An Improved Sample Size Calculation Method for Score Tests in Generalized Linear Models

Yongqiang Tang\textsuperscript{a}, Liang Zhu\textsuperscript{b}, and Jiezhuun Gu\textsuperscript{c}

\textsuperscript{a}Tesaro, Waltham, MA; \textsuperscript{b}Department of Internal Medicine, The University of Texas Health Science Center at Houston, Houston, TX; \textsuperscript{c}Duke Clinical Research Institute, Durham, NC

1. Introduction

Generalized linear models (GLMs) have been commonly used in the analysis of biomedical data (Nelder and Wedderburn 1972; McCullagh and Nelder 1989). Statistical inference in GLMs is often based on the Wald test and the likelihood ratio (LR) test. However, the Wald and LR tests can be liberal in small and moderate samples. In the comparison of two binary proportions, the Wald and LR methods can be anti-conservative under some parameter configurations even when the sample size reaches 200 (Laud and Dane 2014) because the logistic regression overestimates the odds ratio in these studies (Nemes et al. 2009). Similar phenomenon is observed in the analysis of over-dispersed count data using the negative binomial (NB) regression (Aban, Cutter, and Mavinga 2009). The score test has been recommended to control the Type I error rate when the sample size is relatively small. In fact, many widely used methods such as Pearson's chi-squared test, Cochran–Mantel–Haenszel test and Wilcoxon rank sum test are score tests from GLMs.

One concern about the score test is its lower power when compared to the Wald test. In fact, the score test can sometimes be more powerful than the Wald test. Xing et al. (2012) observed that the Wald test from the logistic regression may often miss rare disease-causal variants that can be identified by other asymptotic tests in large case–control association studies. Table 1 presents two scenarios for comparing two binomial proportions on the risk difference metric, in which the score test has higher power than the Wald test. The first scenario tests for superiority when the sample sizes are unbalanced in the two groups. In scenario 2, a noninferiority (NI) test is considered under balanced sample sizes. The results also evidence that it may sometimes be inappropriate to use the power calculation procedure developed for the Wald test to estimate the power of the score test, and vice versa. Technical details on the score test and the exact power calculation can be found in Farrington and Manning (1990) and Tang (2019).

Self and Mauritsen (1988) developed a power and sample size calculation procedure for the score test from GLMs under sequences of contiguous alternatives (Cox and Hinkley 1974). This method generally works well for alternatives close to the null hypothesis. Its accuracy may degrade when the group sample sizes are unbalanced or when the effect size is large (Self, Mauritsen, and O'Hara 1992). Self and Mauritsen’s approach approximates the variance of the score statistic under the null hypothesis by the variance under the alternative hypothesis. This assumption is asymptotically correct under contiguous alternatives, but unlikely to hold at alternatives that are not close to the null hypothesis (Self and Mauritsen 1988).

We propose a modification of Self and Mauritsen’s procedure by taking into account of the variance of the score statistic under both the null and alternative hypotheses. It can greatly improve the performance of the method. For example, Tang (2011) obtained the sample size formula for Wilcoxon rank sum test for ordinal outcomes on basis of the asymptotic variance of the U statistic under both hypotheses, which shows improvements over the formulae derived under contiguous alternatives (Whitehead 1993; Zhao, Rahardja, and Qu 2008). Similar ideas were employed by Farrington and Manning (1990) in the comparison of binary proportions in NI trials. In these simple cases, the score test and its asymptotic distribution can be obtained analytically. In this article, we consider more complex situations where the model contains some nuisance parameters. The score test has been commonly used in the superiority trials. It is less well known how to use the score method to analyze the NI trials. In Section 2, we...
explain how to conduct the NI tests in GLM based on the score method and introduce the modified sample size procedure for both superiority and NI trials via the exemplary dataset approach.

The proposed method is employed to estimate the sample size for the score test from the NB regression in Section 3, and for the score test from the logistic regression with categorical covariates in Section 4. The performance of the proposed method is assessed by numerical examples and compared with some existing procedures.

2. Score Tests in GLM

2.1. Score Test and Score Confidence Interval

In GLMs, the scalar response variables \( y_1, \ldots, y_n \) are assumed to have probability density functions of the form (Nelder and Wedderburn 1972; McCullagh and Nelder 1989)

\[
f(y_i|x_i, \beta, \alpha, \phi) = \exp \left[ \frac{y_i \beta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right],
\]

where \( \theta_i \) is the canonical parameter and \( \phi \) is the dispersion parameter. The mean of \( y_i \) is \( \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i} \), and its variance is \( V_i = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \cdot a(\phi) \). We assume that the covariates are related to the mean \( \mu_i \) via a link function \( g(\mu_i) = \beta_i x_i + \alpha \cdot z_i = g(\mu_i) \), where \( x_i \) is a scalar covariate, the vector \( z_i \) contains other covariates including the intercept, and \( (\beta, \alpha) \) are the regression coefficients. In the analysis of clinical trials, \( x_i = 0 \) or 1 is the treatment status.

Suppose we are interested in testing the hypothesis

\[ H_0 : \beta = \beta_0 \text{ versus } H_1 : \beta \neq \beta_0. \]

The null hypothesis is rejected if \(|Z(\beta_0)| \geq z_{1-\alpha/2}\), where \( z_p \) is the \( p \)th percentile of \( N(0,1) \).

As will be illustrated in Section 3, the score test can be used to test the hypothesis in superiority and NI trials by setting \( \beta_0 \) as the superiority and NI margin. The confidence interval (CI) is often reported to quantify the uncertainty in the estimated effect. The \((1-\alpha)100\%\) score CI for \( \beta \) can be obtained by inverting the score test

\[ \{ \beta : |Z(\beta)| \leq z_{1-\alpha/2}. \]

Statistical decision can be made equivalently based on the score CI. The null hypothesis is rejected if the score CI does not contain the null hypothesis value.

