Research Article

Statistics in Medicine

Received 31 May 2011, Accepted 10 January 2012 Published online 13 March 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5329

The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions

Philip M. Westgate[a*†] and Thomas M. Braun[b]

Generalized estimating equations (GEE) are commonly used for the analysis of correlated data. However, use of quadratic inference functions (QIFs) is becoming popular because it increases efficiency relative to GEE when the working covariance structure is misspecified. Although shown to be advantageous in the literature, the impacts of covariates and imbalanced cluster sizes on the estimation performance of the QIF method in finite samples have not been studied. This cluster size variation causes QIF's estimating equations and GEE to be in separate classes when an exchangeable correlation structure is implemented, causing QIF and GEE to be incomparable in terms of efficiency. When utilizing this structure and the number of clusters is not large, we discuss how covariates and cluster size imbalance can cause QIF, rather than GEE, to produce estimates with the larger variability. This occurrence is mainly due to the empirical nature of weighting QIF employs, rather than differences in estimating equations classes. We demonstrate QIF's lost estimation precision through simulation studies covering a variety of general cluster randomized trial scenarios and compare QIF and GEE in the analysis of data from a cluster randomized trial. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: cluster randomized trial; empirical covariance; estimating equations class; GEE; repeated measures

1. Introduction

Correlated data with imbalanced cluster sizes arise often in practice. Cluster randomized trials (CRTs) and longitudinal studies in which the number of repeated measures is not constant across subjects are two popular examples where data are composed of independent clusters that typically are comprised of varying sizes. With these types of data, individual-level responses within any given cluster are assumed to be correlated.

We particularly focus on CRTs, which are unique from other randomized trials. They typically are comprised of a small number of independent clusters that can be quite large and variable in size. Because of these attributes, statistical power can be quite low, but the study itself can be very costly to conduct. When the desired interpretations for regression parameters are in terms of the population mean, generalized estimating equations (GEE) [1] are a popular tool of choice for the analysis of data arising from these trials. They require working correlation and marginal variance structures to be given, but only the mean structure needs correct specification to obtain consistent parameter estimates.

Because of potentially low power in these studies, the use of a more efficient method would be very beneficial, which is why we focus on the estimation performance of quadratic inference functions (QIFs). With the same limited requirements as GEE, Qu *et al.* [2] proposed QIF as an alternative method, which is a combination of GEE and the generalized method of moments (GMMs) [3]. QIF asymptotically has greater efficiency than GEE when employing the incorrect covariance structure and is as efficient when using the correct structure [2, 4, 5]. This result depends on all cluster sizes being equal if a common

[a] *Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY 40536, U.S.A.*
[b] *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, College of Public Health, University of Kentucky, Lexington, KY 40536, U.S.A.*
[†] *E-mail: philip.westgate@uky.edu*

exchangeable correlation is implemented, such that both procedures' estimating equations are within the same class. Many papers, such as Qu *et al.* [2], demonstrate the utility QIF has over GEE when using a working exchangeable or AR-1 correlation matrix. No paper, however, has studied the finite sample estimation precision, or reliability, of QIF as compared with GEE when cluster sizes vary and the exchangeable structure is reasonably employed, such as in common CRT settings. Additionally, the effect of covariates on QIF's estimation performance has not been considered.

Our motivating dataset comes from Yudkin and Moher [6], who discuss issues with an ongoing CRT dealing with coronary heart disease (CHD) and promoting secondary prevention via two interventions as compared with a control that gives ordinary care to patients. Some concerns, for example, involve the fact that there are a limited number of large clusters, restricted randomization and sample size calculations are difficult, and the choice of an appropriate method for data analysis is not necessarily straightforward. For instance, they utilized a stratified design but indicated that use of an analysis that takes stratification into account may not be best. Furthermore, there is an issue in CRTs on whether subject or cluster-level analyses should be used. Subject-level analyses include random effects or marginal models, whereas cluster-level analyses use a weighted or unweighted t-test or linear regression with cluster-specific summary measures as outcomes and weights that take into account cluster size and possibly the estimated exchangeable correlation value [7, 8]. A thorough discussion of these topics can be found in Donner and Klar [9], and Campbell *et al.* [10] gave an extension of the CONSORT statement to CRTs.

Yudkin and Moher [6] give a table of baseline results on four variables and the size, ranging from 28 to 244 patients, of each of the 21 practices, or clusters, participating in the study. Using the presented data, we found the number of patients in each practice who were recently adequately assessed for three CHD risk factors. Particularly, Yudkin and Moher [6] defined adequate assessment as the recent recording and assessment of blood pressure, smoking habit, and serum cholesterol. The other three variables, practice-level proportions of patients having a record of treatment with aspirin, hypotensives, or lipid-lowering drugs since their diagnosis with CHD, were secondary outcomes.

One specific issue Yudkin and Moher [6] discuss with the baseline data is how to utilize restricted randomization of practices to trial arms such that balance, in terms of adequate assessment and the three records of treatment, is achieved. We use this data, however, to quantify the association between the marginal probability of a practice, which gives ordinary patient care, having recently adequately assessed any given patient and the proportion of patients in that practice having a record of treatment with any of the three drug types. This model was chosen such that we could utilize the available data to demonstrate the differences between QIF and GEE in terms of weighting and estimation, although a limitation is that this model may be slightly implausible. Use of the logistic link would be common for this marginal model, in which the proportion of patients for any one of the three records of drug treatment could simply be used as a continuous covariate. The true, but unobserved, probabilities of adequately assessing any given patient can vary across practices about their marginal means because of unknown factors, thus inducing correlation among patients within the same practice. Therefore, a common exchangeable correlation would be a natural structure to implement with GEE and QIF. As covariate values are the same for each patient in any given practice, a cluster-level weighted linear regression would also be a possible alternative.

We later show that GEE and QIF can produce notably different probability estimates from the analysis of this data, leading to the issue of which method gave more trustworthy estimates and, in general, which of these two methods would be best to use for the analysis of data from any CRT. As QIF theoretically is equally or more efficient than GEE, one may think that its estimates here would be more reliable. For example, if the true exchangeable correlation value for any given practice depends on the proportion of patients with a record of drug treatment in that same practice, QIF should take this into account via its empirical weighting matrix, whereas GEE using a common exchangeable structure does not. The purpose of this paper is to give details into how QIF and GEE can give notably different estimates in this or any other CRT setting, with a direct focus on the impact of cluster-level covariates, and why GEE may actually be better to employ in CRT scenarios.

Section 2 discusses GEE and QIF in more detail, including comparisons of their respective classes of estimating equations when an exchangeable correlation structure is implemented. We discuss in Section 3 how empirical weighting, which is influenced by cluster size imbalance and covariates, can cause QIF and its modified version we present, utilizing estimating equations that are in the same class as GEE, to lose estimation precision relative to GEE when the number of clusters is not large. In Section 4, we present simulation results, with emphasis on our motivating dataset and general CRTs, demonstrating the differences in the precisions of parameter estimates from GEE and both QIF versions. Furthermore, the distinct estimation performances of these methods are shown in application to the motivating dataset. Concluding remarks are given in Section 5.

