HHS Public Access

Author manuscript

J Stat Softw. Author manuscript; available in PMC 2016 October 21.

Published in final edited form as:

J Stat Softw. 2008 November ; 28(7): . doi:10.18637/jss.v028.i07.

GLUMIP 2.0: SAS/IML Software for Planning Internal Pilots

John A. Kairalla,
The University of Florida

Christopher S. Coffey, and
The University of Alabama at Birmingham

Keith E. Muller
The University of Florida

Abstract

Internal pilot designs involve conducting interim power analysis (without interim data analysis) to modify the final sample size. Recently developed techniques have been described to avoid the type I error rate inflation inherent to unadjusted hypothesis tests, while still providing the advantages of an internal pilot design. We present **GLUMIP** 2.0, the latest version of our free SAS/IML software for planning internal pilot studies in the general linear univariate model (GLUM) framework. The new analytic forms incorporated into the updated software solve many problems inherent to current internal pilot techniques for linear models with Gaussian errors. Hence, the **GLUMIP** 2.0 software makes it easy to perform exact power analysis for internal pilots under the GLUM framework with independent Gaussian errors and fixed predictors.

Keywords

sample size re-estimation; power; adaptive designs

1. Internal pilot designs

1.1. Motivation

The greatest barrier to an appropriate choice of sample size in a study is often the choice of valid nuisance parameter values. Internal pilot designs involve conducting interim power analysis (without interim data analysis) to modify the final sample size of a study. When the best nuisance parameter values available for study planning may not reflect the true population parameter values, internal pilot designs become extremely advantageous.

For a linear model with Gaussian errors, the internal pilot is designed to avoid the uncertainty of the error variance. Data collection begins based on the sample size chosen with the best guess for the error variance. Some fraction of the planned observations are then used to re-estimate the variance. An interim power analysis is conducted based on the revised variance estimate and the *initially specified* effect of interest. The interim power

John A. Kairalla, PO Box 100177, Gainesville, FL 32610-0177, United States of America, Telephone: +1/352/265-0111 ext. 85845, Fax: +1/352/265-8047, jak@biostat.ufl.edu, URL: http://biostat.ufl.edu/

analysis allows adjusting the sample size up or down to help achieve the target power and not waste resources. Such designs differ from traditional (external) pilot studies in that the observations used to estimate the variance are included in the final analysis. In contrast to a group sequential design, an internal pilot design involves only an interim power analysis, with no interim hypothesis testing.

The **GLUMIP** 2.0 software package is designed to accurately assist in planning and implementation of internal pilot designs in the General Linear Univariate Model (GLUM) framework with Gaussian errors and fixed predictors using exact theory for power and type I error rate.

1.2. A detailed example

Consider a study of contrast-limited adaptive histogram equalization (CLAHE) to determine which image processing parameters improve observers' abilities to detect breast cancer in mammograms as a function of two within-subject image processing factors (clip value and region size) each having three levels (Pisano *et al.* 1998). One major objective was testing the difference, in log contrast units, between each of the nine processed conditions (clip value×region size combinations) versus the unprocessed condition. Thus, the task reduces to nine paired t tests. In planning the study, we seek the required sample size to ensure some level of target power (P_t) for a specified effect of interest (θ_*) at a target type I error rate (α_t). The true sample size required to meet these conditions also depends on the unknown variance, a nuisance parameter. The design parameters used for sample size determination were chosen from an unpublished earlier study:

Target type I error rate (Bonferroni corrected):	$\alpha_t = 0.01/9 = 0.0011$
Target power:	$P_t = 0.90$
Effect of interest (meaningful difference):	$\theta_* = 0.1$
Variance estimate (from an unpublished study):	$\sigma_0^2 = 0.0065$

With a univariate linear model of Gaussian data (and fixed sample size), the usual test of a general linear hypothesis achieves an actual type I error rate equal to the target type I error rate. However, in many situations, the two differ. For example, a Bonferroni correction often creates a conservative test because it only guarantees achieving an actual type I error rate no more than the target. For the example above, we shall avoid any discussion of type I error rate for the group of nine tests, and consider only the properties of a single test.

Many different programs allow computing power for studies with fixed sample sizes. Choices include free software, such as **POWERLIB** (Johnson *et al.* 2008) or **PS** (Dupont and Plummer 2004), and commercial software, such as **nQuery** (Statistical Solutions 2005) or **PASS** (NCSS 2005). For the CLAHE example, the calculations suggested that 20 readers (subjects) would be required to achieve 0.90 power to detect the effect of interest.

Unfortunately, after the study design had been completed and planning had begun for a study involving 20 readers, an error was discovered in the image processing for the previous

study providing the variance estimate. This error caused concern about the validity of σ_0^2. Since there was no consensus regarding the direction of the possible bias in the estimate, the scientists feared that this uncertainty could lead to a study with incorrect power. Hence, this study could have greatly benefited from the use of an internal pilot design. This example will be further discussed with enumeration in Section 3.2.

1.3. Review of internal pilot literature

The internal pilot design was introduced by Wittes and Brittain (1990) for two group settings with Gaussian outcomes. The method uses the updated variance estimate at the interim stage to re-estimate the sample size. In the final test, a variance estimate using all subjects is used with a usual test statistic and critical value assuming a fixed sample study.

In the case of continuous outcomes, most research in internal pilot designs involves only the independent group *t* test setting (Jennison and Turnbull 2000). However, not all designs, or even all clinical trials involve only two groups. Therefore, Coffey and Muller (1999, 2000a,b, 2001) and Coffey *et al.* (2007) described methods and many exact results, including a computable form of the distribution of the test statistic, for *any* univariate linear model with fixed predictors and Gaussian errors. Many *t* test results are special cases. Good reviews for internal pilot designs have been recently published by Proschan (2005) and Friede and Kieser (2006).

One potential drawback to utilizing an internal pilot design is that the final sample size becomes a random variable. The proposal by Wittes and Brittain (1990), an *unadjusted* test, ignores the randomness of the final sample size and uses the fixed sample test statistic and critical value (`TEST=0` in the **GLUMIP** program). The approach can have great benefits in terms of either increasing power if the original variance value was too small or reducing the expected sample size if the original variance value was too large. One major advantage of the unadjusted test is that it requires no new software for implementation. Existing software for power analysis can be used for the sample size re-estimation and, since the randomness of the final sample size is ignored, most standard statistical software packages can be used at the analysis phase. However, the risk of type I error rate inflation may offset the benefits in the minds of many researchers (Kieser and Friede 2000) and regulatory agencies (ICH Guideline E9 1998). A key motivating reason for using the software described here lies in the need to examine and account for potential inflation of type I error rate when using an internal pilot design.

With internal pilots, the randomness of the final sample size leads to a downwardly biased unconditional final variance estimate which causes type I error rate inflation (Proschan and Wittes 2000; Miller 2005). The amount of type I error rate inflation for the unadjusted test varies directly with the degree of downward bias in the final variance estimate (Coffey and Muller 2000a). Upward bias in the type I error rate results from the downward biased variance estimate residing in the denominator of the unadjusted test statistic. Any approach with comparable power and expected sample size which preserves the type I error rate at or below the target level will be preferred to the unadjusted approach. Consequently, the focus

in the two independent group univariate setting has shifted to retaining most benefits of an internal pilot design while controlling the type I error rate.

