Published in final edited form as:

Stat Biopharm Res. 2009 August 1; 1(3): 291–300. doi:10.1198/sbr.2009.0023..

Robust Non-inferiority Tests for Potency of a Test Drug Against a Reference Drug:

(To appear in *Statistics in Biopharmaceutical Research*)

Michael P. Fay¹, Farzad Noubary², and Allan Saul³

¹ Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 6700B Rockledge Drive MSC 7609, Bethesda, Maryland 20892-7609. (email: mfay@niaid.nih.gov)

² Department of Biostatistics, Harvard School of Public Health, Boston, MA. (email: noubary@fas.harvard.edu)

³ National Institute of Allergy and Infectious Diseases, (current address) Novartis Vaccines Institute for Global Health, Siena, Italy, (email: Allan.Saul@novartis.com)

Summary

In comparing test formulations of a drug product to reference formulations it is necessary to check that the test formulation has as much biologic activity as the reference within some indifference zone. A modern method for testing this relative potency is to fit a nonlinear model of the dose response curve, constrained such that the dose-response of the test is modeled with the same function as the reference except for a multiplicative parameter on the dose. This method works well as long as the nonlinear model fits the data. Although tests of this fit are mentioned in the literature, how relative potency should be measured when those tests fail has not been addressed. We study the use of the sandwich estimator of variance for this non-inferiority potency problem when the model does not fit the data. We focus on the simple linear model. Using the sandwich variance, the inferences are asymptotically valid even if the model does not hold. We discuss interpretation of the results under the misspecified linear model, and show by simulation that the method performs well in some practical cases. Finally, we give advice for designing a study that uses these methods.

Keywords

Assay; Delta Method; Fieller's Theorem; Inverse regression; Sandwich variance estimator

1 Introduction

This paper concerns testing whether a test drug is non-inferior to a reference drug with respect to its potency on some assay. One strategy to address this non-inferiority issue is to show that the dose of the test drug needed to elicit some mean response, say μ_0 , is no more than δ times the dose of the reference drug needed to elicit the same mean response. For many assays the choice of mean response value μ_0 is inherently arbitrary because the standardization of the assay can change for different batches of the assay. For example, in an Enzyme-Linked ImmunoSorbent Assay (ELISA) used to determine the concentration of antibodies that are sensitive to a specific antigen, the units of response (ELISA antibody units) can be quite variable between different batches of the assay. Thus, typically these non-inferiority tests are done with both the test and reference drugs tested in the same batch and assuming a parametric model on the log of the dose. In order for this test not to depend on the arbitrary value of μ_0 , the two parametric curves for the two drugs (or two formulations) are assumed to be “parallel” in the sense that the curves are identical except for a multiplicative parameter for the dose of

the test drug. Following standard practice we use the term parallel to describe two such curves even if they are non-linear models (see e.g., Gottschalk and Dunn, 2005a, Pilaytis and Carlone, 2005). This multiplicative parameter can be used to measure relative potency *if* the parallelism assumption holds. It is possible that a test of parallelism may reject this assumption, even when the degree of non-parallelism is negligible simply because the assay has very little error (Callahan and Sajjadi, 2003, Pilaytis and Carlone, 2005). In this case, since the simplifying assumption has been rejected, the associated simple test of relative potency cannot be justified. Note that even when parallelism is accepted, as response curves are located within some equivalence region (see e.g., Callahan and Sajjadi, 2003), we still have problems formally defining relative potency for two drugs for this equivalence case when the hypothesis of parallelism has been rejected. In other words, even though the two curves are very close within some indifference zone, the definition of relative potency is unclear when the two curves are statistically significantly non-parallel.

A second difficulty with these non-inferiority potency tests is that the variance of the responses are typically related to the expected response, but no simple distribution (e.g., Poisson) can be assumed. The variance function is usually approximated with a power function of the expected response, where the parameters for the variance function are estimated from historical assay data since there typically is not enough data in a single data set to estimate the variances well (Dudley, et al, 1985, Gottschalk and Dunn, 2005b).

In this paper, we discuss how to test non-inferiority of potency without making any parallelism assumptions and without assuming any model on the variance function. In Figure 1 we plot data from an assay which is typical of the kind which motivated this research. We plot both means and confidence intervals for each dose for both test and reference drugs, as well as the model which is both linear and constrained to be parallel on the log dose scale. Since there are only three different doses, more complex models such as a 4-parameter logistic or 5-parameter logistic (see e.g., Gottschalk and Dunn, 2005a) are too complicated for this situation. To solidify the ideas we focus in this paper on the linear model, and we briefly discuss similar strategies for non-linear models in Section 7. Notice in Figure 1 that the test drug appears to have similar responses to the reference, although the linear model may not fit. A likelihood ratio test comparing the linear model to a model which allows different means for each dose/drug combination is barely non-significant ($p=.058$). There are other ways to test for parallelism and other issues related to parallelism for bioassays (see e.g., Plikaytis, et al, 1994), but the focus of this paper is not these tests for parallelism, since we will formulate the problem so that it may be answered with or without parallelism. Further, notice that the variability of the log 10 transformed response for both the test and the reference drug appears slightly larger at the small doses, although not so much larger that weighted least squares regression is vital.

Relative potency may be modeled as a simple multiplicative parameter, but this parameterization makes even the simple model of Figure 1 a non-linear one. Another approach is to fit the usual linear model, so that relative potency is ratio of parameters from that model. Inferences about this ratio have been studied assuming homogeneous normal errors and the problem is known as inverse regression (see e.g., Hubert, Bohidar, and Peace, 1988). Hubert, et al suggested three different methods for estimating the variability of the dose needed to elicit a particular response: the delta method, a method using Fieller's Theorem, and the jackknife method.

