Research Article

Statistics in Medicine

Received 7 April 2010, Accepted 8 September 2010 Published online 5 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4101

Improving small-sample inference in group randomized trials with binary outcomes

Philip M. Westgate*† and Thomas M. Braun

Group Randomized Trials (GRTs) randomize groups of people to treatment or control arms instead of individually randomizing subjects. When each subject has a binary outcome, over-dispersed binomial data may result, quantified as an intra-cluster correlation (ICC). Typically, GRTs have a small number, *n*, of independent clusters, each of which can be quite large. Treating the ICC as a nuisance parameter, inference for a treatment effect can be done using quasi-likelihood with a logistic link. A Wald statistic, which, under standard regularity conditions, has an asymptotic standard normal distribution, can be used to test for a marginal treatment effect. However, we have found in our setting that the Wald statistic may have a variance less than 1, resulting in a test size smaller than its nominal value. This problem is most apparent when marginal probabilities are close to 0 or 1, particularly when *n* is small and the ICC is not negligible. When the ICC is known, we develop a method for adjusting the estimated standard error appropriately such that the Wald statistic will approximately have a standard normal distribution. We also propose ways to handle non-nominal test sizes when the ICC is estimated. We demonstrate the utility of our methods through simulation results covering a variety of realistic settings for GRTs. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: bias correction; correlated data; overdispersion; pseudo-Wald tests; quasi-likelihood

1. Introduction

Many clinical trials involve testing a new treatment or intervention versus a control, with each study participant randomly assigned to one of these study arms. However, group randomized trials (GRTs) are unique in that groups, or clusters, of people are randomized instead of each person individually, but the outcome of interest is still obtained from each subject. Owing to feasibility issues, such as high costs, most likely only a relatively small number of clusters will be involved in a GRT. Additionally, cluster sizes can typically be quite large. To demonstrate, some common groups of randomization are patients with the same healthcare provider, communities, and schools.

An example of a GRT would be the study reported by Atri *et al.* [1]. This study aimed to discover if a 2-h training session for receptionists, who were supposed to later attempt to contact patients, would increase breast screening rates in women who failed to attend an appointment by a certain time point. Twelve practices, and inherently the women their receptionists were to contact, were randomized to this receptionist intervention, while 14 were randomized to be controls. The subject-level outcome of interest was an indicator of whether a given woman had received screening after failing to attend her initial appointment.

It is the purpose of this paper to demonstrate methods for testing for a marginal treatment effect with a Wald statistic that maintains a nominal test size when the outcomes of interest from a GRT are subject-level binary indicators of a desired outcome ('success'). These types of data often lead to overdispersion because of unmeasured group effects that make the probability of success vary between clusters. For example, in the Atri *et al.* [1] study, the receptionists' natural ability to get women to come in for breast screening varies, thus resulting in different success rates for each practice. Statistical methods need to take this overdispersion into account to obtain reliable inference.

In Section 2 we introduce statistical notation and existing methodology. In Section 3, we present drawbacks with a traditional Wald statistic when using a model-based standard error (SE), which motivates the derivation of our psuedo-Wald tests. In Section 4, we examine the performance of our pseudo-Wald tests via simulation as well as in application to an actual GRT. Section 5 contains a discussion and concluding remarks.

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.
*Correspondence to: Philip M. Westgate, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.
†E-mail: pwestgat@umich.edu

2. Notation and existing methodology

2.1. Notation

Throughout this manuscript, we adopt the following notation. Let X_{ij} represent the outcome for subject j in cluster i, $j=1,2,\ldots,n_i$; $i=1,2,\ldots,n$. $X_{ij}=1$ denotes success and $X_{ij}=0$ denotes the absence of the desired outcome ('failure'). We also let ρ denote the intra-cluster correlation (ICC) for any pair of outcomes from individuals within the same cluster. We assume the ICC is constant across clusters.

We let Z_i represent the indicator of treatment assignment for all individuals in cluster i, with $Z_i=1$ indicating new treatment or intervention and $Z_i=0$ indicating control. We let the first u clusters represent the controls, and the last $n-u$ represent groups receiving the new treatment or intervention.

We assume $\pi_i=\text{Prob}(X_{ij}=1)=E[\text{Prob}(X_{ij}=1|p_i)]$ is constant for all clusters in the same treatment arm, where p_i is the unobserved true probability of success for any given subject in the ith cluster. We assume a simple logistic regression model, $\text{logit}(\pi_i)=\beta_0+\beta_1 Z_i$, where $\text{logit}(\pi_i)=\log(\pi_i)-\log(1-\pi_i)$. If the ith cluster is randomized to control, then $\pi_i=\pi_C$; if randomized to the new treatment or intervention, then $\pi_i=\pi_T$. Although this model can be generalized to include cluster-level or individual-level predictors, in the current presentation we assume that such adjustments are unnecessary.

Let $Y_i=\sum_{j=1}^{n_i} X_{ij}$ represent the number of successes in cluster i, which has mean $E(Y_i)=n_i\pi_i$ and variance $\text{Var}(Y_i)=n_i\pi_i(1-\pi_i)[1+(n_i-1)\rho]$. If $\rho=0$, then Y_i has a binomial distribution. In most settings, ρ is assumed to be positive, although negative values are possible (see for example [2]). This positive correlation results in Y_i being overdispersed, i.e. having larger variance than what is predicted by the binomial distribution. The factor $[1+(n_i-1)\rho]$ is known as the variance inflation factor (VIF) for the ith cluster.

The ICC can be viewed as measuring the degree to which responses from subjects within the same cluster tend to respond 'more alike' as compared with subjects from different clusters in the same arm; this effect is due to clusters potentially having different success rates, or $\text{Var}(p_i)>0$ for $i=1,2,\ldots,n$. Mathematically, for $1\leqslant j\neq l\leqslant n_i$,

$$\begin{aligned}\rho &= \frac{\text{Cov}[E(X_{ij}|p_i),E(X_{il}|p_i)]+E[\text{Cov}(X_{ij},X_{il}|p_i)]}{E[\text{Var}(X_{ij}|p_i)]+\text{Var}[E(X_{ij}|p_i)]}\\ &= \frac{\text{Cov}[p_i,p_i]+E(0)}{E(p_i[1-p_i])+\text{Var}(p_i)}\\ &= \frac{\text{Var}(p_i)}{\pi_i(1-\pi_i)}\end{aligned}$$

Owing to the constraint ICC $\leqslant 1$, we have $\text{Var}(p_i)\leqslant\pi_i(1-\pi_i)$.