Proposed methods for controlling the type I error rate generally fall into two categories corresponding to whether or not the blind for treatment group allocation is maintained. For blinded sample size re-estimation, Gould and Shih (1992) and Zucker *et al.* (1999) suggested using the one-sample variance estimator, with a simple adjustment based on the planned treatment effect of interest. When the true treatment difference is close to the prespecified difference, Kieser and Friede (2003) showed that this approach approximately controls the type I error rate. From a regulatory standpoint, methods that keep the treatment group allocation blinded may be preferred to those that require unblinding (ICH Guideline E9 1998). However, as Miller (2005) pointed out, the decision as to whether a blinded or unblinded procedure should be used must be made on a case by case basis. In many instances, an unblinded procedure may be appropriate provided that steps are taken to minimize the number of individuals who have access to the unblinded information, e.g., the use of an independent statistician.

All of the internal pilot methods contained in the **GLUMIP** 2.0 software require treatment group unblinding. Included methods to address the concern of type I error rate inflation fall generally into two categories:

1. Three approaches replace the downward biased unadjusted estimate with an unbiased variance estimate in the denominator of the test statistic to control the unconditional type I error rate. Each uses all of the available data for estimating the numerator of the test statistic (mean effect), but differ in the amount of information used in the denominator (variance estimate). Hence, each method controls the type I error rate by weighting the observations unequally for the purpose of computing a final variance estimate.

   Stein (1945): Use a variance estimate based only on the internal pilot sample.

   Zucker *et al.* (1999): Use a variance estimate based only on that part of the final sums of squares error orthogonal to the internal pilot sample. The approach controls type I error rate both conditionally and unconditionally.

   Proschan and Wittes (2000): Combine the two estimates using fixed weights that are not a function of the observed data. Note that this test is only appropriate if the final sample size is not allowed to decrease below the originally planned sample size.

2. One approach uses the unadjusted test statistic but increases the critical value to ensure that the maximum possible type I error rate is no greater than the target level.

   Coffey and Muller (2001): Approach referred to as a *bounding* method since it guarantees an upper bound on type I error rate. The test may be

conservative, i.e., the observed type I error rate may be less than the target level.

Other unblinded methods for controlling the type I error rate have also been proposed. Miller (2005) proposed a correction to the unadjusted variance estimate that holds the type I error rate at or below the target level. Denne and Jennison (1999) proposed a method that approximately controls the type I error rate using a degrees of freedom adjustment factor to calculate the critical value. These methods are not included in **GLUMIP** 2.0 but may be included in future versions.

Coffey and Muller (2001) compared the performance of three of the approaches above, Stein (1945), Zucker *et al.* (1999), and Coffey and Muller (2001), in terms of maintaining the benefits of the internal pilot design while controlling type I error rate across a range of conditions. They varied 1) the rule for sample size re-estimation, 2) the true variance value, 3) the size of the internal pilot sample, 4) whether or not the final sample size was allowed to decrease if the original variance value was too large, and 5) whether or not a finite maximum sample size was specified. With large samples (~ 500+), the choice of method had little impact on the power and expected sample size. Furthermore, in general, there was minimal (if any) type I error rate inflation with the unadjusted test. Thus, internal pilot designs can be utilized in large randomized clinical trials to assess key nuisance parameters and make appropriate modifications with little cost in terms of an inflated type I error rate. However, in small to moderate samples, the inflation of the type I error rate with the unadjusted test was a concern and the choice of method had a big impact on the power and expected sample size. In general, the bounding test best maintained the benefits of an internal pilot design while controlling type I error rate across the entire range of conditions considered. Coffey and Muller (2000b) demonstrated that the inflation of type I error rate diminishes as the size of the study increases. This suggests that if one were able to determine a priori that the maximum possible inflation of type I error rate was so small as not to cause concern, researchers could easily gain the benefits of the internal pilot using the unadjusted method without having to worry about potential problems with type I error rate inflation.

It is important to distinguish the similarities and differences between internal pilot designs and other common approaches to analyzing and monitoring data from clinical trials. In the past 20 years, the use of group sequential methods to monitor the efficacy and safety data on an interim basis has become common in large clinical trials. Such methods often involve multiple interim analyses and allow early stopping for efficacy or futility, early termination of ineffective arms, and more efficient utilization of limited resources. Hence, group sequential (GS) methods protect against observing a larger or smaller than expected effect during the course of a trial. On the other hand, since internal pilot designs protect against the mis-specification of nuisance parameters at the design stage of a study, they are usually effective with just two stages. Current research is looking at ways to simultaneously obtain the benefits of both approaches, i.e., combine the early stopping benefits of GS methods with sample size re-estimation methods protecting against mis-specification of nuisance parameters (Tsiatis 2006).

More recently there has been great interest in the development of adaptive design (AD) methodology. Adaptive designs offer researchers flexibility to redesign trial procedures and analysis at interim stages if the data disagree with planning assumptions. The large volume of recent research has created a confusion in terminology with many types of study modification referred to as *adaptive*. In Spring 2005, a PhRMA working group on ADs in clinical drug development was formed to investigate and facilitate their acceptance and usage. An Executive Summary of the group's findings (Gallo *et al.* 2006) defines an adaptive design as "a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial." Under this definition, the adaptations only use information from accumulating data as opposed to *flexible* designs which can also incorporate external information (Burman and Sonesson 2006). The PhRMA working group also stresses that the changes should be made "by design" and not undertaken on an ad hoc basis. Adaptive designs often include planned sample size re-estimation based on revised estimates of nuisance parameters and treatment effects. Under the PhRMA definition, it is clear that internal pilots are a special case of adaptive designs with sample size re-estimation based only on updated nuisance parameter estimates. For a more detailed recent discussion comparing various types of adaptive designs and group sequential methods, see Shih (2006).

Various software packages and modules have been released to assist in the planning and analysis for group sequential and adaptive designs. Notable among these packages are **ADDPLAN** (Wassmer and Eisebitt 2007) and **East** (Cytel Inc. 2007). **ADDPLAN** was originally developed to perform the inverse normal combination test proposed by Lehmacher and Wassmer (1999), but currently also supports group sequential methods as well as other methods such as the adaptive inverse normal design which allow for data-driven adaptations. **East** is a well known software package mainly used for planning and analysis of group sequential trials. However, it allows for study adaptation through information based monitoring techniques that can adjust study sample sizes. Wassmer and Vansemeulebroeck (2006) published a review on available group sequential and adaptive software with more details. Very little currently exists in available software to directly assist in the design and implementation of internal pilot designs, however, **East** can be utilized to plan large sample internal pilot studies as a special case of the information based monitoring tools. The underlying theory of the program, however, is based on large samples and trial and error simulations would be needed in order to implement accurate planning in small sample situations. Additionally, a recent SAS macro by Wang *et al.* (2008) uses computationally intensive simulations to assist in the planning and design of blinded internal pilots with the randomization test described by Kieser and Friede (2003).

