12-12-2007

Estimating Sensitivity and Specificity from a Phase 2 Biomarker Study that Allows for Early Termination

Margaret S. Pepe PhD
Fred Hutchinson Cancer Research Center, nnoble@fhcrc.org

Estimating Sensitivity and Specificity from a Phase 2 Biomarker Study that Allows for Early Termination

Margaret Sullivan Pepe[1,*] Ziding Feng[1], Gary Longton[1], Joseph Koopmeiners[1]

[1] *Fred Hutchinson Cancer Research Center*
1100 Fairview Avenue N., M2-B500, Seattle, WA 98109, USA

SUMMARY

Development of a disease screening biomarker involves several phases. In phase 2 its sensitivity and specificity is compared with established thresholds for minimally acceptable performance. Since we anticipate that most candidate markers will not prove to be useful and availability of specimens and funding is limited, early termination of a study is appropriate if accumulating data indicate that the marker is inadequate. Yet, for markers that complete phase 2, we seek estimates of sensitivity and specificity to proceed with the design of subsequent phase 3 studies.

We suggest early stopping criteria and estimation procedures that adjust for bias caused by the early termination option. A novel aspect of our approach is to focus on properties of estimates conditional on reaching full study enrollment. We propose the conditional-UMVUE and contrast it with other estimates, including naïve estimators, the well studied unconditional-UMVUE and the mean and median Whitehead adjusted estimators. The conditional-UMVUE appears to be a very good choice.

1. Introduction

The Early Detection Research Network (EDRN) seeks to develop biomarkers for cancer screening, diagnosis, prognosis and risk prediction. Marker development is a process, a sequence of studies. A 5-phase paradigm for this process has been adopted for the development of screening markers [1]. Briefly, phase 1 concerns marker discovery, phase 2 is retrospective marker validation in specimens from cases concurrent with clinical disease and controls without, phase 3 is retrospective marker validation in specimens taken prior to clinical disease, phase 4 is a prospective population study of test performance and phase 5 is ideally a randomized trial comparing mortality in the presence and absence of screening. Most of the studies conducted by EDRN are phase 1 and 2. Here we consider the design of a phase 2 study.

Stored blood or urine specimens are typically used in a phase 2 study. The marker is measured in specimens from a set of cases with clinical disease and from a set of appropriate

*Correspondence to: Fred Hutchinson Cancer Research Center
1100 Fairview Avenue N., M2-B500, Seattle, WA 98109, USA
email: mspepe@u.washington.edu

Contract/grant sponsor: NIH/NCI; contract/grant number: RO1 GM054438; UO1 CA086368

controls. Considerable effort has been expended to establish high quality specimen repositories for breast, lung and prostate cancer within the EDRN. Other groups have similarly built specimen banks for biomarker evaluation. It is important to use these resources judiciously and efficiently.

There is great enthusiasm in the scientific and business communities about the potential for technology to measure biomarkers [2]. Biomarker discovery studies abound and we anticipate that a large number of candidate biomarkers will be put forward for validation. However, the false discovery rate from phase 1 is likely to be high. That is, we expect that the majority of markers studied in phase 2 will not have adequate performance for proceeding to further development. This, along with concerns about conserving specimen resources and keeping study costs reasonable motivate a group sequential approach to phase 2 study design. In particular, designs that allow early termination when accumulating evidence suggests poor marker performance, are very attractive.

In this paper we consider dichotomous markers, with values denoted by $Y = 1$ for a positive result and $Y = 0$ for a negative result. Let D be the disease indicator, $D = 1$ for cases and $D = 0$ for controls. Marker performance is quantified by the sensitivity, $S = P[Y = 1|D = 1]$, and the false positive rate (or 1−specificity), $F = P[Y = 1|D = 0]$. Higher sensitivities and lower false positive rates indicate better performance.

When a phase 2 study terminates early, the marker is not considered for further development. In contrast when a study completes its full enrollment, estimates of (S, F) will be calculated to determine if and how marker development should proceed further. Our particular interest is in estimating (S, F) with data from completed phase 2 studies, i.e., from studies that do not terminate early.

Group sequential methods have received scant attention in the diagnostic testing literature. Mazumdar and Mazumdar and Liu [3, 4] consider methods for prospective comparative studies with early termination possible for either positive or negative conclusions. The context is geared towards phase 4 studies, not for phase 2 validation studies. There is no existing group sequential methodology for phase 2 biomarker studies.

Phase 2 treatment trials have statistical elements in common with our paradigm for phase 2 biomarker studies. In the prototype phase 2 treatment trial, subjects are classified as responders or not, the parameter of interest is the binomial response probability, and early termination occurs if the observed response rate is low. In our setting there are two binomial probabilities, S in cases and F in controls, and a study terminates early if either is clearly unsatisfactory. For simplicity we will first describe methodology when only one binomial probability is of interest and later address extensions to simultaneous consideration of two independent binomial proportions. We note that our methods are equally relevant to phase 2 treatment trials, although our motivation derived from phase 2 biomarker study design.

Substantial methodology has been developed for estimation following the group sequential design of a phase 2 therapeutic study. A key distinction between previous methods and what we propose here is that we are concerned with estimation only when a study reaches its planned full sample size. We consider properties of estimators conditional on study completion whereas previous methods provide an estimator at the terminating stage, early or not, and their unconditional properties are evaluated. The marginalized mean and variance of estimators are evaluated from all studies, those that terminate early and those that do not. We believe that estimation is less important if a study terminates early since the biomarker is clearly inadequate when that occurs and estimates for planning phase 3 are not needed. Moreover, confidence

intervals would be too wide with the smaller sample sizes for estimates to be meaningful when a study terminates early. Therefore estimation conditional on reaching the full sample size is our focus. We return to this important point in Section 4.

In Sections 2 through 6 we discuss estimation of a single binomial probability. The two-stage group sequential design is described in Section 2 and estimators are defined. Simulation studies described in Section 3 are used to compare them. We contrast unconditional estimation with conditional estimation in Section 4 and argue that our conditional estimators may be useful even if unconditional estimation is required. Inference is the topic of Section 5 where methods to construct confidence intervals with the bootstrap are described along with associated hypothesis testing procedures. Some applications illustrate our approach in Section 6. In Section 7 we return to the context of studying performance of diagnostic tests, illustrating in detail our procedures when two binomial parameters, (S, F), are simultaneously under consideration. Some closing remarks and directions for further work are provided in Section 8.

