IN PRESS, BIOMETRICS

Controlling Test Size While Gaining the Benefits of an Internal Pilot Design

Christopher S. Coffey
A-1124 Medical Center North
Department of Preventive Medicine
Vanderbilt University Medical Center
Nashville, Tennessee 37232-2637
e-mail: Chris.Coffey@mcmail.vanderbilt.edu

and

Keith E. Muller
3105C McGavran-Greenberg Hall, CB #7400
Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7400

SUMMARY. To compensate for a power analysis based on a poor estimate of variance, internal pilot designs use some fraction of the planned observations to re-estimate error variance and modify the final sample size. Ignoring the randomness of the final sample size may bias the final variance estimate and inflate test size. We propose and evaluate three different tests which control test size for an internal pilot in a general linear univariate model with fixed predictors and Gaussian errors. Test 1 uses the first sample plus those observations guaranteed to be collected in the second sample for the final variance estimate. Test 2 depends mostly on the second sample for the final variance estimate. Test 3 uses the unadjusted variance estimate and modifies the critical value to bound test size. We also examine three sample size modification rules. Only Test 2 can control conditional test size, align with a modification rule, and provide simple power calculations. We recommend it if the minimum second (incremental) sample is at least moderate (perhaps twenty). Otherwise the bounding test appears to have the highest power in small samples. Reanalyzing published data highlights some advantages and disadvantages of the various tests.

KEY WORDS: Interim power analysis, Sample size modification, Stein's procedure

1. Introduction

1.1 *Motivation*

In designing a study, researchers want to collect a sample large enough to achieve a target power, while insuring a target test size. To choose a sample size, scientists must often conduct a power analysis based on an educated guess or variance estimate of uncertain validity. For example, consider a study of Contrast Limited Histogram Equalization (CLAHE), by Pisano *et al.* (1998), to determine which image processing parameters (Region Size and Clip) improved observers' ability to detect breast cancer in mammograms. Two sets of hypotheses were tested: 1) the interaction of Region and Clip, with associated simple main effects, and 2) tests of the difference between each of nine processed conditions and the unprocessed condition. We focus on the second set, which reduces to nine paired t-tests. The scientists sought at least 90% power

to detect a difference of 0.1 (in log contrast units) in each of the nine tests. A Bonferroni correction was applied to correct for multiple comparisons. Power calculations based on a variance estimate of 0.0065 from an unpublished earlier study led to the conclusion that twenty subjects would be required. Unfortunately, after completion of the study, an error in the images was discovered which caused concern about the validity of the variance estimate. Since there was no consensus as to the direction of bias in the estimate, the scientists feared that this uncertainty could lead to a study with too little or too much power.

One solution originally explored was to incorporate an internal pilot design, which uses some fraction of the planned observations to estimate error variance and modify the final sample size. This allows increasing power if the original variance value was too small, or reducing the sample size if the original value was too large. Such designs differ from traditional pilot studies in that those observations used to estimate the variance are included in the final analysis. A consequence of using an internal pilot design is that the final sample size becomes a random variable. Ignoring the randomness of the resulting final sample size may inflate test size and reduce coverage of confidence intervals. This bias varies directly with $\gamma = \sigma^2/\sigma_0^2$, the unknown ratio of population variance to the value used for planning.

Jennison and Turnbull's (2000, Chapter 14) recent review of such designs may lead to the impression that internal pilots have little or no effect on test size: "In fact, effects on Type I error rate are usually slight, as we shall see in applications of sample size re-estimation to binary data in Section 14.2.1 and normal data in Section 14.2.2." (p.280). This statement may be true with the moderate to large sample sizes typical of phase III clinical trials. However, Coffey and Muller (1999) demonstrated that there are study designs for which the inflation of test size may cause concern. In addressing this problem Jennison and Turnbull claimed that "The most serious difficulties arise when the variance of a normal response is estimated with fewer than 10 degrees of freedom, and we shall describe adjustments to the simple sample size calculations which protect the Type I error in such situations." (p.280) This is far too optimistic. The claim

was first made by Birkett and Day (1994) based on simulation results. Exact calculations in Coffey and Muller (1999) refute the generality of this claim. Adjustments *are* described later in the book. However, other than Stein's approach (which we generalize in this paper to the General Linear Model), none *guarantee* control of test size. In this paper, we propose and evaluate tests which guarantee control of test size for any internal pilot design using a General Linear Univariate Model (GLUM). We seek a test that controls test size while maintaining most of the benefits in power and expected sample size achieved by the unadjusted test.

