Uncertainty in pilot parameter estimates: A comparison of methods to size full trials

Elizabeth A. Handorf[*] Eric A. Ross[*]

Abstract

Pilot data is a valuable resource for sizing full studies, but parameter estimates based on pilot data can be highly variable. This can lead to substantial under- or overestimation of the necessary sample size, resulting in low power or wasted resources. Several approaches may help decrease the likelihood of underpowering a study, including the use of upper confidence limits of pilot parameter estimates, adaptive approaches with planned interim analyses, and sample size re-estimation without unblinding. In this study, we explore how various methods affect estimated sample size and resulting power. We show that certain re-estimation strategies may substantially improve power, particularly when the pilot study is small. Furthermore, although using upper confidence limits provides increased power, this method leads to large overestimates of sample size. Finally, we show that in most cases considerable benefits come from increasing the number of observations in the pilot study.

Key Words: Sample size, pilot study

1. Introduction

One important reason for conducting pilot studies is to inform the sample size calculations for larger future studies. Especially when prior knowledge of study outcomes is low, pilot data provides a valuable basis for sizing full trials. However, investigators often use pilot parameter estimates in subsequent power calculations as if they are known quantities, when in fact the estimates are themselves variable. We may therefore overestimate or underestimate the needed sample size for the full study, resulting in underpowered analyses or wasted resources.

Limitations of pilot parameter estimates and methods to improve sample size calculations are described in the statistics literature. Browne [1] showed that for continuous outcomes, using the standard deviation estimated from a pilot study ($\hat{\sigma}$) would often result in actual power less than the planned value. He suggested using the one-sided upper 100(1-γ)% confidence limit of $\hat{\sigma}$ in place of $\hat{\sigma}$ in the power calculations. He showed through simulation that the probability of realizing actual power greater than or equal to the planned value is approximately (1-γ) when one uses the 100(1-γ)% upper confidence limit. One drawback of this approach is that the upper confidence limit of $\hat{\sigma}$ is an overestimate of σ. We will therefore overestimate sample size on average, which at best is costly and at worst could make a study infeasible.

Further work in this area includes the analysis of Kieser and Wassmer [7] who investigated the theoretical properties behind the method proposed by Browne. They demonstrated why the upper confidence limit approach led to reliable probabilities of power greater than or equal to the planned value. Julious and Owens [6] provided an alternate method with the goal of obtaining expected power equal to the nominal level chosen. They observed that the variance of the pilot standard deviation leads to expected power less than the chosen value, and provided a method to compensate for this uncertainty.

Adaptive trial designs also address the uncertainty caused by pilot parameter estimates. Planned interim analyses with early stopping rules are popular methods to accommodate

[*] Fox Chase Cancer Center, Temple University Health System, 333 Cottman Ave, Philadelphia, PA 19111

uncertainty in the true treatment effect, but multiple possible sample sizes can also address uncertainty in other parameters such as the standard deviation. [8] The drawback of these methods is the required modification to the type-I error to account for multiple comparisons. [3, 2] Sample size re-estimation techniques more directly address the effect of nuisance parameters in sample size calculation, particularly the method of blinded sample size re-estimation. [5] Once part of the data is collected during the full study, we can obtain more precise estimates of standard deviations or (pooled) proportions, and by masking treatment group type-I error is theoretically preserved. [4]

In this analysis, we consider sample size calculations for continuous outcomes with two groups, suitable for analyzing with the two-sample t-test. We hold the detectable difference of interest as fixed, and focus only on estimation of the standard deviation using pilot data. We consider the effects of increasing the number of pilot samples when we naïvely consider the pilot parameters to be fixed constants in the sample size calculation. We demonstrate that for small or even moderately sized pilots (under 50), there is a substantial probability of meaningful overestimation or underestimation of the required sample size. We then compare the naïve sample sizes to the method of Browne [1] and to adaptive approaches using simulations.

2. Properties of the naïve sample size estimate

We first review standard sample size calculations for a two-sided, two-sample t-test. Let α and β be the desired type I and II error, respectively. We define δ as detectable difference of interest, σ as the true standard deviation, and N as the true sample size per arm needed to power the study given all parameters.

$$N = 2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2/\delta^2 \tag{1}$$

Unfortunately the true value of σ is unknown, so based on pilot data investigators often use

$$N^* = 2\hat{\sigma}_m^2(Z_{1-\alpha/2} + Z_{1-\beta})^2/\delta^2 \tag{2}$$

Where m is the size of the pilot study, $\hat{\sigma}_m$ is the standard deviation estimated in the pilot sample, and N^* is the size of the study estimated using observed pilot data.