2. Program description

2.1. Overview of the program

The **GLUMIP** 2.0 software package uses exact theory to perform power analysis for internal pilot designs under the GLUM framework using results from Coffey and Muller (1999, 2000a,b, 2001) and Coffey *et al.* (2007). Accordingly, while accurate for larger studies, the software package has its primary value in planning small sample studies. Broad classes of

ANOVA and regression problems can be treated with these methods. Key restrictions include the assumption of Gaussian errors, fixed predictor values, common design for all replications, and no missing data. The notation used in the program is based on the model introduced in Coffey and Muller (1999), which includes the two sample t test as a special case:

$$\underset{N_+\times 1}{\boldsymbol{y}_+} = \underset{N_+\times q}{\boldsymbol{X}_+}\,\beta + \underset{N_+\times 1}{\boldsymbol{e}_+}\ . \quad (1)$$

The internal pilot design leads to interest in two different but intimately connected models. The combined model for the final analysis may be written as

$$\begin{bmatrix} \underset{n_1\times 1}{\boldsymbol{y}_1} \\ \underset{N_2\times 1}{\boldsymbol{y}_2} \end{bmatrix} = \begin{bmatrix} \underset{n_1\times q}{\boldsymbol{X}_1} \\ \underset{N_2\times q}{\boldsymbol{X}_2} \end{bmatrix} \beta + \begin{bmatrix} \underset{n_1\times 1}{\boldsymbol{e}_1} \\ \underset{N_2\times 1}{\boldsymbol{e}_2} \end{bmatrix}, \quad (2)$$

with partitioning corresponding to the n_1 and, random, N_2 observations in the internal pilot and second samples, respectively. All rows of e_1 and e_2 are assumed to follow independent Gaussian distributions with mean zero and variance σ^2. Let $Es(\boldsymbol{X})$ represent the *essence* matrix, i.e., the matrix created by deleting duplicate rows from a design matrix (Helms 1988). We restrict $Es(\boldsymbol{X}_1) = Es(\boldsymbol{X}_2)$. Hence, the software applies to designs where the values of the predictors are known in advance and we increment random total sample size, $N_+ = n_1 + N_2$, only in multiples of a constant replication factor, m. Consequently, the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ span the same space and hence $rk(\boldsymbol{X}_1) = rk(\boldsymbol{X}_2) = rk(\boldsymbol{X}_+) = r$. We test H_0: $\theta = C\beta = \theta_0$, with C a fixed $a \times q$ contrast matrix. Without loss of generality assume $\theta_0 = 0$. We seek a sample size large ×enough to ensure target power (P_t) and type I error rate (α_t) for a *scientifically important* effect of interest ($\theta = \theta_*$). See Coffey and Muller (1999, 2001) or Coffey *et al.* (2007) for additional details.

Three sample size re-estimation rules are available within the current **GLUMIP** version:

- `RULE=0` (**Unadjusted**): Sample size re-estimation is based on a fixed sample F approximation (Wittes and Brittain 1990).
- `RULE=1` (**Stein**): Sample size re-estimation is based on an F approximation with $n_1 - r$ denominator degrees of freedom (Stein 1945).
- `RULE=2` (**Second Sample**): Sample size re-estimation is based on an F approximation with n_2 denominator degrees of freedom Zucker *et al.* (1999).

Five testing techniques are available within the current **GLUMIP** version.

- `TEST=0` (**Unadjusted**): Ignores the randomness of the final sample size and uses the fixed sample test statistic and critical value (Wittes and Brittain 1990).

- `TEST=1` (**Stein**): Use a variance estimate based only on the internal pilot sample. This approach is based on Stein's two-stage procedure (Stein 1945).
- `TEST=2` (**Second Sample**): Use a variance estimate based only on that part of the final sums of squares error orthogonal to the internal pilot sample (Zucker *et al.* 1999).
- `TEST=3` (**Bounding**):Use the unadjusted test statistic but increases the critical value to ensure that the maximum possible type I error rate is no greater than the target level (Coffey and Muller 2001).
- `TEST=4` (**Weighted**): Combine the two estimates using fixed weights that are not a function of the observed data (Proschan and Wittes 2000). Note that this test is only appropriate if the final sample size is not allowed to decrease below the originally planned sample size.

For the purposes of this software, an internal pilot *Method* is defined as a combination of sample size re-estimation rule (`RULE`) and testing technique (`TEST`). For example, Method 0/3 implies that `RULE=0` and `TEST=3`. In theory (and in fact **GLUMIP** 2.0 allows for it) one can combine any of the re-estimation rules and tests, however, five natural combinations seem most appealing:

- **Method 0/0 (Unadjusted)**: Sample size re-estimation is based on a fixed sample F approximation. The randomness of the final sample size is ignored and the fixed sample test statistic is used (may inflate type I error rate).
- **Method 1/1 (Stein)**: Sample size re-estimation is based on an F approximation with $n_1 - r$ denominator degrees of freedom. The denominator of the test statistic contains the variance estimate using only the first n_1 observations.
- **Method 2/2 (Second Sample)**: Sample size re-estimation is based on an F approximation with n_2 denominator degrees of freedom. The denominator of test statistic contains the variance estimate using observations orthogonal to the first n_1 observations.
- **Method 0/3 (Bounding)**: The sample size re-estimation and test statistic are the same as for the unadjusted, but the critical value is modified to ensure that the maximum type I error rate as a function of γ, the ratio of the true variance to the variance value used for planning, equals the target type I error rate.
- **Method 0/4 (Weighted)**: The sample size re-estimation is the same as in the unadjusted method. The test statistic is based on combining the variance estimate using only the first n_1 observations and the variance estimate using only those observations orthogonal to the first n_1 observations using fixed, pre-determined weights.

2.2. What's new?

GLUMIP 2.0 includes one new test and one modified test as compared to the previous version. Additionally, a new module (`N2CALC`) has been created to assist in the calculation of the second stage sample size during study implementation.

The weighted test statistic proposed by Proschan and Wittes (2000) has been added (`TEST=4`). For this method, the sample size re-estimation rule is the same as for the unadjusted test, but the test statistic is based on combining the variance estimate using only the first n_1 observations and the variance estimate using only those observations orthogonal to the first n_1 observations using fixed, pre-determined weights. In previous versions of the software, `TEST=1` corresponded to a modified Stein test in which the test statistic was based on a variance estimate using only the minimum number of observations certain to be in the final sample ($n_{+,\min}$). In the current version, `TEST=1` corresponds to a pure Stein-like test for which the test statistic is based on a variance estimate using only the internal pilot sample (n_1).

Coffey and Muller (2001) demonstrated that the bounding method best maintained the benefits of an internal pilot design while controlling type I error rate across a wide range of conditions. Unfortunately, the computational tools utilized in previous versions of the **GLUMIP** software for this method were often very slow and unstable, which limited the practicability of doing a large number of calculations. **GLUMIP** 2.0 has several changes that provide notable improvements in computing stability, speed, and accuracy. The problems addressed and the modifications used to address them in the new version of the software are summarized below. See Coffey *et al.* (2007) for additional theoretical and descriptive details.

1. The old algorithm computed type I error rate or power by nesting two numerical integrations and required the use of Davies' algorithm (Davies 1980). New expressions, based on simple forms for the test statistic density, allow the CDF of the internal pilot test statistic to be written as a series of single integrations with well-behaved function evaluations. The new expressions are faster, more stable, and more accurate than the previous versions.
2. The old algorithm for finding the maximum type I error rate for a given value of α sometimes failed to converge. This problem was eliminated by replacing a derivative-based algorithm with a Fibonacci search (Kiefer 1953) in the `_MAXTS` module.
3. Although theoretically it should be, the computed type I error rate was not always a locally monotone function of γ using the old algorithm. The underlying problem was due to the lack of consistency of stopping values among different γ values (with all other inputs held constant). To correct this, the algorithm in the `_NDIST` module was modified to ensure more consistent stopping values.
4. Although always very close, the desired bound in type I error rate was often not fully achieved using the old algorithm, particularly for examples

with small target type I error rates. This problem was addressed by modifying the stopping rule used in the _FINDADJ module in several ways to guarantee a proper bound on the type I error rate.

