Published in final edited form as:

Clin Trials. 2013 February ; 10(1): 5–18. doi:10.1177/1740774512464725.

Bayesian Adaptive Design for Device Surveillance

Thomas A. Murray¹, Bradley P. Carlin², and Theodore C. Lystig³

Thomas A. Murray: murra484@umn.edu; Bradley P. Carlin: carli002@umn.edu; Theodore C. Lystig: theodore.lystig@medtronic.com

¹Graduate Student, Department of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, MN 55455, (715) 829-0511

²Professor and Head, Division of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, MN 55455, (612) 624-6646

³Lead Corporate Biostatistician, Medtronic Inc., 710 Medtronic Parkway, Minneapolis, MN 55432, (763) 514-0472

Abstract

Background—Post-market device surveillance studies often have important primary objectives tied to estimating a survival function at some future time T with a certain amount of precision.

Purpose—This paper presents the details and various operating characteristics of a Bayesian adaptive design for device surveillance, as well as a method for estimating a sample size vector (determined by the maximum sample size and a pre-set number of interim looks) that will deliver the desired power.

Methods—We adopt a Bayesian adaptive framework which recognizes the fact that persons enrolled in a study report their results over time, not all at once. At each interim look we assess whether we expect to achieve our goals with only the current group, or whether the achievement of such goals is extremely unlikely even for the maximum sample size.

Results—Our Bayesian adaptive design can outperform two non-adaptive frequentist methods currently recommended by FDA guidance documents in many settings.

Limitations—Our method's performance can be sensitive to model misspecification and changes in the trial's enrollment rate.

Conclusions—The proposed design provides a more efficient framework for conducting postmarket surveillance of medical devices.

Keywords

Adaptive trial; Bayesian statistics; futility analysis; interim analysis; Monte Carlo sampling

1 Introduction

Current Bayesian adaptive trial design methods have largely focused on the comparison of two or more samples, typically in the realm of hypothesis testing where the research question generally concerns whether group A differs from group B. By contrast, the one sample problem is largely an estimation problem where the research question concerns the

value of a population parameter. At the outset of a one sample problem, a researcher sets a goal of estimating a parameter with a certain precision, sometimes with an associated acceptable threshold for efficacy. For example, we may wish to measure the percentile of medical device survival at 5 years with a 95% confidence interval having half-width no larger than .03. If the survival rate is significantly below 95%, the device may need to be recalled. Here the precision is the half-width of the interval which improves (shrinks) as the sample size increases, and the efficacy threshold is 95% device survival at 5 years.

The reason this is relevant in surveillance settings stems from the fact that the device should already have been established as having reasonable assurance of safety and efficacy prior to receiving market approval. After receiving approval, the relevant scientific question for the sponsor is often one of estimation (“What is the percentile of freedom from stroke at five years post-implant for device A?”), as opposed to one of testing (“Is the percentile of freedom from stroke at five years post-implant the same for device A and device B?”).

In such scenarios, studies may have important primary objectives tied to estimating quantities with a certain amount of precision, even if they do not necessarily have a hypothesis that is intended to be tested for the primary objective. Sample sizes in such studies should be statistically justified, even if they are not obtained through the traditional routes of power calculations performed under alternative hypotheses. The estimation approach is in agreement with 21 CFR 822 which states that a surveillance plan must include a discussion of the plan objective addressing the surveillance question(s) identified in the order; no mention of a hypothesis test approach is explicitly made in 21 CFR 822. In contrast, the current draft guidance document from the FDA covering postmarket surveillance [4] calls for both study objectives and hypotheses to be included in a postmarket surveillance study plan using a standard (non-adaptive) frequentist sample size calculation. This paper will elucidate the manner in which statistical sample size computation may be performed in the situation where a primary objective exists, but a hypothesis does not.

The frequentist approach can provide an estimated sample size to achieve the precision-based goal mentioned earlier under certain other constraints/assumptions (i.e., drop-out rate, expected device survival percentile at the time of interest, etc.). One frequentist variant starts with the estimation of the Kaplan-Meier survival curve, applies the alternative expression for the standard error of the survival estimate by Peto et al. [11] (useful when the estimate of the survival function is close to zero or one), and performs an ad hoc scaling incorporating a censoring rate out to a given estimation time. The resulting expression was described in a guidance document from the FDA [6]. However, if the survival curve is assumed to arise from a Weibull distribution, then the methods described in Meeker and Nelson [10] can be applied.

This general problem of estimating a single survival curve with a certain amount of precision lends itself well to a Bayesian approach. Even under the frequentist approach to sample size estimation we require a prior guess as to the device survival percentile at the time of interest. Instead of treating this guess as fixed, we can place a prior distribution on this parameter and then attack the problem as a Bayesian. For a general overview of

Bayesian trial design see Berry [1] as well as the most recent FDA guidance document on Bayesian methods [5]. Work summarized in Berry et al. [2] has approached similar problems of sample size in the two sample setting by using interim looks and predictive probabilities of success. That is, one assesses the probability of achieving the given characteristics with the current sample size given the information already obtained from these observations.

Here we too adopt a Bayesian interim look framework which recognizes the fact that persons enrolled in a study report their results over time, not all at once. Our problem consists of predicting whether we will have enough information from the presently enrolled sample to achieve the desired characteristics after follow-up on the current group, or if given the current information, it seems futile that we will ever achieve the desired characteristics even if we continue enrollment to the prespecified maximum sample size. Simply put, at each interim look we assess whether we expect to achieve our goals with only the current group, or whether the achievement of such goals is extremely unlikely even for the maximum sample size. If either of these two situations is present, we stop enrollment at the current sample size. If not, we continue enrollment until the next interim look or the maximum sample size, whichever is dictated by the design. Once we've halted enrollment, we will continue to follow the enrolled cohort to the end of the study (e.g., 5 years minimum follow-up), at which point we evaluate the results.

Because this is a Bayesian adaptive design, each decision is based on a posterior distribution. In this design, we are making decisions at several points in time. We are deciding whether to halt enrollment at each interim sample size using a predictive calculation driven by the *interim posterior* and we are making a final decision about the result of the trial based on the posterior distribution at the very end. The interim looks are set to take place when enrollment hits a prespecified number of persons. Thus, this design can be summarized as a sample size *vector* whose components detail the interim look sample sizes and the maximum sample size. The intent of this paper is to present the general details of this design and its operating characteristics in some settings, as well as a method for estimating a sample size vector that delivers a trial with the desired power.

2 Methods

Before presenting our Bayesian adaptive design we briefly review two standard frequentist designs. We continue in the device surveillance framework, where our aim is to estimate an interval for the percentile of device survival (L) at a time point of interest (T). A trial will be defined as a failure if either the precision, defined as the half-width of the interval and denoted by δ , is larger than ε , or the upper limit of the estimated interval, denoted by U_{CI} , is less than the efficacy threshold (L_{ET}). Otherwise, the trial is defined as a success. Thus, given $\delta < \varepsilon$, a trial is considered successful when at least part of the interval lies above L_{ET} and a trial is considered a failure only when $U_{CI} < L_{ET}$. A trial is always considered a failure when $\delta \geq \varepsilon$. The details of the two frequentist designs, nonparametric Kaplan-Meier and parametric Weibull, and the Bayesian adaptive design are presented in the first two subsections below. Then, Subsection 2.3 describes methods to estimate a sample size vector for the Bayesian adaptive design that will deliver any prespecified power.