2.2. Asymptotic Distribution of the Score Statistic

The score test and its asymptotic distribution usually have explicit analytic expressions in the simple two-group comparison if the model does not contain an unknown dispersion parameter. Please refer to Farrington and Manning (1990) and Tang (2019) for examples. We consider more general cases where the vector of nuisance parameters \( \lambda \) contains other parameters in addition to an intercept term.

In general, \( \hat{\lambda} \) is not a consistent estimate of \( \lambda \) under the restriction of \( \beta = \beta_0 \). It will converge to the limiting value \( \lambda^* \) defined as the solution to the following equation (Self and Mauritsen 1988)

\[
\lim_{n \to \infty} E[n^{-1} S_{\lambda}(\beta_0, \hat{\lambda})] = 0. \tag{3}
\]

We estimate \( \lambda^* \) by adapting the method of Lyles, Lin, and Williamson (2007). We firstly construct an exemplary dataset consisting of records for every possible combination of the covariates and outcomes. Each record has a weight that represents the frequency of the covariate and outcome in the population. A weighted regression is fitted to the exemplary dataset using the standard statistical software. We assume that all covariates are categorical. A continuous covariate can be discretized using a large number of categories. Suppose there are a finite number of distinct covariate configurations \( \{(z_{k}, x_{k}) ; k = 1, \ldots, m\} \), and the proportion of each configuration is \( \pi_k \) in the population. Suppose the response variable takes \( J \) possible values \( \{y_1, \ldots, y_J\} \). We can estimate \( \lambda^* \) by fitting the null model to the following dataset with \( mJ \) observations, where \( w_{ij} = \pi_k \Pr(Y = y_j | z_k, x_k, \beta, \lambda) \) is the weight attached to each observation, and the total weight in all observations is \( \sum_{j=1}^{J} \sum_{i=1}^{m} w_{ij} = 1 \).

Table 1. Two scenarios with higher power in the score test than in the Wald test.

| Group size | True proportion | Hypothesis | Exact power (%) |
|------------|----------------|------------|----------------|
| n1 = 60   | n0 = 30        | p1 = 0.1   | p0 = 0.3       | Scorea | Waldb | Waldc |
| 80         | 80             | 0.35       | 0.4            | 67.33  | 60.81 | 65.28 |

Note: a Score test defined in Farrington and Manning (1990, eq. (2)). b Wald (Z) and Wald2 (Z) are the Wald tests from the binomial regression, respectively, with identity and logit link functions.
\begin{align*}
x_1 & \quad z_1 \quad y_1 \quad w_{11} \\
\ldots & \\
x_m & \quad z_m \quad y_1 \quad w_{1m} \\
\ldots & \\
x_m & \quad z_m \quad y_1 \quad w_{1m} \\
\end{align*}

Lyles, Lin, and Williamson (2007) approach is slightly different. It requires a much larger dataset, and can only estimate the power at a given sample size. The total weight in Lyles, Lin, and Williamson (2007) approach is equal to the total sample size \( N \). Let's give a simple example of comparing two binary proportions with \( \Pr(y = 1 | x = 1) = 0.8 \) and \( \Pr(y = 1 | x = 0) = 0.4 \). Suppose \( N = 100 \), and 75% patients are assigned to the experimental arm (\( x = 1 \)). Then \((m, f) = (2, 2)\). Our approach includes four observations: \((x = 1, y = 1, w = 0.6)\), \((x = 1, y = 0, w = 0.15)\), \((x = 0, y = 1, w = 0.1)\), and \((x = 0, y = 0, w = 0.15)\). In Lyles, Lin, and Williamson (2007) approach, the dataset consists of 200 pseudo-observations with 75 copies of \((x = 1, y = 1, w = 0.8)\) and \((x = 1, y = 0, w = 0.2)\) for subjects in the experimental arm, and 25 copies of \((x = 0, y = 1, w = 0.4)\) and \((x = 0, y = 0, w = 0.6)\) for placebo subjects. If observations with the same \((x, y)\) are combined by adding up their weights, the dataset in Lyles, Lin, and Williamson (2007) approach becomes a dataset with four observations \((x = 1, y = 1, w = 60)\), \((x = 1, y = 0, w = 15)\), \((x = 0, y = 1, w = 10)\), and \((x = 0, y = 0, w = 15)\). The ratio of the weights for observations with the same \((x, y)\) is \(N:1\) between Lyles, Lin, and Williamson (2007) approach and our approach.

Our method is more convenient and potentially more accurate than Lyles, Lin, and Williamson (2007) approach. In Lyles, Lin, and Williamson (2007) method, one needs to guess the sample size, construct the exemplary dataset, fit the null model, and estimate the power at the given sample size. The whole process needs to be repeated if the sample size changes. In theory, \( \lambda^* \) remains unchanged, and the noncentrality parameter of the Wald, score, or LR test or its square change proportionally if we increase or decrease the total sample size. Therefore, the power and sample size calculation can be implemented by first fitting the model at a fixed sample size, and then using analytic methods to adjust the noncentrality parameter and solve the power or sample size equations accordingly. We fit the null model using SAS Proc Genmod with the FREQ option to incorporate the weight. One shall not use the Weight option in the Genmod procedure since it is used to adjust for the dispersion parameter. Because the Genmod procedure truncates the weight to an integer, we multiply all the weights by a large value (say \(10^9\)) to minimize the effect of truncation. Lyles, Lin, and Williamson (2007) fit the model at the sample size for the trial, which is typically small (e.g., below 1000). The weight after truncation in Lyles, Lin, and Williamson (2007) method may no longer represent the frequency of the covariate and outcome in the population, and the estimation of \( \lambda^* \) can be inaccurate.

In Self and Mauritsen (1988), an exemplary dataset contains \( m \) records for all possible combinations of the covariates, where the weight is the frequency of the covariates, and the response outcome is the expected value of response at the covariate configuration. The data structure may not be acceptable by some statistical software packages, and does not allow the estimation of the dispersion parameter \( \phi^* \).

In GLMs, the inference is made by assuming the covariates are known and fixed, but the covariates are typically unobserved at the design stage of a clinical trial. For example, although gender is fixed for each patient, it will be treated as unknown at the design stage since we do not know which patients will be enrolled. We firstly derive the mean and variance of the score statistic given the covariates, which are then averaged over all possible combinations of the covariates.

