Received: 28 May 2019 | Revised: 23 April 2020 | Accepted: 13 May 2020

DOI: 10.1002/sim.8909

RESEARCH ARTICLE

Statistics in Medicine WILEY

Exact confidence limits after a group sequential single arm binary trial

Chris J. Lloyd

Melbourne Business School, University of Melbourne, Carlton, Victoria, Australia

Correspondence
Chris J. Lloyd, Melbourne Business School, University of Melbourne, Carlton, VIC 3053, Australia.
Email: c.lloyd@mbs.edu

Group sequential single arm designs are common in phase II trials as well as attribute testing and acceptance sampling. After the trial is completed, especially if the recommendation is to proceed to further testing, there is interest in full inference on treatment efficacy. For a binary response, there is the potential to construct exact upper and lower confidence limits, the first published method for which is Jennison and Turnbull (1983). We place their method within the modern theory of exact confidence limits and provide a new general result that ensures that the exact limits are consistent with the test result, an issue that has been largely ignored in the literature. Amongst methods based on the minimal sufficient statistic, we propose two exact methods that out-perform Jennison and Turnbull's method across 10 selected designs. One of these we prefer and recommend for practical and theoretical reasons. We also investigate a method based on inverting Fisher's combination test, as well as a pure tie-breaking variant of it. For the range of designs considered, neither of these methods result in large enough improvements in efficiency to justify violation of the sufficiency principle. For any nonadaptive sequential design, an R-package is provided to select a method and compute the inference from a given realization.

KEYWORDS

adaptive trials, Buehler limits, sufficiency principle

1 | INTRODUCTION

In phase II clinical trials, the aim is to determine the efficacy of a treatment, compared with well-established benchmarks. Let p be the probability of a positive response. If p is smaller than a low benchmark p_0 then the treatment will not be further investigated. If p is larger than a high benchmark $p_1 > p_0$ the treatment proceeds to further testing. Thus, it is commonly desired to test the hypotheses

$$\mathcal{H}_0 : p \le p_0, \quad \mathcal{H}_1 : p \ge p_1. \tag{1}$$

Equivalently, one could define $\mathcal{H}_1 : p > p_0$ and target a high power at the specific alternative $p = p_1$. However, the more common formulation is as given in (1).

So-called group sequential designs involve a maximum number of stages K. At each stage, the tester may decide either $\mathcal{H}_0$ or $\mathcal{H}_1$ or reserve judgment to a later trial. At the last stage K, a decision is forced. The design allows early termination if the performance of the treatment is sufficiently clear (ie, poor or good) during the early stages, which results in savings

for the sponsor, as well as benefits to subjects. The standard formulation is to fix the type 1 and type 2 error rate and then minimize the mean total sample size.[1,2] It is well accepted that sequential designs can achieve smaller mean sample size than nonsequential designs.

After the test decision, there is still the issue of summarizing the inference about p in an estimate and confidence interval. The maximum likelihood (ML) estimate is the total positive responses divided by the total sample size. This estimator is biased and several alternative estimators have been proposed.[3-5]

Methods for confidence intervals can be classified according to three criteria. Approximate likelihood methods can be applied[6] though the coverage properties can easily break down for some values of p. Thus, exact methods which guarantee coverage are sought. Several authors have proposed exact one-sided confidence limits.[7-10] One-sided lower limits are of particular interest, particularly if the trial terminates by deciding $\mathcal{H}_1$.

A second criterion is whether the interval is an inversion of a two-sided test[11-14] or the intersection of two one-sided limits. The mean width of two-sided inversion intervals are typically shorter than those based on combining one-sided limits. However, the coverage of the implied lower and upper limits is unknown, yet will tend to be interpreted as if coverage error was equal at each end. Moreover, in the present application, the hypotheses being initially tested are one-sided, and a two-sided interval can easily contradict the test, even grossly. Indeed, this can happen with one-sided methods and we will give a sufficient condition to ensure that the limit and test agree.

Finally, there is the issue of sufficiency, which for the one-arm trial means that inference depends only on the total positive responses and total sample size. Methods based on nonsufficient statistics typically result in exact limits with a finer distribution and modest improvements in mean value.[15,16] In this article, we focus on exact limits based on a one-sided inversion. Our method ensures that the exact upper/lower confidence limits are as small/large as possible, subject to the one-sided coverage constraint. Two of the methods investigated are based on the sufficient statistic and two are based on inversion of a (combination) test which violates sufficiency but which is nevertheless in wide use. While our preference is to create the $1-\gamma$ interval using upper and lower limits both of coverage $1-\gamma/2$, this is a matter for discussion elsewhere.

The plan of the article is as follows. In the following section, we establish the notation for the design, as well as the model and some formulas for calculating probabilities associated with this model. Section 3 describes the modern theory of exact limits which may not be known to many readers. Sections 4 discusses four exact limits, two existing[7] and two new. In Section 5, we assess the performance of these limits on 10 different designs, whence a preferred method emerges. In Section 6, we assess a ranking function based on inverting the Fisher combination test. This and other methods[17] violate the sufficiency principle. The final section summarizes our recommendations.

2 | NOTATION, DESIGN, AND MODEL

2.1 | Design overview

At stage k, a binary response is measured on n_k patients. The number of positive responses is Y_k with binomial(n_k, p) distribution. The decision is based on the cumulative number of responses $S_k = \sum_{j=1}^{k} Y_j$. If $S_k \le a_k$ the hypothesis $H_0 : p \le p_0$ is accepted. If $S_k \ge b_k$ the hypothesis $H_1 : p \ge p_1$ is accepted. If $a_k < S_k < b_k$ then no decision is made and there is another stage. At the final stage K, $a_K = b_K - 1$ so that a decision is forced. The design is determined by K, and three vectors $n = (n_1, \ldots, n_K)$, $a = (a_1, \ldots, a_K)$ and $b = (b_1, \ldots, b_K)$. It will be convenient in the sequel to denote the cumulative sample size by $T_k = \sum_{j=1}^{k} n_j$ and to let $M \le K$ be the stage at which the decision is made. The decision to accept H_j is denoted by $D = j, j = 0, 1$.

