Published in final edited form as:

Clin Trials. 2010 June ; 7(3): 219–226. doi:10.1177/1740774510367525.

Sample size re-estimation in a breast cancer trial

Erinn Hade, David Jarjoura, and Lai Wei

Center for Biostatistics, The Ohio State University, Columbus, Ohio, USA

Abstract

Background—During the recruitment phase of a randomized breast cancer trial, investigating the time to recurrence, we found evidence that the failure probabilities used at the design stage were too high. Since most of the methodological research involving sample size re-estimation has focused on normal or binary outcomes, we developed a method which preserves blinding to re-estimate sample size in our time to event trial.

Purpose—A mistakenly high estimate of the failure rate at the design stage may reduce the power unacceptably for a clinically important hazard ratio. We describe an ongoing trial and an application of a sample size re-estimation method that combines current trial data with prior trial data or assumes a parametric model to re-estimate failure probabilities in a blinded fashion.

Methods—Using our current blinded trial data and additional information from prior studies, we re-estimate the failure probabilities to be used in sample size re-calculation. We employ bootstrap resampling to quantify uncertainty in the re-estimated sample sizes.

Results—At the time of re-estimation data from 278 patients was available, averaging 1.2 years of follow up. Using either method, we estimated an increase of 0 for the hazard ratio proposed at the design stage. We show that our method of blinded sample size re-estimation preserves the Type I error rate. We show that when the initial guess of the failure probabilities are correct; the median increase in sample size is zero.

Limitations—Either some prior knowledge of an appropriate survival distribution shape or prior data is needed for re-estimation.

Conclusions—In trials when the accrual period is lengthy, blinded sample size re-estimation near the end of the planned accrual period should be considered. In our examples, when assumptions about failure probabilities and HRs are correct the methods usually do not increase sample size or otherwise increase it by very little.

Keywords

sample size reestimation; time to event; cancer trial; blinded

I. Introduction

Various methods for re-evaluation of a randomized trial's sample size have been proposed [1–5], based on the recognition that the information used to design the study can be updated

during recruitment to avoid inadequate power. Most of the methodological research involving sample size re-estimation has focused on normal or binary, rather than time to event outcomes. Re-estimation with normal or binary outcomes relies on the availability of final measurements for a sub-sample of patients. In contrast, with survival outcomes the final status of patients is not available for those who have not had an event at an opportune time for re-estimation [1]. In such studies, sample size estimates depend mainly on the proposed effect of treatment on the difference in failure rates, usually measured by the hazard ratio, but sample size is also dependent on the number of events that will be observed during follow up which depends on the overall failure rate. A mistakenly high estimate of the failure rate at the design stage may reduce the power unacceptably for a clinically important hazard ratio. This might not be resolved by increasing follow up time until enough events are observed [6]. Usually a decision is made to dismantle the accrual infrastructure close to the time when the planned accrual number will be achieved. To inform the decision about whether to dismantle or to keep accrual open in an ongoing international breast cancer trial, we developed a method of re-estimating the failure probabilities near the end of the planned accrual period.

Breast Cancer Trial context

In an ongoing breast cancer trial, of premenopausal women with hormone receptor positive operable breast cancer, comparing the timing of surgical oophorectomy and mastectomy (random assignment to immediate surgery versus scheduled to be in luteal phase of the menstrual cycle) followed by five years of tamoxifen treatment, we found a strong suggestion that the estimate of the failure (recurrence) rate at the design stage was too high [7, 8]. It was hypothesized that luteal phase surgeries (scheduled) were associated with a hormonal milieu that protects against recurrence. Adjuvant surgical oophorectomy is not the standard of care in the West; however it has been shown to be superior to the standard of care (mastectomy alone) in Asian populations [8]. In the West women are treated with expensive drug combinations to eliminate estrogen production; these are not available in the healthcare systems in our trial countries. During the design phase before trial initiation, estimates indicated that the five year recurrence probability for controls (immediate follicular phase surgery) should be approximately 40%. It was presumed that the recurrence probability should be reduced to 26% in the scheduled surgery (luteal phase) group. These estimates were supposed to be fairly accurate, because they were based on a prior trial with similar patients and an identical common treatment arm (mastectomy and oophorectomy plus tamoxifen). The prior trial data gave us confidence in the initial sample size until we studied extended follow-up data from the same trial, which was not available at the initial design stage. From these data we found that the recurrence probability at 5 years was 25% in the follicular phase surgery group, rather than 40% as was proposed. Because sample size calculations for survival data depend on the number of events that are expected, sometimes planners will rely on extending follow up until the number of required events is observed [9]. Extending follow up time fails to increase power if hazard functions converge.