As the data from the pilot study are subject to variability, $\hat{\sigma}_m$ and N^* are therefore random variables. As discussed by Browne (1995),

$$\frac{(m-1)\hat{\sigma}_m^2}{\sigma^2} \sim \chi^2_{m-1} \tag{3}$$

It directly follows that

$$\frac{(m-1)\delta^2 N^*}{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2} \sim \chi^2_{m-1}$$

Therefore, the probability that $N^* > N = \Pr(X > (m-1))$, where X is distributed χ^2_{m-1}.

In practice, we would like our estimate of N^* to be close to N. Say that c_L and c_U are some acceptable margins of error we are willing to tolerate in N^*, where c_L is the lower margin $(0 \leq c_L \leq 1)$, and c_U is the upper margin $(1 \leq c_L)$ Let $G(m)$ be the probability that N^* falls within this range.

$$G(m) = \Pr(c_L N \leq N^* \leq c_U N) = \Pr(N^* \leq c_U N) - \Pr(N^* \leq c_L N)$$

Based on the distribution of N^*, this is simply

$$G(m) = F_x((m-1)c_U) - F_x((m-1)c_L)$$

Where $X \sim \chi^2_{m-1}$, and F_x is the cumulative distribution function. Interestingly, this quantity does not depend on σ.

We demonstrate this empirically in Figure 1, where $c_L = 0.8$ and $c_U = 1.2$. We let δ=0.1, α=0.05, and $1-\beta$=0.8. Using a range of values for σ, we drew a sample of size m from the normal distribution to represent pilot data, and calculated the sample size estimate N^* using equation (2). Points represent the proportion of time when $\Pr(0.8N \leq N^* \leq 1.2N)$ based on 1,000 simulations.

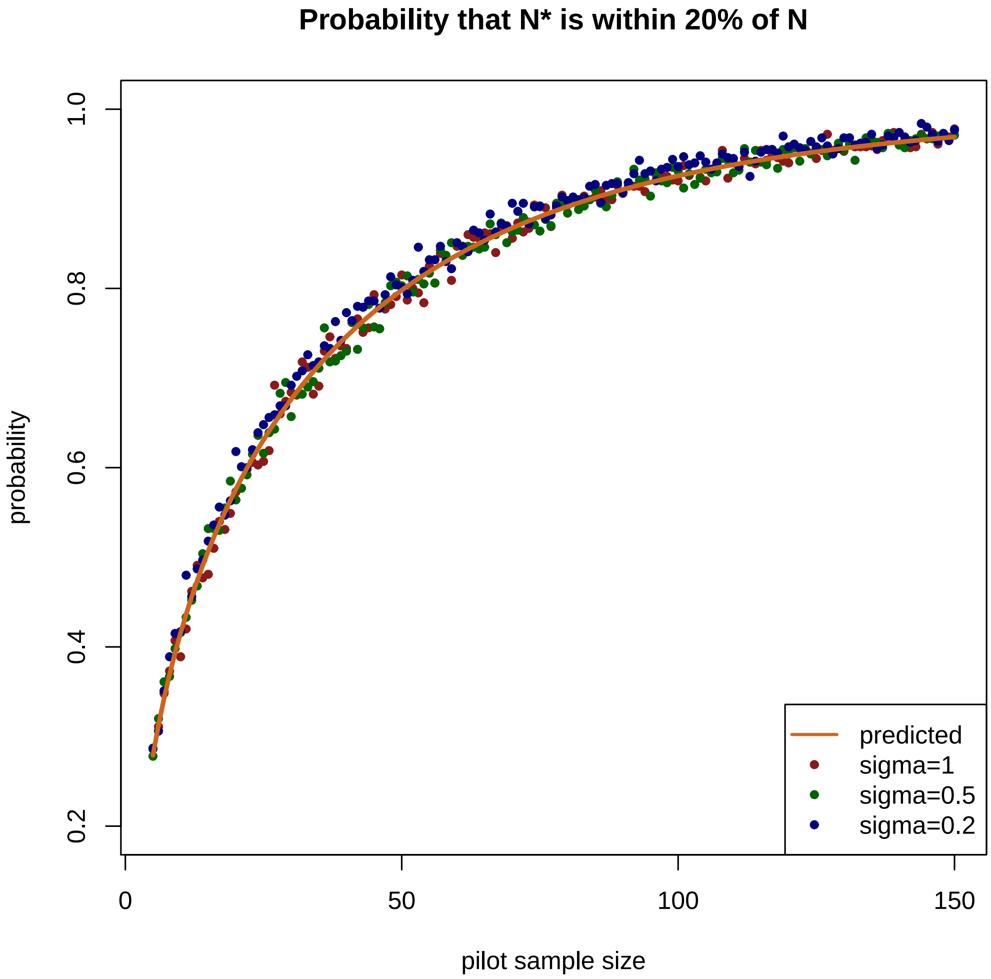

Figure 1: Theoretical and observed probability that estimated sample size is within 20% of true N