Let \( S^k(\beta, \lambda) \) denote the contribution to the score function from a subject with covariate \((x_k, z_k)\). Let \( E^k(\beta, \lambda) \) and \( V^k(\beta, \lambda) \) represent, respectively, the mean and variance of \( S^k(\beta, \lambda) \) under the true model (1). Let \( E = (E_0, E_1)^T = \sum_{k=1}^m \pi_k E^k(\beta, \lambda) \) and \( V = \sum_{k=1}^m \pi_k V^k(\beta, \lambda) \). Note that \( E_0 = 0 \) by Equation (3). The asymptotic distribution of \((S_{\beta_0}(\beta, \lambda), S_{\lambda}(\beta, \lambda)) \) is given by

\[
n^{1/2} \left[ n^{-1} S_{\beta_0}(\beta, \lambda) - E_{\beta_0} \right] \sim N(0, V).
\]

Let \( J^k(\beta, \lambda) = -\left[ \frac{\partial^2 \log f(y | x_k, z_k, \beta, \lambda)}{\partial \beta^2} \right] \) denote the contribution to the observed information matrix from a subject with covariate \((x_k, z_k)\). Let \( \tilde{I}^k(\beta, \lambda) \) be the expectation of \( J^k(\beta, \lambda) \) under the true model (1), and \( \tilde{I}(\beta, \lambda) = \sum_{k=1}^m \pi_k \tilde{I}^k(\beta, \lambda) \). By the Taylor series expansion, we get

\[
n^{-1/2} S_{\beta_0}(\beta, \lambda) \approx n^{-1/2} S_{\beta_0}(\beta, \lambda) - \tilde{I}_{\beta\beta}(\beta, \lambda) \tilde{I}^{-1}_{\beta\beta}(\beta, \lambda) n^{-1} S_{\beta_0}(\beta, \lambda).
\]

Combining Equations (4) and (5) yields the asymptotic distribution of the score statistic

\[
n^{-1/2} S_{\beta_0}(\beta, \lambda) \sim N(0, \sigma_1^2),
\]

where \( \sigma_1^2 = \text{AVA}^\prime A \), \( A = [1, -\tilde{I}_{\beta\beta}(\beta, \lambda) \tilde{I}_{\beta\beta}^{-1}(\beta, \lambda)] \). In the special case considered by Self and Mauritsen (1988), \( \tilde{I}(\beta_0, \lambda) \) is identical to the Fisher information matrix \( I(\beta_0, \lambda) \) and \( \tilde{I}_{\beta\beta}(\beta_0, \lambda) = I_{\beta\beta}(\beta_0, \lambda) \) for the exemplary dataset under \( H_0 \), and therefore \( A = [1, -I_{\beta\beta}(\beta_0, \lambda) I_{\beta\beta}^{-1}(\beta_0, \lambda)] \).

As \( n \to \infty \), the null variance \( n^{-1} V_0 = n^{-1} \{ I_{\beta\beta}(\beta_0, \lambda) - I_{\beta\beta}(\beta_0, \lambda) I_{\lambda\lambda}^{-1}(\beta_0, \lambda) I_{\beta\lambda}(\beta_0, \lambda) \} \) converges in probability to

\[
\sigma_0^2 = I_{\beta\beta}(\beta_0, \lambda) - I_{\beta\beta}(\beta_0, \lambda) I_{\lambda\lambda}^{-1}(\beta_0, \lambda) I_{\beta\lambda}(\beta_0, \lambda).
\]

### 2.3. Power and Sample Size Formulae

The power of the score test (2) is given by

\[
P = \Phi \left( \frac{\sqrt{n} |E_1|}{\sqrt{\sigma_1^2 - z_1 - \frac{\sigma_0^2}{\sigma_1^2}}} \right),
\]
where $\Phi(\cdot)$ is the standard normal distribution function, and $\sigma_0^2$ and $\sigma_1^2$ are defined in Equations (6) and (7). Inverting (8) yields the sample size

$$N_{\text{new}} = \frac{(z_{1-\alpha/2}\sigma_0 + z_p\sigma_1)^2}{E^2_{\phi}}. \tag{9}$$

Self and Mauritsen (1988) method is formulated on basis of the noncentral chi-squared distribution, and the power and sample size estimates can be well approximated by

$$P_{\text{SM}} = \Phi \left( \frac{\sqrt{n}[E_\beta] - z_{1-\frac{\alpha}{2}}}{\sqrt{\sigma_0^2}} \right), \tag{10}$$

$$N_{\text{SM}} = \frac{(z_{1-\alpha/2} + z_p)^2\sigma_0^2}{E^2_{\phi}}.$$  

It assumes $\sigma_0^2 \approx \sigma_1^2$. The assumption holds under a sequence of contiguous alternatives. Self, Mauritsen, and O’Hara (1992) showed that the performance of the Self and Mauritsen (1988) procedure may degrade when the effect size is large or when the group sample sizes are unbalanced.

It is generally easier to compute $\sigma_1^2$ than $\sigma_0^2$. Under contiguous alternatives, the power and sample size can also be calculated as

$$P_{\text{SM}} = \Phi \left( \frac{\sqrt{n}[E_\beta] - z_{1-\frac{\alpha}{2}}}{\sqrt{\sigma_0^2}} \right), \tag{11}$$

$$N_{\text{SM}} = \frac{(z_{1-\alpha/2} + z_p)^2\sigma_0^2}{E^2_{\phi}}.$$  

### 3. Sample Size for NB Regression

The NB regression has been widely used to analyze overdispersed count data and recurrent event data. The NB distribution can be written as a Poisson-gamma mixture. If $Y$ follows a Poisson distribution with mean $\mu \epsilon$, where $\epsilon$ is gamma distributed with mean 1 and variance $\kappa$, the marginal distribution of $Y$ is NB($\mu, \kappa$)

$$\text{Pr}(Y = y|\mu, \kappa) = \frac{\Gamma(y + 1/\kappa)}{y!\Gamma(1/\kappa)} \left[ \frac{\kappa \mu}{1 + \kappa \mu} \right]^y \left[ \frac{1}{1 + \kappa \mu} \right]^{1/\kappa},$$

$$y = 0, 1, 2, \ldots, \tag{12}$$

where $\Gamma(\cdot)$ is the Gamma function.