2.1 Frequentist Models

We compare our Bayesian results to those of two frequentist models: nonparametric Kaplan-Meier estimation and parametric Weibull estimation, both used to obtain a $100(1 - \alpha)\%$ confidence interval for L at a time T . Beginning with the Kaplan-Meier model, the log-transformed point-wise confidence interval is used to decide the result of the trial. Details of the implementation and calculation of a confidence interval for the survival percentile at a particular time for the Weibull model are described in Klein and Moeschberger [9]. Throughout this paper, for computational convenience, the Weibull scale parameter (β) is fixed—often with a value of 1, implying device survival is exponential.

Under the nonparametric Kaplan-Meier design, the researcher calculates the sample size by plugging in hypothesized parameter values and desired characteristics to an equation (shown in the Appendix) described by an FDA guidance document [6]. Alternatively, a parametric Weibull approach discussed by Meeker and Nelson [10] can be used as a guidance, though a direct sample size calculation does not exist in this setting. Meeker and Nelson describe techniques for sample size calculation when the researcher desires an estimate of survival times at a particular percentile. However, here we are concerned with the inverse problem of estimating the survival percentile at a particular time; their methods can be adapted to this setting to a certain extent. The incongruence of the two settings arises in the desired precision: for Meeker and Nelson, the researcher specifies the acceptable degree of error for the desired percentile (e.g., if the 95th percentile is expected to be 5 years, we may willingly accept estimates between 4–6 years, or within 20%). In our setting, we are estimating the survival percentile (L) at a specific time (T) and we are willing to accept $\delta < \varepsilon$.

This inverse problem doesn't translate directly from the methods of Meeker and Nelson. Based on simulations (not shown here), the nonparametric sample size calculation appears to offer a conservative estimate for the parametric setting. When the parametric assumption is approximately correct, the estimated sample size will provide a confidence interval with a smaller half-width relative to the nonparametric interval. Additionally, these existing frequentist methods were developed in the absence of an efficacy threshold, so no previous sample size estimation methods appear to exist for exactly the setting of our concern.

2.2 Bayesian Adaptive Model Details

In the Bayesian setting we no longer estimate confidence intervals, but rather posterior credible intervals. Thus, we determine our Bayesian trial result using the half-width and upper limit of the posterior credible interval. Recall, we also have interim stopping rules which are assessed when enrollment has reached the interim sample size specified by the investigator: we stop enrollment for expected success when the predictive probability of success for the current sample is above a large cut-off value (C_S), and stop enrollment for futility if the predictive probability of success for the maximum sample size (M) is below a small cut-off value (C_F).

During each interim look we must evaluate the predictive distribution arising from the interim posterior, whereas for the final analysis we evaluate the posterior distribution. For flexibility, we assume every device has a survival time that is Weibull with a fixed shape

parameter (β) and each enrollee has a drop-out/censoring time that is exponential with known rate λ . The Weibull parametrization in $\mathbb{R}$ [12] has density

$f(x) = \beta \left(\frac{x}{\eta}\right)^{\beta} \exp\left(-\left(\frac{x}{\eta}\right)^{\beta}\right)$. Zhang and Meeker [15] show that the conjugate prior for a transformed scale parameter ($\theta = \eta^{\beta}$) of the Weibull distribution with a fixed β is the inverse gamma distribution, which we notate $\theta \sim IG(a, b)$, and which has density function

$$\pi(\theta|a, b) = \frac{b^a}{\Gamma(a)} \theta^{-a-1} \exp\left(-\frac{b}{\theta}\right).$$

Let $x_1, \dots, x_M$ be the set of observation times, R denote the set of observations with an observed device failure, and C the set of right-censored observations. If we assume

$X_i|\theta \stackrel{iid}{\sim} Weibull(\beta, \theta)$, the likelihood given $\mathbf{x} = (x_1, \dots, x_m)$ is given by

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{x_i \in R} f(x_i|\beta, \theta) \prod_{x_i \in C} (1 - F(x_i|\beta, \theta)) \\ &\propto \prod_{x_i \in R} \theta^{-1} \exp\left(-\frac{x_i^{\beta}}{\theta}\right) \prod_{x_i \in C} \exp\left(-\frac{x_i^{\beta}}{\theta}\right) = \theta^{-r} \exp\left(-\frac{\sum_{i=1}^M x_i^{\beta}}{\theta}\right) \end{aligned}$$

where r is the number of device failures. The conditional posterior for θ emerges as

$$\pi(\theta|\mathbf{x}, a, b, \beta) \propto \theta^{-a-r-1} \exp\left(-\frac{b + \sum_{i=1}^M x_i^{\beta}}{\theta}\right) \equiv IG\left(a+r, b + \sum_{i=1}^M x_i^{\beta}\right).$$

The last design element we must specify is the prior distribution on θ ; however, we are

primarily concerned with the quantity $L = S(T)$. Since $L = \exp\left(-\left(\frac{T}{\eta}\right)^{\beta}\right) = \exp\left(-\frac{T^{\beta}}{\theta}\right)$, or

equivalently, $\theta = g(L) = \frac{-T^{\beta}}{\log(L)}$, placing a prior distribution on either parameter implicitly defines a prior on the other. For instance, if we chose $L \sim p_L(L)$, because $g(L)$ is monotonically increasing on $L \in (0, 1)$, the derivation of $p_{\theta}(\theta)$ is a straightforward

transformation of random variables calculation. Namely, $p_{\theta}(\theta) = p_L(g^{-1}(\theta)) \left| \frac{dg^{-1}(\theta)}{d\theta} \right|$

where $\frac{dg^{-1}(\theta)}{d\theta} = \exp\left(\frac{-T^{\beta}}{\theta}\right) \frac{T^3}{\theta^2}$. Taking $L \sim Beta(\gamma, \xi)$, we have that

$$\begin{aligned} p_{\theta}(\theta) &\propto \theta^{-2} \exp\left(\frac{-T^{\beta}}{\theta}\right) \exp\left(\frac{-(\gamma-1)T^{\beta}}{\theta}\right) \left(1 - \exp\left(\frac{-T^{\beta}}{\theta}\right)\right)^{\xi-1} \\ &\propto \theta^{-1-1} \exp\left(\frac{-\gamma T^{\beta}}{\theta}\right) \left(1 - \exp\left(\frac{-T^{\beta}}{\theta}\right)\right)^{\xi-1}. \end{aligned}$$

We would like a noninformative prior distribution, which for a proportion often manifests as $U(0, 1) = Beta(1, 1)$. Plugging in these noninformative hyperparameter values, the term

raised to the $(\xi - 1)$ power cancels and we find that $p_{\mathcal{A}}(\theta) \equiv IG(1, T^{\beta})$. Thus, the conditional posterior for θ emerges as $IG\left(1+r, T^{\beta} + \sum_{i=1}^M x_i^{\beta}\right)$ with a prior mean of corresponding to 50% device survival at time T and a 95% prior confidence of (2.5%, 97.5%) survival at time T .

