An Overview of Variance Inflation Factors for Sample-Size Calculation

F. Y. Hsieh, Philip W. Lavori, Harvey J. Cohen and John R. Feussner

Eval Health Prof 2003 26: 239
DOI: 10.1177/0163278703255230

AN OVERVIEW OF VARIANCE INFLATION FACTORS FOR SAMPLE-SIZE CALCULATION

F. Y. HSIEH
Department of Veterans Affairs Palo Alto Health Care System

PHILIP W. LAVORI
Department of Veterans Affairs Palo Alto Health Care System and Stanford University

HARVEY J. COHEN
Veterans Affairs Medical Center

JOHN R. FEUSSNER
Department of Veterans Affairs

For power and sample-size calculations, most practicing researchers rely on power and sample-size software programs to design their studies. There are many factors that affect the statistical power that, in many situations, go beyond the coverage of commercial software programs. Factors commonly known as design effects influence statistical power by inflating the variance of the test statistics. The authors quantify how these factors affect the variances so that researchers can adjust the statistical power or sample size accordingly. The authors review design effects for factorial design, crossover design, cluster randomization, unequal sample-size design, multiarm design, logistic regression, Cox regression, and the linear mixed model, as well as missing data in various designs. To design a study, researchers can apply these design effects, also known as variance inflation factors to adjust the power or sample size calculated from a two-group parallel design using standard formulas and software.

Keywords: *clinical trials; design effect; power; sample size; variance inflation factor*

AUTHORS' NOTE: This work was supported by the Department of Veterans Affairs Cooperative Studies Program. Address correspondence to F. Y. Hsieh, CSPCC, Department of Veterans Affairs Palo Alto Health Care System (151-K), Palo Alto, CA 94304.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 26 No. 3, September 2003 239-257
DOI: 10.1177/0163278703255230

When planning a study, practicing researchers specify an expected (standardized) effect size and desired Types I and II errors to calculate the required sample size. The process of power calculation can be reversed to compute the sample size from estimates of effect size. In addition to these well-known parameters for power and sample-size calculations, there are many other factors, such as the types of design and analytical models and endpoint variables that are crucial to the calculations. Because it is important for an analytical model to be consistent with the study design, the sample size should be appropriate to the design and also sufficient for the final analytical model. However, the calculations of power and sample size can be complicated for various designs and models. To simplify the calculations, we propose a method of easy adjustments from the basic design.

The basic design for a comparative study is a randomized, balanced, two-group parallel design that has power and sample-size formulas available in many software programs. Once we start shifting away from the basic design, the statistical power changes. Most of these changes affect the statistical power through the change of the variance of test statistics. We call this change a "design effect," "relative efficiency," or "variance inflation factor" (VIF). If we can quantify the VIF, we can adjust power or sample size accordingly. The following sections describe VIFs for various designs, models, and endpoints.

SAMPLE-SIZE FORMULAS FOR A BASIC DESIGN

The sample-size formulas for a two-group parallel design are well-known and are implemented in most sample-size software programs. For a continuous endpoint using a two-sample *t* test, a normal approximation is commonly used in the formula. Assuming equal variances of the two groups, the formula has the following form:

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(f\,(1-f)\mu^2/\sigma^2), \qquad (1.A)$$

where N is the total sample size, f is the fraction of total sample size assigned to one treatment group; μ^2/σ^2 is the squared standardized effect size, and $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are standard normal deviates at the desired two-sided significance level α and power $1-\beta$, respectively. For a balanced design, the fraction $f = 0.5$, formula 1.A reduces to

$$N = 4\,(Z_{1-\alpha/2} + Z_{1-\beta})^2/(\mu^2/\sigma^2). \quad (1.B)$$

Snedecor and Cochran (1989, p. 104) suggested a simple sample-size adjustment for the *t* test using normal approximation:

> For tests of significance at 5 per cent or 10 per cent levels, increase the number of pair n by 2 with paired samples and the size n of each sample by 1 with independent samples. For tests at the 1 per cent level, change these increases to 3 and 2, respectively.

For a binary outcome using the chi-square test for a 2×2 table, the sample-size formula (Lachin, 1981) is

$$N = (Z_{1-\alpha/2}\,\sigma\,(f^{-1} + (1-f)^{-1})^{1/2} + Z_{1-\beta}(\sigma_1 f^{-1} + \sigma_2 (1-f)^{-1})^{1/2})^2/\mu^2, \quad (2.A)$$

where P_1 and P_2 are the proportion of events in the two groups, $P = (P_1 + P_2)/2$, $\sigma^2 = P(1-P)$, $\sigma_1^2 = P_1(1-P_1)$, $\sigma_2^2 = P_2(1-P_2)$ and $\mu = P_1 - P_2$. Because $2\sigma^2 \geq \sigma_1^2 + \sigma_2^2$ by using the upper bound, when $f = 0.5$, formula 2.A simplifies to

$$N = 4\,(Z_{1-\alpha/2} + Z_{1-\beta})^2/(\mu^2/\sigma^2). \quad (2.B)$$

For a survival endpoint using a logrank test, a similar formula is available (Schoenfeld, 1983):