2. Design and Estimation

2.1. Design

We consider a single binomial probability, $S = P[Y = 1]$, which could denote the response rate in a phase 2 therapeutic study, though we use terminology from diagnostic studies here and let S be the sensitivity, i.e., the rate of positive biomarker responses in cases. Suppose that sensitivities below γ_0 are undesirable while values at or above γ_1 are desirable. In particular in phase 2 we will need to show that $S > \gamma_0$, the maximal undesirable sensitivity, in order to proceed with phase 3 development. On the other hand, γ_1 is minimally desirable in the following sense: if $S > \gamma_1$ we certainly want to proceed with development while for $S\epsilon(\gamma_0, \gamma_1)$, the equivocal region of sensitivities, there is little enthusiasm. In terms of hypotheses upon which to base study design, we write

$$H_0 : S \leq \gamma_0 \text{ versus } H_1 : S = \gamma_1.$$

High power is sought only if $S \geq \gamma_1$. As an example, for detection of ovarian cancer sensitivities below $\gamma_0 = 0.6$ would be undesirable since existing markers reach at least this level of detection while we seek markers with sensitivities of at least $\gamma_1 = 0.8$, since this would be a substantial improvement and worth investing resources for further research.

A single stage study will enroll n cases with disease and reject H_0 if the lower two-sided $(1-\alpha)$ confidence limit for S exceeds γ_0. For the purposes of study monitoring after m samples are evaluated, we propose to construct a two-sided $(1-\delta)\times 100\%$ confidence interval, and if the upper limit is less than γ_1, the study terminates. That is, if there is strong evidence that the sensitivity is below the minimally desirable level, the study will not continue to completion. Otherwise, the study continues to evaluate the remaining $n - m$ samples. This stopping rule is reasonable and easy to explain to investigators. Moreover, under H_1 there is only a small chance, $\delta/2$, of stopping early, suggesting that it will maintain statistical power relative to a single stage study.

2.2. Estimation at Study Completion

We now consider how to estimate the sensitivity, S, at the end of a completed phase 2 study. The data are denoted by $\{Y_i, i = 1, \ldots n\}$ with the index i indicating the order in which samples are evaluated. One option is to calculate the naïve estimator that ignores the early stopping procedure

$$\widehat{S}(all) = \sum_{i=1}^{n} Y_i/n.$$

However this is likely to be biased upward since it is contingent upon an adequately high response rate amongst the first m samples to result in completing the study. An unbiased estimator that is unaffected by the early stopping option uses only the second stage samples,

$$\widehat{S}(stage2) = \sum_{i=m+1}^{n} Y_i/(n-m).$$

Because of the relatively small sample size this estimator is likely to suffer from imprecision.

We now propose an unbiased estimator that incorporates data from both stages. Having used $\widehat{S}$ to denote simple proportions we write this more complicated estimator as

$$\widehat{U} = E(\widehat{S}(stage2)|\widehat{S}(all), C = 1)$$

where $C = 1$ indicates that the criterion for continuation past the first stage was passed.

Result 1

Conditional on $C = 1$, $\widehat{S}(all)$ is a complete and sufficient statistic for the distribution of $\widehat{S}(stage2)$.

Proof

For sufficiency we need to show that the conditional distribution of $\widehat{S}(stage2)$ given $\widehat{S}(all)$ and $C = 1$ does not depend on the parameter S. But conditional on $\widehat{S}(all)$, the distribution of $\widehat{S}(stage2)$ is hypergeometric $(n, n-m, \widehat{S}(all))$. Moreover, since $\widehat{S}(stage1) = m^{-1}\{n\widehat{S}(all) - (n-m)\widehat{S}(stage2)\}$, C can be determined from $\widehat{S}(all)$ and $\widehat{S}(stage2)$. The distribution of $\widehat{S}(stage2)$ conditioning on $C = 1$ in addition to $\widehat{S}(all)$ can be derived from the distribution of $\widehat{S}(stage2)$ conditioning on $\widehat{S}(all)$:

$$P(\widehat{S}(stage2)|\widehat{S}(all), C = 1) = I(C = 1)\frac{P(\widehat{S}(stage2)|\widehat{S}(all))}{P(C = 1|\widehat{S}(all))}.$$

Therefore, since $P(\widehat{S}(stage2)|\widehat{S}(all))$ does not depend on S, neither does $P(\widehat{S}(stage2)|\widehat{S}(all), C = 1)$. The proof of completeness follows from detailed tedious arguments given in Appendix A of Jung and Kim [5].

■

Corollary

$\widehat{U}$ is the unique minimum variance unbiased estimator of S among all estimators that are unbiased conditional on $C = 1$.

Proof

This follows from the fact that $\widehat{S}(stage2)$ is independent of C and hence conditionally unbiased

$$E(\widehat{S}(stage2)|C=1) = E(\widehat{S}(stage2)) = S$$

and the Rao-Blackwell theorem [6].

■

Two other estimators are inspired by Whitehead [**?**]. They adjust $\widehat{S}(all)$ for bias caused by the early termination option. The median adjusted estimator is

$$\widehat{W}_{\text{med}} = \gamma : P_\gamma(\widehat{S}^*(all) > \widehat{S}(all)|C^*=1) = 0.5, \quad (1)$$

where the $*$ superscript denotes random variables generated from our study design using γ as the binomial response probability, $\gamma = P_\gamma(Y=1)$. Intuitively, $\widehat{W}_{\text{med}}$ is the response probability for which the observed naïve proportion is the median naïve proportion in studies that continue to completion. A mean adjusted estimator is similarly defined

$$\widehat{W}_{\text{mean}} = \gamma : E_\gamma(\widehat{S}^*(all)|C^*=1) = \widehat{S}(all), \quad (2)$$

2.3. Calculations

We calculate $\widehat{U}, \widehat{W}_{\text{med}}$ and $\widehat{W}_{\text{mean}}$ numerically using simulations. For $\widehat{U}$, we noted earlier that the conditional distribution of $\widehat{S}(stage2)$ given $\widehat{S}(all)$ is hypergeometric. Therefore, in each of K simulations we sample m cases at random from the n available to simulate the first stage data, and the remaining $n-m$ simulate the second stage data. Accordingly, in the k^{th} simulation, values of $\widehat{S}^k(stage2), \widehat{S}^k(stage1)$ and C^k are calculated. Averaging $\widehat{S}^k(stage2)$ across simulations where $C^k = 1$ yields $\widehat{U}$. Exact calculations using the hypergeometric distribution are also possible.