2. Marginal models

2.1. Generalized estimating equations

We have N independent clusters of data, and cluster i, $i = 1, 2, \dots N$, has n_i observations, outcome vector $\boldsymbol{Y}_i = [Y_{i1}, \dots, Y_{in_i}]^T$, and mean vector $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$. The marginal mean structure is specified as $h(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{x}_i\boldsymbol{\beta}$, where the jth row of $\mathbf{x}_i$, $j = 1, 2, \dots n_i$, is $\mathbf{x}_{ij} = [x_{ij0}, x_{ij1}, \dots, x_{ij(p-1)}]$, the vector of covariate values for the jth observation in cluster i, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_{p-1}]^T$ is a $p \times 1$ vector of corresponding regression parameters. The estimates for $\boldsymbol{\beta}$ are obtained by setting the GEE equal to zero:

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{1}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\boldsymbol{V}_i$ is the working covariance structure for $\boldsymbol{Y}_i$. $\boldsymbol{V}_i$ can be written as $\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i$ is a diagonal matrix of the working marginal variances for the n_i observations, and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is their working correlation structure with parameter(s) $\boldsymbol{\alpha}$. When the covariance structure is correctly specified and a consistent estimate for $\boldsymbol{\alpha}$ is employed, GEE as given in Equation (1) are optimal estimating equations [11]. If $\boldsymbol{V}_i$ is misspecified, the parameter estimates, $\hat{\boldsymbol{\beta}}$, are still consistent when the mean structure is correct.

When implementing an exchangeable correlation structure, Equation (1) can be rewritten as

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2}(\gamma_{1i}\boldsymbol{M}_{1i} + \gamma_{2i}\boldsymbol{M}_{2i})\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{2}$$

where $\gamma_{1i} = -[(n_i - 2)\rho_i + 1]/k_i$, $\gamma_{2i} = \rho_i/k_i$, $k_i = (n_i - 1)\rho_i^2 - (n_i - 2)\rho_i - 1$, $\boldsymbol{M}_{1i}$ is an $n_i \times n_i$ identity matrix, $\boldsymbol{M}_{2i}$ is an $n_i \times n_i$ matrix composed of zeros on the diagonal and ones elsewhere, and ρ_i is a function of $\boldsymbol{\alpha}$ and is the assumed common correlation within the ith cluster [2]. If cluster sizes are all equal and a constant correlation is assumed across clusters, Equation (2) is easily seen as being in the class of estimating equations given by

$$\sum_{r=1}^{2} \boldsymbol{B}_r \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2}\boldsymbol{M}_r\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3}$$

where $\boldsymbol{B}_r$, $r = 1, 2$, are $p \times p$ arbitrary nonrandom matrices. $\boldsymbol{M}_{1i}$ and $\boldsymbol{M}_{2i}$, $i = 1, 2, \dots N$, do not change across clusters and therefore are denoted here as $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. They can be thought of as basis matrices because all other quantities inside the two sums over the N clusters are the same [2]. With respect to GEE, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are identity matrices multiplied by γ_1 and γ_2, respectively, where $\gamma_r = \gamma_{ri}$, $r = 1, 2$; $i = 1, 2, \dots N$. When clusters vary in size, or a common correlation is no longer used across all clusters, GEE belongs to the class given by

$$\sum_{r=1}^{2} \boldsymbol{O}_r \sum_{i=1}^{N} \gamma_{ri}\boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{ri}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4}$$

which is distinct from Equation (3) in that the basis matrices depend on size and the two sums over the N clusters now include the unique values for γ_{ri}, $r = 1, 2$; $i = 1, 2, \dots N$. Here, $\boldsymbol{O}_r$, $r = 1, 2$, are $p \times p$ arbitrary nonrandom matrices equal to the identity matrix for GEE.

2.2. Quadratic inference functions

The QIF proposed by Qu *et al.* [2] combines the methods of GMMs and GEE. It assumes $\boldsymbol{R}_i^{-1}(\boldsymbol{\alpha}) = \sum_{r=1}^{m} \gamma_{ri}\boldsymbol{M}_{ri}$, where $\boldsymbol{M}_{ri}$, $r = 1, 2, \dots m$, are known basis matrices and γ_{ri}, $r = 1, 2, \dots m$, are functions of $\boldsymbol{\alpha}$ that we will refer to as correlation weights. An exchangeable structure is a specific case where the inverse of the correlation matrix can be written as the sum of weighted basis matrices, as shown in

Equation (2). Unstructured, AR-1, and independence are the other correlation structures QIF currently support via this assumption, each of which have inverses that can at least be approximated by using two basis matrices [5]. We do not focus on these in this paper because QIF and GEE lead to identical estimating equations when using independence [12], and in CRT settings unstructured is not plausible, whereas exchangeable is more commonly employed than AR-1. However, although simulation settings focus on the exchangeable structure, the discussion in Section 3 on the variable empirical weighting utilized by QIF, and its impact on estimation precision, holds for unstructured and AR-1 as well.

Equation (2) can be viewed as the sum of two unbiased estimating equations, each of which are used to build extended score equations defined as

$$\bar{g}_N(\boldsymbol{\beta}) = \frac{1}{N} g_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} g_{1i}(\boldsymbol{\beta}) \\ \frac{1}{N}\sum_{i=1}^{N} g_{2i}(\boldsymbol{\beta}) \end{bmatrix}. \qquad (5)$$

The number of extended score equations is twice the number of regression parameters and therefore cannot be set equal to zero to obtain parameter estimates, as is carried out for GEE, because no identifiable solution exists. The extended score equations are used in Hansen's [3] GMMs to create the QIF, defined as

$$Q_N(\boldsymbol{\beta}) = N\bar{g}_N^T(\boldsymbol{\beta})C_N^{-1}(\boldsymbol{\beta})\bar{g}_N(\boldsymbol{\beta}) = \left[\sum_{i=1}^{N} g_i^T(\boldsymbol{\beta})\right]\left[\sum_{i=1}^{N} g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta})\right]^{-1}\left[\sum_{i=1}^{N} g_i(\boldsymbol{\beta})\right].$$

The estimate for $\boldsymbol{\beta}$ can now be found by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$, which is asymptotically equivalent to solving

$$N\nabla\bar{g}_N^T(\boldsymbol{\beta})C_N^{-1}(\boldsymbol{\beta})\bar{g}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N}\nabla\bar{g}_N^T(\boldsymbol{\beta})C_N^{-1}(\boldsymbol{\beta})g_i(\boldsymbol{\beta}) = \mathbf{0} \qquad (6)$$

for $\boldsymbol{\beta}$, where ∇ denotes the gradient with respect to $\boldsymbol{\beta}^T$. Here, $C_N(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N} g_i(\boldsymbol{\beta})g_i^T(\boldsymbol{\beta})$ is used to estimate the optimal weight matrix $\boldsymbol{\Sigma}_N = (1/N)\sum_{i=1}^{N} Cov[g_i(\boldsymbol{\beta})]$. Results using $C_N(\boldsymbol{\beta})$ are asymptotically equivalent to using $\boldsymbol{\Sigma}_N$ because $C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N \overset{p}{\to} 0$ [2, 13].