Suppose in a trial, $n$ subjects are assigned randomly to either the experimental ($g = 1$) or control ($g = 0$) treatment group. Let $n_g$ be the number of subjects in group $g$. We assume the planned treatment duration is $\tau_c$ for each subject, but subjects may discontinue the study with the loss-to-follow-up distribution $G(t)$. Let $t_i$ be the follow-up time, and $y_i$ the number of events for subject $i$. Then $y_i|g_i = g, t_i \sim \text{NB}(\lambda_g t_i, \kappa)$, where $\lambda_g$ is the event rate in group $g$.

Suppose a lower event rate indicates better health status. In a superiority trial, the purpose is to demonstrate that the experimental treatment can reduce the event rate relative to the control treatment. The hypothesis can be written as

$$H_0 : \frac{\lambda_1}{\lambda_0} = 1 \text{ versus } H_1 : \frac{\lambda_1}{\lambda_0} < 1. \tag{13}$$

In a NI trial, the objective (Tang 2017, 2018b) is to show that the test treatment is not materially less efficacious than a standard control treatment by proving $\lambda_1/\lambda_0 < M_0$, where $M_0$ is the prespecified margin that is bigger than 1, but close to 1. The hypothesis can be written as

$$H_0 : \frac{\lambda_1}{\lambda_0} = M_0 \text{ versus } H_1 : \frac{\lambda_1}{\lambda_0} < M_0. \tag{14}$$

Mathematically, the superiority trial can be viewed as a special case of the NI trial by setting $M_0 = 1$.

Let $\beta = \log(\lambda_1/\lambda_0)$, $\beta_0 = \log(M_0)$, and $\alpha = \log(\lambda_0)$. Since $\mu_1 = \alpha + \log(t_i) + g_i\log(M_0)$ under $H_0$, the null model can be easily fitted using standard software packages (e.g., SAS Proc Genmod) by setting the offset as $\log(t_i)$ for subjects in the control group, and $\log(M_0 t_i)$ for subjects in the experimental arm. The score test can be written as

$$Z = \frac{\sum_i (y_i - \bar{\mu}_i) \bar{\mu}_i}{\tilde{d}_0 + d_1}, \tag{15}$$

where $\bar{\mu}_i = \exp(\hat{\alpha} + \beta_0 g_i) t_i$, and $\tilde{d}_g = \sum_i (y_i = g) \bar{\mu}_i/(1 + \hat{\kappa} \hat{\mu}_i)$ for $g = 0$ and 1.

The power and sample size can be calculated using the procedure described in Section 2. The expressions for the score function, observed and expected information matrix are given in Lawless (1987, eqs. (2.3)–(2.8)). In our implementation, the continuous time to follow-up is approximated by a categorical variable with $L = 100$ levels.

$$t = \begin{cases} t_L = \tau_c & \text{with probability } p_L = 1 - G(\tau_c), \\ t_l = G^{-1} \left( \frac{p_l(0.5)}{L-1} \right) & \text{with probability } (1 - p_L)/(L-1) \text{ for } l = 1, \ldots, L-1. \end{cases} \tag{16}$$

The final result is insensitive to the choice of $L$ if $L$ is not too small. We allow the loss-to-follow-up distribution to differ by the treatment group. There are $m = 2L$ possible combinations of the values for the treatment and time to follow-up. We truncate the number of response categories at a large number $J = 200$ so that $\Pr(y_j \geq J) < 10^{-5}$. The full exemplary dataset consists of $mL$ observations. The weights for the $mL = 2LJ$ observations are calculated according to the treatment allocation ratio and the true distribution defined in Equation (12). It requires the specification of the dispersion parameter $\kappa$ and the event rates for each group. As mentioned in Section 2.2, we multiply all the weights by a large value (say $10^5$) to minimize the effect of truncation since the SAS Genmod procedure truncates the weight to an integer.

When all subjects have equal follow-up time ($t_{ij} \equiv t$), the method can be slightly simplified with $m = 2$. In this article, we focus on the analysis of recurrent events. The simplified procedure is also suitable for other types of overdispersed counts such as the number of magnetic resonance imaging lesions in multiple sclerosis trials. Let $\mu_g$ and $\lambda_g$ be, respectively, the expected and observed mean count in group $g$. For recurrent events, $\mu_g = \lambda_g t$. The score test (15) reduces to

$$W_i = \frac{\tilde{y}_1 - M_0 \tilde{y}_0}{\sqrt{n_1^{-1}(\hat{\mu}_1 + \hat{\kappa} \hat{\mu}_1^2) + n_0^{-1}(\hat{\mu}_0 + \hat{\kappa} \hat{\mu}_0^2)}}, \tag{17}$$

where $\tilde{y}_g = \sum_i (y_i = g) \hat{\mu}_i/(1 + \hat{\kappa} \hat{\mu}_i)$ for $g = 0$ and 1.
where $\hat{\mu}_1 = M_0\hat{\mu}_0$ is the MLE under $H_0$. Test (17) is similar to the test of Farrington and Manning (1990, eq. (7)) for assessing the relative risk between two binomial proportions. In this special case, the sample size in the control arm is

$$n_0 = \left[ \frac{z_{1-\alpha/2} \sqrt{V(\kappa^*, \mu_0^*, \mu_1^*)} + z_\theta \sqrt{V(\kappa, \mu_0, \mu_1)}}{(\mu_1 - M_0\mu_0)^2} \right]^2,$$