Although we only work with the ICC in this paper, it is important to note that another popular method is using pairwise odds ratios for modeling the association among subject-level outcomes within the same cluster. See Carey *et al.* [3] for more detail.

2.2. Quasi-likelihood

Wedderburn [4] developed the theory of quasi-likelihood (QL), which is used when a generalized linear model (GLM) [5, 6] is desired but the true distribution for the observed data is unknown. This is important since the true distribution for over-dispersed binomial data will be unknown in practice. Using QL, only the mean and variance structures for the proportion of successes in each cluster need correct specification. More specifically, the link function with its linear predictor, $g(\mu_i)=\eta_i=\mathbf{z}_i'\boldsymbol{\beta}$, and $\text{Var}(Y_i)$ need correct specification. Here, $\mathbf{z}_i'=[1,z_{i,1},\ldots,z_{i,p-1}]$ is a $p\times 1$ vector of covariate values for the ith independent observation and $\boldsymbol{\beta}=[\beta_0,\ldots,\beta_{p-1}]'$ is a $p\times 1$ vector of corresponding regression parameters. Maximum quasi-likelihood (MQL) involves setting the following quasi-score equations equal to zero and solving for $\boldsymbol{\beta}$, where $\mathbf{D}_i=\partial\mu_i/\partial\boldsymbol{\beta}=[\partial\mu_i/\partial\beta_0,\ldots,\partial\mu_i/\partial\beta_{p-1}]'$ and $V_i=\text{Var}(Y_i)$:

$$U(\boldsymbol{\beta})=\sum_{i=1}^{n}\mathbf{D}_i V_i^{-1}(Y_i-\mu_i).$$

In our setting, the quasi-score equations simplify to

$$\sum_{i=1}^{n}[1,Z_i]'\frac{Y_i-n_i\pi_i}{1+(n_i-1)\rho}.$$

As long as a consistent estimate for the true ICC is used, the MQL estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{MQL}}=[\hat{\beta}_{0\mathrm{MQL}},\ldots,\hat{\beta}_{(p-1)\mathrm{MQL}}]'$, converges in probability to $\boldsymbol{\beta}$ and has an asymptotic normal distribution with covariance matrix $(\sum_{i=1}^{n}\mathbf{D}_i V_i^{-1}\mathbf{D}_i')^{-1}$, simplifying to

$$\left(\sum_{i=1}^{n}[1,Z_i]'[1,Z_i]\frac{n_i\pi_i(1-\pi_i)}{1+(n_i-1)\rho}\right)^{-1} \tag{1}$$

in our scenario.

Our main concern was that MQL, which is popular for analyzing over-dispersed binomial data, gives statistical results relying upon asymptotic theory. Unfortunately, since GRTs generally do not have a large number of independent clusters, asymptotic theory may not hold. We discuss three potential problems resulting from this, all of which may cause a test size to not be at its nominal value when using a Wald statistic. First, when the data analyst is unsure of the correct variance or ICC structure, the sandwich, or empirical, covariance matrix estimator can be used and is given by

$$\left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i\hat{V}_i^{-1}\hat{\boldsymbol{D}}_i'\right)^{-1}\left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i\hat{V}_i^{-1}(Y_i-n_i\hat{\pi}_i)^2\hat{V}_i^{-1}\hat{\boldsymbol{D}}_i'\right)\left(\sum_{i=1}^{n}\hat{\boldsymbol{D}}_i\hat{V}_i^{-1}\hat{\boldsymbol{D}}_i'\right)^{-1},$$

with consistent estimates used in place of unknown parameters [7]. This gives a consistent estimate for the covariance of the MQL parameter estimates. Unfortunately, the variance of this covariance estimate, along with its tendency to underestimate the true SEs for small n, can cause test size to be too large (see [8, 9]). Although there is no formal definition for what small n is, Mancl and DeRouen [9] and Murray *et al.* [10] suggest $n<50$ and $n<40$, respectively. Bootstrap and jackknife methods are alternatives to the sandwich estimator, but can also be problematic, particularly when the number of successes in each cluster are zero or small [9]. Second, MQL estimates of the regression parameters tend to be biased away from zero when using the logistic link and n is small, with bias increasing as the true parameter values move further from zero (presented later). Third, since normality of the Wald statistic is an asymptotic result, combined with the need to estimate the ICC, the correct distribution from which to obtain critical values is unknown.

Methods to help reduce test size toward the nominal value when using the sandwich SE estimator have been introduced by Kauermann and Carroll [8], Mancl and DeRouen [9], Pan [11], Fay and Graubard [12], Pan and Wall [13], Morel *et al.* [14], and McCaffrey and Bell [15]. Drum and McCullagh [16] argued that using the model-based SE instead of the sandwich estimate is best when n is small and there is no reason to believe the assumed variance structure is 'substantially incorrect'. Liang and Hanfelt [17] also expected the model-based SE to be more stable.

2.3. *Bias correction for the MQL parameter estimates*

Similar to the formulas for the bias in maximum likelihood estimates from GLMs given by Cordeiro and McCullagh [18], Cordeiro and Demetrio [19] gave formulas for the bias to order n^{-1} for the MQL estimates. These formulas are given by

$$\mathbf{Bias}(\hat{\boldsymbol{\beta}}_{\mathrm{MQL}})=-0.5(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{M}_d\boldsymbol{F}\mathbf{1}.$$

Here, $\mathbf{W}=\mathrm{diag}(n_i\pi_i(1-\pi_i)/[1+(n_i-1)\rho])$, $\mathbf{Z}'=[\mathbf{z}_1,\ldots,\mathbf{z}_n]$, $\mathbf{M}=\mathbf{Z}(\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'$, $\mathbf{M}_d=\mathrm{diag}[M_{ii}]$, where M_{ii} is the ith diagonal element of $\mathbf{M}$,

$$\mathbf{F}=\mathrm{diag}\left[\frac{n_i\pi_i(1-\pi_i)[2(1-\pi_i)-1]}{1+(n_i-1)\rho}\right],$$

and $\mathbf{1}$ is an $n\times 1$ vector of ones. We now write π_C and π_T as $\pi_C(\boldsymbol{\beta})$ and $\pi_T(\boldsymbol{\beta})$, respectively, since marginal probabilities are functions of $\boldsymbol{\beta}$. In our settings,

$$\mathrm{Bias}(\hat{\beta}_{0\mathrm{MQL}})=\frac{2\pi_C(\boldsymbol{\beta})-1}{2\pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})]\sum_{i=1}^{u}q_i} \tag{2}$$

and

$$\mathrm{Bias}(\hat{\beta}_{1\mathrm{MQL}})=\frac{2\pi_T(\boldsymbol{\beta})-1}{2\pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})]\sum_{i=u+1}^{n}q_i}-\mathrm{Bias}(\hat{\beta}_{0\mathrm{MQL}}), \tag{3}$$

where $q_i=n_i/[1+(n_i-1)\rho]$. Bias increases with decreases in n and cluster sizes and increases in ρ, $|\pi_C-0.5|$, and $|\pi_T-0.5|$.