The new algorithms provide much faster and more stable calculations of type I error rate and power. As a consequence, the bounding method becomes practical and easy to implement even in small sample studies, where it has the most value.

With knowledge of the sample size re-estimation rules, current software can be employed to calculate the second stage sample size based on the revised variance estimate. However, the module N2CALC was created to simplify and automate this process when desired. The module uses the given sample size update rule (RULE), a variance input, and the study design to calculate the additional sample size needed for the second stage. See Section 2.5 for more details.

2.3. Running the program

The **GLUMIP** program consists of a set of modules written in SAS/IML. Because the program is run from within the SAS/IML package, the user must first issue the PROC IML; statement. The **GLUMIP** program code can then be attached to the user's program with a %INCLUDE statement. **GLUMIP** can be called from the user's program by four different commands: RUN GLUMIP;, RUN MAXTS;, RUN FINDADJ;, or RUN N2CALC;. The RUN GLUMIP; command accesses the main module and is used for power, type I error rate, and expected sample size calculations for all of the possible RULE/TEST combinations. The RUN MAXTS; command can be used to determine the maximum type I error rate for the unadjusted test (TEST=0) as a function of $\gamma=\sigma^2/\sigma_0^2$ (the ratio of the true variance to the variance value used for planning). The RUN FINDADJ; command can be used with the bounding test (TEST=3) to determine an adjusted level of that can be used to solve for a critical value such that the maximum type I error rate (over possible values of γ) is no larger than α_t. The RUN N2CALC; command can be used to determine the second stage sample size needed during the interim analysis for a given first stage variance estimate, $\hat{\sigma}_1^2$ (SIGHAT1), and a given sample size re-estimation rule (RULE).

The following module names are used by the software, and hence should not be used by the user: _MAXTSB, _FINDADJB, _GLUMIP, _FCTS1, _FCTS2, _FCTT, _GETPOW, _QUAD2D01, _NDIST, _FINDADJ, _MAXTS, and _N2CALC. With 27 exceptions, all matrices used within the modules are locally defined, so there is no possibility of accidentally redefining a user-defined matrix. The exceptions are the required input matrices (SIGMA0, SIGHAT1, ALPHAT, POWERT, C, N1, BETA_PLN, ESSENCEX, BETA_ALT, GAMLIST, NPLUSMIN, RULE, TEST), the optional input matrices (NPLUSMAX, ROUND, WEIGHTS), the output matrices (_IPCALCS, _IPNAMES, _MAXTS, _MAXTSNM, _FINDADJ, _FINDADJNM, _N2CALCS, _N2CALCNM) and the matrices of parameters necessary for invoking the algorithms (GLOBAL1, GLOBAL2, _ZERO_).

Table 1 summarizes the notation. See Coffey and Muller (1999, 2001) or Coffey *et al.* (2007) for additional details. The table also provides the corresponding variable names in the

GLUMIP program. The user must specify the design parameters for the initial power calculation. These include scalars for the initial variance estimate (`SIGMA0`), target type I error rate (`ALPHAT`), and target power (`POWERT`), a column vector of primary planning parameters used for sample size calculations (`BETA_PLN`), an *essence* matrix (`ESSENCEX`), and contrast matrix (`C`) for the initial power calculation. Note that these parameters are simply those that would be required for a power calculation in a fixed sample setting. The user must also specify a column vector containing the true matrix of primary parameters (`BETA_ALT`) and a list of gamma values (`GAMLIST`, where gamma is the ratio of the true variance to the variance value used for planning) to be used in the power calculations. The program tolerates defining `GAMLIST` as either a row or column vector. Finally, the user must specify the internal pilot specifications to be utilized for the chosen study. These consist of scalars for the choice of internal pilot sample size (`N1`), the minimum allowable size of the final sample (`NPLUSMIN`), the sample size re-estimation rule to apply (`RULE`), and the method for hypothesis testing (`TEST`).

By default the program performs computations assuming a balanced design with no finite upper bound on the final sample size. The user may choose to modify this by supplying one or more optional matrices. The user can specify a scalar (`NPLUSMAX`) to define a maximum allowable final sample size. In addition, the user can specify `ROUND`, a value between 1 and 10. This value determines when probabilities are sufficiently small to stop calculating the pdf of the final sample size for the largest value in `GAMLIST`. To allow for purposefully unbalanced designs, the user can specify a vector of weights (`WEIGHTS`) for each row of `ESSENCEX` per replication. For example, in a two group comparison where it is desired to maintain a 2:1 ratio in group 1 vs. group 2, one would choose a 2×2 identity matrix for `ESSENCEX` along with the following weight vector: `WEIGHTS={2 1}';`. Note that the replication factor m is calculated by the program as the sum of the elements of the `WEIGHTS` vector. For this unbalanced design, this implies that $m = 3$ while in a corresponding balanced design, $m = 2$ is the number of independent groups. This factor does not ever have to be specified and `WEIGHTS` does not have to be specified for balanced designs.

The current version of the software has a limited error checking capacity. **GLUMIP** 2.0 will check to make sure that all required inputs are valid. If an invalid input is encountered (e.g., a negative variance), the program will stop and warn the user.

2.4. How the program works

The **GLUMIP** package can be called from the user's SAS/IML program by any of four different commands: `RUN GLUMIP;`, `RUN MAXTS;`, or `RUN FINDADJ;` or `RUN N2CALC;`. The `RUN GLUMIP;` command accesses the software's main module, which calls on other modules to perform various tasks and functions.

If the `RUN GLUMIP;` command is given, the program first determines what value to use for critical value calculations associated with the final test statistic. For all tests (`TEST=0, 1, 2`, or `4`) other than the bounding test (`TEST=3`), this is simply the target level input to the program (`ALPHAT`). If the bounding test is selected, the program uses the `_FINDADJ` module to determine the adjusted level of alpha to use for determining the critical value such that the

maximum type I error rate as a function of γ will equal ALPHAT. The _FINDADJ module begins by calling the _MAXTS module to determine maximum type I error rate at the specified value of ALPHAT. Linear interpolation is then used as an initial guess for the required α adjustment, along with +/− 10% on either side. The algorithm then uses a doubly iterative search to find the desired adjustment. The inner iteration calls the _MAXTS module to determine the maximum type I error rate at a specified value of α (which changes during each loop). The outer iteration searches for the adjusted $\alpha*$ value such that the maximum type I error rate across γ equals α_t. Success of the outer iteration depends directly on the accuracy in achieving type I error rate bounded at or below α_t. The outer iteration continues until either 1) $\alpha* \leq \alpha_t$ is found with maximum type I error rate sufficiently close in relative difference to α_t, or 2) the lower and upper limits of $\alpha*$ are sufficiently close to each other relative to α_t. In the first case, the $\alpha*$ value is returned. In the latter case, the lower limit is returned.