1.2 *Notation*

We study the same model as in Coffey and Muller (1999), which includes the two sample t-test as a special case. See the Appendix and Table 1 for a description of the model and notation. Let $\chi^2(\nu, \omega)$ indicate a noncentral χ^2 and $F(\nu_1, \nu_2, \omega)$ a noncentral F. For fixed ψ, which may or may not depend on n_+, let $f(\alpha, \psi) = F_F^{-1}(1-\alpha; a, \psi)$, let $\omega(\psi)$ represent the noncentrality which satisfies $P_t = 1 - F_F[f(\alpha_t, \psi); a, \psi, \omega(\psi)]$, and define

$$q(n_+, \gamma, \psi) = \frac{\nu_1 \boldsymbol{\theta}_*' [\boldsymbol{C}(\boldsymbol{X}_+'\boldsymbol{X}_+)^- \boldsymbol{C}']^{-1} \boldsymbol{\theta}_*}{\gamma \sigma_0^2 \omega(\psi)} \,. \tag{1}$$

1.3 *Known Results*

Stein (1945) derived a two-stage procedure to determine the required size of the second sample such that the width of the final $(1-p)\%$ confidence interval around the mean will be no larger than some specified amount. The method uses all n_+ observations for estimating the mean, but only the first n_1 observations to estimate σ^2.

Wittes and Brittain (1990) introduced the concept of an internal pilot study for the two sample t-test. Based on simulations of test size, power, and expected sample size for an example with $n_{+,\min} = 86$, they suggested ignoring the randomness of N_+ and using fixed sample methods for testing. We refer to this approach as the unadjusted test (Test 0). Wittes *et al.* (1999) examined test size and other analytic properties of the unadjusted t-test.

Coffey and Muller (1999) described an exact algorithm for computing test size, power, and expected sample size for the unadjusted test in any GLUM with fixed predictors and Gaussian errors. They demonstrated that various design choices have a wide range of effects on test size, power, and expected sample size. Furthermore, if a fixed sample size modification rule is used, Coffey and Muller proved that

$$\Pr\{N_+ = n_+\} = F_{\chi^2}[q(n_+, \gamma, \nu_+); \nu_1] - F_{\chi^2}[q(n_+ - m, \gamma, \nu_+); \nu_1] \,. \tag{2}$$

For testing under an internal pilot design with a two sample t-test, Zucker *et al.* (1999) considered an approach similar to Stein's, thus controlling the *unconditional* test size. They compared the power of the modified Stein test to both the unadjusted test, and an idealized version assuming an outside source reveals σ^2 at the end of study. For a variety of designs, the Stein test usually provided greater power. However, the modified Stein test applied a sample size modification rule based on a noncentral t with ν_1 degrees of freedom, while the unadjusted test modified sample size based on a noncentral t with ν_+ degrees of freedom. The concomitant larger $\mathcal{E}N_+$ of the modified Stein approach may explain the larger power. In the notation of this paper, Zucker *et al.* also considered using $\widehat{\sigma}^2_*(n_+) = SSE_*(n_+)/n_2$ in the denominator of the test statistic, thus controlling *conditional* test size. They compared the procedure to a permutation-based test proposed by Gould and Shih (1992).

Denne and Jennison (1999) considered using the unadjusted test statistic along with the Stein-based rule for modifying sample size. They suggested modifying the critical value with a heuristic reduction in the error degrees of freedom. On average, the method appears to correct for test size inflation (based on simulations) but does not *ensure* either an unbiased test or bounding of test size at the target level.

Kieser and Friede (2000) described an algorithm for insuring bounded test size in the t-test setting. However, their bound is based on an approximation known to have test size greater than or equal to that of the unadjusted test.

