TECHNICAL REPORT #57
April 2012

Internal Pilot Studies: An Annotated Bibliography

Aniko Szabo, Tao Wang, Peng He, and Sergey Tarima
Division of Biostatistics, Medical College of Wisconsin

1. Introduction

The appropriate sample size for a study depends on many parameters. In most cases, in addition to the desired type I error rate and the size of the treatment effect, other parameters not specified in the null or alternative hypotheses affect the power. Examples of such parameters include the variance of a normally distributed outcome for the power of the t-test, or the baseline probability of success for the power of the z-test of proportions. These nuisance parameters have to be specified before the sample size calculation can proceed. This is usually done through literature review, a pilot study, or just ad-hoc.

The main idea of *internal pilot* studies is to estimate these nuisance parameters during the study in a way that the data collected for this purpose can still be used in the final efficacy analysis. A typical internal pilot study has two stages: in the first, pilot, stage n_0 observations are collected. At the end of this internal pilot, the nuisance parameters are estimated, and power calculations are performed to determine the overall sample size for the study. At the end of the study, either usual inference is applied ("naive" approach), or some sort of adjustment is made to the test statistic, critical value, or significance level.

Internal pilot studies are a special case of the much more general "adaptive", or "flexible" designs, they also resemble group-sequential designs. They are set apart by their tight restrictions of the allowed adaptations - only the sample size of the study is modified, deliberate ignorance of the treatment effect attained at the end of the pilot, and the lack of intent of stopping after the first stage. In fact, many internal pilot designs conduct "blinded" sample size re-estimation, in which the group assignment of the observations is unknown, and so current effect size cannot be estimated. A study with an internal pilot can stop immediately after the pilot stage if the power calculation indicates that sufficient power has been achieved for the targeted effect size, however there is no direct connection to the outcome of the significance test.

2. Normally distributed outcome: one- or two-sample superiority test

Most of the theoretical development concentrates on designing a two-arm randomized clinical trial with normally distributed outcomes. In this case the unknown common variance is the nuisance parameter.

2.1 Unblinded sample size re-estimation

The most natural estimate of the variance is the pooled variance that is the weighted average of within-group variances. This calculation requires knowing which observations belong to the same group, though not the identity of the groups. In the context of internal pilots such a setting is considered "unblinded".

Stein, C. (1945) A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16 (3), 243-258.
Proposes an internal pilot design with the goal of developing a 2-sample test the power of which is independent of the variance, or equivalently, which gives a confidence interval of prespecified width. The variance estimate (and degrees of freedom) used at the end of the study are based on the first stage data only. Extension to a linear hypothesis (eg ANOVA) are developed.
TAGS: F-test, continuous, fixed effect size, methodology, nuisance: SD, t-test; unadjusted alpha

Wittes, J.T., Brittain, E. (1990) The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9 (1-2), 65-71; discussion 71-2.
Introduces the idea of an internal pilot study. Shows example of recalculation of variance in the middle of a trial if the preliminary estimate is lower than the observed variability. Argues that the effect on type I error rate is minimal. Authors suggest to recalculate variances (and, possibly, event rates) at the interm analysis and recalculate the total sample size. Two group comparison is considered. The treatment effect (difference between group means) stays unchanged at the interim analysis. The procedure is based on calculating the total sample size n_0 and using the (p n_0) observations as the internal pilot, $0 < p < 1$. Their procedure is similar to Stein's sample size recalculation, but the new recalculated sample size (n_1) can only be higher than n_0. In their simulations, the inflation of the type I error is minimal.
TAGS: continuous, methodology, nuisance: SD, t-test, unadjusted alpha

Birkett, M.A., Day, S.J. (1994) Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13 (23-24), 2455-2463.
The authors use simulations to show that an internal pilot study with about 20 degrees of freedom for estimating the variance provides type I error and power characteristics similar to larger pilot studies. They also argue against the restriction of no decrease below the originally planned sample size proposed by [WB90].
TAGS: methodology, continuous, nuisance: SD, t-test, unadjusted alpha

Proschan, M.A., Hunsberger, S.A. (1995) Designed extension of studies based on conditional power. *Biometrics*, 51 (4), 1315-24.
Proposes a sample size reestimation method based on conditional power. The type 1 error is controlled with a conditional error function. Authors focus on two group comparisons with a z-test and use the "circular" conditional error function through the manuscript. At the interim analysis, the study may be stopped due to futility if the observed test statistic is below some lower bound (for example, the lower bound associated with P=15%) or stopped due to high efficacy (the interim test statistic is above some upper bound on critical values, for example, with P=3.4%). Otherwise, the study continues with a reestimated sample size to reach a predefined conditional power (for example, 50%).