We found from this trial and other breast cancer trials, that there is evidence of converging hazard functions approximately 5 years post treatment. This was not known at the initial design stage [10–12]. So we chose not to rely on extending follow up time, but rather to

increase accrual. Information from the prior trial might have been enough to re-estimate sample size. However, we decided that as we approached the end of the accrual period in the current trial, it was possible to use blinded estimates of early recurrence probabilities in combination with the prior data to re-estimate sample size.

In re-estimation research, many have suggested using only information not directly related to the effect size as a method to control Type I error [4, 13, 14]. Blinded variance re-estimation is an example. It has been made clear that with unblinded variance estimation, Type I error can be adequately controlled when sample size is not allowed to decrease and when sample sizes are not small [4]. Others have suggested estimating the observed effect of treatment at an interim period for sample size re-estimation including necessary adjustments to control Type I error [2–4]. To preserve the blinding and integrity of study design, to avoid suspicion of manipulating the trial to obtain ‘statistical significance’, and to avoid the related necessary adjustment to critical values, we did not estimate the treatment difference from the ongoing trial data to modify the clinically meaningful alternative hypothesis for testing [6]. Moreover, as would be expected in a typical cancer trial with a time to event outcome, the confidence interval on the treatment effect hazard ratio estimate will inevitably be very wide at the point in time that re-estimation is needed, near the end of the planned accrual period. In addition, in a trial testing a bold hypothesis about survival benefit, it is critical to avoid any suspicion that the protocol for treatment or measurements on the target population has changed as a consequence of finding a smaller effect than was initially hypothesized from the trial data. We note however, it is possible at the time of re-estimation (end of accrual) that information about effect size will be available from independent sources which may impact the accrual decisions. In this trial and many other time-to-event trials the accrual period is lengthy. Given a long enough accrual period re-estimation can take place near the end of this period. Then, follow up on a subset of patients may be long enough to estimate early survival probabilities sufficiently accurately for re-estimation. It makes sense to re-estimate sample size at the end of accrual in our case because we needed to know whether additional subjects would be required before the accrual infrastructure was dismantled.

We show below that the blinded data on failure probabilities, which is commonly made available to a Data and Safety Monitoring Committee, can be used from the ongoing trial to re-estimate sample size.

In this trial:

- There is no imaginable differential risk of adverse events for the two treatment arms
- All patients receive the same surgery (oophorectomy and mastectomy) along with tamoxifen therapy, so that only the timing of the surgery differs and follow up treatment is the same
- The common therapy for both groups has been shown to be better than the standard of care, mastectomy alone (trial patients receive better treatment than what was previously available to them) [8]. Once we accrue the planned number of patients, stopping the trial only means stopping follow-up.

- With this bold hypothesis we would like a sufficiently narrow confidence interval, if the hazard ratio is not significant, to alleviate doubt that non-significant results are due to an under powered study.

Given this, it makes little sense to stop the trial early because of futility. Moreover, with large survival probabilities at five years and a meaningful effect size for luteal surgery benefit, the chance of achieving significance at an interim point with short average follow-up time is extremely low.

We propose that in the context described above, it is possible to focus on re-estimating the recurrence probabilities from the overall trial data, without unblinding. From prior research we expect that focusing on blinded estimation of the survival distribution should preserve an appropriate Type I error rate [4, 14]. Varying reports of Type I error control have been reported in the literature for binary outcomes. Shih and Zhao report that type I error can be increased even under blinded re-estimation, while Gould and Shih (using a different method) report adequate Type I error control when estimating the response rate in the control group [5, 15]. Therefore, we tested this assumption with simulations with parameters chosen to resemble the context of our trial.