1.4 *Overview of New Results*

Our new results may be summarized as follows. 1) In §2.1 and §2.2, we generalize internal pilot methods due to Stein (1945) and Zucker *et al.* (1999) for a t-test to the GLUM. 2) In §2.3 we describe a new bounding method which allows using *all* of the data in tests which have an upper bound on test size of α_t (the target level chosen by the analyst). The bounding method described in §2.3 is based on the *exact* distribution of the unadjusted test statistic and hence is a slight improvement and substantial generalization over the method for t-tests described by Kieser and Friede (2000). 3) Most previous research on internal pilot designs has focused on the choice of test statistic. In §3 we address implications of the choice of sample size modification rule. We believe that separating the discussion of sample size modification and testing rules provides a new and superior framework for evaluating internal pilot analysis methods.

2. Tests For Controlling Test Size of the General Linear Model

2.1 *Test 1: Use First Sample Plus Those Additional Observations Guaranteed to be Taken*

In the two sample t-test setting, Zucker *et al.* (1999) controlled test size by modifying Stein's approach. Stein's approach extends to the GLUM, but can be improved by using the first $n_{+,\min} \geq n_1$ observations for the final variance estimate. Consider the (conditional) test statistic

$$F_1(n_+) = \frac{SSH(n_+)/a}{\widehat{\sigma}^2(n_{+,\min})}\,, \tag{3}$$

with $\widehat{\sigma}^2(n_{+,\min})$ the estimate using the first $n_{+,\min}$ observations. Due to the dependence of n_+ on $\widehat{\sigma}_1^2$, $F_1(n_+)$ will not follow an F distribution. Write the true (conditional) noncentrality as $\omega(n_+;\gamma,\boldsymbol{\theta}) = \boldsymbol{\theta}'[\boldsymbol{C}(\boldsymbol{X}_+'\boldsymbol{X}_+)^{-}\boldsymbol{C}']^{-1}\boldsymbol{\theta}/(\gamma\sigma_0^2)$. For $d \in \{0,1\}$, let $q_d = q(n_+ - dm, \gamma, \psi)$. For a critical value, f, and $n_{2,\min} = n_{+,\min} - n_1$, the conditional power of Test 1 may be written

$$P_1(\gamma,\boldsymbol{\theta}|n_+) = 1 - \int_{q_1}^{q_0} \Pr\left\{ \frac{\nu_{+,\min}}{fa} \cdot \chi^2[a, \omega(n_+;\gamma,\boldsymbol{\theta})] - \chi^2(n_{2,\min}) \leq t \right\} \times \frac{f_{\chi^2}(t;\nu_1)}{\Pr\{N_+ = n_+\}} dt\,. \tag{4}$$

In turn, the unconditional power of Test 1 is

$$P_1(\gamma,\boldsymbol{\theta}) = 1 - \sum_{N_+=n_{+,\min}}^{n_{+,\max}} \int_{q_1}^{q_0} \Pr\left\{ \frac{\nu_{+,\min}}{fa} \cdot \chi^2[a, \omega(n_+;\gamma,\boldsymbol{\theta})] - \chi^2(n_{2,\min}) \leq t \right\} \times f_{\chi^2}(t;\nu_1)dt \,. \tag{5}$$

Note that the integrand in (5) depends on n_+ only through $\omega(n_+;\gamma,\boldsymbol{\theta})$. Under the null $\omega(n_+;\gamma,\mathbf{0}) = 0$, $F_1(N_+) \sim F(a,\nu_{+,\min})$, and using a critical value of $f(\alpha_t,\nu_{+,\min})$ preserves unconditional test size.

2.2 Test *2: Depend Mostly on the Second Sample (Ignore the First Sample)*

In the two sample setting, Zucker *et al.* (1999) suggested using $\widehat{\sigma}^2_*(n_+) = SSE_*(n_+)/n_2$ in the denominator of the t. Here $SSE_*(n_+)$ equals the part of $SSE(n_+)$ which is orthogonal to SSE_1. Generalized to the GLUM, the (conditional) test statistic becomes

$$F_2(n_+) = \frac{SSH(n_+)/a}{\widehat{\sigma}^2_*(n_+)} \,. \tag{6}$$

Note that $\widehat{\sigma}^2_1$ is not present in $F_2(n_+)$. Thus $F_2(n_+) \sim F[a, n_2, \omega(n_+;\gamma,\boldsymbol{\theta})]$, and using $f(\alpha_t, n_2)$ as the critical value controls both conditional and unconditional test size. For Test 2, conditional power equals $P_2(\gamma,\boldsymbol{\theta}|n_+) = 1 - F_F[f_2; a, n_2, \omega(n_+;\gamma,\boldsymbol{\theta})]$, and unconditional power is