TAGS: adjusted alpha, application, conditional power, continuous, methodology, nuisance: SD, reestimated effect size, z-test

Betensky, R., Tierney, C. (1997) An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine*, 16 (22), 2587-98.
The paper considers a one-sample t-test, and concentrates on finding the correct sample size and not on inference. The sequential method of Anscombe (1953) is adapted to an internal pilot setup. After the pilot data is gathered, bootstrap resampling is used to generate potential realizations of future data, and the fully sequential approach is applied to find the sample size at stopping. The average of all these potential stopping sample sizes is used as the targeted sample size. Simulations show that this procedure has better MSE for estimating the correct sample size than Stein's procedure.
TAGS: continuous, methodology, t-test, unadjusted alpha, nuisance: SD

Denne, J.S., Jennison, C. (1999) Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine*, 18 (13), 1575-85.
Considers an internal pilot for a two-sample t-test (paired or unpaired) using an unmodified t-statistic at the end of the study with modified degrees of freedom. For the degrees of freedom calculations observations from the second stage have a weight 1/8 determined based on simulation studies). This method is shown to provide good control of both type I and II errors even for values of n_0 as low as 5.
TAGS: adjusted alpha, continuous, methodology, nuisance: SD, t-test

Coffey, C.S., Muller, K.E. (1999) Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine*, 18 (10), 1199-214.
Internal pilot studies for a linear hypothesis in a regression framework are considered, and exact power function is derived. The formulas apply if all the regression covariates are set by design. Both increase and decrease compared to the planned sample size are considered. An inflation of type I error is seen especially for small sample size for the internal pilot, and when reduction in the planned sample size is allowed. SAS/IML code is available at
`http://www.soph.uab.edu/coffey/`
TAGS: F-test,SAS, continuous, methodology, nuisance: SD, software, unadjusted alpha, GLUM

Proschan, M.A., Wittes, J.T. (2000) An Improved Double Sampling Procedure Based on the Variance. *Biometrics*, 56 (4), 1183-1187.
Stein's (1945) [Ste45] procedure is improved by modification of the variance estimate (and df) in the final test statistic to use not just data from the first stage, but up to the originally planned sample size.
TAGS: continuous, methodology, nuisance: SD, t-test, unadjusted alpha

Kieser, M., 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 (7), 901-11.
The authors derive an upper limit on the type I error inflation, and an iterative method for finding the correct adjusted significance level in the context of an internal pilot with a two-sample t-test. The ideas is based on 1) replacing the variance in the final t-test by a pooled variance estimate from the two stages; 2) obtaining an explicit formula for type I error (requires numerical integration); 3) showing that this provides an upper limit for the type I error inflation for the t-

test. The authors also suggest using an upper one-sided confidence limit of the interim variance estimate for the sample size update to increase the chances of achieving adequate power.
TAGS: adjusted alpha, continuous, methodology, t-test, nuisance: SD

Coffey, C.S., Muller, K.E. (2001) Controlling test size while gaining the benefits of an internal pilot design. *Biometrics*, 57 (2), 625-31.
Under the GLUM framework, the F-test statistics were applied in three cases: 1) use pilot data plus additional guaranteed observations to estimate the SD; 2) mainly use the second sample and ignore the pilot data to estimate SD; 3) bound test size $\alpha^* \leq \alpha$.
TAGS: F-test, conditional power, continuous, GLUM, methodology, nuisance: SD, adjusted alpha

Li, G., Shih, W.J., Xie, T., Lu, J. (2002) A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics (Oxford, England)*, 3 (2), 277-87.
This paper revised Proschan and Hunsberger's (1995) [PH95] conditional power method. They treat the critical value c as a constant which does not depend on the intermediate statistic z1. As a result, no need to choose the conditional error function. But still, as the PH method, they assume that the variance estimate from the pilot data is accurate.
TAGS: conditional power; continuous; methodology; reestimated effect size; z-test; nuisance: SD; adjusted alpha

Miller, F. (2005) Variance estimation in clinical studies with interim sample size re-estimation. *Biometrics*, 61 (2), 355-61.
The author considers the problem of end-of-study variance estimation after an internal pilot and shows that the usual estimator is negatively biased. Tight theoretical bounds are derived for this bias, and an "almost unbiased" estimator is proposed.
TAGS: methodology; continuous; nuisance: SD; unadjusted alpha; t-test; fixed effect size

2.2 Blinded sample size re-estimation

In the context of randomized clinical trials any breaking of the blinding is considered undesirable. The following papers describe internal pilots where no information about the treatment group membership is available for sample size re-estimation after the interim pilot stage.

