Bayesian stopping guidelines for heart valve premarket approval studies

Gary L. Grunkemeier, PhD,[a] YingXing Wu, MD, MS,[a] Lian Wang, MS,[a] and Cody Hamilton, PhD[b]

Objectives: The Data Monitoring Committee (DMC) for the premarket approval (PMA) study of a new heart valve prosthesis convenes periodically to review the accumulating results of the study, and determines, among other things, whether there is enough concern with safety to stop the study. Their deliberations are largely subjective, based on their combined experience and expertise, but an objective aid to evaluating complication rates, usually called a stopping rule, is desirable.

Methods: The US Food and Drug Administration has designated objective performance criteria (OPC) for 7 heart valve complications. At the end of the PMA study, when approximately 800 patient-years have been accumulated, the complication rates must compare favorably with the OPC. Given the results to date at an interim review of the data, we use a Bayesian approach to compute the probability of passing the OPC test by the end of study.

Results: We provide a method that the DMC can use to predict the probability of passing the OPC test for each complication, and a graphical aid for each number of events, observed at 100 patient-year intervals.

Conclusions: Although the DMC ultimately uses combined experience and expertise to make the decision to stop a PMA valve study, we have provided an objective assessment of the probability of the valve ultimately passing the OPC test to aid in making that decision. (J Thorac Cardiovasc Surg 2014;148:2813-7)

Supplemental material is available online.

US Food and Drug Administration (FDA) premarket approval (PMA) studies of new heart valve prostheses are single-armed observational studies in which the observed complication rates are compared with historical values known as objective performance criteria (OPC).[1] The OPC are average complication rates, derived from thousands of patient-years of published results, and range from 0.2 to 3.5 events per 100 patient-years (%/y) for 7 different heart valve complications (Table 1). The sample size requirement of 800 patient-years was derived using a hypothesis test for an OPC of 1.2%/y, with the assumption that the complication rate (hazard function) is constant during the late follow-up period. "The null hypothesis for a complication rate can be rejected at the one-sided significance level of 0.05 if the upper 95% confidence limit for the complication rate is less than 2 times the OPC for that complication."[2] The null hypothesis is that the observed rate is greater than twice the OPC, so to reject it is to pass the OPC test.

INTERIM DATA REVIEWS

In an ongoing PMA study, the data are periodically reviewed by a group, independent from the conduction of the study, called a Data Monitoring Committee (DMC) or Data Safety and Monitoring Board (DSMB). The DMC/DSMB reviews the accumulating evidence with the new valve, and, based on the results, makes a recommendation on whether to continue the study or to stop the study for safety concerns. Much of their deliberation is subjective, using their combined clinical experience and wisdom, but they can also benefit from an objective statistical measure to reinforce their clinical judgment in the event of unusually high complication rates. Such objective measures are usually called stopping rules,[3] but we prefer to refer to them as stopping guidelines, to convey the idea that the collective clinical judgment of the DMC members can overrule the rules. It is worth remembering that if a trial was important enough to be started based on the knowledge available at that time, then caution should be exercised before concluding that there is sufficient evidence to terminate the trial; hence *P* values alone are unlikely to suffice for decision making about the future of a clinical trial, although they may be an important consideration.[4]

Bayesian Persuasion

Stopping rules/guidelines are often formulated using conventional statistical methods, called frequentist to

From the Medical Data Research Center,[a] Providence Health & Services, Portland, Ore; Department of Global Clinical Operations,[b] Edwards Lifesciences, Irvine, Calif.

Disclosures: Authors have nothing to disclose with regard to commercial support.

Received for publication March 29, 2014; accepted for publication June 13, 2014; available ahead of print Aug 16, 2014.

Address for reprints: YingXing Wu, MD, MS, 9205 SW Barnes Rd Suite 33, Portland, OR 97225 (E-mail: yingxing.wu@providence.org).