In this paper we explore relaxing assumptions for this problem. We essentially use the linear model estimator but calculate the variance such that the inferences do not require assuming that the linear model fits the data. We use a sandwich estimator of variance to calculate the variability of the linear model estimator, and explore the finite properties of using the sandwich estimator together with either the delta method or Fieller's theorem through simulations. We

show that the jackknife method is not applicable when the model of the mean is misspecified. Using a misspecified model of the variance has been widely studied in the biostatistics literature especially through the generalized estimating equations (see e.g., Diggle, Heagerty, Liang and Zeger, 2002). Models with a misspecified mean have been studied more thoroughly in the econometric literature (see e.g., White, 1994). The contribution of this paper is combining the use of the sandwich estimator with the inverse regression problem to this specific application of testing relative potency. The beauty of this approach to this problem is that we need not make any parallelism assumptions, nor assumptions about the variance model. Similar applications of sandwich variance estimators to non-linear models are possible but not studied in this paper.

In Section 2 we outline the simple linear model approach to this problem. In Section 3 we describe how the linear model may be used for this problem even when the mean does not follow that linear model and when the variances are heterogeneous across different dose/drug combinations. In Section 4 we show through simulation that our new approach retains a nominal size and has better power than the linear model approach when the linear model does not fit the data. In Section 5 we give some practical advice on both the design of such studies (Section 5.1) and on a modification needed when there are response values which are beyond the limit of detection of the assay (Section 5.2). We apply the methods to the data in Figure 1 in Section 6 and end with a discussion in Section 7.

2 Non-inferiority Tests for Potency using the Linear Model

We start by reviewing the linear model and the associated non-inferiority tests. The specific motivation for this work is for comparing malaria vaccine formulations (see Giersing, et al, 2006), but throughout the paper we will use the more general terms: test drug and reference drug. Let Y_{ijk} be the response of the i th drug (reference=1 and test=2) and j th dose and the k th observation within that drug/dose combination. The linear model is

$$E(Y_{ijk}) = \theta_1 + \theta_2 I(i=1) + \theta_3 x_j \quad (1)$$

where $x_j = \log_{10}(d_j)$, d_j is the j th dose, and $I(i=1) = 1$ for the reference drug and 0 for the test drug. In matrix notation we write, $\mathbf{Y} = \mathbf{X}\theta + \varepsilon$, where $\mathbf{Y}$ is an $n \times 1$ vector of all responses, $\theta = [\theta_1, \theta_2, \theta_3]$, and $\mathbf{X}$ is a $n \times 3$ matrix with the first column all ones, the second column all $I(i=1)$ and the third column the x_j values corresponding to the associated Y_{ijk} response in $\mathbf{Y}$. Then $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the least squares estimator of the parameters, and

$$\text{Var}(\hat{\theta}) \equiv \sum = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (2)$$

where $\text{Var}(\mathbf{Y}) = \text{Var}(\varepsilon) = \mathbf{V}$.

Let $D_1(\mu_0)$ and $D_2(\mu_0)$ be the doses of the reference and test respectively that give a mean response of μ_0 . Under the linear model, $\mu_0 = \theta_1 + \theta_2 I(i=1) + \theta_3 \log_{10}\{D_i(\mu_0)\}$, and by algebra,

$$D_i(\mu_0) = 10^{\frac{\mu_0 - \theta_1 - \theta_2 I(i=1)}{\theta_3}}.$$

We want to reject the null hypothesis that $D_2(\mu_0) > \delta D_1(\mu_0)$ which under the linear model can be equivalently written

$$10^{\frac{\mu_0 - \theta_1}{\theta_3}} > \delta 10^{\frac{\mu_0 - \theta_1 - \theta_2}{\theta_3}} \Rightarrow \frac{\theta_2}{\theta_3} > \log_{10}(\delta)$$

Let $r(\theta) = \theta_2/\theta_3$, so that the hypotheses can be written

$$\begin{aligned} H_0: r(\theta) &> \log_{10}(\delta) \\ H_1: r(\theta) &\leq \log_{10}(\delta). \end{aligned}$$

For the homogeneous linear model we estimate $\text{Var}(\mathbf{Y}) = \mathbf{V}$ with $\sigma^2 \mathbf{I}$ where $\mathbf{I}$ is an identity matrix and σ^2 is the mean squared error from the model, so that $\mathbf{\Sigma}$ is estimated with $\hat{\mathbf{\Sigma}} = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$. Then we can test whether any linear combination of the parameters θ_i is different from a constant by comparing the associated linear combination of parameter estimates divided by its estimated standard deviation to a t distribution with $n - 3$ degrees of freedom (see e.g., Graybill, 1976). However, in this case we are interested in the ratio of two parameters, $r(\theta) = \theta_2/\theta_3$, so we do not have a linear combination of parameters. There are three traditional choices for conducting inferences about $r(\theta)$ under this situation (see e.g., Hubert et al, 1988).