$$P_2(\gamma,\boldsymbol{\theta}) = 1 - \sum_{N_+=n_{+,\min}}^{n_{+,\max}} F_F[f(\alpha_t,n_2); a, n_2, \omega(n_+;\gamma,\boldsymbol{\theta})] \times \Pr\{N_+ = n_+\} \,. \tag{7}$$

Note that when $n_2 \gg n_1$, $\widehat{\sigma}^2_*(n_+)$ dominates the final variance estimate, and we do not lose much information by ignoring the first sample. At the other extreme, if $n_2 \ll n_1$, this test ignores the majority of the information and may have less power than the other tests we consider.

2.3 Test *3: Bound Test Size*

For the unadjusted test (Test 0), actual test size is never less than α_t, for $\gamma \in (0,\infty)$ (see Appendix). Under H_0, power depends on γ only through the regions of integration. Simple iterative techniques allow solving $\partial P(\gamma_*,\mathbf{0})/\partial\gamma = 0$ for γ_*. If γ_* is the unique maximum value

of γ, then $P(\gamma_*, \mathbf{0})$ equals the corresponding maximum test size. We have not been able to prove conclusively that the equation has a single solution, although we believe this to be true. The smoothness of the functions allows using a grid search to check.

We used a SAS IML® program (SAS Institute, 1989) to compute maximum test size for a variety of designs based on the CLAHE example. The results (not shown) led us to conclude that the choice of π strongly affects maximum test size, particularly if we allow a decrease from the sample size originally planned.

Changing the critical value used with Test 0 changes the maximum test size. Test 3 bounds test size by rejecting H_0 if $F(n_+) > f(\alpha_*, \nu_+)$, with $\alpha_* \leq \alpha_t$ chosen such that the maximum test size across γ equals α_t. The appeal lies in using the likelihood ratio test statistic and all available data in estimating the variance. However, Test 3 may be conservative since we merely bound test size. Furthermore, two nested iterative searches are required to find α_* for each design. Thus, Test 3 requires far more computations to implement than do Tests 0-2.

3. Rules for Modifying the Sample Size

Implementing an internal pilot design requires choosing a sample size modification rule as well as a test. Most previous work focused on Test 0 with sample size modification based on an $F[a, \nu_+, \omega(n_+; \gamma, \boldsymbol{\theta})]$ approximation. Call this sample size modification Rule 0. Similarly, an $F[a, \nu_{+,\min}, \omega(n_+; \gamma, \boldsymbol{\theta})]$ approximation defines sample size modification Rule 1, and an $F[a, n_2, \omega(n_+; \gamma, \boldsymbol{\theta})]$ approximation defines sample size modification Rule 2. See equation 2 for the probability density function (pdf) of N_+ using Rule 0. Likewise, the pdf of N_+ for Rules 1 and 2 can be obtained easily by changing the truncation regions to $\{q(n_+, \gamma, \nu_{+,\min})\}$ and $\{q(n_+, \gamma, n_2)\}$, respectively. Rule 0 will always have the smallest $\mathcal{E}N_+$, while $\mathcal{E}N_+$ for Rule 1 may be larger or smaller than for Rule 2, depending upon γ. Note that any other rule may be chosen, provided that it maps mutually exclusive and together exhaustive regions of $\widehat{\sigma}_1^2$ into distinct values of N_+.

The choice of modification rule and test interact with each other. We describe the combination of a modification rule and test as defining a method. For example, Method 0/0 refers to using Rule 0 with Test 0. The most natural combinations are 0/0, 1/1, and 2/2. Furthermore, since Test 3 merely involves an adjustment to account for the bias in Test 0, Method 0/3 seems natural for the bounding test. However, since only $F_2(n_+)$ follows an F distribution, only Method 2/2 aligns the test and modification rule. We see no statistical basis for preferring any of the other obvious combinations. However, to simplify explaining the design to non-statisticians, we recommend using one of the natural combinations, unless there is a compelling reason to use an alternate method.