(18)

where $V(\kappa, \mu_0, \mu_1) = \frac{\theta'}{\tau(\theta')^2} + (\mu_0 + \kappa \mu_1^2)$, $\theta = n_1/n_0$ and $(\mu_0^*, \mu_1^*, \kappa^*)$ can be estimated by the method of moments,

$$\mu_0^* = \mu_1^* = \bar{\mu} = \frac{\theta \mu_1 + \mu_0}{\theta + 1} \text{ and } \kappa^* = \frac{\kappa(\theta \mu_1^2 + \mu_0^2)}{(\theta + 1)\bar{\mu}^2} + \frac{\theta(\mu_1 - \mu_0)^2}{(\theta + 1)^2\bar{\mu}^2},$$

(19)

where $\kappa^* > \kappa$ if $\mu_1 \neq \mu_0$, and $\kappa^*$ is the solution to

$$n_1 \frac{E(y_1 - \mu_1^*)^2}{\mu_1^*(1 + \kappa^* \mu_1^*)} + n_0 \frac{E(y_0 - \mu_0^*)^2}{\mu_0^*(1 + \kappa^* \mu_0^*)} = n_1 + n_0.$$

It would be interesting to compare the proposed method with that recently developed by Zhu and Lakkis (2014) and Zhu (2017) for superiority and NI trials because they use a similar idea to the score test. The approaches of Zhu and Lakkis (2014) and Zhu (2017) are based on the statistic $\log(\hat{\mu}_1/\hat{\mu}_0) - \log(M_0)$ instead of the score statistic

$$n_0 = \left[ \frac{z_{1-\alpha/2} \sqrt{V(\kappa^*, \mu_0^*, \mu_1^*)} + z_\theta \sqrt{V(\kappa, \mu_0, \mu_1)}}{[\log(\mu_1/\mu_0) - \log(M_0)]^2} \right]^2,$$

(20)

where $h(\kappa, \mu_0, \mu_1) = \frac{1}{\mu_0^*} + \frac{1}{\mu_1^*} + \frac{1+\theta}{\theta} \kappa$, $\kappa$ is assumed to be known, $a = -\kappa M_0(1 + \theta)$, $b = \kappa(\mu_0 M_0 + \theta \mu_1) - (1 + \theta M_0)$, $c = \mu_0 + \theta \mu_1$, and $\mu_0^{**} = (\mu_1 - \sqrt{b^2 - 4ac})/2a$ and $\mu_1^{**} = M_0\mu_0^{**}$ are the limiting values of the restricted MLE at given $\kappa$. Zhu and Lakkis (2014) and Zhu (2017) implicitly make two approximations. First, the follow-up time is set to their mean values (i.e., $t_i = \bar{t}$) for all individuals, leading to underestimated variance of $\log(\hat{\mu}_1/\hat{\mu}_0) - \log(M_0)$ under both $H_0$ and $H_1$ (this can be proved by using the inequality in Tang (2015, Appendix A.2)). Second, it approximates $\kappa^*$ by $\kappa$, and the null variance of $\log(\hat{\mu}_1/\hat{\mu}_0) - \log(M_0)$ is usually underestimated since $\kappa^*$ obtained under the null hypothesis in the score approach tends to be larger than $\kappa$ particularly when the treatment effect is large. This is shown in Equation (19) for superiority trials when all subjects have equal follow-up time. The phenomenon is analogous to the comparison of two groups with continuous outcomes, in which the variance estimate based on the pooled outcomes $y_0 \sim N(\mu_0, \sigma^2)$ and $y_1 \sim N(\mu_1, \sigma^2)$ tends to overestimate the true variance if the mean difference is ignored. Therefore, Zhu–Lakkis’s approach tends to underestimate the sample size. In superiority trials ($M_0 = 1$) with equal treatment allocation, Zhu–Lakkis’s sample size estimate is strictly smaller than the lower sample size bound of Tang (2015) for the Wald test from the NB regression (Tang 2017).

Below we present several examples to illustrate the proposed method.

**Example 1.** Chronic granulomatous disease (CGD) is a rare inherited disorder of the immune system, characterized by recurrent pyogenic infections. Suppose we plan to design a two-arm CGD trial to assess the effect of an experimental treatment on the infection rate. Some parameters are estimated from a CGD trial analyzed by Matsui (2005) and Tang (2018a). The historical trial enrolled $n = 128$ eligible patients. It was terminated early for efficacy based on an interim analysis. In the trial, $14$ (22.2%) out of 63 patients in the gamma interferon group and $30$ (46.2%) out of 65 patients on placebo had at least one serious infection. We analyze the number of repeated infections using the NB regression. The event rate ratio between two treatments based on the Wald statistic is $0.3566$ (95% CI: [0.1934, 0.6575]). Figure 1 plots the chi-square statistic (i.e., $Z^2(\beta)$ from the score test) as a function of $M_0$. The score CI is [0.1957, 0.6681], which corresponds to the region $\{\beta : Z^2(\beta) \leq 3.814\}$. The score CI is slightly wider than the Wald CI.

We estimate the sample size at the following parameter values. The infection rate is $\lambda_0 = 1.1$ infections per year in the

**Figure 1.** Plot of the chi-square statistic ($Z^2(\beta)$ from the score test) as a function of $M_0$ in a CGD trial.
control arm, and \( \kappa = 0.9 \), which are close to the unconstrained MLE \((\lambda_0 = 1.07, \hat{\kappa} = 0.91)\) from the analysis of the historical CGD trial. Suppose the experimental treatment can reduce the infection rate by 60% (i.e., \( \lambda_1/\lambda_0 = 0.4 \)). The target power is 80% or 90%, and the two-sided significance level is \( \alpha = 0.05 \). The treatment allocation ratio is 1 : 1. The planned treatment duration is \( \tau_c = 1 \) or 3 years for each subject, but subjects may discontinue the trial early with a \( w_c = 25\% \) chance and the loss to follow-up is exponentially distributed. We also assess the performance of the proposed method at other parameter values \((\lambda_0 = 0.8, \kappa = 1.2, w_c = 0)\).