More extensive computations are required for calculating $\widehat{W}_{\text{med}}$ and $\widehat{W}_{\text{mean}}$, because they involve searching for γ to satisfy (1) and (2), respectively. For each value of γ considered we simulate two stage studies with binomial probability equal to γ and select $\widehat{S}^k(all)$ for studies that satisfy $C^k = 1$. We calculate $P_\gamma(\widehat{S}^*(all) > \widehat{S}(all)|C^*=1)$ as the proportion of $\widehat{S}^k(all)$ exceeding the observed $\widehat{S}(all)$, and $E_\gamma(\widehat{S}^*(all)|C^*=1)$ is calculated as the mean of $\widehat{S}^k(all)$. $\widehat{W}_{\text{med}}$ is the γ for which $P_\gamma(\widehat{S}^*(all) > \widehat{S}(all)|C^*=1)$ is closest to 0.5 and $\widehat{W}_{\text{mean}}$ is the γ for which $E_\gamma(\widehat{S}^*(all)|C^*=1)$ is closest to $\widehat{S}(all)$. In our applications we used $K = 5000$ simulations to calculate $\widehat{U}$. Also for each γ, $P_\gamma(\widehat{S}^*(all) > \widehat{S}(all)|C^*=1)$ and $E_\gamma(\widehat{S}^*(all)|C^*=1)$ were calculated with $K = 5000$ simulations, we selected $\widehat{W}_{\text{med}}$=γ if $P_\gamma(\widehat{S}^*(all) > \widehat{S}(all)|C^*=1)$ was within 0.005 of 0.5, and $\widehat{W}_{\text{mean}}$=$\gamma$ if $E_\gamma(\widehat{S}^*(all)|C^*=1)$ was within 0.005 of $\widehat{S}(all)$.

3. Performance of Estimators

3.1. Initial Assessment

A single stage study to test $H_0 : S \leq \gamma_0 = 0.6$ with 90% power at $H_1 : S = \gamma_1 = 0.8$ and allowing type 1 error rate $\alpha = 0.05$ requires 42 cases according to asymptotic theory formulas

[9]. We simulated 1000 studies with $n = 40$ allowing for early termination after responses from $m = 20$ are observed if the upper two-sided 95% confidence limit for S [10] does not exceed $\gamma_1 = 0.8$. Results in Table 1 show estimates calculated from studies that complete enrollment of all 40 cases. If the true sensitivity is low, it is likely that the study will terminate early. For example, when $S = 0.6$, 59.2% of studies stop early while 40.8% continue to full enrollment of $n = 40$. Thus the means and standard deviations in the corresponding row of Table 1 relate to $40.8\% \times 1000 = 408$ studies. Consider first the naïve estimator, $\widehat{S}(all)$, that ignores the early stopping option. The anticipated upward bias is evident, and most pronounced when S is small. For example, when $S = 0.55$ the mean is 0.62, a substantial bias. When S is at the null hypothesis value, $S = .60$, the bias leads to a type 1 error rate that is twice the nominal value. The other naïve estimator using only second stage data, $\widehat{S}(stage2)$ is unbiased. However, its precision is low, a problem that is evident when the probability of early stopping is very small (i.e., S is large). Indeed when no studies terminate early (e.g., $S = 0.85$), we note that $\text{var}(\widehat{S}(stage2)) = \frac{n}{n-m}\text{var}(\widehat{S}(all))$ in general.

The conditional UMVUE, $\widehat{U}$, appears to maintain the best properties of both naïve estimators. Like $\widehat{S}(stage2)$, it is unbiased across all values of S. In addition, when early stopping is unlikely, its precision is comparable with $\widehat{S}(all)$. These results are encouraging.

The performances of the mean and median adjusted estimators are comparable with that of $\widehat{U}$. They substantially adjust for bias when S is low and are relatively precise when S is large. Despite their good performance we will not study them further here for the following reasons: (i) there is no theory to support them, unlike $\widehat{U}$, which is theoretically unbiased. A close look at Table 1 indicates some residual bias in $\widehat{W}_{\text{med}}$; (ii) Their computation is more difficult than that for $\widehat{U}$; and (iii) Our preliminary simulation studies in Table 1 indicate no particular improvement in their performances over that of $\widehat{U}$.

3.2. Additional Scenarios

Table 2 shows additional simulation results for studies with larger sample sizes. The top panel is motivated by the context of ovarian cancer screening where a very high specificity is desired. False positive screening tests result in subjects undergoing laproscopic surgery, so the rate must be kept very small. Specificity values at or above 0.98 are desired while values below 0.95 would be considered unacceptable. A single stage study would require $n = 230$ specimens from non-diseased subjects, and we consider early termination after evaluating half that number, $m = 115$. The bottom panel shows a setting similar to Table 1, but with $\gamma_1 = 0.70$ rather than $\gamma_1 = 0.80$. The results corroborate those in Table 1.

We also investigated choices of m other than $n/2$ (Table 3). Since the criterion for early stopping is based on the upper confidence limit for S not exceeding γ_1, the probability of early stopping for $S < \gamma_1$ is larger when more data is available at stage 1. On the other hand the bias in the naïve estimator $\widehat{S}(all)$ for studies that complete is larger with larger value of m. For example, with $m/n = 27/40$, if $S = 0.55$, 84% of studies terminate early and the expectation of $\widehat{S}(all)$ is 0.653. In contrast with $m = 13/40$, 59% of studies terminate early and the expectation of $\widehat{S}(all)$ is 0.592.

The conditional UMVUE, $\widehat{U}$, is by definition conditionally unbiased, regardless of m, as is borne out again by Table 3. Its variance, however, is larger with larger values of m, a point we return to in section 6.

4. Unconditional Estimation

Estimation following group sequential designs for phase 2 therapeutic trials has been studied at least since 1958 [11]. We refer to Jennison and Turnbull [12] and Emerson and Fleming [13] as key papers. The UMVUE for binary response data was studied recently by Jung and Kim [5], although related results for the mean of a normal distribution have long been available [**?**]. For a two stage study with binary response, the UMVUE is easy to calculate and is likely the popular choice so we consider it here.