For all tests (TEST=0, 1, 2, 3, or 4), following selection of the α value for critical value calculations, the program determines the distribution of the final sample size for the maximum value in the GAMLIST vector. This process uses the _NDIST module and depends on the specified re-estimation rule (RULE). The distribution of the final sample size is obtained by noting the monotone relationship between the final sample size, n_+, and the variance estimate obtained from the internal pilot sample, $\hat{\sigma}_1^2$. For each possible value of n_+, the program determines the cut point of $\hat{\sigma}_1^2$ that, if used in a sample size calculation, would lead to a sample size less than or equal to n_+ (Coffey and Muller 1999, 2001). This process continues until either the maximum final sample size is encountered or the probability of observing a sample size greater or equal to the current value of n_+ falls below some threshold defined by $\alpha_t \cdot 10^{-\text{ROUND}}$ (by default ROUND is 3). The program then begins looping through each of the other specified values in GAMLIST. For each γ value, the process described above is applied until the algorithm first encounters the maximum final sample size specified by the user, the sample size at which the algorithm stopped for the maximum value in GAMLIST, or the current value of n_+ falls below some threshold defined by $\alpha_t \cdot 10^{-(\text{ROUND}+2)}$.

The distribution obtained from _NDIST is used to calculate the expected final sample size and is used in the _GETPOW module to compute the exact power under an internal pilot design for the specified testing method. For TEST=0, 1, 2, or 4, the critical value is obtained by choosing the $1-\alpha_t$ percentile of an appropriate F distribution. If TEST=3 (the bounding test), the program uses the adjusted alpha level returned by the _FINDADJ function for determining the critical value. The power calculation consists of two stages. In the first stage, we condition on n_+ and perform appropriate calculations. For TEST=0 or 3, the SAS/IML **QUAD** function is used along with either the _FCTT module (if $n_+ = n_1$) or the _FCTS1 and _FCTS2 modules to integrate the CDF of the test statistic across the cut points for each possible n_+ value determined from _NDIST. For TEST=1, the **QUAD** function is used along with the _FCTT module to integrate the CDF. For TEST=2, the test statistic conditionally follows an F distribution and only the SDF function is required for the conditional calculation. For TEST=4, the _QUAD2D01 module is used to invoke the Davis

and Rabinowitz (1984) method of numerical integration to perform 2D numerical integration of the quantile transformed integrand of the test statistic. Finally, the _GETPOW module returns the unconditional power determined by applying the law of total probability and summing over all possible values for the final sample size.

The _GLUMIP module repeats the above for each value specified in GAMLIST. After looping through all values, the module returns a matrix, _IPCALCS, that has one row for each value in GAMLIST and eleven columns corresponding to the values for ALPHAT, the alpha value used to compute the critical value (equal to ALPHAT unless the bounding test is used), POWERT, GAMMA, N1, NPLUSMIN, NPLUSMAX (the SAS missing value .I is returned if no finite maximum sample size is specified), RULE, TEST, expected sample size, and power (or type I error rate if under null hypothesis). The program also returns a matrix, _IPNAMES, which contains column labels useful when printing the _IPCALCS matrix. These matrices are available to the user for further manipulation until the next RUN GLUMIP; command is issued. If the user wants to keep these values after an additional invocation of **GLUMIP**, they should be stored in matrices with distinct names before reissuing the RUN GLUMIP; command, or the results will be lost.

In some extreme cases, all of the probability may lie in one value for the final sample. For example, consider a case where no reduction in the planned sample size is allowed but we have dramatically overstated the true variance during planning. It may follow that the original planned sample size will be observed with a probability of 1. In such situations, we effectively have a fixed sample size and calculations can be simplified and based on an appropriate F distribution.

This software was written and has run successfully under SAS 9.1.3 using the Microsoft PC-Windows operating system.

2.5. Additional useful functions

The **GLUMIP** program provides the RUN MAXTS; command as a stand alone option for determining the maximum type I error rate as a function of γ when using the unadjusted method. This command requires the same design parameters and internal pilot rules (except for RULE and TEST) as the RUN GLUMIP; command. The RUN MAXTS; command accesses the main module, which calls on the _MAXTSB and _MAXTS modules to implement an optimal Fibonacci algorithm to determine the maximum type I error rate for the given set of inputs. This command produces a list: _MAXTS, which contains the value of γ at which the maximum type I error rate occurs, the maximum type I error rate, and the ratio of the maximum type I error rate to the target type I error rate. The module also produces _MAXTSNM, a list of column labels useful when printing _MAXTS. See Section 3.1 or 3.2 for use of the MAXTS module in examples.

If the user wants to implement a bounding test to protect against type I error rate inflation, the **GLUMIP** program provides the RUN FINDADJ; command as a stand alone option. The command finds an adjusted alpha level to ensure that the maximum type I error rate (over possible values of γ) is no larger than ALPHAT. The same design parameters and internal

pilot rules as the `RUN GLUMIP;` command are required except for `RULE` and `TEST`. The `RUN FINDADJ;` command accesses the main module, which calls on the `_FINDADJB` and `_FINDADJ` modules. This command returns a scalar value, `_FINDADJ`, which contains the adjusted alpha value and `_FINDADJNM`, a column label useful when printing the returned value. See Section 3.2 for use of the `FINDADJ` module in an example.

The **GLUMIP** program provides the `RUN N2CALC;` command as a stand alone option for determining the second stage sample size needed during study implementation. This command requires the same design parameters and internal pilot rules (except for `TEST` and `SIGMA0`) as the `RUN GLUMIP;` command. Additionally, the first sample error variance estimate (`SIGHAT1`) is a necessary input. The `RUN N2CALC;` command accesses the main module, which calls on the `_N2CALC` module to implement the necessary sample size calculations.

The code determines the second stage sample size in the following manner. For each candidate value for the second stage sample size (increasing by replication factor m), a critical value is determined using the given sample size re-estimation rule (`RULE`). Next, the minimum noncentrality factor needed for significance of the final test is calculated. From this noncentrality, assuming the effect size of interest, a cut-off for a sufficiently low variance value is determined and the program stops if the first sample variance estimate is less than or equal to this value. Otherwise the program increases the sample size until either the condition is met, or the maximum sample size (`NPLUSMAX`) is attained.

This command produces a list: `_N2CALCS`, which contains `SIGHAT1`, `N1`, `RULE`, `N2CALC` (second stage sample size), `NPLUSCALC` (total sample size), and the projected power of the study using the given total sample size. The module also produces `_N2CALCNM`, a list of column labels useful when printing `_N2CALCS`. See Section 3.5 for use of the `N2CALC` module in an example.

2.6. Calling the program from a user written module

The software can also be run by using a `CALL` statement. This allows using the **GLUMIP** software within other modules. Note that the **GLUMIP** input matrices must all be specified as global variables in the modules referring to them. The following incomplete code (Program 1) illustrates how to define a user module that calls the **GLUMIP** software.

3. Example programs

We illustrate the usefulness of our code by reproducing results reported in four journal articles and one textbook. With the exception of our own works, the original published results were based on simulations. However, power calculation via simulation should only be a last resort since simulations have many practical disadvantages when compared to exact or approximate methods. Specifically, simulations typically require large amounts of computer time and each slight design change requires waiting for a new set of simulations to run. These limitations are worsened by the additional dimensions of study design created by the introduction of an internal pilot design. (Due to the fact that we are re-estimating the sample size at the internal pilot phase, a power calculation is embedded in the internal pilot phase.

Hence simulating the power of an internal pilot design conditional on a particular first stage sample size requires, in essence, a simulation embedded in a simulation.) The output given are results using exact theory embedded in the **GLUMIP** program that are fast, stable, and accurate, thus removing the need to conduct simulations.