$$D = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(f(1-f)\log^2\Delta), \quad (3.A)$$

where D is the total number of events and Δ is the hazard ratio of the two groups. The "hazard" can be described as the instantaneous probability of an event at time *t* and the "hazard rate" can be described as the number of events per interval of time. The required sample size is then equal to the number of events divided by the overall event rate P, which can be calculated by combining the event rates of the two groups $fP_1 + (1-f)P_2$. For a balanced design, $f = 0.5$, formula 3.A reduces to

$$D = 4\,(Z_{1-\alpha/2} + Z_{1-\beta})^2/\log^2\Delta. \quad (3.B)$$

The sample-size ratio of an unbalanced versus a balanced design is $1/(4f(1-f))$ obtained from formula 1.A versus formula 1.B or from formula 3.A versus formula 3.B. This quantity is called the VIF for an unbalanced design. The same VIF can be derived for a binary endpoint,

from formula 2.A versus formula 2.B, through a normal approximation using the arcsine transformation. We discuss the details in the next section.

UNBALANCED DESIGN

In a two-arm study, it is a standard practice that both arms have the same sample size. It is known that an equal sample-size (balanced) design provides the most efficient comparison of the two arms and therefore requires the smallest total sample sizes. An unequal sample-size design is relatively uncommon. However, in a trial to compare an experimental treatment against a standard, researchers would rather put more sample size into an experimental treatment because the outcome of a standard treatment is often better known. For example, in a pilot study for the long-term treatment of Tardive Dyskinesia conducted at Department of Veterans Affairs medical centers, patients were randomized in 3:2 ratio for vitamin E versus control so that more experience could be obtained for vitamin E treatment (Adler et al., 1998). If the ratio of the sample sizes between the two arms is controllable as in the above example, Pocock (1983, p. 89) suggested that "a 2:1 or 3:2 ratio for new: standard treatment is a realistic proposition." However, in many observational studies of the effects of treatments or other exposures, the sample-size ratio is predetermined by the study population, especially in a cross-sectional study, a cohort study, or a case control study of risk factors for a rare disease. If we keep the total sample size constant, the power decreases when the sample-size ratio of exposed and unexposed subjects starts moving away from the equal sample design. To accommodate the loss of power due to the unequal "allocation" of the sample sizes, researchers have to quantify the increased sample sizes needed to achieve the same power as in the equal sample-size design. Let us assume that the sample-size ratio between the two groups is K (i.e., $K = 1$ for a balanced design) with equal variances and the VIF for an unbalanced design is $1/(4f(1-f))$ or $(K+1)^2/4K$ (Hsieh, 1987; Lee, 1984). This VIF, as the sample-size ratio required for an unbalanced design versus a balanced design, applies to the two-sample t test, chi-square test for binary endpoints, and logrank test for survival endpoints. For example, when $K = 2$ and $\text{VIF} = (K+1)^2/4K = 1.125$, we will need a total sample size of 675 to

achieve the same power as an equal sample-size design with a total sample size of 600. Notice that in logrank test for survival endpoints, a most efficient design lies between the balance of sample sizes and the balance of numbers of events (Hsieh, 1992). How to allocate patients into different treatment groups to produce the desired ratio of events is a practical issue to be addressed before randomization.

FACTORIAL DESIGN

Although factorial designs are known to be efficient, they have not been used frequently in clinical trials. A 2×2 factorial design proposes to study two factors (or treatments) in the same study where each study participant is simultaneously randomized twice. Therefore, in the absence of interaction that depends on the assumptions of the model, researchers can conduct two studies for the price of one. Notice that a factorial design is different from a randomized block design with a single treatment factor and a block factor where the block factor is not of interest for testing. For example, in a recently published 2×2 factorial study of geriatric evaluation and management (GEM) conducted at 11 Department of Veterans Affairs medical centers, inpatients were randomized either to a control group or to a GEM unit and then, at discharge, randomized either to a control group or to an outpatient GEM clinic (Cohen et al., 2002). The study evaluated inpatient GEM units and outpatient GEM clinics in the same trial. Most factorial trials are designed on the assumption of no interaction. Therefore, the power and sample size for a 2×2 factorial design can be calculated using formulas for a two-group parallel study. Suppose for a 2×2 factorial design, the sample size required is n for each of the four arms, with a total sample size of $4n$ for the entire study. If we conduct trials separately for each factor, we will need $4n$ for each parallel study or a total of $8n$ for the two parallel studies (Piantadosi, 1997). For an unbalanced 2×2 factorial design, we extend the VIFs from the above parallel designs to simplify the calculations. Let us assume that a 2×2 factorial design has a randomization ratio of K:1 for factor A versus non-A and a ratio of L:1 for factor B versus non-B with equal variances. The VIFs extended from two-group parallel designs will be $(K + 1)^2/4K$ for factor A and $(L + 1)^2/4L$ for factor B assuming no interaction. If interaction is of interest, a 2×2 factorial design is the only

way to study interaction and the VIFs will be $(L + 1)(K + 1)^2/4K$ for factor A, $(K + 1)(L + 1)^2/4L$ for factor B, and $(K + 1)^2 (L + 1)^2/4KL$ for testing interaction. When $K = L = 1$ as in a balanced factorial design, the total sample size needed is $8n$ for the main effects and $16n$ for the interaction effect, four times the required sample size if no interaction is assumed. This implies that only one quarter of sample size contributes to the analysis of interaction effect as compared with main effects. The VIF for an unbalanced 2×2 factorial design with interaction is derived in the appendix.