Subject to error restrictions $\Pr(D = 1|p) \le \alpha$ for $p \le p_0$ and $\Pr(D = 0|p) \le \beta$ for $p \ge p_1$, designs are sought that minimize expected sample size $\mathrm{E}(T_M|p)$ at p_0,[1,2] at both p_0 and p_1[18] or for the worst case.[19] Jennison and Turnbull[7] listed a large number of military standard sequential designs for attribute testing. Our focus in this article, however, is not on the designs, but on inference at the conclusion of the trial.

The design can be generalized to allow the three design vectors to be data dependent, so-called *adaptive* designs. Several authors have given adaptive designs that lead to slightly lower mean sample size.[20-24] The methods and results in this article apply to adaptive designs, but computations are more complex and will not be examined in detail.

2.2 | Statistical model

The observed data is a vector of counts $\mathbf{y}=(y_1, \ldots, y_M)$ up to the last stage M. The sample space $\mathcal{Y}$ comprises all those count vectors that are possible under the design. While the cardinality of this set is typically smaller than $\prod_k(n_k+1)$, it can still be very large. For instance, in the seventh listed model of Table 1, $\mathcal{N} := \text{card}(\mathcal{Y}) = 52{,}251$.

Under the standard assumption that all responses are independent and that p is constant over the study, the probability of an outcome $\mathbf{y}=(y_1, \ldots, y_m)$ is a product of success and failure probabilities and so the log-likelihood is

$$\ell(p;\mathbf{y}) = \sum_{k=1}^{m} y_k \log p + (n_k - y_k)\log(1-p) = s_m \log p + (T_m - s_m)\log(1-p), \tag{2}$$

where n_k is a function of $y_1, \ldots, y_{k-1}$ for an adaptive design. Nevertheless, the minimal sufficient statistic (MSS) for p is $(S_M(\mathbf{y}), T_M(\mathbf{y}))$. For a nonadaptive design, $T_M(\mathbf{y})$ is a monotonic function of $M(\mathbf{y})$ only and so the MSS is $(S_M(\mathbf{y}), M(\mathbf{y}))$. In either case, the ML estimator is $\hat{p} = s_m(\mathbf{y})/T_m(\mathbf{y})$ which is known to be biased for p but can be easily corrected.[3] Dependence of these statistics on the data $\mathbf{y}$ will be suppressed where this causes no confusion.

Despite the familiar form of the ML estimator and log-likelihood (2), the distribution of S_M is not binomial(T_m, p), because of the restriction on the observable vectors $(y_1, \ldots, y_m)$. Nevertheless, the distribution of the sufficient statistic (S_M, M) can be derived[6] by noting that for each outcome $\mathbf{y}$ such that $S_M(\mathbf{y})=s$ and $M(\mathbf{y})=m$ there were s responses from T_m possible. Therefore,

$$\Pr(M=m, S_m=s|p) = Q_m(s)p^s(1-p)^{T_m-s}, \tag{3}$$

where $Q_m(s)$ is the number of ways of obtaining s responses from T_m subjects without exceeding the decision boundaries during trials $1, \ldots, m-1$. At least for nonadaptive designs, computing these numbers is not difficult for the kinds of sample sizes common in practice. This is discussed in Supplementary Materials Appendix C where we prove the recursive formula

$$q_k(s) = \sum_{j=\max(a_{k-1}+1, s-n_k)}^{\min(b_{k-1}-1, s)} q_{k-1}(s) \binom{n_k}{s-j}, \tag{4}$$

where $q_k(s)$ is the number of ways of obtaining s positive responses at the kth trial, that is, along each sample path. So $Q_m(s)$ equals $q_k(s)$ when the sample path hits the decision boundary, that is, when $k=m$.

For nonadaptive designs, the number $\mathcal{N}_{SM}$ of distinct values of (S, M) is typically small in practice. As an example, in the previously mentioned example where $\mathcal{N} = 52{,}251$, there are only 351 elements in the set $\mathcal{Y}_{SM}$ of possible values for (S, M). Thus, once the counts $Q_m(s)$ have been determined on $\mathcal{Y}_{SM}$ via (4), the model (3) is very easy to deal with numerically for nonadaptive designs.

3 | EXACT UPPER LIMITS

Buehler[25] was the first to properly formulate the problem of a smallest frequentist upper limit based on discrete data. Let $Y \in \mathcal{Y}$ be the data and $f(y; p, \phi)$ be any probability model where p is the parameter of interest and ϕ a nuisance parameter (of any dimension). In the context of a single arm trial there is no nuisance parameter, however, the theory is presented here in its full generality.

Select a function $R : \mathcal{Y} \to \mathfrak{R}$ that ranks datasets $\mathbf{y}$ with respect to p, that is, if $R(\mathbf{y}_1) > R(\mathbf{y}_2)$ then $\mathbf{y}_1$ supports a higher value of p than $\mathbf{y}_2$. Buehler defined

$$U^*(\mathbf{y}) = \sup\{p : \Pr(R(\mathbf{Y}) \le R(\mathbf{y}); p, \phi) > \gamma\} \tag{5}$$

and suggested that $U^*(\mathbf{Y})$ is the smallest possible upper limit that is a nondecreasing function of $R(\mathbf{Y})$. This was not formally proven until 35 years later.[26] The absolute optimality of these limits once a ranking function has been chosen does not seem to have been appreciated in the literature on sequential trials.