The literature on group sequential designs considers that estimation occurs at the end of the study, i.e., at stage 1 if the study terminates there or at stage 2 if it continues. The unconditional UMVUE is defined as

$$\widetilde{U} = E(\widehat{S}(stage1)|\widehat{S}, stage)$$

where *stage* denotes the stopping stage and $\widehat{S}$ denotes the response rate calculated with all data collected in the study by the stopping stage. Thus,

$$\begin{aligned} \widetilde{U} &= \widehat{S}(stage1) && \text{if } C = 0 \\ \widetilde{U} &= E(\widehat{S}(stage1)|\widehat{S}(all), C = 1) && \text{if } C = 1. \end{aligned}$$

Averaging over all studies, including those that terminate at stage 1, $\widetilde{U}$ is unbiased because $\widehat{S}(stage1)$ is unbiased. However, if interest is in estimation only for studies that complete both stages, then $\widetilde{U}$ is biased upward, i.e., $E(\widetilde{U}|C = 1) > S$. Intuitively this follows from the fact that since $\widetilde{U}$ is marginally unbiased

$$S = E(\widetilde{U}) = E(\widetilde{U}|C = 0)P(C = 0) + E(\widetilde{U}|C = 1)P(C = 1)$$

and $E(\widetilde{U}|C = 0) = E(\widehat{S}(stage1)|C = 0)$, is the mean response in stage 1 restricted to studies that terminate early for lack of response, which, by definition, is biased low. Therefore $E(\widetilde{U}|C = 1)$ is biased high. For the scenarios considered in Tables 1 and 2 we calculated the conditional mean and sd of $\widetilde{U}$, shown in Table 4.

The estimates calculated from studies that complete stage 2 have substantial bias. In these settings the bias is at least as large as that of the naïve uncorrected estimator $\widehat{S}(all)$. In conclusion, if one is primarily interested in estimates of the response rate for studies that complete evaluation of all n samples, we suggest using the conditional UMVUE, $\widehat{U}$, over the traditional unconditional UMVUE, $\widetilde{U}$.

We focus on estimation in studies that do not terminate early because our purpose is to determine if and how to design the next study. In particular, they will be used in sample size calculations. If a study terminates early due to lack of response, we conclude that $S < \gamma_1$ and the biomarker is considered inadequate for further development. The estimate and its sampling variability are usually not of great interest.

Nevertheless, we believe that there may also be a role for the conditional UMVUE in the traditional group sequential design settings where estimation at the terminating stage is required, be it early or not. Define

$$\begin{aligned} U^* &= \widehat{S}(stage1) && \text{if } C = 0 \\ U^* &= \widehat{U} && \text{if } C = 1 \end{aligned}$$

The estimator U^* is equal to the traditional UMVUE if the study stops early and equal to the conditional UMVUE if the study completes. It is unbiased conditional on completing both stages, but is not marginally unbiased. Observe that

$$\begin{aligned}
&E(\widetilde{U}-S)^2 - E(U^*-S)^2 \\
&= P(C=1)\{E(\widetilde{U}-S)^2|C=1) - E((\widehat{U}-S)^2|C=1)\} \\
&= P(C=1)\{\mathrm{var}(\widetilde{U}|C=1) + \mathrm{bias}^2(\widetilde{U}|C=1) - \mathrm{var}(\widehat{U}|C=1))\}.
\end{aligned}$$

From Tables 1 and 4 we see that when S is low the bias in $\widetilde{U}$ dominates and U^* has smaller (unconditional) mean squared error than $\widetilde{U}$. However, when the response rate is high there is little bias in any of the estimates, including $\widetilde{U}$. In these cases the small conditional variance of $\widetilde{U}$ is attractive. In summary in terms of mean squared error, $\widetilde{U}$ performs better than U^* when the response rate is high but worse than U^* when the response rate is low. In phase 2 biomarker development studies we anticipate that low response rates will be more common. Hence we recommend $\widehat{U}$ and U^* for conditional and unconditional estimation, respectively.

5. Inference with the Conditional UMVUE

5.1. Confidence Intervals

We seek not only an estimate of S at the end of a completed study, but a confidence interval as well. For this we propose two resampling methods. Note that simple bootstrapping, resampling at random from $\{Y_i, i = 1, \ldots, n\}$ is not valid under a group sequential design. The responses in the observed data are biased due to having passed the early stopping criterion.

In the first resampling approach we use the estimated population response rate, $\widehat{U}$, to simulate $b = 1, \ldots, B$ group sequential studies with our design. Selecting those for which the continuation criterion is satisfied, $C^b = 1$, and calculating the corresponding statistics, $\widehat{U}^b$, we use their empirical distribution as an estimate of the sampling distribution of $\widehat{U}$, conditional on $C = 1$. The $\alpha/2$ and $\{1-\alpha/2\}$ empirical quantiles are used as confidence limits. We call this approach the parametric bootstrap because data are simulated with response probability $\widehat{U}$, though we note that no parametric assumptions are made.

We call the second approach the nonparametric bootstrap. Here in the b^{th} resampling, we resample n responses with replacement from the n observed, and calculate $\widehat{U}^b = E(\widehat{S}^b(stage2)|\widehat{S}^b(all), C = 1)$. Again, quantiles of the distribution of $\widehat{U}^b$ are used as confidence limits.

Table 5 shows coverage of confidence intervals under the scenarios and design of Table 1 ($n = 40, m = 20$) and Table 2 ($n = 220, m = 110$). Due to the extensive computation involved, we used $K = 500$ (rather than $K = 5000$) in calculating $\widehat{U}$. We see that coverage is reasonably close to the nominal 95% level for both bootstrap methods, but somewhat lower for the parametric bootstrap than for the nonparametric bootstrap. Correspondingly, the standard deviation tends to be slightly underestimated with the parametric methods but overestimated with the nonparametric bootstrap.

5.2. Power

We can use the confidence interval to formally test $H_0 : S \leq \gamma_0$. Recall that only values of S at or above γ_1 are considered desirable so we study power for $S \geq \gamma_1$. Compared to a fixed sample size study of n samples, power is reduced by the group sequential design for two reasons. First, by allowing studies to stop at stage 1, power is lost if some fraction of those would have proceeded to yield a positive conclusion had they not been terminated. Second, power is lost if the width of the confidence interval for S is wider when it is based on an adjusted estimator than when it is based on the naïve estimator.

The stopping criterion used plays a large role in regards to the first power loss mechanism (although the discussion so far in this paper does not rely on it). Our proposed criterion is to stop after evaluating m subjects if the upper two-sided $(1 - \delta)$ confidence limit lies below γ_1. Therefore the associated power loss at $S \geq \gamma_1$ is no more than $\delta/2$. It is likely to be less than $delta/2$ even when $S = \gamma_1$ because some of those terminated studies would presumably be in the fraction of studies deemed to be negative even if enrollment continued to n samples.

