Published in final edited form as:

Stat Med. 2018 February 10; 37(3): 375–389. doi:10.1002/sim.7524.

Internal Pilot Design for Balanced Repeated Measures

Xinrui Zhang¹, Keith E. Muller², Maureen M. Goodenow³, and Yueh-Yun Chi^{*,4}

¹Novartis Pharmaceuticals Corporation, East Hanover, NJ

²Department of Health Outcomes and Policy, University of Florida, Gainesville, FL

³National Institutes of Health, Bethesda, MD

⁴Department of Biostatistics, University of Florida, Gainesville, FL

Abstract

Repeated measures are common in clinical trials and epidemiological studies. Designing studies with repeated measures requires reasonably accurate specifications of the variances and correlations in order to select an appropriate sample size. Underspecifying the variances leads to a sample size that is inadequate to detect a meaningful scientific difference, while over-specifying the variances results in an unnecessarily large sample size. Both lead to wasting resources and placing study participants in unwarranted risk. An internal pilot design allows sample size recalculation based on estimates of the nuisance parameters in the covariance matrix. We provide the theoretical results that account for the stochastic nature of the final sample size in a common class of linear mixed models. The results are useful for designing studies with repeated measures and balanced design. Simulations examine the impact of misspecification of the covariance matrix and demonstrate the accuracy of the approximations in controlling the Type I error rate and achieving the target power. The proposed methods are applied to a longitudinal study assessing early antiretroviral therapy for youth living with HIV.

Keywords

sample size; linear mixed model; UNIREP; power; fixed effects inference

1 Introduction

1.1 Motivation

Sample size determination plays a critical role in planning a clinical trial and epidemiological study. A sample size too small can lead to imprecise estimates and lack of power to detect clinically relevant effects. On the other hand, a sample size unnecessarily large can waste time and resources. For Gaussian outcomes, one of the required inputs for sample size determination is the nuisance parameters in the covariance matrix. For example, a study comparing mean responses of two groups requires an accurate guess of the error variances. Values of the nuisance parameters are generally uncertain in the planning phase.

*Correspondence to: Yueh-Yun Chi, Department of Biostatistics, University of Florida, Gainesville, FL. yychi@ufl.edu.

Common practices include using values obtained from the literature, a pilot study, or experts' opinions.

To avoid the uncertainty in finding valid values for the nuisance parameters, a variety of two-stage procedures have been proposed.^{1,2} One of the two-stage procedures uses the stage 1 (internal pilot stage) data to reassess the nuisance parameters and the sample size. An internal pilot design does not allow evaluating treatment effects using interim data and a decision to stop the study early. Sample size recalculation is based on the revised estimates of the nuisance parameters. The recalculation makes the final sample size a random variable. In turn, classical results assuming a fixed sample size no longer hold. Reference distributions of a test statistic have to be adjusted to account for the distribution of the sample size.

This paper contains theoretical results for fixed effects inference when an internal pilot design is used to measure outcomes repeatedly over time. With a design balanced within independent sampling units, e.g., across time, the underlying general linear mixed model can be written as a general linear multivariate model (Muller and Stewart,³ Chapter 12). A balanced design requires no missing or mistimed response data. Ringham *et al.*⁴ and Muller *et al.*⁵ highlighted common scenarios in imaging research and biomarker studies, in which such a design is appropriate for assessments taken at pre-defined time points. In the simulations, the theoretical results are evaluated by examining the empirical power and Type I error rate.

1.2 Internal Pilot Design

An internal pilot design has two stages for studies with balanced repeated measures of Gaussian outcomes. Stage 1 is known as the “internal pilot stage” when the initial power analysis, based on a prior guess of the error covariance matrix Σ_0 , is performed, and the internal pilot data are collected. Here, n_0 is the initial sample size, and π ($0 < \pi < 1$) is the pre-specified proportion to determine the size of the internal pilot sample ($n_1 = \pi n_0$). At the

end of the stage 1, the error covariance matrix $\hat{\Sigma}_1$ can be estimated and used to perform the second power analysis at stage 2. The second power analysis is based on the same specifications of the initial power analysis, except with a revised error covariance matrix $\hat{\Sigma}_1$.

Two rules are commonly used to determine the final sample size based on the second power analysis. Wittes and Brittain² required the final sample size to be at least as large as the initially planned sample size (i.e., $N_+ = n_0$). Birkett and Day⁶ noted that the resulting “restricted rule” may lead to an unnecessarily large sample size and high power if the prior guess overestimates the error variance. In turn, they recommended using the “unrestricted rule,” to only require the final sample size to be no less than the size of the stage 1 sample (i.e., $N_+ = n_1$).

Once the final sample size N_+ is determined, $N_2 = N_+ - n_1$ additional subjects are to be enrolled during stage 2. At the end of study, all N_+ observations will be used to compute the final test statistic and test the primary hypothesis. We note that both N_+ and N_2 are random

variables. Classic results for the final test statistic no longer hold because of the randomness in N_{j+} .

1.3 Literature Review

Most research for the internal pilot design has been focused on the two-sample t test. Wittes and Brittain² used a “naïve” t test with the restricted rule. It was called “naïve” because the distribution of t -statistic was not modified to account for the randomness in N_{j+} . Birkett and Day⁶ examined the same test with the unrestricted rule. Both reported that the naïve t -test inflates the Type I error rate, especially when the size of the internal pilot sample is small ($n_1 < 20$). The inflation of the Type I error rate is caused by the systematic underestimation of the variance.⁷ Various approaches have been proposed to control the Type I error rate using alternative variance estimators,^{8–10} a downweighted α -level,⁷ or an adjusted critical value.¹¹

Coffey and Muller^{12,13} applied the internal pilot design to the general linear univariate model, of which the two-sample t -test is a special case. They proposed using a bounding test, which adjusts the critical value to control the Type I error rate to be no greater than the targeted level. They provided extensive enumeration results to examine the effects of the restricted and unrestricted rule, the size of the initial sample and internal pilot sample. Their results suggested the bounding test protects power and controls the Type I error rate, especially when the sample size is small or moderate.

Few authors have discussed internal pilots for studies with repeated measures. Zucker and Denne¹⁴ developed a procedure for two-arm repeated measures studies under the general linear mixed model framework. Their results are restricted to highly structured covariance patterns. Coffey and Muller¹⁵ examined misspecified covariances for fixed and internal pilot designs for the general linear multivariate model. They showed that the internal pilot design preserves power better than the fixed design when the covariance matrix is misspecified. They also pointed out that, as in the univariate setting, internal pilot designs with repeated measures can inflate the Type I error rate. All internal pilot results reported by Coffey and Muller¹⁵ were based on simulations. Using simulations for the internal pilot power calculations can be computationally intensive.

The work here fills a gap in the literature by providing analytical and numerical results for an internal pilot design with balanced repeated measures. The rest of the paper is organized as follows. In Section 2, we introduce the motivating example, a clinical trial investigating early treatment effect in youth living with HIV. In Section 3, we review the existing methods that can be used to approximate power when fixed effects inference for a common class of linear mixed models is of interest. In Section 4, we extend the results and derive approximate distributions when the final sample size becomes a random variable under an internal pilot design. In Section 5, we use simulations to evaluate the accuracy of our proposed methods. In Section 6, we apply the new methods to the motivating example. In Section 7, we discuss strengths and limitations of the current work and provide areas of further research. All mathematical proofs are included in the Appendix.

2 Motivating Example

Human immunodeficiency virus type 1 (HIV-1) infection causes depleted CD4+ T cells and altered immune function. With the advent of antiretroviral therapy (ART), HIV-1 infection has become a manageable chronic disease. The threshold of CD4+ count used to decide when to begin ART in youth living with HIV (YLWH) has gradually increased in the last five years.¹⁶ Although several randomized studies have established the benefit of early ART in reducing long-term HIV-related events,¹⁷ the impact of early treatment on inflammatory biomarkers remains unclear. A planned study will investigate the impact of early treatment on inflammatory biomarkers soluble CD14 and CD163 within YLWH. Information about mean differences and correlations for the two investigated biomarkers are obtained from a previous study sponsored by the Adolescent Trials Network (ATN) for HIV/AIDS.

Figure 1 displays the observed means and correlations of the two biomarkers for 76 YLWH. The existing data will be used to guide the planning and power analysis. The observed means (solid lines, Figure 1A) are used to determine the treatment differences of interest (dashed lines). The observed correlation matrices (Figure 1B) are used to permit examining the effects of misspecified covariances. We will revisit this example in section 6.

3 Balanced Linear Mixed Models

3.1 General Linear Multivariate Model as an Equivalent Model

Considering a study with N subjects and p repeated measures for each subject, we follow the notation used in Muller and Stewart³ to define the “population-average” form of a balanced

linear mixed model as $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i$, where $i \in \{1, 2, \dots, N\}$. The aggregated error term $\mathbf{e}_i \sim \mathcal{N}(0, \sum_i)$ is a function of the subject-specific random effects and the pure error term. Muller *et al.*⁵ pointed out a linear mixed model is equivalent to a general linear multivariate model when the following conditions are satisfied: (i) no missing or mistimed data, (ii) no time-varying covariates, and (iii) a common unstructured error covariance matrix for all subjects, i.e., $\sum_i = \Sigma$ for all i . The equivalent multivariate model can be written as $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, with the i th row of $\mathbf{Y}$ and $\mathbf{E}$ being $\mathbf{y}_i$ and $\mathbf{e}_i$ respectively. The distributional assumption that $\mathbf{e}_i \sim \mathcal{N}(0, \sum_i)$ independently leads to a matrix Gaussian distribution for $\mathbf{Y}$ with $\mathbf{Y} \sim \mathcal{N}_{N,p}(\mathbf{XB}, \mathbf{I}_N, \sum_i)$.³ Model degrees of freedom are denoted by $\nu_e = N - \text{rank}(\mathbf{X})$.