To make use of all the information available when making decisions about enrollment, we update the interim posterior when the next group is ready to enroll. So when we check the trial at the first interim look of size m_1 enrollees, we will have accumulated some amount of follow-up in the most recent group enrolled, some amount more for the previously enrolled group, and so on back to the greatest accumulation of follow-up for the first group enrolled. For an interim look of size m_k enrollees ($k = 1, \dots, K$ where K is the potential number of interim looks), we can use this partial information to get the number of events already unmasked (r_k), and the sum of the adjusted follow-up times $\left(\sum_{i=1}^{m_k} x_{k_i}^{\beta}\right)$. Then the interim posterior at interim look k under our noninformative prior is

$$P(\theta|\mathbf{x}_k) = IG\left(1+r_k, T^{\beta} + \sum_{i=1}^{m_k} x_{k_i}^{\beta}\right) \quad (1)$$

Working the other direction, since $(\theta|\mathbf{x}) \sim IG(\kappa, \zeta)$, by taking $\tau = -\log(L)$ so that

$\tau = h(\theta) = \frac{T^{\beta}}{\theta}$, we see $p_{\tau}(\tau) \propto \tau^{\kappa-1} \exp\left(-\frac{\zeta}{T^{\beta}}\tau\right) \equiv \text{Gamma}\left(\kappa, \frac{\zeta}{T^{\beta}}\right)$. Plugging in the

relevant parameter values reveals $(\tau|\mathbf{x}) \sim \text{Gamma}\left(r+1, 1 + \sum_{i=1}^m \left(\frac{x_i}{T}\right)^{\beta}\right)$. Since $L = \exp(-\tau)$, the distribution of $(L|\mathbf{x})$ is not common, but an application of the delta method provides an approximation of $V[L|\mathbf{x}]$, and $E[L|\mathbf{x}]$ can be calculated directly. Having the first two posterior moments will provide insight into the distribution of $(L|\mathbf{x})$. Now,

$$\begin{aligned} E[L|\mathbf{x}] &= \zeta^{\kappa} \int_0^{\infty} \Gamma(\tau)^{-1} \exp(-\tau) \tau^{\kappa-1} \exp(-\zeta\tau) \\ &= \frac{\zeta^{\kappa}}{(\zeta+1)^{\kappa}} \int_0^{\infty} \frac{(\zeta+1)^{\kappa}}{\Gamma(\tau)} \tau^{\kappa-1} \exp(-(\zeta+1)\tau) \\ &= \left(\frac{\zeta}{\zeta+1}\right)^{\kappa} = \left(\frac{1 + \sum \left(\frac{x_i}{T}\right)^{\beta}}{2 + \sum \left(\frac{x_i}{T}\right)^{\beta}}\right)^{r+1} \end{aligned}$$

Applying the delta method,

$$\begin{aligned} V[L|\mathbf{x}] &\approx g'(E[\tau|\mathbf{x}])^2 V[\tau|\mathbf{x}] = \exp\left(-\frac{\kappa}{\zeta}\right)^2 \frac{\kappa}{\zeta^2} \\ &= \exp\left(\frac{-2(r+1)}{1 + \sum \left(\frac{x_i}{T}\right)^{\beta}}\right) \frac{r+1}{\left(1 + \sum \left(\frac{x_i}{T}\right)^{\beta}\right)^2} \end{aligned}$$

Note that $V[L|\mathbf{x}]$ depends upon the fixed shape parameter β as well as the total transformed

follow up time $\left(\sum \left(\frac{x_i}{T}\right)^\beta\right)$ and number of failures (r). Holding all else constant, $V[L|\mathbf{x}]$ increases with β when follow-up times are predominately greater than T (i.e., $S(T) > .5$), otherwise the relationship flips, and $V[L|\mathbf{x}]$ decreases as follow-up increases. The relationship between the number of failures and $V[L|\mathbf{x}]$ is more complex, although, in our setting $V[L|\mathbf{x}]$ increases as the number of failures observed increases.

The complexity of this problem lies primarily in the computation of the expected success and futility at each interim look, via Bayesian predictive probability distributions. At each interim look, we have partial information about the current group of enrollees, and no information from potential future enrollees. We use the partial information at hand to update the prior to the interim posterior, and then calculate two predictive distributions. Rather than calculating the exact probabilities of expected success and futility from their corresponding predictive distributions, we estimate them using Monte Carlo sampling. First, we draw a large sample $(\theta_k^{(1)}, \dots, \theta_k^{(g)}, \dots, \theta_k^{(G)})$ from the interim posterior described in equation (1). With this sample on hand, we can simulate the remaining follow-up for the currently enrolled individuals and potential *future* enrollees.

For current enrollees, we only need to simulate remaining follow-up for those that have not yet experienced device failure nor dropped out (i.e., those remaining in the risk set at the time of the interim look) via composition from a left truncated Weibull distribution with the sampled scale parameter $\theta_k^{(g)}$ and left truncation time corresponding to each individual's accumulated follow-up at the time of interim look k . For potential future enrollees, we simulate survival times from a Weibull distribution with sampled scale parameter $\theta_k^{(g)}$. or each sample g , we also simulate dropout/censoring times from an exponential distribution with known rate λ . Lastly, if we assume a constant enrollment rate (e.g., 10 persons each 3-month period), and carry out each interim look when the next group/person is ready to enroll, the follow-up accumulated for each enrollee is a straightforward calculation.

Expected success at interim look k ($\hat{E}_{S_k}$) is calculated with respect to the currently enrolled cohort. Assuming no more persons will be enrolled, the study will end when the most recent enrollee has accumulated follow-up T . So for each simulated trial, every current enrollee's status and follow-up can be determined at study end. Armed with G trials containing simulated full information, we update each predictive posterior distribution, and record each result. We take $(\hat{E}_{S_k})$ to be the proportion of predicted trials for currently enrolled cohorts that had a successful outcome. Now, futility at interim look k ($1 - \hat{E}_{F_k}$) is calculated relative to the prespecified maximum sample size M . Assuming enrollment continues to the maximum sample size M , the study will end when the last potential future enrollee has accumulated T follow-up. So for each simulated trial of both current and potential future enrollees, the simulated status and follow-up can be determined at study end for each enrollee. We take $\hat{E}_{F_k}$ as the proportion of predicted trials for maximum-sized cohorts that meet the final criterion for success, so $1 - \hat{E}_{F_k}$ is the proportion of predicted trials that fail.

The expected success and futility calculations outlined above provide us with a one-off method of calculating $\hat{E}_S$ and $\hat{E}_F$ at each interim look, and thus making a sensible decision about halting enrollment. The calculated values $\hat{E}_S$ and $\hat{E}_F$ dictate whether to halt enrollment, or continue to the next interim look. If $\hat{E}_S$ is larger than C_S , we halt enrollment because given the current information it is very likely that we will achieve the desired characteristics with the currently enrolled cohort. If instead $\hat{E}_F$ is below C_F , the trial is defined as futile, because given the current information it is very unlikely that we will achieve the desired characteristics when enrolling M persons. Once enrollment has finished, we follow the cohort to study end, update the posterior, and record the final trial result. When an actual trial operating under this design has enrolled m_k persons and the next group is ready to be enrolled, the calculation of expected success and futility and subsequently the enrollment decision can be carried out using Algorithm 1.

To implement this design, there are a number of parameters that need to be specified. The investigator needs to define a maximally acceptable half-width (ε), credible interval coverage level ($100(1 - \alpha)\%$), efficacy threshold (L_{ET}), time point of interest (T), drop-out/censoring rate (λ), enrollment rate information, cut-offs for expected success (C_S) and futility ($1 - C_F$), and in our current Weibull model, the shape parameter β . The enrollment information is broken down into the size of each enrollment group (n) and the time lag between groups (e.g., 10 persons are enrolled each quarter year). The investigator will also need to decide the number of Monte Carlo iterations (G) used to calculate the expected success ($\hat{E}_S$) and futility ($1 - \hat{E}_F$) at each interim look. Lastly, the investigator needs a sample size vector, determined by the maximum sample size (M) and a predetermined number of interim looks (K). The sample size vector is denoted by $(m_1, \dots, m_k, \dots, m_K, M)$, where m_k is the cumulative sample size at the k^{th} interim look.