0022-5223/$36.00

http://dx.doi.org/10.1016/j.jtcvs.2014.06.085

Abbreviations and Acronyms

DMC	= Data Monitoring Committee
DSMB	= Data safety and monitoring board
FDA	= US Food and Drug Administration
NB	= negative binomial
OPC	= objective performance criteria
PMA	= premarket approval
TE	= thromboembolism

distinguish them from an alternative statistical approach called Bayesian. There are many resources available to understand the differences between these statistical approaches.[5,6] Frequentist approaches to stopping rules/guidelines can be found in Jennison and Turnbull's book.[7] One desirable property of a Bayesian analysis is that it allows a true probability statement to be made about the current trial (producing a credible interval), whereas the frequentist can only make a convoluted statement about many similar hypothetical trials that could be undertaken in the future (leading to a confidence interval). The FDA/Center for Devices and Radiological Health has endorsed the use of Bayesian methods for medical device trials,[2] and we use this approach here to derive stopping guidelines, following the method proposed by Lee and Liu.[8]

The Situation

Suppose that during a particular DMC review of the data for a new heart valve, a certain complication (event) has been observed E_1 times during T_1 late (>30 days after surgery) patient-years of observation. The reason for the subscript 1 is that the study is only partially completed and we can anticipate that, in the remaining T_2 years of the study, E_2 additional events will be observed. The duration of the entire study, $T_1 + T_2$, will be about 800 patient-years[9]; so, given T_1 we know what T_2 will be, but we do not yet know E_2. But E_2 will be critical in determining whether the valve passes the OPC test at the conclusion of the study. We do have a sense of what E_2 might be, based on the fact that the complication rate so far is E_1 events occurring in T_1 patient-years. Using the Bayesian approach, we wish to parlay that information into an estimate of the distribution of the possible values of E_2, which in turn will provide an estimate of the probability of passing the OPC test when the trial is complete. If that estimated probability is extremely low, because the observed number of a particular complication so far is very high, the DMC might want to consider stopping the study.

The Mechanics

To illustrate this method, we use thromboembolism (TE) as the complication and assume that the new valve under study is a biological valve that has an OPC for TE of 2.5 events per 100 patient-years (Table 1). We assume that the number of late (30 days after implant) TEs follows a Poisson distribution, which implies that the late TE risk (hazard) is constant over time.

Prior and Posterior Distributions

In a frequentist analysis, the TE risk λ (lambda) is regarded as an unknown fixed number, to be estimated. In the Bayesian approach, λ is considered to be a random variable and, hence, it must be assigned a prior distribution. This distribution will influence the analysis, because the observed data will be combined with it to produce the final answer, contained in the posterior distribution (see Appendix E1). The major criticism of the Bayesian approach is the allegation of subjectivity in the assignation of the prior distribution. For the purposes of constructing a DMC stopping guideline, however, we choose a weak (vague, diffuse) prior distribution, in the sense that it does not influence the result very much (Appendix E1). And we still get the benefit of being able to make direct probability statements about λ, and hence about the results of the study.

EXAMPLES

Example 1: Low Interim TE Occurrence

Suppose a DMC is reviewing data after the first 400 patient-years ($T_1 = 400$) of the study, and 11 TE ($E_1 = 11$) have occurred. At the end of this study, there will be about 400 more patients-years ($T_2 = 400$) and E_2 more events. At that time, the final TE rate will pass the OPC test if the upper one-sided 95% confidence limit is less than $2 \times \text{OPC} = 5\%/\text{y}$. Figure 1, *A*, illustrates that, under these circumstances, the OPC test will be passed if $E_2 \leq 18$.

Posterior Predictive Distribution

Now comes the tricky part. What is the probability that, given the data so far, E_2 will be ≤ 18, so that the TE rate will pass the OPC test at the end of the study? This can be determined using the Bayesian posterior predictive distribution (Appendix E1). Figure 1, *B*, shows that the probability of observing ≤ 18 events is 92%. Thus, the probability is very high that, with $E_1 = 11$ events at $T_1 = 400$ patient-years, the TE rate criterion will be satisfied at the conclusion of the study.