With model equivalence, a general linear hypothesis about fixed effects in a linear mixed model can be expressed as a multivariate general linear hypothesis when testing between-subject effect, within-subject effect, or their interaction. With secondary parameters

$\boldsymbol{\Theta} = \begin{bmatrix} \mathbf{C} & \mathbf{B} & \mathbf{U} \end{bmatrix}$, a multivariate general linear hypothesis can be stated as $H_0: \boldsymbol{\Theta} = \mathbf{C}\mathbf{B}\mathbf{U} = \boldsymbol{\Theta}_0$. Matrices $\mathbf{C}$ and $\mathbf{U}$ contain the between-subject and within-subject contrasts, respectively. Each row of $\mathbf{C}$ corresponds to a contrast among the predictors in $\mathbf{X}$. Each column of $\mathbf{U}$ corresponds to a contrast among the p repeated measures, which collectively defines the transformed responses $\mathbf{Y}_U = \mathbf{Y}\mathbf{U}$, with $\mathbf{Y}_U \sim \mathcal{N}_{N,b}(\mathbf{X}\mathbf{B}\mathbf{U}, \mathbf{I}_N, \sum_*)$ and $\Sigma_* = \mathbf{U}$

ΣU . We focus on testable hypotheses (Muller and Stewart,³ Chapter 16). Model parameter estimates are described in the Appendix.

3.2 Existing Results for Multivariate Tests Equivalent to Mixed Model Tests

Recognizing the equivalence between the linear mixed model and multivariate model allows us to focus on the univariate approach to repeated measures (UNIREF) test statistic for the multivariate model. A detailed account of the UNIREF tests can be found in the Appendix. The UNIREF test statistic is a ratio of traces, i.e., $T_u = [\text{tr}(\mathbf{S}_h)/a]/[\text{tr}(\mathbf{S}_e)/\nu_e]$, where $\mathbf{S}_h$ is the hypothesis sum of squares matrix and $\mathbf{S}_e$ is the error sum of squares matrix. The five UNIREF tests, Box, Geisser-Greenhouse (GG), Chi-Muller (CM), Hunyh-Feldt (HF) and uncorrected (UN), all use the test statistic T_u . The distribution of T_u depends on the

sphericity parameter $\varepsilon = \text{tr}^2(\sum_*) / [b \text{tr}(\sum_*^2)]$, which measures the spread of eigenvalues of the hypothesis covariance matrix $\Sigma_* = U' \Sigma U$. When all the eigenvalues are equal, ε achieves its upper bound ($\varepsilon = 1$), and the assumption of sphericity is met. Conversely, minimum sphericity occurs when only one eigenvalue is nonzero, and ε achieves its lower bound ($\varepsilon = 1/b$). With sphericity ($\varepsilon = 1$), test statistic T_u follows an F distribution exactly. When the sphericity assumption is not met ($\varepsilon \neq 1$), the test statistic T_u follows an F distribution approximately with $T_u \sim F[ab\varepsilon, b\nu_e\varepsilon; \omega_T \cdot \varepsilon]$.

The UNIREF tests differ by their approaches to estimate ε , which result in differences in the Type I error rate and power (see Appendix for details). The Box test, which is the most conservative of the five, uses the lower bound of ε . The UN test, which is the least conservative, uses the upper bound of ε . When sphericity ($\varepsilon = 1$) is present, the UN test is size α and uniformly most powerful (among similarly invariant tests). The GG, HF and CM tests estimate the sphericity parameter from the data. The GG test uses the MLE of ε . The HF test uses a ratio of unbiased estimators. Gribbin *et al.*¹⁸ proved the unbiasedness holds only when $\text{rank}(X) = 1$ and proposed using the rank-adjusted HF test to apply to the $\text{rank}(X)$

1 cases. In this paper, the HF test is referred as the rank-adjusted HF test. The CM test was developed to extend the UNIREF approach to the high dimension, low sample size setting. The Type I error rate and power for the five UNIREF tests are ordered (Box $\leq$ GG $\leq$ CM $\leq$ HF $\leq$ UN).

Gribbin *et al.*¹⁸ approximated power of the UNIREF tests for the fixed design when the covariance Σ_* is known:

$$P \approx 1 - F_F \left[F_F^{-1}(1 - \alpha; ab \cdot e, b\nu_e \cdot e); ab \cdot \varepsilon_n, b\nu_e \cdot \varepsilon, \omega_* \right] \quad (1)$$

Here, α and P denote the Type I error rate and power. The different choice of multiplier e leads to different critical values for the UNIREF tests (see Appendix for details). The numerator degree of freedom ($ab \cdot \varepsilon_n$) is a function of the parameter $\varepsilon_n = \bar{\lambda}/\lambda_*$. Here $\bar{\lambda} = \text{tr}(\sum_*)/b$ is the average eigenvalue of Σ_* , and λ_* is defined as

$\lambda_* = [tr(\sum_*^2) + 2tr(\sum_* \Delta/a)] / [tr(\sum_*) + 2tr(\Delta/a)]$. The noncentrality $\omega_* = tr(\lambda_*)/\lambda_*$ is a function of known covariance and means.

3.3 Model Properties Under an Internal Pilot Design

By satisfying the conditions specified in Section 3.1, multivariate model equivalent to the mixed model of interest can be found under the two-stage internal pilot design. Table 1 summarizes the notation. Equation (2) expresses the equivalent multivariate model with subscripts 1, 2 and +, representing the stage 1, 2 and final samples, respectively.

$$\begin{bmatrix} \mathbf{Y}_1 \\ (n_1 \times p) \\ \mathbf{Y}_2 \\ (N_2 \times p) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ (n_1 \times q) \\ \mathbf{X}_2 \\ (N_2 \times q) \end{bmatrix} \mathbf{B}_+ + \begin{bmatrix} \mathbf{E}_1 \\ (n_1 \times p) \\ \mathbf{E}_2 \\ (N_2 \times p) \end{bmatrix} \quad \mathbf{Y}_+ = \mathbf{X}_+ \mathbf{B}_+ + \mathbf{E}_+ \quad (2)$$

We assume the design matrix in different stages of an internal pilot design has the same rank, i.e., $\text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X}_2) = \text{rank}(\mathbf{X}_+) = r$. Consequently, $\text{Es}(\mathbf{X}_1) = \text{Es}(\mathbf{X}_2) = \text{Es}(\mathbf{X}_+)$, in which the essence matrix $\text{Es}(\mathbf{X}_+)$ is defined by deleting any duplicate rows from the design matrix $\mathbf{X}_+$. Following Coffey and Muller,¹² we restrict attention to the final sample size being an integer multiple of the number of rows in the essence matrix. With $N_+ = n_+$,

$\mathbf{X}_+ \mathbf{X}_+ = n_+ [\text{Es}(\mathbf{X}_+)' \mathbf{W}_+ \text{Es}(\mathbf{X}_+)]$ and $\text{tr}(\mathbf{X}_+ \mathbf{X}_+) = n_+ \text{tr}(\text{Es}(\mathbf{X}_+))$. Here, $\text{Es}(\mathbf{X}_+)' \mathbf{W}_+ \text{Es}(\mathbf{X}_+)$ and $\text{Es}(\mathbf{X}_+) = (\Theta - \Theta_0)' \{C[\text{Es}(\mathbf{X}_+)' \mathbf{W}_+ \text{Es}(\mathbf{X}_+)]^{-1} C'\}^{-1} (\Theta - \Theta_0)$ are functionally independent of n_+ .

Given $N_+ = n_+$, $\mathbf{W}_+$ ($r \times r$) is the diagonal weight matrix with elements $(n_{+g}/n_+, g \in \{1, \dots, r\})$, the proportions of the subgroup sample sizes to the total sample size. For a balanced design, $\mathbf{W}_+ = (1/r) \cdot \mathbf{I}_r$. We also assume the subgroup sample size proportions are the same between stage 1 and 2 samples, i.e., $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}_+$.