TABLE 1 Mean value of exact upper and lower limits and interval width

Design			lr	cp	jt	ml
$n=(5,6,5,9)$	$p_0=0.40$	$\alpha=0.095$	0.781	0.773	0.810	0.806
$a=(2,4,5,12)$	$p_1=0.75$	$\beta=0.106$	0.496	0.489	0.526	0.526
$b=(5,9,11,13)$			0.286	0.285	0.284	0.280
$n=(19,35)$	$p_0=0.10$	$\alpha=0.048$	0.387	0.385	0.403	0.399
$a=(4,15)$	$p_1=0.30$	$\beta=0.094$	0.123	0.123	0.122	0.120
$b=(20,16)$			0.264	0.263	0.281	0.279
$n=(18,14)$	$p_0=0.70$	$\alpha=0.050$	0.925	0.925	0.926	0.925
$a=(13,26)$	$p_1=0.90$	$\beta=0.099$	0.670	0.670	0.669	0.670
$b=(19,27)$			0.254	0.254	0.256	0.255
$n=(15,15,10)$	$p_0=0.05$	$\alpha=0.046$	0.262	0.262	0.264	0.263
$a=(-1,2,4)$	$p_1=0.20$	$\beta=0.087$	0.214	0.214	0.217	0.216
$b=(4,5,5)$			0.0478	0.0479	0.0471	0.0473
$n=(15,15,10)$	$p_0=0.08$	$\alpha=0.045$	0.311	0.312	0.313	0.314
$a=(0,3,6)$	$p_1=0.25$	$\beta=0.099$	0.240	0.240	0.242	0.244
$b=(5,6,7)$			0.0718	0.0718	0.0708	0.0702
$n=(20,15,15)$	$p_0=0.40$	$\alpha=0.064$	0.658	0.657	0.666	0.665
$a=(9,16,24)$	$p_1=0.65$	$\beta=0.091$	0.308	0.307	0.319	0.317
$b=(16,21,25)$			0.350	0.350	0.348	0.348
$n=\text{rep}(50,7)$	$p_0=0.02$	$\alpha=0.043$	0.0889	0.0907	0.0901	0.0895
$a=(0,1,3,5,7,10,13)$	$p_1=0.07$	$\beta=0.037$	0.0701	0.0717	0.0721	0.0714
$b=(4,6,8,10,11,12,14)$			0.0188	0.0190	0.0180	0.0180
$n=\text{rep}(80,7)$	$p_0=0.05$	$\alpha=0.077$	0.127	0.128	0.129	0.128
$a=(2,7,13,19,25,31,37)$	$p_1=0.10$	$\beta=0.026$	0.0754	0.0754	0.0784	0.0775
$b=(9,14,19,25,29,33,38)$			0.0520	0.0521	0.0505	0.0508
$n=\text{rep}(13,7)$	$p_0=0.08$	$\alpha=0.043$	0.317	0.322	0.319	0.319
$a=(0,1,3,5,7,10,13)$	$p_1=0.25$	$\beta=0.036$	0.237	0.242	0.243	0.243
$b=(4,6,8,10,12,14)$			0.0794	0.0802	0.0764	0.0765
$n=\text{rep}(8,7)$	$p_0=0.08$	$\alpha=0.026$	0.370	0.366	0.375	0.375
$a=(0,3,6,8,11,14,18)$	$p_1=0.30$	$\beta=0.040$	0.281	0.277	0.288	0.288
$b=(5,8,10,13,15,17,19)$			0.0889	0.0892	0.087	0.0871

Note: Ordering functions are lr, likelihood ratio limit; cp, clopper-pearson limit; jt, jennison and turnbull ordering; ml, maximum likelihood ordering. Upper values are conditional mean value of exact upper limit (with smallest underlined). Lower values are conditional mean value of exact lower limit (with largest underlined). Central values are mean width of interval (with smallest underlined). All mean values are calculated at $p=(p_0+p_1)/2$.

Solving (5) requires evaluating $\Pr(R(\mathbf{Y}) \le R(\mathbf{y}); p, \phi)$ across many parameter values. Each evaluation is a computation of order $\mathcal{N} := \text{card}(\mathcal{Y})$. Finding the largest solution for p in (5) is facilitated if $\Pr(R(\mathbf{Y}) \le R(\mathbf{y}); p, \phi)$ is a monotone function of p for each ϕ, though this is not essential and the algorithm we employ does not assume such. For both steps, there are large computational savings in defining the ranking on the much smaller space $\mathcal{Y}_{SM}$, over and above arguments based on sufficiency. Regardless, for practical designs computation of an exact upper limit does not take more than a few seconds.

There is an omission in this powerful theory, namely that the condition $\Pr(R(\mathbf{Y}) \le R(\mathbf{y}); p, \phi) > \gamma$ may have no solution. In this case, let E be the subset of $\mathcal{Y}$ for which a solution exists, and let $u^*_{\min} = \min(u^*(\mathbf{y}) : \mathbf{y} \in E)$. It was shown[27] that for $\mathbf{y} \notin E$ the upper limit $U^*(\mathbf{y})$ can be assigned an arbitrary value so long as it is no larger than $u^*_{\min}$. They speculated that the

set $\bar{E}$ being nonempty may indicate a defect in the ranking function. This does occur for apparently reasonable ranking functions in the present context of sequential trials. Further details are given in Web Appendix A where we also give the theory for exact lower limits.

There is one final aspect of the theory worth mentioning. If the ranking function $R(\mathbf{y})$ has any ties then breaking these ties results in an exact upper limit $U^*(\mathbf{y})$ that is no larger for all $\mathbf{y}$ and almost always smaller for at least one $\mathbf{y}$.[28] It follows that exact limits based on a ranking function R defined on the sufficient statistic (s, m) can be improved by refining R to further order datasets $\mathbf{y}$ that give the same value of (s, m). This, of course, means violating the sufficiency principle. We present two such methods in Section 7.

4 | ALTERNATIVE RANKING FUNCTIONS

To define a Buehler upper limit, it is sufficient to give a ranking function. If the ranking function is defined on the sufficient statistic (S, M) then it should be intuitively clear that $R(s, m)$ should be increasing in s for fixed m. This leaves quite a lot of flexibility on how to rank datasets with different values of m.

The limits proposed by Jennison and Turnbull[7] are Buehler limits based on a particular ranking function, though Buehler's theory was unknown to them. It is simpler to explain this ranking, hereafter referred to as JT, in terms of a particular example.