4. Numerical Evaluations

We illustrate our results with the CLAHE example [Pisano *et al.* (1998)], described previously. Consider using an internal pilot design with $\pi = 0.5$, which implies $n_1 = 10$, $n_{+,\min} = 10$, and $n_{+,\max} = 30$. The unadjusted test and sample size allocation rules can reduce $\mathcal{E}N_+$ for $\gamma < 1$ or increase power for $\gamma > 1$. Unfortunately, for some values of γ the design may nearly double α_t. The new methods described in this paper eliminate the problem.

An ideal method would control test size, achieve target power, and minimize $\mathcal{E}N_+$. Choosing the best method requires examining $\mathcal{E}N_+$ and power simultaneously, since we can always increase power by taking a larger sample. Therefore, to characterize the tradeoffs, for each method we examined $\mathcal{E}N_+$ and the power to detect $\theta_* = 0.10$ in the CLAHE example, for $\gamma \in \{0.5, 1.0, 2.0\}$ and $\pi \in \{0.25, 0.5, 0.75\}$. All computations were performed on a 550 MHz Pentium Pro with 256 Mb RAM, SAS IML® 6.12, and Windows NT® 4.0. For each rule, computations in Table 2 for Tests 0-3 required approximately 15 min., 2 min., 1 sec., and 40 minutes, respectively. Test 2 is the fastest since its power computations require only the distribution function of an F, while the other tests require numerical integration for power. Accurate approximations for the integrals would greatly reduce the discrepancies in computation times.

Table 2 contains a rich array of information. For sake of brevity, we highlight only a few of the conclusions. In this small sample setting, the choice of method has a big impact on power and $\mathcal{E}N_+$. When γ is such that Rules 0 and 1 lead to small n_2 with high probability, Rule 2 is penalized. For small $n_{+,\min}$, Rule 1 requires a much larger $\mathcal{E}N_+$. However, as $n_{+,\min}$ increases, differences between Rules 0 and 1 become smaller.

Despite many desirable properties, Method 2/2 does not provide acceptable power in the small samples of Table 2. Furthermore, the other methods using Test 2 can lead to power below that of a fixed sample design, suggesting Test 2 is not appropriate for this example. Among the other tests that control test size, Test 3 appears to be have the most power. However, regardless of the modification rule chosen, the loss of power for Tests 1 or 3 compared to using the biased Test 0 becomes smaller as π, and hence n_1, increases. Fortunately, researchers often have great flexibility in choosing π. If so, the choice of method can be based on convenience.

We also examined a range of examples with roughly one hundred observations (results not shown). In such cases, the choice of method appears to have little impact on power and $\mathcal{E}N_+$. With sample sizes this large or larger, the desirable properties of Method 2/2 make it most appealing.

5. Example Revisited

For the CLAHE study, the results of Table 2 illustrate that our new methods achieve most of the benefits of an internal pilot design while controlling test size. Regardless of the rule chosen, Test 3 appears to have the most power among the tests which control test size. Furthermore, there appears to be little advantage to choosing Methods 1/3 or 2/3, and hence a larger $\mathcal{E}N_+$, over the natural combination of Method 0/3. Therefore we reanalyzed the data using Method 0/3.

Table 3 contains the estimated differences and associated hypothesis tests for the nine comparisons from the CLAHE study, both as originally reported in Pisano, *et al.* (1998), and also analyzed with Method 0/3. Note that positive differences correspond to improved performance

as a result of processing. With Method 0/3 and $n_1 = 10$, we obtain $\widehat{\sigma}_1^2 = 0.0028$ ($\widehat{\gamma}_1 = \widehat{\sigma}_1^2/\sigma_0^2 = 0.43$), which leads to a final sample of only $n_+ = 11$. The fixed sample size design, with 20 observations, finds two conditions with significantly improved performance and one condition with significantly worse performance. However, with Method 0/3 we find no statistically significant results. The different conclusions occur because all observed differences are less than the value for which the study was designed to detect ($\boldsymbol{\theta}_* = 0.10$). Thus the significance observed for the fixed sample design is a consequence of taking a sample larger than required. This highlights the importance of insuring that $\boldsymbol{\theta}_*$ truly is the smallest, scientifically meaningful difference, which holds for any power analysis, not just those for internal pilot designs.

