomial sampling scheme conditional on $(X, N)$ it takes on distinct values for every distinct combination of the sufficient statistics $S_1 = \sum Y_i$ and $S_2 = \sum X_i Y_i$. For an assumed true value of $\theta$, it is computationally feasible to exhaustively enumerate the entire sample space for $Y$ given $(X, N)$, compute the probability of that outcome, and thus determine the exact distribution of $(S_1, S_2)$ and $T(Y, X, N|\gamma)$. The exact probability that $T$ would equal or exceed a given realized value can thus be computed. One then determines whether $\gamma$ is a member of the set (7) by evaluating this probability over the range of the nuisance parameter $\beta$.

The up-and-down sampling scheme places constraints on the realizations of $Y$ that are possible, given $(X, N)$. We know of no simple algorithm that will determine whether a possible realization of $Y$ under the fixed binomial sampling scheme is also realizable under the true sampling scheme. Since one does not in fact condition on $(X, N)$ in sampling, it is not clear that it is even desirable to proceed on this basis, even if it were feasible. However, it is possible to work with the unconditional distribution of $T(Y, X, N|\gamma)$, wherein only the starting dose value $x_1$ and the second-stage sample size $n_2$ are fixed. Though exact computations are theoretically possible, in this case it is simpler to estimate the probability in (7) via Monte Carlo simulation.
Under both sampling schemes, significant computational simplification is obtained if the probability in (7) can be evaluated only at the restricted maximum estimate of $\beta$ given $\gamma$, rather than over the range $\beta > 0$. This interval is defined as
\[ \{ \gamma : \Pr[T(Y,X,N|\gamma) \geq t(y,x,n|\gamma)] > \alpha \} . \tag{8} \]

In a different setting, that of testing the equality of two binomial proportions, Storer and Kim (1990) found that a test statistic defined similarly to that inside (8) performed nearly as well as the exact unconditional test (Suissa and Shuster, 1985), which would be defined similarly to that inside (7). For both sampling schemes this is the procedure that will be evaluated in the next section.

3. Evaluation of Small-Sample Behavior

Using Monte Carlo simulations we evaluated some aspects of the performance of intervals (8) under the two sampling schemes, as well as that of intervals (4)-(6). The coverage probability of these intervals was estimated by generating data under a true model $\theta_0$ and observing whether or not the true value $\gamma_0$ was included in the confidence set, i.e., whether one accepted or rejected $H_0 : \gamma = \gamma_0, \beta > 0$.

For determining membership of $\gamma_0$ in various confidence sets each replication comprised the following steps:

[a] at $\theta_0$, generate sample data $y, x, n$;

[b] obtain maximum likelihood estimate $\hat{\theta}$ and restricted maximum likelihood estimate $\hat{\theta}$;

[c] determine large-sample $\chi^2$ statistics for:
   [c1] delta method interval;
   [c2] Fieller's method interval;
   [c3] likelihood ratio interval;

[d] at $\hat{\theta}$ from [b], generate 250 realizations of each $T(Y,X,N|\gamma_0)$ using the sampling scheme:
   [d1] $Y$ conditional on $(x,n)$ from [a];
   [d2] unconditional on $(Y,X,N)$;
[e] note membership of $\gamma_0$ in interval by reference to either nominal $\chi^2$ [c], or empirical probability of a larger or equal value [d].

Although above we imply the feasibility of exactly evaluating the probability in (8) for the fixed binomial sampling scheme, for this extensive set of simulations it proved more efficient and sufficiently accurate to use Monte Carlo simulation here as well.

The settings of $\theta_0$ in these simulations comprised $\gamma_0 = 0(2)6$ by $\beta_0 = 0.1(0.1)2.0$, defined with respect to dose levels $a_i = i$, $i = -11, \ldots, 20$, with $x_1 = a_0 = 0$, so that the settings of $\gamma_0$ define increasing distances of the true MTD from the starting value, ranging from 0 to 6 dose levels. For each setting, we obtained 1000 realizations for which $0 < \hat{\beta} < \infty$; the probabilities estimated from the sub-simulations in [d] were based on the same restriction. Results are presented here for design 'BD', with $m_2 = 24$; results with design 'BC' and different sample sizes are similar.