Example 1. Consider testing $\mathcal{H}_0 : p \leq 0.4$ vs $\mathcal{H}_1 : p \geq 0.75$ with a four-stage group sequential trial. Let $n = (5, 6, 5, 9)$, $a = (2, 4, 5, 12)$, and $b = (5, 9, 11, 13)$. It is easy to calculate that the type 1 and type 2 errors α, β of this design are close to 0.10. The left panel of Figure 1 displays the sufficiency reduced sample space $\mathcal{Y}_{SM}$, the decision variable $D(\mathbf{y})$ and the JT ranking to the left of each point. The ranking increases with s for each fixed m, separately on sets $D(\mathbf{y}) = 0$ and $D(\mathbf{y}) = 1$. Superficially, it appears reasonable, however, it is not obvious that $(s, N[m]) = (0, 5[1])$ should be ranked lowest. It is even less clear that the outcome $(6, 16[3])$ should be ranked lower than $(7, 25[4])$, since the estimate of p is 0.375 in the former and 0.280 in the latter case.

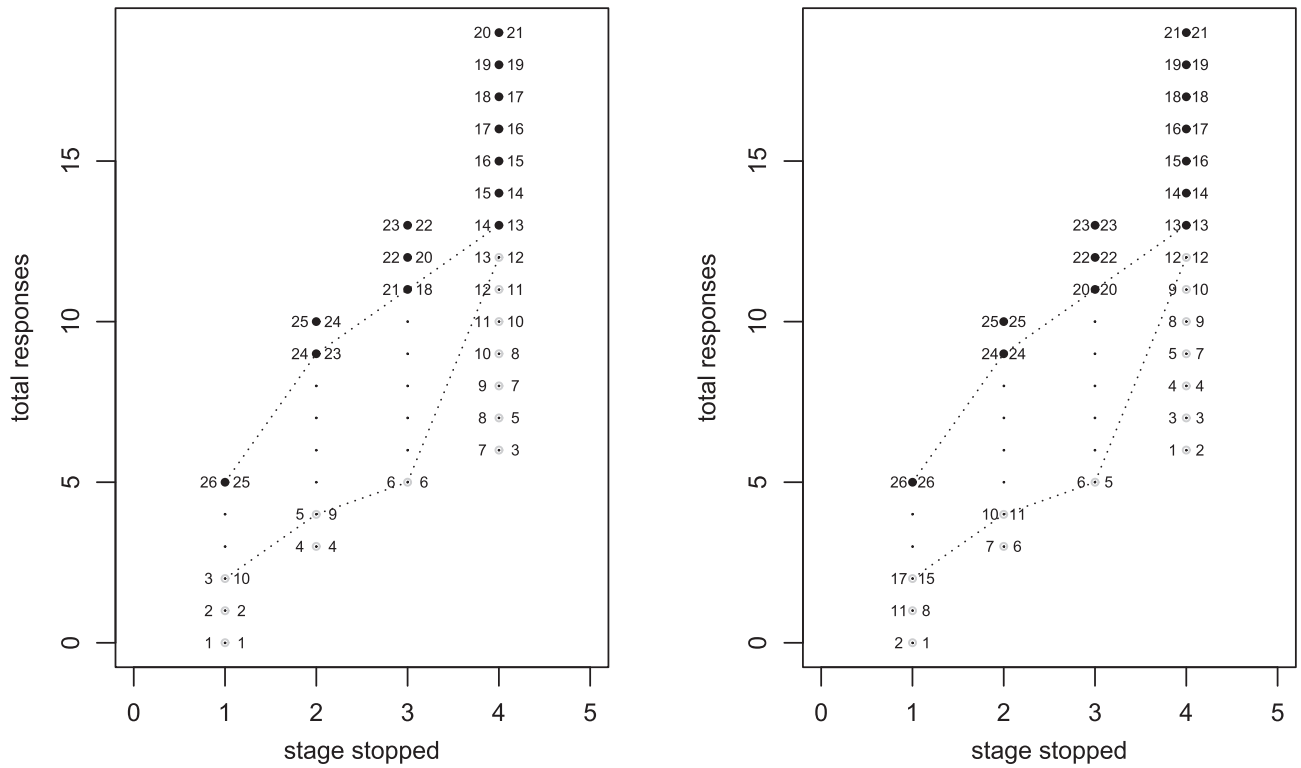

FIGURE 1 Representation of sample space $\mathcal{Y}_{SM}$ and four alternative rankings for design $n = (5, 6, 5, 9)$, $a = (2, 4, 5, 12)$, and $b = (5, 9, 11, 13)$. Solid plotting symbol indicates decision variable $d = 1$. *Left.* JT ranking to left of each point, ML ranking to right. *Right.* CP ranking to left of each point, LR ranking to right

Thus, the same authors also suggested that outcomes might be ranked by the ML estimator $\hat{p} = S_M/T_M$. This ranking is indicated to the right of each point of the sample space. The two rankings do not agree. Moreover, two datasets give the same estimate $\hat{p} = 0.4$, namely, (2, 5[1]) and (10, 25[4]). It was noted in the previous section that ties can only lead to larger upper limits. Jennison and Turnbull claimed[7] that the two rankings largely agree which does not seem to be true in this instance. They also made the weaker claim that the resulting Buehler limits are extremely close. The ranking function $R(s, m) = s$ has also been suggested.[8] We do not include it in our study as it is logically implausible and does not perform well. We now propose two alternative ranking functions. Kabaila and Lloyd[29-31] show that the ranking function should itself be based on an approximate $(1 - \gamma)$ upper limit, rather than an estimator, or something more ad hoc. The intuition for this is that if we start with an approximate confidence limit then the sample space is appropriately ordered for the task of defining a confidence limit, and solving (5) merely adjusts it to be exact, while preserving this ordering. Our first proposal is to rank datasets based on the upper limit of Clopper and Pearson[32] assuming that S_m is binomial from T_m trials. This was recently suggested[20] though not evaluated. If $B(x, a, b)$ is the distribution function of the beta distribution with parameters (a, b) then an explicit formula is

$$U_{1-\gamma}(s_m, T_m) = B^{-1}(1 - \gamma, s_m + 1, t_m - s_m)$$

when $s_m < t_m$ and $U(s_m, T_m) = 1$ when $s_m = T_m$. Of course, S_m is not binomial so this upper limit is not correct but it can still be used as a ranking function. We will denote this ranking by CP.