The bias-corrected estimates (BCEs) are given by $\hat{\boldsymbol{\beta}}_{\mathrm{BC}}=[\hat{\beta}_{0\mathrm{MQL}}-\widehat{\mathrm{Bias}}(\hat{\beta}_{0\mathrm{MQL}}),\hat{\beta}_{1\mathrm{MQL}}-\widehat{\mathrm{Bias}}(\hat{\beta}_{1\mathrm{MQL}})]'=[\hat{\beta}_{0\mathrm{BC}},\hat{\beta}_{1\mathrm{BC}}]'$, where $\pi_C(\hat{\boldsymbol{\beta}}_{\mathrm{MQL}})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{\mathrm{MQL}})$ are used in place of $\pi_C(\boldsymbol{\beta})$ and $\pi_T(\boldsymbol{\beta})$, respectively, in equations (2) and (3). Owing

to using estimated marginal probabilities, biases can be slightly overestimated as found in Bull and Greenwood [20] and our simulations (not shown), especially when marginal probabilities are near the boundary of the parameter space and the number of clusters is small. The following iterative procedure was therefore used, which produces better bias approximations:

1. Estimate bias as just mentioned to obtain $\widehat{\text{Bias}}^{(1)}(\hat{\beta}_{0\text{MQL}})$ and $\widehat{\text{Bias}}^{(1)}(\hat{\beta}_{1\text{MQL}})$. From these, denote the current BCEs as $\hat{\boldsymbol{\beta}}^{(1)}_{\text{BC}}$.
2. Next, use $\pi_C(\hat{\boldsymbol{\beta}}^{(1)}_{\text{BC}})$ and $\pi_T(\hat{\boldsymbol{\beta}}^{(1)}_{\text{BC}})$ in equations (2) and (3) to obtain updated bias estimates, $\widehat{\text{Bias}}^{(2)}(\hat{\beta}_{0\text{MQL}})$ and $\widehat{\text{Bias}}^{(2)}(\hat{\beta}_{1\text{MQL}})$. Use these to update the BCEs: $\hat{\boldsymbol{\beta}}^{(2)}_{\text{BC}}=[\hat{\beta}_{0\text{MQL}}-\widehat{\text{Bias}}^{(2)}(\hat{\beta}_{0\text{MQL}}),\hat{\beta}_{1\text{MQL}}-\widehat{\text{Bias}}^{(2)}(\hat{\beta}_{1\text{MQL}})]'$.
3. Keep repeating until $|\hat{\beta}^{(s)}_{0\text{BC}}-\hat{\beta}^{(s-1)}_{0\text{BC}}|+|\hat{\beta}^{(s)}_{1\text{BC}}-\hat{\beta}^{(s-1)}_{1\text{BC}}|<\varepsilon$, for some ε close to 0 and $s\geqslant 1$. We used $\varepsilon=10^{-7}$.

Equation (2) shows $0\leqslant|\beta_0|\leqslant|E(\hat{\beta}_{0\text{MQL}})|$, with β_0 and $E(\hat{\beta}_{0\text{MQL}})$ having the same sign, so $\hat{\beta}_{0\text{MQL}}$ is positively biased when $\beta_0>0$ and negatively biased when $\beta_0<0$. This implies that the BCE for β_0 will take on a value between 0 and $\hat{\beta}_{0\text{MQL}}$, giving $0\leqslant\hat{\beta}_{0\text{BC}}/\hat{\beta}_{0\text{MQL}}\leqslant 1$. Equation (3) shows that the same relationship typically occurs for $\hat{\beta}_{1\text{MQL}}$ and $\hat{\beta}_{1\text{BC}}$. When π_C and π_T are close in value, though, there is a chance that this will not occur if $|\sum_{i=1}^{u}q_i-\sum_{i=u+1}^{n}q_i|$ is not 'small'. Although it makes little difference, we chose to set $\hat{\beta}_{1\text{BC}}=\hat{\beta}_{1\text{MQL}}$ if $\hat{\beta}_{1\text{BC}}/\hat{\beta}_{1\text{MQL}}$ was originally greater than 1.

3. Developing a pseudo-Wald test with nominal size

3.1. Quantifying the impact of model-based SEs on traditional Wald test size

From Equation (1), the model-based SE for $\hat{\beta}_{1\text{MQL}}$ is $\text{SE}(\hat{\beta}_{1\text{MQL}})=\text{SE}_{\hat{\beta}_{1\text{MQL}}}[\pi_C(\boldsymbol{\beta}),\pi_T(\boldsymbol{\beta})]=$

$$\sqrt{\left[\sum_{i=1}^{u}\frac{n_i\pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})]}{1+(n_i-1)\rho}\right]^{-1}+\left[\sum_{i=u+1}^{n}\frac{n_i\pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})]}{1+(n_i-1)\rho}\right]^{-1}} \tag{4}$$

This is estimated by $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})=\text{SE}_{\hat{\beta}_{1\text{MQL}}}[\pi_C(\hat{\boldsymbol{\beta}}_{\text{MQL}}),\pi_T(\hat{\boldsymbol{\beta}}_{\text{MQL}})]$. Simulations (not shown) indicated that when using an unbiased estimate, $\hat{\rho}$, for the ICC, $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ is also fairly unbiased. For now, assume ρ is known.

The Wald statistic $W_{\text{Reg}}=\hat{\beta}_{1\text{MQL}}/\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ is regularly used in practice to test for a marginal treatment effect. Using $\pi_C=\pi_T=\rho=0.05$, Figure 1(a) shows the empirical distribution for W_{Reg} from 100 000 simulations, along with the density of the $N(0,1)$ distribution. Cluster sizes varied uniformly from 25 to 150 subjects, 10 clusters were randomized to each treatment arm, and outcomes were generated by the beta-binomial distribution. Figure 1(a) reveals that the $N(0,1)$ distribution has heavier tails since the variance of W_{Reg}, $\text{Var}(W_{\text{Reg}})$, is less than 1 (0.905 in these simulations). This example implies that using W_{Reg} in conjunction with critical values from the $N(0,1)$ distribution can result in a test size that is smaller than desired, as will using any heavy-tailed distribution, such as a t-distribution.