Test 3 controls unconditional test size, but does not directly control conditional test size. Figure 1 plots conditional test size and power as a function of γ for the CLAHE example. The vertical line represents $\widehat{\gamma}_1$, while the horizontal lines represent the test size and power targets. Near $\widehat{\gamma}_1$, Test Method 0/3 controls both conditional test size and power. For large values of γ, conditional power falls below the target, while conditional test size exceeds the target. Thus conditional test size and power of Method 0/3 depend on how well $\widehat{\gamma}_1$ estimates γ.

6. Discussion

This work extends many internal pilot design concepts from the t-test to the classic General Linear Univariate Model setting. Such an improvement may serve as a basis later for cases such as logistic regressions or time-to-event models, where group sequential methods are currently most commonly used. The overall conclusions may be summarized as follows. 1) The benefits of an internal pilot design may be gained while controlling test size. 2) With current methods, controlling test size exactly requires ignoring some information to obtain an unbiased estimate of σ^2, while using all of the information only allows bounding test size. 3) Implementing an internal pilot design requires choosing a sample size modification rule as well as a test. 4) Although many combinations of test and rule misalign testing and power, we recommend

using the natural combinations unless there is a compelling reason not to. 5) For moderate to large samples, Method 2/2 appears best, due to the alignment of the test and modification rule, control of conditional test size, and ease of power computations. 6) In small samples, no single choice dominates, although the bounding test appears to cost the least power to control test size. 7) Differences in power among the methods become smaller as n_1 increases.

Our results suggest that, in most situations, one can control test size with minimal impact on power and expected sample size. We caution the reader that only calculations specific to each particular design can insure that the design does not inflate test size. Hence we recommend always using a test which controls test size to avoid any risk of test size inflation.

Many features of internal pilot designs merit further research. Sample size modification rules should be sought which align the test statistic with the interim power analysis. The implications of using a quantile estimate of variance to determine the final sample size (Taylor and Muller, 1995) merit separate study, as does using confidence interval width as the measure of merit for internal pilot designs. Extensions to repeated measures and other multivariate models are needed (as for the example). Finally, fast and accurate approximations of test size and power of the bounding test would greatly increase its appeal.

ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful comments from Dr.'s R. J. Carroll, D. H. Glueck, and an anonymous reviewer. Muller's work supported in part by NCI program project grant P01 CA47 982-04. Coffey's work supported in part by NIEHS grant 5-T32-ES07018.

REFERENCES

Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, **13**, 2455-2463.

Coffey, C. S. and Muller, K. E. (1999). Exact test size and power of a gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199-1214.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29**, 323-333.

Denne, J. S. and Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* **18**, 1575-1585.

Gould, A. L. and Shih, W. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods* **21**, 2833-2853.

Helms, R. W. (1988). Comparisons of parameter and hypothesis definitions in a general linear model. *Communications in Statistics - Theory and Methods* **17**, 2725-2753.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC.

Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901-911.

Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K. E., Brauening, M. P., and Pizer, S. M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging* **11**, 193-200.

SAS Institute (1989). *SAS/IML® Software: Usage and Reference, Version 6.* Cary, NC: SAS Institute.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243-258.

Taylor, D. J. and Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician* **49**, 43-47.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., and Proschan, M. (1999). Internal pilot studies I: type I error rate of the naive t-test. *Statistics in Medicine*, **18**, 3481-3491.

Zucker, D. M., Wittes, J. T., Schabenberger, O., and Brittain, E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*, **18**, 3493-3509.