Example 2: High Interim TE Occurrence

Figure 2 displays the same calculations for the situation in which 19 TEs have been observed after 400 patient-years. In this case, in order to pass the OPC test, a maximum of 10 events can be allowed in the remainder of the study (Figure 2, *A*), and the probability of this happening is only 6% (Figure 2, *B*). Thus, the probability is quite low that with $E_1 = 19$ events at $T_1 = 400$ patient-years, the study

TABLE 1. Current OPC values, given as the number of events per 100 patient-years (%/y), for the 2 types of valves

Complication	Biological	Mechanical
Thromboembolism	2.5	3.0
Valve thrombosis	0.2	0.8
Hemorrhage		
All	1.4	3.5
Major	0.9	1.5
Paravalvular leak		
All	1.2	1.2
Major	0.6	0.6
Endocarditis	1.2	1.2

will fulfill the TE criterion. The DMC may wish to take this into consideration in making their recommendation about continuation of the study.

Complete TE Example: Other Values of E_1 and T_1

Repeating the above computations at $T_1 = 400$ for all values of E_1 between 11 and 19 produces the series of blue circles in Figure 3. The examples in Figures 1 and 2 are indicated by their corresponding digits (1 and 2) superimposed onto the highest and lowest of the circles for 400 patient-years. Figure 3 also contains the values for other intermediate assessment times, from $T_1 = 200$ to 600 patient-years, thus providing a visual tool for associating eventual OPC success probabilities based on intermediate results.

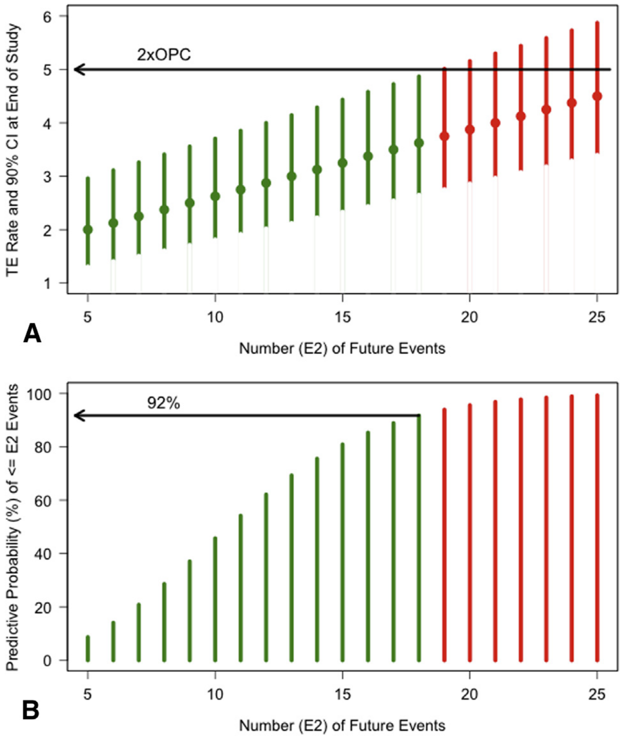

FIGURE 1. A, Point estimates and 90% two-sided confidence intervals for the TE rate at the end of the study, for various values of E_2 occurring after the current data review, where $E_1 = 11$ events have been observed in the first 400 (T_1) patient-years. The upper confidence limit is the same as a one-sided 95% limit, so cases with an upper limit less than 2 × OPC (= 5%) will pass the OPC test. B, Cumulative distribution of the predictive probabilities for the number E_2 of future TEs. *TE*, Thromboembolism; *CI*, confidence interval; *OPC*, objective performance criteria.