3.1. Example from Wittes and Brittain (1990) paper

A two group comparison previously described in Wittes and Brittain (1990) illustrates the use of the software. For this example with $\alpha_t = 0.05$, $P_t = 0.90$, $\theta_* = 1$ and $\sigma_0^2 = 2$, a fixed sample size power calculation suggests 43 subjects per group ($n_0 = 86$). We consider an internal pilot design with 22 subjects per group in the internal pilot sample ($n_1 = 44$), no finite maximum sample size, and no reduction of the final sample size if the variance was originally overstated ($n_{+,\min} = 86$). Program 2 invokes the `RUN GLUMIP;` command to reproduce the power calculations reported in Table I of the Wittes and Brittain (1990) paper and Table II of the Coffey and Muller (1999) paper. For all examples displayed here, the **GLUMIP** code is attached using a SAS macro variable (`&PROGPATH`) to point to the directory where the **GLUMIP** 2.0 program is saved. For instance, all of the examples displayed here assume that the **GLUMIP** source code is saved in the H:\GLUMIPZIP\IML directory on the user's hard drive. Please note that this line of code will need to be changed to point to the directory where the **GLUMIP** 2.0 program is saved on the user's computer. The user must then provide values for all required inputs and any optional inputs desired. Using the `/NOSOURCE2` option prevents the hundreds of lines of **GLUMIP** code from appearing in the SAS log. The `RUN GLUMIP;` command begins execution of the **GLUMIP** program. Program 2 generates the following output, indicating that the internal pilot design successfully maintains power at the target level of 0.90 regardless of the true value of the variance.

Program 3 invokes the `RUN GLUMIP;` command to reproduce the type I error rate calculations for this same example. Note that `BETA_ALT`, the vector of true parameters, takes the null vector value of 0_2 for this calculation. Program 3 generates the following output.

While the results above suggest that type I error rate inflation is not a significant issue, we only observed the type I error rate for 5 specific values of γ. If one were able to determine a priori that the maximum inflation of type I error rate was so small as not to cause any concern, researchers could gain the benefits of the internal pilot using the unadjusted method without having to worry about potential problems with type I error rate inflation. Program 4 invokes the `RUN MAXTS;` command to determine the extent of potential type I error rate inflation for this example.

Program 4 generates the following output, indicating the maximum type I error rate of 0.0518 occurs at $\gamma = 1.4425$. Hence, for this example, the maximum amount of type I error rate inflation is a little under 4% and no adjustment may be required. As a consequence, researchers could ignore the randomness of the final sample size and use most standard statistical software packages at the analysis phase.

(n_1 = 44), no finite maximum sample size, and no reduction of the final sample size if the variance was originally overstated ($n_{+,\min}$ = 86)

In Program 5, we slightly alter the example from Wittes and Brittain (1990) in order to exemplify the use of unbalanced groups for power and expected sample size calculations with the **GLUMIP** 2.0 software. We will assume a 2:1 unbalanced two group design, which might be considered if a new treatment were more expensive than the control treatment. With the same effect size and planning variance used, the fixed sample, unbalanced design requires 96 subjects (n_0 = 96). The internal pilot sample will use 48 subjects (n_1 = 48) with the minimum final sample size not allowed to decrease ($n_{+,\min}$ = 96). Note the addition of the `WEIGHTS = {2 1}';` line to the program to specify the non-default unbalanced design preference. Program 5 generates the following output for the unbalanced design.

3.2. Example from Coffey and Muller (2001) paper

We revisit the CLAHE example (Pisano *et al.* 1998) described in Section 1.2 and used in Coffey and Muller (2001) to demonstrate the performance of various methods of controlling type I error rate while gaining the benefits of an internal pilot design. Specifically, we consider an internal pilot design with the following specifications: 1) a pre-planned sample size of n_0 = 20 radiologists, 2) the first n_1 = 10 radiologists comprise the internal pilot sample (`N1=10`), 3) the final sample size is allowed to decrease if the original variance value overestimates the true variance (`NPLUSMIN=10`), and 4) a finite upper bound of $n_{+,\max}$ = 30 radiologists (`NPLUSMAX=30`). Program 6 invokes the `RUN MAXTS;` command to determine the maximum type I error rate using the unadjusted method for the CLAHE example in this setting. Program 6 generates the following output, indicating the maximum type I error rate of 0.0019 occurs at γ = 1.70:

Hence, for this example, ignoring the randomness of the final sample size introduced by using an internal pilot design could lead to as much as a 70% inflation of the target type I error rate. In such situations, one should attempt to maintain the benefits of an internal pilot design while controlling type I error rate using one of the methods described in Coffey and Muller (2001).

The bounding method provides one option for controlling the type I error rate for this example. Program 7 invokes the `RUN _FINDADJ;` command to determine the adjusted alpha value to use for determining the critical value. Program 7 generates the following output, indicating that the critical value for the bounding test should be computed using α_* = 0.0006.

In general, if the analysis calls for a final sample size n_+, the critical value for a one group t test for a given α can be written $t_{crit} = F_t^{-1}(1-\alpha/2, n_+-1)$. For this example, if n_+ = 20 is determined, this translates to using an adjusted critical value of $t_{crit} = F_t^{-1}(1-0.0006/2, 19)$ = 4.11 while the unadjusted test with α = 0.0011 would use $t_{crit} = F_t^{-1}(1-0.0011/2, 19)$ = 3.84.

However, the bounding test is not the only option for controlling the type I error rate. Rows 4–6 of Table 2 in Coffey and Muller (2001) provided some results for expected sample size and power when applying Methods 0/0, 1/1, 2/2, and 0/3 to the CLAHE example with these internal pilot rules. Program 8 uses **GLUMIP** 2.0 to reproduce the calculations in addition to including calculations for Method 2/2. In the current version of the software, `RULE` and `TEST` must be scalar values. Hence, we had to write our own `DO-END` loop to compute the power and expected sample size for multiple combinations of the two. In addition, note that when using Method 2/2, we must set `NPLUSMIN=12` to ensure that we have enough observations in the second sample to compute an estimate of the variance. Program 8 produces the following results, which clearly agree with the results shown in rows 4–6 of Table 2 in Coffey and Muller (2001):

3.3. Example from Proschan and Wittes (2000) paper

A two group comparison previously described in Proschan and Wittes (2000) with $\alpha_t = 0.05$, $P_t = 0.90$, $\sigma^2 = 1$, and $\theta_* = 1$ illustrates the performance of the weighted test. The authors reported simulation results to evaluate the type I error rate and power of Stein's test, the weighted test, and the unadjusted test for three different scenarios, corresponding to whether the prior estimate of the standard deviation was 25% low, on target, or 25% high ($\sigma_0^2 \in$ 0.5625, 1.0, 1.5625). For these three scenarios, a fixed sample size power calculation suggests 12, 22, and 34 subjects per group, respectively ($n_0 \in$ 24, 44, 68). They considered an internal pilot design with one-half of the originally planned subjects included in the internal pilot sample ($n_1 \in$ 12, 22, 34), no finite maximum sample size, and no reduction of the final sample size if the variance was originally overstated. Program 9 invokes the `RUN GLUMIP;` command to reproduce the type I error rate calculations reported for Table 1 in the Proschan and Wittes (2000) paper. Note that when `TEST=4` is used, **GLUMIP** issues the following caution (not included in output): `WARNING: TEST=4 is only appropriate when NPLUSMIN = N_not, the originally planned sample size. Otherwise interpret results with caution`. Program 9 generates the following output.