For example, suppose $N = 600$ is needed from both main effects in a balanced 2×2 factorial design assuming no interaction. If we decide to randomize 2:1 ratio ($K = 2$) to factor A versus non-A and 3:2 ratio to factor B versus non-B ($L = 1.5$), we will need 675 and 625 participants for the same power to test factors A and B, respectively. If the interaction effect is assumed in the model, we will need sample sizes of 1,688, 1,875, and 2,813 for testing factors A, B, and interaction, respectively, with the same effect size. The following table illustrates the calculations:

	Without Interaction	*With Interaction*
Factor A	$N(K + 1)^2/4K = 675$	$N(L + 1)(K + 1)^2/4K = 1{,}688$
Factor B	$N(L + 1)^2/4L = 625$	$N(K + 1)(L + 1)^2/4L = 1{,}875$
Interaction		$N(K + 1)^2(L + 1)^2/4KL = 2{,}813$

In a balanced design where $L = K = 1$, we need $N = 2{,}400$ to have enough power to test interaction as illustrated in the following table:

	Without Interaction	*With Interaction*
Factor A	$N = 600$	$N(L + 1)(K + 1)^2/4K = 1{,}200$
Factor B	$N = 600$	$N(K + 1)(L + 1)^2/4L = 1{,}200$
Interaction		$N(K + 1)^2(L + 1)^2/4KL = 2{,}400$

Pocock (1983) also provided a discussion of using an unbalanced factorial design to study the combination of aspirin and an antihypertensive drug.

MULTIARM DESIGN

A multiarm design is needed when researchers would like to compare more than two treatment groups. In a cardiovascular study where a standard treatment is usually available, it is common to compare an experimental treatment with a standard treatment, a combination treatment, and/or a placebo. For example, in a published three-arm vascular-bypass study (VA Cooperative Study #141 Study Group, 1988) conducted at 18 Department of Veterans Affairs medical centers, the efficacy of vascular bypass materials in two active treatment groups was compared with a control group. Unlike a 2×2 factorial design where an additional treatment factor does not cost additional sample size, the power in a multiarm design depends on the sample size per arm. In a multiarm design, power calculation for testing the overall treatment difference has to involve the noncentrality parameter (NCP) of a test statistic (Lachin, 1997). In addition to the degrees of freedom, the NCP is a parameter of noncentral chi-square or noncentral F distribution for the specified alternative hypothesis. A common sample-size formula for a multiarm design using an F test, chi-square test for a $r \times 2$ table, or multiarm logrank test for testing the overall treatment difference is, approximately,

$$N = \text{NCP}/(\Sigma(\mu_i - \mu)^2/g\ \sigma^2),$$

where $\Sigma(\mu_i - \mu)^2/g\ \sigma^2$ is known as the squared standardized effect size, $(\mu_i - \mu)^2/g$ is the variance of group means (or proportions), g is the number of groups, μ is average of means (or proportions), and σ^2 is common variance or $\mu(1 - \mu)$ for proportions. Notice that this sample size does not take into account of the pattern of treatment means or any linear contrast among them.

If the desired effect size remains the same, the ratio of total sample size required for a multiarm study versus a two-arm study is the ratio of their noncentrality parameters. Researchers can use these ratios to expand their sample sizes from a two-arm study to a multiarm study assuming equal sample size per arm. For a multiarm logrank test, the NCP ratios refer to the number of events, not the sample sizes (Ahnn & Anderson, 1995). The NCP for a specific power in testing the null hypothesis at level alpha can be obtained through an iterative method using SAS (2001) or S-Plus functions (Mathsoft, 1999). Makuch and Simon (1982) has a table listed some commonly used NCP values for

chi-square statistic with alpha = .05. When the sample size approaches infinity, the NCP of *F* statistic decreases and approaches to that of chi-square statistic. Therefore, if the total sample size is not too small, the NCP ratio of *F* statistics approximate that of chi-square statistics. We provide the following ratios of two noncentrality parameters of chi-square statistics, for three- and four-arm studies versus a two-arm study. For example, for alpha = .05 and power = 0.95, the sample-size ratio for a four-arm design versus a two-arm design is 17.17/12.995 = 1.321.