In practice, $g_{ri}(\boldsymbol{\beta}) = \boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{ri}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, 2$; $i = 1, 2, \ldots N$, are regularly implemented in QIF's extended score equations. These ignore the correlation weights, implying that $\boldsymbol{\alpha}$ does not need to be estimated. When using an exchangeable structure and cluster sizes do not vary, Equation (6) is in the class of estimating equations given by Equation (3). When cluster sizes vary, Equation (6) is in the class of estimating equations given by

$$\sum_{r=1}^{2}\boldsymbol{O}_r\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{ri}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

This is not the class given by Equation (4) to which GEE belongs in this scenario, as the correlation weights are not included. For Equation (6) and GEE to be in the same class when cluster sizes vary, the extended score equations need to incorporate the correlation weights, that is, use $g_{ri}(\boldsymbol{\beta}) = \gamma_{ri}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_{ri}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $r = 1, 2$; $i = 1, 2, \ldots N$, inducing the need to estimate $\boldsymbol{\alpha}$.

From Lindsay [14], Hansen [3], and Small and McLeish [11], Qu *et al.* [2] show that because $C_N(\boldsymbol{\beta}) - \boldsymbol{\Sigma}_N \overset{p}{\to} 0$, the estimating equations given in Equation (6) are fully efficient when the covariance structure is correctly specified and they are in the same class as GEE, and always optimal in the Löwner ordering among estimating equations within their given class. When cluster sizes are constant and assuming a common correlation with an exchangeable structure, Equation (6) and GEE are in the class of estimating equations given by Equation (3); therefore, QIF has the theoretical advantage of asymptotically producing parameter estimates having equal or less variance than estimates from GEE. When cluster sizes vary, QIF loses this theoretical advantage unless the correlation weights are used inside the extended score equations. In this case, GEE and Equation (6) both belong to the class given by Equation (4).

3. Empirical weighting, quadratic inference function estimation precision, and the impacts of imbalanced cluster sizes and covariates

The QIF method uses an estimated empirical weighting matrix, $C_N(\hat{\boldsymbol{\beta}})$, to estimate $\boldsymbol{\Sigma}_N$, which is optimal. Asymptotically, using this matrix is the source of QIF's efficiency advantage. Even when clusters vary in size, causing QIF and GEE to not be directly comparable in terms of efficiency theory when using a common exchangeable correlation, QIF still has an advantage in the sense that it is composed of a weighting matrix that is consistent for the true covariance structure. However, we soon explain that because of its variability, the use of $C_N(\hat{\boldsymbol{\beta}})$ can lead to lost, rather than gained, estimation precision as compared with GEE, in small to moderately sized samples, such as the CRT dataset containing only 21 practices. Both covariates and imbalance in cluster sizes impact the empirical information used from C_N to obtain the working weights inside QIF's estimating equations. The weight(s) used by QIF for any given cluster's outcomes can therefore be quite variable when N is not large. Conversely, GEE uses a fixed correlation, and thus weighting, structure for each cluster's outcome(s). The only variability in this weighting strategy is due to estimating a common correlation parameter.

Equation (4) is helpful for distinguishing the different sources of weighting variability for both methods. As stated for GEE, $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ are identity matrices, and the variability in each scalar, $\gamma_r = \gamma_{ri}$, $r = 1, 2$; $i = 1, 2, \ldots N$, from estimating the correlation parameter will not have a large impact on weighting. However, along with $\nabla \bar{g}_N^T$, QIF use C_N to estimate $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ and utilizes $\gamma_r = \gamma_{ri} = 1$, $r = 1, 2$; $i = 1, 2, \ldots N$. As N decreases, the variability in C_N, and thus $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$, increases. This creates increasing variances in the weights given to $\boldsymbol{Y}_i$, $i = 1, 2, \ldots N$, because $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ are random matrices, rather than scalars, that influence the entire weighting structure. We present via simulation an example of the difference in variances of weights utilized by QIF and GEE in Section 4.1.1. We note that a modified QIF implementing the correlation weights, and therefore employing estimating equations that are in the same class as GEE, incorporates both sources of weighting variability. Therefore, our discussion on the variable empirical weighting nature utilized by QIF also pertains to this newly defined QIF version.

Qu and Song [12] exhibited the sensitive, or variable, nature of weighting used by QIF via C_N, as they show that QIF is robust to outliers and contaminated data because of its use of empirical estimates for $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$. Specifically, they prove $||\nabla \bar{g}_N^T(\boldsymbol{\beta}) C_N^{-1}(\boldsymbol{\beta}) g_i(\boldsymbol{\beta})||^2 \rightarrow 0$ as $||\boldsymbol{Y}_i - \boldsymbol{\mu}_i|| \rightarrow \infty$, where $||\boldsymbol{K}|| = [tr(\boldsymbol{K}^T \boldsymbol{K})]^{1/2}$ for some arbitrary matrix $\boldsymbol{K}$. This result indicates that the amount of weight given to outcomes from any given cluster depends on the empirical covariances of these same outcomes. Alternatively, GEE uses the identity matrix for $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$, implying that it does not have this sensitive weighting nature as an attribute. However, $C_N(\hat{\boldsymbol{\beta}})$ is asymptotically equivalent to $\boldsymbol{\Sigma}_N$, which does not depend on empirical covariances. Therefore, $C_N(\hat{\boldsymbol{\beta}})$ becomes less sensitive to individual estimated empirical covariances as N increases, implying that QIF's weighting strategy becomes less variable.

When QIF's weights are quite variable because of small N, $Var(\hat{\boldsymbol{\beta}})$ can be large as well. Therefore, the estimated empirical covariances $(g_i(\hat{\boldsymbol{\beta}}) g_i^T(\hat{\boldsymbol{\beta}})$, $i = 1, 2, \ldots N)$ used in practice can be notably different than the true empirical covariances $(g_i(\boldsymbol{\beta}) g_i^T(\boldsymbol{\beta})$, $i = 1, 2, \ldots N)$ in some samples. This is a drawback because $E[C_N(\hat{\boldsymbol{\beta}})|\hat{\boldsymbol{\beta}}] \neq E[C_N(\boldsymbol{\beta})] = \boldsymbol{\Sigma}_N$ for $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ and finite N. Particularly, because of the variable weighting nature used by QIF, these differences between estimated and true empirical covariances can lead to notable differences between the estimated and true weights given to outcomes from any given cluster, which affect parameter estimation. As GEE does not have this small-sample problem, it can produce parameter estimates having smaller variances. Additionally, Windmeijer [15] notes in the Econometrics literature for GMM that for finite N, the variance of $\hat{\boldsymbol{\beta}}$ can be larger than expected because of the use of an estimate for $\boldsymbol{\beta}$ inside the empirical weighting matrix.

To discuss the impacts of imbalanced cluster sizes and covariates, we note that QIF has a variable weighting nature that uses averaged information from sensitivities [4] and empirical covariances via $\nabla \bar{g}_N^T C_N^{-1}$ inside its estimating equations to determine how much individual weight should be given to each $g_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$. Specifically, QIF uses $\nabla \bar{g}_N^T C_N^{-1}$ to obtain $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ to weight $g_{1i}(\boldsymbol{\beta})$ and $g_{2i}(\boldsymbol{\beta})$, $i = 1, 2, \ldots N$, respectively. With the use of information from all N practices, QIF takes into account how sensitivities and empirical covariances change on average with respect to size and covariate values to determine the weights. Therefore, when clusters vary in size, the average sensitivity and empirical covariance trends have to be determined with respect to the numerous combinations of size and covariates. As covariates are added to the model, the dimension of C_N increases, and there is greater complexity, and thus variability, in weighting.