The remainder of the article is organized as follows. Section II describes the relationship between sample size, the recurrence probabilities and the hazard function in our example. Two strategies for using ongoing study data for blinded sample size re-estimation are discussed in Section III. Simulations describing the effects of re-estimation on Type I error and on the size of the increase in sample size are provided. Section IV concludes with remarks on what can be learned from our experience.

II. Sample Size as a Function of Failure Rate

Table 1 illustrates a large difference in sample size which occurs when the recurrence rate changes, but the hazard ratio, accrual and minimum follow up time, stay the same. The first row mimics the conditions expected before the trial began. This was based on limited data from the prior trial. Row two, a 35% reduction in the failure probabilities of Row 1, is consistent with what we obtained with extended follow up of the prior trial. With this reduction, the sample size must be increased by 56% to maintain the same power of 81%. Many have designed similar trials using the option to extend planned follow up until the needed number of events is observed (110 in this case). We show in this example how waiting until events accumulate will reduce power when hazard functions converge.

Implications of the Log Rank Test

To compare hazard rates between two groups, we planned to use the log-rank test outlined by Peto and Peto [16]. For two groups, the test statistic can be expressed as:

$$Z = \frac{\sum_{i=1}^D [d_{i1} - Y_{i1}(\frac{d_i}{Y_i})]}{\sqrt{\sum_{i=1}^D \frac{Y_{i1}}{Y_i} (1 - \frac{Y_{i1}}{Y_i}) (\frac{Y_i - d_i}{Y_i - 1}) d_i}},$$

where Y_i is the total number at risk at each failure time, sorted and indexed by i ; Y_{i1} is the number at risk in the treatment group at each of these failure times. The d_i are the total observed number of events and d_{i1} represent the number of events in the treatment group at these failure times. The variable D indexes the last time at which an event occurs. The power of the log rank test is optimized under alternatives of proportional hazard functions ($h_1(t)=\lambda h_2(t)$) where λ is the hazard ratio, and the null hypothesis assumes $\lambda=1$ and that Z follows a standard normal distribution asymptotically [16].

As suggested by previous data (Figure 1), the hazard functions were found to converge at 6 years of follow up, so that the proportional hazards assumption may not be met in the ongoing trial if follow up extends beyond 6 years [8, 10–12]. To maintain power for the restricted hypothesis in the current trial, we intend to limit follow up of patients to a maximum of 6 years. Loss in power from converging hazards is clear from the log rank test equation. After hazard functions converge, the numerator of the log rank statistic's expected value does not increase while the denominator continues to increase, causing the test statistic to become smaller. Under the convergence of hazards, the survival curves maintain separation but the test statistic decreases. It is important therefore to enroll enough patients into a trial to avoid having to extend follow-up to times when the hazard functions converge. An accurate accrual number depends heavily on a confident re-estimate of failure probabilities. An additional practical concern is the minimum amount of time patients are to be followed. In order to have adequate opportunity to assess treatment benefit for all trial patients by allowing them to contribute to the failure distribution, we intend to follow the last enrolled patient for approximately 3.5 years. In this disease and therapy, the majority of failures occur before this time. We proposed a minimum of 3.5 years follow up and a maximum of 6 years follow up for censored observations.

In patient data from the previous trial of mastectomy and oophorectomy followed by tamoxifen, we observed convergence in the hazards for recurrence between those patients who happened to (not randomized to) received oophorectomy and tamoxifen during the luteal phase versus the follicular phase of their cycle (Figure 1) [8]. Calculating the log rank statistic at various time points we see how it was reduced past 5.5 years. At 3.5, 5.5, 7 and 9 years the log rank test statistics were 1.95, 2.06, 1.97, and 1.90 respectively. Using all follow up data through 12 years produced a log rank statistic of 1.73. Although the number of events in our study is sparse past year 6, other breast cancer trials have shown similar convergence and shape of hazards during follow up extended beyond 6 years. Investigators have speculated about causes including that standard treatment with tamoxifen is limited to 5 years and the treatment benefit may disappear after this time, so that the hazard functions become co-incident [8, 10–12].