Equation (1) shows that the variance of $\hat{\boldsymbol{\beta}}_{\text{MQL}}$ increases with decreases in n and cluster sizes and increases in ρ, $|\pi_C-0.5|$, and $|\pi_T-0.5|$. In practice, since $\hat{\boldsymbol{\beta}}_{\text{MQL}}$ is actually used to estimate its own variability, $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ is not a fixed quantity. Its variance, $\text{Var}[\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})]$, will increase as $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{MQL}})$ increases, and therefore is also a function of n, cluster sizes, ρ, π_C, and π_T.

Owing to the variability in $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$, $\text{Var}(W_{\text{Reg}})$ will depend on the variances and covariance of $\hat{\beta}_{1\text{MQL}}$ and $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$. As $\text{Var}(\hat{\beta}_{1\text{MQL}})$ increases, there are more extreme values for $\hat{\beta}_{1\text{MQL}}$, and these large values are associated with values for $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ that are larger than the true SE. Hauck and Donner [21] demonstrated this tendency in logistic regression. This relationship can cause W_{Reg} to be smaller than desired, therefore reducing $\text{Var}(W_{\text{Reg}})$ and making the tails in the distribution of W_{Reg} to become lighter, thus diminishing test size. Test size decreases as $\text{Var}[\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})]$ increases, implying test size decreases away from its nominal level with decreases in n and cluster sizes and increases in ρ, $|\pi_C-0.5|$, and $|\pi_T-0.5|$.

This phenomenon of test size being smaller than its nominal level is rather minor when marginal probabilities are not near the edge of the parameter space. Owing to the curvature of equation (4) with respect to the marginal probabilities, the impact from variation in the estimate of $\boldsymbol{\beta}$ used in this formula leads to increasingly larger variations in the estimated SE as $|\pi_C-0.5|$ and $|\pi_T-0.5|$ approach 0.5. Empirical evidence (not shown) indicates that having a large number of clusters, say 30 or more per treatment arm, will typically combat this problem quite well. For a small to moderate

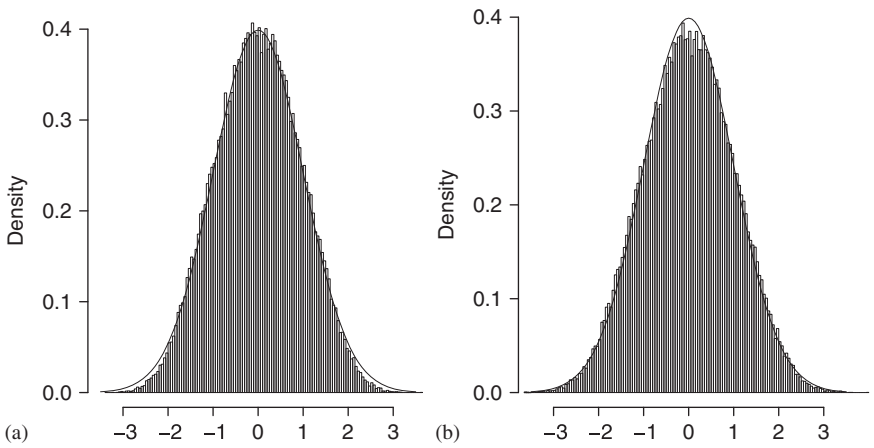

Figure 1. $N(0,1)$ density and empirical distributions for W_{Reg} (a) and $\widetilde{W}_{1.5}$ (b), where 10 clusters were randomized to each treatment arm, and marginal probabilities and the ICC were 0.05.

number of clusters, the decrease in test size may become important as $|\pi_C - 0.5|$ and $|\pi_T - 0.5|$ rise to 0.3 or higher, especially when ρ is almost as large, if not larger, than π_C and π_T. Additionally, some GRTs may not have large cluster sizes. An example of this would be a study where each cluster is actually an individual subject contributing a small number, or group, of binary outcomes. Having smaller sizes for a fixed number of clusters will increase the variation in the estimate of $\boldsymbol{\beta}$, causing a greater impact on test size. This impact may be negligible, however, unless the differences in cluster sizes are large. For instance, a study in which only a small number of observations on each subject are observed can have larger test size problems than a GRT where there are a large number of outcomes in each cluster.

3.2. Deriving a pseudo-Wald statistic with known ICC

In practice, the ICC will need to be estimated, but the goal of this section is to show how the size of the Wald test can be adjusted closer to a nominal level, α, when ρ is known. As mentioned previously, $\text{Var}(W_{\text{Reg}})$ can be less than 1, resulting in test sizes smaller than α when using $N(0,1)$ critical values. Two possible ways of fixing this would be to find critical values that will consistently produce a test size equal to α, or modify $\widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ such that the resulting SE estimate is smaller by an amount depending on n, cluster sizes, ρ, π_C, and π_T, with the goal of producing a Wald statistic with a variance of 1. We take the latter approach, but utilize the idea of changing the critical values when we later incorporate estimation of ρ.

Test size, $\text{Bias}(\hat{\beta}_{0\text{MQL}})$, and $\text{Bias}(\hat{\beta}_{1\text{MQL}})$ are functions of n, cluster sizes, ρ, π_C, and π_T. Incorporating these relationships, we define $\tilde{\beta}_k^N = (\hat{\beta}_{k\text{BC}}/\hat{\beta}_{k\text{MQL}})^N \hat{\beta}_{k\text{MQL}} = (\hat{\beta}_{k\text{BC}})^N/(\hat{\beta}_{k\text{MQL}})^{N-1}$, $k=0,1$, for any non-negative real number N. Our proposed pseudo-SE estimate is $\widetilde{\text{SE}}_N(\hat{\beta}_{1\text{MQL}}) = \text{SE}_{\hat{\beta}_{1\text{MQL}}}[\pi_C(\tilde{\boldsymbol{\beta}}^N), \pi_T(\tilde{\boldsymbol{\beta}}^N)]$, in which $\tilde{\boldsymbol{\beta}}^N = [\tilde{\beta}_0^N, \tilde{\beta}_1^N]'$. Since N is non-negative, $(\hat{\beta}_{k\text{BC}}/\hat{\beta}_{k\text{MQL}}) \in [0,1]$ implies $|\tilde{\beta}_k^N| \leqslant |\hat{\beta}_{k\text{MQL}}|$, which in practice will give $\widetilde{\text{SE}}_N(\hat{\beta}_{1\text{MQL}}) \leqslant \widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ since $\pi_C(\tilde{\boldsymbol{\beta}}^N)$ and $\pi_T(\tilde{\boldsymbol{\beta}}^N)$ will be no further in value from 0.5 than $\pi_C(\hat{\boldsymbol{\beta}}_{\text{MQL}})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{\text{MQL}})$, respectively, in realistic settings. Typically, $\widetilde{\text{SE}}_N(\hat{\beta}_{1\text{MQL}}) < \widehat{\text{SE}}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$. The pseudo-Wald statistic $\widetilde{W}_N = \hat{\beta}_{1\text{MQL}}/\widetilde{\text{SE}}_N(\hat{\beta}_{1\text{MQL}})$ will then be larger in absolute value than W_{Reg}. In addition, decreases in n and cluster sizes and increases in ρ, $|\pi_C - 0.5|$, and $|\pi_T - 0.5|$ correspond to increases in test size when using $\widetilde{W}_N$ as the test statistic as compared to using W_{Reg}.