4 New Approximations for an Internal Pilot Design

4.1 Distribution of the Final Sample Size

In this section, we first provide an approximation for the distribution of the final sample size N_+ . Second, we derive the approximate conditional distribution of the final test statistic for a particular realization of $N_+ = n_+$. Finally, we obtain the marginal distribution of the final test statistic by using the law of total probability. As summarized in Table 1, estimates from stage 1, 2 and final samples are denoted by subscript 1, 2 and +, respectively. All proofs are included in the Appendix. In an internal pilot study, observations in the stage 1 sample

provide an estimate of the covariance matrix $\sum_{*1} = \mathbf{U}' \sum_1 \mathbf{U}$. After stage 1, the second power analysis is conducted to determine the final sample size N_+ . For a particular realization of $N_+ = n_+$, we can find the noncentrality $\omega_*(n_+)$ which satisfies

$$P_t \approx 1 - F_F \left[F_F^{-1} (1 - \alpha_t; ab \cdot e_1, b\nu_+ \cdot e_1); ab \cdot \varepsilon_n, b\nu_+ \cdot \varepsilon, \omega_*(n_+) \right] \quad (3)$$

with e_1 defined as in Section 3.2 but with degrees of freedom $\nu_1 = n_1 - r$. Here, α_t and P_t denote the target Type I error rate and target power. Since the cumulative distribution function of an F random variable is monotonically decreasing in terms of the noncentrality, a unique solution of $\omega_*(n_+)$ can be obtained given $N_+ = n_+$. In turn, $\lambda_*(n_+)$, a function of $\omega_*(n_+)$, can be determined by

$$\lambda_*(n_+) = \frac{tr(\Delta)}{\omega_*(n_+)} = \frac{n_+ tr(\Delta_{Es})}{\omega_*(n_+)}. \quad (4)$$

We propose using the distribution of the λ_* estimator to approximate the distribution of N_+ . This approach is in parallel to the univariate setting presented by Coffey and Muller,¹² which leads to $Pr\{N_+ \leq n_+\} = Pr\{\hat{\sigma}_1^2 \leq \sigma^2(n_+)\}$. In the univariate testing ($b = 1$), $\lambda_*(n_+) = \sigma^2(n_+)$.

Results in Gribbin *et al.*¹⁸ lead to using $\tilde{\lambda}_{*1} = \hat{\lambda}_1 / \tilde{\varepsilon}_{n1}$, a ratio obtained from the stage 1

sample, to estimate $\lambda_* = \bar{\lambda} / \varepsilon_n$. The numerator $\hat{\lambda} = tr\left(\sum_{*1}\right) / b$ is an unbiased estimator of $\bar{\lambda}$, while the denominator

$\tilde{\varepsilon}_{n1} = [tr^2(\sum_{*1}) + 2tr(\sum_{*1})tr(\Delta/a)] / \{b[tr(\sum_{*1}^2) + 2tr(\sum_{*1}\Delta/a)]\}$ is a biased estimator of ε_n . The distribution of $\tilde{\lambda}_{*1}$ can be approximated by a chi-square distribution with $\tilde{\lambda}_{*1}\tilde{\nu}_* / \lambda_* \sim \chi^2(\tilde{\nu}_*)$ using the Satterthwaite approximation and degrees of freedom $\tilde{\nu}_* = (b\nu_1) \hat{\varepsilon}_1 / \tilde{\varepsilon}_{n1}$.

For power calculation, we consider approximating $\tilde{\nu}_*$ by $\nu_* = (b\nu_1) E(\hat{\varepsilon}_1) / E(\tilde{\varepsilon}_{n1})$. Muller *et al.*⁵ provided approximation of $E(\hat{\varepsilon}_1)$ (refer to Appendix for details), we use the same approach to approximate $E(\tilde{\varepsilon}_{n1})$ by

$$E(\tilde{\varepsilon}_{n1}) \approx \frac{E[tr^2(\sum_{*1})] + 2E[tr(\sum_{*1})tr(\Delta/a)]}{b\{E[tr(\sum_{*1}^2)] + 2E[tr(\sum_{*1}\Delta/a)]\}} = \frac{E[tr^2(\sum_{*1})] + 2tr(\sum_{*1})tr(\Delta/a)}{b\{E[tr(\sum_{*1}^2)] + 2tr(\sum_{*1}\Delta/a)\}}. \quad (5)$$

The equality in Equation (5) holds because Gribbin *et al.*¹⁸ proved $tr\left(\sum_{*1}\right)$ and $tr\left(\sum_{*1}\Delta\right)$ are unbiased estimators of $tr(\Sigma_*)$ and $tr(\Sigma_* \Delta)$.

In an internal pilot study, the n_1 observations of stage 1 sample are used to obtain the

covariance estimate $\sum_{*1}$, and subsequently compute $\tilde{\lambda}_{*1} = \hat{\lambda}_1 / \tilde{\varepsilon}_{n1}$. If the value of $\tilde{\lambda}_{*1}$ is less or equal to $\lambda_*(n_+)$, the final sample size would be no greater than n_+ . Hence the distribution of N_+ can be approximated by the distribution of $\tilde{\lambda}_{*1}$, namely

$$Pr\{N_+ \leq n_+\} \approx Pr\{\tilde{\lambda}_{*1} \leq \lambda_*(n_+)\} \approx Pr\left\{\chi^2(\nu_*) \leq \frac{\nu_* n_+ \text{tr}(\Delta_{E_s})}{\lambda_* \omega_*(n_+)}\right\}. \quad (6)$$

The discrete nature of N_+ leads to

$$Pr\{N_+ = n_+\} = Pr\{N_+ \leq n_+\} - Pr\{N_+ \leq n_+ - 1\} \approx Pr\left\{\lambda_*(n_+ - 1) < \tilde{\lambda}_{*1} \leq \lambda_*(n_+)\right\}. \quad (7)$$

Considering the special case in which the testing is univariate ($b = 1$), we prove Lemma 1 to show Equation (6) reduces to Equation (17) in Coffey and Muller¹² and the proposed methods reduce to the exact distribution derived by Coffey and Muller for the univariate case.

Lemma 1—For univariate testing, i.e., $\text{rank}(\mathbf{U}) = b = 1$, $\lambda_* = \sigma^2$, $\lambda_*(n_+) = \sigma^2(n_+)$, and $\nu_* = \nu_1 = n_1 - r$.

4.2 Conditional Distribution of the Final Test Statistic

The results throughout this subsection are derived under the condition that $N_+ = n_+$, i.e., when the final sample size equals a particular realization. Let $I(n_+)$ represent the probability of failing to reject H_0 given $N_+ = n_+$:

$$I(n_+) = Pr\{T_u \leq f_{\text{crit}}(n_+) | N_+ = n_+\} = Pr\left\{\frac{\text{tr}(\mathbf{S}_{h+})/a}{\text{tr}(\mathbf{S}_{e+})/\nu_+} \leq f_{\text{crit}}(n_+) | N_+ = n_+\right\}. \quad (8)$$

Here $f_{\text{crit}}(n_+) = F_F^{-1}(1 - \alpha_t; ab \cdot e_+, b\nu_+ \cdot e_+)$ is the critical value and $\nu_+ = n_+ - r$ is the degrees of freedom of the final sample.

For the fixed design, Chi *et al.*¹⁹ proved that the hypothesis sum of squares matrix $\mathbf{S}_h$ and error sum of squares matrix $\mathbf{S}_e$ are independent. We prove Lemma 2 to show a variety of independent properties for matrices in an internal pilot study.

Lemma 2—For an internal pilot design, given $N_+ = n_+$, the following pairs of matrices (and their traces) are independent: $\mathbf{S}_{h+}$ and $\mathbf{S}_{e+}$; $\mathbf{S}_{h+}$ and $\mathbf{S}_{e1}$; $\mathbf{S}_{h+}$ and $(\mathbf{S}_{e+} - \mathbf{S}_{e1})$; $\mathbf{S}_{e1}$ and $(\mathbf{S}_{e+} - \mathbf{S}_{e1})$.

For the fixed design, Muller *et al.*⁵ reported accurate approximations of $\text{tr}(\mathbf{S}_h)$ and $\text{tr}(\mathbf{S}_e)$ with $(\varepsilon_n/\bar{\lambda}) \text{tr}(\mathbf{S}_h) \dot{\sim} \chi^2(ab\varepsilon_n, \omega_*)$ and $(\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_e) \dot{\sim} \chi^2(b\nu_e\varepsilon)$. The two-stage nature of an internal pilot design imposes a constraint on the stage 1 error sum of squares matrix $\mathbf{S}_{e1}$

when $N_+ = n_+$. Lemma 3 provides approximate distributions for $\text{tr}(\mathbf{S}_{h+})$, $\text{tr}(\mathbf{S}_{e+} - \mathbf{S}_{e1})$, and $\text{tr}(\mathbf{S}_{e1})$.

Lemma 3—For an internal pilot design, given $N_+ = n_+$, $\text{tr}(\mathbf{S}_{h+})$ and $\text{tr}(\mathbf{S}_{e+} - \mathbf{S}_{e1})$ can be approximated by independent chi square variables with $(\varepsilon_n/\bar{\lambda}) \text{tr}(\mathbf{S}_{h+}) \sim \chi^2(ab\varepsilon_n, \omega_*)$ and $(\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e+} - \mathbf{S}_{e1}) \sim \chi^2(bn_2\varepsilon)$, where $n_2 = n_+ - n_1$. In addition, $(\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e1}) \sim \chi^2(b\nu_1\varepsilon)$ with a lower bound $q_L(n_+) \approx \varepsilon_n(b\nu_1\varepsilon/\bar{\lambda}) \cdot \lambda_*(n_+ - 1)$ and an upper bound $q_U(n_+) \approx \varepsilon_n(b\nu_1\varepsilon/\bar{\lambda}) \cdot \bar{\lambda}_*(n_+)$.