APPENDIX

Model Description and Notation

Write a general linear univariate model with Gaussian errors and fixed predictors as

$$\begin{bmatrix} \underset{n_1 \times 1}{\boldsymbol{y}_1} \\ \underset{N_2 \times 1}{\boldsymbol{y}_2} \end{bmatrix} = \begin{bmatrix} \underset{n_1 \times q}{\boldsymbol{X}_1} \\ \underset{N_2 \times q}{\boldsymbol{X}_2} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \underset{n_1 \times 1}{\boldsymbol{e}_1} \\ \underset{N_2 \times 1}{\boldsymbol{e}_2} \end{bmatrix},$$

with partitioning corresponding to the internal pilot and second samples. Here N_2 and $N_+ = n_1 + N_2$ are random. Let $\text{Es}(\boldsymbol{X})$ represent the matrix created by deleting duplicate rows from a design matrix (Helms, 1988). We restrict $\text{Es}(\boldsymbol{X}_1) = \text{Es}(\boldsymbol{X}_2) = \text{Es}(\boldsymbol{X}_+)$, with $\boldsymbol{X}_+$ the vertical concatenation of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We wish to test H_0: $\boldsymbol{\theta} = \boldsymbol{C\beta} = \boldsymbol{\theta}_0$. Without loss of generality assume $\boldsymbol{\theta}_0 = \boldsymbol{0}$ (Coffey and Muller, 1999).

For the unadjusted test, compute $F(n_+)$, the fixed sample F statistic and reject H_0 if $F(n_+) > f(\alpha_t, \nu_+)$. Coffey and Muller (1999) proved that $SSE(n_+)$ can be written as the sum of two independent, quadratic forms, SSE_1 and $SSE_*(n_+)$.

Proof that Actual Test Size Using the Unadjusted Method is Never Less Than Target

As $\gamma \to 0$, $\Pr\{N_+ = n_{+,\text{min}}\} \to 1$, the dependence of N_+ on $\widehat{\sigma}_1^2$ disappears, and there will be no test size inflation. As $\gamma \to \infty$ with a finite maximum sample size, $\Pr\{N_+ = n_{+,\text{max}}\} \to 1$, and the dependence of N_+ on $\widehat{\sigma}_1^2$ disappears. As $\gamma \to \infty$ without a finite maximum sample size, the probability mass at $n_+ = \infty$ goes to 1, which implies $\widehat{\sigma}^2(N_+) \to \sigma^2$. Furthermore, in the two sample setting, Wittes *et al.* (1999) showed that $\widehat{\sigma}^2(N_+)$ is biased downward. Their reasoning is easily extended to the general linear model. There is a direct relationship between the downward bias in $\widehat{\sigma}^2(N_+)$ and the potential test size inflation.

TABLE 1

Internal Pilot Study Notation for Testing $H_0 : \boldsymbol{\theta} = \boldsymbol{C\beta} = \boldsymbol{\theta}_0$

Symbol	Definition
Notation	
a	rank($\boldsymbol{C}$)
ν_i	n_i − rank[Es($\boldsymbol{X}_+$)]
Design Parameters	
Es($\boldsymbol{X}_+$)	Essence Design Matrix
m	Observations taken per replication
α_t	Target test size
P_t	Target Power
$\boldsymbol{\theta}_*$	'Scientifically Important' value of $\boldsymbol{\theta}$
σ_0^2	Variance estimate used for planning
n_0	Pre-planned sample size based on $(\alpha_t, P_t, \boldsymbol{\theta}_*, \sigma_0^2)$
Sample Size Allocation	
π	Proportion of N_0 used in internal pilot
$n_1 = \pi n_0$	Internal pilot sample size
$n_{+,\min}$	Minimum size of *final* sample
$n_{+,\max}$	Maximum size of *final* sample
Fixed, Unknown Parameters	
σ^2	True variance
$\gamma = \sigma^2/\sigma_0^2$	Ratio of true to estimated variance
$\boldsymbol{\theta}$	$a \times 1$, True value of secondary parameter
Random Variables	
$\widehat{\sigma}_1^2$	Internal pilot variance estimate
N_2	Second sample size
$N_+ = n_1 + N_2$	Final sample size
$\widehat{\boldsymbol{\theta}}(N_+)$	Final estimate of secondary parameter

TABLE 2

*Expected Sample Size and Power*100, CLAHE Example,* $n_{+,\min} = n_1$, $n_{+,\max} = 30$
Compare to $n_0 = 20$ *fixed power of >99,93, and 55 for* $\gamma \in \{0.5,1,2\}$, *respectively*