Gould, A.L., Shih, W.J. (1992) Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods*, 21 (10), 2833-2853.
The paper considers blinded sample size reestimation in the context of a two-arm trial with normally distributed outcomes. An exact expression for the effect of an internal pilot on the type I error is derived, and some specific values are shown to be close to the nominal error rate. Two blinded variance estimates are proposed: a "simple" estimate that adjust the overall variance using the mean difference under the alternative and an EM estimation procedure based on mixture of two normal distributions. The authors argue that the EM-based estimate is better, because its performance does not depend on the alternative hypothesis.
TAGS: methodology; continuous; t-test; nuisance: SD; unadjusted alpha; blinded

Kieser, M., Friede, T. (2003) Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, 22 (23), 3571-81.
The authors derive explicit formulas (requiring numeric integration) for power analysis in case of blinded sample size reestimation. They show that the type I error is essentially maintain with either restricted or unrestricted designs, with slightly better power characteristics for the blinded estimate.
TAGS: blinded; continuous; review; t-test; permutation test; fixed effect size; unadjusted alpha

Xing, B., Ganju, J. (2005) A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, 24 (12), 1807-14.
The authors propose a procedure for blinded variance estimation in a randomized clinical trial that uses block randomization. They show that the variance of the within-block sums scaled by the block-size equals to the within-group variance, and thus can be used for the sample size recalculation. The relative efficiency of this estimator compared to the unblinded estimator is proportional to the square-root of the block size, so lower block sizes are preferable.
TAGS: methodology; blinded; continuous; nuisance: SD; t-test; unadjusted alpha; fixed effect size

Shih, W.J. (2009) Two-Stage Sample Size Reassessment Using Perturbed Unblinding. *Statistics in Biopharmaceutical Research*, 1 (1), 74-80.
Proposes a method to partially unblind a trial so that within-group variance can be estimated. This involves adding fixed but hidden constants a and b to each of the treatment groups. Uses an adjusted estimate of variance in the final test per Miller (2005) [Mil05].
TAGS: blinded, continuous, methodology, nuisance: SD, t-test, unadjusted alpha

Ganju, J., Xing, B. (2009) Re-estimating the sample size of an on-going blinded trial based on the method of randomization block sums. *Statistics in Medicine*, 28 (1), 24-38.
This paper extends in several ways the randomized block-sum approach of Xing and Ganju, 2005 [XG05] for blinded estimation of variance after the internal pilot phase. The major extension is the incorporation of covariate adjustment: the within-block sums of the outcome are regressed on the within-block sums of the covariates to obtain a blinded estimate of the residual variance. The authors recommend a block size of 2, and propose a method to provide a blinded block size of 2 for the statistician while maintaining a higher block size for the other aspects of the trial.
TAGS: blinded; continuous; methodology; nuisance: SD; regression; unadjusted alpha; fixed effect size

3. Binomial outcome: one- or two-sample superiority test

The power of a two-sample test for a binary outcome depends on the probabilities of success in both groups. The nuisance parameter is usually specified as either the probability of success in the control group, or the overall probability of success in the study.

Gould, A.L. (1992) Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine*, 11 (1), 55-66.

The paper considers an interim pilot study for a two-group trial with a binary outcome. The average response rate is treated as the nuisance parameter, the hypothesis of interest could be defined as a difference, ratio, or odds ratio of the response probabilities. At the interim examination the overall response rate is estimated in a blinded fashion. Simulation studies show no substantial increase of type I error rate over a study without an interim recalculation.
TAGS: binary, blinded, methodology, nuisance: other, unadjusted alpha, chi-square test

Herson, J., Wittes, J.T. (1993) The use of interim analysis for sample size adjustment. *Drug Information Journal*, 27, 753-760.
The authors present an example of a two-arm clinical trial comparing the rate of a binary outcome. The effect size is given as a relative risk. The response rate in the placebo group is treated as the nuisance parameter. Simulation studies are used to examine the effect of various design parameters such as the size of the internal pilot as a proportion of the originally planned sample size, the minimal number of additional samples, and the maximal sample size. The type I error is found to be well controlled in most situations.
TAGS: application; binary; nuisance: other; unadjusted alpha; z-test

Shih, W.J., Zhao, P.L. (1997) Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine*, 16 (17), 1913-23.
Proposes sample-size reestimation at the half-way point of a trial with two groups with a binary outcome without unblinding individual assignments. Main idea is using two dummy strata with opposite allocation probabilities, so that success rates can be estimated as weighted sums of the stratum-specific rates. Both increase and decreased sample size is allowed, but the extent of change is limited.
TAGS: binary, blinded, methodology, reestimated effect size, unadjusted alpha, z-test, nuisance: other