From Lemma 3, $(\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e1})$ follows a doubly truncated chi-square distribution when $N_+ = n_+$. Coffey *et al.*²⁰ derived the density of the sum of two independent random variables, with one following a doubly truncated chi-square distribution, and the other following a standard chi-square distribution, with different degrees of freedom. Using the result, we can

approximate the distribution of $Z = (\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e+})$ for $z > q_L(n_+)$ as

$$f_z(z) \approx f_{\chi^2}(z; b\nu_1\varepsilon) \frac{F_{\beta}[q_U(n_+)/z; b\nu_1\varepsilon/2, bn_2\varepsilon/2] - F_{\beta}[q_L(n_+)/z; b\nu_1\varepsilon/2, bn_2\varepsilon/2]}{F_{\chi^2}[q_U(n_+); b\nu_1\varepsilon] - F_{\chi^2}[q_L(n_+); b\nu_1\varepsilon]}. \quad (9)$$

When $z < q_L(n_+)$, $f_z(z) = 0$. From Equation (8), the conditional distribution of the test statistic can then be expressed as

$$I(n_+) = \Pr \left\{ \text{tr}(\mathbf{S}_{h+}) \leq \frac{af_{\text{crit}}(n_+)}{\nu_+} \text{tr}(\mathbf{S}_{e+}) \mid q_L(n_+) < (\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e1}) \leq q_U(n_+) \right\} \approx \Pr \left\{ \chi^2(ab\varepsilon_n, \omega_*) \leq \frac{Z}{c(n_+)} \mid z > q_L(n_+) \right\} \approx . \quad (10)$$

where $\alpha(n_+) = (\nu_+ \varepsilon) / [af_{\text{crit}}(n_+) \varepsilon_n]$. If the final sample size is allowed to be smaller than n_0 (unrestricted rule), then the size of the stage 2 sample could be zero. When $n_2 = 0$, any value of $(\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e1})$ in the interval $(0, q_U(n_1)]$ would lead to choice of a final sample size of n_1 .

In this case, $Z = (\varepsilon/\bar{\lambda}) \text{tr}(\mathbf{S}_{e1})$ is a doubly truncated chi-square variable with a lower bound 0 and an upper bound $q_U(n_1)$. The conditional probability $I(n_+)$ then reduces to $I(n_1)$ in Equation (11) in which $\alpha(n_1) = (\nu_1 \varepsilon) / [af_{\text{crit}}(n_1) \varepsilon_n]$.

$$I(n_1) \approx \int_0^{q_U(n_1)} F_{\chi^2} \left[\frac{z}{c(n_1)}; ab\varepsilon_n, \omega_* \right] \frac{f_{\chi^2}(z; b\nu_1\varepsilon)}{F_{\chi^2}[q_U(n_1); b\nu_1\varepsilon]} dz. \quad (11)$$

4.3 The Marginal Distribution of the Final Test Statistic

Let n_L represents the smallest possible value of N_+ greater than n_1 . The marginal (unconditional) probability of failing to reject H_0 can be derived by using the law of total probability

$$Pr\{T_u \leq f_{crit}\} = I(n_1) Pr\{N_+ = n_1\} + \sum_{n_+ = n_L}^{\infty} I(n_+) Pr\{N_+ = n_+\}. \quad (12)$$

The first term on the right-hand side of Equation (12) corresponds to $N_+ = n_1$. The term will only appear with the unrestricted rule, i.e., allowing the final sample size to decrease and be smaller than the initial n_0 . If the restricted rule is used, the first term will disappear and $n_L = n_0$. Theoretically, there is no upper limit for the final sample size. Practically, resource limitations always impose a finite maximum feasible sample size.

5 Simulations

5.1 Overview of Simulations

We conducted two sets of simulations: one with a one-group profile design and the other with a two-group longitudinal design. The simulations examined the impact of misspecifying the covariance matrix, and evaluated the performance of the new approximation over multiple plausible covariance patterns. The initial sample size n_0 , was chosen to achieve target fixed-design power for the CM test. Observed Type I error rates and power values were reported for both a one-stage fixed design and a two-stage internal pilot design. The internal pilot calculations used half of the initial observations (i.e., $\pi = 0.5$, $n_1 = 0.5n_0$, rounded to integer) as the stage 1 sample to initiate sample size recalculation. The unrestricted rule (i.e., $N_+ = n_1$) was used to allow the final sample size to decrease from the initial sample size. SAS/IML® (SAS 9.4, SAS Institute Inc.) was used for all calculations. The software POWERLIB version 2.2 <http://samplesizeshop.org/software-downloads/other/> was used to determine the final sample size N_+ at the conclusion of the internal pilot phase. Summaries of all simulated results are based on 10,000 replications.

In Sections 5.2 and 5.3, we present the results for the GG, CM, and (rank-adjusted) HF tests. Results for the uncorrected and Box tests show similar patterns, and are omitted for presentation. The uncorrected test has exact control of the Type I error rate when $\varepsilon = 1$ (all eigenvalues are equal). The Box test controls the Type I error rate only when $\varepsilon = 1/b$.

5.2 Simulation 1: One-Group Profile Design

Parameters in Simulation 1 were chosen to examine the operating characteristics (power and the Type I error rate) of the proposed methods in small samples. Data were generated assuming four outcome variables ($p = 4$), with four distinct covariance patterns for $\Sigma_* = \mathbf{U}'\Sigma\mathbf{U}$. The objective was to make inference about the means of the outcomes with $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$. Specifically, we have $\mathbf{C} = \mathbf{I}$, $\mathbf{U} = \mathbf{I}_4$ and $\text{Es}(\mathbf{X}) = \mathbf{1}$. Orthonormal invariance of T_u allows using diagonal Σ_* , i.e., $\Sigma_* = \text{diag}(\lambda_k)$ for $k = 1, 2, 3, 4$. Following the design in Coffey and Muller (Table II, condition 5–8),¹⁵ we choose $\{\lambda_k\}$ to vary in degree of

deviation from a spherical structure, and $\varepsilon \in \{0.282, 0.505, 0.72, 1\}$. The target Type I error rate and power were $\alpha_t = 0.04$ and $P_t = 0.9$, with the effects of interest $B = 0.25 \times [0.5 \ 1 \ -1 \ 0.5]$.

To examine the impact of misspecified covariances, we assumed Σ_{*0} (the prior guess of Σ_* for the initial power calculation) to have $\{\lambda_k\}$ for which $\varepsilon_0 = 0.72$. Subsequently, when $\varepsilon = 0.72$, the power analysis would be based on an incorrectly specified covariance matrix. We considered $n_0 = 17$, which resulted from achieving the targeted power P_t using the CM test at the α_t level when Σ_* was correctly specified. In turn, $n_1 = 9$.

Table 2 displays the observed and predicted power and Type I error rate for the GG, CM and HF tests. The close agreement between predicted and observed power for all three tests suggests our proposed methods performed well in approximating power for an internal pilot design. Moreover, an internal pilot design yields power much closer to the target than the fixed design when the covariance structure is misspecified (i.e., when $\varepsilon \in \{0.282, 0.505, 1\}$). Table 2 shows our proposed approximations are robust against misspecified covariances (spherical covariance structure and covariance matrices far away from sphericity) even when the initial sample size was as small as 9.

An internal pilot design protects power but may inflate the Type I error rate. The CM test was shown to provide accurate p-values without being either too liberal or too conservative for fixed design settings,¹⁹ but its good control of the Type I error rate was not preserved in an internal pilot design. Even when the covariance matrix was correctly specified ($\varepsilon = 0.72$), the observed Type I error of 0.0505 was greater than the targeted 0.04. The GG test controlled the Type I error rate when the covariance was correctly specified, but was overly conservative when the sphericity assumption was met ($\varepsilon = 1$). This phenomenon was observed by Coffey and Muller.¹⁵ We noticed the largest difference between the observed and predicted Type I error rate when $\varepsilon = 1$. When strong evidence suggests a spherical underlying covariance structure, for example, exchangeability of observations is present and verifiable, none of the three UNIREP tests should be used. Instead, the uncorrected test is the natural choice and uniformly most powerful.

5.3 Simulation 2: Two-Group Longitudinal Design

Simulation 2 was planned to assess the performance of the proposed methods in designing studies in which common covariance structures for the longitudinal data are considered. Data were generated assuming five repeated measures ($p = 5$) at time 0, 0.25, 0.5, 0.75, 1.0. The objective was to compare the difference in changes over time between two groups. Thus, $E(\mathbf{X}) = \mathbf{I}_2$, $\mathbf{C} = [1 \ -1]$, and $\mathbf{U}$ is a 5×4 matrix with columns for orthonormal polynomial contrasts of the five time points. The target Type I error rate and power were $\alpha_t = 0.05$ and $P_t = 0.9$, with the effects of interest $\Theta = [-1.13 \ -0.05 \ 0.35 \ 0.22]$.