III. Methods of Re-estimation

Blinded Early Estimation of Survival Probabilities

In our trial, as is common, we expected that the study objectives were not to change and we were willing to assume the hazard of failure to be proportional between randomized groups within the planned maximum follow up time. We used the *blinded* event information ($S_{blinded}(t)$) to contribute to a re-estimate of the sample size based on:

$$S_{\text{blinded}}(t) = \frac{[S_f(t)^{HR} = S_f(t)]}{2}. \quad (1)$$

In the above equation $S_f(t)$ is the estimated survival probability in the control group (follicular phase/immediate surgery group) at time t and $S_f(t)^{HR}$ is the alternative hypothesis survival probability of luteal phase/scheduled surgery. The divisor of 2 comes by assuming an equal number of subjects in each arm. Using this, one can estimate the survival function for either group under the assumption that the alternative HR is correct.

Bootstrapped Intervals for the Re-estimated Sample Size

A second important aspect to our procedure is quantifying the uncertainty in our sample size estimates. We quantified this uncertainty using bootstrap intervals. We repeatedly estimated the baseline survival function developed using equation 1 and from a fixed value of the hazard ratio in order to repeatedly estimate sample size. We used the upper 80th percentile of the bootstrapped interval on sample size as an updated conservative sample size re-estimate [4, 17]. Several authors have suggested calculating confidence intervals for sample size (or on power for a fixed sample size), understanding that sample size will vary as a function of sampling error of point estimates from prior data [4, 18, 19]. In our case, taking account of sampling error of the survival function seems particularly valuable, because of limited information on failure at the end of the planned enrollment period. We propose using the 80th percentile of the sample size bootstrapped interval as a satisfactory upper bound on sample size.

Sample size algorithm—All sample size calculations and bootstrap re-sampling were performed using SAS software, Version 9.1.3 of the SAS System Copyright © 2003 SAS Institute Inc. In particular we utilized the PROC POWER software to estimate sample size allowing specification of the estimated survival functions, our clinically relevant hazard ratio, accrual, and follow up time, and consideration of loss to follow up. This procedure is based on work by Lakatos, requires specification of the control group survival probabilities at multiple time points (i.e. it is not simply based on assuming an exponential survival function) [20]. Sample size was re-estimated for several HRs to show the sensitivity of sample size to the HR.

Methods of Estimating the Survival Function

In order to use equation 1, and to obtain bootstrap estimates of sample size we developed two ways to estimate the blinded survival function through 6 years. Method A relies on prior data in combination with current trial data. For this method the bootstrap intervals rely on the assumption that the models for making the data relevant to our context are not misspecified. Method B estimates the survival function under a parametric Weibull survival time distribution; which replaces the assumption that prior trial data are relevant with a distribution assumption. Convergence of the two methods added confidence to our decisions. A method like either described above that estimates survival probabilities to the end of follow up is required for this sample size algorithm.

Method A: Based on Prior Trial Data Combined with Current Trial Data—

As stated above, from data in our prior trial, the 5 year failure probability was expected to be close to 25% in the immediate surgery/follicular phase, rather than the 40% originally proposed in the initial trial design. If we would have used this information alone to increase the sample size, the proposed increase in sample size would have been 230 to detect a HR of 0.58. To confirm the 25% failure probability at 5 years, we believed it necessary to combine the prior and current trial data.

Taking advantage of relevant data collected from our prior trial of oophorectomy, mastectomy and tamoxifen [8], this method models current and prior data. Limiting the previous trial data to patients who were of the appropriate age range and disease types, we used a Cox regression model and combined the prior trial data and the current trial blinded data to estimate the survival distribution through year 5. The Cox model included an indicator for study (prior/current) and one for luteal phase surgery (1 for luteal, 0 for follicular surgery in the prior trial, and 0.5 commonly assigned to the randomized groups in the current trial). For each bootstrap iteration data was sampled and, from the Cox model, estimates were obtained of the blinded survival probability at each year. Equation (1) is used to solve for survival probabilities in the follicular phase surgery (control) group, $S_f(t)$, using the SAS's PROC NLP software. These control group survival estimates, combined with fixed effect size (HR) and drop out estimates were then used to estimate sample size with the SAS algorithm for each bootstrap iteration.