Alpha	*Power (%)*	*3-Arm Versus 2-Arm (NCP)*	*4-Arm Versus 2-Arm (NCP)*
.05	80	1.228 (9.9635/7.849)	1.389 (10.9025/7.849)
.05	90	1.204 (12.654/10.507)	1.349 (14.1715/10.507)
.05	95	1.188 (15.443/12.995)	1.321 (17.17/12.995)
.01	80	1.189 (13.8807/11.679)	1.324 (15.4577/11.679)
.01	90	1.171 (17.4267/14.8794)	1.294 (19.2474/14.8794)
.01	95	1.159 (20.65/17.8142)	1.273 (22.6743/17.8142)

In addition to testing the overall treatment differences, paired comparisons of treatment effects are usually performed in a multiarm design. Sample size needs to be adjusted for the overall *p* value according to the number of paired comparisons.

CROSSOVER DESIGN

Crossover designs are very popular in clinical pharmacology research because they reduce variation by using each study participant as his own control. The crossover design is similar to the factorial design except that one of the factors is the study time period, the calendar time that a participant receives the other factor, the treatment factor. For example, in a two-period two-treatment crossover design, study participants are randomized to Groups 1 or 2. Group 1 receives treatments A and B in Periods 1 and 2, respectively, and Group 2 receives the treatments in the reverse order, treatments B then A. Most crossover designs assume no residual or carryover effect of treatments. If

the assumption of no residual effect is invalid, a crossover design is uneconomical. The ratio of variances of a crossover study versus a parallel design, assuming no baseline adjustment, is $1 - \rho$ where ρ is the intrasubject correlation coefficient (ICC), the ratio of variances of within-subject versus the sum of within- and between-subjects (Brown, 1980; Chassan, 1970). In a two-period two-treatment crossover design, the ICC is the simple correlation coefficient between pairs of measurements in Periods 1 and 2 taken on randomly selected participants. The VIF for paired comparisons in a balanced *M*-period *M*-treatment crossover design is $(1 - \rho)/m$, where m is the number of measurements per participant. For example, if 100 participants are needed for a completely randomized two-group parallel design with $\rho = .1$, the number of participants needed for a two-period two-treatment crossover design will be $100 (1 - \rho)/2 = 45$ participants. Here, the parallel design assumes no baseline measurement. Because each participants takes two measurements in this crossover design, the total number of measurements needed is $45 \times 2 = 90$ in comparison with the 100 measurements needed in a completely randomized design. In addition, for an unbalanced crossover design, researchers may use the analogy of an unbalanced 2×2 factorial design for sample-size adjustments.

In a published three-period, three-group crossover study conducted at eight Department of Veterans Affairs medical centers, the efficacy of three commonly used hearing aid circuits were compared (Larson et al., 2000). Patients were randomized to 1 of 6 sequences, from two Latin squares balanced for carryover effects, of the three hearing aid circuits:

A B C	A C B
B C A	C B A
C A B	B A C

A list of Latin squares for a crossover design balanced for carryover effects is available in some publications (Fleiss, 1986). If 1,000 participants are needed for a completely randomized two-group parallel design with $\rho = .1$, the number of participants needed for a two-group comparison will be 300, or 50 participants per sequence.

PRETEST AND POSTTEST DESIGN

In most studies, measurements of baseline variables are standard procedure. Adjustment to baseline measurements, as in a pretest and posttest design, may change the power of treatment comparison. If the correlation of pre- and posttreatment measurements is ρ, the variance of the change from baseline (i.e., the difference between pre- and posttreatments) needs to be multiplied by $2\ (1 - \rho)$ assuming equal variances between pretest and posttest measurements. For example, if the correlation of pretest and posttest measurements is .65 and the variance of the measurement for both pretest and posttests is 2.56, the variance of the change from baseline will reduce to $2.56 \times 0.7 = 1.792$. However, if the correlation is only .35, the variance of the change will increase to $2.56 \times 1.3 = 3.328$. Therefore, if the total sample size required for a two-arm design is 100, due to baseline adjustment, the sample size will need to decrease to 70 or increase to 130, respectively, for correlation .65 and .35. In sum, if the pretest and posttest correlation is smaller than .5, adjustment for pretest or baseline measurements will reduce the power.

MULTIPLE REGRESSION, LOGISTIC REGRESSION, AND COX REGRESSION

Multivariate analyses, such as analysis of variance (ANOVA), multiple linear regression, multiple logistic regression, and Cox regression (for survival analysis), are known to improve the power from their univariate counterparts by reducing the residual error. However, the variance of a parameter that measures an intervention effect is jointly influenced by the residual error and by other variance components in the multivariate model, including the covariance of the intervention variable and the other covariates. Recent studies show that the addition of covariates may also inflate the variance of the estimated parameter and thus reduce the power. The reduction of power occurs especially in a nonlinear regression model such as logistic regression (Hsieh, 1989, 1998; Robinson & Jewell, 1991) and Cox regression (Hsieh & Lavori, 2000). The reduction of residual error in a regression model is given by

$$\sigma_e^2 = \sigma_y^2 \left(1 - R_{yx}{}^2\right)$$

where $R_{yx}{}^2$ is the squared multiple correlation coefficient indexing the proportion of variance of endpoint Y explained by the covariate X. Therefore, the impact of the regression adjustment for additional covariates $X_2, \ldots X_p$ is the ratio