We considered combinations of covariance structures commonly used in longitudinal studies: compound symmetry (CS) and first-order autoregressive (AR(1)). Gurka *et al.*²¹ noted that medical researchers often assume a CS structure, while Simpson *et al.*²² stated AR(1) is the most commonly used structure in repeated measures settings. Four covariance structures for Σ were considered following the specifications of Gurka *et al.*²¹:

1,IID: random intercept and i.i.d. within-subject errors ($\epsilon = 1$);

1,AR: random intercept and AR(1) within-subject errors ($\epsilon = 0.94$);

2,IID: random intercept and slope (with correlation = 0.25) and i.i.d. within-subject errors ($\epsilon = 0.95$);

2,AR: random intercept and slope (with correlation = 0.25) and AR(1) within-subject errors ($\epsilon = 0.83$).

The variances of the random intercept, random slope, and the error term were 2, 1, and 1, respectively. The correlation for AR(1) was 0.25.

To examine the effect of misspecifying the covariance structure, we assumed Σ_0 (the prior guess of Σ) to be compound symmetric (**1,IID**, $\epsilon_0 = 1$). Consequently, when $\epsilon \neq 1$, the power analysis was conducted using a misspecified covariance matrix. An initial sample size of $n_0 = 44$ ($n_1 = 22$) is required to achieve the targeted power P_t at the α_t level using the CM test. Table 3 displays the observed and predicted power and Type I error rate for the GG, CM and HF tests.

We observed similar patterns as seen in the first simulation and Table 2. An internal pilot design preserves power better than the fixed design, particularly when the covariance matrix is misspecified (i.e., $\epsilon \in \{0.94, 0.95, 0.83\}$ corresponds to **1,AR**, **2,IID**, and **2,AR**, respectively), but has a disadvantage of potentially inflating the Type I error rate. In Table 3, an internal pilot design increased the sample size and maintained targeted power when the covariance matrix underspecified the true structure (**2,IID** and **2,AR**), where the fixed design led to power lower than the target.

The covariance structures examined in simulation 2 have similar values for the sphericity parameter ϵ . Additional covariance structures were considered (data not shown) to examine mild, moderate and severe misspecifications of the variances and correlations. The results show that 1) predicted and observed power and Type I error rate agree, 2) an internal pilot design protects power, and 3) inflation of the Type I error rate is reduced when the misspecifications are mild. The results demonstrate our proposed methods are robust to covariance structures often seen in a mixed model analysis.

The Type I error rate inflation in Simulation 2 is much smaller than in Simulation 1. The result appears to be due to the less extreme misspecification of covariances in Simulation 2. Specifically, the misspecified covariances led to using a spherical Σ_{*0} ($\epsilon_0 = 1$) when the true Σ_* has $\epsilon \in \{0.94, 0.95, 0.83\}$. In Simulation 1 however, the population ϵ ranges from 0.282 to 1, while the correct covariance matrix has $\epsilon = 0.72$. In addition, the large sample size could also have helped alleviate the inflation of Type I error rate. In Table 4, we examined the effect of internal pilot sample size using $\pi = 0.25$ ($n_1 = 10$), $\pi = 0.75$ ($n_1 = 32$), and $\pi = 1$ ($n_1 = 44$). The choice of π has little impact on power, but does affect the Type I error rate, with greater inflation when n_1 is small in absolute size, not relative size.

6 Application to the Motivating Example

The proposed methods were applied to design a longitudinal study investigating the impact of early treatment on inflammatory biomarkers over time for YLWH. Concentrations of two contributing biomarkers soluble CD14 and CD163 will be measured to characterize the immune function for each study participant. Data will be collected at study entry, year 0.5, year 1 and year 3 (end of study). A 3:1 randomization scheme will be used to allocate 75% of the study participants to the early treatment arm and 25% of the study participants to standard of care. The target Type I error rate was $\alpha_t = 0.05$, and the target power was $P_t = 0.9$. The underlying population correlation structure (Figure 1C) was assumed to be a smoothed version of the observed correlation matrix (Figure 1B), i.e., AR(1) for CD14 with $\rho = 0.5$, and compound symmetry for soluble CD163 with $\rho = 0.62$. Correlation values are based on six-month lags. The mean observed variances were assumed to be the true variances. The CM test option in software POWERLIB version 2.2 was used to determine the initial sample size n_0 , in achieving the target power for the fixed design. In turn, $n_0 = 204$ and $n_0 = 112$ for soluble CD14 and CD163, respectively. We used half of the initial sample to form the stage 1 sample. The 3:1 randomization ratio led to considering n_1 to be a multiple of 4, i.e., $n_1 = 104$ and $n_1 = 56$.

After determining the stage 1 sample size n_1 , the probability of every possible realization of N_+ was evaluated using our methods discussed in section 4. Following an unrestricted rule, the smallest possible N_+ has the value of n_1 . Figure 2 shows the distributions of the final sample size N_+ of two biomarkers using the CM test. The expected final sample sizes for soluble CD14 and CD163 are $E(N_+) = 110$ and $E(N_+) = 105$, respectively. An internal pilot design protects power to reach the targeted 0.9 for both biomarkers. We note the expected final sample size of soluble CD14 is much smaller than the initial sample size, i.e., $E(N_+) < n_0$. Power analysis based on n_0 subjects would lead to an unnecessarily high power of more than 0.99. Here, an internal pilot design can save sample size and maintain power when the initially planned total sample size is too big (same as soluble 163). For the planned study with two investigated biomarkers, we choose the maximum calculated sample size to ensure the target power for both biomarkers.

7 Discussion

An internal pilot design can protect against a poor guess of error covariance. The advantage becomes more important in planning a study with repeated measures that require specification of multiple covariance parameters. The potential inflation of the Type I error rate, however, may compromise the benefit. Although large sample size could minimize or even eliminate the inflation, we suggest always using the proposed approximation to examine a range of plausible covariance patterns and document the possible Type I error rate inflation before starting an internal pilot study. Simulations could also provide observed values of the Type I error rate and power, but simulations of an internal pilot design are computationally intensive.

The new analytic results enable straightforward implementation of the power and sample size analysis when an internal pilot design is considered for repeated data collection. The

agreement between predicted and observed power in the simulations highlights the good performance of our methods. The methods apply to a common class of linear mixed model, i.e., mixed models which can be written as equivalent general linear multivariate models. The equivalence allows making fixed effects inference using multivariate tests and extending the well-established power and sample size methods. The approach, however, does not permit missing data or time-varying covariates. Methods that account for missing data while capitalizing on the benefits of an internal pilot design are areas for further research.

Acknowledgments

This work was supported in part by NIH/NIDA R01-DA031017; Stephany W. Holloway University Chair for HIV/AIDS Research (Goodenow); and Laura McClamma Fellowship (Zhang). Chi's support included NIDCR R01-DE020832, NIDDK R01-DK072398, NICHD PO1-HD065647-03, and NCI 1U10CA180899. Muller's support included NCRR K30-RR022258, NICHD PO1 HD065647-03, NIGMS 1R25GM111901-01, NIGMS 9R01GM121081-05, and NLM 1G13LM011879-01.

Appendix

Model Parameter Estimates

A linear mixed model that satisfy the conditions specified in Section 3.1 is equivalent to a general linear multivariate model. Least square estimators of the equivalent model are

$\tilde{B} = (X'X)^{-1}X'Y$ and $\hat{\Sigma} = \{Y' [I_N - X(X'X)^{-1}X'] Y\} / \nu_e$. The estimators allow calculating the hypothesis sum of squares matrix $S_h = \hat{\Delta} = (\hat{\Theta} - \Theta_0)' M^{-1} (\hat{\Theta} - \Theta_0)$ and error sum of squares matrix $S_e = \nu_e \sum_* = \nu_e U' \sum U$. Here $\hat{\Theta} = C\tilde{B}U$, and the middle matrix is $M = C(X'X)^{-1}C'$.

UNIREF Tests

The univariate approach to repeated measures (UNIREF) test statistic is a ratio of traces, i.e., $T_u = [\text{tr}(S_h)/a] / [\text{tr}(S_e)/\nu_e]$. With sphericity ($\epsilon = 1$), test statistic T_u follows an F distribution exactly, $T_u \sim F(ab, b\nu_e; \omega_T)$, with the noncentrality parameter $\omega_T = \text{tr}(\hat{\Delta}) / \text{tr}(\hat{\Sigma}_*)$ and $\hat{\Delta} = (\hat{\Theta} - \Theta_0)' M^{-1} (\hat{\Theta} - \Theta_0)$. When the sphericity assumption is not met ($\epsilon \neq 1$), the test statistic T_u follows an F distribution approximately with $T_u \sim F[ab\epsilon, b\nu_e\epsilon; \omega_T \cdot \epsilon]$.

The five UNIREF tests differ by their approaches to estimate the sphericity parameter

$\epsilon = \text{tr}^2(\hat{\Sigma}_*) / [b \text{tr}^2(\hat{\Sigma}_*)]$. The Box test uses the lower bound of $\epsilon = 1/b$, while the UN test uses the upper bound of $\epsilon = 1$. The GG test uses the MLE, $\tilde{\epsilon} = \text{tr}^2(\hat{\Sigma}_*) / [b \text{tr}(\hat{\Sigma}_*)^2]$. The original HF test uses a ratio of unbiased estimators, $\tilde{\epsilon} = (Nb\hat{\epsilon} - 2) / [b(\nu_e - b\hat{\epsilon})]$. The rank-adjusted HF test uses $\tilde{\epsilon}_r = [(\nu_e + 1)b\hat{\epsilon} - 2] / [b(\nu_e - b\hat{\epsilon})]$. The CM test uses a multiple of $\tilde{\epsilon}_r$ with $\tilde{\epsilon}_{CM} = \tilde{\epsilon}_r(\nu_a - 2)(\nu_a - 4) / \nu_a^2$ and $\nu_a = (\nu_e - 1) + \nu_e(\nu_e - 1) / 2$. The five ϵ estimates are ordered: $1/b \leq \hat{\epsilon} \leq \tilde{\epsilon}_{CM} \leq \tilde{\epsilon}_r \leq 1$. Consequently, the Type I error rate and power are in the same order (Box GG CM HF UN).