2.1 Delta Method

The delta method is a very general method used to give asymptotic distribution for functions of normal random variables (see e.g., Lehmann, 1999). The delta method estimate of variance applied to this situation is

$$\widehat{\text{Var}}[r(\hat{\theta})] = \left(\frac{\partial r(\hat{\theta})}{\partial \hat{\theta}} \right)' \widehat{\Sigma} \left(\frac{\partial r(\hat{\theta})}{\partial \hat{\theta}} \right) = \left(\frac{1}{\hat{\theta}_3} \right)^2 \widehat{\Sigma}_{22} - 2 \left(\frac{\hat{\theta}_2}{\hat{\theta}_3^2} \right) \widehat{\Sigma}_{23} + \left(\frac{\hat{\theta}_2}{\hat{\theta}_3^2} \right)^2 \widehat{\Sigma}_{33} \quad (3)$$

where $\hat{\Sigma}_{22}$, $\hat{\Sigma}_{23}$, and $\hat{\Sigma}_{33}$, denote the estimated variance of $\hat{\theta}_2$, the covariance of $\hat{\theta}_2$, and $\hat{\theta}_3$, and the variance of $\hat{\theta}_3$, respectively. To better control the type I error, we use the t distribution with $n - 3$ degrees of freedom in calculating the confidence intervals since that distribution would be used if, for example, θ_3 was a known constant and the ratio could be written as a linear combination of parameters. This will be asymptotically equivalent to the usual delta method using the normal distribution and should perform better for finite samples.

2.2 Fieller's Theorem

We state Fieller's confidence intervals without derivation (for details see e.g., Buonaccorsi, 2005). The $100(1 - \alpha)\%$ confidence interval is finite only when $\hat{\theta}_3$ is significantly different

from zero, i.e., when $|\hat{\theta}_3| / \sqrt{\widehat{\Sigma}_{33}} \geq t$, where t is the $1 - \alpha/2$ quantile of the t distribution with $n - 3$ degrees of freedom. In that case the $100(1 - \alpha)\%$ confidence interval is

$$\left[\frac{f_1 - (f_1^2 - f_0 f_2)^{1/2}}{f_2}, \frac{f_1 + (f_1^2 - f_0 f_2)^{1/2}}{f_2} \right]. \quad (4)$$

where

$$\begin{aligned} f_0 &= \widehat{\theta}_2^2 - t^2 \widehat{\Sigma}_{22}, \\ f_1 &= \widehat{\theta}_2 \widehat{\theta}_3 - t^2 \widehat{\Sigma}_{23}, \text{ and} \\ f_2 &= \widehat{\theta}_3^2 - t^2 \widehat{\Sigma}_{33}. \end{aligned}$$

2.3 Jackknife Variance Estimate

The application of the jackknife to our situation is detailed by Hubert, Bohidar, and Peace (1988). We do not give the details here, but note that the calculation of the jackknife involves sequentially deleting each observation and recalculating $r(\hat{\theta})$ without that observation.

We show in the following section that this strategy does not work with a misspecified model because under the misspecified model the $r(\hat{\theta})$ depends on the points in the design space. We confirm this suspicion in the simulations, where the jackknife performs poorly for the misspecified models.

3 Relaxing the Linear Model Assumption

Recall that in the linear model the ratio of the test dose needed to elicit a response, μ_0 , to the reference dose needed to elicit that same response was a constant regardless of μ_0 : $D_2(\mu_0)/D_1(\mu_0) = 10^{\theta_2/\theta_3}$ for all μ_0 . In this section, we propose a test when the dose responses of the two drugs do not necessarily follow a linear model. The idea is to continue to base the test on $r(\hat{\theta})$ but calculate the variability on $r(\hat{\theta})$ without assuming the linear model.

3.1 Formulating the Test

In this section we assume that

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}^* \quad (5)$$

where μ_{ij} represents a separate mean for each drug/dose combination, and assume ε_{ijk}^* are distributed with mean 0 and possibly different but finite variances σ_{ij}^2 . Let $\mu(x_{ij}) = \mu_{ij}$ where $x_{ij} = [1, I(i = 1), x_j]$ is the row of $\mathbf{X}$ from the linear model associated with Y_{ijk} for any k . We extend the notation in a natural way by letting $\mu(\mathbf{X})$ be an $n \times 1$ vector where the ℓ th element of $\mu(\mathbf{X})$ is equal to μ_{ij} when the ℓ th row of $\mathbf{X}$ equals x_{ij} . Then equation 5 in matrix notation is $\mathbf{Y} = \mu(\mathbf{X}) + \varepsilon^*$. Let

$$\theta^*(\mathbf{X}) = E(\hat{\theta}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mu(\mathbf{X}),$$

where $\hat{\theta}$ is the same as in Section 2, i.e., $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. We test the hypotheses,

$$\begin{aligned} H_0: r\{\theta^*(\mathbf{X})\} &> \log_{10}(\delta) \\ H_1: r\{\theta^*(\mathbf{X})\} &\leq \log_{10}(\delta), \end{aligned}$$

where $r\{\theta^*(\mathbf{X})\}=\theta_2^*/\theta_3^*$ and $\theta^*(\mathbf{X})=[\theta_1^*, \theta_2^*, \theta_3^*]$. Notice that unlike the linear model, the parameter of interest and the hypotheses change as the design matrix, $\mathbf{X}$, changes.