$$\left(1 - R_{y,12\ldots p}{}^2\right) / \left(1 - R_{y,1}{}^2\right).$$

The squared multiple correlation coefficients $R_{y,1}{}^2$ and $R_{y,12\ldots p}{}^2$ can be computed from most regression software programs when Y is the dependent variable in the regression on X_1 alone and $X_1, X_2, \ldots X_p$, respectively. However, this impact is very unpredictable due to the changes of variance components with different variables and thus is rarely applied in power and sample-size calculation. To ensure that we have enough power or sample size, a conservative approach is to adjust the variance inflation from collinearity between the intervention variable and the other explanatory variables. The VIF for adjusting the collinearity is given as the reciprocal of $1 - R_{1,2\ldots p}{}^2$ where $R_{1,2\ldots p}{}^2$ is the squared multiple correlation coefficient of X_1 with other covariates $X_2, \ldots X_p$ and can be computed from a regression software program when X_1 is the dependent variable in the regression on $X_2, \ldots X_p$.

It is known that the sample size for a simple logistic regression model can be calculated from the formulas for two-sample t test and chi-square test when the dependent variable X is a continuous variable and binary variable, respectively (Hsieh, Bloch, & Larsen, 1998). For comparing two treatment groups using a Cox regression model, the sample size can be obtained from the formula for the logrank test. After calculating the sample size required for an univariate analysis to study the effect of X_1 on endpoint Y, we decide to adjust for some confounding variables such as baseline variables $X_2, \ldots X_p$ in a multivariate model. We can then inflate the sample size by a factor of

$$VIF = 1 / \left(1 - R_{1,2\ldots p}{}^2\right)$$

so that the study has a sufficient power. For example, in a study (psychophysiological study of chronic post-traumatic stress disorder [PTSD]) conducted at 15 Department of Veterans Affairs medical centers, a logistic regression model is fitted to use four psychophysiological measurements in the diagnosis of PTSD (Keane et al., 1998). A sample size of $N = 580$ was calculated from a two-sample t test (comparing PTSD with non-PTSD groups) for one primary psychophysiological variable, heart rates. The prevalence rate of PTSD among the Vietnam veterans was assumed 20% (i.e., a sample-size ratio of 4:1 for non-PTSD vs. PTSD groups). By adjusting for VIF for unbalanced design, we obtained a sample size of $N = 580 \times 1.5625 = 906$. The squared multiple correlation coefficient $R_{1,2...4}^2$ of the primary variable versus other three psychophysiological variables was estimated to be .1. Therefore, the final sample size was $906/(1 - .1) = 1{,}007$ for fitting a multiple logistic regression model. This VIF has been used in sample-size calculation for multiple logistic regression (Hsieh, 1989; Hsieh et al., 1998) and Cox regression (Hsieh & Lavori, 2000) and implemented in nQuery Advisor software (Elashoff, 2001).

CLUSTER RANDOMIZATION AND LINEAR MIXED MODELS

Instead of randomizing individuals, researchers may choose to randomize clusters of participants, because clusters or groups of participants are sensible or natural units for interventions. Although the authors of many articles have discussed the adjustments for sample size for cluster-randomized designs (e.g., Murray, 1998), the need for variance inflation may be independent of whether participants were randomized in clusters. Instead, it comes from the lack of independence of outcomes within treatment group. For example, in a recently published treatments-of-PTSD (TOP) study conducted at 10 Department of Veterans Affairs medical centers, each cohort of 12 participants were randomized individually into either standard treatment or experimental treatment for group therapy of PTSD (Schnurr, Friedman, Lavori, & Hsieh, 2001). Treatment interventions were delivered to each group of six participants.

In the TOP study and in some other studies, the outcomes of patients are not independent within treatment group. For example,

community intervention studies (in which the entire community groups are randomized) or studies in which a treatment is delivered in a group setting (such as the TOP study) involve a correlation of measures due to clustering. It is known that the power is reduced by positive ICC, the ratio of the within-group variance to the total variance, which is the sum of the between- and within-group variance. Suppose $n = mk$ participants are randomized to k groups, m per group, then the VIF is known as $1 + (m - 1)\rho$, where ρ is the ICC. If the group size is just one participant, m is 1 and there is no inflation of variance (i.e., VIF = 1). If the ICC is high, adding participants to groups does not help with variance as much as adding groups does. When the ICC is greater than .30 and the number of participants per group is greater than 10, the variance is insensitive to the addition of new participants to groups and thus the power of the test (Hsieh, 1988). In most community intervention studies where the cluster sizes are large, such as a work site or a classroom, the addition of the number of participants is insensitive to the power when the ICC greater than .30. Fortunately, the values of ICC in most community intervention studies are small, which make the studies affordable.

Although there are formulas available for a two-sample t test for cluster randomized data (Hsieh, 1988), researchers have to use statistical software for linear mixed model, such as SAS PROC MIXED (SAS, 2001), to perform a valid two-sample t test. For example, specifying the treatment variable or intervention variable as the only fixed-effect and specifying the group variable as the only random effect gives a valid test, if the variance component model is correct. Beyond this t test, any additional covariates in the model require additional power or sample-size adjustments such as the collinearity adjustment of $1/(1 - R^2)$ discussed in the previous section.