Table 6 displays the power of the standard analysis based on $\widehat{S}(all)$ in a fixed sample size design. That is, the power if all studies continued to $n = 40$ regardless of interim results. Also shown are the powers associated with designs that allow early stopping and use confidence intervals based on $\widehat{U}$ at the end of stage 2 for testing $H_0 : S \leq \gamma_0$. We see that two-stage studies using the parametric bootstrap confidence interval have power comparable with the fixed sample size power. That is, their benefit, which is to terminate early those studies in which markers have poor performance, is gained without substantial loss in their capacity to identify good markers as such. The nonparametric bootstrap confidence interval seems to not achieve the same power, due presumably to their over conservative nature.

Our focus here is on power achieved when $S \geq \gamma_1$. We defined γ_1 as the minimum desirable value of S meaning that values of S less than γ_1 are not desirable. We therefore do not seek high power for S in the range (γ_0, γ_1). The two-stage design in fact ensures that power in this range is reduced relative to a single stage study and we view this as a good attribute. Nevertheless, it underscores that the choice of γ_1 should be made judiciously and must be the minimum desirable value. Similarly the choice of γ_0 is crucial, γ_0 is the maximal unacceptable values. Values in the equivocal range (γ_0, γ_1) may be reluctantly acceptable but are not desirable. Specifying (γ_0, γ_1) is often a difficult challenge in practice.

6. Illustrations

To fix ideas we now provide in Table 7 a few simple illustrations using simulated data. For each we use the design of Table 1, i.e., $n = 40, m = 20, \gamma_0 = 0.6, \gamma_1 = 0.8$. In the first illustration, at the interim analysis only 5 of 20 samples have a positive response. The 95% confidence interval for S is (0.11,0.47). Since the upper limit is below $\gamma_1 = 0.80$ the study terminates early.

In the second illustration, the response rate at the interim analysis is much higher, with 18 of 20 responses positive and 95% confidence interval for S, (0.70,0.97). The study continues to accrue responses from 20 more subjects, of which 17 responses are positive, yielding $\widehat{S}(all) = 35/40 = 0.88$. The estimates that adjust for the early stopping option $\widehat{U}$, $\widehat{W}_{\text{med}}$, and $\widehat{W}_{\text{mean}}$, are all equal to 0.88. We calculate 95% confidence intervals for S based on $\widehat{U}$ as (0.75,0.98) with the nonparametric bootstrap and (0.77,0.97) with the parametric bootstrap.

In either case we conclude that the response rate exceeds the unacceptable level of 0.60. In fact it appears to be within the desirable range and deliberations about the next phase of biomarker development ensue.

Six further illustrations are shown in Table 7. Two, studies 3 and 8, terminate early. Two, studies 4 and 5, continue to completion but do not yield positive conclusions about marker performance. Study 6 is inconclusive. Unfortunately when the design stipulates only 90% power, even with a fixed sample size design inconclusive studies can occur. Study 7 indicates a 100% response rate (CI=(0.84,1.00)) in the initial stage. One might be tempted to terminate at that point. However a more prudent approach is to collect additional data, and indeed the second stage data tempers enthusiasm somewhat, providing adjusted estimates of 0.85 for the response rate.

The results in Table 7 suggest relationships between $\widehat{U}$, $\widehat{S}(all)$ and $\widehat{S}(stage2)$. In particular when $\widehat{S}(all)$ is large, we find that $\widehat{U} \approx \widehat{S}(all)$. This is reasonable since $\widehat{U} = E(\widehat{S}(stage2)|\widehat{S}(all), C = 1)$, and when $\widehat{S}(all)$ is large it follows that $C = 1$ with high probability so that $\widehat{U} \approx E(\widehat{S}(stage2)|\widehat{S}(all)) = \widehat{S}(all)$. On the other hand, when $\widehat{S}(all)$ is small, $\widehat{U} \approx \widehat{S}(stage2)$. This makes sense because a small value of $\widehat{S}(all)$ together with the knowledge that the continuation criterion was passed indicates that $\widehat{S}(stage1)$ was close to the critical value for continuation. This in turn informs about $\widehat{S}(stage2)$, which is equal to $(n-m)^{-1}\{n\widehat{S}(all) - m\widehat{S}(stage1)\}$.

These observations also have implications for the performance of $\widehat{U}$ relative to $\widehat{S}(all)$ and $\widehat{S}(stage2)$ in general. When the true response rate is small, $\widehat{U}$ behaves similarly to $\widehat{S}(stage2)$, while $\widehat{U}$ behaves more like $\widehat{S}(all)$ when the response rate is high. The conditional standard deviations reported in Tables 1 and 2 bear this out. In addition, we see in Table 3 that differing values of m have little impact on the conditional performance of $\widehat{U}$ when S is large, but greater impact when S is small. In the former case, $\widehat{U}$ is similar to $\widehat{S}(all)$, which is unaffected by m. In the latter case, $\widehat{U}$ is similar to $\widehat{S}(stage2)$, which is more variable when the second stage sample size $n-m$ is small.

7. Simultaneous Inference for Sensitivity and Specificity

We now return to the context of evaluating a diagnostic or screening marker where considerations of both sensitivity (S) and specificity ($1-F$) must be made simultaneously. Let γ_0 and η_0 denote maximal unacceptable values of sensitivity and specificity, respectively, while γ_1 and η_1 denote minimum desirable values. The design and analysis of a fixed sample size study are described in detail in Pepe, pages 218–220 [9].

Briefly, using subscripts D and $\bar{D}$ to denote cases and controls, a fixed sample size study enrolls n_D cases and $n_{\bar{D}}$ controls. A joint confidence $(1-\alpha)$ rectangle for $(S, 1-F)$ is calculated as the Cartesian product of $(1-\alpha^*)$ confidence intervals for S and $1-F$ where $(1-\alpha^*) = (1-\alpha)^{\frac{1}{2}}$. A positive conclusion is drawn about marker performance if the lower limit for S exceeds γ_0 *and* the lower limit for $1-F$ exceeds η_0. The sample sizes are chosen so that when $S = \gamma_1$ and $1-F = \eta_1$ the probability is high, $1-\beta$, that both lower confidence limits exceed the thresholds γ_0 and η_0. To illustrate, with $(\gamma_0, \gamma_1) = (0.6, 0.8)$ and $(\eta_0, \eta_1) = (0.95, 0.98)$, values appropriate for an ovarian cancer screening marker, the sample size formulae (Pepe equations (8.2) and (8.3)) [9] yield $n_D = 78$ and $n_{\bar{D}} = 572$ to achieve size

$\alpha = 0.05$ and power $1 - \beta = 0.90$.