In this situation, we estimate $\theta^*(\mathbf{X})$ with $\hat{\theta}^*(\mathbf{X}) \equiv \hat{\theta}$ and $Var\{\hat{\theta}^*(\mathbf{X})\}$ with equation 2, except now we estimate $\mathbf{V}$ with the sample variance at each dose/drug combination. In other words, we use $\hat{\mathbf{V}}$ which is diagonal with terms

$$\widehat{Var}(Y_{ijk})=\hat{\sigma}_{ij}^2=\frac{1}{m_{ij}-1}\sum_{k=1}^{m_{ij}}(Y_{ijk}-\bar{Y}_{ij})^2$$

where m_{ij} is the number of observations for the i th drug at dose x_j , and $\bar{Y}_{ij}=m_{ij}^{-1}\sum_{k=1}^{m_{ij}}Y_{ijk}$. Notice that we use the unbiased estimate of the variance, which is not always done in applications using the sandwich estimator to avoid some of the usual problems with the liberalness of the sandwich estimator (see e.g., Fay and Graubard, 2001). If we assume normality of the variances then this is equivalent to the recommendation of Diggle, Heagerty, Liang, and Zeger (2002): fit a saturated model to the data and use the restricted maximum likelihood estimator (REML) of the variance components. Specifically, we estimate $Var[r\{\hat{\theta}^*(\mathbf{X})\}]$ using $\hat{\Sigma}$, which is equation 2 with $\hat{\mathbf{V}}$ replacing $\mathbf{V}$.

Both the delta method and Fieller's theorem are applied as before except with $\hat{\Sigma}$ replacing Σ and with $n-2J$ replacing the degrees of freedom for the t distribution, where J is the number of different doses in the design.

3.2 Interpreting the Null Hypothesis

Recall that when the linear model with parallel lines on the log dose fits the data, then $r\{\hat{\theta}^*(\mathbf{X})\}$ may be interpreted as $\log_{10}\{D_2(\mu_0)/D_1(\mu_0)\}$ so that the null hypothesis: $r\{\hat{\theta}^*(\mathbf{X})\} > \log_{10}(\delta)$ may be equivalently expressed as $D_2(\mu_0) > \delta D_1(\mu_0)$. When the linear model does not fit the data, the former expression is still correct but the latter expression is not. Nevertheless, if we let

$$D_i^*(\mu_0)=10^{\frac{\mu_0-\theta_1^*-I(i=1)\theta_2^*}{\theta_3^*}}$$

be the estimated dose for drug i from the best fitting linear model (by ordinary least squares with constrained parallel lines on the log dose), then the expression $D_2^*(\mu_0) > \delta D_1^*(\mu_0)$ is correct and describes a null hypothesis which does not depend on μ_0 . Thus, even when the linear model does not fit the data, δ still has an interpretation related to a factor increase in the test drug as estimated by the best fit linear model on the log dose.

We now give some details on interpreting $r\{\hat{\theta}^*(\mathbf{X})\}$ in some special cases. For the rest of this section we restrict ourselves to the case where both drugs have the same number of observations at each dose. Specifically, let m_j be the number of observations for each drug at x_j and let

$m=\sum_{j=1}^J m_j$. Let $\mu_i=\sum_{j=1}^J m_j \mu_{ij}$ be the sum of means of all the observations in the i th drug group. In the Appendix we show that

$$r\{\theta^*(X)\} = \frac{\mu_{1\cdot} - \mu_{2\cdot}}{m\theta_3^*} \quad (6)$$

where θ_3^* is the least squares slope through all the means (with each mean weighted by the number of observations at that mean), i.e.,

$$\theta_3^* = \frac{\sum_{j=1}^J m_j (x_j - \bar{x}) \left(\frac{\mu_{1j} + \mu_{2j}}{2} \right)}{\sum_{j=1}^J m_j (x_j - \bar{x})^2}, \quad (7)$$

where $\bar{x} = m^{-1} \sum_{j=1}^J m_j x_j$. So $r\{\theta^*(\mathbf{X})\}$ is the average difference in means standardized by the slope of the least squares line through all the means.

Recall that in the original parallel linear model $r\{\theta^*(\mathbf{X})\}$ is interpreted as the log base 10 of the ratio of the dose of the test over the dose of the reference needed to elicit the same mean response. Thus, when $r(\theta^*) = 0$ both drugs elicit the same response.

Now consider a different linear model, the case where both drugs have the same response at the lowest dose, say μ_1 , but each drug has its own possibly different linear response. Let β_1 and β_2 be the slopes of the reference and test drugs respectively. Then $\mu_{ij} = \mu_1 + \beta_i(x_j - x_1)$ so that $m^{-1}(\mu_{1\cdot} - \mu_{2\cdot}) = (\beta_1 - \beta_2)$. Further, $\theta_3^* = \frac{\beta_1 + \beta_2}{2}$, so that

$$r(\theta^*) = \frac{2(\beta_1 - \beta_2)(\bar{x} - x_1)}{(\beta_1 + \beta_2)} = 2(\bar{x} - x_1) \left(\frac{1 - f}{1 + f} \right) \quad (8)$$

where we have let $f = \beta_2/\beta_1$. Notice that as $f \rightarrow 1$ then $r(\theta^*) \rightarrow 0$ as we would wish, since that limiting case denotes equivalent dose response curves for test and reference. As $f \rightarrow 0$ then $r(\theta^*) \rightarrow 2(\bar{x} - x_1)$. We return to this important special case in Section 5.1 where we discuss designing these potency experiments.

4 Simulations

For the simulations we consider only designs with five equally-spaced values of $\log_{10}(\text{Dose})$ (in particular $-1, -0.5, 0, 0.5$, and 1). We considered sample sizes of 20 or 50 corresponding to $m_j = 4$ and $m_j = 10$ replicates at each dose, respectively. We used three different models for the means of each drug group:

Linear: $\mu_{ij} = 4 + 2(x_j + R_i)$

Logistic: $\mu_{ij} = 6 - 4(1 + 10^{2(x_j + R_i)})^{-1}$

Log function: $\mu_{ij} = \log_{10} \left(10^6 - \frac{10^6}{1 + 10^{2(x_j + R_i)}} \right)$

where $R_i = \log_{10}(\Delta) I(i = 1)$, so that for the linear case when $\Delta = \delta$ then the mean model is on the boundary between the null and alternative hypotheses. We tested the alternative that $r(\theta^*) \leq \log_{10}(1.25)$ at the one-sided $\alpha = 0.025$ level. For each mean model we tested at the boundary

between the null and alternative hypotheses as well as the case when $R_i = 0$ such that both groups are the same. We plot the means where $r(\theta^*) = \log_{10}(1.25)$ in Figure 2.