ADJUSTMENTS FOR MISSING DATA

Adjusting for missing data is an important procedure in sample-size calculation. When data are missing at random (MAR) or missing completely at random (MCAR) and occur on continuous covariates (Little & Rubin, 1987), researchers can use multiple imputation software to impute missing data and to combine the results from the analyses of imputed data sets. Notice that the assumption of MAR or not is

not testable in the data, but the assumption of MAR versus MCAR is testable. Software programs for imputations available from SAS Institute (SAS, 2001) include SAS PROC MI and PROC MIANALYZE. The MI procedure produces a "fraction of missing information" (Little & Rubin, 1987) on the parameter, which is not the same as the fraction of missing data unless the data are MCAR. Researchers can use the inverse of this fraction as a VIF if data are MCAR. If the fraction of missing information is P and the data are MCAR, the sample size needs to be inflated by a factor of $(1 - P)^{-1}$ to compensate the loss of power. If the data are MAR, the correct VIF can be larger or smaller than $(1 - P)^{-1}$.

The power of survival data is a function of the number of events. However, a MAR or MCAR survival data can occur either on an expected censored data or an event. A compromise solution is to assume half of the missing data, assuming MCAR, are expected events that contribute to the power. So if the proportion of MCAR is P, we will inflate by a factor of $(1 - P/2)^{-1}$, instead of $(1 - P)^{-1}$. For example, if the number of events needed for a study is 100 and the proportion of MCAR is 20%, we will need to increase the sample size to 100/.9 = 112.

For a cluster-randomized design, adjustments for missing data need to be made at each level of the linear mixed model because applying the same VIF to an individual level may have a very different impact than applying it to a cluster level. It may be that missing data in a 2×2 factorial design is heavier on one factor than the other, because the intervention in that factor retains fewer study participants. In this situation, a balanced 2×2 factorial design may turn into an unbalanced design. Adjustments for missing data should follow the same VIFs for an unbalanced 2×2 factorial design and should begin the study with an unbalanced randomization.

Although a two-period, two-treatment crossover design is more efficient than a completely randomized two-treatment parallel design when there is no residual effect, the effect of missing data, especially from dropouts, in a crossover design can be more severe than a regular parallel design. Assume the proportion of missing data is P for each treatment. The VIF for missing data in a completely randomized two-treatment parallel design versus a two-period, two-treatment crossover design is $(1 - P)^{-1}$ versus $(1 - P)^{-2}$.

DISCUSSION AND CONCLUSION

The VIF or design effect provides a simple method to adjust sample size for various designs or analytical models from a basic two-group design. The following is a summary table of VIFs applied to either continuous, binary, or survival endpoints.

Type of Design/Model	*VIF*
Two-group unbalanced	$(K + 1)^2/4K$, where K or L is sample-size ratio of the two groups
2×2 factorial	$(K + 1)^2/4K$ and $(L + 1)^2/4L$ for main effect without interaction $(L + 1)(K + 1)^2/4K$ and $(K + 1)(L + 1)^2/4L$ for main effect with interaction $(K + 1)^2(L + 1)^2/4KL$ for interaction effect
Multiple arms/ANOVA	Ratio of noncentrality parameters
Balanced $M \times M$ crossover	$(1 - \rho)/m$, where ρ is the ICC
Pretest and posttest	$2(1 - \rho)$, where ρ is pretest and posttest correlation or ICC
Regressions	$1 / \left(1 - R_{1,2\ldots p}{}^2\right)$, where $R_{1,2\ldots p}{}^2$ is the squared multiple correlation coefficient
Cluster randomization	$1 + (m - 1)\rho$, where m is the number of observations within a randomized unit
Missing data	$(1 - P)^{-1}$, where P is fraction of missing information if the data is MCAR $(1 - P/2)^{-1}$ for survival data

By using the above VIFs, researchers can design a more efficient study and simplify sample-size calculation.

APPENDIX

In a 2×2 factorial design assuming no interaction, the effect of factors A and B, in an additive model, are estimated by linear contrasts of treatment means of the four groups:

$$\theta_A = (Y_A - Y_0 + Y_{AB} - Y_B)/2$$

and

$$\theta_B = (Y_B - Y_0 + Y_{AB} - Y_A)/2.$$

The tests of main effects can be viewed as comparisons of two groups, for example, $Y_A + Y_{AB}$ versus $Y_B + Y_0$ for factor A. Therefore, the VIF for an unbalanced two-group parallel design can be applied for an unbalanced factorial design.

If the interaction effect exists in an univariate analysis, the main and interaction effects can be estimated by $\theta_A = Y_A - Y_0$, $\theta_B = (Y_B - Y_0)$, and $\theta_{AB} = (Y_{AB} - Y_A - Y_B + Y_0)/2$. If a and b are fractions of total sample size N for factors A and B, respectively, the loss of information due to interaction are fractions of b and a for factors A and B, respectively. Therefore, the VIF due to interaction are $(1 - b)^{-1}$ for factor A and $(1 - a)^{-1}$ for factor B.