For example, we have p cluster-level covariates, and $h(.)$ is the canonical link, allowing Equation (6) to simplify to

$$\sum_{i=1}^{N}\begin{bmatrix} \sum_{j=0}^{p-1}\kappa_{j0}x_{ij} + (n_i-1)\sum_{j=p}^{2p-1}\kappa_{j0}x_{i(j-p)} \\ \vdots \\ \sum_{j=0}^{p-1}\kappa_{j(p-1)}x_{ij} + (n_i-1)\sum_{j=p}^{2p-1}\kappa_{j(p-1)}x_{i(j-p)} \end{bmatrix}(Y_i - n_i\mu_i) = \mathbf{0}, \qquad (7)$$

in which $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, $E(Y_{ij}) = \mu_i$, and x_{ij} is the value of the jth covariate, $j = 0, \ldots p-1$, for the ith cluster. All kappas are estimated using functions of estimated parameters, cluster size, covariate values, and the N empirical covariances, which have the most influence. The number of kappa parameters increases with the number of covariates, and the amount of weight given to any cluster's outcomes relies upon linear combinations of its size and covariate values. For fixed N, the amount of empirical information from our sample remains unchanged. However, if the number of kappa parameters increases, this fixed amount of information utilized from empirical covariances results in greater weighting variability because of the complexity of the numerous combinations of size and covariates. For example, when only using one drug treatment percentage as a covariate, QIF estimates eight kappas, whereas this number increases to 32 when using all three treatments in the model.

To clearly demonstrate the variability in QIF's estimating equations' weights due to imbalanced cluster sizes and covariates, we first use an intercept-only model example in which clusters are one of two possible sizes. There are two kappas to estimate, and Equation (7) reduces to $\sum_{i=1}^{N}[\kappa_{00} + \kappa_{10}(n_i - 1)]$ $(Y_i - n_i\mu) = \sum_{i=1}^{N} w_{\mathrm{QIF}i}(Y_i - n_i\mu)$, where μ is the marginal mean shared by all outcomes. Here, the weight given to the ith cluster will actually be estimated using only the empirical covariances of the extended score equation components from clusters of that vary same size, rather than using the empirical covariances from all N clusters. This nature of weighting is advantageous for QIF when N is arbitrarily large and the true covariances do depend on cluster size in some misspecified manner, but for small N, it can lead to increased variability in the working weights that can more than offset this advantage in terms of estimation performance.

If we keep N fixed and extend the model to resemble a general CRT in which there is only one covariate, a cluster-level intervention indicator, then there are eight unknown kappas, each estimated with a larger variability. In this situation, QIF carries out estimation in a manner equivalent to fitting an intercept-only model for each trial arm. This implies that $\nabla\bar{g}_N^T C_N^{-1}$ accounts for combinations corresponding to study arm and cluster size, and so the weight given to $(Y_i - n_i\mu_i)$, $i = 1, 2, \ldots N$, is obtained using only the empirical covariances from equivalently sized clusters within the same trial arm.

In practice, as is the case with our motivating dataset, there typically is much larger imbalance in size across clusters. When this occurs, $\nabla\bar{g}_N^T C_N^{-1}$ has to determine if the weight given to outcomes should increase or decrease with size, which is carried out separately for each trial arm. For instance, suppose outcomes from clusters larger in size have averaged empirical covariances smaller than their true covariances. Although clusters cannot be combined into distinct groups as was the case when there were only two possible sizes, QIF's estimating equations will overweight larger clusters here, whereas the weight given to smaller clusters may also be influenced as a result of the linear trend shown in the intercept-only model's estimating equation.

With respect to our motivating dataset, we first use an intercept-only model. Here, $\nabla\bar{g}_N^T(\hat{\boldsymbol{\beta}})C_N^{-1}(\hat{\boldsymbol{\beta}})$ estimates how the weight, given to the number of adequately assessed patients in any given practice, should change with respect to practice size. If we were to expand this model even further by using aspirin treatment percentage, Equation (7) shows that there will be more unknown kappa parameters to estimate. If we were to include the other two drug treatment percentages in the model as well, there are even more combinations of size and covariates that QIF has to account for while using empirical covariances from only 21 practices, potentially resulting in large weighting variances.

4. The impacts of cluster sizes and covariates

4.1. *Shown via simulation study*

4.1.1. Intercept-only simulations. Employing a common exchangeable correlation, we first demonstrate the difference between both QIF versions and GEE in the context of an intercept-only model, representing the setting in which we are only interested in estimating the marginal probability of adequate assessment. Results from 10 random simulations, with outcomes generated from a beta-binomial distribution, are presented in Table I, including the intercept estimates and ratios equaling the estimated weight given to a cluster of size 50 divided by the estimated weight given to a cluster of size 150. We generated data using the model $\text{logit}(\pi) = \log(\pi) - \log(1-\pi) = \beta_0$, in which π is the marginal probability for any given outcome. Values for the marginal probability were 0.25 (0.05) for the first (last) five simulations, implying $\beta_0 = -1.10$ ($\beta_0 = -2.94$), whereas the common correlation was 0.05. Each simulated dataset consisted of 21 practices, with corresponding sizes generated by a normal distribution with mean 100 and standard deviation 50, approximately representing the empirical distribution of sizes contained in the motivating dataset. We rounded generated sizes to the nearest integer and forced them to take on values between 25 and 250. In this setting, the optimal weight is given as $w_i = [1 + (n_i - 1)0.05]^{-1}$ inside the estimating equation $\sum_{i=1}^{N} w_i (Y_i - n_i\pi) = 0$. The optimal ratio is therefore 2.45.

In these 10 simulations, the weights used by GEE were much less variable than those used by either QIF version. For GEE, clusters of size 50 were given anywhere from 2.19 to 2.66 times more weight than clusters of size 150. QIF, however, once gave more weight to larger clusters, and clusters of size 50 were given anywhere from 38% less to 6.74 times more weight than clusters of size 150. The QIF with corresponding estimating equations in the same class as GEE produced notably different weight ratios but performed similarly to QIF with respect to the variability in its working weights. Because of this variability in relative weighting, estimation precision is lost as compared with GEE here.

For each of the two examined settings, we also performed 1000 additional simulations generated in the same manner. When the marginal probability was 0.25 (0.05), the empirical mean square error (MSE) for GEE's intercept estimate was only 82% (62%) and 86% (65%) as large as the MSEs produced by QIF and the newly defined QIF, respectively, implying that GEE was more precise. The last five simulation results in Table I show that the weights implemented by both QIF versions were more variable when the marginal probability was 0.05, leading to the smaller MSE ratios in this setting. Additionally, the newly defined QIF only slightly increased precision over the typical QIF.

4.1.2. Description of general simulation settings. Song *et al.* [5] discussed a SAS macro that implements QIF. We utilized its R counterpart, Sister R Package QIF, both of which can currently be found at Dr. Peter X.K. Song's website: http://www-personal.umich.edu/~pxsong/. As this QIF function does not require the value for the QIF to decrease at each iteration of the estimation procedure, we modified it accordingly as suggested by Loader and Pilla [16]. We also adjusted this function to implement our QIF version that utilizes the correlation weights and thus estimating equations that are in the same class as GEE. These functions, in addition to a GEE function that produces bias-corrected standard errors [17] as mentioned in Section 4.2, are available upon request to the author at philip.westgate@uky.edu.