We compare several sample size procedures for the NB regression. In Zhu and Lakkis (2014), three methods were proposed to evaluate the variance of the test statistic under \( H_0 \). We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE. We evaluate only the approach recommended by the authors, which the null variance is calculated based on the approximate restricted MLE.

Because the treatment effect is quite large \((\lambda_1/\lambda_0 = 0.4)\), the variances of the score statistic under \( H_0 \) and \( H_1 \) are not close. Self and Mauritsen's method (Equation (10)) underestimates the required size while formula (11) overestimates the sample size. Tang (2017, 2018b) demonstrated that the method of Zhu and Lakkis (2014) underestimates the required size if the follow-up time varies across patients. Table 2 indicates that Zhu–Lakkis’s formula still underestimates the sample size and overestimates the power even if all patients have equal follow-up time. At the target 90% power, the sample size estimates are quite close for the Wald and score tests. But at the target 80% power, the score test requires about 7.0%–8.5% more subjects than the Wald test when \( \lambda_0 = 1.1, \tau_c = 3, \) and \( w_c = 0 \) or 25% in this example.

| Table 2. Estimated sample size at the target 90% power and estimated power for the score test from NB regression in superiority trials. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( w_c \) (%)   | \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( \text{Nominal power} \) |
| \( w_c \) (%)   | \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \) |
| \( \lambda_0 \) | \( \kappa \) | \( \tau_c \) | \( N_{\text{new}} \) | \( N_{\text{SM}} \) | \( N_{\text{IO}} \) | \( ZL^a \) | \( \text{Wald}^b \) | \( P_{\text{new}} \) | \( P_{\text{SM}} \) | \( P_{\text{IO}} \) | \( ZL^a \)|
where exp \( S \) with Self and Mauritsen (1988) investigated the sample size estimation for comparing two binomial proportions \((g_1 = 0 \ or \ 1)\) using logistic regression while controlling for a categorical covariate \(z_i\) with \(S = 2\) levels. We call \(z_i\) a stratum variable, and revisit the problem with \(S \geq 2\) strata. Suppose for subjects in stratum \(s\), \(y_i\) follows a Bernoulli distribution with the probability of success

\[
Pr(y_i = 1|g_s, z_i = s) = \frac{\exp(\alpha_0 + \alpha_s + \beta g_s)}{1 + \exp(\alpha_0 + \alpha_s + \beta g_s)},
\]

for \(s = 1, \ldots, S\), \(g_s\) is the odds ratio associated with the group status \(g_s\) among subjects from the same stratum, and \(\alpha_s\) is the odds ratio for subjects in stratum \(s\) relative to subjects with the same \(g_s\) from stratum \(1\) (\(\alpha_1 = 0\)).

\( \alpha \) Model (21) can be used to analyze data from both prospective clinical trials and retrospective case–control studies. The objectives are different, but the underlying statistical problems are similar in the two types of studies. Table 4 displays the data format for both studies. In a clinical trial, we compare the proportion of responders between two treatment groups, where \(x_{g_s}\) is the number of responders among \(n_{g_s}\) subjects assigned to treatment group \(g\) in stratum \(s\). In the case–control study, the aim is to compare the proportion of exposed between the case and control groups, where \(x_{g_s}\) is the number of exposed subjects among \(n_{g_s}\) case \((g = 1)\) or control \((g = 0)\) subjects in stratum \(s\).

In fact, the score test for testing \(H_0: \beta = 0\) has explicit analytic expression

\[
Z = \frac{\sum_{s=1}^{S} \frac{n_{1g_s}n_{0g_s}}{n_s} (\hat{p}_{g_s} - \hat{p}_0)}{\sqrt{\sum_{s=1}^{S} \frac{n_{1g_s}n_{0g_s}}{n_s} \hat{p}_s(1 - \hat{p}_s)}},
\]

where \(\hat{p}_g = x_{g_s}/n_{g_s}\), \(n_s = n_{0g_s} + n_{1g_s}\), and \(\hat{p}_s = (x_{0s} + x_{1s})/n_s\). The power and sample size formulae in Section 2.3 can be used by setting

\[
E_\beta = \sum_{s=1}^{S} t_s \rho_s (1 - \rho_s) (p_{1s} - p_{0s}),
\]

\[
\sigma_0^2 = \sum_{s=1}^{S} t_s \rho_s (1 - \rho_s) \rho_{s1}^2 (1 - \rho_{s0}^2),
\]

\[
\sigma_1^2 = \sum_{s=1}^{S} t_s \rho_s (1 - \rho_s) [(1 - \rho_s) p_{1s} (1 - p_{0s}) + \rho_s p_{0s} (1 - p_{0s})],
\]

where \(p_{g} = E(\hat{p}_g)\) is the true response rate, \(n = \sum_{s=1}^{S} n_s\) is the total sample size, \(t_s = \frac{n_s}{n}\) is the proportion of subjects contributed by stratum \(s\), \(\rho_s = \frac{n_{1s}}{n_s}\) is the proportion of subjects from group \(g = 1\) in stratum \(s\), and \(p_{s1}^* = \rho_s p_{1s} + (1 - \rho_s) p_{0s}\). The technical details are omitted here. We will extend the method to sample size determination for the stratified score tests in superiority, NI and equivalence trials on basis of the risk difference, relative risk or odds ratio effect measures in Tang (2019), and a general proof will be presented in that article. The score statistic (22) is identical to Cochran (1954) statistic, and the power formula (8) is identical to that derived by Nam (1992) for Cochran’s test although Nam (1992) considered only the case–control studies.
We conduct two simulation studies to compare several methods.