Our second proposal is to rank datasets based on the approximate likelihood ratio upper limit. Explicitly, the upper limit for p is the solution (greater than $\hat{p}$) of

$$2s_m \log\left(\frac{\hat{p}}{p}\right) + 2(T_m - s_m)\log\left(\frac{1-\hat{p}}{1-p}\right) = z^2_{1-\gamma}, \tag{6}$$

where z_q is the standard normal q% quantile. In the case $s_m = 0$, the first term is interpreted as zero. When $s_m = T_m$, the upper limit is set to 1.0. We denote this ranking by LR. Two-sided continuity corrected inversion limits based on the LR test have been proposed elsewhere.[6] The distinction here is that we are using the one-sided inversion as a ranking function for an exact procedure.

The right panel of Figure 1 compares the CP and LR rankings. Both take 26 distinct values on the 26 points in $\mathcal{Y}_{\text{SM}}$: there are no ties. They both give much lower rank to the points in the bottom right of the sample space than JT. It is worth noting that, using the CP ranking, the lowest ranked point $(S, T_m) = (6, 25[4])$ is in the set $\bar{E}$ when $\gamma = 0.05$. While $U^*(\mathbf{y})$ can be defined anywhere in the interval $[0, u^*_{\min}]$, we will always use the conservative convention of defining $U^*(\mathbf{y}) = u^*_{\min}$ when $\mathbf{y} \in \bar{E}$.

5 | COMPATIBILITY OF LIMITS WITH TEST

Confidence limits should agree with the test result. Most importantly, if $D(\mathbf{y}) = 1$ so that the drug continues to development and further testing, we would expect the lower limit $L^*(\mathbf{y})$ to be larger than p_0. Less critical in practice, if $D(\mathbf{y}) = 0$ then we would expect the upper limit $U^*(\mathbf{y})$ to be less than p_1. If this is true we will say the limits and the test are *compatible*. This problem has hardly been mentioned in the literature.[5] All the confidence limits described in this article (as well as others not considered) are based on inverting a ranking function R of the sample space with respect to p. So the question becomes what conditions on R ensure that the generated limit is compatible with the test.

There is a simple condition proven in Appendix D. Let $D_1 = \{y \in \mathcal{Y} : D(\mathbf{y}) = 1\}$ and $D_0 = \overline{D}_1$. Then the limits generated by a ranking R will be compatible with the test provided $R(y_1) > R(y_0)$ for all $\mathbf{y}_1 \in D_1$ and $\mathbf{y}_0 \in D_0$. This is an extremely natural condition. In words, it requires that a higher ranking cannot lead us to accept instead of reject the null. If a ranking fails this condition then we would plausibly suspect that there is something wrong either with the test itself or with the ranking. Looking at the CP and LR rankings in the right of Figure 1, both violate this condition because of the point $(s, m) = (2, 1)$. The test boundaries here have not been chosen to optimize any standard criteria. They are deliberately ad hoc. So the fact that a natural ranking of points is incompatible with the test is not such a surprise and suggests that the test boundary might be modified. Indeed, we could remove the point $(2, 1)$ from acceptance region D_0 and perhaps add another point such as $(s, m) = (6, 3)$. This apparently results in

both CP and LR rankings being compatible, though we would need to recheck the size and power of the modified test.

If we decide there is something wrong with the ranking, then we can easily modify it to be compatible. For rankings like CP and LR that are based on a numeric measure (an approximate upper limit), we can simply add a sufficiently large constant to this numeric measure on the rejection set D_1 only. This ensures all points in D_1 are ranked higher than D_0 while preserving the rankings within D_0 and D_1. In the following sections, where we compare several methods, our algorithm first checks the consistency condition and makes an adjustment if necessary, though this was not activated in any of the examples.

6 | PERFORMANCE OF DIFFERENT METHODS

Exact upper/lower limits are as small/large as possible subject to the ordering imposed by the ranking function R and the coverage requirement. To compare rankings, it is natural to look at some measure of the size of the exact upper/lower limits, smaller/larger being better. If $L^*(\mathbf{Y})$, $U^*(\mathbf{Y})$ are exact lower/upper limits for p and $\hat{p}(\mathbf{Y})$ is the ML estimator then we measure their quality by

$$m_L(p) = \mathrm{E}\{L^*(\mathbf{Y}) - \hat{p}(\mathbf{Y}); p\}, \quad m_U(p) = \mathrm{E}\{U^*(\mathbf{Y}) - \hat{p}(\mathbf{Y}); p\}. \tag{7}$$

Subtraction of $\hat{p}$ has no effect on comparisons of two alternative limits, since the conditional mean of the estimator will cancel. We prefer this measure since without this normalization $m_L(p)$ and $m_U(p)$ increases from $p=0$ to $p=1$ and makes graphical comparison difficult. The interval formed by $(L^*(\mathbf{Y}), U^*(\mathbf{Y}))$ is assessed by $m_I(p) = m_U(p) - m_L(p)$.

We will investigate 10 different designs. The first is the toy design of Example 1. The next two are taken from table 1 of Simon (1989) and are commonly used in practice. These designs only allow stopping for futility so $b_1 = n_1 + 1$. The next three are three-stage Fleming designs[1] taken from table 5.5. The last four are all seven-stage designs taken from tables 1 and 2 in Jennison and Turnbull,[7] specifically those designated L1.5, M4.0, H6.5, G6.5. These are all approximate designs with target error $\alpha = \beta = 0.05$. Details of the designs are in the left columns of Table 1. The performance measures are in the right section of the table. The upper, lower, and middle sets of numbers refer to the performance of the upper limit, lower limit, and interval, respectively. For each design, the optimal value is underlined. All mean values are calculated at the point $p = (p_0 + p_1)/2$. In all 10 designs, the JT and ML rankings produce quite similar results as asserted.[7] More importantly, in every case exact limits based on CP or LR are better, though the differences are often quite modest. While it is possible to find particular designs and values of p for which JT or even ML will very slightly outperform CP/LR the overall pattern is clear. Apart from the first toy example whose design was ad hoc, for the other nine published designs the limits based on all four rankings satisfied the compatibility condition of Section 5.