To assess the operating characteristics of a particular design and situation, the investigator will further need to specify the true device survival percentile L at time T (or equivalently, $\eta = T/(-\log L)^{1/\beta}$) and the number of trials to simulate (J). With these additional specifications, an assessment of the operating characteristics for a particular setting and design can be carried out using Algorithm 2.

2.3 Bayesian Adaptive Design Sample Size Estimation Technique

To implement this Bayesian adaptive design in practice, a common question would involve the maximum sample size (M) to be used to obtain a desired power (B) for a given number of interim looks (K). Power in this setting might be the proportion of trials that will result in a sufficiently precise credible interval that contains or is completely above the efficacy threshold. This design allows for the specification of both a maximum sample size and interim sample sizes that allow for early stopping of enrollment. A sample size calculation should thus return a sample size vector, $\mathbf{V} = (m_1, \dots, m_K, M)$, as opposed to a single number, that will deliver the desired power for a particular prior. In this section, by specifying enrollment rate information, an efficacy threshold, and the same parameter inputs as the nonparametric sample size equation, we indicate how a sample size vector for this design can be selected via simulation.

The vector should have certain properties which motivate the sample size calculation methodology. The maximum sample size should deliver the desired power if prior expectations come to fruition. Early interim looks provide a safeguard for uncertainty about L ; thus, these looks will ideally halt enrollment when L is very favorable. It would be inefficient to conduct an interim look with so little information that no matter the actual value of L , enrollment will always continue. Hence, the first interim look should be related to the “largest plausible” value of L , which will often halt enrollment when L is in fact large. Lastly, the interim looks $(m_1, m_2, \dots, m_K)$ should not be tightly clustered, but rather reasonably spread between the m_1 and M , with each m_K a multiple of enrollment group size (n).

This design provides significant flexibility regarding the placement of interim looks, but using the above principles as a guidance we suggest the following method for determining a sample size vector that will deliver the desired power for a particular prior. Here we simulate trials under a point-mass prior on the investigator's expected value of L to mirror frequentist methods. The investigator needs to specify two values of L , the “expected” value and the “largest plausible” value (e.g., the design has 80% power for $L = .95$, and the investigator believes the largest plausible value of L is .98). The “largest plausible” value will dictate the placement of m_1 , with the goal of halting enrollment when L is in fact extreme.

Once the investigator has provided these two hypothesized values of L , m_1 is set to the sample size which is a multiple of n and delivers the desired power for this Bayesian design when no interim looks are conducted, and device survival is fixed at the “largest plausible” L . Similarly, the initial maximum sample size M_1 is set to the sample size which is a multiple of n and delivers the desired power for this Bayesian design when no interim looks are conducted and device survival is fixed at the expected L . Lastly, the initial sample size vector $\mathbf{V}_1$ is constructed by spacing the remaining $K - 1$ interim looks evenly between m_1 and M_1 at multiples of n . Finding initial values for m_1 and M_1 can be done quickly by simulating a large number of trials (J^*) with the appropriate fixed value of L and an array of sample sizes. These calculations are fast since there are no interim decisions, so there are no additional Monte Carlo draws required, which vastly increases computational efficiency. The smallest sample size that is a multiple of n with estimated power ($\hat{B}$) at least B is used as the initial value.

With the introduction of interim looks, some of the trials will halt early and finish with a trial result that is discordant with the counterfactual trial in which enrollment had not been halted. Because we are concerned with settings in which the expected L is greater than or equal to the efficacy threshold, and the cut off for expected success is less stringent than the cut off for futility, we will see trials halting enrollment for expected success more often than futility. When a trial stops for expected success, the predictive probability of success is high, at least $C_S < 1$, which means the predictive probability of failure is low, no more than $1 - C_S > 0$, but not zero. Some of the trials that halt enrollment may finish with a credible interval slightly larger than ε ; whereas, had these failed trials continued enrollment to the maximum size, some would have been sufficiently powered. Following this counterfactual argument, a non-adaptive trial with sample size M will necessarily have greater power than an adaptive

design with interim looks and maximum sample size M . This does not mean that an adaptive trial reduces power, but rather that a non-adaptive trial with a sample size set to the maximum sample size of the adaptive trial will have greater power (i.e., $B > \tilde{B}_1$ where B_1 is the power of the adaptive trial). In fact, in a fair comparison of power, meaning a non-adaptive trial with a sample size equal to the average sample size of the adaptive trial, the non adaptive trial will have much less power than the adaptive trial.

By design, $B_1 < \tilde{B}$, but we are limited to working with a crude estimate $\hat{B}_1$ from $J = 250$ MCMC draws. Furthermore, due to forcing sample sizes to be a multiple of n , $\tilde{B}$ may be quite a bit larger than B , and introducing adaptivity to the design may result in either $B < \hat{B}_1 < \tilde{B}$ or $B_1 < B < \tilde{B}$. With these limitations in mind, we devise the following strategy for obtaining the appropriately powered sample size vector $\mathbf{V}_i$. We initialize the algorithm with the sample size vector $\mathbf{V}_1$, and implement Algorithm 2 for J Bayesian adaptive trials with L fixed at the expected value, and take the proportion of trials that are successful as B_1 . If $\hat{B}_1 > B$, we assume power was not exceedingly reduced by the introduction of interim looks so that $B < \hat{B}_1 < \tilde{B}$, and we select $\mathbf{V}_1$ for our design. However, if $\hat{B}_1 < B$, we assume introduction of interim looks resulted in $B_1 < B < \tilde{B}$, and we increase M_1 by n to recover the lost power. With the now larger maximum sample size, we reconstruct the sample size vector by spacing the remaining interim looks between m_1 and M_2 to obtain $\mathbf{V}_2$. We continue this process until $\hat{B}_i > B$, at this point we assume B_i is greater than B and we select the sample size vector $\mathbf{V}_i$ for our design. This process leaves m_1 fixed at the initial value, but allows flexibility in the value of M and the remaining interim looks.

Here the enrollment information is again broken down into the size of each enrollment group (n), and the time lag between groups (say, 10 persons every quarter year). Given the number of trials to simulate at each iteration (J), the number of Monte Carlo samples (G) used to calculate the expected success and futility at each interim look, the cut-off values for expected success (C_S) and futility (C_F), the total number of interim looks (K), and a maximum number of algorithm iterations (I), the sample size vector estimation can be summarized by Algorithm 3.

3 Results

3.1 Operating Characteristics

We now report the operating characteristics of our two interim look design with sample size vector (140, 190, 240), and compare them with those of the two frequentist designs, using $J = 1,000$ trials. Though we defer the calculation until later, this sample size vector corresponds to a design with 80% power when device survival is expected to be 95% at 5 years with the largest plausible device survival at 5 years is 98%. Our main investigation below looks at the characteristics when device survival varies between 92% and 98% at 5 years. The parametric models require a fixed value for the Weibull shape parameter β ; for now we defer investigation of performance when the shape parameter is misspecified. Lastly, the enrollment rate is an important factor, since the faster persons are enrolled, the less information the data will contain during the interim looks. The degenerate case is when every person is enrolled at once, where no adaptation is possible. As such we investigate the

effects of three enrollment rates; 10 persons every tenth of a year, 10 persons every quarter year, and 10 persons every half year.