We had 94 patients from the prior study (averaging 6.8 years of follow up) and combined those with 278 patients from the current trial, which had an average of 1.2 years of follow up during the accrual period (maximum 4 years). These numbers were enough to keep the difference between the median and 80th percentile of bootstrap sampling distribution small. Remember that we suggested that re-estimation occurs when accrual is near completion so that this method, for practical reasons, can only result in not changing or in an increase in sample size. At the time of re-estimation we found that we had more failures than expected based on our prior trial data. The blinded 5 yr failure probability was approximately 45% (an estimated 50% in the immediate/follicular phase group), which consequently resulted in a median and 80th percentile sample size increase of 0 for a HR of 0.58. Adjustment of the Cox model for prognostic characteristics such as nodal status was explored. However, we found that the proportion of node positive patients to be similar in the two studies, and this did not explain differences between prior and current trial failure probabilities.

Method B: Assuming a Parametric Form to Extrapolate beyond current data—

Given that follow up time is limited towards the end of planned enrollment, extrapolation of the survival function is needed if we do not have or have doubts about the relevance of prior trial data. As a means to compare the estimates obtained from Method A, we estimated parameters of a Weibull distribution and used equation (1) to obtain survival probabilities out to 6 years. From previous work, we expected the Weibull to be able to mimic the hazard function appropriately.

As we observed in Method A, our survival estimates at 3 and 5 years, near the end of accrual, were closer to the originally planned failure rates than to those expected when we

initiated plans for re-estimation, resulting in a sample size increase of zero for a HR of 0.58. Note that the difference between the medians and 80th percentile of the bootstrap sampling distribution is small, suggesting that the information from these patients along with the Weibull distribution assumption was enough to keep the sampling error small for the sample size increase. Both the median and bootstrapped upper 80% increase in sample size were 0. Later we provide a simulation example in which this does not hold because the information from accrual data was too limited.

Updated Sample Size Estimate for Current Trial

At the time of re-estimation, we propose that investigators should take advantage of all relevant information newly available since the time of the initial design. Recent adjusted analysis of prior trial data which included longer follow-up, indicated that the alternative hypothesis HR=0.58 was too optimistic. A HR of 0.65 is more similar to estimates obtained from analyses on data unavailable at the time of the initial study design and was not derived from any current trial data (adjusted HR estimates ranged from 0.61 (95% CI 0.32–1.15) to 0.73 (95% CI 0.42–1.29). Furthermore, a HR of 0.65 is associated with an absolute reduction of 8% in the five-year failure probability. An 8% difference between treatment groups is clinically important since the only change in the treatment regime is the timing of surgery, and this distinction adds no cost or known risk to the patient. Given the failure probabilities from the current trial, combined with prior trial data, and a proposed HR of 0.65, we increased the accrual number by 170 patients for a total of 510 randomized patients. We think such considerations might be unusual, and were made here only because additional observational data were available to us near the end of accrual.

Simulation studies

Control of Type I Error—Table 2 describes results of a simulation study that investigated the impact of blinded re-estimation on the Type I error probability, when failure probabilities are incorrectly assumed to be larger than is true at trial initiation. In this simulation, the initial sample size calculation incorrectly assumed that the 3 and 5 year control group survival probabilities were 0.7 and 0.6, whereas the true values were 0.85 and 0.8; 2500 datasets were created using the Weibull distribution. For each dataset, we generated accrual times and used these to truncate survival information past this time for re-estimation in order to mimic what we would see in our trial. Accrual times were generated by using a uniform distribution of accrual between 0 and 2.5 years and we assumed a maximum of six years of follow-up to fit within the context of our trial. For each simulated subject, survival time up to the end of accrual was determined from their full follow-up survival time. Survival time up to the end of accrual was then used for re-estimation, and the full follow-up time was used in the hypothesis testing. For each simulation dataset, 500 bootstrap samples were selected. Using the data at the end of accrual, equation (1) and the sample size algorithm, sample size is re-estimated for each bootstrap repetition. Each bootstrap sample is summarized by both the median and 80% percentile of the 500 samples. We report on the median of the 50th and 80th percentiles of the bootstrap samples increases in sample size. The simulation was repeated for four alternative values of hazard ratios (0.70, 0.67, 0.65, 0.58). From Table 2 we observe that the Type I error rates are very close to the nominal value, 0.05, with SE = 0.0044.