Using $\widetilde{W}_N$ as our test statistic, we needed to find a value for N such that the variance of $\widetilde{W}_N$ is always approximately 1, with a resulting test size equal to α when using $N(0,1)$ critical values. Increasing N causes $\widetilde{\text{SE}}_N(\hat{\beta}_{1\text{MQL}})$ to decrease, and therefore test size will increase. Simulations were conducted to find an appropriate solution. Each setting was examined in 10 000 simulations, and cluster sizes varied uniformly from 25 to 150 subjects. Empirical test size was compared for W_{Reg} and $\widetilde{W}_N$ with $N \in \{1, 1.25, 1.5, 1.75, 2\}$. The results from using a 5 per cent significance level can be seen in Table I and show that $N = 1.5$ performed best. However, allowing N to be any value from 1.25 to 2 would be adequate. Figure 1(b) shows the empirical distribution for $\widetilde{W}_{1.5}$ from the same set of simulations used to produce Figure 1(a), along with the density of the $N(0,1)$ distribution. The tails from this empirical distribution match the tails from the $N(0,1)$ density, indicating that test size is at its nominal level.

Although we are dealing with scenarios involving small n, it is important to show how $\widetilde{W}_N$ performs asymptotically. Given $N < \infty$ and $(n-u)/u$ is a constant (typically around one), both $\tilde{\boldsymbol{\beta}}^N \xrightarrow{p} \boldsymbol{\beta}$ and $\widetilde{W}_N \xrightarrow{d} N(0,1)$ as $n \to \infty$. Therefore,

Table I. Empirical test sizes using the given Wald statistic and $N(0,1)$ critical values. $n/2$ clusters were randomized to each treatment arm, and had a marginal success rate π. The ICC is *known*. Bold values have corresponding 95 per cent confidence intervals covering 0.05, the nominal level. Outcomes from the first 10 sets of simulations were generated using the beta-binomial distribution, whereas outcomes from the last 10 sets of simulations were generated using a log-gamma mixture distribution, such that $-\ln(p_i)\sim\text{Gamma}(\theta_i,\phi_i)$.

$n/2$	π	ICC	W_{Reg}	$\widetilde{W}_1$	$\widetilde{W}_{1.25}$	$\widetilde{W}_{1.5}$	$\widetilde{W}_{1.75}$	$\widetilde{W}_2$
10	0.05	0.05	0.0360	**0.0459**	**0.0482**	**0.0509**	**0.0540**	0.0557
20	0.05	0.05	0.0407	0.0440	0.0453	**0.0464**	**0.0472**	**0.0485**
10	0.10	0.05	0.0436	**0.0461**	**0.0471**	**0.0478**	**0.0482**	**0.0491**
20	0.10	0.05	**0.0470**	**0.0490**	**0.0494**	**0.0499**	**0.0505**	**0.0509**
10	0.10	0.10	0.0390	**0.0462**	**0.0483**	**0.0502**	**0.0521**	**0.0534**
20	0.10	0.10	0.0429	**0.0458**	**0.0466**	**0.0471**	**0.0487**	**0.0490**
10	0.20	0.05	**0.0481**	**0.0492**	**0.0495**	**0.0495**	**0.0496**	**0.0500**
20	0.20	0.10	**0.0496**	**0.0502**	**0.0502**	**0.0503**	**0.0504**	**0.0505**
10	0.30	0.05	**0.0518**	**0.0523**	**0.0524**	**0.0525**	**0.0527**	**0.0528**
20	0.30	0.10	**0.0514**	**0.0518**	**0.0520**	**0.0520**	**0.0520**	**0.0520**
10	0.05	0.05	0.0350	**0.0464**	**0.0491**	**0.0517**	0.0543	0.0565
20	0.05	0.05	0.0422	**0.0464**	**0.0471**	**0.0479**	**0.0492**	**0.0502**
10	0.10	0.05	0.0429	0.0452	0.0456	**0.0470**	**0.0483**	**0.0492**
20	0.10	0.05	**0.0477**	**0.0493**	**0.0496**	**0.0502**	**0.0507**	**0.0513**
10	0.10	0.10	0.0375	0.0450	**0.0476**	**0.0496**	**0.0514**	**0.0536**
20	0.10	0.10	0.0440	**0.0460**	**0.0467**	**0.0472**	**0.0483**	**0.0492**
10	0.20	0.05	**0.0479**	**0.0485**	**0.0488**	**0.0492**	**0.0496**	**0.0500**
20	0.20	0.05	**0.0483**	**0.0488**	**0.0488**	**0.0488**	**0.0488**	**0.0488**
10	0.20	0.10	**0.0459**	**0.0474**	**0.0480**	**0.0483**	**0.0490**	**0.0497**
20	0.20	0.10	**0.0492**	**0.0502**	**0.0505**	**0.0507**	**0.0511**	**0.0512**

$\widetilde{W}_N$ and W_{Reg} should give very similar results for large n, and both will approximately behave as standard normal random variables.