		Rule 0					Rule 1					Rule 2				
		$\mathcal{E}N_+$	Power Test				$\mathcal{E}N_+$	Power Test				$\mathcal{E}N_+$	Power Test			
π	γ		*0*	1	2	*3*		0	*1*	2	3		0	1	*2*	3
0.25	0.5	12	***91***	41	51	***84***	23	99	***90***	86	97	15	98	89	***80***	97
	1.0	18	***84***	21	64	***77***	28	98	***57***	93	97	20	91	63	***76***	89
	2.0	25	***73***	11	65	***64***	29	85	***25***	82	78	26	76	33	***68***	72
0.50	0.5	13	***97***	93	24	***94***	15	98	***98***	41	97	18	>99	>99	***86***	>99
	1.0	19	***91***	77	56	***87***	24	97	***90***	77	95	23	97	92	***82***	97
	2.0	26	***79***	49	66	***73***	29	85	***55***	77	79	28	82	62	***73***	81
0.75	0.5	16	***>99***	>99	02	***99***	16	>99	***>99***	03	99	22	>99	>99	***90***	>99
	1.0	20	***95***	93	29	***94***	21	96	***95***	39	95	26	99	99	***86***	99
	2.0	27	***83***	72	59	***80***	28	85	***74***	65	82	29	86	78	***71***	85

The "logical" combinations of test and rule are in bold italics.

TABLE 3

Results of CLAHE Example with $n_1 = 10$

		Fixed ($n = 20$)		Method 0/3 ($n_+ = 11$)	
Clip	Region	Mean	F	Mean	$F(n_+)$
2	2	-0.002	0.1	-0.001	0.0
2	8	-0.007	0.4	-0.007	0.2
2	32	0.061	51.2*	0.062	24.3
4	2	-0.019	3.6	-0.025	2.3
4	8	0.008	0.5	-0.006	0.1
4	32	0.053	28.4*	0.053	9.7
16	2	-0.039	18.8*	-0.041	11.9
16	8	-0.036	7.7	-0.042	6.6
16	32	-0.031	5.0	-0.049	4.9

* indicates statistical significance.
Critical values were 14.8 (Fixed) and 24.3 (Method 0/3)

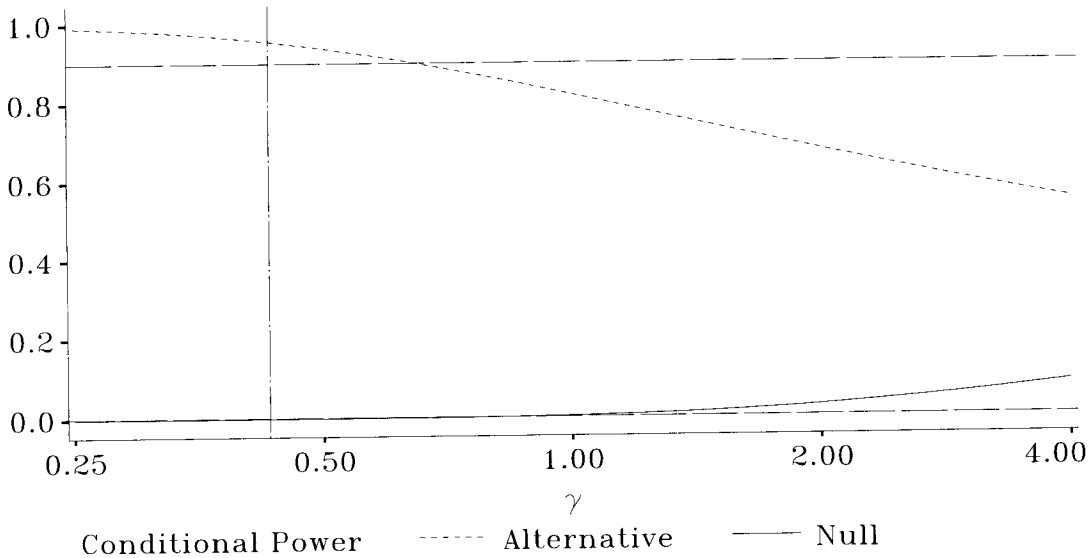

Figure 1. Conditional test size and Power for the CLAHE example with Method 0/3, $n_1 = 10$, and $n_+ = 11$. The horizontal reference line represents $\widehat{\gamma}_1$, while the vertical reference lines represent the target test size (0.0011) and target power (0.9).