Figure 2 presents the coverage probabilities associated with the three statistics based on a nominal $\chi^2$ distribution (top panels), an exact conditional binomial distribution (middle panels), and an exact unconditional distribution (bottom panels). Since these results did not depend greatly on $\gamma_0$ (see Figure 1 in this regard), we present the results as a function of $\beta_0$ for $\gamma_0 = 4$. This figure dramatically illustrates the impact of the nuisance parameter $\beta$ on coverage using the large-sample $\chi^2$ approximation, and reveals that only the exact method conforming with the true sampling scheme attains approximately the nominal coverage over the range of this parameter.

We performed two other sets of simulations to further elucidate the impact of the sampling scheme and sample size on inference. In the first set, we used the same sequence of realizations of $(Y, X, N)$ from above, discarded the originally realized $y$, and then generated a new realization of $Y$ using fixed binomial sampling conditional on $(x, n)$. The coverage results for the intervals (4)-(6) are shown in the bottom panels of Figure 3. Except for the delta method, the large-sample procedures now attain coverage probabilities close to nominal, which reaffirms the result of Williams (1986) for likelihood ratio intervals.

In the second set, we investigated the question of how large the sample has to be before large sample methods provide satisfactory results for small values of $\beta$ under the stochastic sampling scheme. For these simulations we fixed $\beta_0 = 0.2$ and $\gamma_0 = 4$, and used second stage sample sizes $m_2 = 24(2^k)$, $k = 0, \ldots, 5$, or a maximum sample size of 768. The results for the intervals (4)-(6) are presented in the top panels of Figure 3, and indicate that even at the largest sample size the coverage levels are still somewhat below nominal.
In addition to coverage probability, it is also of interest to consider the width of confidence intervals. Although it would be feasible to evaluate (8) to any desired precision for a single interval, as would be done in a practical implementation, it is not practical to do so in extensive simulations. Evaluation of the single point \( H_0 : \gamma_0 = \pm \infty (\beta = 0) \) conveys some information, but does not provide information about relative width per se and does not distinguish completely infinite intervals from the case of disjoint infinite half-intervals, similar to those that occur when using (5). For this reason, Figure 4 presents the results of testing \( H_0 : \gamma = \gamma_0(q) = [\logit(q) - \logit(q_0)]/\beta_0 \), for \( q = 0.03(0.03)0.81 \) and \( \beta_0 = .1, .5, \) and 1, for the exact intervals using the true sampling scheme. The modification to the ‘B’ stage noted earlier in Section 1 improves the efficiency of the design when simulating a situation where the starting dose is above the true MTD, a situation which would be unusual, though not impossible, in practice.

Figure 4 reveals the cost in accurately accounting for the inherent bias of the up-and-down design. For the smallest setting of \( \beta \) the intervals are essentially all infinite, though for the largest setting it is apparent that at least some portion of the finite range of \( \gamma \) is excluded a significant fraction of the time. Under conditions where finite intervals can be provided, it appears that those based on the likelihood ratio criterion are likely to be somewhat narrower than others, particularly for the lower bound of the interval. It should also be noted that the likelihood ratio statistic can be computed in situations where \( \hat{\beta} = \infty \), whereas the statistics in (4) and (5) are undefined. This occurs whenever there exists a value \( x_8 \), not necessarily unique, such that \( y_i = 0 \) for any \( x_i < x_8 \) and \( y_i = n_i \) for any \( x_i > x_8 \). The top panels of Figure 1 indirectly indicate the frequency with which this situation occurs, as it accounts for nearly all the instances where \( \hat{\beta} \) is not positive and finite. When the simulations presented in Figures 2 and 4 for the likelihood ratio based intervals are repeated without the restriction \( \hat{\beta} < \infty \) in the main and sub simulations, very similar results are obtained.

As noted above, we obtained very similar results with design ‘BC’, which differs in the second stage sampling scheme (Storer, 1989). In general, the performance of the intervals (8) should be examined explicitly with any candidate sampling design. For example, with a unidirectional traditional design (‘A’ in Storer (1989)), the coverage of the intervals (8) is similar and reasonably adequate using both the true and conditional binomial sampling schemes. Because the design cannot de-escalate, however, the lower bound tends to be relatively much tighter than the upper bound. Also, the data configurations generated by the design yield a positive finite estimate of \( \beta \) as little as 20% of the time, which as a practical
matter restricts one to the use of likelihood ratio procedures.