We now compare the MSE of both QIF versions and GEE, all implementing a common exchangeable correlation structure, in a variety of simulations representing CRT scenarios. Table II presents empirical

Table I. Intercept estimates and weight ratios, equaling the estimated weight given to a cluster of size 50 divided by the estimated weight given to a cluster of size 150, from GEE and both QIF versions.

	GEE		QIF		QIF in GEE class	
Simulation	$\hat{\beta}_0$	Ratio	$\hat{\beta}_0$	Ratio	$\hat{\beta}_0$	Ratio
1	−1.18	2.49	−1.18	1.50	−1.18	2.01
2	−0.95	2.59	−0.99	0.62	−0.97	1.00
3	−1.12	2.47	−1.17	1.10	−1.15	1.52
4	−1.15	2.66	−1.13	1.87	−1.13	3.13
5	−0.99	2.64	−0.99	1.51	−0.98	4.03
6	−3.23	2.34	−3.32	2.19	−3.32	4.39
7	−3.09	2.25	−3.06	1.44	−3.09	2.20
8	−2.88	2.19	−3.02	6.74	−3.07	17.00
9	−3.26	2.50	−3.22	3.94	−3.31	5.33
10	−2.93	2.33	−2.92	1.22	−2.93	2.09

The first (last) five simulation results come from the analyses of randomly generated datasets in which outcomes had a marginal probability of 0.25 (0.05) and exchangeable correlation of 0.05.
QIF, quadratic inference functions; GEE, generalized estimating equations.

Table II. Empirical MSEs for both QIF versions and ratios comparing GEE's empirical MSEs to these respective quantities.

				QIF		QIF in GEE Class	
Scenario	N	Range for n_i	ρ_C, ρ_T	MSE	MSE ratio	MSE	MSE ratio
1.1	20	5, 20	0.10, 0.10	0.293	0.660	0.297	0.650
1.2	20	5, 20	0.15, 0.05	0.268	0.710	0.264	0.720
1.3	20	25, 150	0.10, 0.10	0.183	0.745	0.175	0.782
1.4	20	25, 150	0.15, 0.05	0.177	0.722	0.174	0.735
1.5	100	5, 20	0.10, 0.10	0.043	0.889	0.043	0.891
1.6	100	5, 20	0.15, 0.05	0.043	0.906	0.043	0.909
1.7	100	25, 150	0.10, 0.10	0.029	0.927	0.029	0.926
1.8	100	25, 150	0.15, 0.05	0.027	0.881	0.026	0.925
2.1	20	5, 20	0.10, 0.10	1.306	0.537	1.290	0.544
2.2	20	5, 20	0.15, 0.05	1.212	0.562	1.214	0.561
2.3	20	25, 150	0.10, 0.10	0.666	0.595	0.639	0.621
2.4	20	25, 150	0.15, 0.05	0.641	0.630	0.631	0.640
2.5	100	5, 20	0.10, 0.10	0.136	0.866	0.136	0.869
2.6	100	5, 20	0.15, 0.05	0.126	0.897	0.125	0.901
2.7	100	25, 150	0.10, 0.10	0.082	0.852	0.081	0.868
2.8	100	25, 150	0.15, 0.05	0.083	0.879	0.082	0.894
3.1	20	25, 150		0.362	0.708	0.384	0.668
3.2	20	25, 150		0.093	1.123	0.091	1.144
3.3	100	25, 150		0.007	1.165	0.007	1.161
3.4	100	25, 150		0.030	1.032	0.030	1.029
4.1	21	25, 250		3.1×10^{-4}	0.670	2.6×10^{-4}	0.808
4.2	100	25, 250		5.7×10^{-5}	0.648	4.3×10^{-5}	0.858
4.3	21	25, 250		3.6×10^{-4}	0.658	3.0×10^{-4}	0.798
4.4	100	25, 250		6.4×10^{-5}	0.724	5.2×10^{-5}	0.891
5.1	21	25, 250		8.0×10^{-4}	0.696	8.3×10^{-4}	0.669
5.2	100	25, 250		1.2×10^{-4}	0.784	1.1×10^{-4}	0.876
5.3	21	25, 250		7.9×10^{-4}	0.682	7.8×10^{-4}	0.690
5.4	100	25, 250		1.2×10^{-4}	0.794	1.1×10^{-4}	0.894

We employed common exchangeable correlation structures with these methods. The scenarios presented are general representations of CRTs and the CRT dataset of interest. We denote by ρ_C and ρ_T the correlation values for control and intervention clusters, respectively, in Scenarios 1 and 2.

QIF, quadratic inference function; GEE, generalized estimating equations; MSE, mean square error; CRT, cluster randomized trial.

MSEs for each QIF version, in addition to ratios comprised of the MSEs from GEE and the respective QIF version in the numerator and denominator, respectively. The presented MSE quantity for any given method is the sum of the empirical MSEs from all non-intercept regression parameter estimates in the respective model. We examined five different scenarios comprised of four or eight settings each in 1000 simulations.

As with simulated data from the intercept-only models, a beta-binomial distribution was used to generate outcomes. For instance, see Ridout *et al.* [18]. Specifically, cluster i, $i = 1, 2, \ldots N$, has true probability, p_i, such that $E[p_i] = \pi_i$ and $var(p_i) = \rho_i \pi_i (1-\pi_i)$. First, p_i was randomly generated from the corresponding beta distribution with these properties, and then Y_i was drawn from $Binomial(n_i, p_i)$. As our functions require vector outcomes, we let $\boldsymbol{Y}_i$ contain $n_i - Y_i$ zeros and Y_i ones.

In the first three scenarios, which represent general CRTs, the true model in each scenario is a logistic regression with only cluster-level covariates. Scenarios one and three only use an intervention indicator, with $N/2$ clusters in each trial arm. The second scenario uses an additional indicator and a continuous covariate, with corresponding parameters $\beta_2 = \beta_3 = 0$. There were $N/4$ clusters for each of the four possible combinations for the two indicators, and we independently drew the values for the continuous covariate from *Uniform*$(-1, 1)$. Settings for Scenarios 1 and 2 were exactly the same, although we adjust for two covariates that have no impact on marginal probabilities in Scenario 2 to demonstrate their influence on QIF's estimation performance relative to GEE. Cluster sizes varied uniformly and independently from 5 to 20 or 25 to 150 in the first two scenarios, and 25 to 150 in Scenario 3. Marginal probabilities were 0.3 and 0.2 for control and intervention clusters, respectively, in Scenarios 1 and 2. In Scenario 3, marginal probabilities were constant across clusters, equaling 0.05 in 3.1 and 3.4 and 0.5 in 3.2 and 3.3.

The models in the last two scenarios are representative of the analyses we later carry out that use the percentages of patients with records of drug treatment to predict adequate assessment. Scenario 4 represents the logistic regression in which the proportion of patients having a record of treatment with aspirin is used as a predictor, whereas the fifth scenario is representative of using proportions from all three drug treatment records as covariates. We generated cluster sizes in the same manner as for the intercept-only model simulations. The percent with a record of aspirin treatment varied uniformly and independently from 66 to 96 in both scenarios, whereas we generated the percents of patients having a record of treatment with hypotensives or lipid-lowering drugs uniformly and independently on the set of integers ranging from 37 to 75 and 14 to 50, respectively, in the last scenario. In the first two settings of Scenarios 4 and 5, we set $\beta_0 = -1$ and all other parameter values were zero. In the last two settings, $\boldsymbol{\beta} = [-2, 0.015]^T$ in Scenario 4 and $\boldsymbol{\beta} = [-3, 0.01, 0.02, 0.04]^T$ in Scenario 5. Table II indicates the number of practices, N.