The drop-out/censoring rate (λ) is .08, the expected success cut-off is .8, the futility cut-off is .05, the efficacy threshold is 95%, and the acceptable precision (ε) is .03, throughout. For each simulated trial, $G = 1,000$ Monte Carlo samples were used to estimate the interim expected success and futility.

With both frequentist designs (denoted “K-M” and “Weibull”) having sample sizes fixed at $N = 206$ persons throughout, Table 1 provides a comparison of the three designs when five-year device survival is a range of values. Persons were enrolled at a rate of 10 per quarter year and the shape parameter was fixed at 1. The mean sample size decreases as device survival deviates from the expected value, if device survival were even lower than 92%, the mean sample size would again decrease as more trials would stop enrollment for futility. Table 1 also reports the percentages of trials under each design that terminate successfully, that achieve the desired half-width ($\delta < .03$), and that have an upper interval estimate above the efficacy threshold ($U_{CI} > 95\%$). We see the Bayesian adaptive design makes large gains in achievement of the desired half-width which is driving the increase in power. We mention that, though rare, it is possible for a trial to result in a $\delta < \varepsilon$ and $U_{CI} < L_{ET}$, making all three “% of Trials” columns in Table 1 different. As device survival decreases to 92%, 3% below the efficacy threshold, both frequentist designs become underpowered, achieving the desired half-width less than 5% of the time while correctly excluding the efficacy threshold less than 50% of the time; in contrast, the Bayesian adaptive design achieves the desired half-width 20% of the time while correctly excluding the efficacy threshold nearly 60% of the time. When device survival is at the efficacy threshold of 95%, the Bayesian design enjoys greater success probability of 80.3%, versus below 60% for the two frequentist designs.

Table 1 also reports the average confidence/credible interval bounds, which verify that all of the models are estimating intervals centered about the actual device survival at time T . Though the Bayesian design does have a larger mean sample size in two of the three settings, the design could be tweaked by adjusting the prior distributions on θ , C_S and C_F , or increasing/decreasing K and M to deliver an average sample size of roughly 206 while still maintaining greater power (e.g., changing $C_S = .6$ and $C_F = .2$ reduces the average sample size for the 95% device survival setting to 204 while maintaining 67.4% power). When device survival is 98%, the present Bayesian design enjoys both a similar power and an average sample size 25 persons smaller than the frequentist designs. The frequentist Weibull model fails when there are no device failures observed, this happens because the maximum

likelihood estimate of theta ($\hat{\theta} = \frac{\sum x_i^\beta}{r}$) doesn't exist when $r = 0$.

Table 2 contains a breakdown of the reasons for halting enrollment in the Bayesian adaptive trials. That is, reported are the number of times the trial was stopped for futility ($\hat{E}_F < C_F$) and expected success ($\hat{E}_S > C_S$), as well as the percentage of times these halted trials were ultimately successful (Success %). When a trial stops for expected success, we expect these trials to succeed at a high frequency, whereas when a trial stops for futility we expect the trial to succeed only rarely. Table 2 shows that when stopping for futility, a trial never

succeeds, and when stopping for expected success, a trial succeeds between 15% and 100% of the time depending on actual underlying device survival. For lower rates of survival, trials halt for futility more often; at 92% device survival we see 258 of 1,000 trials halt for futility. However, the number of trials halting for futility drops to 27 and then to 0 as actual device survival increases to 95% and 98%, respectively. As device survival increases, the frequency of trials stopping for expected success increases as well, with actual success occurring in all but 1 trial for 98% device survival. Adjusting the cut-off for expected success and futility to reflect the proportion of information collected would help stabilize the percent of trials halted early that are ultimately successful. For example, at 98% device survival, we see only 43.7% of the trials halted for expected success at the first interim look having a successful final result. Making the halting condition more stringent for earlier interim looks would increase this percentage but also increase the average sample size.

Next, we investigate the robustness of the three designs to shape parameter misspecification, where we expect device survival to follow $Weibull(\beta, \theta)$ for a range of β values while in reality device survival is distributed $Weibull(\beta = 1, \theta)$. The adaptive design has been calibrated to deliver 80% power in the absence of misspecification and the Weibull design is given a sample size equal to the average sample size from the adaptive design. Recall that the frequentist Weibull design and Bayesian adaptive design are both operating with a fixed shape parameter β . If the assumed shape parameter, $\tilde{\beta}$, is incorrect (i.e., $\tilde{\beta} \neq \beta$) there may arise a bias in estimates for $S(T)$. Table 3 explores the effect of various misspecifications with $S(T) = 0.95$. The average bounds of the intervals show that a misspecification can result in under- or over-estimation of device survival at the time point of interest. The direction of the bias in the resulting intervals has much to do with the hazard function, which for the

Weibull model is given by $\lambda(t) = \frac{\beta * t^{\beta-1}}{\theta}$. Table 3 shows that a misspecification can cause a particular design to be either conservative or anti-conservative; therefore, an investigation into the effect of misspecification should be conducted before a particular design is implemented. Table 3 shows that the adaptive design is more robust to misspecification than the frequentist Weibull design. For instance, when $\beta = 0.75$ the limits on the interval as well as the power shift less in the presence of misspecification for the adaptive design relative to the frequentist Weibull design. The performance of the nonparametric Kaplan-Meier design, while poor, is of course unaffected and therefore omitted from Table 3.

Figure 1 shows the fitted survival curves for three $\tilde{\beta}$ values from a Weibull regression model on one simulated trial where $\beta = 1$ and $S(5) = 0.95$. As illustrated by Figure 1, a misspecified $\tilde{\beta}$ implying a hazard increasing faster (or decreasing slower) across time than in reality (i.e., $\tilde{\beta} > \beta$) will under-estimate survival in the later survival times. The maximum likelihood

estimate under a misspecified model is $\tilde{S}(t) = \exp\left(-r \sum \left(\frac{t}{x_i}\right)^{\tilde{\beta}}\right)$; thus, the direction and magnitude of the bias in the resulting survival estimates depends upon the direction and magnitude of the misspecification, as well as the survival time of interest relative to the follow-up time covered by the data. For a thorough discussion shape parameter misspecification in Weibull regression see Xie et al. [13].

Table 4 assesses the effect of enrollment rates. As discussed previously, the adaptive design relies on interim information to predict whether more persons should be enrolled. Faster enrollment means less information will be available at each interim look to make predictions, since information here is essentially follow-up time. The Kaplan-Meier design is again unaffected by this change as the estimate for 5 year survival is independent of anything that happens beyond 5 years. Since interim looks are driven by enrollment count and not calendar dates, the frequentist Weibull design and the Bayesian adaptive design see increased power as the enrollment rate slows, implying longer average follow-up and more information. By contrast, faster enrollment rates result in less power for these parametric designs. The Bayesian adaptive design also shows a shrinking mean sample size as enrollment rates slow. In fact, when the enrollment rate is 10 persons per half year, the Bayesian adaptive design has a similar mean sample size as both the frequentist designs and still enjoys greater power.

3.2 Sample Size Vector Estimation

Here we address the sample size vector estimation problem in a few different settings. Each implementation of Algorithm 3 returns a sample size vector that will deliver the desired power under the design constraints and hypothesized parameter values specified by the investigator. Recall the primary objective is to estimate device survival at $T = 5$ years using a 95% credible interval with a half width less than $\varepsilon = .03$ and an efficacy threshold $L_{ET} = .95$. This subsection reports the sample size vectors estimated and average trial duration for combinations of enrollment rates, device survival percentiles, and numbers of interim looks. For each expected device survival value a corresponding “largest plausible” value is specified to obtain m_1 .