We see that even with large pilot studies, there is still a fairly high probability that N^* will be outside the ± 20% range. We note that a 20% margin for sample size does not translate into a 20% margin for power. For example, if $\sigma = 1$, ± 20% in N^* translates into a range of 1,254-1,882, yielding 71-87% power.

We conducted an empirical exploration of power using simulation methods. Using similar methods as above, with δ=0.1 and $\sigma = 1$, we again drew m pilot observations from the normal, and calculated N^*. We then estimated power as $1-\beta = \Phi^{-1}\left(\sqrt{N^*}\delta/2\sigma - Z_{1-\alpha/2}\right)$. The resulting power was categorized as too low ($\leq$ 70 %), too high ($\geq$90%), or within an acceptable range (70-90%). We then repeated this procedure 1,000 times for each value of m. Results are shown in Figure 2. We note that range of power acceptable to an investigator may vary, but the 70-90% range should be practical for most purposes when the desired power is 80%.

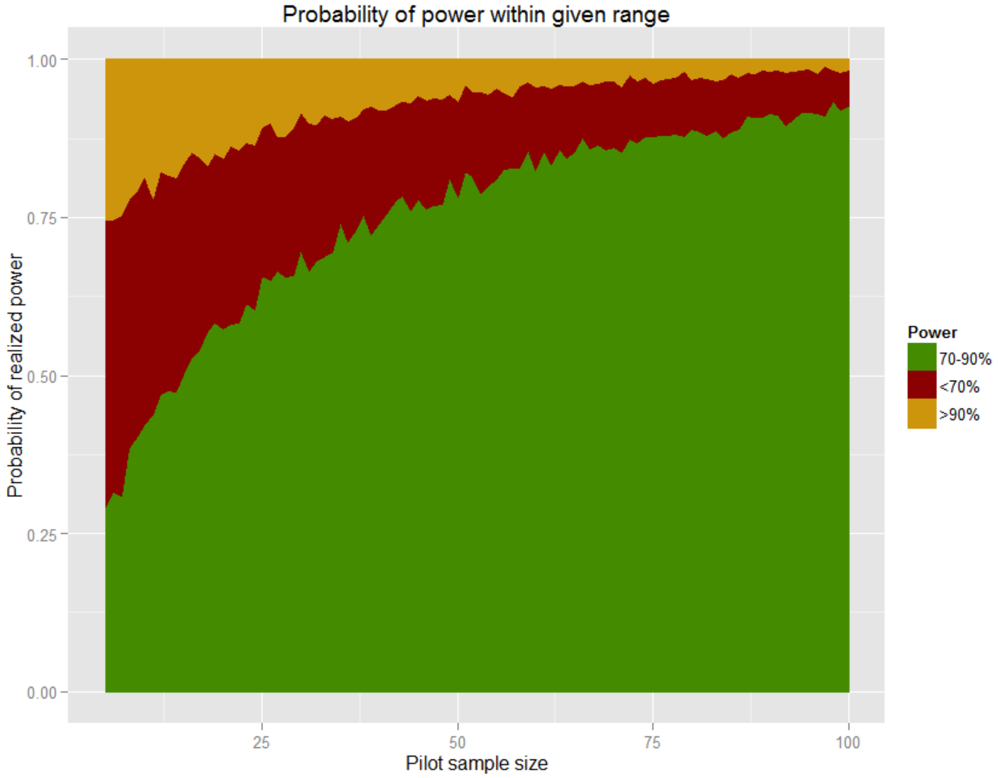

Figure 2: Probability that naïve sample size estimate will produce acceptable power

Based on this simulation, we see that for small pilot studies, it is common to obtain power lower or higher than the desired value. As the pilot sample size increases it becomes less common, with 50 appearing to be a practical cutoff. Larger pilots do improve the probability of acceptable power, but the probability of reasonable power grows more slowly as m increases.