Friede, T., Kieser, M. (2004) Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, 3 (4), 269-279.
The paper follows up on the work of Gould (1992) [Gou92] in proposing blinded sample size reestimation for two-sample trials with a binary endpoint. An unbalanced but known allocation ratio is allowed. The power and type I error are evaluated with exact binomial computations, and control comparable to that of a fixed sample size chi-square test is demonstrated.
TAGS: binary; blinded; chi-square test; methodology; nuisance: other; unadjusted alpha; fixed effect size

Friede, T., Mitchell, C., Müller-Velten, G. (2007) Blinded sample size reestimation in non-inferiority trials with binary endpoints. *Biometrical journal. Biometrische Zeitschrift*, 49 (6), 903-16.
The authors derive exact expression for the power function of a non-inferiority trial with binary endpoint after blinded interim-pilot based sample size reassessment. They use numeric examples to demonstrate that the power is stabilized and the type I error is essentially maintained.
TAGS: methodology; binary; z-test; blinded; unadjusted alpha; nuisance: other; non-inferiority; fixed effect size

4. Generalizations, other distributions, nuisance parameters

Moving beyond simple randomized two-arm studies with normal or binary outcomes introduces a wide variety of nuisance parameters. These are often even harder to prespecify than the variance or baseline probability of success, because they can be very trial-specific and are rarely published even if similar studies have already been conducted.

4.1 Multiple stages

Gould, A.L., Shih, W.J. (1998) Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*, 17, 89-100.
Considers blinded sample size re-estimation in the context of group-sequential trials. The re-estimation occurs before the first interim analysis. Several strategies for allocating additional samples among group-sequential stages are compared. For continuous outcomes variance is estimated by fitting a normal mixture to the blinded data; for binary outcomes the update is based on the observed pooled event rate and the odds ratio specified in Ha (per Gould, 1992, [Gou92]).
TAGS: methodology; binary; blinded; continuous; multi-stage; unadjusted alpha; t-test; chi-square test; fixed effect size; nuisance: SD; nuisance: other

Denne, J.S., Jennison, C. (2000) A group sequential t-test with updating of sample size. *Biometrika*, 87 (1), 125-134.
The paper describes a generalization of Stein's two-stage interim analysis procedure (Stein, 1945 [Ste45]) to a group sequential design with multiple stages. The procedure is then modified to allow data beyond the first stage to contribute to the variance estimation by adjusting the degrees of freedom of the subsequent test statistics by giving lower weight (1/4) to later samples as in Denne and Jennison(1999) [DJ99].
TAGS: continuous; methodology; nuisance: SD; t-test; unadjusted alpha; multi-stage

4.2 Other alternative hypothesis

Friede, T., Kieser, M. (2003) Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine*, 22 (6), 995-1007.
The paper discusses blinded interim sample size recalculation in non-inferiority trials. They show that the type I error inflation is larger than in superiority trials.
TAGS: application, blinded, continuous, methodology, non-inferiority, nuisance: SD, t-test, unadjusted alpha; fixed effect size

4.3 Other outcome distributions / nuisance parameters

Lake, S., Kammann, E., Klar, N., Betensky, R. (2002) Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, 21 (10), 1337-50.
The "naive" internal pilot approach is extended to cluster randomized trials. In addition to the variance, the sample size of these trials is affected by other nuisance parameters: the within-cluster correlation, the average cluster size, and the variance of cluster size. The authors propose (re)estimating these parameters after a relatively large internal pilot. Simulation are used to show that the type I error inflation is not severe, and more appropriate power can be realized.
TAGS: methodology, continuous, nuisance: other, unadjusted alpha; Wald test

Coffey, C.S., Muller, K.E. (2003) Properties of internal pilots with the univariate approach to repeated measures. *Statistics in Medicine*, 22 (15), 2469-85.
The paper examines the use of internal pilots with repeated measures outcomes. They find that type I error inflation is more severe than in the univariate case. However since the commonly used Muller-Barton power formula is found to be conservative (gives more than the designed power), these effects can offset each other.
TAGS: F-test; continuous; methodology; nuisance: SD; nuisance: other; repeated measures; unadjusted alpha

Wüst, K., Kieser, M. (2003) Blinded Sample Size Recalculation for Normally Distributed Outcomes Using Long- and Short-term Data. *Biometrical Journal*, 45 (8), 915-930.
The paper proposes sample size recalculation from an interim pilot study by augmenting the interim variance estimate based on the outcome of interest with a short-term outcome that is correlated with the actual outcome and would be available for more subjects. An example would be using a 2-week change score available for 50 subjects in addition to the main outcome of an 8-week change score available only for 25 subjects (the size of the interim trial).
TAGS: methodology; blinded; fixed effect size; continuous; nuisance: SD; unadjusted alpha; t-test