For each sample size vector estimation in Table 5, the desired power is 80% with a dropout/censoring rate of $\lambda = .08$ and Weibull shape parameter fixed at $\beta = 1$. We used $J^* = 3000$ trials to obtain initial values and used $J = 250$ trials for each algorithm iteration with $G = 1,000$ Monte Carlo samples to estimate $\tilde{E}_S$ and $\tilde{E}_F$. Recall that each element of an estimated sample size vector is constrained to be a multiple of enrollment group size $n = 10$. Power in each setting is calculated with respect to a point-mass prior on the expected value of L specified in the first column. The “largest plausible” value of L corresponding to an expected device survival of .95 was set to .98; for expected device survival of .97, the “largest plausible” value of L was set to .99. Both of these “largest plausible” values dictate the placement of m_1 , which is held constant for each unique combination of L and enrollment rate.

One clear pattern in Table 5 is that both slowing enrollment and greater device survival at time T shift the sample size vector downward. Slower enrollment means longer observed survival times, but also increases average trial duration. An obvious effect of increasing K is that interim looks become more frequent. Recall the approximation for $V[L|\mathbf{x}]$ and its dependence upon β , though not illustrated in the table here, as β increases power increases in this setting as follow-up times are predominately greater than T . With the same specifications in Table 5 and device survival of .95/.98, shifting β to 1.5 and 0.5 requires maximum sample sizes of 200 and 260 to maintain 80% power, respectively. Changing the

number of potential interim looks appears to have relatively little effect on the range of the sample size vectors, as the maximum sample sizes are nearly all the same for each particular enrollment rate and device survival percentile combination. Although, for a time lag of .1 between enrollment groups and device survival of .95/.98, we see the maximum sample size jumps from 270 to 290 when the interim looks go from 2 or 3 to 4, respectively. Trial duration is reduced slightly when the number of interim looks increases and the maximum sample size stays the same. Clear patterns from increasing the number of interim looks are hard to discern with the admittedly small number of MCMC draws ($J = 250$) and lack of granularity due to the large enrollment group size ($n = 10$).

Further investigation into the effect of the number of interim looks was conducted by increasing the number of trials used in each iteration of Algorithm 3 to $J = 750$, and adding granularity to the sample sizes by assuming an enrollment rate of $n = 3$ persons every .05 years. The latter change allows each element of the sample size vector to be a multiple of 3, whereas, increasing J reduces the Monte Carlo error in the power estimate each iteration. With these specifications, Algorithm 3 was run for 2, 6, and 10 potential interim looks. Resulting sample size vectors are omitted, but the maximum sample size needed for 80% power increases from 183 to 186 to 189, respectively. Yet, the average trial duration remained nearly constant across the three choices of K . This verifies that a larger maximum sample size becomes necessary to maintain power as the number of interim looks increases. Recall, this pattern is due to a higher frequency of trials halting enrollment for expected success with a growing number of interim looks, and once a trial has halted enrollment, it may finish with a half-width larger than ε due to a random change in device failure following the interim decision.

4 Discussion

The Bayesian adaptive approach provides us with a more flexible design which results in greater power and potentially smaller mean sample sizes when actual survival at time T is different from the assumed value used to power the design. The adaptive design becomes more powerful as the enrollment rate slows and, even when our prior expectations are exactly correct, can offer a smaller average sample size while maintaining greater power. Any $Beta(\gamma, 1)$ prior could be employed on L while retaining conjugacy, a noninformative choice was made here. The Bayesian adaptive design could be improved in settings where we have strong opinions about device survival through a more informative prior, perhaps informed by commensurate historical information [2, 5]. The R code to reproduce the results in this paper or to investigate design characteristics in other settings is available on the second author's software page (<http://www.biostat.umn.edu/~brad/software.html>).

The cost of adding more interim looks while maintaining the same power is the need for a larger maximum sample size. This arises because most of the halted trials do so for expected success with promising outlooks in the first enrollees. These trials with early promise regress to the true device survival rate, which occasionally results in a half-width slightly larger than ε at study end. Employing cut off values for expected success and futility that reflect the amount of information collected at each interim look would reduce the proportion of times a halted trial finishes with an unacceptably large half width (see e.g., Yin, 2012,

Section 6.6 [14]). Though the maximum sample size may increase with more interim looks, the average trial *duration* is left largely unaffected. The adaptivity increases power relative to a non-adaptive trial with the same average sample size by reducing waste. The adaptive trial reduces average sample size and trial duration by balancing the FDA's demand for estimate precision and the company's desire for efficiency. A beauty of this design is that it provides a safeguard against errant prior beliefs about device survival. That is, if the maximum sample size is chosen in a conservative manner, we rarely suffer from insufficient power and still avoid wasting resources via automatic calibration to a smaller sample size when interim results are persuasive.

The current frequentist sample size calculation does not incorporate the idea of an efficacy threshold, whereas the Bayesian framework allows us to calculate the sample size vector for a trial that delivers the desired power in the presence of an efficacy threshold as well as a credible interval half-width requirement. Our simulations show that the frequentist sample size formula often returns a sample size with insufficient power in this setting. Admittedly, there are frequentist approaches not discussed here that use interim looks to re-estimate sample size (see e.g., Yin, 2012, Section 6.7 [14] or Jennison and Turnbull, 2000, [8]). For the Bayesian design, we have shown that a sample size vector can be reliably estimated to deliver a desired power in the presence of an efficacy threshold and a precision requirement; related applications appear in Berry et al., 2010 [2]. This provides an appropriate and detailed statistical justification for the sample sizes of device surveillance studies that are estimation driven, instead of hypothesis driven. The gains in lowered sample size and/or increased power indicate that the approach outlined here would be least burdensome for many practical scenarios.

An important limitation of both the Bayesian adaptive design and the frequentist Weibull design studied above is the fixed shape parameter. As seen in Table 3, misspecification of this parameter causes these designs to give biased results, in either a conservative or anti-conservative direction. The Bayesian adaptive design shows greater robustness to misspecification than the frequentist Weibull design. Though, in the fixed sample frequentist setting, the Weibull design can be extended to allow for simultaneous estimation of the shape and scale parameters; see Meeker and Nelson [10] or Klein and Moeschburger [9]. A corresponding Bayesian design that remains computationally feasible is a direction for future work. In our current setting, the predictive calculations can be more easily (and speedily) conducted using a one-off sampling technique. In the two parameter non-conjugate setting, however, these draws must be made in a vastly more computationally intensive manner (e.g., by iteratively calling an MCMC sampler) since the two-parameter Weibull distribution does not admit a closed form posterior. In principle, trial operating characteristics could still be simulated by calling BUGS or Jags repeatedly from R using the `bugs()` or `rjags()` functions, respectively, and perhaps the `multicore()` function as well if using a multiple core processor. Simulating operating characteristics in this way for a non-conjugate model might take over 500 hours with the current computing capacities and similar numbers of iterations used in Section 3. One potential alternate solution is to use a bivariate normal approximation to the joint posterior of β and θ (see e.g., Carlin and Louis, 2009, Sec 3.1 [3]). Another route would be to abandon the Weibull for a piecewise

exponential model of device survival (see e.g., Ibrahim, Chen, and Sinha, 2001, Sec 3.1 [7]). Regardless, as computational advances inevitably arrive, a more general adaptive model in settings where both parameters are unknown will become feasible.

Acknowledgments

The work of the first two authors was supported by a grant from the Medtronic Corporation.