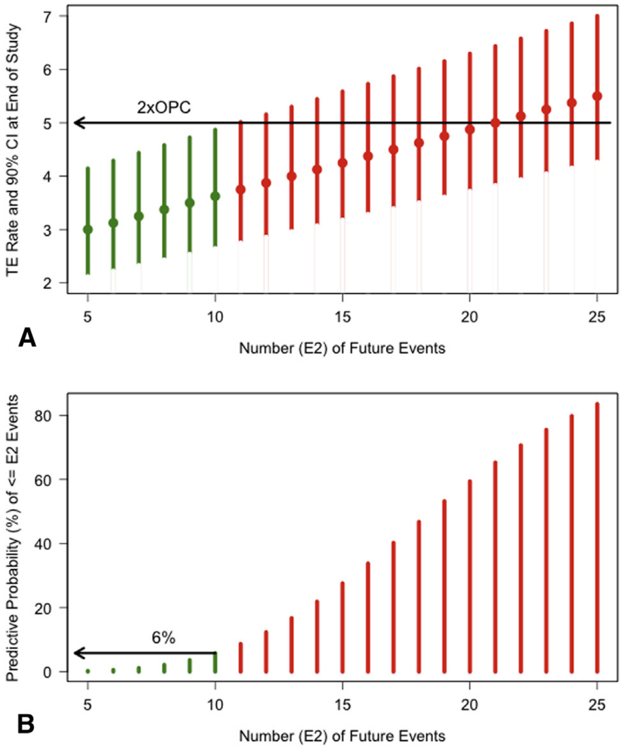

FIGURE 2. A, Point estimates and 90% two-sided confidence intervals for the TE rate at the end of the study, for various values of E_2 occurring after the current data review, where 19 events have been observed in the first 400 patient-years. The upper confidence limit is the same as a one-sided 95% limit, so cases with an upper limit less than 2 × OPC (= 5%) will pass the OPC test. B, Cumulative distribution of the predictive probabilities for the number E_2 of future TEs. *TE*, Thromboembolism; *CI*, confidence interval; *OPC*, objective performance criteria.

Complete Example for Other Complications

In practice, a visual stopping guideline tool such as Figure 3 would be produced for all the OPC values in Table 1. For example, Figure 4 contains the corresponding plot for an OPC of 1.2% per patient-year, which corresponds to the events of paravalvular leak and endocarditis for biological valves. Alternatively exact probabilities for any OPC could be calculated for any values of T_1, as described in Appendix E1.

COMMENT

We have produced a DMC stopping guideline tool for biological valve TE (Figure 3) and paravalvular

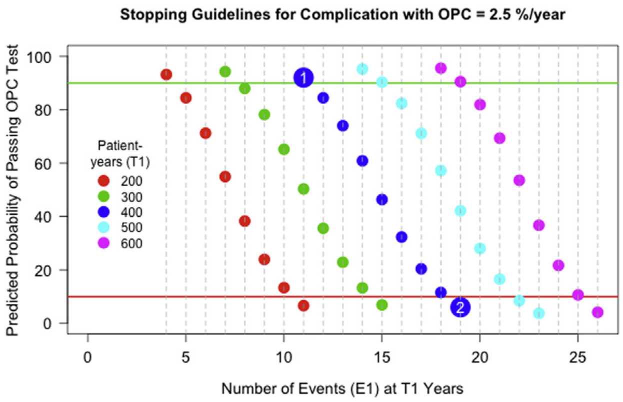

FIGURE 3. Predicted probabilities of passing the OPC test at the end of the trial for a complication with an OPC of 2.5%/y (TE for biological valves), computed for various numbers of events (E_1) observed at various intermediate data review times (T_1). The 2 examples depicted in Figures 1 and 2, after 400 patient-years were observed, are shown by the slightly larger circles at the top and bottom of the 400 patient-years grouping, with the corresponding digits (1, 2) superimposed. Symbols below the red horizontal gridline at 10% indicate performance that may influence the Data Monitoring Committee members to consider stopping the trial. These probabilities were derived using a noninformative (Jeffreys) prior distribution. *OPC*, Objective performance criteria.

