



Monitored Progress

Bruce Stouch at Premier Research assesses the value of monitoring for futility in fixed and adaptive designs using conditional and predictive power

Bruce Stouch is the Vice President of Strategic Product Development at Premier Research and has over 25 years of experience in pharmaceutical research, medical devices, and *in vitro* diagnostics and is an expert statistical resource and US FDA liaison for clinical trial designs. Prior to joining Premier Research, Bruce was the Director of Biostatistics and Scientific Data Management at Johnson & Johnson for 12 years and the principle statistician for two major drugs for Sanofi-Synthelabo for three years. Bruce is currently the Director, Biostatistics and Clinical Epidemiology at the Philadelphia College of Osteopathic Medicine. Bruce received his PhD in Biostatistics and Epidemiology through a cooperative programme with Johnson & Johnson with post-graduate studies at Temple University, Jefferson Medical College, and Harvard School of Public Health.

Statistical monitoring procedures are used in a broad range of clinical trials to ensure the early availability of efficacious treatments while at the same time preventing spurious early termination of trials for apparent benefit that may later diminish. When properly designed, these procedures can also provide support for stopping a trial early when results do not appear promising, conserving resources and affording patients the opportunity to pursue other treatment options and avoid regimens which may have known and unknown risks while offering little benefit. By weighing these considerations against each other in the specific study situation at hand, a satisfactory monitoring procedure can be chosen.

Formal statistical monitoring rules serve as guidelines for data monitoring committees to provide for possible early trial termination. The pre-specified monitoring rules allow for the possibility of early stopping in response to positive trends that are sufficiently strong to substantiate the treatment differences the clinical trial was designed to detect. At the same time, the monitoring boundaries must guard against prematurely terminating a trial on the basis of initial positive results that may not be maintained with additional follow-up. Premature termination may also be recommended if the current trends in the data indicate that eventual positive findings are highly unlikely. Early termination for negative results may be called for if the data collected to date are sufficient to rule out the possibility of improvements in efficacy that are large enough to be clinically relevant. Alternatively, it may have become clear that study accrual, drug compliance, or follow-up compliance, among other factors have rendered the study incapable of discovering a difference, whether or not one exists. The focus of this article is to explore the construct of asymmetrical boundaries for interim monitoring of a fixed design and the design requirements for monitoring using a Bayesian approach. Additionally, the complement to monitoring using fixed boundaries by establishing optimistic and sceptical priors is also discussed relative to the complexity associated with modelling clinical assumptions and expectation.

DESIGNING AN INTERIM MONITORING STRATEGY

In designing an interim monitoring strategy, it is preferable to include a plan for stopping in the face of negative results at the

time the study protocol is developed. In particular, it is important to know what effect the plan will have on the power and significance level of the overall design. In calculating the asymmetric monitoring boundaries for a trial, many factors need to be estimated, and the expected results need to be modelled to ensure the boundary is clinically and ethically acceptable. Specifically, attrition, mortality (if appropriate), severity of the disease, the expected accrual rate, and the potential impact of the results from other concurrent studies, all need to be considered. The requirement that the monitoring boundaries actually join at the final analysis is not definitive, however the assumptions underlying the formation of the individual boundaries need to be understood and accepted both statistically and clinically.

Serious and thoughtful consideration needs to be given to the establishment of asymmetric monitoring rules. If consideration has not been given to establishing the monitoring boundary for evaluating a negative effect and the data monitoring committee requests guidance for evaluating such an outcome during a pre-planned interim assessment, the impact on the original planning estimates of power and significance will be difficult to quantify and defend. Caution should be exercised in invoking early stopping rules when only a small positive treatment effect is evident during the examination of the interim data. The legitimacy of this concern cannot be overstated, given that many trials are planned based on an overly favourable alternative outcome, rather than the minimum clinically relevant treatment effect. The genesis for this flaw originates during the planning phase of the trial, when the sample size to evaluate the minimum clinical effect is considered operationally or financially too great, leading to a greater hypothesised effect. Although there may be special circumstances that require analysis for monitoring to not be pre-specified *a priori*, it is preferable to determine in advance if the considerations for stopping for a positive or negative effect are asymmetric, and to include an appropriate asymmetric stopping rule in the initial trial design.