3.3. *Incorporating estimation of ICC*

The following two issues still need to be handled in practice: (i) finding a consistent estimate, $\hat{\rho}$, for the ICC and (ii) finding appropriate critical values and/or adjust equation (4) to deal with the effect estimating ρ has on the distribution of the Wald statistic that is utilized. With regard to the first issue, there are many papers that have dealt with the topic of estimating the ICC, with Ridout *et al.* [22] presenting an overview with simulations comparing many different estimation procedures. Some of the more well-known ways to estimate the ICC are Williams' method [23], Pseudolikelihood (PL) [24, 25], Extended Quasi-Likelihood (EQL) [26], and ANOVA [27–29]. These are useful in that they can handle regression models with multiple covariates, both categorical and continuous, at the cluster level. Preliminary simulations (not shown) were done to find the most appropriate method in our settings in terms of mean squared error (MSE). Of the previously mentioned methods, ANOVA performed best. If the assumption of a common ICC is correct, a consistent estimate is given by

$$\hat{\rho}_{\text{ANOVA}}=\frac{\text{MSB}-\text{MSE}}{\text{MSB}+(K-1)\text{MSE}}$$

where

$$\text{MSB}=\frac{1}{n-2}\sum_{i=1}^{n}n_i\left(\frac{Y_i}{n_i}-\hat{\pi}_i\right)^2,\quad \text{MSE}=\frac{1}{\sum_{i=1}^{n}n_i-n}\sum_{i=1}^{n}Y_i\left(1-\frac{Y_i}{n_i}\right),$$

$$K=\frac{1}{n-2}\left[\sum_{i=1}^{n}n_i-\sum_{i=1}^{u}\frac{n_i^2}{m_C}-\sum_{i=u+1}^{n}\frac{n_i^2}{m_T}\right],\quad m_C=\sum_{i=1}^{u}n_i,\quad m_T=\sum_{i=u+1}^{n}n_i.$$

With regard to the second issue, estimating the ICC causes $\text{Var}(W_{\text{Reg}})$ and $\text{Var}(\widetilde{W}_N)$ to increase, leading to an inflation of test size. One method to deal with increased test size would be to use critical values that are larger in absolute value; however, these values would need to depend on the bias and variance of $\hat{\rho}$, and therefore n, since increasing n will decrease these quantities, causing test size to reduce back toward α. One possible solution would be to obtain critical

values from a t-distribution with $f(n)$ degrees of freedom (df), denoted as $t_{f(n)}$, assuming $f(\cdot)$ is a 'correctly' chosen function. We later demonstrate the utility of $f(n)=n$.

Another way to shrink test size back toward α after estimating the ICC would be to use leverage values to inflate the estimated versions of equation (4) for W_{Reg} and $\widetilde{W}_N$ while continuing to use $N(0,1)$ critical values. Leverage values, $0 \leqslant h_i \leqslant 1,\ i=1,2,\ldots,n$, are the diagonal elements of the $n \times n$ matrix

$$\boldsymbol{H}=\boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}^{1/2},$$

such that $\sum_{i=1}^{n} h_i = p$, where p is the number of regression parameters ($p=2$ in our setting). Note that leverage values are estimated using $\pi_C(\hat{\boldsymbol{\beta}}_{\text{MQL}})$ and $\pi_T(\hat{\boldsymbol{\beta}}_{\text{MQL}})$. As $n \to \infty$, we have $h_i \to 0,\ i=1,\ldots,n$. We multiply the ith term in equation (4) by $(1-h_i)$, giving

$$\sqrt{\left[\sum_{i=1}^{u}\frac{n_i\pi_C(\boldsymbol{\beta})[1-\pi_C(\boldsymbol{\beta})](1-h_i)}{1+(n_i-1)\rho}\right]^{-1}+\left[\sum_{i=u+1}^{n}\frac{n_i\pi_T(\boldsymbol{\beta})[1-\pi_T(\boldsymbol{\beta})](1-h_i)}{1+(n_i-1)\rho}\right]^{-1}} \tag{5}$$

Using this function of the leverages inside equation (4) will cause the denominator of the Wald statistic to increase as n decreases, offsetting at least part of the elevation in the variance of the Wald statistic due to the increase in variance and bias of $\hat{\rho}$.

4. Assessing the utility of the pseudo-Wald statistic

4.1. Via simulation study

We compare the test size of W_{Reg} and $\widetilde{W}_{1.5}$ to those of two other Wald statistics. The first, $W_S = \hat{\beta}_{1\text{MQL}}/\widehat{\text{SE}}_S(\hat{\beta}_{1\text{MQL}})$ is the traditional Wald statistic using the sandwich form for the SE of $\hat{\beta}_{1\text{MQL}}$. The second, $W_{\text{SBC}} = \hat{\beta}_{1\text{MQL}}/\widehat{\text{SE}}_{\text{SBC}}(\hat{\beta}_{1\text{MQL}})$ replaces the sandwich estimate for $\text{SE}(\hat{\beta}_{1\text{MQL}})$ with the bias-corrected version proposed by Mancl and DeRouen [9]. All four statistics are compared to both $N(0,1)$ and t_n critical values. We also compute versions of W_{Reg} and $\widetilde{W}_{1.5}$ implementing equation (5) which are compared to $N(0,1)$ critical values. As mentioned previously, ANOVA is used to estimate a common ICC. Therefore, our results apply to using any estimator of ρ with bias and variance similar to that of $\hat{\rho}_{\text{ANOVA}}$. Empirical test sizes under five settings using a significance level of 0.05 are displayed in Table II.

Results show that using W_{Reg} leads to inconsistent test sizes. When compared to $N(0,1)$ critical values, the test size is too large in scenarios with marginal probabilities of at least 0.20, unless we use the inflated SE. When using the inflated SE or comparing to t_n critical values, the null hypothesis is not rejected enough if marginal probabilities are 0.10 or less. W_S is even less desirable to use. It gives inconsistent test sizes, all tending to be too large. W_{SBC} fares better, but also leads to inconsistent inference. W_{SBC} is more reliable when compared to t_n critical values, but test size is too large when group probabilities are 0.05 and too small when group probabilities range from 0.2 to 0.5. Comparing $\widetilde{W}_{1.5}$ and its inflated SE version to t_n and $N(0,1)$ critical values, respectively, test size is consistently at its nominal level, and so there is no need to be concerned whether inference will be liberal or conservative. This gives a valid test, which will not reject the null hypothesis too often as the previously mentioned tests do in some settings. Additionally, these two proposed pseudo-Wald tests will produce greater power over all scenarios where other tests are conservative, such as when comparing W_{SBC} to t_n critical values and the marginal probabilities, although not necessarily equal, range from 0.20 to 0.50.

Table II. Empirical test sizes using the given Wald statistic and distribution for obtaining critical values when *estimating* a common ICC. Bold values have corresponding 95 per cent confidence intervals covering 0.05, the nominal level. Each setting used 10 000 replications, with cluster sizes generated as in previous simulations, while data came from a beta-binomial distribution with a common ICC. Equation (5) denotes the use of equation (5) and $N(0,1)$ critical values.