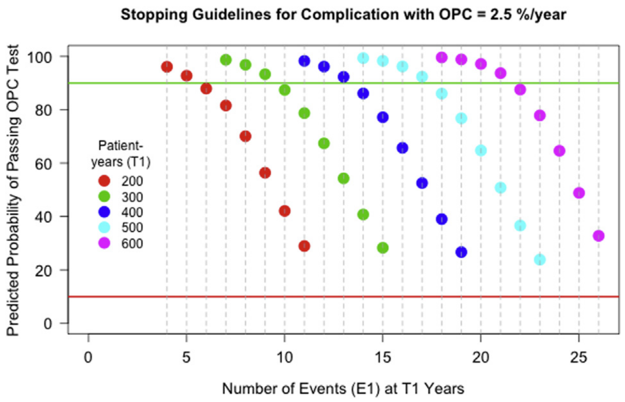

FIGURE 5. A reproduction of Figure 3, but with a mildly informative prior distribution, based on the OPC value. *OPC*, Objective performance criteria.

leak/endocarditis (Figure 4) using a noninformative Bayesian prior distribution, which allows the posterior distribution to be predominantly determined by the data. This seems appropriate because the DMC should be neutral, and avoids other assumptions that might favor the new valve under investigation. But others may choose to use other prior distributions. The prior distribution is subjective, that is the essence of the Bayesian approach, and it should be informed by whatever the investigator knows or believes, based on evidence or experience.

For example, a valve manufacturer may possess clinical data from a previous generation of the investigational device.[4,10] Or, an informative prior could incorporate known information about biological valve complication rates: the current FDA OPC complication rate would seem to be a reasonable mean value for the prior distribution, and the variance could be made very large, to incorporate uncertainty, so that the study valve data contribute a preponderance of the influence on the posterior distribution. For example, the OPC for biological valve TE is 2.5 events per 100 patient-years. In this case, a logical choice for the TE rate prior would be a gamma (α, β) distribution with $\alpha = 2.5$ and $\beta = 1$. This will result in a prior mean of 2.5 with 95% of the probability density between 0.4 and 6.4, which seems reasonable, and makes some use of the vast amount of information already known about TE rates with biological valves.

With this (slightly) more informative prior, the predicted probabilities of passing the OPC test are greatly increased for all values of E_1 at each value of T_1 (Figure 5). But the

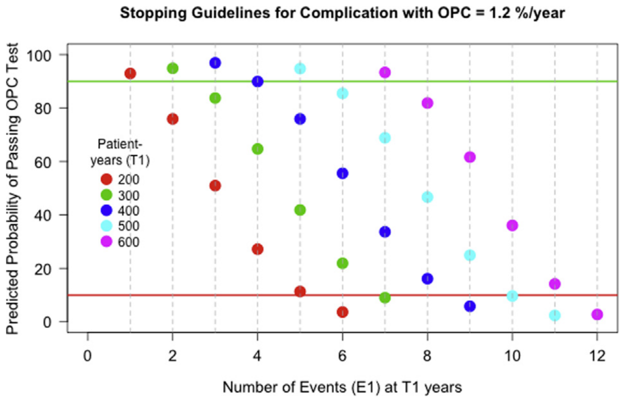

FIGURE 4. Predicted probability of success at the end of the trial for a complication with an OPC of 1.2%/y (endocarditis and leak for biological valves), computed for various numbers of events observed at various intermediate data review times. Symbols below the red horizontal gridline at 10% indicate performance that may influence the Data Monitoring Committee members to consider stopping the trial. These probabilities were derived using a noninformative (Jeffreys) prior distribution. *OPC*, Objective performance criteria.

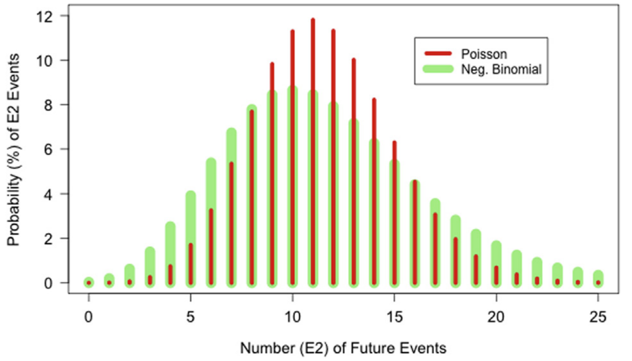

FIGURE 6. Comparison of 2 different estimates of the probabilities of the numbers of future events: (1) the naive Poisson distribution, based on a fixed value of λ; (2) the negative binomial distribution, which takes into account the imprecision in the estimate of λ, and thus has more variability.