Cook, R.J., Bergeron, P.J., Boher, J.M., Liu, Y. (2009) Two-stage design of clinical trials involving recurrent events. *Statistics in Medicine*, 28 (21), 2617-2638.
Interim pilot designs for two-sample prospective studies with counts of recurrent events as primary outcome are considered. The counts are modeled by a negative binomial distribution, and the nuisance parameters are the event rate in the control group and the overdispersion parameter. Extensions to serial patient entry are considered. Three methods for interim sample size reestimation are considered: unblinded, blinded, and partially blinded; in the last setting group membership is known, but the actual treatments are not known. An EM-algorithm based estimation is used for the two blinded settings. The partially blinded models suffer from identifiability issues with respect to the baseline event rate but not the overdispersion parameter. The authors claim good type I error control, however this is not supported by the simulation studies which show substantial type I error inflation for the blinded approaches when the baseline event rate was originally underestimated. They also have a tendency to overpower studies when the overdispersion is underestimated.
TAGS: methodology; count; Wald test; blinded; nuisance: other; unadjusted alpha; fixed effect size

Friede, T., Schmidli, H. (2010) Blinded Sample Size Reestimation with Negative Binomial Counts in Superiority and Non-inferiority Trials. *Methods of Information in Medicine*, 49 (6), 618-624.
The paper considers internal pilots for prospective trials with outcomes that are counts of recurrent events modeled using negative binomial distribution. They provide a blinded sample size estimation formula that depends on the average event rate and the common overdispersion parameter. The main difference from [FS10a] is the explicit use of the negative binomial likelihood, and extension to non-inferiority settings. Simulations are used to show that the "naive" approach with reestimation at half of the originally planned sample size controls the type I error rate well, and provides a robust design at a price of 5 extra subjects per group on average.

TAGS: count; methodology; nuisance: other; quasi-likelihood test; unadjusted alpha; non-inferiority

Friede, T., Schmidli, H. (2010) Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statistics in Medicine*, 29 (10), 1145-1156.
Blinded interim sample size recalculation is considered for two-sample studies with Poisson or overdispersed Poisson outcomes. Sample size formulas that depend only on the response rate in the pooled data(and the hypothesized effect size) are derived, and this nuisance parameter is estimated at the end of the interim pilot in a blinded fashio. If needed, an overdispersion parameter is estimated from the variance of the pooled data using the method of moments. Simulation studies show good control of type I error rates and power.
TAGS: count; methodology; nuisance: other; unadjusted alpha; quasi-likelihood test

Gurka, M.J., Coffey, C.S., Gurka, K.K. (2010) Internal pilots for observational studies. *Biometrical journal. Biometrische Zeitschrift*, 52 (5), 590-603.
The paper describes the use of interim pilot studies for two-group comparison when the group proportions are not known (eg, they are defined by genetic testing). They show through simulations that interim reestimation of the group proportion can maintain designed power with at most slight effect on type I error rate.
TAGS: methodology, binary, continuous, nuisance: other, t-test, unadjusted alpha, z-test

Jensen, K., Kieser, M. (2010) Blinded sample size recalculation in multicentre trials with normally distributed outcome. *Biometrical journal. Biometrische Zeitschrift*, 52 (3), 377-99.
The paper describes the use of an internal pilot study for a multicenter trial. Both weighted (random effect model) and unweighted analyses are considered. In the latter case in addition to the standard deviation, the variability of the center sample sizes is also a nuisance parameter. Simulation studies are used to show that combining blinded within-sample variance estimates provides the best power characteristics while maintaining the type I error rate.
TAGS: F-test; blinded; methodology; continuous; nuisance: SD; nuisance: other; unadjusted alpha

Friede, T., Kieser, M. (2011) Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical statistics*, 10 (1), 8-13.
Considers interim pilot studies for ANCOVA-based randomized trial which adjust for one normally distributed "random" covariate. The sample size reestimation is based on estimating the nuisance parameter $(1-r^2)s^2$ under the null hypothesis of no treatment effect. This Gould-Shih (1998) [GS98]-type approach controls the type I error rate without adjustment (even the conditional type I error).
TAGS: blinded, continuous, methodology, unadjusted alpha, regression; nuisance: SD; nuisance: other