In Table 1 we consider the case where all the variances are the same. In Table 2 we consider the case where the variances are larger for the smaller means as is the case with the data in Figure 1. From the simulations we see that when the linear model of equation 1 holds then all methods give similar results. When that linear model does not hold, then the first three methods (the delta method and Fieller's theorem method on the linear model, and the jackknife method) give very poor power, while the methods that use the sandwich estimator give much better power. From these simulations, we can see no practical difference between the delta method and Fieller's method. In addition, we did simulations where there was no dose response for the test drug, and all methods had zero simulated power (results not shown).

5 Practical Issues

5.1 Designing Studies

In designing studies to measure potency, it is important to ensure that the test does not show non-inferiority when the test drug has no effect at all. Consider the linear model that motivated equation 8. In that case, the reference drug has a linear dose-response and gives the same response as the test drug at the lowest dose. This would be the case if there was some threshold of dose needed to give a response above background, and the lowest dose was just at this threshold for the reference drug. When the test drug gives no response at any dose, its mean is the same as that background response and the slope of the test drug is $\beta_2 = 0$. In this case from equation 8, $r(\theta^*) = 2(\bar{x} - x_1) > 0$. Thus, we want to make sure that the design is such that $2(\bar{x} - x_1) > \log_{10}(\delta)$ otherwise as the number of observations goes to infinity you will always show that the test drug is non-inferior even though the test drug has no dose-response. Similar statements could be made for different values of f . For example, for some $f < 1$, say f_0 , one may want to design the study such that $2(\bar{x} - x_1) \left(\frac{1-f_0}{1+f_0} \right) > \log_{10}(\delta)$, otherwise one may show that the test drug is non-inferior when the slope of the test drug is only f_0 times the slope of the reference (i.e., $\beta_2 = f_0 \beta_1$). In practice, one will want to have $2(\bar{x} - x_1) \left(\frac{1-f_0}{1+f_0} \right) \gg \log_{10}(\delta)$.

For the purposes of calculating sample sizes, we assume that the linear model of equation 1 with homogeneous variance is true. If there is sufficient information about the dose response so that we suspect that the linear model is not approximately true, then it may make sense to work with a transformation of the response for which the linear model approximately holds. Then we could proceed as with the usual linear model but all the parameters in the following equations would refer to the transformed responses.

For the choice of the x_j values, there is a large literature on experimental design (see e.g., Ryan, 2007). From that literature we could choose among the optimal designs. For example, the D-optimal design puts half the observations at the smallest dose and half at the largest within the range of possible doses. We do not recommend using this or other optimal designs, since we want to allow the possibility that the linear model of equation 1 is misspecified. For example if the dose response is logistic and the smallest and largest doses gave similar responses with both reference and test drugs, that does not mean that the responses in the middle range will be similar. Thus, we do not make specific suggestions about the choice of the doses except to say that the doses should be spread out throughout the range of possible doses (see for example, the design for the simulations).

Let σ^2 be the variance at each point. We use the delta method estimate of variance. Let J be the number of unique doses, and suppose there are m_j observations at each dose/drug combination. To keep things simple, we assume that $m_1 = m_2 = \dots = m_J$. Then equation 3 with

the parameters replacing their estimates gives an asymptotic variance expression which we write as $m_1^{-1}\Omega\{r(\theta)\}$, where

$$\Omega\{r(\theta)\} = \sigma^2 \left(\frac{\{r(\theta)\}^2 + 4S_x^2}{2J\theta_3^2 S_x^2} \right), \quad (9)$$

$$\text{and } S_x^2 = J^{-1} \sum_{j=1}^J (x_j - \bar{x})^2.$$

For the sample size calculations we require the variance of $r(\hat{\theta})$ under both the null and the alternative hypotheses. For the null hypothesis we use $r(\theta) = \log_{10}(\delta)$ which is on the boundary between the two hypotheses, and for the alternative hypotheses we use $r(\theta) = 0$ (which happens when the test and reference are the same). Then we can use the usual sample size derivation (see e.g., Lachin, 1981); we get

$$m_1 \geq \frac{(Z_\alpha \sqrt{\Omega\{\log_{10}(\delta)\}} + Z_\beta \sqrt{\Omega\{0\}})^2}{[\log_{10}(\delta)]^2} = \frac{2\sigma^2 \left(Z_\alpha \sqrt{1 + \frac{[\log_{10}(\delta)]^2}{4S_x^2}} + Z_\beta \right)^2}{J\theta_3^2 [\log_{10}(\delta)]^2}$$

where $Z_q = \Phi^{-1}(1 - q)$, e.g., $Z_{0.025} = 1.96$, α is the one-sided significance level, and $1 - \beta$ is the power. This can also be applied to equivalence trials where we usually use two one-sided hypothesis tests, with one test for each direction (see for example, Berger and Hsu, 1996, Chow and Shao, 2002, Chapter 8).