To reproduce the power calculations for Table 1 in the Proschan and Wittes (2000) paper, the only change that is required to the above code is to replace the statement `BETA_ALT = {0 0}';` with the statement `BETA_ALT = {0 1}';`. The code (Program 10) generates the following output.

3.4. Example from Jennison and Turnbull (2000) textbook

In Table 14.4 of their textbook, Jennison and Turnbull (2000) reported simulation results to examine the mean total sample size, type I error rate, and power of the unadjusted test for comparing two means. For their example, they set $\alpha_t = 0.05$, $P_t = 0.90$, and $\theta_* = 1$. They considered five different scenarios for the variance, but in each case they assume that the prior estimate of the variance is equal to the true, unknown value. For these five scenarios, a fixed sample size power calculation suggests 8, 12, 23, 44, and 86 subjects per group, respectively. They considered an internal pilot design with varying fractions of the originally planned subjects included in the internal pilot sample, no finite maximum sample size, and allowed the final sample size to be reduced if the variance was originally overstated.

Program 11 invokes the `RUN GLUMIP;` command to reproduce the type I error rate calculations reported for Table 14.4 in the Jennison and Turnbull (2000) text. Program 11 generates the following output.

To reproduce the power calculations for Table 14.4 in Jennison and Turnbull (2000), the only change that is required to the above code is to replace the statement `BETA_ALT = {0 0}';` with the statement `BETA_ALT = {0 1}';`. The code (Program 12) generates the following output.

3.5. 3-Group ANOVA Example from Coffey and Muller (1999) paper

To illustrate the flexibility and usefulness of the **GLUMIP** software package, we consider a three-group analysis of variance (ANOVA) design described as Example C in Coffey and Muller (1999). For the two degree of freedom test of differences among groups with α_t = 0.05, P_t = 0.90, σ^2 = 1, and θ_* = [0.5 1.0]′, a fixed sample calculation suggests 27 observations per group. To mirror the results found in Coffey and Muller (1999), we consider the three-group ANOVA example for an internal pilot with the following specifications: 1) a pre-planned sample size of n_0 = 81, 2) the first n_1 = 39 (13 per group) observations comprise the internal pilot sample (`N1=39`), 3) both allowing and not allowing the final sample size to decrease if the original variance value overestimates the true variance (`NPLUSMIN=39` and `NPLUSMIN=81`, respectively), 4) no finite upper bound of observations, 5) using Unadjusted Method for sample size re-estimation and testing (`RULE=0` and `TEST=0`). Program 13 invokes the `RUN GLUMIP;` command to reproduce the type I error rate results found in Coffey and Muller (1999, Table V).

Program 13 generates the following output.

To reproduce the power calculations for Table V in Coffey and Muller (1999), the only change that is required to the above code is to replace the statement `BETA_ALT = {0 0 0} ';` with the statement `BETA_ALT = {0 0.5 1.0}';`. The code (Program 14) generates the following output.

The following program (Program 15) illustrates the use of the new (`N2CALC`) module for the three-group ANOVA example. The program will consider sample size calculations for various first sample variance estimates (`SIGHAT1`) using either the unadjusted or the Stein-like sample size re-estimation rules (`RULE=0 or 1`, respectively).

Program 15 generates the following output, giving needed second stage sample sizes.

4. Discussion

The results in Coffey and Muller (2001) extend many internal pilot design concepts from the *t* test setting to the classic general linear univariate model setting. The introduction of the **GLUMIP** program further extends the practicability of internal pilot designs in two ways:

1. Researchers can examine the maximum inflation of type I error rate possible if an internal pilot is used with the unadjusted test. This can allow

a very quick and early determination as to whether type I error rate inflation should be a concern in any particular study.

2. When there is reason to be concerned about type I error rate inflation, investigators have several options as to how to control type I error rate and maintain the benefits of an internal pilot design. **GLUMIP** can allow investigators to examine the impact on power and sample size for each choice and make a decision to maximize the benefits of the internal pilot design for their particular study.

Furthermore, the new and much simpler density and cumulative distribution function forms included in this updated version make the bounding method more practical and convenient even in very small samples where it has the most value. The speed also allows quickly plotting power and expected sample size over wide ranges of design parameters. The ability to produce such plots in a timely manner has many advantages. For example, there is an obvious trade-off between choosing a large enough value of n_1 to ensure a reliable estimate of σ^2 and choosing the smallest possible value of n_1 such that modifications to sample size, and any corresponding logistical changes to the on-going study, can be implemented as early as possible. Historically, completely arbitrary values, such as $n_1 = n_0/2$, have been chosen. Examining plots of power over a wide range of design parameters can be used to determine the smallest value of n_1 which retains the desired benefits of an internal pilot.

The **GLUMIP** program is intended to aid in study planning when incorporating an internal pilot design. While the inclusion of the `N2CALC` facilitates study implementation, the current software does not fully support study analysis. However, with **GLUMIP** and careful use of other standard software, analysis can be implemented as follows.

1. Use **GLUMIP** to explore possible design parameters with respect to sample size, power, and type I error rate properties in order to determine appropriate design.
2. Use software such as **POWERLIB** (Johnson *et al.* 2008) to determine necessary fixed study sample size
3. Sample subjects needed for first stage according to design
4. Use N2CALC to calculate necessary second stage sample size according to sample size update rule (`RULE`) chosen at design stage
5. Perform usual analysis at final stage using desired variance estimate and testing procedure (`TEST`) chosen at design stage

   Step 5 is straightforward with the unadjusted (`TEST=0`) or bounding (`TEST=3`) tests since they use the usual test statistic and critical value calculations involve only traditional forms using either the preplanned α (`TEST=0`) or the adjusted α_* predetermined using the `RUN _FINDADJ;` command (`TEST=3` in **GLUMIP**; see examples in Section 3.2). Implementation for the other three testing methods included for planning purposes in **GLUMIP** 2.0 (`TEST=1`, `TEST=2` or `TEST=4`) is more difficult.

Since they do not use the usual test statistic in final analysis, the correct variance and degree of freedom components must be carefully picked off from appropriate locations and combined to create the necessary statistic and critical values. These testing methods would have the most to gain from the availability of additional analysis software.

Although we provide valid arguments for utilizing the **GLUMIP** 2.0 software for internal pilot designs, users must be cautioned that this is still a work in progress. The user should consult the appropriate articles in the reference list to understand the limitations of the program.

There are several ways in which we hope to improve the software in the future:

- The current version includes only a subset of proposed methods for controlling the type I error rate in internal pilot designs. The versatility of the package will continue to grow by incorporating alternative methods into the software. This would especially include methods employing blinded sample size re-estimation techniques that have great appeal when mitigation of potential bias is crucial.
- The software (and internal pilot results in general) are limited somewhat in that they have only been studied in the univariate setting. We are currently exploring ways to extend the internal pilot results to the repeated measures setting. As these methods are developed and published, we envision that a more general version of the software will become available with the univariate algorithms occurring as special cases.
- The **GLUMIP** program does not currently perform full data analysis for internal pilot designs. We plan to simplify this task by including additional software for data analysis under an internal pilot in later versions of **GLUMIP**.
- Allow additional flexibility of designs. For one example, the program could allow for potential changes in variance over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NCI R01 CA095749. The authors would like to thank the anonymous reviewers for comments during the review process which greatly improved the manual and software.