			W_{Reg}			$\widetilde{W}_{1.5}$			W_S		W_{SBC}	
$n/2$	π	ICC	$N(0,1)$	t_n	Equation (5)	$N(0,1)$	t_n	Equation (5)	$N(0,1)$	t_n	$N(0,1)$	t_n
10	0.05	0.05	**0.0506**	0.0376	0.0387	0.0662	**0.0478**	**0.0503**	0.1029	0.0836	0.0715	0.0548
20	0.10	0.10	**0.0509**	0.0443	0.0453	0.0557	**0.0486**	**0.0494**	0.0713	0.0635	0.0586	**0.0509**
10	0.20	0.10	0.0614	**0.0462**	**0.0484**	0.0645	**0.0491**	**0.0518**	0.0836	0.0693	0.0589	0.0452
10	0.30	0.05	0.0655	**0.0514**	**0.0537**	0.0660	**0.0520**	**0.0537**	0.0806	0.0662	0.0550	0.0418
20	0.50	0.10	0.0567	**0.0499**	**0.0511**	0.0567	**0.0499**	**0.0511**	0.0635	0.0556	**0.0501**	0.0437

Table III. Estimates, p values, and 95 per cent confidence intervals (CIs) resulting from the analysis of the breast screening data. Critical values were obtained from the $N(0,1)$ and t_{26} distributions. P values and CIs correspond to the Wald test using the SE estimate in the corresponding row. $\widetilde{\mathrm{SE}}^{*}_{1.5}(\hat{\beta}_{1\mathrm{MQL}})$ indicates the use of equation (5) with our pseudo-SE method.

		$N(0,1)$		t_{26}	
	Estimate				
$\hat{\beta}_{1\mathrm{MQL}}$	1.138	p-value	95 per cent CI	p-value	95 per cent CI
$\widehat{\mathrm{SE}}_{\mathrm{MQL}}(\hat{\beta}_{1\mathrm{MQL}})$	0.499	0.023	(0.160, 2.116)	0.031	(0.113, 2.163)
$\widetilde{\mathrm{SE}}_{1.5}(\hat{\beta}_{1\mathrm{MQL}})$	0.479	0.017	(0.200, 2.076)	0.025	(0.154, 2.122)
$\widetilde{\mathrm{SE}}^{*}_{1.5}(\hat{\beta}_{1\mathrm{MQL}})$	0.498	0.022	(0.162, 2.114)		
$\widehat{\mathrm{SE}}_{S}(\hat{\beta}_{1\mathrm{MQL}})$	0.417	0.006	(0.322, 1.955)	0.011	(0.282, 1.995)
$\widehat{\mathrm{SE}}_{\mathrm{SBC}}(\hat{\beta}_{1\mathrm{MQL}})$	0.448	0.011	(0.261, 2.015)	0.017	(0.218, 2.058)

4.2. Via application to actual GRT

We illustrate an application of our pseudo-Wald tests using the study reported by Atri *et al.* [1], the data for which are presented in Turner *et al.* [30]. The number of women failing to attend appointments in a given practice ranged from 19 to 201. The ICC was estimated to be 0.064, indicating a small variation in estimated success rates between practices in the same treatment arm. Using the MQL regression parameters, the marginal probabilities of breast screening for intervention and control practices were estimated to be 0.101 and 0.035, respectively. Parameter and SE estimates, along with p-values and 95 per cent confidence intervals, for the various methods presented in this paper are given in Table III. The model-based SE gave the largest SE estimate, while the sandwich SE estimate was the smallest. The use of any combination of SE estimate and distribution to obtain critical values from resulted in the rejection of the null hypothesis, implying that there is strong enough evidence at the 5 per cent significance level to conclude that the intervention was effective. The pseudo-Wald statistic appeared to give slightly stronger evidence supporting a treatment effect as compared with the use of W_{Reg}. The use of the sandwich SEs gave the strongest support for a treatment effect, although these SE estimates may be biased downward.

5. Concluding remarks

Many GRTs randomize a relatively small number of clusters. When the data to be analyzed from this setting is in the form of a binary observation from each study participant, our proposed pseudo-Wald statistic, $\widetilde{W}_{1.5}$, outperforms existing Wald statistics using model-based or sandwich SEs. The Wald statistic using model-based SEs can produce a test size smaller than the nominal value, and therefore will produce less power than our pseudo-Wald statistic under the alternative hypothesis. Additionally, test size can be too large and is inconsistent when using the sandwich-based methods. Therefore, we recommend that $\widetilde{W}_{1.5}$ should be utilized with t_n critical values, or with the proposed inflated SE given in equation (5) and $N(0,1)$ critical values, for hypothesis testing and obtaining confidence intervals.

One may be interested in the validity of the use of $\widetilde{W}_{1.5}$ when the nominal level is, say, 0.01 or 0.10, rather than a traditional value of 0.05. The density corresponding to $\widetilde{W}_{1.5}$ is just a widening of the bell-shaped density of W_{Reg}, especially in the tails. $\widetilde{W}_{1.5}$ produces test sizes at the nominal level of 0.05 and also has a bell-shaped density, making it unlikely that choosing a nominal level smaller than 0.05 will yield a larger test size. With levels larger than 0.05, where the density corresponding to W_{Reg} is very similar to the $N(0,1)$ density, there is a possibility that $\widetilde{W}_{1.5}$ may lead to a slightly inflated realized test size. However, this possible increase in realized size will be of little concern for fixed nominal levels regularly used in practice.

Throughout this manuscript, we have assumed that the ICC is equal for all clusters. If the correlation varies from cluster to cluster, our bias and covariance formulas for the MQL estimates take this variable correlation into account, and our method is still valid when the varying correlations are known. However, the correlations will need to be estimated in practice. If one were to incorrectly assume a common ICC, the size of the proposed pseudo-Wald test may not be nominal if the quantities $\hat{q}_i = n_i/[1+(n_i-1)\hat{\rho}_{\mathrm{ANOVA}}]$, $i=1,2,\ldots n$, are not close to the values that would have resulted in correctly specified cluster-specific correlation estimates. This is more likely to occur in scenarios in which cluster sizes are large and correlation varies moderately, or with smaller clusters with large variations in ICC.