The five UNIREP tests also have different choices of the multiplier e for power approximation (Equation 1), i.e.,

$e(\text{Box})=1/b$, $e(\text{GG}) = E(\hat{\varepsilon})$, $e(\text{HF})=E(\tilde{\varepsilon}_r)$, $e(\text{CM})=E(\tilde{\varepsilon}_{CM})$, and $e(\text{UN}) = 1$. Muller *et al.*⁵ proposed using the first-order Taylor series to approximate the expectations of e estimators.

They proved $E(\hat{\varepsilon}) \approx E[tr^2(\hat{\Sigma}_*)]/\{bE[tr(\hat{\Sigma}_*)^2]\}$, with

$$E[tr^2(\hat{\Sigma}_*)]=(1/\nu_e)[2tr(\hat{\Sigma}_*)^2+\nu_e tr^2(\hat{\Sigma}_*)] \text{ and}$$

$E[tr(\hat{\Sigma}_*)^2]=(1/\nu_e)[(\nu_e+1)tr(\hat{\Sigma}_*)^2+tr^2(\hat{\Sigma}_*)]$. The same Taylor series approximation can be used to obtain

$$E(\tilde{\varepsilon}_r) \approx \{(\nu_e+1)E[tr^2(\hat{\Sigma}_*)]-2E[tr(\hat{\Sigma}_*)^2]\}/\{b\nu_e E[tr(\hat{\Sigma}_*)^2]-bE[tr^2(\hat{\Sigma}_*)]\}, \text{ and}$$

$$E(\tilde{\varepsilon}_{CM})=[(\nu_a-2)(\nu_a-4)/\nu_a^2] \cdot E(\tilde{\varepsilon}_r).$$

Proof of Lemma 1

Proof

Coffey and Muller¹² derived the distribution of the final sample size for an internal pilot design with univariate outcomes. They adopted the general linear univariate model

framework with $\mathbf{y}_+ = \mathbf{X}_+\beta_+ + \mathbf{e}_+$ and $\mathbf{e}_+ \sim \mathcal{N}_{N_+}(0, \sigma^2 \mathbf{I})$. The associated general

linear hypothesis is $H_0: \theta = C\beta = \theta_0$. For multivariate model with univariate testing, i.e., $b = \text{rank}(\mathbf{U}) = 1$, the transformed response $\mathbf{Y}_{U+} = \mathbf{Y}_+\mathbf{U}$ is univariate, and the underlying covariance matrix reduces to the underlying variance, i.e., $\Sigma_* = \mathbf{U}'\Sigma\mathbf{U} = \sigma^2$. Equivalently, the estimated covariance matrix from the stage 1 sample reduces to the estimated variance from

the stage 1 sample, i.e., $\hat{\Sigma}_{*1} = \mathbf{U}'\hat{\Sigma}_1\mathbf{U} = \hat{\sigma}_1^2$.

Coffey and Muller¹² used SSH_{Es} to denote the essence hypothesis sums of squares with $\text{SSH}_{Es} = (\theta - \theta_0)' \{C[\text{Es}(\mathbf{X}_+)' \mathbf{W}_+ \text{Es}(\mathbf{X}_+)]^{-1} C'\}^{-1} (\theta - \theta_0)$, where the essence matrix $\text{Es}(\mathbf{X}_+)$ and the diagonal weight matrix $\mathbf{W}_+$ have the same definition, regardless of whether the outcome is univariate or multivariate. Consequently, $\text{tr}(\hat{\Sigma}_{Es}) = \text{SSH}_{Es}$ when $b = 1$. In turn, λ_* reduces to σ^2

$$\lambda_* = \frac{tr(\hat{\Sigma}_*) + 2n_+ tr(\hat{\Sigma}_* \Delta_{Es}/a)}{tr(\hat{\Sigma}_*) + 2n_+ tr(\Delta_{Es}/a)} = \frac{(\sigma^2)^2 + 2n_+ \sigma^2 (\text{SSH}_{Es}/a)}{\sigma^2 + 2n_+ (\text{SSH}_{Es}/a)} = \sigma^2. \quad (13)$$

To demonstrate $\lambda_*(n_+) = n_+ \text{tr}(\hat{\Sigma}_{Es})/\omega_*(n_+)$ reduces to $\sigma^2(n_+) = n_+ \text{SSH}_{Es}/\omega(n_+)$ for the univariate testing, we need to prove that the noncentrality $\omega_*(n_+)$ reduces to $\omega(n_+)$. Given $N_+ = n_+$, $\omega_*(n_+)$ needs to satisfy Equation (3). When $b=1$, the stage 1 e estimators for five UNIREP tests are all equal to 1. In turn, the expectations of these estimators are all equal to 1, i.e., the multiplier $e_1 = 1$. Furthermore,

$\varepsilon = tr^2(\hat{\Sigma}_*)/\{btr(\hat{\Sigma}_*)^2\} = (\sigma^2)^2/(\sigma^2)^2 = 1$, $\bar{\lambda} = tr(\hat{\Sigma}_*)/b = \sigma^2$, and $\varepsilon_n = \bar{\lambda}/\lambda_* = 1$. Equation

(3) reduces to $P_t \approx 1 - F_F[F_F^{-1}(1 - \alpha_t; a, \nu_+); a, \nu_+, \omega_*(n_+)]$, which is the condition $\omega(n_+)$ needs to satisfy in the univariate setting when $N_+ = n_+$, i.e., Equation (13) in Coffey and Muller.¹² Hence, when $b = 1$, $\lambda_*(n_+) = \sigma^2(n_+)$. In addition,

$$\tilde{\varepsilon}_{n1} = \frac{tr^2(\hat{\Sigma}_{*1}) + 2n_+ tr(\hat{\Sigma}_{*1}) tr(\Delta_{Es}/a)}{b\{tr(\hat{\Sigma}_{*1})^2 + 2n_+ tr(\hat{\Sigma}_{*1}) tr(\Delta_{Es}/a)\}} = \frac{(\hat{\sigma}_1^2)^2 + 2n_+ \hat{\sigma}_1^2 (SSH_{Es}/a)}{(\hat{\sigma}_1^2)^2 + 2n_+ \hat{\sigma}_1^2 (SSH_{Es}/a)} = 1. \quad (14)$$

which leads to $E[\tilde{\varepsilon}_{n1}] = 1$, and $\nu_* = (bv_1)E[\hat{\varepsilon}_1]/E[\tilde{\varepsilon}_{n1}] = \nu_1$. Equation (6) reduces to

$$Pr\{N_+ \leq n_+\} \approx Pr\left\{\chi^2(\nu_1) \leq \frac{\nu_1}{\sigma^2} \frac{n_+ SSH_{Es}}{\omega(n_+)}\right\}. \quad (15)$$

which is the distribution of the final sample size in the univariate setting, i.e., Equation (17) in Coffey and Muller.¹² In sum, when the testing is univariate ($b = 1$), the distribution of the final sample size determined by the multivariate model is the same as Coffey and Muller derived for the univariate outcome.

Proof of Lemma 2

Proof

Let $\mathbf{Y}_{U+} = \mathbf{Y}_+ \mathbf{U}$ represent the transformed response, which follows a matrix Gaussian distribution with $\mathbf{Y}_{U+} \sim \mathcal{N}_{N_+,b}(\mathbf{X}_+ \mathbf{B}_+ \mathbf{U}, \mathbf{I}_{N_+}, \sum_*)$. Without loss of generality, we consider $\Theta_0 = 0$ for testing $H_0 : \Theta = 0$.¹⁵ To prove the matrix independences when $N_+ = n_+$, we first show S_{h+} , S_{e+} , S_{e1} , and $S_{e+} - S_{e1}$ can all be expressed as multivariate Gaussian quadratic forms in $\mathbf{Y}_{U+}$, i.e., $S_{h+} = \mathbf{Y}_{U+}' \mathbf{A}_h \mathbf{Y}_{U+}$, $S_{e+} = \mathbf{Y}_{U+}' \mathbf{A}_e \mathbf{Y}_{U+}$, $S_{e1} = \mathbf{Y}_{U+}' \mathbf{A}_1 \mathbf{Y}_{U+}$, and $S_{e+} - S_{e1} = \mathbf{Y}_{U+}' (\mathbf{A}_e - \mathbf{A}_1) \mathbf{Y}_{U+}$. Then we prove $\mathbf{A}_h$, $\mathbf{A}_e$, $\mathbf{A}_1$, and $(\mathbf{A}_e - \mathbf{A}_1)$ are all idempotent matrices. If any pairs of these idempotent matrices has a product of 0, we can apply theorem 10.8 in Muller and Stewart³ to show the independence of the corresponding multivariate Gaussian quadratic forms (and their traces).