Although its structure was not misspecified, we allowed the exchangeable correlation value to vary from cluster to cluster in some settings because this will be accounted for, at least asymptotically, by $C_N(\hat{\boldsymbol{\beta}})$. In the first two scenarios, correlations were dependent on whether a cluster was in the control or intervention arm and were 0.05, 0.10, or 0.15. $exp(\varphi_1 + \varphi_2 n_i)/[1 + exp(\varphi_1 + \varphi_2 n_i)]$ gave the correlations in the third scenario, where we chose φ_1 and φ_2 such that $\rho_i, i = 1, 2, \ldots N$, were in the ranges (0.024, 0.079), (0.022, 0.337), (0.011, 0.119), and (0.008, 0.067) for the first through fourth settings, respectively. The correlation was fixed at 0.05 in the first two settings of Scenarios 4 and 5 but was equal to $\log[\rho_i/(1-\rho_i)] = -2.94 + 0.075(x_i - 81)$ and $\log[\rho_i/(1-\rho_i)] = -2.25 - 5|\pi_i - 0.5|$ in the last two settings of Scenarios 4 and 5, respectively. Here, x_i represents the practice level proportion of CHD patients having a record of aspirin treatment.

4.1.3. Description of results. Empirical variances dominated empirical MSEs, as squared bias was negligible, and we discuss MSE results in terms of precision. In the majority of settings, both QIF versions led to larger MSEs than GEE because of the impacts of imbalance in cluster sizes and the use of covariates. Even when the number of clusters was 100, which is larger than typically seen for CRTs, QIF were notably less precise than GEE in many settings.

In the first two scenarios, it is clear that QIF's estimation performance relative to GEE can worsen as N decreases because of a lack of empirical information. Similarly, for $N = 20$, the use of two additional covariates in Scenario 2 yielded notably smaller MSE ratios than in Scenario 1. However, this result was only slightly evident when $N = 100$, as there was much more empirical information to handle the complexity of the numerous combinations of size and covariates. Additionally, when $N = 20$, there was evidence that the distribution from which cluster sizes were generated had a slight impact on MSE ratios, although there appeared to be negligible impact for $N = 100$. The importance of this result is that at least in these settings with cluster-level covariates, the important factor is that any imbalance in the size of clusters influences QIF's estimation performance, whereas differences in how clusters vary in size may only play a minor role. Finally, except for settings 3, 4, 7, and 8 in Scenario 1, QIF's performance relative to GEE improved slightly when correlations were not equivalent for both trial arms.

Although GEE assumes a common correlation here, allowing the true correlation value to vary across clusters did not make QIF more reliable, except in three settings of the third scenario. This is an example where QIF may be advantageous over GEE. It appears, however, that this was only the case when marginal probabilities were not near zero or the number of clusters was large. The degree of dependency correlation had on cluster size also was influential to the relative performances of QIF and GEE here. Additionally, the differences in precisions of QIF and GEE shown in this scenario are small compared with the overall results from the other scenarios, deeming GEE as a more reliable method in the majority of scenarios we considered.

Both QIF versions performed approximately the same in the first three scenarios. However, in all four settings of Scenario 4, along with the settings consisting of 100 clusters in the last scenario, the QIF with estimating equations in the same class as GEE performed notably better than the regular QIF. This may imply that estimating equations class may influence estimation precision in some scenarios, whereas the empirical weighting employed by both QIF versions still is the major influence on the differences between these two methods and GEE. In all presented simulation results, we took the estimated correlation used for the QIF with estimating equations in the same class as GEE as the estimate for the common correlation from GEE. We also estimated correlation iteratively inside this QIF version in the same manner as GEE, which led to almost identical results.

4.2. *Shown via application to the motivating example*

We now return to our motivating dataset. By multiplying the size and adequate assessment percentage for a given practice and then rounding this quantity to the nearest integer, our dataset had a total of 629 adequately assessed patients, whereas Yudkin and Moher [6] reported 627. However, this slight difference is not notable in terms of the regression results, which are presented in Table III. We fit three models of the form $\text{logit}(\pi_{ij}) = \text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, in which x_i represents the proportion of patients within the ith practice who have a record of treatment with the corresponding drug type, and one model in which $\text{logit}(\pi_{ij}) = \text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$. x_{1i}, x_{2i}, and x_{3i} correspond to the percentages having a record of treatment with aspirin, hypotensives, and lipid-lowering drugs, respectively. π_i is the marginal probability of any given CHD patient in Practice i being adequately assessed.

We presented the bias-corrected standard errors, using the method of Mancl and DeRouen [17], in Table III for the GEE results. Another popular alternative that could also have been employed was the bias-corrected standard errors of Kauermann and Carroll [19]. See these manuscripts and Lu *et al.* [20] for more details. We also presented the typical empirical standard errors with QIF [4, 5] and similarly the QIF with estimating equations in the same class as GEE. In some preliminary simulations (not shown), we observed that QIF's standard errors were negatively biased; however, more assessment is needed. As the presented standard errors were possibly produced from a negatively biased procedure, we caution their use for obtaining confidence intervals and p-values in this particular example.

We first demonstrate the difference in weighting between GEE and both QIF versions by estimating the overall marginal probability of adequate assessment, corresponding to an intercept-only model. The estimated weights (marginal probabilities) used (produced) by GEE, QIF, and the newly defined QIF were $[1 + 0.058(n_i - 1)]^{-1}$ (0.276), 0.273-0.001$(n_i - 1)$ (0.276), and $2.510\hat{\gamma}_{1i} + 2.697(n_i - 1)\hat{\gamma}_{2i}$ (0.257), respectively, when using the logistic link. There is no difference between the probability estimates from GEE and QIF, although the newly defined QIF does give a slightly smaller value. Obtaining parameter estimates this close in value from these three methods appears to coincide with the simulation results presented earlier when the true marginal probability was 0.25. However, when the true marginal probability is closer to zero or the number of clusters is smaller, there is a greater chance in obtaining a sample from which probability estimates can be notably different across these three methods because of a larger variability in weighting used by both QIF versions.

The proportion of patients having a record of treatment with aspirin, hypotensives, or lipid-lowering drugs ranged from 66 to 96, 37 to 75, and 14 to 50, respectively. In each model, all three methods estimated the marginal probability of adequate assessment to be larger for practices having a greater proportion of patients with a record of drug treatments. Table III shows the range of estimated marginal probabilities over all practices used in the corresponding model. Results clearly show the difference between GEE and both QIF versions. The range in marginal probability estimates was always smaller for GEE than either QIF version, except when excluding Practice 21 from the analysis. Additionally, of the two QIF versions, parameter estimates from the version with estimating equations in the same class as GEE were notably closer to the estimates from GEE for the first two models. We especially saw this type of result in Scenario 4 of our simulation results, in which the newly defined QIF performed better than the regular QIF but not as precisely as GEE. This gives some indication that the GEE estimates here may be most reliable. Furthermore, when we used the proportion of patients with a record of aspirin or lipid-lowering drug treatments in the model, QIF estimated the strongest association between these covariates and adequate assessment.