Appendix

Frequentist Sample Size Calculation for One Sample Survival Curve [11]:

α : 100(1- α)% CI

δ : Half-width of Interval

T : Time of Survival Estimate

L : Expected Device Survival Percentile at Time T

λ : Censoring Rate

Then the required sample size N is given by

$$N = \frac{L^2(1-L)z_{1-\alpha}^2}{\delta^2(1-\lambda)^T}$$

References

1. Berry DA. Bayesian clinical trials. *Nature Reviews Drug Discovery*. 2006; 5(1):27–36.
2. Berry, SM.; Carlin, BP.; Lee, JJ.; Müller, P. *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall; 2010.
3. Carlin, BP.; Louis, TA. *Bayesian Methods for Data Analysis*. third. Chapman & Hall; 2009.
4. Center for Devices & Radiological Health. Draft Guidance for Industry and Food and Drug Administration Staff: Procedures for Handling Section 522 Postmarket Surveillance Studies. Rockville, MD: Food and Drug Administration; 2001.
5. Center for Devices & Radiological Health. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Rockville, MD: Food and Drug Administration; 2010.
6. Center for Devices & Radiological Health: Scientific & Technical Review Committee for Cardiac Pacemaker Leads. Guidance to Sponsors on the Development of a Discretionary Post-market Surveillance Study for Permanent Implantable Cardiac Pacemaker Electrodes. Rockville, MD: Food and Drug Administration; 1992.
7. Ibrahim, JG.; Chen, MH.; Sinha, D. *Bayesian Survival Analysis*. Springer; 2001.
8. Jennison, C.; Turnbull, BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2000.
9. Klein, JP.; Moeschberger, ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer; 2003.
10. Meeker WQ, Nelson W. Weibull percentile estimates and confidence limits from singly censored data by maximum likelihood. *Reliability, IEEE Transactions on*. 1976; R-25(1):20–24.

11. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British Journal of Cancer*. 1977; 35(1):1–39. [PubMed: 831755]
12. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2011.
13. Xie M, Yang Z, Gaudoin O. More on the mis-specification of the shape parameter with weibull-to-exponential transformation. *Quality and Reliability Engineering International*. 2000; 16(4):281–290.
14. Yin, G. *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. Hoboken, NJ: John Wiley & Sons Inc; 2012.
15. Zhang Y, Meeker WQ. Bayesian life test planning for the weibull distribution with given shape parameter. *Metrika*. 2005; 61:237–249.

Figure 1. Parametric Weibull Regression Survival Curves Under Shape Parameter Misspecification ($\beta = 1$)

Table 1
Operating Characteristics for Various Device Survival Values at 5 Years (Counts are out of J=1,000 trials)

Actual Device Survival	Design	Sample Size Mean(SD)	% of Trials			Bounds		
			Success	$U_{CI} > 95\%$	$\delta < .03$	Mean δ	Lower	Upper
92%	K-M	206(-)	2.3	78.4	2.3	0.041	0.880	0.961
	Weibull	206(-)	4.7	48.7	4.7	0.036	0.877	0.949
	Adaptive	214.1(40.1)	18.7	41.1	18.7	0.036	0.874	0.946
95%	K-M	206(-)	32.2	99.2	32.2	0.032	0.919	0.983
	Weibull	206(-)	57.2	97.5	57.2	0.029	0.913	0.971
	Adaptive	220.8(34.2)	80.3	96.3	80.3	0.028	0.916	0.971
98%	K-M	206(-)	95.1	100.0	95.1	0.019	0.960	0.998
	Weibull*	206(-)	99.4	100.0	99.4	0.022	0.947	0.992
	Adaptive	180.8(38.6)	97.9	100.0	97.9	0.021	0.951	0.992

* Weibull Model Failed to Converge 4 times for this set-up

Table 2
Composition of Early Stopping Reason and Final Result by Interim Sample Sizes of the Adaptive Design Various Device Survival Values at 5 Years (Counts are out of J=1,000 trials)

Actual Device Survival at 5 Years	Stopping Reason	Statistic	Halted At Interim			Maximum Sample Size Reached
			140	190	Overall	
92%	$\hat{E}_F < C_F$	Count	179	108	287	678 26.4
		Success %	0.0	0.0	0.0	
	$\hat{E}_S > C_S$	Count	18	17	35	
		Success %	16.7	29.4	22.9	
95%	$\hat{E}_F < C_F$	Count	28	15	43	731 88.8
		Success %	0.0	0.0	0.0	
	$\hat{E}_S > C_S$	Count	87	139	226	
		Success %	43.7	82.7	67.7	
98%	$\hat{E}_F < C_F$	Count	0	0	0	121 100.0
		Success %	-	-	-	
	$\hat{E}_S > C_S$	Count	406	371	777	
		Success %	95.1	99.7	97.3	

Table 3
Operating Characteristics of the Designs in the Presence of Shape Parameter Misspecification (Counts are out of J=1,000 trials)

β	$\tilde{\beta}$	Design	Average Sample Size	% of Trials			Bounds		
				Success	$U_{CI} > 95\%$	$\delta < .03$	Mean δ	Lower	Upper
1.4	1.25	Weibull	217	87.0	98.2	87.0	0.026	0.919	0.971
		Adaptive	216.6	88.8	97.2	88.8	0.026	0.919	0.971
1.25	1.25	Weibull	215	77.7	96.6	77.7	0.027	0.915	0.970
		Adaptive	214.3	79.7	95.7	79.7	0.027	0.915	0.970
1.1	1.25	Weibull	212	64.1	94.8	64.1	0.029	0.912	0.969
		Adaptive	211.1	71.3	95.4	71.3	0.028	0.913	0.970
1.25	1	Weibull	223	91.3	97.8	91.3	0.026	0.921	0.972
		Adaptive	222.9	90.7	97.3	90.8	0.026	0.921	0.973
1.1	1	Weibull	223	81.5	97.2	81.5	0.027	0.917	0.971
		Adaptive	222.2	85.6	97.0	85.6	0.027	0.918	0.971
1	1	Weibull	221	71.3	96.2	71.3	0.028	0.915	0.970
		Adaptive	220.8	80.3	96.3	80.3	0.028	0.916	0.971
0.9	1	Weibull	222	67.7	96.0	67.7	0.028	0.913	0.970
		Adaptive	221.7	74.5	95.4	74.5	0.028	0.913	0.970
0.75	1	Weibull	218	49.4	95.7	49.4	0.030	0.909	0.969
		Adaptive	217.9	67.2	95.6	67.2	0.029	0.913	0.970
0.9	0.75	Weibull	231	81.7	97.1	81.7	0.027	0.917	0.971
		Adaptive	230.6	86.6	96.5	86.6	0.027	0.918	0.972
0.75	0.75	Weibull	229	67.4	96.8	67.4	0.028	0.914	0.971
		Adaptive	228.8	80.9	95.8	80.9	0.028	0.916	0.971
0.6	0.75	Weibull	222	54.7	97.1	54.7	0.029	0.911	0.970
		Adaptive	221.7	69.1	95.7	69.1	0.029	0.914	0.971

Table 4
Operating Characteristics of the Designs with Various Enrollment Rates (Counts are out of J=1,000 trials)