**Example 3.** Suppose there are two strata. Let \((\Pi_1, \Pi_2, \Pi_3, \Pi_4)\) denote, respectively, the proportion of subjects with \((g, z_i) = (0,1), (0,2), (1,1), \) and \((1,2)\). Thus, \(t_1 = \Pi_1 + \Pi_3, t_2 = \Pi_2 + \Pi_4, \rho_1 = \Pi_1 / T_1, \) and \(\rho_2 = \Pi_2 / T_2\). We set \((\Pi_1, \Pi_2, \Pi_3, \Pi_4) = (0.25, 0.25, 0.25, 0.25)\) and \((0.4, 0.1, 0.1, 0.4)\). The odds ratio associated with the stratum is \(\exp(\alpha_1) = 2\), and the odds ratio for the exposure is \(\exp(\beta) = 2\) or 3. The overall response rate in the study population

\[
\hat{\mu} = \frac{\Pi_1}{1 + \exp(-\alpha_0)} + \frac{\Pi_2}{1 + \exp(-\alpha_0 - \alpha_1)} + \frac{\Pi_3}{1 + \exp(-\alpha_0 - \beta)} + \frac{\Pi_4}{1 + \exp(-\alpha_0 - \alpha_1 - \beta)}
\]

is set to 0.15 and 0.5, which is used to derive \(\alpha_0\). The setup is similar to that reported in Self and Mauritsen (1988, Table 2).

Table 5 displays the power and sample size results at the target power 80%, 90%, and 95%, and two-sided Type I error 0.05. The analytic expression (10) gives the same sample size estimates as that reported in Self and Mauritsen (1988, Table 2) in all cases at \(\hat{\mu} = 0.5\), but slightly larger estimates in all cases at \(\hat{\mu} = 0.15\) possibly due to rounding errors. This verifies the validity of the power and sample size calculation based on the simpler equation (23) in the logistic regression. The simulated power is estimated at the sample size \(N_{\text{new}}\) from the proposed method based on \(10^6\) simulated datasets. There is more than 95% chance that the simulated power lies within \(2\sqrt{0.8 \times 0.2/10^6} = 0.08\%\) of the true power.

All methods perform well possibly because the sample sizes are balanced overall between two groups \((Pr(g = 1) = Pr(g = 0) = 0.5)\) although when \((\Pi_1, \Pi_2, \Pi_3, \Pi_4) = (0.4, 0.1, 0.1, 0.4)\), the sample sizes are highly unbalanced between two groups within each stratum. We compare the methods by assessing how close the estimated nominal power is to the empirical power at a given sample size. There are more cases with >1% difference between the nominal and simulated power estimates by formulas (10) and (11) than by formula (8). In nearly all cases at \((\Pi_1, \Pi_2, \Pi_3, \Pi_4) = (0.4, 0.1, 0.1, 0.4)\), the nominal power by formula (8) is closer to the simulated power than that by Equation (10).

**Example 4.** Self, Mauritsen, and O’Hara (1992) observed that the Self and Mauritsen (1988) method degrades when the sample sizes are highly unbalanced between two groups. In this simulation, the setup is similar to that reported in Self, Mauritsen, and O’Hara (1992, Table 1). We set \(\sigma = Pr(g = 1) = 0.05, 0.5, 0.75\), \(Pr(Z = 2|x_1 = 1) = 0.8\), and \(Pr(Z = 2|x_1 = 0) = 0.2\). Thus, \((\Pi_1, \Pi_2, \Pi_3, \Pi_4) = (0.8(1 - \pi), 0.2(1 - \pi), 0.2\pi, 0.8\pi)\). Note that \(\pi = 0.5(0.75)\) corresponds to the 1:1 (3:1) treatment allocation ratio, which is commonly used in clinical trials. The scenario \(\sigma = 0.05\) may arise in case–control studies or in genetic studies when a small proportion of subjects carry the risk genotypes (Tang 2011). The true odds ratio is \(\exp(\alpha_1) = 2\) for stratum and \(\exp(\beta) = 2\) for exposure. The overall response rate in the study population

\[
\hat{\mu} = \frac{\Pi_1}{1 + \exp(-\alpha_0)} + \frac{\Pi_2}{1 + \exp(-\alpha_0 - \alpha_1)} + \frac{\Pi_3}{1 + \exp(-\alpha_0 - \beta)} + \frac{\Pi_4}{1 + \exp(-\alpha_0 - \alpha_1 - \beta)}
\]

is set to 0.02 and 0.15. Because the sample size estimates vary greatly by methods, we evaluate the nominal power and empirical power based on \(10^6\) simulations at both the sample sizes from the proposed method and Self and Mauritsen (1988) methods. We repeat

---

**Table 4.** Binary outcomes from stratified clinical trials and case–control studies.

| Event | Placebo g = 0 | Active g = 1 | Placebo g = 0 | Active g = 1 |
|-------|---------------|-------------|---------------|-------------|
| Yes   | \(x_{10}\)    | \(x_{11}\)   | \(x_{00}\)    | \(x_{01}\)   |
| No    | \(n_{10}\)    | \(n_{11}\)   | \(n_{00}\)    | \(n_{01}\)   |

| Exposure status | Control g = 0 | Case g = 1 | Control g = 0 | Case g = 1 |
|-----------------|---------------|------------|---------------|------------|
| Yes             | \(x_{10}\)    | \(x_{11}\)   | \(x_{00}\)    | \(x_{01}\)   |
| No              | \(n_{10}\)    | \(n_{11}\)   | \(n_{00}\)    | \(n_{01}\)   |

---

**Table 5.** Power and sample size estimate for the score test from logistic regression with equal group sample sizes.

| Target | Estimated size | Nominal power (%) | Estimated size | Nominal power (%) |
|--------|----------------|-------------------|----------------|-------------------|

---

NOTE: Simulated power (SIM) are evaluated at \(N_{\text{new}}\) based on 1,000,000 simulated datasets.
Table 6. Power and sample size estimate for the score test from logistic regression with unequal group sample sizes.