5.2 Assays Below the Limit of Detection

For many assays, there is a limit of detection of the response such that the exact value of some responses are unknown and only known to be smaller than the limit of detection (LD). Many times the LD is the same for each response and some value, say ℓ , is substituted for any response below the LD. In that case, if there is more than one response within a dose/drug combination that is below the LD, then the estimate of the variance will be underestimated. To avoid this underestimation of variance, we suggest the following *ad hoc* fix: if the ij th dose/drug

combination has more than one response below the LD, let $\tilde{\sigma}_{ij}^2$ be given by $\max(\tilde{\sigma}_{ij}^{*2}, \bar{\sigma}^2)$, where $\tilde{\sigma}_{ij}^{*2}$ is the usual sample variance calculated after substituting ℓ for values below the LD, and where $\bar{\sigma}^2$ is the homogeneous variance estimate calculated from the saturated mean model (i.e., the model of equation 5 except with $\varepsilon_{ijk}^* \sim N(0, \sigma^2)$) from the data with all values below the LD excluded from the calculation. In some cases, some of the μ_{ij} parameters will not be estimated during the calculation of $\bar{\sigma}^2$ because all of the responses within that dose/drug combination are below the LD.

6 Application

We now apply the method to the data in Figure 1. First notice that $2(x - x_1) = 1.015 \gg \log_{10}(1.25) = 0.0969$. Thus, it is reasonable to perform tests with $\delta = 1.25$ and we need not worry that a test drug with no dose-response will appear significantly non-inferior (see Section 5.1). Our estimate of $r\{\theta^*(X)\}$ is -0.222 . If we assume that the linear model fits then the interpretation is that the dose of the test needs to be only $10^{-222} = .600$ times the dose of the

reference to obtain the same response. Regardless of whether the linear model fits, since $r\{\hat{\theta}^*(X)\}$ is negative, we estimate that the standardized sum of the expected responses from the reference group is less than the standardized sum of the expected responses from the test group.

Assuming the linear model and using the methods of Section 2, the 95% confidence intervals on $r(\theta)$ are $(-0.534, 0.091)$ for the delta method and $(-0.595, 0.081)$ for Fieller's method both using a t-distribution with $n - 3 = 57$ degrees of freedom. Using the sandwich estimator the 95% confidence intervals on $r\{\theta^*(X)\}$ are $(-0.510, 0.066)$ for the delta method and $(-0.558, 0.074)$ for Fieller's method, where here we use $n - 6 = 54$ degrees of freedom. Notice that the confidence intervals are slightly smaller for the sandwich estimators because these estimators do not assume a linear model and so the lack-of-fit to the linear model does not unduly increase those variance estimates. We can invert the confidence intervals to obtain one-sided p-values (i.e., find the level α such the $100(1 - 2\alpha)\%$ upper confidence limit exactly equals $\log_{10}(1.25) = .0969$). The one-sided p-values are: for the delta method ($p = .023$ for the linear model or $p = .015$ for the sandwich) and for Fieller's method ($p = .020$ for the linear model or $p = .018$ for the sandwich). For both methods the p-value is slightly smaller using the sandwich method. All methods show significant non-inferiority of the test compared to the reference at the 0.025 one-sided level, but the sandwich methods do not require the assumption that expected values of the responses follow a linear model.

7 Discussion

We have shown how the sandwich estimator of variance may be used in potency tests. The parameter that measures potency, $r(\theta^*)$, is defined for all mean models and is interpreted as the average difference in mean responses standardized by the slope of the least squares line on the log dose. We showed how the sandwich estimator of variance may be used to calculate confidence intervals which hold asymptotically regardless of whether the linear model holds. Both the delta method and Fieller's theorem appear to perform adequately, and we see no reason to prefer one over the other.

We have discussed that since $r(\theta^*)$ is a standardized average difference in mean responses care must be made in designing the study so that our proposed method does not show non-inferiority for a test drug with no dose-response. An additional caution in interpretation is that although the proposed method can show "averaged" non-inferiority, this does not imply that the responses to the test drug is non-inferior at each dose. If the test drug is superior to the reference for low doses but inferior for high doses (or vis versa), it is possible that the proposed method can have a high power to show significant averaged non-inferiority. In order to help in the interpretation of the results from the proposed test, the test could be supplemented by comparing the mean difference between the two groups at each dose with pointwise confidence intervals.

Another approach to this problem is to use a profile likelihood method (Ghosh, Datta, Kim, and Sweeting, 2006). The problem with that method (as with the usual application of the linear model through Fieller's theorem and the delta method) is that it assumes that the linear model fits and we want to relax that assumption. Thus, there is no reason to expect the profile likelihood method to work well when the mean model is misspecified, and we have not explored it in this paper.

When the linear model does not hold then the parameter of interest, $r\{\theta^*(X)\}$, is a function of the design matrix X . This means that methods such as the jackknife fail to give good estimates of variance. This failure is due to the nature of the jackknife calculation which repeatedly recalculates $r\{\hat{\theta}^*(X)\}$ after changing the design matrix by eliminating a different row for each jackknife sample. By changing the design matrix for each jackknife sample, each of these

samples is actually estimating a different parameter when the mean model is misspecified, and the jackknife has poor power. We expect similar problems with other methods such as a bootstrap where the design matrix changes with each bootstrap sample data set.

Other models can be put into a linear model by the appropriate transformation. For example, the 4-parameter logistic can put into a linear model form by using a logit transformation (see e.g., Plikaytis, et al 1994); however, the transformation depends on unknown parameters. Further, the details of non-linear models when dealing with misspecification are left to future work.

Acknowledgments

We thank Laura Martin for providing the data listed in Figure 1, and Michael Hughes, the associate editor, and referees for comments on the paper.