References

Burman CF, Sonesson C. Are Flexible Designs Sound? Biometrics. 2006; 62(3):664–669. [PubMed: 16984302]

Coffey CS, Kairalla JA, Muller KE. Practical Methods for Bounding Type I Error Rate with an Internal Pilot Design. Communications in Statistics – Theory and Methods. 2007; 36:2143–2157.

Coffey CS, Muller KE. Exact Test Size and Power of a Gaussian Error Linear Model for an Internal Pilot Study. Statistics in Medicine. 1999; 18(10):1199–1214. [PubMed: 10363340]

Coffey CS, Muller KE. Properties of Doubly-Truncated Gamma Variables. Communications in Statistics – Theory and Methods. 2000a; 29(4):851–857. [PubMed: 24465079]

Coffey CS, Muller KE. Some Distributions and Their Implications for an Internal Pilot Study with a Univariate Linear Model. Communications in Statistics – Theory and Methods. 2000b; 29(12): 2677–2691.

Coffey CS, Muller KE. Controlling Test Size While Gaining the Benefits of an Internal Pilot Design. Biometrics. 2001; 57(2):625–631. [PubMed: 11414593]

Cytel Inc. **East** 5: Clinical Trial Design System. 2007. URL http://www.cytel.com/Products/East/

Davies RB. The Distribution of a Linear Combination of Chi-Squared Random Variables. Applied Statistics. 1980; 29:323–333.

Davis, R.; Rabinowitz, P. Methods of Numerical Integration. Academic Press; New York: 1984.

Denne JS, Jennison C. Estimating the Sample Size for a t-test Using an Internal Pilot. Statistics in Medicine. 1999; 18(13):1575–1585. [PubMed: 10407230]

Dupont, WD.; Plummer, WD. **PS**: Power and Sample Size Calculation. 2004. URL http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

Friede T, Kieser M. Sample Size Recalculation in Internal Pilot Study Designs: A Review. Biometrical Journal. 2006; 48(4):537–555. [PubMed: 16972704]

Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive Designs in Clinical Drug Development – An Executive Summary of the PhRMA Working Group. Journal of Biopharmaceutical Statistics. 2006; 16:275–283. [PubMed: 16724485]

Gould AL, Shih WL. Sample-Size Reestimation Without Unblinding for Normally Distributed Outcomes with Unknown Variance. Communications in Statistics – Theory and Methods. 1992; 21(10):2833–2853.

Helms RW. Comparisons of Parameter and Hypothesis Definitions in a General Linear Model. Communications in Statistics – Theory and Methods. 1988; 17(8):2725–2753.

ICH Guideline E9. Statistical Principles for Clinical Trials. The Federal Register. 1998; 63(179): 49583–49598. [PubMed: 10185190]

Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC; Boca Raton: 2000.

Johnson JL, Muller KE, Slaughter JC, Gurka MJ, Gribbin MJ, Simpson SL. **POWERLIB**: SAS/IML Software for Computing Power in Multivariate Linear Models. Journal of Statistical Software. 2008 Forthcoming.

Kiefer J. Sequential Minimax Search for a Maximum. Proceedings of the American Mathematical Society. 1953; 4:502–506.

Kieser M, Friede T. Re-Calculating the Sample Size in Internal Pilot Study Designs with Control of The Type I Error Rate. Statistics in Medicine. 2000; 19(7):901–911. [PubMed: 10750058]

Kieser M, Friede T. Simple Procedures for Blinded Sample Size Adjustment that Do not Affect the Type I Error Rate. Statistics in Medicine. 2003; 22(23):3571–3581. [PubMed: 14652861]

Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. Biometrics. 1999; 55(4):1286–1290. [PubMed: 11315085]

Miller F. Variance Estimation in Clinical Studies with Interim Sample Size Reestimation. Biometrics. 2005; 61(2):355–361. [PubMed: 16011681]

NCSS. **PASS**: Power Analysis and Sample Size. 2005. URL http://www.ncss.com/pass.html

Pisano ED, Zong SQ, Hemminger BM, DeLuca M, Johnston RE, Muller K, Braeuning MP, Pizer SM. Contrast Limited Adaptive Histogram Equalization Image Processing to Improve the Detection of Simulated Spiculations in Dense Mammograms. Journal of Digital Imaging. 1998; 11(4):193–200. [PubMed: 9848052]

Proschan MA. Two-Stage Sample Size Re-Estimation Based on a Nuisance Parameter: A Review. Pharmaceutical Statistics. 2005; 15:559–574.

Proschan MA, Wittes J. An Improved Double Sampling Procedure Based on the Variance. Biometrics. 2000; 56(4):1183–1187. [PubMed: 11129477]

Shih WJ. Group Sequential, Sample Size Re-Estimation and Two-Stage Adaptive Designs in Clinical Trials: A Comparison. Statistics in Medicine. 2006; 25(4):933–941. [PubMed: 16220505]

Statistical Solutions. **nQuery** Advisor. 2005. URL http://www.statsol.ie/

Stein C. A Two-Sample Test for a Linear Hypothesis whose Power is Independent of the Variance. Annals of Mathematical Statistics. 1945; 16:43–58.

Tsiatis AA. Information-Based Monitoring of Clinical Trials. Statistics in Medicine. 2006; 25:3236–3244. [PubMed: 16927248]

Wang S, Xia J, Yu L, Li C, Xu L. A SAS Macro for Sample Size Adjustment and Randomization Test for Internal Pilot Study. Computer Methods and Programs in Biomedicine. 2008; 90:66–88. [PubMed: 18192069]

Wassmer, G.; Eisebitt, R. **ADDPLAN**: Adaptive Designs – Plans and Analyses (Release 4.0). 2007. URL http://www.addplan.com/

Wassmer G, Vansemeulebroeck M. A Brief Review on Software Developments for Group Sequential and Adaptive Designs. Biometrical Journal. 2006; 48(4):732–737. [PubMed: 16972726]

Wittes J, Brittain E. The Role of Internal Pilot-Studies in Increasing the Efficiency of Clinical-Trials. Statistics in Medicine. 1990; 9(1-2):65–72. [PubMed: 2345839]

Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal Pilot Studies II: Comparison of Various Procedures. Statistics in Medicine. 1999; 18(24):3493–3509. [PubMed: 10611621]

Table 1

Program inputs (with notation from Coffey and Muller 2001).

GLUMIP	Symbol	Dimensions	Description
Required design parameters:			
ESSENCEX	$E_s(\boldsymbol{X}_+)$	$g \times q$	Essence design matrix
ALPHAT	α_t	1×1	Target type I error rate
POWERT	P_t	1×1	Target power
C	C	$a \times q$	Contrast matrix
BETA_PLN	β_*	$q \times 1$	Primary parameters for power calculations
SIGMAO	σ_0^2	1×1	Planning variance estimate
Required fixed parameters:			
GAMLIST	γ	$1 \times h$	List of gamma values: $\gamma = \sigma^2 \Big/ \sigma_0^2$
BETA_ALT	β	$q \times 1$	True primary parameters
Required internal pilot rules:			
N1	n_1	1×1	Internal pilot sample size
NPLUSMIN	$n_{+,\min}$	1×1	Minimum final sample size
RULE		1×1	Sample size re-estimation rule
TEST		1×1	Testing method
Optional inputs:			
WEIGHTS		$g \times 1$	Weights for rows of Es $(\boldsymbol{X}_+)$ per replication
NPLUSMAX	$n_{+,\max}$	1×1	Maximum final sample size
ROUND		1×1	Places to round N_+ pdf relative to α_t (1-10)