Figure 2 gives more detailed results for the first Simon two-stage model in the second row of Table 1, specifically how relative performance varies as a function of the unknown p. The left plot give the conditional mean value $m_U(p)$ of the upper limits, relative to the ML based limit. Since this is a plot for an upper limit, lower is better and any value smaller than zero means that limit outperforms the ML based upper limit. Clearly, the CP and LR based rankings (in green and blue, respectively) produce better results, especially between the clinically relevant values of $p \in [0.1, 0.3]$. The central plot is of $m_L(p)$ for the lower limit (so higher is better) and again, CP and LR outperform the ML and JT methods for all clinically relevant values. The right plot shows $m_I(p)$ for all four methods. Apparently, JT and ML (in red and black, respectively) have very similar performance but are outperformed by CP and LR (green and blue) which are also very similar.

7 | INSUFFICIENT RANKINGS

The MSS is (S, M). If we are prepared to violate the sufficiency principle, then there are many more ranking functions we could consider. This has been suggested in cases where there is auxiliary information[17] in which case (S, M) is not really sufficient. It has been noted[16] that methods that violate sufficiency can deliver small gains in efficiency.

Our first proposal starts with Fisher's combination test for adaptive multistage trials, which is based on the statistic

$$T(Y, p_0) = -\sum_{k=1}^{M} \log P_k$$

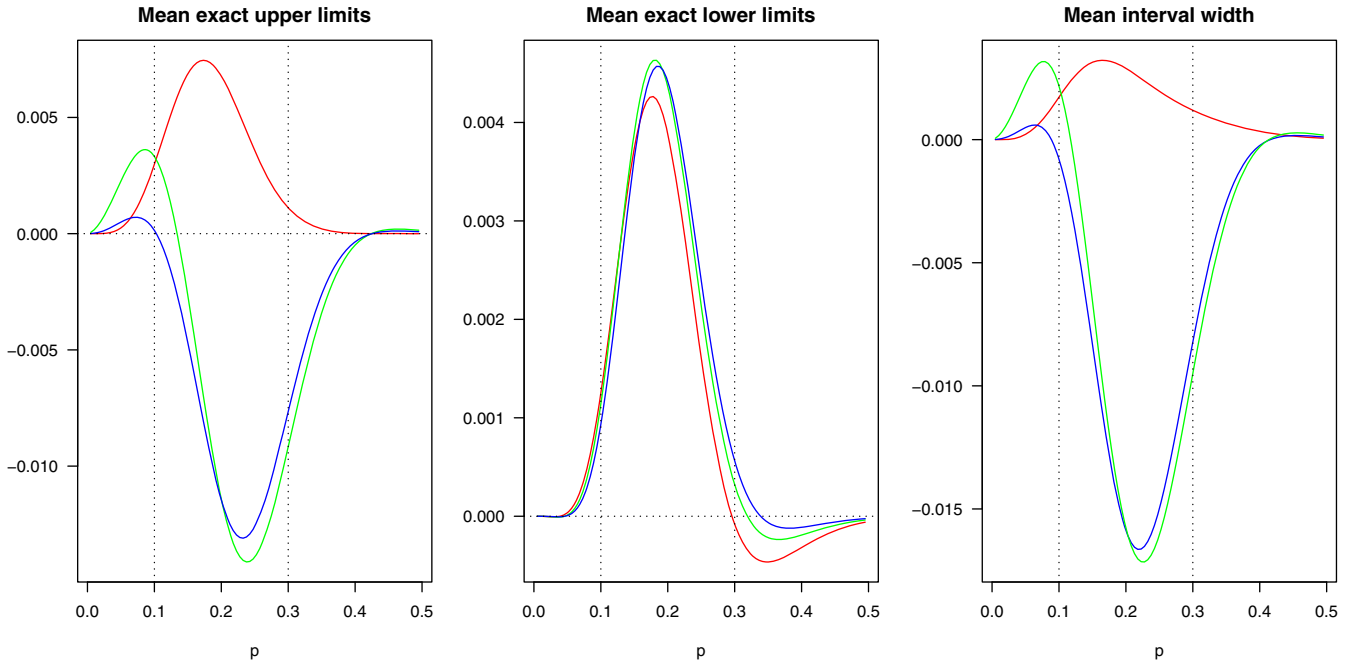

FIGURE 2 Performance for Simon design Table 1 (row 2). Red is for JT ranking, green is for CP ranking, blue for LR ranking. *Left.* Mean value measure (7) of exact upper limit minus same measure for ML method as function of p. Lower is better. *Center.* Mean value measure (7) of exact lower limit minus same measure for ML method as function of p. Higher is better. *Right.* Mean width of interval minus same measure for ML. Lower is better [Colour figure can be viewed at wileyonlinelibrary.com]

where $P_k = \Pr(Y_k \geq y_k; p_0)$ is the P-value for testing $p \leq p_0$ vs $p > p_0$. This is calibrated by the gamma distribution with shape parameter m which generates a combined P-value. It is a simple matter to then invert this test with respect to p_0 to obtain a combined upper limit for p. We will denote this Fisher inversion ranking by FI. While this ranking, and the exact limits based on it, violate the sufficiency principle, the same is true of the combination test itself which has not stopped it being widely used in the analysis of adaptive designs. Note that since the FI method is based on a P-value which is always less when the null is rejected than when it is accepted, the generated exact limits are compatible with the test by our main result in Section 5.

The FI ranking actually contradicts the CP ranking for many datasets. A modification of the FI method is to break the ties in the CP ranking, based on the value of the FI statistic, without directly contradicting the CP ranking. I will call this tie-break ranking TB. According to the earlier mentioned theory,[28] the Buehler upper limits based on this compromise ranking will never be larger and will sometimes be smaller than Buehler limits based on the CP ranking.