In a multivariate analysis, a linear model for a 2×2 factorial design with interaction has the following form (Piantadosi, 1997):

$$E\{Y\} = \theta_0 + \theta_A X_A + \theta_B X_B + \theta_{AB} X_A X_B,$$

where X_A and X_B are both binary indicators of factors A and B and θ_A, θ_B, θ_{AB} are the effects of factors A, B, and interaction, respectively. For example, $X_A = 1$ for factor A and 0 otherwise; $X_B = 1$ for factor B and 0 otherwise. Notice that the indicator $X_A X_B$ provides a comparison of treatment means Y_{AB} with $Y_A + Y_B + Y_0$ instead of $Y_{AB} + Y_0$ with $Y_A + Y_B$. The unbalance of sample sizes for testing the interaction effect has reduced the power dramatically. Inferences for both main effects θ_A and θ_B and for the interaction effect θ_{AB} can be simultaneously performed in this three-parameter model. Under this three-parameter model, the null hypotheses to be tested for each parameter are

Hypothesis 1: $[\theta_A, \theta_B, \theta_{AB}] = [0, \theta_B, \theta_{AB}]$ for testing θ_A main effect.
Hypothesis 2: $[\theta_A, \theta_B, \theta_{AB}] = [\theta_A, 0, \theta_{AB}]$ for testing θ_B main effect.
Hypothesis 3: $[\theta_A, \theta_B, \theta_{AB}] = [\theta_A, \theta_B, 0]$ for testing θ_{AB} interaction effect.

The design matrix has dimension $N \times 4$ and is

$$X'_{4\times N} = \begin{array}{c} \begin{matrix} (1-a)(1-b)N & a(1-b)N & b(1-a)N & abN \end{matrix} \\ \begin{bmatrix} 1\ldots\ldots & 1\ldots\ldots & 1\ldots\ldots & 1\ldots\ldots \\ 0\ldots\ldots & 1\ldots\ldots & 0\ldots\ldots & 1\ldots\ldots \\ 0\ldots\ldots & 0\ldots\ldots & 1\ldots\ldots & 1\ldots\ldots \\ 0\ldots\ldots & 0\ldots\ldots & 0\ldots\ldots & 1\ldots\ldots \end{bmatrix} \end{array}.$$

The covariance matrix of parameter estimates is $(X'X)^{-1}\sigma^2$, where σ^2 is the variance of each observation and

$$X'X = N\begin{bmatrix} 1 & a & b & ab \\ a & a & ab & ab \\ b & ab & b & ab \\ ab & ab & ab & ab \end{bmatrix},$$

$$(X'X)^{-1} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1/a & 1 & -1/2 \\ -1 & 1 & 1/b & -1/b \\ 1 & -1/a & -1/b & 1/ab \end{bmatrix} / N(1-a)(1-b).$$

The variance of estimates of factors A, B, and interaction are $\sigma^2/Na(1 - a)(1 - b)$, $\sigma^2/Nb(1 - a)(1 - b)$, and $\sigma^2/Nab(1 - a)(1 - b)$, respectively. Therefore, the sample-size formulas for a 2×2 factorial design with interaction in a linear model are

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(a(1 - a)(1 - b)\,\theta_A^2/\sigma^2), \quad \text{(A.1)}$$

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(b(1 - a)(1 - b)\,\theta_B^2/\sigma^2), \quad \text{(A.2)}$$

and

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(a(1 - a)\mathrm{b}(1 - b)\,\theta_{AB}^2/\sigma^2) \quad \text{(A.3)}$$

for factors A, B, and interaction, respectively. In comparison, the sample size for a balanced two-sample parallel design, assuming equal variances, has the following formula

$$N = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(\theta^2/\sigma^2), \quad \text{(A.4)}$$

where θ^2/σ^2 is the squared standardized effect size.

For a 2×2 factorial design with survival endpoints, the Cox regression model assumes that the hazard function $\lambda(t)$ for the survival time T given the treatment indicators X_A and X_B has the following regression formulation:

$$\log(\lambda(t|X)/\lambda_0(t)) = \theta_A X_A + \theta_B X_B + \theta_{AB} X_A X_B,$$

where $\lambda_0(t)$ is the baseline hazard.