DMC should probably be more skeptical; it is worse to approve an inferior valve than to reject an acceptable valve, because there are already so many acceptable valves available. Thus, we believe that the objective decision aids using the noninformative prior (Figures 3 and 4) are more appropriate (more impartial, neutral) for the DMC to use.

False-Negative Rates

To recap, we have described a method of computing, for any number of events E_1 observed after T_1 patient-years, the (Bayesian) probability of passing the OPC test at the end of the study (800 patient-years). We plotted these probabilities for 5 values of T_1 and several values of E_1 in Figures 3 and 4, and suggested in the figure legend that symbols below the red horizontal gridline at 10% indicate performance that may influence the DMC members to consider stopping the trial.

An observer might wonder: Assuming that the true TE rate for a good-performing aortic biological valve is 2.5%/y (the situation illustrated in Figure 3), what is the probability that this valve would ever fall below the red line (ie, be considered a failure) in Figure 3? Assuming that the DMC performs exactly 5 interim looks at the values of T_1 plotted in Figure 3, the probability of wrongly rejecting a valve based on this event can be computed using elementary probability. This probability turns out to be 2.6%. For lower-risk events, such as the 1.2%/y considered in Figure 4, this probability is 13.4%. However, in practice, these probabilities would depend on when and how often the data reviews are actually done by the DMC as well as the assumed true rate of events (any possible rate up to the highest that could pass the OPC test). More fundamentally, this probability relies on the selection of the cutoff for the predicted passing probability (arbitrarily set at 10% in Figures 3 and 4); in fact, it might be better to have different cutoffs for the different interim looks, with more liberal cutoffs for earlier looks, and more stringent cutoffs for later looks, when more patient-years are available. Even more fundamental is the distribution taken as the Bayesian prior (Figure 5 vs Figure 4). For all these reasons, we stress that these are stopping guidelines rather than rules.

References

1. Johnson DM, Sapirstein W. FDA's requirements for in-vivo performance data for prosthetic heart valves. *J Heart Valve Dis*. 1994;3:350-5.
2. Draft Guidance for Industry and FDA Staff: Heart Valves—Investigational Device Exemption (IDE) and Premarket Approval (PMA) Applications. January 20, 2010. Available at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM198043.pdf. Accessed August 10, 2014.
3. Stallard N, Whitehead J, Todd S, Whitehead A. Stopping rules for phase II studies. *Br J Clin Pharmacol*. 2001;51:523-9.
4. Fayers PM, Ashby D, Mahesh KB, Parmar MK. Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Stat Med*. 1997;16:1413-30.
5. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov*. 2006;5:27-36.
6. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Sci*. 1988;76:159-65.
7. Jennison C, Turnbull BW. *Group Sequential Methods with Applications in Clinical Trials*. Boca Raton, Fla: Chapman and Hall/CRC; 2000:205-20.
8. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin Trials*. 2008;5:93-106.
9. Grunkemeier GL, Johnson DM, Naftel DC. Sample size requirements for evaluating heart valves with constant risk events. *J Heart Valve Dis*. 1994; 3:53-8.
10. Spiegelhalter DJ. Incorporating Bayesian ideas into health-care evaluation. *Stat Sci*. 2004;19:156-74.
11. Cox DR. Some simple approximate test for Poisson variates. *Biometrika*. 1953; 40:354-60.
12. Grunkemeier GL, Anderson WN Jr. Clinical evaluation and analysis of heart valve substitutes. *J Heart Valve Dis*. 1998;7:163-9.