The study could be designed to terminate after half the cases and half the controls are evaluated if the joint confidence rectangle does not contain both minimally desirable values for sensitivity and specificity (γ_1, η_1). Otherwise the study continues to complete enrollment at which time the conditional UMVUE estimates of S and $1-F$ are calculated. Corresponding $(1-\alpha^*)$ level confidence intervals yield a joint $(1-\alpha)$ confidence rectangle. A positive conclusion about marker performance ensues if the $(1-\alpha^*)$ confidence intervals for S and $1-F$ exclude γ_0 and η_0 respectively. Table 8 shows the results of some simulation studies.

We see that the study is likely to stop early if the true sensitivity or the true specificity is low but likely to continue if both are at the minimally desirable value. Coverage for the 95% parametric bootstrap confidence rectangle was slightly lower than the nominal rate, although, four of the five scenarios achieved at least 93% coverage. We observe that the study has very low rejection rate when $S < \gamma_0$ or $1 - F < \eta_0$, as desired. When $S \geq \gamma_1$ and $1 - F \geq \eta_1$, we desire high power. We observe that the 81% unconditional power when $S = \gamma_1$ and $1 - F = \eta_1$ represents a 9% decrease from the fixed sample size power.

There are many variations on study design that could be explored. Our choice of interim analysis when both $m_D = n_D/2$ cases and $m_{\bar{D}} = n_{\bar{D}}/2$ controls are evaluated is arbitrary. One need not enroll cases and controls at the same relative rates. In fact one option would be to enroll all $m_{\bar{D}}$ controls first before using samples from cases. If the study terminates early because of poor specificity, precious samples from cases are saved. Yet inference is the same. In practice however, one may want to mix up the order of cases and controls somewhat in order to expose testers to heterogeneous samples and to aid with blinding. In a similar vein, for S and F we have chosen equal adjusted significance levels α^* for construction of their joint confidence rectangle. Unequal values can be employed. Letting α^*_D and $\alpha^*_{\bar{D}}$ denote adjusted values for S and F, respectively, the requirement for joint $1 - \alpha$ coverage is that

$$(1 - \alpha^*_D)(1 - \alpha^*_{\bar{D}}) = 1 - \alpha.$$

However, arguments leading to particular choices of $(\alpha^*_{\bar{d}}, \alpha^*_D)$ that are unequal have not been developed yet.

8. Conclusions and Remarks

We have proposed the conditional UMVUE, $\widehat{U}$, for estimation at the end of a phase 2 group sequential study that does not terminate early. It is appropriate when unbiased estimation is required from studies that reach full enrollment. In our experience with phase 2 biomarker studies, calculation of estimates is of less concern in studies that terminate early, where the conclusion is simply that the biomarker is inadequate for further development and sufficient data for precise estimation is not available in any case. Hence we focused on estimators with good properties conditional on full enrollment. These considerations seem equally relevant for phase 2 group sequential therapeutic studies and we suggest $\widehat{U}$ for application in that context too. We noted that the standard unconditional UMVUE, $\widehat{U}$, can show considerable conditional bias. The naive unadjusted estimator is also conditionally biased, although in terms of mean squared error and coverage of confidence intervals it performed reasonably well in most scenarios we considered.

Conditional inference has been discussed from a decision theoretic point of view [16] and was recently applied to group sequential designs [17]. In particular, Strickland and Casella considered the conditional confidence interval, (γ_L, γ_U), where limits are defined in a similar vein to Whitehead's median adjusted estimator but using target probabilities of $\alpha/2$ and $1-\alpha/2$ instead of 0.5 in equation (1). For normally distributed data they proved an optimally result for these intervals. This suggests that they be examined for binary data and compared to the confidence intervals based on $\widehat{U}$ that were proposed here. They also noted for normally distributed data that the conditional performance of unconditional confidence intervals can be very poor.

The design of a group sequential study requires choosing values for the confidence level at the interim analysis, $1-\delta$, and for the stage 1 sampling fraction, m/n. The probability of early stopping when $S = \gamma_1$ is $\delta/2$. Since this should be small, we chose $\delta = 0.05$ in our illustrations. Another attractive feature of the choice $\delta = 0.05$ is that the practice of calculating 95% confidence intervals is familiar to our collaborators and they can easily accept abandoning a biomarker study if the 95% confidence interval does not contain γ_1. That is the early stopping criterion makes sense to collaborators. Observe that one can also consider δ as a type 1 error for testing $H : S = \gamma_1$ based on m observations. The corresponding power is the probability of early stopping under $H : S = \gamma_0$. Larger values of m give rise to higher power. An optimal choice of m might be based on minimizing the expected sample size, which requires postulating a prior probability distribution for S.

This paper considered biomarkers with dicohotomous values. However, most biomarkers are measured on a continuous scale and performance is evaluated with the receiver operating characteristic(ROC) curve. Methods for estimating the ROC curve following a group sequential phase 2 study would be worthy of research.

REFERENCES

1. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson M, Thornquist M, Winget M and Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 2001; **93** 1054–1061.
2. Institute of Medicine. Workshop Summary: Developing Biomarker-based Tools for Cancer Screening, Diagnosis, and Therapy–The State of the Science, Evaluation, Implementation, and Economics. National Academies Press, 2006.
3. Mazumdar M. Group Sequential Design for Comparative Diagnostic Accuracy Studies: Implications and Guidelines for Practitioners. *Medical Decision Making* 2004; **24** 525–533.
4. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. *Statistics in Medicine* 2003; **22** 727–739.
5. Jung S-H and Kim K-M. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* 2004; **23** 881–896.
6. Bickel PJ and Doksum KA. *Mathematical Statistics: basic ideas and selected topics.* Holden-Day, San Francisco, 1977, page 121.
7. Whitehead J. *The Design and Analysis of Sequential Clinical Trials.* Elis Horwood Limited, Chichester, 1983.
8. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986; **73** 461–471.
9. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, 2003.
10. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**:101-133.
11. Armitage P. Sequential methods in clinical trials. *American Journal of Public Health* 1958; **48** 1395–402.

12. Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multi-stage test with application to MIL-STD 105D and medical trials. *Technometrics* 1983; **25** 49–58.
13. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77** 875-892.
14. Emerson SS. Computation of the uniform minimum variance unbiased estimator of a normal mean following a group sequential trial. *Computational Biomedical Research* 1993; **26** 68–73.
15. Emerson SS, Kittelson JM. A computationally simpler algorithm for the UMVUE of a normal mean following a group sequential trial. *Biometrics* 1997; **53** 365–359.
16. Kiefer J. Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association* 1977; **72**:789–827.
17. Strickland PA, Casella G. Conditional Inference Following Group Sequential Testing. *Biometrical Journal* 2003; **45**:515–526.

Table I. Results of simulation studies with $n = 40$ and early termination option at $m = 20$. Shown are mean (sd) of estimated sensitivities in studies that reached completion. One thousand simulations per true sensitivity, S.

True S	% early stopping	$\widehat{S}(all)$	$\widehat{S}(stage2)$	$\widehat{W}_{\text{med}}$	$\widehat{W}_{\text{mean}}$	$\widehat{U}$
55	73%	0.623	0.547	0.576	0.550	0.553
		(0.062)	(0.112)	(0.095)	(0.098)	(0.102)
60	59%	0.654	0.606	0.619	0.599	0.604
		(0.061)	(0.109)	(0.091)	(0.094)	(0.096)
65	41%	0.685	0.647	0.662	0.644	0.650
		(0.062)	(0.102)	(0.086)	(0.090)	(0.091)
70	22%	0.720	0.698	0.709	0.693	0.699
		(0.062)	(0.099)	(0.081)	(0.085)	(0.084)
75	8%	0.761	0.756	0.760	0.746	0.751
		(0.061)	(0.097)	(0.073)	(0.077)	(0.075)
80	3%	0.804	0.799	0.809	0.798	0.8000
		(0.060)	(0.090)	(0.067)	(0.070)	(0.067)
85	1%	0.850	0.851	0.858	0.848	0.849
		(0.057)	(0.083)	(0.059)	(0.061)	(0.059)

Table II. Results of additional simulation studies with larger sample sizes. 1000 studies were simulated for each scenario.

True S	% early stopping	$\widehat{S}(all)$	$\widehat{S}(stage2)$	$\widehat{W}_{\text{med}}$	$\widehat{W}_{\text{mean}}$	$\widehat{U}$
		$\gamma_0 = 0.95$, $\gamma_1 = 0.98$, $n = 230$, $m = 115$				
0.90	98%	0.924	0.889	0.890	0.891	0.888
		(0.015)	(0.029)	(0.029)	(0.027)	(0.028)
0.95	50%	0.957	0.948	0.950	0.948	0.948
		(0.012)	(0.022)	(0.019)	(0.022)	(0.020)
0.965	22%	0.968	0.965	0.966	0.965	0.965
		(0.011)	(0.017)	(0.014)	(0.016)	(0.015)
0.98	3%	0.981	0.981	0.981	0.981	0.980
		(0.008)	(0.012)	(0.009)	(0.009)	(0.009)
0.99	0%	0.990	0.991	0.990	0.990	0.990
		(0.007)	(0.009)	(0.006)	(0.007)	(0.007)
		$\gamma_0 = 0.60$, $\gamma_1 = 0.70$, $n = 220$, $m = 110$				
0.55	91%	0.595	0.553	0.560	0.553	0.555
		(0.021)	(0.036)	(0.033)	(0.037)	(0.037)
0.60	61%	0.621	0.597	0.600	0.594	0.597
		(0.028)	(0.050)	(0.042)	(0.043)	(0.044)
0.65	21%	0.659	0.652	0.652	0.649	0.651
		(0.028)	(0.045)	(0.036)	(0.038)	(0.036)
0.70	2%	0.700	0.699	0.700	0.698	0.698
		(0.030)	(0.044)	(0.032)	(0.033)	(0.032)
0.75	0%	0.750	0.751	0.751	0.750	0.750
		(0.030)	(0.040)	(0.030)	(0.030)	(0.030)

Table III. Results of additional simulation studies with various choices for m/n, the fraction of total sample size that enters into the first stage evaluation. Data were simulated using the same context as Table 1, $\gamma_0 = 0.60$, $\gamma_1 = 0.80$, $n = 40$. 1000 simulated studies per scenario.

True S	m	% early stopping	$\widehat{S}(all)$	$\widehat{S}(stage2)$	$\widehat{U}$
0.55	13	59%	0.592	0.549	0.550
			(0.065)	(0.090)	(0.085)
0.60	13	41%	0.628	0.597	0.597
			(0.071)	(0.098)	(0.090)
0.65	13	26%	0.668	0.646	0.647
			(0.067)	(0.091)	(0.082)
0.70	13	19%	0.709	0.693	0.695
			(0.069)	(0.090)	(0.081)
0.75	13	7%	0.756	.747	0.749
			(0.066)	(0.084)	0.073)
0.80	13	2%	0.804	0.800	0.801
			(0.061)	(0.078)	(0.065)
0.85	13	0%	0.852	0.852	0.851
			(0.057)	(0.068)	(0.059)
0.55	27	84%	0.653	.563	0.560
			(0.049)	(0.138)	(0.112)
0.60	27	68%	0.670	0.588	0.593
			(0.056)	(0.136)	(0.119)
0.65	27	50%	0.698	0.641	0.651
			(0.054)	(0.130)	(0.101)
0.70	27	28%	0.728	0.696	0.700
			(0.057)	(0.127)	(0.091)
0.75	27	14%	0.761	0.748	0.748
			(0.060)	(0.120)	(0.081)
0.80	27	3%	0.803	0.793	0.798
			(0.059)	(0.112)	(0.068)
0.85	27	1%	0.852	0.852	0.851
			(0.056)	(0.099)	(0.058)

Table IV. Performance of the traditional unconditional UMVUE in studies that complete evaluation of all n subjects. The scenarios and simulations are the same as in Tables 1 and 2.