We now take a closer look into the strength of the estimated marginal association between lipid-lowering drugs and adequate assessment. In one practice (Practice 1), only 14% of patients were adequately assessed, which is much smaller than any of the corresponding marginal probability estimates, given in the third model presented in Table III, from any of the three methods. This practice had the maximum proportion of patients with a record of being treated with lipid-lowering drugs and

Table III. Estimated logistic regression results when analyzing the cluster randomized trial dataset using the given record of drug treatment proportions as covariates inside the model.

Covariate(s)	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Min	Max
Aspirin	GEE	−2.050	0.014			0.246	0.332
		(0.960)	(0.013)				
	QIF	−4.025	0.039			0.192	0.435
		(1.331)	(0.018)				
	QIF in GEE Class	−2.365	0.016			0.217	0.311
		(1.204)	(0.016)				
Hypotensives	GEE	−1.793	0.015			0.226	0.343
		(0.812)	(0.015)				
	QIF	−2.431	0.025			0.183	0.370
		(1.079)	(0.020)				
	QIF in GEE Class	−2.158	0.019			0.191	0.331
		(0.967)	(0.018)				
Lipid-lowering drugs	GEE	−1.736	0.031			0.214	0.453
		(0.717)	(0.031)				
	QIF	−2.459	0.063			0.172	0.670
		(0.240)	(0.009)				
	QIF in GEE Class	−2.535	0.063			0.161	0.652
		(0.209)	(0.007)				
All three	GEE	−2.845	0.006	0.014	0.028	0.193	0.401
		(1.104)	(0.012)	(0.016)	(0.031)		
	QIF	−5.884	0.021	0.035	0.052	0.117	0.515
		(1.688)	(0.018)	(0.015)	(0.010)		
	QIF in GEE Class	−3.746	0.004	0.019	0.057	0.139	0.555
		(1.137)	(0.012)	(0.010)	(0.007)		
Results after removing Practice 1							
Lipid-lowering drugs	GEE	−2.409	0.061			0.173	0.518
		(0.241)	(0.009)				
	QIF	−2.462	0.062			0.170	0.525
		(0.224)	(0.008)				
	QIF in GEE Class	−2.520	0.063			0.163	0.519
		(0.198)	(0.007)				
Results after removing Practice 21							
Aspirin	GEE	−2.247	0.016			0.235	0.334
		(0.970)	(0.013)				
	QIF	−2.029	0.011			0.219	0.283
		(1.121)	(0.014)				
	QIF in GEE Class	−1.920	0.010			0.222	0.278
		(1.015)	(0.013)				

We give the minimum and maximum predicted marginal probabilities from each model and method. We give the standard error estimates in parentheses below their corresponding parameter estimates.
QIF, quadratic inference function; GEE, generalized estimating equations.

therefore had the largest marginal probability estimate. The first plot in Figure 1 clearly shows the difference in this practice's observed and estimated probabilities. GEE does not directly take into account how far the observed proportion of adequately assessed patients is from the marginal mean, and therefore, the estimated association between adequate assessment and lipid-lowering drugs is not as strong as it would be without using data from this practice. QIF, however, does directly take into account the large empirical variability and downweights this practice's outcomes, allowing the estimated association to be stronger. Explicitly, if we were to estimate these models without the first practice, QIF and GEE

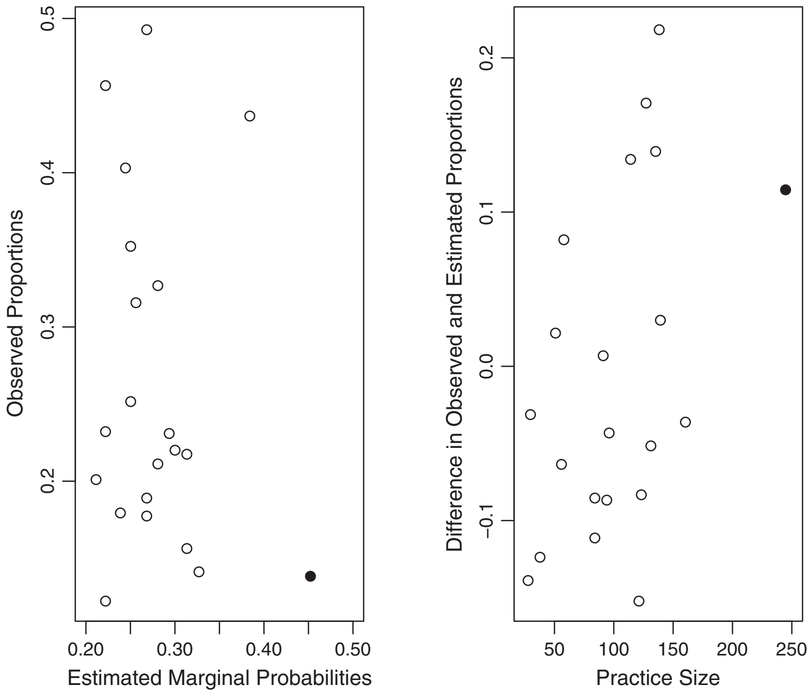

Figure 1. Estimated marginal probabilities in the left plot are from using GEE to estimate the model in which the proportion of patients having a record of treatment with lipid-lowering drugs is the only covariate. The bold dot corresponds to Practice 1. In the right plot, estimated marginal probabilities used to obtain differences are from using GEE and the model in which the proportion having a record of treatment with aspirin is the only covariate. The bold dot corresponds to Practice 21.

would produce $\hat{\boldsymbol{\beta}} = [-2.462, 0.062]$ and $\hat{\boldsymbol{\beta}} = [-2.409, 0.061]$, respectively. These estimates are only slightly different than the estimates given from QIF when including this practice, distinctly showing the robustness property of QIF.

In the model in which the proportion of patients having a record of treatment with aspirin is the only covariate, the difference in estimates from GEE and QIF were not due to an outlier. Rather, the sensitivity, with respect to empirical covariances, of the estimated weights implemented inside QIF's estimating equations led to the notable differences. For instance, the largest practice (Practice 21) had a notable influence on the estimates produced by QIF, but not GEE. Table III presents the differences in parameter estimates before and after removing this practice. The second plot in Figure 1 shows that the overall empirical variation increases with practice size, excluding the largest practice, in our sample. This one practice actually brings down the averaged covariance trend with respect to size, estimated via $\nabla \bar{g}_N^T(\hat{\boldsymbol{\beta}}) C_N^{-1}(\hat{\boldsymbol{\beta}})$, as it is notably bigger than the other large practices, but its empirical variation is only moderate relative to these same practices. Therefore, when we remove Practice 21, larger practices receive less weight, whereas the weight given to smaller clusters increases, because of a rise in the averaged empirical covariances in larger practices.