4. Discussion

We have presented here a method for constructing valid confidence sets for the MTD in combination with certain sampling algorithms associated with Phase I clinical trials. In principle the methodology is adaptable to a variety of settings, but with a number of caveats. Although the restricted intervals (8) performed well for the designs considered, their suitability must be examined on a case by case basis. The intervals (7) will be valid in any case, but are computationally far more intensive and likely to be very conservative. The validity of the coverage probability must also be weighed against the width of the intervals, which can be significantly greater than that of large-sample intervals. Though the smallest setting of $\beta$ examined in Figure 4 is likely not realistic in practice, the larger settings also were associated with an undesirably high proportion of infinite or very wide intervals.

Choi (1990) discusses a method of interval estimation of the LD$_{50}$ in an up-and-down design that is associated with a point estimator computed from the dose levels associated with the turning points of the design. A variance estimate is derived that requires some inflation in small samples to achieve nominal coverage probabilities. Although the inflation factor did not depend on the slope of the dose response curve, it should be noted that the shallowest situation examined (a $N(0,1)$ tolerance distribution with dose spacings of 0.5) corresponds to $\beta \approx 1$ in the present discussion. Also, the number of turnings of design ‘BD’ (typically 1-3) is much smaller than even the smallest values he considered.

A larger problem is the dependence of inference on the true sampling scheme. One of our goals in considering alternative designs for Phase I clinical trials has been to establish procedures in which the “design” of the trial may be reasonably separated from the “analysis”. That is, we would like to be able to view the sampling algorithms ‘BC’ and ‘BD’ or others merely as recipes for choosing the dose levels that have desirable properties with respect to conservativeness and efficiency, if followed with reasonable fidelity. Ideally, the data that are generated by following these recipes would then be subject to analysis as if they had been generated by sampling under a fixed model. Although this may be a reasonable approach to point estimation, it seems clear that it does not carry through for interval estimation, and that the true sampling scheme must be accounted for. In practice, however, it is inevitable that violations of the sampling protocol will occur. In most cases it will then be difficult to specify what actually constitutes the true sampling scheme. Even if this could be done in
some circumstances, it might be cumbersome as a practical matter to accommodate alterations to computational procedures which were designed with a particular ideal scheme in mind.

ACKNOWLEDGEMENTS

This work was supported by NIH grants R29 CA45313, R01 AI29131, and P01 CA20432.

REFERENCES


\textbf{Figure Legends}

\textit{Figure 1.} Estimation results based on 1000 realizations of sampling scheme ‘BD’ with $m_2 = 24$ for which $0 < \hat{\beta} < \infty$. Results in the middle panel concern estimation of $\gamma$ are expressed in terms of the associated percentile of response for the true model. The legend in the top panel indicates the different settings of $\gamma_0$ relative to the starting dose level.

\textit{Figure 2.} Coverage probabilities based on 1000 realizations of sampling scheme ‘BD’ with $m_2 = 24$ for which $0 < \hat{\beta} < \infty$. Reference lines provide pointwise 95\% confidence limits on the observed coverage frequency if the true probability were exactly equal to nominal. The legend in the bottom middle panel indicates the different intervals under consideration.

\textit{Figure 3.} Coverage probabilities based on 1000 realizations of sampling scheme ‘BD’ for which $0 < \hat{\beta} < \infty$. Both top and bottom panels present coverage associated with the large-sample intervals (4)-(6). The top panels are presented as a function of increasing sample size; the bottom panels use resampling of $Y$ from the originally observed $(x,n)$ as described in Section 3, with $m_2 = 24$. The latter can be compared to the top panels of Figure 2, where the same intervals are applied directly to the data from sampling scheme ‘BD’. The legend in the bottom middle panel indicates the different intervals under consideration.

\textit{Figure 4.} Coverage probabilities based on 1000 realizations of sampling scheme ‘BD’ with $m_2 = 24$ for which $0 < \hat{\beta} < \infty$. The x-axis is expressed as the response percentile associated with the value of $\gamma$ being tested. The legend in the bottom middle panel indicates the different intervals under consideration.