DERIVING SYMMETRICAL VERSUS ASYMMETRICAL MONITORING BOUNDARIES

The different operating characteristics of group sequential monitoring rules need to be carefully examined to ensure that

the boundaries accurately align with the goals and design of the clinical trial. The options regarding symmetric monitoring boundaries differ greatly in terms of both the number of events required to trigger an interim analysis and the overall efficiency in terms of the target number of patients. For a more complete understanding of the differences ascribed to early stopping rules, the published work of Haybittle (1), Pocock (2,3), O'Brien and Fleming (4), Wang and Tsatis (5), and Fleming, Harrington, and O'Brien (6) all provide guidance on crafting the monitoring boundaries.

In pivotal research where the event time distribution is attenuated and enrolment is essentially completed prior to a sufficient number of events occurring for the interim analysis, the propriety of the results is essential. Additionally, hierarchical testing of secondary and tertiary endpoints argues for the accumulation of a substantial number of events before a definitive analysis is conducted. For this reason, the monitoring boundaries are traditionally rather conservative in terms of early stopping. The decision of which rule to use should be weighted against several factors that include the likelihood that patients will still be receiving the treatment at the time of the interim assessments if it is likely that the experimental treatment will be used outside the setting of the clinical trial before the results are presented.

There are several commercially available computer packages available to help with the design and monitoring of studies. Alternatively, one may modify a symmetric rule by retaining the upper boundary for early termination due to positive results, but replacing the lower boundary to achieve a more appropriate rule for stopping due to negative results. Prior research suggests that this modification will have little effect on the operating characteristics of the original plan.

Examining the potential negative effect for an experimental therapy at the time of an interim analysis may be more accurately assimilated based on conditional power calculations, predictive power, or Bayesian methods. These methods are practical, have enjoyed a careful mathematical development, and have well-studied operating characteristics. Bayesian monitoring plans based on the use of the sceptical and optimistic prior distributions as described earlier have been used in several cancer clinical trials in the UK. This approach has been found to aid data monitoring committees in focusing on the treatment effect size and the question of whether interim results had effectively established or ruled out the possibility of a clinically relevant benefit (rather than significance levels for a given alternative) for the therapy in question.

EVALUATING FUTILITY BASED ON CONDITIONAL POWER

Predictive modelling to assess futility using a stochastic curtailment approach requires a computation of conditional power, defined as:

$$\gamma_i = \Pr(Z \in R \mid D, H_i),$$

where Z represents a test statistic to be computed at the end of the trial, R is the rejection region of this test, D represents current data, and H_i denotes either the null hypothesis H_0 or an alternative hypothesis H_A . If this conditional probability is relatively large and exceeds a pre-specified threshold under H_0 , the recommendation from the data monitoring committee would most likely be to halt enrolment and immediately reject H_0 in favour of the alternative hypothesis. Alternatively if under H_A the probability is relatively small, the data monitoring committee may recommend stopping the trial given continued enrolment is futile because H_0 will not be rejected under the current fixed sample size. The recommendation to stop based on futility is a declaration that the possibility is highly unlikely that the trial results will reverse from early interim results that favour the control group.

EVALUATING FUTILITY BASED ON PREDICTIVE POWER

The substantive issue with stochastic curtailment is that it requires conditioning on the current data and at the same time an alternative hypothesis that may be unlikely to have given rise to the data. Regardless since H_A must be pre-specified, the method always depends on unknown information at the time of the decision. Alternatively, methods that take an unconditional predictive approach in assessing the consequences of continuing a trial are not saddled with this requirement. Predictive power procedures use weighted averages of conditional power over values of the alternative. A Bayesian framework for adopting this approach is a natural setting for this methodology. If non-informative prior distribution is used for the distribution of the parameter of interest, then the posterior distribution is a weighted average of conditional power with equal prior weight for all alternatives. Alternatively, the current (observed) alternative could be used in the conditional power formulation to project power resulting from further follow-up according to the pattern of observations recorded to date. Prior distributions that assign higher weight to a specific range of alternatives may be used in a fully Bayesian approach to assessing interim analysis results.

INTERIM MONITORING USING A BAYESIAN APPROACH

An important consideration in using a Bayesian approach is that if the monitoring strategy is Bayesian then the final analysis should be conducted within a Bayesian framework. Additionally, there are a number of components that are routinely requested by regulatory agencies prior to a Bayesian design being sanctioned. A list of the required elements is presented below, although not all points will apply to all studies.

- ◆ Detailed accounting of all prior information used to formulate the assumptions.
- ◆ The criteria for success of the study.
- ◆ The derived sample size using a pre-specified level of power and simulations, calculated by finding the

posterior distribution of the primary outcome parameter. This posterior distribution is used in calculating the posterior probability of the study claim for the chosen sample size and true parameter value.