Specifically, Equation (7) reduces to

$$\sum_{i=1}^{21} \begin{bmatrix} 0.004 + 0.109x_i + 0.243(n_i - 1) - 0.004(n_i - 1)x_i \\ 0.245 + 7.917x_i + 15.030(n_i - 1) - 0.233(n_i - 1)x_i \end{bmatrix} (Y_i - n_i \pi_i) = \mathbf{0}$$

when using all practices and becomes

$$\sum_{i=1}^{20} \begin{bmatrix} 0.005 + 0.257x_i + 0.183(n_i - 1) - 0.004(n_i - 1)x_i \\ 0.354 + 21.030x_i + 10.274(n_i - 1) - 0.279(n_i - 1)x_i \end{bmatrix} (Y_i - n_i \pi_i) = \mathbf{0}$$

when deleting Practice 21. By plugging in values for x_i and size from the dataset, it can be seen that the proportion of weight given to smaller (larger) clusters typically increased (decreased) after removing this practice. For instance, Practice 1 (20) consisted of 28 (160) CHD patients, and 79% (67%) of patients had a record of aspirin treatment. The weight matrix given to the residual from the first practice increased from $[7.26, 534.51]^T$ to $[16.53, 1344.01]^T$ after removing Practice 21, whereas the weight matrix corresponding to Practice 20 decreased from $[6.53, 438.31]^T$ to $[2.65, 70.74]^T$. In comparison, GEE is given as

$$\sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \frac{Y_i - n_i \pi_i}{1 + (n_i - 1)\hat{\rho}} = \mathbf{0}$$

for this example, in which $\hat{\rho}$ is estimated using the empirical variabilities from all practices, and the influence via $\hat{\rho}$ from the empirical variability of one cluster on the weights used by GEE is very minor. Before (after) removing Practice 21, ρ was estimated to be 0.055 (0.057), giving weights that were only negligibly affected.

The GEE estimates do not differ by a large amount when including or excluding Practice 21, and so it appears that GEE produced more reliable estimates. The estimates produced by QIF were largely influenced by the empirical variabilities in larger clusters, especially Practice 21, showing that this method can be sensitive in settings consisting of a small number of clusters due to the variability in $C_N(\hat{\boldsymbol{\beta}})$. We note that if more practices were included in this study, the average of the empirical covariances would lessen the influence from a single cluster on weights used inside QIF's estimating equations, making them more reliable. Additionally, although we do not know the true covariances for outcomes in this dataset, implying we cannot say for sure that GEE produced more appropriate estimates than QIF, we do see here that QIF can be sensitive even to a single cluster's empirical covariance, which is the type of scenario in which QIF can be less reliable than GEE.

When using all three covariates in the same model, there were notable differences in parameter estimates across the three methods. Both QIF versions were similar in terms of their predicted ranges of marginal probabilities, which were approximately twice as wide as the range given from GEE. As with the two previously discussed models, the influences from Practices 1 and 21 were the major factors in the differences between GEE and both QIF versions. Taking into account how these practices influenced QIF in these two models, in addition to the simulation results presented in Scenario 5, it is likely that the GEE estimates are more reliable here. These estimates show a notable association between adequate assessment and the three covariates but not nearly as strong of a relationship estimated by either QIF version due to the influence of only two practices.

5. Concluding remarks

The QIF method has the theoretical advantage of producing regression estimates with equal or greater efficiency than GEE [2]. We have given details and evidence that the class of estimating equations realistically has less to do with differences in estimation precision between QIF and GEE than the empirical nature of $C_N(\hat{\boldsymbol{\beta}})$ and how this matrix is used to weight outcomes inside QIF's estimating equations. In small to moderately sized samples, and for unequal cluster sizes, and models with only cluster-level covariates, as is common with CRTs, QIF can produce estimates with lower precision than GEE, even when the incorrect covariance structure is implemented. We also showed via our motivating dataset that the weights QIF implements in its estimating equations can be sensitive to the empirical variability of even a single cluster, as the number of practices was small.

Although we focused on CRTs scenarios with binary outcomes, we also performed simulations (to appear in future work) in which outcomes were continuous or counts. We studied as well a variety of general repeated measures scenarios with time-dependent and cluster-level covariates. When the number of independent clusters was not large, simulations showed that QIF did not necessarily perform as well as GEE. We observed these results for a variety of true and working correlation structures.

This paper focused on an exchangeable structure, as unstructured typically is used with a balanced design and QIF and GEE are equivalent estimation procedures under an independence assumption. Although less common when clusters vary in size, especially in CRT scenarios, an AR-1 structure would be appropriate, for example, in a setting resembling administrative censoring in which patients contribute data at the same distinct time points that are equally spaced but with the allowance that they can drop out of the study any time before their final scheduled visit. The inverse of an AR-1 structure is the weighted sum of three basis matrices, the last of which is usually not implemented with QIF as it contains little information [2,5]. The corresponding estimating equations from QIF are in the same class as GEE whether clusters vary in size or not, as γ_{ri}, $r = 1, 2, 3$; $i = 1, 2, \dots N$, are not functions of size [2]. However, just as we showed in this paper that the variable empirical nature of $C_N(\hat{\boldsymbol{\beta}})$ can lead to a loss in estimation reliability for both QIF versions, relative to GEE, when using an exchangeable structure, the same result can occur for the AR-1 structure as well.

Further research is required to improve the estimation performance of QIF in small to moderately sized samples, as well as to address the possibility of negatively biased standard error estimates. However, if estimation precision and standard errors are not concerns when deciding between QIF and GEE for data analysis, QIF has distinct advantages. For example, the QIF can itself be used as a statistic in goodness-of-fit and likelihood ratio score tests [2,5]. Although we do not suggest a numerical value, as we have argued that the impact of empirical information used to estimate weights given to any cluster's outcomes depends on the setting, QIF may be a better method to employ when N is arbitrarily large. This is particularly true when the actual covariance structure is believed to possibly deviate largely from the chosen working structure. In this situation, $C_N(\hat{\boldsymbol{\beta}})$ may have greater accuracy in modeling the entire true covariance structure on average, potentially leading to an improved estimation performance, in addition to the ability to make use of QIF's other advantages.

Acknowledgements

We would like to thank the anonymous associate editor and reviewers for their constructive comments that helped improve this paper.

References

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
2. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**:823–836.
3. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica* 1982; **50**:1029–1054.
4. Song PXK. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer: New York, 2007.
5. Song PXK, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine* 2009; **28**:3683–3696.
6. Yudkin PL, Moher M. Putting theory into practice: a cluster randomized trial with a small number of clusters. *Statistics in Medicine* 2001; **20**:341–349.
7. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized tirals. *Statistics in Medicine* 2007; **26**:3415–3428.
8. Bland JM, Kerry SM. Statistics notes: weighted comparison of means. *British Medical Journal* 1998; **316**:129.
9. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000.
10. Campbell MK, Elbourne DR, Altman DG, for the CONSORT Group. CONSORT statement: extension to cluster randomised trials. *British Medical Journal* 2004; **328**:702–708.
11. Small CG, McLeish DL. *Hilbert Space Methods in Probability and Statistical Inference*. Wiley: New York, 1994.
12. Qu A, Song PXK. Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika* 2004; **91**:447–459.
13. Han P, Song PXK. A note on improving quadratic inference functions using a linear shrinkage approach. *Statistics and Probability Letters* 2011; **81**:438–445.
14. Lindsay B. Conditional score functions: some optimality results. *Biometrika* 1982; **69**:503–512.
15. Windmeijer F. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 2005; **126**:25–51.
16. Loader C, Pilla RS. Iteratively reweighted generalized least squares for estimation and testing with correlated data: an inference function framework. *Journal of Computational and Graphical Statistics* 2007; **16**:925–945.
17. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
18. Ridout MS, Demetrio CGB, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; **55**:137– 148.
19. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; **96**:1387–1396.
20. Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, Wolfson M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* 2007; **63**:935–941.