5. Reviews

5.1 Reviews

Wittes, J.T., Schabenberger, O., Zucker, D.M., Brittain, E., Proschan, M.A. (1999) Internal pilot studies I: type I error rate of the naive t-test. *Statistics in Medicine*, 18 (24), 3481-91.
The paper presents an extensive simulation study of the effect of an internal pilot on the type I error with a t-test. They find that the restricted design, in which decrease of the planned sample size is not allowed, induces very little type I error inflation; if very strict control is desired a final significance level of 0.047 could be used. The unrestricted design can lead to substantial type I error inflation especially if the internal pilot is done early (eg at 1/4 of the planned sample size). In this case a significance level as low as 0.038 might be needed.
TAGS: methodology, continuous, t-test, unadjusted alpha, nuisance: SD; fixed effect size

Zucker, D.M., Wittes, J.T., Schabenberger, O., Brittain, E. (1999) Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*, 18 (24), 3493-509.
The paper reviews and compares several ways of constructing the test statistic at the end of the study with an interim trial: using only first stage variance (Stein, 1945 [Ste45]), no adjustment (Wittes and Brittain, 1990 [WB90]), using only the second stage variance, and blinded interim variance estimation Gould, 1992 [Gou92]. In simulation studies, Stein's procedure performs well, though the Wittes-Brittain procedure has better power at effect sizes lower than the designed effect. They also note that only the Gould-Shih and second-stage variance procedures control the conditional type I error rate.
TAGS: continuous; nuisance: SD; review; t-test

Gould, A.L. (2001) Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine*, 20 (17-18), 2625-43.
The paper reviews blinded and unblinded sample size reestimation procedures that do not consider the effect size, as well as conditional power based approaches that do depend on the observed effect size. Both normal and binary data are considered. They conclude that type I error inflation is not substantial if the pilot sample size is above 10 (per group), and that blinded estimation is feasible and has comparable characteristics to unblinded methods. The authors suggest that the conditional power based approach might be converted to a group-sequential setup instead.
TAGS: binary, blinded, conditional power, continuous, review, unblended

Bauer, P., Brannath, W. (2004) The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today*, 9 (8), 351-7.
Discusses adaptive clinical trials in general, and mentions internal-pilot type recalculations as a special case. Also notes that conditional power based recalculations should keep the original treatment effect estimate.
TAGS: review

Lancaster, G.A., Dodd, S., Williamson, P.R. (2004) Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10 (2), 307-12.
This paper discusses pilot studies in general, with most attention to external pilot studies. The authors note that internal pilot studies could have advantages, but do not allow testing feasibility, and have to be part of the complete protocol.
TAGS: review

Proschan, M.A. (2005) Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics*, 15 (4), 559-74.
This paper provides a review with examples of internal pilot studies. Continuous and binary outcomes, as well as blinded and unblinded estimators are considered.
TAGS: review

Friede, T., Kieser, M. (2006) Sample Size Recalculation in Internal Pilot Study Designs: A Review. *Biometrical Journal*, 48 (4), 537-555.
This is a fairly comprehensive review of internal pilot studies. It focuses mostly on the two-sample normal comparison, but touches on binary outcomes as well. Blinded versus un-blinded reestimation, as well as non-inferiority/equivalence testing are discussed.
TAGS: binary, continuous, review

5.2 Regulatory considerations

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J. (2006) Adaptive Designs in Clinical Drug Development: An Executive Summary of the PhRMA Working Group. *Journal of Biopharmaceutical Statistics*, 16 (3), 275-283.
Several types of adaptive designs, including internal pilot studies are discussed. The need for pre-specifying sample size reestimation procedures, ensuring proper control of type I error rates, blinding study sponsors, and addressing logistical issues are stressed.
TAGS: regulatory; blinded; reestimated effect size

Hung, H.M.J., O'Neill, R.T., Wang, S.J., Lawrence, J. (2006) A regulatory view on adaptive/flexible clinical trial design. *Biometrical journal. Biometrische Zeitschrift*, 48 (4), 565-73.
The paper discusses potential problems with the use of adaptive trials, especially in phase III settings. They argue that key parameters such as (unstandardized) effect size or equivalence margin should not be modified. They warn of problems with analysis of secondary outcomes when a trial/arm is stopped/modified based on a primary outcome.
TAGS: regulatory

Coffey, C.S., Kairalla, J.A. (2008) Adaptive clinical trials: progress and challenges. *Drugs in R&D*, 9 (4), 229-42.
Authors provide a nontechnical review of current literature on adaptive designs. They discuss definitions, challenges, controversies and specifically focus on contrasting sample size reestimating procedures based on the single interim analysis with and without treatment effect reestimation. In their opinion, reestimation of treatment effect is a controversial measure often leading to less efficient designs as compared to group sequential designs with sample size reestimation. Authors often rely on executive summary of Pharmaceutical Researches and Manufacturers of America (PhRMA) working group.
TAGS: adjusted alpha, fixed effect size, nuisance: SD, reestimated effect size, review

6. Software

While simple internal pilot designs can be implemented with the use of standard statistical software, more advanced methods require extensive numerical computations. Here we summarize papers describing free publicly available software for internal pilot studies.