- ◆ Frequentist power tables of the probability of satisfying the study claim given various 'true' parameter values (such as event rates) and various sample sizes for the new trial. These tables give the probabilities of observing data that allow the study claim to be met given the indicated true parameter value. They provide an estimate of the type I error rate in cases where there are true parameter values consistent with a null hypothesis.
- ◆ Evaluation of the predictive probability of the study claim prior to seeing any new data. The predictive probability should not be as high as the simulated posterior probability of the study claim and ideally should be substantially lower so that there is little chance of satisfying the claim before even running the proposed trial. This is recommended in order to ensure that the prior information does not take precedent over the current data (unfavourable results from the proposed study could be masked by favourable prior results). In an evaluation of the prior probability of the claim, how informative the prior is will be balanced against the efficiency gained from using prior information as opposed to using non-informative priors.

DERIVING RELIABLE PRIORS

Bayesian monitoring based on the log of the hazard ratio provides a relative measure of treatment efficacy. The formation of a normal prior distribution is defined by a mean and variance and typically reflects the enthusiasm of the medical community regarding the treatment. In the formation of the prior distribution, sceptical and optimistic views are used to generate two separate distributions that are both considered to be clinically meaningful and relatively probable. From the sceptic, a nominal probability that a positive effect could be achieved should exist. For the optimist, a nominal probability that a negative effect might actually be the outcome should also be characterised.

For a given prior distribution, a posterior density can be computed and the current weight of evidence for benefit can be quantified from the probability values. From the posterior distribution, a predictive distribution can be derived to assess the consequences of continuing the trial for some fixed additional number of failures.

Given the difficulty of accurately and reliably eliciting prior distributions, a commonly agreed upon approach is to generate a prior that represents either no prior belief or a sceptical prior position. The argument in favour of representing no prior information is that this avoids any criticism about subjectivity of the prior. Numerous attempts have been made to derive a formula for representing prior ignorance; however, none have gained any true consensus. Regardless, the various iterations

for a trial can be useful for representing relatively weak prior information. When the new data are strong (relative to the prior information), the prior information is not expected to make any appreciable contribution to the posterior. The rationale behind establishing a sceptical prior is that if a sceptic can be persuaded by the data then anyone with a less sceptical prior position would also be persuaded. Thus, if one begins with a sceptical prior position with regard to some hypothesis and are nevertheless persuaded by the data, so that their posterior probability for that hypothesis is high, then someone else with a less sceptical prior position would end up giving that hypothesis an even higher posterior probability. In that case, the data are strong enough to reach a firm conclusion. If on the other hand, when we use a sceptical prior the data are not strong enough to yield a high posterior probability for that hypothesis, we should not yet claim any definite inference about it.

Specifying prior distributions for monitoring the observed response can be a laborious and tedious task. However if the objective of the trial is to contribute to the knowledge-base ascribed to treating a disease, the results must be considered robust and compelling to the medical community, whose prior opinions and experiences are often quite diverse. Therefore, an interim analysis that incorporates current clinical thinking and models the *a priori* beliefs of the medical community should be interpreted as relevant and considered for adoption in clinical practice. Unfortunately in a number of instances, early termination of a pivotal trial has resulted in the findings having a diminished impact perpetuating concern regarding the clinical utility of a treatment. This realistic and well-documented concern draws into sharp relief the importance of a well designed and implemented interim monitoring strategy to help ensure the overall success of the trial. ◆

*The author can be contacted at
bruce.stouch@premier-research.com*

References

1. Haybittle JL, Repeated assessment of results in clinical trials in cancer treatments, *Br J Radiol*, 44: pp793-797, 1971
2. Pocock SJ, Group sequential methods in the design and analysis of clinical trials, *Biometrika*, 64: pp191-199, 1977
3. Pocock SJ, Interim analyses for randomized clinical trials: the group sequential approach, *Biometrics*, 38: pp153-162, 1982
4. O'Brien, PC and Fleming TR, A multiple testing procedure for clinical trials, *Biometrics*, 35: pp549-556, 1970
5. Wang SK and Tsatis AA, Approximately optimal one-parameter boundaries for group sequential trials, *Biometrics*, 43: pp193-199, 1987
6. Fleming TR, Harrington DP and O'Brien PC, Designs for group sequential tests, *Control Clin Trials*, 5: pp348-361, 1984