Here, $\mathbf{A}_h = \mathbf{X}_+ (\mathbf{X}_+' \mathbf{X}_+)^{-1} \mathbf{C}' \mathbf{M}^{-1} \mathbf{C} (\mathbf{X}_+' \mathbf{X}_+)^{-1} \mathbf{X}_+'$ is idempotent of rank a , and $\mathbf{A}_e = \mathbf{I}_{N_+} - \mathbf{X}_+ (\mathbf{X}_+' \mathbf{X}_+)^{-1} \mathbf{X}_+'$ is idempotent of rank ν_+ . We define $\mathbf{A}_1$ as the following form. The idempotency of $\mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$ indicates that $\mathbf{A}_1$ is an idempotent matrix with rank ν_1 .

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{I}_{n1} - \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' & 0 \\ 0 & 0 \end{bmatrix} \quad (16)$$

$(N_+ \times N_+)$ $\begin{matrix} (n_1 \times n_1) & (n_1 \times n_2) \\ (n_2 \times n_1) & (n_2 \times n_2) \end{matrix}$

To prove $(\mathbf{A}_e - \mathbf{A}_1)$ is an idempotent matrix with rank n_2 , we first show

$$\mathbf{X}'_+ \mathbf{A}_1 = [\mathbf{X}'_1 \mathbf{X}'_2] \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_1 - \mathbf{X}'_1 \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

(17)

In turn, $\mathbf{A}_e \mathbf{A}_1 = [\mathbf{I}_{N_+} - \mathbf{X}_+(\mathbf{X}'_+ \mathbf{X}_+)^{-1} \mathbf{X}'_+] \mathbf{A}_1 = \mathbf{A}_1$ and also $\mathbf{A}_e \mathbf{A}_1$ is symmetric. Hence, $(\mathbf{A}_e - \mathbf{A}_1)(\mathbf{A}_e - \mathbf{A}_1) = \mathbf{A}_e - \mathbf{A}_1$. Furthermore, based on matrix algebra, we have $\mathbf{A}_h \mathbf{A}_e = \mathbf{A}_h \mathbf{A}_1 = \mathbf{A}_h(\mathbf{A}_e - \mathbf{A}_1) = \mathbf{A}_1(\mathbf{A}_e - \mathbf{A}_1) = 0$. Consequently, $\mathbf{S}_{h+}$ and $\mathbf{S}_{e+}$, $\mathbf{S}_{h+}$ and $\mathbf{S}_{e1}$, $\mathbf{S}_{h+}$ and $\mathbf{S}_{e+} - \mathbf{S}_{e1}$, $\mathbf{S}_{e1}$ and $\mathbf{S}_{e+} - \mathbf{S}_{e1}$ are independent, so as their traces.³

Proof of Lemma 3

Proof

Given $N_+ = n_+$, $\mathbf{S}_{h+}$ follows a non-central Wishart distribution, while $\mathbf{S}_{e1}$ and $\mathbf{S}_{e+} - \mathbf{S}_{e1}$ follow a central Wishart distribution with $\mathbf{S}_{h+} \sim W_b(a, \sum_*, \Delta \sum_*^{-1})$, $\mathbf{S}_{e1} \sim W_b(\nu_1, \sum_*, 0)$, and $\mathbf{S}_{e+} - \mathbf{S}_{e1} \sim W_b(n_2, \Sigma^*, 0)$, respectively. Glueck and Muller²³ proved trace of a Wishart equals positively weighted sum of independent chi-square random variables. Kim *et al.*²⁴ proposed matching two moments and using a single non-central chi-square variable to approximate the weighted sum of independent chi-square random variables. Applying their results, $tr(\mathbf{S}_{h+}) \approx (\bar{\lambda}/\varepsilon_n) \cdot \chi^2(ab\varepsilon_n, \omega_*)$, $tr(\mathbf{S}_{e1}) \approx (\bar{\lambda}/\varepsilon) \cdot \chi^2(b\nu_1\varepsilon)$ and $tr(\mathbf{S}_{e+} - \mathbf{S}_{e1}) \approx (\bar{\lambda}/\varepsilon) \cdot \chi^2(bn_2\varepsilon)$.

Equation (7) implies $\bar{\lambda}_{*1} = \hat{\lambda}_1 / \tilde{\varepsilon}_{n1}$ has a lower bound $\lambda_*(n_+ - 1)$, and an upper bound $\lambda_*(n_+)$ when $N_+ = n_+$. Accordingly, we can approximate the lower bound $q_L(n_+)$ and upper bound

$q_U(n_+)$ of $(\varepsilon/\bar{\lambda})tr(\mathbf{S}_{e1}) = (\varepsilon/\bar{\lambda}) \cdot b\nu_1 \hat{\lambda}_1 \approx (\varepsilon/\bar{\lambda}) \cdot b\nu_1 \varepsilon_n \cdot \bar{\lambda}_{*1}$ with $q_L(n_+) \approx \varepsilon_n(b\nu_1\varepsilon/\bar{\lambda}) \cdot \lambda_*(n_+ - 1)$ and $q_U(n_+) \approx \varepsilon_n(b\nu_1\varepsilon/\bar{\lambda}) \cdot \lambda_*(n_+)$.

References

1. Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann Math Stat.* 1945; 16(3):243–258.
2. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med.* 1990; 9(1–2):65–72. [PubMed: 2345839]
3. Muller, KE., Stewart, PW. *Linear model theory: univariate, multivariate, and mixed models.* John Wiley & Sons; 2006.
4. Ringham BM, Kreidler SM, Muller KE, Glueck DH. Multivariate test power approximations for balanced linear mixed models in studies with missing data. *Stat Med.* 2016; 17(35):2921–2937.
5. Muller KE, Edwards LJ, Simpson SL, Taylor DJ. Statistical tests with accurate size and power for balanced linear mixed models. *Stat Med.* 2007; 26(19):3639–3660. [PubMed: 17394132]
6. Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Stat Med.* 1994; 13(23–24): 2455–2463. [PubMed: 7701146]
7. Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: type I error rate of the naive t-test. *Stat Med.* 1999; 18(24):3481–3491. [PubMed: 10611620]

8. Proschan MA, Wittes J. An improved double sampling procedure based on the variance. *Biometrics*. 2000; 56(4):1183–1187. [PubMed: 11129477]
9. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Stat Med*. 2000; 19(7):901–911. [PubMed: 10750058]
10. Miller F. Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*. 2005; 61(2):355–361. [PubMed: 16011681]
11. Denne JS, Jennison C. Estimating the sample size for a t-test using an internal pilot. *Stat Med*. 1999; 18(13):1575–1585. [PubMed: 10407230]
12. Coffey CS, Muller KE. Exact test size and power of a Gaussian error linear model for an internal pilot study. *Stat Med*. 1999; 18(10):1199–1214. [PubMed: 10363340]
13. Coffey CS, Muller KE. Controlling test size while gaining the benefits of an internal pilot design. *Biometrics*. 2001; 57(2):625–631. [PubMed: 11414593]
14. Zucker DM, Denne J. Sample-size redetermination for repeated measures studies. *Biometrics*. 2002; 58(3):548–559. [PubMed: 12229989]
15. Coffey CS, Muller KE. Properties of internal pilots with the univariate approach to repeated measures. *Stat Med*. 2003; 22(15):2469–2485. [PubMed: 12872303]
16. World Health Organization. Guideline on when to start antiretroviral therapy and on preexposure prophylaxis for HIV. Geneva, Switzerland: World Health Organization; 2015.
17. Lundgren JD, Babiker AG, Gordin F, et al. Initiation of antiretroviral therapy in early asymptomatic HIV infection. *N Engl J Med*. 2015; 373(9):795. [PubMed: 26192873]
18. Gribbin MJ, Chi YY, Stewart PW, Muller KE. Confidence regions for repeated measures ANOVA power curves based on estimated covariance. *BMC Med Res Methodol*. 2013; 13(1):57. [PubMed: 23586676]
19. Chi YY, Gribbin M, Lamers Y, Gregory JF, Muller KE. Global hypothesis testing for high-dimensional repeated measures outcomes. *Stat Med*. 2012; 31(8):724–742. [PubMed: 22161561]
20. Coffey CS, Kairalla JA, Muller KE. Practical methods for bounding type I error rate with an internal pilot design. *Commun Stat Theory Methods*. 2007; 36(11):2143–2157.
21. Gurka MJ, Edwards LJ, Muller KE. Avoiding bias in mixed model inference for fixed effects. *Stat Med*. 2011; 30(22):2696–2707. [PubMed: 21751227]
22. Simpson SL, Edwards LJ, Muller KE, Sen PK, Styner MA. A linear exponent AR (1) family of correlation structures. *Stat Med*. 2010; 29(17):1825–1838. [PubMed: 20658550]
23. Glueck DH, Muller KE. On the trace of a Wishart. *Commun Stat Theory Methods*. 1998; 27(9): 2137–2141.
24. Kim HY, Gribbin MJ, Muller KE, Taylor DJ. Analytic, computational, and approximate forms for ratios of noncentral and central gaussian quadratic forms. *J Comput Graph Stat*. 2006; 15(2)

A – Observed and planned treatment differences **B – Observed correlation matrices** **C – Assumed true correlation matrices** **Fig. 1.**

Fig. 2.