Slud (1994) proposed an adjusted logrank statistic for a factorial survival design involving interaction effect. The total number of events has the following similar forms (Schmoor, Sauerbrei, & Schumacher, 2000; Slud, 1994):

$$D = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(a(1 - a)(1 - b)\,\theta_A^2), \quad \text{(A.5)}$$

$$D = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(b(1-a)(1-b)\,\theta_B^2), \quad \text{(A.6)}$$

$$D = (Z_{1-\alpha/2} + Z_{1-\beta})^2/(a(1-a)b(1-b)\,\theta_{AB}^2), \quad \text{(A.7)}$$

for factors A, B, and interaction, respectively. In comparison, the formulas for a balanced two-sample logrank test is

$$D = 4\,(Z_{1-\alpha/2} + Z_{1-\beta})^2/\theta^2, \quad \text{(A.8)}$$

where θ is the log hazards ratio of the two groups. By taking the sample-size ratios of (A.1), (A.2), and (A.3) versus (A.4) or ratios of (A.5), (A.6), and (A.7) versus (A.8), the VIF for an unbalanced factorial design versus a balanced two-sample parallel design are $[4a(1-a)(1-b)]^{-1}$, $[4b(1-a)(1-b)]^{-1}$, and $[4a(1-a)b(1-b)]^{-1}$, for factors A, B, and interaction, respectively. By substituting sample-size ratios $K = a/(1-a)$ and $L = b/(1-b)$, the VIFs for an unbalanced design have the equivalent forms $(L+1)(K+1)^2/4K$ for factor A, $(K+1)(L+1)^2/4L$ for factor B, and $(K+1)^2(L+1)^2/4KL$ for interaction.

REFERENCES

Adler, L. A., Edson, R., Lavori, P., Peselow, E., Duncan, E., Rosenthal, M., et al. (1998). Long-term treatment effects of vitamin E for Tardive Dyskinesia. *Biological Psychiatry, 43*(12), 868-872.

Ahnn, S., & Anderson, S. J. (1995). Sample size determination for comparing more than two survival distributions. *Statistics in Medicine, 14*, 2273-2282.

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics, 36*, 69-79.

Chassan, J. B. (1970). A note on the relative efficiency in clinical trials. *Journal of Clinical Pharmacology , 10*, 359-360.

Cohen, H. J., Feussner, J. R., Weinberger, M., Carnes, M., Hamdy, R. C., Hsieh, F., et al. (2002). A controlled trial of inpatient and outpatient geriatric evaluation and management. *New England Journal of Medicine, 346*(12), 905-912.

Elashoff, J. D. (2001). *nQuery advisor* (version 4.0 user's guide). Boston: Statistical Solution.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: John Wiley.

Hsieh, F. Y. (1987). A simple method of sample calculation for unequal-sample-size designs that use the logrank or *t*-test. *Statistics in Medicine, 6*, 577-581.

Hsieh, F. Y. (1988). Sample size formulas for intervention studies with the cluster as unit of randomization. *Statistics in Medicine, 7*, 1195-1201.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine, 8*, 795-802.

Hsieh, F. Y. (1992). Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine, 11*, 1091-1098.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17*, 1623-1634.

Hsieh, F. Y., & Lavori, P. W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials, 21*, 552-560.

Keane, T. M., Kolb, L. C., Kaloupek, D. G., Orr, S. P., Blanchard, E. B., Thomas, R. G., et al. (1998). Utility of psychophysiological measurement in the diagnosis of posttraumatic stress disorder: Results from a Department of Veterans Affairs cooperative study. *Journal of Consulting and Clinical Psychology, 66*(6), 914-923.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, *2*, 93-113.

Lachin, J. M. (1997). Sample size determination for rxc comparative trials. *Biometrics*, *33*, 315-324.

Larson, V. D., Williams, D. W., Henderson, W. G., Luethke, L. E., Beck, L. B., Noffsinger, D., et al. (2000). Efficacy of 3 commonly used hearing aid circuits: A crossover trial. *Journal of the American Medical Association, 284*(14), 1806-1813.

Lee, Y. J. (1984). Quick and simple approximation of sample sizes for comparing two independent binomial distribution: Different-sample-size case. *Biometrics*, *40*, 239-241.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Makuch, R. W., & Simon, R. M. (1982). Sample size requirements for comparing time-to-failure among k treatment groups. *Journal of Chronic Diseases*, *35*, 861-867.

Mathsoft Inc. (1999). *S-Plus 2000 guide to statistics* (Vol. 1). Seattle, WA: Author.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Piantadosi, S. (1997). *Clinical trials—A methodological perspective*. New York: John Wiley.

Pocock, S. J. (1983). *Clinical trials—A practical approach*. New York: John Wiley.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review, 58*(2), 227-240.

SAS Institute Inc. (2001) *SAS/STAT software: Changes and enhancements* (Release 8.2). Gary, NC: Author.

Schmoor, C., Sauerbrei, W., & Schumacher, M. (2000). Sample size consideration for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine*, *19*, 441-452.

Schnurr, P. P., Friedman, M. J., Lavori, P. W., & Hsieh, F. Y. (2001). Design of Department of Veterans Affairs Cooperative Study No. 420: Group treatment of posttraumatic stress disorder. *Controlled Clinical Trials*, *22*, 74-88.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, *39*, 499-503.

Slud, E. V. (1994). Analysis of factorial survival experiments. *Biometrics*, *50*, 25-38.

Snedecor, G., & Cochran, W. (1989). *Statistical methods*. Ames: Iowa State University Press.

VA Cooperative Study #141 Study Group. (1988). Comparative evaluation of prosthetic, reversed and in situ vein bypass grafts in distal popliteal and tibial/peroneal revascularization. *Archives of Surgery, 123*(4), 434-438.