Zellner, D., Zellner, G.E., Keller, F. (2001) A SAS macro for sample size re-estimation. *Computer Methods and Programs in Biomedicine*, 65 (3), 183-90.
The paper describes the implementation of the EM-algorithm based blinded sample size reestimation procedure proposed by Gould and Shih (1992) [GS92]. The procedure is implemented as a SAS macro SSEM that is provided as an appendix to the paper.
TAGS: blinded; continuous; software; t-test; unadjusted alpha; nuisance: SD; fixed effect size

Coffey, C.S., Kairalla, J.A., Muller, K.E. (2007) Practical Methods for Bounding Type I Error Rate with an Internal Pilot Design. *Communications in Statistics- Theory and Methods*, 36 (11), 2143-2157.
The "alpha-bounding" test for internal pilot studies proposed by Coffey and Muller (2001) [CM01] was shown to good operating characteristics in terms of controlling type I error rate while not sacrificing power. However its practical implementation suffered from numerical instability and common lack of convergence. This paper develops a computational approach that allows a faster and more stable computation of the bound for the type I error rate. The method is implemented as a SAS macro GLUMIP 2.0.
TAGS: F-test, SAS, continuous, fixed effect size, GLUM, methodology, nuisance: SD, software, adjusted alpha, marginal power

Muller, K.E., Coffey, C.S., Kairalla, J.A. (2008) GLUMIP 2.0: SAS/IML Software for Planning Internal Pilots. *Journal of Statistical Software*, 28 (7).
A comprehensive implementation as SAS/IML modules of multiple methods for internal pilot studies continuous outcomes is described. The framework of linear models with fixed predictors includes t-tests and ANOVAs as special cases. Three sample-size reestimation rules (unadjusted [WB90], Stein's [Ste45], and second sample [ZWSB99]), and five testing approaches are implemented. The testing methods include the unadjusted `naive approach', three test-statistic modifications that control the type I error rate ([Ste45], [ZWSB99], [PW00]), and an alpha-bounding approach with the original test statistic ([CM99]).
TAGS: F-test; GLUM; PH; SAS; adjusted alpha; continuous; software; t-test; unadjusted alpha; nuisance: SD

Wang, S., Xia, J., Yu, L., Li, C., Xu, L. (2008) A SAS macro for sample size adjustment and randomization test for internal pilot study. *Computer Methods and Programs in Biomedicine*, 90 (1), 66-88.
A SAS macro implementing blinded sample size reestimation for a two-sample t-test setup is described. A permutation test is used for inference, which preserves the type I error rate both unconditionally and conditional on the results of the first stage.
TAGS: SAS; blinded; continuous; nuisance: SD; permutation test; software; unadjusted alpha

Wang, S., Xia, J., Yu, L., Li, C., Xu, L., Zheng, L. (2009) Realization of simulations for blinded internal pilot study based on web. *Journal of Biomedical Informatics*, 42 (2), 262-71.

The paper describes a browser/server mode software conducting internal pilot studies. Both blinded and non-blinded variance estimation are implemented for sample size recalculation. A randomization test is used for inference. Some code is available in the appendix, but the overall software does not seem to be available.
TAGS: blinded; continuous; nuisance: SD; permutation test; software; unadjusted alpha; unblinded; web

7. Application

The following papers describe actual studies that used an internal pilot design.

Bolland, K., Sooriyarachchi, M.R., Whitehead, J. (1998) Sample size review in a head injury trial with ordered categorical responses. *Statistics in Medicine*, 17 (24), 2835-2847.
The paper describes the design of a randomized trial with a 3-level ordinal outcome. The baseline probability of each outcome within strata defined by prognostic factors affects the power of the study in addition to the odds ratio in a proportional odds model (the quantity of interest), and thus are nuisance parameters. The paper describes how an interim pilot study was planned and executed.
TAGS: application, nuisance: other, ordinal, unadjusted alpha, Wald test

Friede, T., Stammer, H. (2010) Blinded Sample Size Recalculation in Noninferiority Trials: A Case Study in Dermatology. *Drug Information Journal*, 44 (5), 599-607.
The authors present a case study of an internal pilot design for a non-inferiority test with normally distributed data. The sample size recalculation uses the pooled variance, and thus blinded. Simulation studies are used to select a nominal significance level that bounds the maximum type I error rate over the range of the unknown standard deviation at the desired level. Both restricted and unrestricted designs are considered, and this choice if found to be the main predictor of the maximal type I error inflation. The restricted design is chose due to the logistics of implementation (recruitment can proceed while the interim recalculation is done), and a smaller significance level adjustment needed.
TAGS: adjusted alpha, continuous, nuisance: SD, t-test, application, blinded