The left panel of Figure 3 plots exact FI and TB limits against exact CP limits for toy example 1. The insufficient limits take many more distinct values than the CP based limits, which gives the potential for greater efficiency. As demanded by theory, the TB based limits are all smaller than the CP limits. The FI based limits are mostly much smaller than the CP limits but sometimes they are also larger.

To compare the methods systematically, we again use the measures in (7). The right panel of Figure 3 plots the conditional mean for the two insufficient methods minus the conditional mean for the CP method. The FI method is better (smaller) than CP for some values of p and worse (larger) for others. As theory implies, the tie break method is uniformly better than CP but the advantage is practically small.

This analysis has been repeated for the three-stage Fleming designs of Table 1 in Figure 4. The clinically relevant values of p are vertical dotted lines. Only for design 8 is there any really good news for the FI method. The upper limits based on FI is smaller than CP, but only for values of p much lower than the lower benchmark. For values within the relevant range, and especially for values above the upper benchmark when full inference would be of most interest, there is very little advantage.

It is computationally difficult to numerically evaluate the insufficient methods for the other seven-stage designs in Table 1, see Web Appendix B for a detailed explanation.

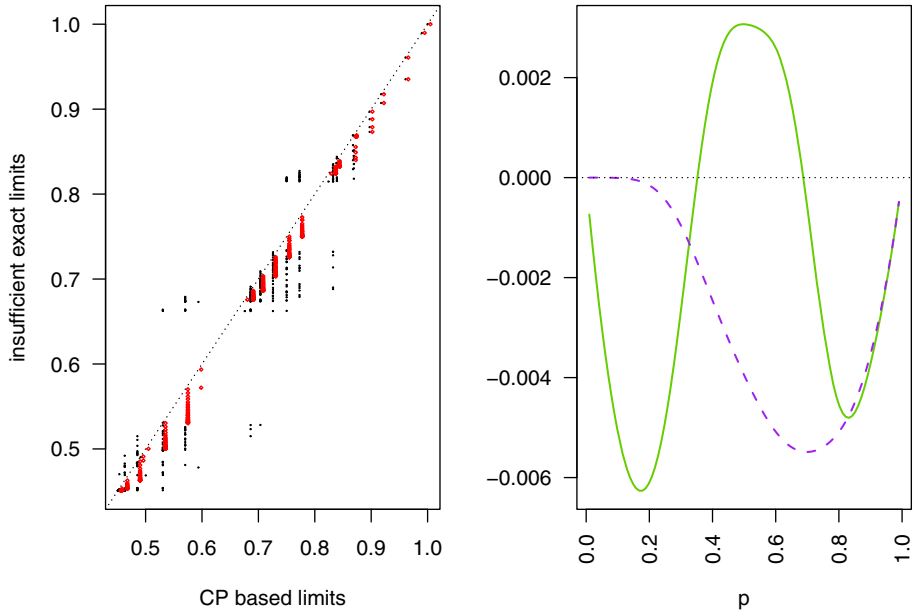

FIGURE 3 Insufficient methods for toy Example 1. *Left.* Exact upper limits based on insufficient ordering FI (black points) and TB (red triangles) vs CP exact upper limits. *Right.* Conditional mean value of FI (green solid) and TB (purple dashed) minus conditional mean value of CP. Smaller is better [Colour figure can be viewed at wileyonlinelibrary.com]

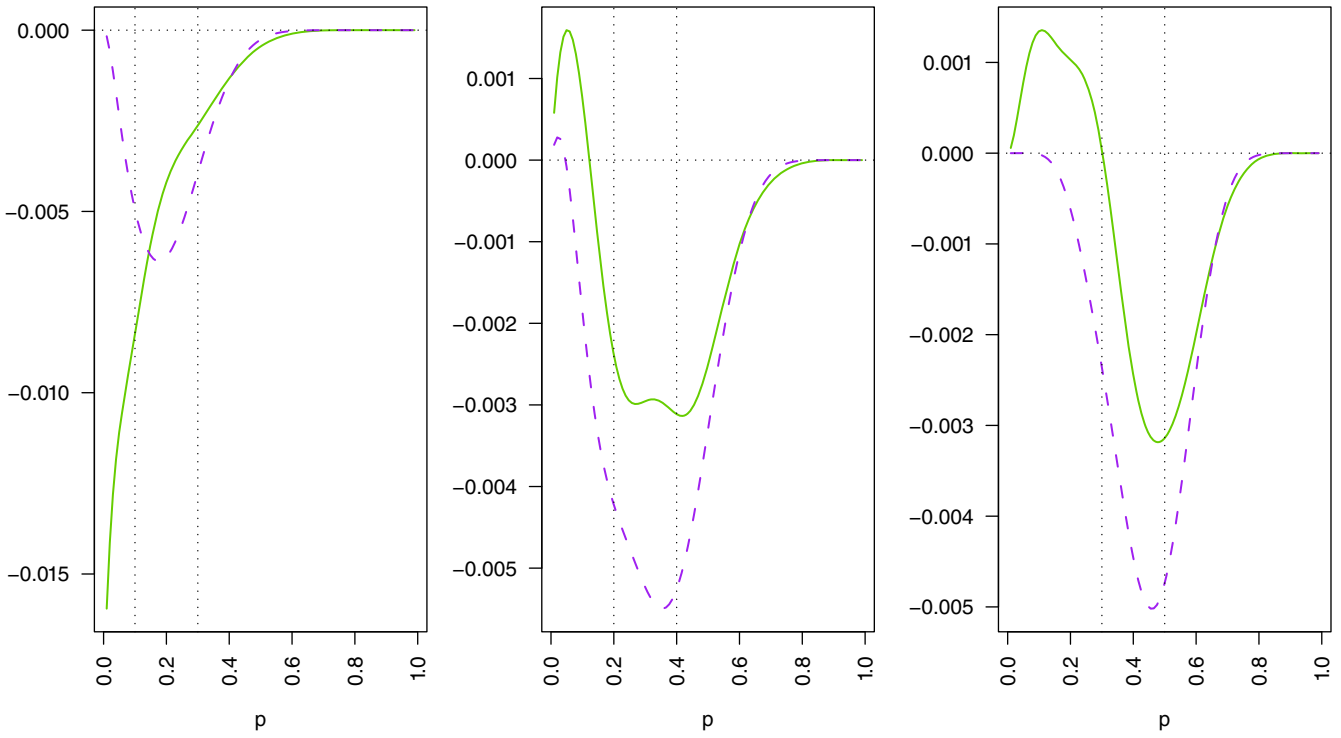

FIGURE 4 Performance of insufficient methods for designs 4 to 6. Each plot shows conditional mean value of FI (green solid) and TB (purple dashed) method minus conditional mean value of exact CP method. Lower is better. *Left.* Design 4. *Center.* Design 5. *Right.* Design 6 [Colour figure can be viewed at wileyonlinelibrary.com]