3. Improvements to standard power calculations

Clearly, larger pilot studies will produce more reliable parameter estimates, but even in large pilot studies, error in these estimates can still lead to suboptimal sample size estimates for full trials. Furthermore, large pilot studies will not be feasible in all cases. This motivates methods which improve final study power over the naïve calculation.

3.1 Use upper confidence intervals

As described in the introduction, Browne [1] proposed using the (1-γ)*100% Upper Confidence Limit (UCL) for $\hat{\sigma}_m$. This is easily calculated from the distribution of $\hat{\sigma}_m$. (See equation 3) Browne notes that using the (1-γ)*100% UCL should result in at least a (1-γ)*100% chance of obtaining greater than or equal to the desired power. He suggests using the 80% UCL, but the best choice for γ is not necessarily clear. We note that a major drawback of this procedure is that, on average, the sample size will be overestimated.

3.2 Interim analysis for efficacy

It is widely recognized that an intervention may be demonstrably beneficial (or harmful) before the full trial is complete, and that under such circumstances it is desirable to terminate the study early, motivating methods for interim analyses. If the data are analyzed at multiple timepoints as the study progresses, this can lead to inflation of type-I error, so several authors propose using modified values of α to preserve overall type I error, with different "α spending" functions. [3] Study designers can choose how many analyses are desirable and which spending function to use, but all methods must be pre-specified. [2]

The well-known methods for interim analyses may also help overcome the uncertainty of pilot parameter estimates as having multiple planned analysis points effectively allows for multiple sample sizes. They are also appealing as they are well-understood and widely accepted. However, the analysis points we pre-specify will determine the set of possible final sample sizes, and choosing the best ones presents a challenge. An early analysis (sample size $< N^*$) would be beneficial if we overestimated N, and extending the planned final sample size (final size $> N^*$) would be beneficial if N^* was too low. The more analysis points we allow, the more likely one of them is going to be close to the true N. However, at each analysis point we must spend part of our type-I error, so using a limited number of points is important or it may be difficult to detect differences at the modified values of α.

3.3 Sample size re-estimation

Sample size re-estimation directly addresses the uncertainty of N^* by allowing some portion of the trial to be conducted, and then using this data to calculate an adjusted sample size. Blinded sample size re-estimation is particularly appealing, as treatment indicators are not revealed, which should preserve type-I error without the need for adjustment. [5] In this procedure, the nuisance parameter (σ) is estimated from partial data, but treatment indicators are blinded and the observed effect size is not calculated. This estimate of σ can then be used for a new sample size calculation.

For continuous measures, we must estimate σ from combined treated and untreated patients without knowing treatment indicators. This can be accomplished via an expectation-maximization algorithm, or by the approximation [5]

$$\hat{\sigma}^2 \approx \frac{n-1}{n-2}(\hat{\sigma}_B^2 - \delta^2/4),$$

where $\hat{\sigma}_B$ is the estimated standard deviation from the blinded sample and δ is assumed to be the true difference. In this procedure, we must pre-specify when the re-estimation will occur, and how different the new estimate can be from our original $\hat{\sigma}_m$ before we modify the final study size.

4. Performance of methods to improve attained power

We conducted simulation experiments to evaluate the likelihood of under- or overpowering a study using the designs discussed above. Our baseline method is simply the use of the naïve sample size estimate, N^*. We compared the naïve power to that of the three strategies discussed above. Although each strategy entails its own set of choices for parameters and analysis points, we chose simple designs for illustrative purposes.

Our first alternative method followed Browne, [1], where we used the 75% UCL of $\hat{\sigma}_m$ to estimate N. (denoted N^*_{UCL}) In our second alternative method, the planned final sample size was N^*_{UCL}, allowing for one interim analysis at N^* with stopping for efficacy using the Pocock boundary. In our third alternative, we conducted blinded re-estimation of σ at $N^*/2$, which we denote $\hat{\sigma}_B$. If $\hat{\sigma}_B$ was more than 20% different from $\hat{\sigma}_m$, we re-estimated the final sample size (denoted N^*_{RE}) .