Two further sets of simulations were explored, in which the true values of the survival probabilities were equal to what was proposed (0.70 and 0.60 at 3 and 5 years) and lower than expected (0.55 and 0.45 at 3 and 5 years). As was seen in the previous set of simulations, the Type I error rates were again very close to the nominal value, 0.05. For example, with a HR of 0.58, simulations showed a median increase of zero in sample size, and Type I error rates of 0.0483 and 0.0429 for the 50th and 80th percentiles, respectively for the case in which the survival probabilities were *equal* to what was expected; Type I error rates were 0.0476 and 0.0476 for the 50th and 80th percentiles, for the case in which survival probabilities were *lower* than what was expected. We chose parameters for the simulations that were relevant to our trial. Different situations would require modifying parameter choices and additional simulations.

Sample Size Increases under the Alternative—Further simulations were conducted to investigate the impact on sample size re-estimation for four alternative (true) values of hazard ratios when the proposed HR is 0.65. This is used to understand how misspecification of the HR at initiation would affect re-estimation. We assumed that the proposed survival probabilities of the control group were correct, and only allowed the true HR to be closer to 1 than proposed. The original required sample size was estimated to be 480. When the HR proposed at initiation was correct at 0.65, the median increase and 80th percentile increase from blinded sample size re-estimation were 0 and 74, respectively. Based on the simulation above, when the null hypothesis is true (HR=1) and the true control group survival probabilities are those used to calculate the initial sample size, the median and 80th percentile increases were 0 using our procedure. A lower increase in accrual number happens when using the blinded survival function because of equation (1). When the true value of hazard ratio is closer to 1 than proposed, the re-estimated sample sizes tend to be smaller than the proposed sample sizes, because the test group survival probabilities will be assumed (incorrectly) to be larger than they are under the null, or under any true HR closer to 1 than proposed. In summary, the blinded re-estimation method was robust to overly optimistic proposed HRs.

Impact of a Shorter Accrual Period—In another simulation study, we used the 80th percentile of sample size to judge if there is enough information to re-estimate sample size accurately when accrual time is shortened. Recall that in our data example and in previous simulations, the accrual period was 2.5 years. If accrual time is reduced to half (1.25 years), the re-estimated sample size at end of the accrual period, has much fewer events with which to estimate survival probabilities. With shorter accrual time, re-estimated sample size showed more variability as expected, i.e. higher 80th percentiles of the increase in sample size. For example, with HR=0.65 and accrual of 1.25 years, the 80th percentile increased to 140, which is double the value when the accrual time is 2.5 years. We found in other simulation examples that reducing the accrual time by half (to 1.25 yrs) doubles the 80th percentile increase in sample size. The effect of accrual time on re-estimation accuracy must be considered in planning.

IV. Conclusions

Other trials that rely on accruing events through extending follow-up rely on the assumptions that hazards that do not converge over time. We had evidence of convergence from the prior trial data and other previous reports. Therefore we expect that the increase in sample size near the planned end of enrollment will maintain power by ensuring the needed number of events will be observed during planned follow up. Two important features of our methods are that they do not un-blind the data and they do not decrease the enrollment number. We showed that a substantial increase in sample size would be proposed using either method when the initial sample size was based on too high an estimate of failure probabilities. We showed that in a situation in which the sample sizes were increased substantially, no increase in Type I error occurred. Moreover, when the initial guess of the failure probabilities were correct, we showed that the median increase in sample size was zero and the 80th percentile was not much higher; indicating not much additional cost on average due to these methods and usually little justification to increase the original enrollment number. Our example provided enough information about failure probabilities so that the difference between the median sample size increase and the 80th percentile increase was fairly small. When the information appeared intuitively to be inadequate (where we allowed simulations to have only have ½ the accrual time when re-estimation occurred), the difference between the median and 80th percentile was large indicating too large sampling error. In summary, the proposed method preserves treatment group blinding, sample size is increased when initial guesses of failure probabilities are wrong and Type I error is protected. When initial assumptions about failure rates and HRs are correct, the methods usually do not increase sample size or otherwise increase it by very little.

Acknowledgments

Funding

This work was supported by the National Institutes of Health [P30CA16058-30 and R01CA097375].