8 | CONCLUSION

This study has focused on exact upper and lower confidence intervals for the response probability p after a binary group sequential trial. We have placed existing methods within a wider theory and generated two new methods which have consistently better performance in terms of average size. Of the two suggested methods, we prefer the one based on the LR ordering to the one based on the CP ordering for two reasons. First, the ordering is based on the likelihood function rather than the false assumption of a binomial distribution. Second, for all the designs we investigated, as well as some not

reported here, the tail-set in (5) was never the null-set whereas for the CP ranking there were always datasets for which it was. Nevertheless, for some designs and parameter values CP can perform better than LR.

The second contribution is to give simple conditions under which exact limits will be compatible with the test result, as well as a method of adjusting the limits when they are incompatible. This was not, however, a practical issue for any of the designs and limits evaluated.

The third contribution is to investigate methods that are not based on the MSS. Such methods involve more computation, but are completely feasible for standard designs on modern computers. Heuristically, if any method were to outperform CP or LR we might expect it to be the method based on Fisher inversion (or some other combination) since this is breaking datasets $\mathbf{y}$ with the same values of (S, M) according to a statistically meaningful quantity. However, we find that the performance is mixed. The tie-break method is theoretically guaranteed to improve any sufficient method such as CP, however, the ranking is harder to justify inferentially. The modestly improved performance is purely based on the finer distribution of the insufficient ranking function which is a form of randomization by stealth. Therefore, it is not recommended.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ORCID


Chris J. Lloyd https://orcid.org/0000-0001-5745-1951

REFERENCES

1. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982;38:143-151.
2. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10:1-10.
3. Jung S, Kim MK. On the estimation of the binomial probability in multistage clinical trials. *Stat Med*. 2004;23:881-896.
4. Kunzmann K, Kieser M. Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint. *Stat Med*. 2017;36:971-984.
5. Nhacolo A, Brannath W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*. 2018;27(1):1-14.
6. Chang MN, O'Brien PS. Confidence intervals following group sequential tests. *Control Clin Trials*. 1986;7:18-26.
7. Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics*. 1983;25:49-58.
8. Atkinson EN, Brown BW. Confidence limits for the probability of response in multistage phase II clinical trials. *Biometrics*. 1985;41:741-744.
9. Tsai WY, Chi Y, Chen CM. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Stat Med*. 2008;27:15-35.
10. Porcher R, Desseaux K. What inference for two-stage phase II trials? *BMC Med Res Methodol*. 2012;bf 12(1):117-130.
11. Sterne TE. Some remarks on confidence of fiducial limits. *Biometrika*. 1954;41:275-278.
12. Crow EL. Confidence intervals for a proportion. *Biometrika*. 1956;43:423-435.
13. Blyth CR, Still HA. Binomial confidence intervals. *J Am Stat Assoc*. 1983;78:108-116.
14. Duffy DE, Santner TJ. Confidence Intervals for a binomial parameter based on multistage tests. *Biometrics*. 1987;43:81-93.
15. Englert S, Kieser M. Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics*. 2012;68:886-892.
16. Wang W. A "paradox" in confidence interval construction using sufficient statistics. *Am Stat*. 2018;72(4):315-320.
17. Chang MN. Improved confidence intervals for a binomial parameter following a group sequential phase II clinical trial. *Stat Med*. 2004;23:2817-2826.
18. Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics*. 1987;43:865-874.
19. Shuster J. Optimal two-stage designs for single arm phase II cancer trials. *J Biopharm Stat*. 2002;12:39-51.
20. Wang W, Yin H, Zhang Z. On construction of smallest one-sided confidence intervals for the response rate in adaptive two- or multi-stage designs. *Stat Methods Med Res*. 2020;29(6):1682-1699. https://doi.org/10.1177/0962280219873359.
21. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. *Stat Med*. 2006;25:3382-3395.
22. Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemp Clin Trials*. 2006;29:32-41.
23. Englert S, Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biom J*. 2013;55:955-968.
24. Zhou H, Lee JJ, Yuan Y. BOPs: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Stat Med*. 2017;36(21):3302-3314.
25. Buehler RJ. Confidence intervals for the product of two binomial probabilities. *J Am Stat Assoc*. 1957;52:482-493.
26. Jobe JM, David HT. Buehler confidence bounds for a reliability maintainability measure. *Technometrics*. 1992;34:214-222.
27. Kabaila PV, Lloyd CJ. On the optimality and limitations of Buehler bounds. *Austr NZ J Stat*. 2003;45:167-174.
28. Kabaila PV, Lloyd CJ. The efficiency of Buehler confidence limits. *Stat Probab Lett*. 2003;65:21-28.

29. Kabaila PV, Lloyd CJ. A computable confidence upper limit from discrete data with good coverage properties. *Stat Probab Lett*. 2000;47:189-198.
30. Kabaila PV, Lloyd CJ. The importance of the designated statistic on Buehler upper limits on a system failure probability. *Technometrics*. 2002;44(4):390-395.
31. Kabaila PV, Lloyd CJ. Buehler limits from approximate upper limits: a numerical investigation of the role of nominal coverage. *J Appl Stat Sci*. 2005;13:217-230.
32. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26: 404-413.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lloyd CJ. Exact confidence limits after a group sequential single arm binary trial. *Statistics in Medicine*. 2021;40:2389–2399. https://doi.org/10.1002/sim.8909