To assess realized power for each method, we conducted simulation studies as follows using R software (version 3.0.1). For each value of $m = 5$ to $m = 100$:

1. Draw m pilot observations from the Normal distribution with a mean of 0 and standard deviation of 1
2. Calculate N^* and N^*_{UCL} based on $\hat{\sigma}_m$
3. Calculate empirical power given $\hat{\sigma}_m$
   (a) Draw 1,000 sets of full data (with the sample size determined by the method under consideration) from Normal(0,1) and Normal(δ,1) to represent the two different study arms
   (b) In adaptive approaches, follow rules stated above to set final sample size
   (c) Conduct a 2-sided 2-sample t-test (with 5% type I error) in each of the 1,000 data sets
   (d) Empirical power is therefore the proportion of times the null hypothesis was rejected.
4. Repeat 1,000 times for each value of m

We found that the sample size re-estimation strategy was generally superior to all other methods explored. This alternative method often provided power within 70-90%, even when the pilot sample sizes were small. Other methods improved as the pilot size increased, but the re-estimation strategy had the highest probability of acceptable power for every pilot size tested (see Figure 3).

Use of the 75% UCL for $\hat{\sigma}_m$ was superior to the other methods at preventing underpowering (see Figure 4), but had the drawback of often overpowering, and hence oversizing the final study (see Figure 5). The interim look similarly prevented underpowering at the expense of often oversizing the study. Interestingly, considering only obtaining acceptable (70-90%) power, both the 75% UCL and interim look methods performed slightly worse than the naïve method (see Figure 3).

5. Conclusion

We have shown that the use of parameter estimates from small pilot studies when sizing full trials can often result in poor estimates of the required sample size, which leads to underpowered studies or wasted resources. Interestingly, when the true distribution of the data is Normal, error in the estimated sample size follows a predictable pattern that does not depend on the true underlying standard deviation but only the size of the pilot study. We

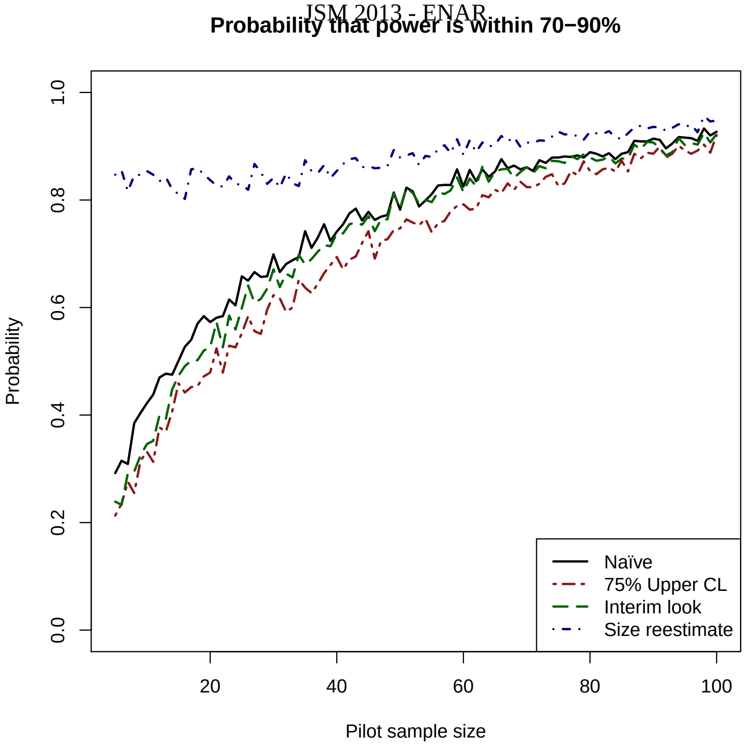

Figure 3: Comparison of probabilities that realized power is within an acceptable range