Enroll 10 Persons Per	Design	Sample Size Mean(SD)	% of Trials			Bounds		
			Success	$U_{CI} > 95\%$	$\delta < .03$	Mean δ	Lower	Upper
.1 Years	K-M	206(-)	32.2	99.2	32.2	0.032	0.919	0.983
	Weibull	206(-)	31.6	97.0	31.6	0.032	0.909	0.973
	Adaptive	222.6(30.4)	57.6	96.6	57.6	0.030	0.911	0.971
.5 Years	K-M	206(-)	32.2	99.2	32.2	0.032	0.919	0.983
	Weibull	206(-)	84.6	97.5	84.6	0.027	0.916	0.970
	Adaptive	208.6(34.3)	91.2	97.0	92.4	0.026	0.919	0.971

Table 5

Estimated Sample Size Vectors That Provide 80% Power

<i>L</i> (Expected/Largest Plausible)	10 Persons Enrolled Each	2	Number of Potential Interim Looks		
			Mean Duration	3	Mean Duration
.95/.98	.1 Years	(160, 220, 270)	7.49 Years	(160, 200, 230, 270)	7.40 Years
	.25 Years	(140, 190, 240)	10.37 Years	(140, 170, 210, 240)	10.19 Years
	.5 Years	(130, 170, 210)	14.16 Years	(130, 160, 180, 210)	14.01 Years
.97/.99	.1 Years	(120, 160, 200)	6.83 Years	(120, 150, 170, 200)	6.84 Years
	.25 Years	(110, 150, 190)	9.24 Years	(110, 130, 160, 180)	8.89 Years
	.5 Years	(100, 130, 160)	11.65 Years	(100, 120, 150, 170)	11.86 Years
		4 Mean Duration			
.95/.98	.1 Years	(160, 190, 220, 260, 290)	7.47 Years		
	.25 Years	(140, 170, 200, 220, 250)	10.19 Years		
	.5 Years	(130, 150, 170, 190, 210)	13.81 Years		
.97/.99	.1 Years	(120, 140, 160, 180, 200)	6.82 Years		
	.25 Years	(110, 130, 140, 160, 180)	8.82 Years		
	.5 Years	(100, 120, 130, 140, 160)	11.48 Years		

Algorithm 1

(Interim Enrollment Decision for Interim Look k)

Step 1: Draw a large sample random sample $(\theta_k^{(1)}, \dots, \theta_k^{(g)}, \dots, \theta_k^{(G)})$ from the interim posterior described in equation (1).

For $g = 1, \dots, G$, repeat Steps 2 through 5:

Step 2: There are $m_k - r_k - d_k$ persons remaining in the risk set at interim look k , where r_k and d_k are the cumulative number of device failures and drop outs by interim look k , respectively. Simulate survival times from a truncated $Weibull(\beta, \theta_k^{(g)})$ and drop-out/censoring times from a truncated $Exp(\lambda)$ for these persons, with each person's truncation time equal to their current follow up accrued.

Step 3: Calculate the amount of follow up accrued for each currently enrolled person at study end assuming no additional individuals are to be enrolled. Determine if each current enrollee's simulated event or drop-out/censoring time has been unmasked by study end. Update the predictive posterior and record the predicted credible interval.

Step 4: In addition to the truncated survival times simulated in Step 2, simulate survival and drop-out/censoring times for the $M - m_k$ future enrollees from $Weibull(\beta, \theta_k^{(g)})$ and $Exp(\lambda)$ distributions, respectively.

Step 5: Calculate the amount of follow up accrued for each current and future person at study end assuming M individuals are to be enrolled. Note that in addition to a larger cohort than Step 3, because future persons are enrolled over time, the follow up accrual will be greater. Determine if each individual's simulated event or drop-out/censoring time has been unmasked by study end. Update the predictive posterior and record the predicted credible interval.

Step 6: $\hat{E}_S$ is taken to be the proportion of predicted credible intervals from Step 3 with an acceptable half-width and upper limit above L_{ET} . $\hat{E}_F$ is taken to be the proportion of predicted credible intervals from Step 5 with an acceptable half-width and upper limit above L_{ET} .

Step 7: If $\hat{E}_S > C_S$, halt enrollment for expected success. If $\hat{E}_F < C_F$, halt enrollment for futility. Otherwise ($\hat{E}_S < C_S$ and $\hat{E}_F > C_F$), continue enrollment.

Algorithm 2

(Operating Characteristics for the Bayesian Adaptive Trial)

Step 1: Enroll m_1 persons into the trial by simulating survival times as *Weibull* ($\beta, \eta = T/(-\log L)^{1/\beta}$) and drop-out/censoring times as *Exp*(λ). Calculate the amount of follow up accrued for each person and determine if each individual's event or drop-out/censoring time has been unmasked at the time of the first interim look. Update the interim posterior described in equation (1).

For $k = 1, \dots, K$, repeat Steps 2 and 3:

Step 2: Implement Algorithm 1 to determine the enrollment decision. If enrollment will continue, enroll $m_{k+1} - m_k$ more persons by simulating their survival times as *Weibull* ($\beta, \eta = T/(-\log L)^{1/\beta}$) and drop-out/censoring times as *Exp*(λ). If $m_{k+1} = M$ or enrollment is halted at m_k , proceed to Step 4.

Step 3: Recalculate the amount of follow up accrued for each person and determine whether additional event or drop-out/censoring times have been unmasked at the $(k + 1)^{th}$ interim look. Update the interim posterior, let $k = k + 1$, and return to Step 2.

Step 4: Now that enrollment has halted, follow the currently enrolled cohort to study end. Determine the amount of follow up accrued by each person and whether their event or drop-out/censoring time has been unmasked at study end. Update the posterior distribution and record the sample size, credible interval, result of the trial and any other statistics that may be of interest.

Step 5: (Frequentist Designs) Fit each frequentist model to the first N persons enrolled in the Bayesian adaptive trial, where N is the fixed frequentist sample size given in the appendix. If N is larger than the Bayesian trial sample size, simulate survival and drop-out/censoring times for the additional persons needed. Follow up times may need to be adjusted based on assumptions about enrollment. Fit each frequentist model to the simulated trial and record the confidence interval, result of the trial, and any other statistics of interest.

Step 6: Repeat Steps 1-5 for J independent simulated trials. Calculate the proportion of successful trials under each design, average Bayesian trial sample size, and any other operating characteristics of interest.

Algorithm 3

(Sample Size Vector Estimation for the Bayesian Adaptive Trial)

Step 1: Elicit from the investigator the expected value and “largest plausible” value of the device survival percentile L at time T . Fixing L at the “largest plausible” value, find the minimum sample size that is a multiple of n and delivers a power of at least B under the Bayesian design with no interim looks; take this value as m_1 . Similarly, fixing L at the expected value, find the minimum sample size that is a multiple of n and delivers a power of at least B under the Bayesian design with no interim looks; take this value as M_1 . Lastly, space the remaining interim looks evenly between m_1 and M_1 , each a multiple of n , to obtain an initial sample size vector $\mathbf{V}_1$.

For $i = 1, \dots, I$, repeat Steps 2 and 3:

Step 2: Implement Algorithm 2 for J trials with L fixed at the expected value and sample size vector $\mathbf{V}_i$. Save the proportion of trials that are successful and take that value as $\hat{B}_i$.

Steps 3: If $i < I$ and $\hat{B}_i < B$, take $M_{i+1} = M_i + n$. Space the remaining interim looks evenly between m_1 and M_{i+1} , each a multiple of n , to obtain $\mathbf{V}_{i+1}$ and return to Step 2. Else, proceed to Step 4.

Step 4: If $\hat{B}_i \geq B$, the appropriately powered sample size vector is $\mathbf{V} = \mathbf{V}_i$. Else, the maximum number of iterations (I) has been reached without finding an appropriately powered sample size vector.