APPENDIX E1. TECHNICAL DETAILS

Prior and Posterior Distributions

As mentioned in the text, to enjoy the Bayesian benefits, we must assume a prior distribution for the parameter of interest, in our case the complication rate. We assume a gamma (α, β) distribution as the prior distribution for the Poisson rate parameter λ. The gamma distribution is very flexible, fitting a wide range of shapes. It is also conjugate to the Poisson distribution, meaning that the incorporation of the accumulated data (E_1 events in T_1 patient-years), via the Bayes theorem, yields a new, more accurate posterior distribution for λ, which is also a gamma distribution, namely, gamma ($\alpha + E_1$, $\beta + T_1$). A commonly used weak (noninformative) Bayesian prior for the Poisson distribution is the Jeffreys prior, which in this case is a gamma distribution with $\alpha = 0.5$ and $\beta = 0$, so that the posterior distribution for λ is gamma($0.5 + E_1$,T_1).

The credible interval for λ using the Jeffreys prior gamma distribution turns out to be identical to the confidence interval for λ proposed by Cox.[11] Several formulas have been proposed for Poisson confidence intervals. A simulation study, using data typical of heart valve PMA studies, found that the formula recommended by Cox had the best coverage properties among 7 alternative methods, and was advocated for testing the relationship of PMA results to the OPC.[12] The above gamma distribution, using the Jeffreys prior, is identical to $\chi^2_{2E+1}/2T$, where χ^2_{2E+1} is a χ^2 distribution with 2E + 1° of freedom. This is the formula proposed by Cox,[11] and was used to produce the confidence intervals displayed in Figures 1, *A*, and 2, *A*.

Posterior Predictive Distribution

The critical next step, and the beauty that this Bayesian approach brings, is the ability to determine the true probability of observing the number E_2 of events by the end of the study. This is accomplished using the posterior predictive distribution.[6] A naive approach to estimating the probabilities of the possible future values of E_2 would be to simply use the current point estimate of λ at the interim review (E_1/T_1) to predict the number of future events based on the Poisson distribution. That approach would be OK if we knew for certain that the current estimate of λ is exactly correct. But, in fact, our knowledge of λ consists only of a posterior (gamma) distribution of values. To acknowledge this variation, the Bayesian approach is to take a weighted average of the estimates of E_2 values over the range of probable values of λ, weighted by the probability assigned to various values of λ.[8]

For the case in hand, this is easy, because it turns out that, after observing E_1 events in the first T_1 patient-years, the number of events, E_2, in the remaining $800-T_1$ patient-years has a negative binomial (NB) distribution with parameters $\alpha + E_1$ and $(\beta + T_1)/(\beta + 800)$, where α and β are, as before, 0.5 and 0, respectively. This is the distribution used to compute the cumulative probabilities of $\leq E_2$ future events in Figures 1, *B*, and 2, *B*.

Thus, the posterior predictive distribution incorporates (1) the uncertainty in the estimation of λ, and (2) the statistical variation of the number of events given each potential value of λ. The naive method, using the Poisson distribution, ignores (1). Figure 6 shows the difference between these 2 distribution for the case of Figure 1, *B* (where $T_1 = 400$, $E_1 = 11$ and $T_2 = 400$); the 2 distributions have the same mean value (11.5), but the predictive NB distribution has more variance (less precision) than the Poisson distribution. The Poisson distribution has its mean equal to its variance, but the NB distribution has 1 more parameter that allows the variance to be larger than the mean. Hence, the NB distribution is often used to model count data when there is extra-Poisson variation.

STATISTICAL SOFTWARE AND CODE

Statistical analysis was performed using R 3.0 (http://www.R-project.org). We used the R function pnbinom to calculate the predicted probability (pp) of passing the OPC tests in Examples 1 and 2.

```
pp<-pnbinom(E2, 0.5 + E1, T1/800)
```

Example 1:

E1 = 11, E2 = 18, T1 = 400; then pp = .917 ~ 92%

Example 2:

E1 = 19, E2 = 10, T1 = 400; then pp = .058 ~ 6%