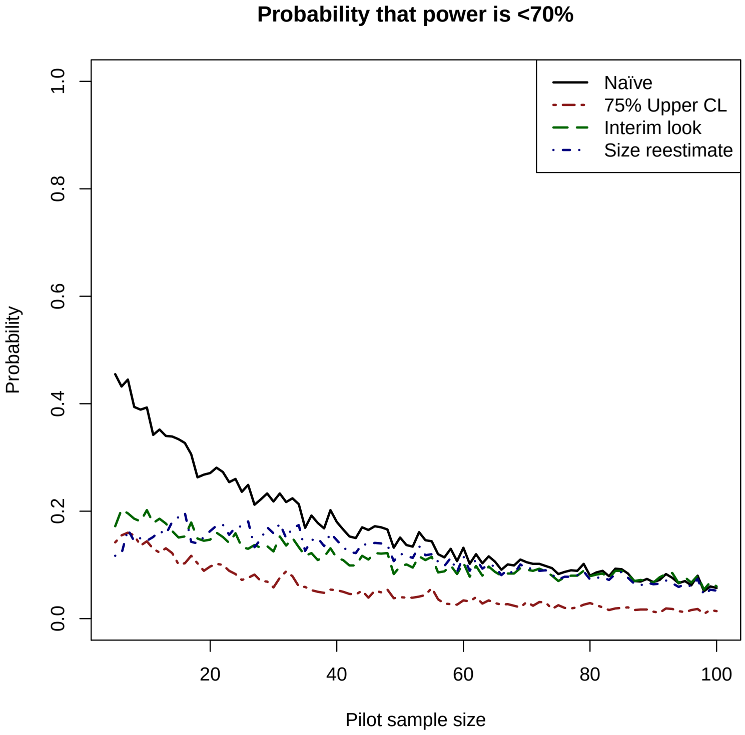

Figure 4: Comparison of probabilities that realized power is too low

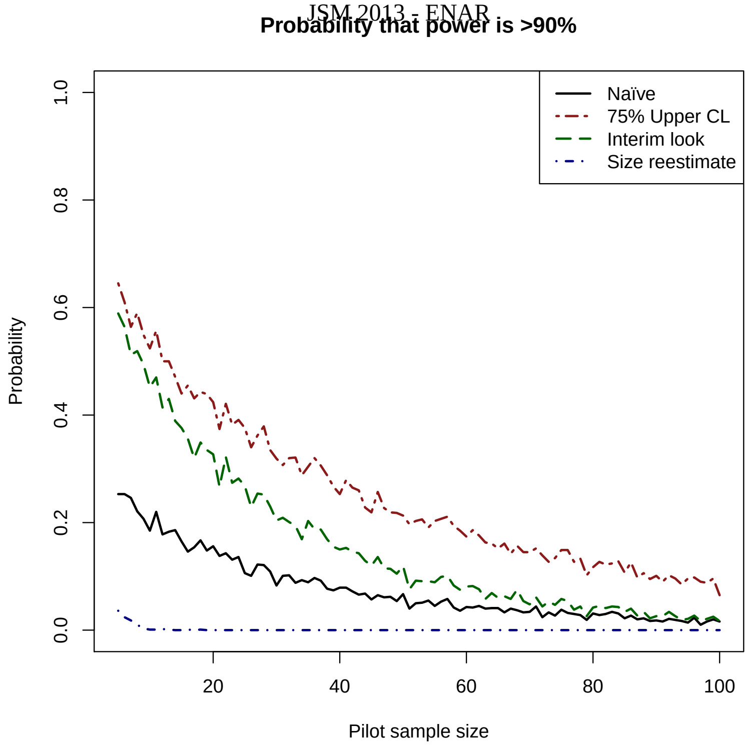

Figure 5: Comparison of probabilities that realized power is too high

demonstrated that large pilot studies produce more robust estimates, with sizes over 50 often resulting in acceptable power. We also found that sample size re-estimation can substantially improve realized power, but that strategies using upper confidence limits or interim analyses were sub-optimal. Although a single interim analysis and the Browne method [1] were not very successful as tested here, we note that there are many different ways that an investigator could use these approaches. Different choices of percent confidence limits or additional interim analyses may improve the performance of these approaches.

The main drawbacks of the blinded sample size re-estimation method are practical considerations. Without knowing a priori what study size is needed, it is difficult to determine whether the necessary recruitment is feasible or whether financial resources are sufficient to fund the required sample size. Although such limitations are beyond the scope of this analysis, they do provide a barrier to implementation in practice.

We have shown that for Normally distributed outcomes, sample size re-estimation should be considered if limited pilot data is available to size a full trial. In future work, we will more thoroughly evaluate the UCL, interim look, and re-estimation methods. Furthermore, we will investigate the performance of these methods for binary outcomes.

Acknowledgements

This work was supported by the Fox Chase Cancer Center core grant (USPHS P30 CA 06927).

References

[1] Richard H. Browne. On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940, 1995.

[2] David L. DeMets and Gordon Lan. *Recent advances in clinical trial design and analysis*. Kluwer Academic Publishers, 1995.

[3] David L Demets and KK Lan. Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352, 1994.

[4] A Lawrence Gould. Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine*, 20(17-18):2625–2643, 2001.

[5] Lawrence A. Gould and Weichung Joseph. Shih. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics-Theory and Methods*, 21.10:2833–2853, 1992.

[6] Steven A. Julious and Roger J. Owen. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics*, 5(1):29–37, 2006.

[7] Meinhard Kieser and Gernot Wassmer. On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*, 38(8):941–949, 1996.

[8] SJ Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199, 1977.