| π   | μ     | α₀   | Target power(%) | Estimated size | Power (%) at N_new | Power (%) at N_sw | Power (%) at N_sm |
|-----|-------|------|-----------------|----------------|--------------------|-------------------|-------------------|
|     |       |      | N_new | N_sw | N_sm | P_new | P_sm | P_S | P_o |
| 0.05| 0.02  | --3.9398| 80   | 11,661 | 17,232 | 9601 | 91.07 | 90.97 | 90.00 | 96.35 | 80.08 | 80.00 | 63.48 | 87.03 |
| 0.15| --1.7776 | 2035 | 2626 | 1805 | 88.37 | 88.04 | 88.00 | 92.22 | 88.02 | 88.00 | 69.37 | 84.51 |
| 0.50| 0.02  | --4.2951 | 80   | 3587 | 3581 | 3589 | 90.63 | 79.95 | 80.01 | 79.92 | 80.69 | 80.01 | 80.07 | 79.99 |
| 0.75| 0.02  | --4.4502 | 80   | 5879 | 4651 | 6451 | 90.55 | 89.97 | 80.00 | 89.94 | 90.39 | 90.00 | 90.04 | 89.98 |
| 0.15| --2.1230 | 536  | 531  | 524  | 80.48 | 79.64 | 80.03 | 89.62 | 80.42 | 80.03 | 80.28 | 89.86 |
| 0.50| 0.02  | --2.2845 | 80   | 840  | 697  | 906  | 72.33 | 71.50 | 80.01 | 69.08 | 80.52 | 80.02 | 86.81 | 76.99 |
| 0.75| 0.02  | --2.8345 | 80   | 1097 | 933  | 1213 | 84.85 | 84.33 | 90.03 | 81.17 | 90.24 | 90.00 | 94.02 | 86.95 |

Confounding: p(y = 1|z, g) = 1 / [1 + exp(−zα₀ − g log(z))] + 1.

The simulation for unstratified score tests without adjustment for the stratum effect when there is no confounding effect (α₀ = 0), where α₀ is the solution to

μ = π / [1 + exp(−α₀ − β)] + (1 − π) / [1 + exp(−α₀)].

Table 6 displays the power and sample size results. The nominal power by formula (8) is very close to the simulated power, but formulae (10) and (11) may produce very poor power estimates which can deviate from the simulated power by 16% in some cases.

5. Discussion

We propose a modification of the Self and Mauritsen (1988) method for sample size calculation for score tests from GLMs, and extend it to the N1 trials. The modification takes into account the fact that the variance of the score statistic differs under H₀ and H₁. The proposed method is also suitable for other regression models. For example, the binary outcome is often analyzed by the logistic regression on basis of the odds ratio between two groups. Now suppose the parameter of interest is the relative risk instead of the odds ratio. The method is still suitable if the model is reparameterized in terms of the response rate in the control group and the relative risk parameter (Tang 2019).

The proposed method shows a marked improvement over the Self and Mauritsen (1988) formula in logistic and NB regressions when either the treatment effect is large or sample sizes are unbalanced in the two groups. In these situations, the Self and Mauritsen (1988) method degrades because the variance of the score statistic can be quite different under H₀ and H₁. As illustrated in Section 3, the approaches of Zhu and Lakiss (2014) and Zhu (2017) for NB regression tend to underestimate the size (1) when there is a large variation in the patients' follow-up time, and/or (2) when there is a large treatment effect.

The sample size calculation for the score test requires the construction and analysis of an exemplary dataset if the model is complex, and there is no analytic solution for the restricted MLE under the null hypothesis. The main computation time lies in the analysis of the exemplary data, and this can usually be done within few minutes. It is much quicker than the simulation method, which requires the generation and analysis of at least thousands of datasets to get a quite precise power estimate at a given sample size. As evidenced by the results in Tables 1 and 2, the sample size procedure shall be consistent with the test used for the analysis (Zhu and Lakiss 2014). Otherwise, the study may be either underpowered or overpowered. We recommend using the proposed procedure (or a simplified version if it exists) to determine the sample size if one plans to analyze the trial using the score test.

A future research direction is to extend the exemplary dataset approach to the Wald test and score test (Liu and Liang 1997) from the generalized estimating equations (GEE) in the analysis of repeated measurements. The sample size calculation is complicated even for the Wald test in GEE because there are missing outcomes, and the working correlation structure may be different from the true correlation structure. Even if the analytic formula exists by using the independent or true correlation structure, the calculation can still be complex to account for missing data. In the exemplary dataset approach, one can get the noncentrality parameter for the Wald test directly through the

NOTE: Simulated power (SIM) are evaluated at N_new based on 1,000,000 simulated datasets.
analysis of the exemplary data. The generation of the exemplary data also provides an opportunity to verify whether the sample size assumption is correct for correlated outcomes.

References

Aban, I. B., Cutter, G. R., and Mavinga, N. (2009), “Inferences and Power Analysis Concerning Two Negative Binomial Distributions With an Application to MRI Lesion Counts Data,” *Computational Statistics & Data Analysis*, 53, 820–833. [415]

Cochran, W. G. (1954), “Some Methods for Strengthening the Common Chi-Square Tests,” *Biometrics*, 10, 417–451. [421]

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall. [415,416]

Farrington, C. P., and Manning, G. (1990), “Test Statistics and Sample Size Formulae for Comparative Binomial Trials With Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk,” *Statistics in Medicine*, 9, 1447–1454. [415,416,419]

Laud, P. J., and Dane, A. (2014), “Confidence Intervals for the Difference Between Independent Binomial Proportions: Comparison Using a Graphical Approach and Moving Averages,” *Pharmaceutical Statistics*, 13, 294–308. [415]

Lawless, J. F. (1987), “Negative Binomial and Mixed Poisson Regression,” *The Canadian Journal of Statistics*, 15, 209–225. [418]

Liu, G., and Liang, K. Y. (1997), “Sample Size Calculations for Studies With Correlated Observations,” *Biometrics*, 53, 937–947. [423]

Matsui, S. (2005), “Sample Size Calculations for Comparative Clinical Trials With Over-Dispersed Poisson Process Data,” *Statistics in Medicine*, 24, 1339–1356. [419]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall. [415,416]

Nam, J. (1992), “Sample Size Determination for Case-Control Studies and the Comparison of Stratified and Unstratified Analyses,” *Biometrics*, 48, 389–395. [421]

Nelder, J. A., and Wedderburn, R. W. M. (1972), “Generalized Linear Models,” *Journal of the Royal Statistical Society*, Series A, 135, 370–384. [415,416]