8. Conclusions

Research related to internal pilot studies has been progressing steadily since their introduction by Wittes and Brittain in 1990 [WB90] (Table 1). Most of the theoretical work has been done in the context of a prospective trial analyzed with a two-sample t-test and in the more general setting of a linear model with fixed covariates (Table 2). In these cases the residual variance is the nuisance parameter. Multiple methods for exact control of the type I error rate have been proposed either through adjusting the significance level and/or the critical value of the final test, or by modifying the test statistic. The modification is typically done by using a non-standard estimate of the variance parameter. Many of the methods are computationally simple, so no specialized software is provided. Some of the computationally intensive alpha-adjustment procedures are implemented as freely available SAS macros.

Most of the work with non-normally distributed outcome variables has a more applied nature. Methods for binary, ordinal, and count outcomes have been proposed, however most of the emphasis has been on developing the appropriate sample size formulas. In most of these papers the "naive" approach of no adjustment to significance level or test statistic is used, with simulation studies demonstrating reasonable type I error rate control and good power characteristics. Just as with normally distributed outcomes, the emphasis has been on prospective two-sample comparisons. There is no dedicated software for these methods; however they can be easily implemented.

Guided by regulatory considerations, a major thrust of the internal pilot research has been in ensuring blinding during the sample-size re-estimation step. Surprisingly, this appears to be simpler in situations where the nuisance parameters are strongly related to the treatment effect; for example, for binary data the required sample size depends on the probability of response in either group and not just their difference. While blinding is more important in these cases, since the estimate of the nuisance parameter could provide information about the current effect size, the required adjustments are fairly straightforward. For normally distributed outcomes the nuisance parameter is not informative with respect to the effect size, yet its blinded estimation is more complicated. Nonetheless, multiple approaches have been developed.

Despite the well established methodology, and generally favorable regulatory aspects, we found little evidence of internal pilots being used in practice. This may be due to the lack of awareness by both statisticians and applied researchers about internal pilot studies. The literature and methodology has a relatively narrow focus on prospective randomized trials, especially for drug development. In this area internal pilot designs "compete" with more complicated adaptive multi-stage designs with interim testing. The main strengths of internal pilots - their simplicity and minimal need for expert statistical support - could come into play outside the world of pharmaceutical clinical trials. We believe it would be extremely useful for early stage clinical and translational research, when large uncertainty about nuisance parameters exists, and external pilot studies are not feasible. By extending the methodology to incorporate a wider range of experimental designs, such as allowing for random covariates and observational data, and publicizing it in a wider range of fields, we hope that this valuable methodology can improve research.

Table 1: Article type and journal category by year of publication

	1945-1991	1992-1996	1997-2001	2002-2006	2007-2011
	N=2	N=5	N=14	N=15	N=15
Type					
application		20% (1)	7% (1)		7% (1)
application/methodology		20% (1)		7% (1)	
methodology	100% (2)	60% (3)	64% (9)	47% (7)	60% (9)
methodology/software			7% (1)		7% (1)
regulatory				13% (2)	
review			14% (2)	33% (5)	7% (1)
software			7% (1)		20% (3)
Journal category					
Clinical Medicine				7% (1)	7% (1)
Computer Science			7% (1)		20% (3)

Mathematics	100% (2)	80% (4)	93% (13)	67% (10)	53% (8)
Pharmacology & Toxicology		20% (1)		27% (4)	20% (3)

Numbers after percents are frequencies.

Table 2: Methods used for non-review articles

Alpha	
adjusted	17% (7)
adjusted/unadjusted	2% (1)
unadjusted	80% (33)
Power	
conditional	5% (2)
conditional/marginal	2% (1)
marginal	93% (38)
Nuisance	
other	27% (11)
SD	63% (26)
SD/other	10% (4)
Data	
binary	12% (5)
binary/continuous	5% (2)
continuous	73% (30)
count	7% (3)
ordinal	2% (1)
Effect	
fixed	93% (38)
reestimated	7% (3)
Test	
χ-square	5% (2)
F-test	12% (5)
F-test/t-test	5% (2)
permutation	5% (2)
quasi-likelihood	5% (2)
regression	5% (2)
t-test	39% (16)
t-test/χ-square	2% (1)
t-test/z-test	2% (1)
Wald	7% (3)
z-test	12% (5)
Blinded	

blinded	41%	(17)
blinded/unblinded	2%	(1)
unblinded	56%	(23)

Numbers after percents are frequencies.

File translated from TEX by TTH, version 4.03.
On 01 Mar 2012, 12:57.