Table 1

Notation for the equivalent multivariate model with an internal pilot design. Elements are grouped into 5 categories: Design, Known constants, Sample size, Covariance and Hypothesis. Random quantities and estimates are highlighted in gray.

Category	Symbol	Dimension	Definition
Design	π	1×1	Proportion for the stage 1 sample
	α_t	1×1	Target Type I error rate
	P_t	1×1	Target power
	$\Theta_t = \mathbf{C}\mathbf{B}_t\mathbf{U}$	$a \times b$	Clinically relevant effects
	$\text{Es}(\mathbf{X}_1) = \text{Es}(\mathbf{X}_2) = \text{Es}(\mathbf{X}_+)$	$r \times q$	Essence matrix
	r	1×1	Rank of design matrix
Known constants	$\mathbf{C}$	$a \times q$	Between-subject contrast matrix
	$\mathbf{U}$	$p \times b$	Within-subject contrast matrix
	a	1×1	Rank of C
	b	1×1	Rank of U
Sample size	n_0	1×1	Initial sample size
	$n_1 = \pi n_0$	1×1	Stage 1 sample size
	$\nu_1 = n_1 - r$	1×1	Stage 1 sample degrees of freedom
	N_+	1×1	Final sample size
	$\nu_+ = N_+ - r$	1×1	Final sample degrees of freedom
	$N_2 = N_+ - n_1$	1×1	Stage 2 sample size
Covariance	Σ	$p \times p$	Outcome covariance matrix
	Σ_0	$p \times p$	Prior guess of covariance matrix
	$\hat{\Sigma}_1(\hat{\Sigma}_+)$	$p \times p$	Stage 1 (final) covariance matrix estimates
	$\Sigma_* = \mathbf{U}'\Sigma\mathbf{U}$	$b \times b$	Hypothesis covariance matrix
	$\hat{\Sigma}_{*1}(\hat{\Sigma}_{*+})$	$b \times b$	Stage 1 (final) hypothesis covariance matrix estimates
	λ	$b \times 1$	Eigenvalue vector of Σ_*
	$\mathbf{\lambda}$	$b \times b$	Eigenvector matrix of Σ_*
	$\varepsilon = \text{tr}^2(\hat{\Sigma}_*) / [\text{btr}(\hat{\Sigma}_*^2)]$	1×1	Sphericity parameter
	$\hat{\varepsilon}_1(\hat{\varepsilon}_+)$	1×1	Stage 1 (final) ε estimates for GG test
	$\tilde{\varepsilon}_{r1}(\tilde{\varepsilon}_{r+})$	1×1	Stage 1 (final) ε estimates for rank-adjusted HF test

Category	Symbol	Dimension	Definition
	$\tilde{\epsilon}_{CM1}(\tilde{\epsilon}_{CN+})$	1×1	Stage 1 (final) e estimates for CM test
Hypothesis	B	$q \times p$	Primary regression parameters
	$\tilde{B}_1(\tilde{B}_+)$	$q \times p$	Stage 1 (final) primary parameter estimates
	$\Theta = \mathbf{CBU}$	$a \times b$	Secondary parameters
	Θ_0	$a \times b$	Secondary parameters under the null
	$\hat{\Theta}_1(\hat{\Theta}_+)$	$a \times b$	Stage 1 (final) secondary parameter estimates
	Es	$b \times b$	Essence unscaled noncentrality
	$\hat{\Delta}_1(\hat{\Delta}_+)$	$b \times b$	Stage 1 (final) unscaled noncentrality estimates
	$S_{hl}(S_{h+})$	$b \times b$	Stage 1 (final) hypothesis sum of squares matrices
	$S_{el}(S_{e+})$	$b \times b$	Stage 1 (final) error sum of squares matrices

Simulation 1 (one-group profile design) results. Observed values for the fixed and an internal pilot design are based on 10,000 replications. Predicted values for an internal pilot design are calculated from our proposed approximation. The target power $P_I = 0.9$ and the target Type I error rate $\alpha_I = 0.04$. The average final sample size of an internal pilot design is denoted by $E(N_+)$.

Table 2

ϵ		Power $\times 100$				Type I error rate $\times 100$			
		0.282	0.505	0.720	1	0.282	0.505	0.720	1
Geisser-Greenhouse	Fix Observed	73	87	90	93	4.17	4.41	3.31	2.93
	IPS Observed	93	93	93	93	5.97	4.71	4.10	2.92
	IPS Predicted	90	93	93	93	5.20	4.60	3.91	3.40
	$E(N_+)$	22.2	18.9	18.4	17.3	22.4	18.9	18.4	17.3
Chi-Muller	Fix Observed	72	88	92	94	4.11	4.70	4.27	3.97
	IPS Observed	94	93	93	92	5.90	5.48	5.05	4.05
	IPS Predicted	91	93	93	93	5.20	4.60	3.91	3.40
	$E(N_+)$	23.5	18.7	17.6	16.0	22.4	18.9	18.4	17.3
Rank-adjusted Huyuh-Feldt	Fix Observed	74	89	92	95	4.35	5.10	4.39	4.15
	IPS Observed	93	90	92	92	5.65	5.32	6.11	5.19
	$E(N_+)$	22.0	17.5	16.6	15.6	22.0	17.6	16.6	15.5

Table 3

Simulation 2 (two-group longitudinal design) results. Observed values for the fixed and an internal pilot design are based on 10,000 replications. Predicted values for an internal pilot design are calculated from our proposed approximation. The target power $P_t = 0.9$ and the target Type I error rate $\alpha_t = 0.05$. The average final sample size of an internal pilot design is denoted by $E(N_+)$.

	Power $\times 100$				Type I error rate $\times 100$				
	1,IID	1,AR	2,IID	2,AR	1,IID	1,AR	2,IID	2,AR	
Geisser-Greenhouse	Fix Observed	89	89	79	78	4.62	4.41	4.44	4.47
	IPS Observed	91	91	91	90	4.93	4.80	4.59	5.12
	IPS Predicted	92	91	91	91	4.69	4.63	4.71	4.73
	$E(N_+)$	47.6	47.4	61.6	62.4	47.7	47.5	61.8	62.4
Chi-Muller	Fix Observed	90	90	80	80	5.06	5.14	4.65	5.23
	IPS Observed	91	90	90	89	5.33	5.09	4.83	5.36
	IPS Predicted	91	91	90	90	5.29	5.17	5.14	5.09
	$E(N_+)$	45.7	45.6	59.4	60.4	45.7	45.7	59.3	60.3
Rank-adjusted Huyuh-Feldt	Fix Observed	91	90	81	79	5.19	5.18	5.06	4.93
	IPS Observed	91	90	90	89	5.81	5.41	4.89	5.34
	IPS Predicted	91	91	90	90	5.34	5.21	5.17	5.11
	$E(N_+)$	45.4	45.5	59.0	59.9	45.5	45.4	58.9	60.1

Simulation 2 (two-group longitudinal design) results for different sizes of the internal pilot sample, i.e., $\pi \in \{0.25, 0.75, 1\}$ corresponds to $n_1 = 10, 32$, and 44, respectively. Observed values are based on 10,000 replications, predicted values are calculated from our proposed approximation. Both are for an internal pilot design. The target power $P_t = 0.9$ and the target Type I error rate $\alpha_t = 0.05$.

Table 4

Power $\times 100$	Geisser-Greenhouse		Chi-Muller		Rank-adjusted Huyuh-Feldt	
	IPS Obs	IPS Pred	IPS Obs	IPS Pred	IPS Obs	IPS Pred
1, IID	10	92	91	90	89	89
	32	91	91	91	92	91
	44	92	92	92	92	92
1, AR	10	90	91	90	89	89
	32	91	91	91	91	91
	44	92	92	92	92	92
2, IID	10	91	89	90	88	89
	32	91	90	91	91	91
	44	91	90	91	91	91
2, AR	10	89	91	88	90	87
	32	90	91	91	90	91
	44	91	91	91	91	91

Type I error rate $\times 100$	Geisser-Greenhouse		Chi-Muller		Rank-adjusted Huyuh-Feldt	
	IPS Obs	IPS Pred	IPS Obs	IPS Pred	IPS Obs	IPS Pred
1, IID	n_1					
	10	4.77	4.74	5.02	5.35	5.43
	32	4.79	4.67	4.98	5.27	5.31
1, AR	44	4.73	4.51	4.75	5.04	5.07
	10	4.69	4.70	5.08	5.24	5.31
	32	4.92	4.59	5.08	5.13	5.17
2, IID	44	4.38	4.57	4.94	5.09	5.13
	10	5.09	4.76	4.97	5.19	5.23
	32	4.70	4.69	5.52	5.11	5.14

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Type I error rate $\times 100$	Geisser-Greenhouse		Chi-Muller		Rank-adjusted Huynh-Feldt	
	IPS Obs	IPS Pred	IPS Obs	IPS Pred	IPS Obs	IPS Pred
2,AR	44	4.84	5.26	5.09	5.25	5.12
	10	5.10	5.37	5.14	5.19	5.18
	32	4.70	5.39	5.06	5.44	5.08
	44	4.80	5.60	5.03	5.30	5.05