References

- Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996;11:283–319.
- Buonaccorsi, JP. Fieller's Theorem. In: Armitage, P.; Colton, T., editors. *Encyclopedia of Biostatistics*. Vol. 2. Vol. 3. Wiley; New York: 2005.
- Callahan JD, Sajjadi NC. Testing the null hypothesis for a specified difference- the right way to test for parallelism. *Bioprocessing Journal* 2003;2:71–77.
- Chow, SC.; Shao, J. *Statistics in Drug Research: Methodologies and Recent Developments*. Marcel Dekker, Inc; New York: 2002.
- Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. *Analysis of Longitudinal Data*. Vol. 2. Oxford University Press; Oxford: 2002.
- Dudley RA, Dewards P, Ekins RP, Finney DJ, McKenzie IGM, Raab GM, Rodbard D, Rogers RPC. Guidelines for immunoassay data processing. *Clinical Chemistry* 1985;31:1264–1271.
- Fay MP, Graudbard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001;57:1198–1206. [PubMed: 11764261]
- Ghosh M, Datta GS, Kim D, Sweeting TJ. Likelihood-based Inference for the Ratios of Regression Coefficients in Linear Models. *Annals of the Institute of Statistical Mathematics* 2006;58:457–473.
- Gottschalk PG, Dunn JR. Measuring parallelism, linearity, and relative potency in bioassay and immunoassay data. *Journal of Biopharmaceutical Statistics* 2005;15:437–463. [PubMed: 15920890]
- Giersing BK, Dubovsky F, Saul A, Denamur F, Minor P, Meade B. Potency assay design for adjuvanted recombinant proteins as malaria vaccines. *Vaccine* 2006;24:4264–4270. [PubMed: 16767804]
- Graybill, FA. *Theory and Application of the Linear Model*. Wadsworth & Brooks/Cole; Pacific Grove, California: 1976.
- Hubert, JJ.; Bohidar, NR.; Peace, KE. Assessment of Pharmacological Activity. In: Peace, KE., editor. *Biopharmaceutical Statistics for Drug Development*. Marcel Dekker, Inc; New York: 1988. Chapter 3
- Lachin JM. Introduction to Sample Size Determination and Power Analysis for Clinical Trials. *Controlled Clinical Trials* 1981;2:93–113. [PubMed: 7273794]
- Lehmann, EL. *Elements of Large Sample Theory*. Wiley; New York: 1999.
- Plikaytis BD, Holder PF, Pais LB, Maslanka SE, Gheesling LL, Carlone GM. Determination of parallelism and nonparallelism in bioassay dilution curves. *Journal of Clinical Microbiology* 1994;32:2441–2447. [PubMed: 7814480]
- Plikaytis BD, Carlone GM. Statistical considerations for vaccine immunogenicity trials Part 1: introduction and bioassay design and analysis. *Vaccine* 2005;23:1596–1605. [PubMed: 15694512]
- Ryan, TP. *Modern Experimental Design*. Wiley; New York: 2007.
- White, H. *Estimation, Inference and Specification Analysis*. Cambridge University Press; Cambridge: 1994.

Appendix

Appendix: Expression for $r\{\theta^*(\mathbf{X})\}$ when both Drugs have the Same Doses

Recall that $\mathbf{X}$ is an $n \times 3$ matrix, where the first column is all ones, the second column is half ones and half zeros with the ones corresponding to the rows associated with reference drug, and the third column is the $x_j = \log_{10}(d_j)$ values associated with each row. We have assumed

that at dose d_j ; $j = 1, \dots, J$ there are m_j observations from each drug. Then $\sum_{j=1}^J m_j \equiv m = n/2$. Let $\bar{x} = m^{-1} \sum_{j=1}^J m_j x_j$ and $w^2 = m^{-1} \sum_{j=1}^J m_j x_j^2$. Thus,

$$\mathbf{X}'\mathbf{X} = m \begin{bmatrix} 2 & 1 & 2\bar{x} \\ 1 & 1 & \bar{x} \\ 2\bar{x} & \bar{x} & 2w^2 \end{bmatrix}$$

and using Mathematica or some other computer algebra software we can show

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{m} \begin{bmatrix} \frac{2w^2 - \bar{x}^2}{2w^2 - 2\bar{x}^2} & -1 & \frac{-\bar{x}}{2w^2 - 2\bar{x}^2} \\ -1 & 2 & 0 \\ \frac{-\bar{x}}{2w^2 - 2\bar{x}^2} & 0 & \frac{1}{2w^2 - 2\bar{x}^2} \end{bmatrix}.$$

Further,

$$\mathbf{X}'\boldsymbol{\mu}(\mathbf{X}) = \begin{bmatrix} \mu_{1\cdot} + \mu_{2\cdot} \\ \mu_{1\cdot} \\ \sum_{i=1}^2 \sum_{j=1}^J m_j x_j \mu_{ij} \end{bmatrix}.$$

Thus, since $\theta^*(\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\mu}(\mathbf{X})$, we have

$$\theta_2^* = \frac{\mu_{1\cdot} - \mu_{2\cdot}}{m}$$

and

$$\theta_3^* = \frac{1}{2m(w^2 - \bar{x}^2)} \left\{ -\bar{x}(\mu_{1\cdot} + \mu_{2\cdot}) + \sum_{i=1}^2 \sum_{j=1}^J m_j x_j \mu_{ij} \right\}$$

which by simplification reduces to equation 7. Equation 6 follows from $r\{\theta^*(\mathbf{X})\} = \theta_2^* / \theta_3^*$.