Our proposed method can be useful in the analysis of rare or common events data where $|\pi_C - 0.5|$ and $|\pi_T - 0.5|$ are very close to 0.5. Owing to the marginal probabilities being near the boundary of the parameter space, test size can be smaller than the nominal level even if there is no correlation. Here, our method has no additional limitations as

compared to the traditional Wald statistic. Furthermore, our methods are applicable to any non-GRT setting that produces cluster-correlated binary data. These settings include teratology experiments and studies collecting repeated measures on the same subjects.

The results of this paper focus on Wald tests for a marginal model parameter. Another popular approach for the analysis of correlated binary data is the use of a generalized linear mixed effects model (GLMM), in which the correlation is modeled as a random cluster effect, thereby making interpretation of mean parameters conditional for a given cluster. In the Atri *et al.* [1] study, the interest was in increasing breast screening in the population, and so a marginal model and our methods would be more suitable than a conditional or random effects approach. A marginal interpretation may not be as suitable, though, if we had a scenario where a binary outcome were measured repeatedly on each patient and a subject-specific interpretation of a mean parameter were of primary interest. In this setting, our methods would not be used, although generalization of our methods to random effects model parameters is certainly worthy of research.

Our presentation did not implement $\hat{\beta}_{1\mathrm{BC}}$ in the numerator for any of the presented Wald statistics, although it will contain less, if any, bias than $\hat{\beta}_{1\mathrm{MQL}}$; additionally, its variance is smaller. Neither Cordeiro and McCullagh [18] nor Cordeiro and Demetrio [19] proposed a variance estimator for $\hat{\boldsymbol{\beta}}_{\mathrm{BC}}$. Cordeiro and McCullagh showed that for logistic regression with no overdispersion, the bias in the maximum likelihood estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, is approximately $p\boldsymbol{\beta}/n$ for small $\boldsymbol{\beta}$. From this, King and Zeng [31] proposed estimating the variance of $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ by multiplying the model-based variance by $[n/(n+p)]^2$. We utilize a similar approach; by taking into account that the BCE is approximately a fraction of the MQL estimate, we suggest estimating the SE for $\hat{\beta}_{1\mathrm{BC}}$ by multiplying the estimated SE for $\hat{\beta}_{1\mathrm{MQL}}$ by $\hat{\beta}_{1\mathrm{BC}}/\hat{\beta}_{1\mathrm{MQL}}$. One can then incorporate this with the results of our paper, i.e. use $(\hat{\beta}_{1\mathrm{BC}}/\hat{\beta}_{1\mathrm{MQL}})\widetilde{\mathrm{SE}}_{1.5}(\hat{\beta}_{1\mathrm{MQL}})$ as the pseudo-SE estimate implemented inside a Wald statistic with $\hat{\beta}_{1\mathrm{BC}}$ in the numerator. This quantity is equivalent to $\widetilde{W}_{1.5}$; therefore, test size remains unchanged. Simulations (not shown) demonstrated that using $\hat{\beta}_{1\mathrm{BC}}$ with this SE estimate yields approximately the same coverage probability as if $\hat{\beta}_{1\mathrm{MQL}}$ were utilized; however, due to $\hat{\beta}_{1\mathrm{BC}}$ being approximately unbiased and less variable, it will yield a more desirable confidence interval.

In further research on testing for a marginal treatment effect, we will study test size resulting from the typical Wald test using the model-based SE when the outcomes of interest are not binary responses. We will also extend our model to include other covariates. Further study is also needed to find more exact SE formulas for the BCEs, and to determine if these would carry more accuracy and utility than our proposed formula.

An R function that implements our proposed pseudo-Wald tests, and also outputs our suggested 95 per cent confidence intervals using $\hat{\beta}_{1\mathrm{BC}}$ as the point estimate, can be obtained by contacting the author at pwestgat@umich.edu.

References

1. Atri J, Falshaw M, Gregg R, Robson J, Omar RZ, Dixon S. Improving uptake of breast screening in multiethnic populations: a randomised controlled trial using practice reception staff to contact non-attenders. *British Medical Journal* 1997; **315**:1356–1359.
2. Prentice RL. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 1986; **81**:321–327.
3. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 1993; **80**:517–526.
4. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 1974; **61**:439–447.
5. Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society*, *Series A* 1972; **135**:370–384.
6. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
7. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
8. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; **96**:1387–1396.
9. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
10. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* 2004; **94**:423–432.
11. Pan W. On the robust variance estimator in generalised estimating equations. *Biometrika* 2001; **88**:901–906.
12. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; **57**:1198–1206.
13. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 2002; **21**:1429–1441.
14. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of gee estimators. *Biometrical Journal* 2003; **45**:395–409.
15. McCaffrey DF, Bell RM. Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. *Statistics in Medicine* 2006; **25**:4081–4098.
16. Drum M, McCullagh P. Comment on 'regression models for discrete longitudinal responses'. *Statistical Science* 1993; **8**:300–301.
17. Liang KY, Hanfelt J. On the use of the quasi-likelihood method in teratological experiments. *Biometrics* 1994; **50**:872–880.
18. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society*, *Series B* 1991; **53**:629–643.
19. Cordeiro GM, Demetrio CGB. Corrected estimators in extended quasi-likelihood models. *Communications in Statistics—Theory and Methods* 2008; **37**:873–880.

20. Bull SB, Greenwood CMT. Jackknife bias reduction for polychotomous logistic regression. *Statistics in Medicine* 1997; **16**:545–560.
21. Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 1977; **72**:851–853.
22. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; **55**:137–148.
23. Williams DA. Extra-binomial variation in logistic-linear models. *Journal of the Royal Statistical Society*, *Series C* 1982; **31**:144–148.
24. Carroll RJ, Ruppert D. Robust estimation in heteroscedastic linear models. *The Annals of Statistics* 1982; **10**:429–441.
25. Davidian M, Carroll RJ. Variance function estimation. *Journal of the American Statistical Association* 1987; **82**:1079–1091.
26. Nelder JA, Pregibon D. An extended quasi-likelihood function. *Biometrika* 1987; **74**:221–232.
27. Donner A, Donald A. The statistical analysis of multiple binary measurements. *Journal of Clinical Epidemiology* 1988; **41**:899–905.
28. Reed JF. Eliminating bias in randomized cluster trials with correlated binomial outcomes. *Computer Methods and Programs in Biomedicine* 2000; **61**:119–123.
29. Jung SH, Kang SH, Ahn C. Sample size calculations for clustered binary data. *Statistics in Medicine* 2001; **20**:1971–1982.
30. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* 2001; **20**:453–472.
31. King G, Zeng L. Logistic regression in rare events data. *Political Analysis* 2001; **9**:137–163.