	$\gamma_0 = 0.60$	$\gamma_1 = 0.80$	$n = 40$	$m = 20$			
True S	0.550	0.600	0.650	0.700	0.750	0.800	0.850
mean $(\widetilde{U})$	0.692	0.705	0.720	0.741	0.771	0.808	0.851
sd $(\widetilde{U})$	0.023	0.029	0.035	0.043	0.050	0.054	0.055
	$\gamma_0 = 0.95$	$\gamma_1 = 0.98$	$n = 230$	$m = 115$			
True S	0.900	0.950	0.965	0.980	0.990		
mean $(\widetilde{U})$	0.960	0.966	0.972	0.981	0.990		
sd $(\widetilde{U})$	0.002	0.005	0.007	0.008	0.007		
	$\gamma_0 = 0.60$	$\gamma_1 = 0.70$	$n = 220$	$m = 110$			
True S	0.550	0.600	0.650	0.700	0.750		
mean $(\widetilde{U})$	0.635	0.646	0.667	0.701	0.750		
sd $(\widetilde{U})$	0.005	0.013	0.021	0.028	0.029		

Table V. Estimated mean and sd of $\widehat{U}$ and coverage of 95% confidence intervals (CI) based on the 2.5^{th} and 97.5^{th} percentiles of the nonparametric and parametric bootstrap distributions of $\widehat{U}$. Shown are results for completed studies in 500 simulations with $\gamma_0 = 0.60$ and $\gamma_1 = 0.80$ or $\gamma_1 = 0.70$. The number of bootstrap samples per simulated study was chosen to be $min(n_b, 5000)$ where n_b yielded 500 resampled datasets that satisfied C=1

True S	number of completed studies	mean ($\widehat{U}$)	sd ($\widehat{U}$)	Parametric Bootstrap mean ($\widehat{\text{sd}}$)	Parametric Bootstrap CI coverage (%)	Nonparametric Bootstrap mean ($\widehat{\text{sd}}$)	Nonparametric Bootstrap CI coverage (%)
			$\gamma_0 = 0.60$ $\gamma_1 = 0.80$ $n = 40$ $m = 20$				
0.55	131	0.562	0.092	0.097	92.4	0.123	98.5
0.60	207	0.595	0.111	0.094	92.7	0.116	95.7
0.65	285	0.649	0.092	0.089	93.7	0.107	97.9
0.70	394	0.697	0.083	0.083	95.4	0.097	96.2
0.75	453	0.749	0.077	0.075	93.6	0.085	94.7
0.80	479	0.801	0.068	0.066	95.0	0.072	96.0
0.82	491	0.818	0.065	0.063	93.1	0.067	93.5
0.85	497	0.850	0.059	0.057	93.8	0.059	94.8
			$\gamma_0 = 0.60$ $\gamma_1 = 0.70$ $n = 220$ $m = 110$				
0.55	48	0.542	0.048	0.044	89.4	0.053	97.9
0.60	204	0.602	0.041	0.041	94.6	0.049	97.5
0.65	378	0.649	0.038	0.037	93.7	0.042	97.6
0.70	486	0.702	0.032	0.032	95.5	0.034	95.7
0.72	499	0.721	0.028	0.031	96.6	0.032	96.6
0.75	499	0.751	0.030	0.029	94.2	0.030	95.0

Table VI. Power based on $\widehat{S}(all)$ in a fixed sample size study of n subjects and power based on $\widehat{U}$ in studies that allow early termination. Early stopping uses $1-\delta$ confidence interval at the interim analysis. Power for $\widehat{U}$ is the proportion of studies that reach complete enrollment and 95% confidence interval does not include γ_0. Scenarios of Tables 1 and 2 (lower panel) are employed, $m = n/2$, $\gamma_0 = 0.60$ and $\delta = 0.05$. 500 simulated studies. Values of $\widehat{U}$ calculated with $K = 500$.

S	n	Early Stopping (%)	P-BS($\widehat{U}$)	NP-BS($\widehat{U}$)	Logit($\widehat{S}(all)$)
0.80	40	4.2%	0.722	0.638	0.710
0.82	40	1.8%	0.804	0.724	0.808
0.85	40	0.6%	0.918	0.870	0.920
0.70	220	2.8%	0.802	0.708	0.872
0.72	220	0.2%	0.942	0.904	0.974
0.75	220	0.2%	0.992	0.982	1.000

NP-BS($\widehat{U}$) nonparametric bootstrap; P-BS($\widehat{U}$) parametric bootstrap; Logit($\widehat{S}(all)$) normal approximation to the distribution of logit($\widehat{S}(all)$).

Table VII. Eight simulated studies with $n = 40$, $m = 20$, $\gamma_0 = 0.6$, $\gamma_1 = 0.8$. 95% confidence intervals are calculated based on $\widehat{U}$ with the parametric (pCI) or nonparametric (npCI) method.

Study	$\widehat{S}(stage1)$	CI($stage1$)	$\widehat{S}(stage2)$	$\widehat{S}(all)$	$\widetilde{U}$	$\widehat{W}_{\text{med}}$	$\widehat{W}_{\text{mean}}$	$\widehat{U}$	npCI	pCI
1	0.25	(0.11,0.47)	–	–	–	–	–	–	–	–
2	0.90	(0.70,0.97)	0.85	0.88	0.88	0.88	0.88	0.87	(0.77,0.98)	(0.78,0.98)
3	0.60	(0.39,0.78)	–	–	–	–	–	–	–	–
4	0.70	(0.48,0.86)	0.55	0.63	0.69	0.58	0.57	0.56	(0.28,0.77)	(0.34,0.75)
5	0.75	(0.53,0.89)	0.40	0.58	0.67	0.50	0.46	0.47	(0.15,0.72)	(0.25,0.68)
6	0.65	(0.43,0.82)	0.85	0.75	0.76	0.75	0.74	0.74	(0.56,0.88)	(0.56,0.88)
7	1.00	(0.84,1.00)	0.70	0.85	0.85	0.86	0.85	0.85	(0.71,0.95)	(0.72,0.95)
8	0.50	(0.30,0.70)	–	–	–	–	–	–	–	–

Table VIII. Simulated studies using a two-stage design with ($n_D = 78, m_D = 39, \gamma_0 = 0.6, \gamma_1 = 0.8$) and ($n_{\bar{D}} = 572, m_{\bar{D}} = 286, \eta_0 = 0.95, \eta_1 = 0.98$). Coverage and power shown for the conditional UMVUE estimators of S and F with parametric bootstrapped confidence intervals. Nominal coverage probability=95%

S	F	% Early Termination	Conditional† Joint Coverage	Conditional† Power	Unconditional Power	Fixed Sample* Power
0.6	0.95	95%	92%	0.00	0.00	0.00
0.6	0.98	75%	90%	0.02	0.01	0.02
0.8	0.95	77%	96%	0.01	0.00	0.00
0.8	0.98	2%	94%	0.83	0.81	0.90
0.7	0.97	38%	94%	0.09	0.06	0.13

† restricted to studies that complete both stages; * no option for early termination.