References

1. Shen Y, Cai J. Sample size reestimation in clinical trials with censored survival data. *J Am Stat Assoc.* 2003; 98:418–426.
2. Wassmer G. A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics.* 1998; 54:696–705. [PubMed: 9629649]
3. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994; 50:1029–1041. [PubMed: 7786985]
4. Gould A. Sample size re-estimation: recent developments and practical considerations. *Stat Med.* 2001; 20:2625–2643. [PubMed: 11523073]
5. Shih W, Zhao P. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Stat Med.* 1997; 16:1913–1923. [PubMed: 9304763]
6. Korn E, Simon R. Data monitoring committees and problems of lower-than-expected accrual or event rates. *Control Clin Trials.* 1996; 17:526–535. [PubMed: 8974211]
7. Love RR, et al. Oophorectomy and Tamoxifen Adjuvant Therapy in Premenopausal Vietnamese and Chinese Women With Operable Breast Cancer. *J Clin Oncol.* 2002; 20(10):2559–2566. [PubMed: 12011136]

8. Love RR, et al. Survival After Adjuvant Oophorectomy and Tamoxifen in Operable Breast Cancer in Premenopausal Women. *J Clin Oncol*. 2008; 26(2):253–257. [PubMed: 18086800]
9. Gould A. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Stat Med*. 1992; 11:55–66. [PubMed: 1557576]
10. Anderson W, Matsuno R. Breast Cancer Heterogeneity: A mixture of at least two main types? *J Natl Cancer Inst*. 2006; 98:948–951. [PubMed: 16849671]
11. Saphner T, Tormey D, Gray R. Annual hazard rates recurrence for breast cancer after primary therapy. *J Clin Oncol*. 1996; 14(10):2738–2746. [PubMed: 8874335]
12. Retsky MW, et al. Computer simulation of a breast cancer metastasis model. *Breast Cancer Research and Treatment*. 1997; 45(2):193–202. [PubMed: 9342444]
13. Coffey C, Muller K. Controlling test size while gaining the benefits of an internal pilot design. *Biometrics*. 2001; 57:625–631. [PubMed: 11414593]
14. Coffey C, Muller K. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Stat Med*. 1999; 18:1199–1214. [PubMed: 10363340]
15. Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*. 1998; 17(1):89–100. [PubMed: 9463852]
16. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc A*. 1972; 135:185–206.
17. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
18. Dudewicz E. Confidence intervals for power with special reference to medical trials. *Australas J Stat*. 1972; 14:211–216.
19. Taylor D, Muller K. Computing confidence bounds for power and sample size of the general linear univariate model. *Am Stat*. 1995; 49(43–44)
20. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*. 1988; 44:229–241. [PubMed: 3358991]

Figure 1. Disease Free Survival and Hazard of Recurrence Estimates from a breast cancer trial with extended follow up

- a) Solid line is the follicular phase surgery group; the dashed line is the luteal phase surgery group.
- b) Solid lines indicate the smoothed hazard of recurrence, and dashed lines the pointwise 95% confidence intervals for each curve. The black lines are the follicular phase surgery groups; the grey lines are for the luteal phase surgery group.

Table 1

Sample size estimates for a hazard ratio of 0.58, 2 yrs of accrual and a minimum of 4 years of follow up.

Failure probability follicular (control) group (3 years, 5 years)	Failure probability luteal (TRT) group (3 years, 5 years)	Total sample size
0.28, 0.40	0.17, 0.26	340
0.16, 0.26	0.10, 0.16	532

Note: Log rank test with 81% power, 5% type i error, requires 110 total events.

Table 2

Increase in enrollment numbers under the null hypothesis (type i errors using the log rank test)

Proposed <i>HR</i>	Type I error	Proposed sample size	Estimate	Increase in sample size (se) ^{a,b}
0.58	0.0484	314	50th percentile	142 (2.4)
	0.0490		80th percentile	280 (3.0)
0.65	0.0480	480	50th percentile	206 (4.1)
	0.0517		80th percentile	418 (4.6)
0.67	0.0522	548	50th percentile	266 (6.1)
	0.0524		80th percentile	478 (6.2)
0.70	0.0518	678	50th percentile	474 (5.9)
	0.0493		80th percentile	696 (5.8)

Standard error of type i error is ± 0.0044 based on 2500 simulations.^aMedian of 2500 simulations estimate from bootstrap samples^bStandard error of median of sample size estimates from 2500 simulations.