Figure 1.

Results from an assay comparing a test vaccine to a reference vaccine. Each point represents the ELISA antibody response from a different animal given one of the two drugs. The ELISA detects antibody response to the *MSP1₄₂* protein of the FVO strain of the plasmodium falciparum (malaria) parasite. Test and reference are given at the same three doses but are plotted at slightly higher (reference) or lower (test) doses for differentiation. Bars are the means and 95% confidence intervals by t-test. The two sloping lines are from the linear model.

Figure 2. Simulation means for three different models, where reference is dotted lines and triangular points and test is solid lines and circular points. Plots all have $r(\theta^*) = \log_{10}(1.25) = 0.0969$.

Table 1

Percent rejected at the (one-sided) $\alpha = .025$ level (and average length of 95% confidence intervals) when testing the alternative that $r(\theta^*) \geq \log_{10}(1.25) = .0969$ from 10,000 simulations using the 5 different methods. Variance is homogeneous normal with standard deviation, $\sigma = .3$.

Mean Model	m	$\log_{10}(\Delta)$	$r(\theta^*)$	Delta, Linear Model	Fidler, Linear Model	Jackknife	Delta, Sandwich	Fidler, Sandwich
Linear, size	4	0.0969	0.0969	2.54 (0.192)	2.52 (0.1925)	2.28 (0.1969)	2.60 (0.1934)	2.54 (0.1939)
	10			2.66 (0.1191)	2.64 (0.1192)	2.56 (0.1203)	2.60 (0.1192)	2.66 (0.1193)
Linear, equal	4	0	0	51.37 (0.1914)	51.15 (0.1918)	49.31 (0.1962)	50.68 (0.1926)	50.42 (0.1931)
	10			88.77 (0.1189)	88.67 (0.119)	88.41 (0.1201)	88.78 (0.119)	88.7 (0.1191)
Logistic, size	4	0.1341	0.0969	0.06 (0.2913)	0.06 (0.2929)	0.06 (0.2954)	2.51 (0.1745)	2.47 (0.1749)
	10			0.07 (0.1779)	0.06 (0.1783)	0.06 (0.1781)	2.26 (0.1078)	2.22 (0.1079)
Logistic, equal	4	0	0	17.14 (0.2806)	17.05 (0.282)	15.54 (0.2873)	59.97 (0.1728)	59.78 (0.1732)
	10			65.77 (0.1721)	65.52 (0.1724)	64.64 (0.1737)	94.75 (0.1071)	94.73 (0.1072)
Log function, size	4	0.0954	0.0969	0.36 (0.5698)	0.37 (0.5824)	0.26 (0.583)	2.55 (0.4098)	2.63 (0.4147)
	10			0.33 (0.3488)	0.33 (0.3515)	0.28 (0.3506)	2.49 (0.2524)	2.50 (0.2535)
Log function, equal	4	0	0	3.94 (0.5367)	4.09 (0.5471)	3.36 (0.5519)	15.97 (0.3868)	16.01 (0.391)
	10			13.6 (0.3281)	13.54 (0.3304)	12.80 (0.3317)	35.59 (0.2385)	35.42 (0.2394)

Table 2

Percent rejected at the one-sided $\alpha = .025$ level (and average length of 95% confidence intervals) when testing the alternative that $r(\theta^*) \leq \log_{10}(1.25) = .0969$ from 10,000 simulations using the 5 different methods. Variance is heterogeneous normal with standard deviation, $\sigma = 1.2/\mu_{ij}$.

Mean Model	m	R_i	$r(\theta^*)$	Delta Model	Fieller Model	Jackknife	Delta Sandwich	Fieller Sandwich
Linear, size	4	0.0969	0.0969	2.35 (0.2305)	2.34 (0.2314)	2.09 (0.2371)	2.65 (0.2318)	2.73 (0.233)
	10			2.57 (0.1441)	2.55 (0.1443)	2.48 (0.1453)	2.70 (0.1438)	2.73 (0.1441)
Linear, equal	4	0	0	36.95 (0.2377)	36.72 (0.2386)	34.44 (0.2453)	35.96 (0.2398)	36.07 (0.2412)
	10			72.97 (0.1488)	72.88 (0.149)	72.12 (0.1507)	72.70 (0.149)	72.49 (0.1493)
Logistic, size	4	0.1341	0.0969	0.19 (0.3233)	0.19 (0.3255)	0.18 (0.3292)	2.51 (0.2224)	2.52 (0.2235)
	10			0.24 (0.1978)	0.24 (0.1983)	0.25 (0.1984)	2.65 (0.1377)	2.60 (0.138)
Logistic, equal	4	0	0	14.35 (0.3174)	14.28 (0.3195)	12.83 (0.3258)	38.73 (0.2276)	38.63 (0.2287)
	10			49.95 (0.1946)	49.61 (0.1951)	48.78 (0.1967)	76.54 (0.1406)	76.42 (0.1408)
Log function, size	4	0.0954	0.0969	0.1 (0.506)	0.1 (0.5145)	0.08 (0.5176)	2.26 (0.3146)	2.29 (0.3172)
	10			0.09 (0.3086)	0.09 (0.3105)	0.07 (0.3098)	2.38 (0.1937)	2.42 (0.1942)
Log function, equal	4	0	0	3.1 (0.4783)	3.24 (0.4855)	2.77 (0.4932)	24.26 (0.3005)	24.22 (0.3028)
	10			15.11 (0.2921)	14.97 (0.2937)	14.39 (0.2956)	53.32 (0.1859)	53.06 (0.1864)