Biometrika (2001), **88**, 4, *pp.* 1121–1134

Printed in Great Britain

Within-cluster resampling

By ELAINE B. HOFFMAN
7 Janson Court, Westport, Connecticut 06880, U.S.A.
meahoffman@aol.com

PRANAB K. SEN
Department of Biostatistics, CB #7400, University of North Carolina, Chapel Hill,
North Carolina 27599-7400, U.S.A.
pksen@bios.unc.edu

AND CLARICE R. WEINBERG
National Institute of Environmental Health Sciences, P.O. Box 12233, MD A3-03,
Research Triangle Park, North Carolina 27709, U.S.A.
weinberg@niehs.nih.gov

Summary

Within-cluster resampling is proposed as a new method for analysing clustered data. Although the focus of this paper is clustered binary data, the within-cluster resampling asymptotic theory is general for many types of clustered data. Within-cluster resampling is a simple but computationally intensive estimation method. Its main advantage over other marginal analysis methods, such as generalised estimating equations (Liang & Zeger, 1986; Zeger & Liang, 1986) is that it remains valid when the risk for the outcome of interest is related to the cluster size, which we term nonignorable cluster size. We present theory for the asymptotic normality and provide a consistent variance estimator for the within-cluster resampling estimator. Simulations and an example are developed that assess the finite-sample behaviour of the new method and show that when both methods are valid its performance is similar to that of generalised estimating equations.

Some key words: Clustered binary data; Generalised estimating equations; Generalised linear model; Marginal model; Nonignorable cluster size; Resampling; Within-cluster correlation.

1. Introduction

Correlated binary outcomes arise in many research settings: in a clinical study, a particular outcome may be assessed at repeated visits; in reproductive toxicology, birth defect status may be observed for the pups in each litter. It is well known that methods of analysis that treat such positively correlated outcomes as independent will tend to produce standard errors that are too small and hypothesis tests that have an inflated Type I error rate.

A wide range of statistical techniques have been developed to handle dependent binary outcomes, some of which estimate subject-specific measures of effect, e.g. random-effects models (Laird & Ware, 1982), and some of which estimate marginal, population-averaged measures of effect (Pendergast et al., 1996). Marginal approaches based on generalised estimating equations have recently become widely used (Liang & Zeger, 1986; Zeger & Liang, 1986). Their popularity can be attributed in part to their robustness against misspecification of the correlation structure and to their relative ease of use.

Despite the power, robustness and flexibility of the generalised estimating equations method, there are some limitations to its use. First, while the method is asymptotically robust against misspecification of the correlation structure, for a finite dataset quite different results can arise with different correlation specifications (Mancl & Leroux, 1996). The data analyst may have little a priori or data-based justification for preferring one correlation structure over another, and thus this choice can in practice be both important and subjective. Secondly, the generalised estimating equations method implicitly presumes that the size of the cluster is unrelated to the parameters under study, i.e. is ignorable, and in many applications this may not be true. Consider the above examples. In the clinical setting, the number of visits to the clinic may depend in part on how sick the study subject perceives himself to be. In the toxicology setting, the dams that are particularly susceptible to effects of a toxicant may produce more pups with birth defects and may simultaneously experience more foetal resorptions, thereby reducing the corresponding litter size. This problem, which we term 'nonignorable cluster size', can be thought of as closely related to violation of the assumption (Liang & Zeger, 1986; Zeger & Liang, 1986) that observations missing from a cluster must be missing completely at random.

The purpose of this paper is to describe an alternative marginal approach, 'within-cluster resampling'. We first describe the method and discuss the subtle difference in the marginal parameter it estimates, as compared to that estimated with generalised estimating equations. The method is computer intensive, but yields consistent and asymptotically normal estimators of cluster-based marginal parameters, and we provide a consistent estimator for the variance. Within-cluster resampling offers two important advantages over generalised estimating equations for the analysis of clustered binary data. First, within-cluster resampling handles the correlation structure in a fully implicit way, so that the analyst does not need to specify a correlation structure in applying the method. Secondly, within-cluster resampling remains valid in the presence of nonignorable cluster size, whereas the generalised estimating equations method does not. For scenarios where both approaches would be valid, we present results of simulations in order to compare their operating characteristics.

Section 2 describes the resampling and analysis strategy of within-cluster resampling. Section 3 provides the theory for the asymptotic normality of the estimator and the consistent variance estimator. In § 4, methods and results of simulations are detailed. In the simulations the exposure assignment is either cluster-specific, the same for all units in each cluster, or observation-specific. We also provide results based on simulating data where cluster size is nonignorable, revealing that the generalised estimating equations approach can yield statistically incompatible parameter estimates under two different specifications of the correlation structure. This serves as a counterexample to the usual consistency theory for generalised estimating equations, in the context of nonignorable cluster size. A periodontal study of gingivitis is analysed in § 5. Finally, we discuss strengths and limitations of within-cluster resampling in § 6.

2. Within-cluster resampling

We randomly sample one observation from each of the I clusters, with replacement. The resampled dataset of size I can then be analysed by a generalised linear model, such as logistic regression, since the I observations are independent. This process is repeated a large number of times, Q, say, where each of the Q analyses provides a consistent estimator of the parameter of interest. The within-cluster resampling estimator is constructed as the average of the Q resample-based estimates. Since the resampled datasets contain correlated and overlapping observations, the Q estimates based on resampled datasets are dependent. However, this dependency can be implicitly taken into account by the variance formula. Let $\hat{\beta}(R; q)$ denote the resample-based estimate for the qth resampled dataset, $q = 1, \ldots, Q$. As will be shown, the within-cluster resampling estimator,

$$\bar{\beta} = Q^{-1} \sum_{q=1}^{Q} \hat{\beta}(R; q),$$

is asymptotically normal as $I \rightarrow \infty : I^{\frac{1}{2}}(\bar{\beta} - \beta) \sim N_p(0, \Sigma)$, where Σ is a finite and positive-definite matrix. Let y_{ij} denote the binary outcome and let x_{ij} denote the covariate vector for the (i, j)th observation, with $i = 1, \ldots, I, j = 1, \ldots, n_i$.

Within-cluster resampling is a natural method for analysing data with nonignorable cluster size, because of the sampling scheme. We distinguish sampling where one selects a random observation from a randomly-sampled cluster, 'cluster-based sampling', from the unit-based sampling that is implicit in the generalised estimating equations approach. All clusters are given equal weight in a within-cluster resampling analysis, so the marginal parameter will have a cluster-based interpretation. This is in direct contrast to generalised estimating equations, where large clusters are weighted more than small clusters. This difference in relative weighting does not affect the asymptotic parameter except in scenarios where cluster size is nonignorable.

Nonignorable cluster size can be defined as any violation of the property that $E(Y_{ij}|n_i, X_{ij}) = E(Y_{ij}|X_{ij})$. Nonignorability occurs when the size of the cluster is related to the risk for the outcome of interest. Suppose smaller clusters are at greater baseline risk, such as was described in the toxicological example. In generalised estimating equations, litters with more pups are weighted more than litters with fewer pups. If the baseline risk for birth defect in a pup is negatively related to the litter size, then any analysis that weights larger litters more, such as generalised estimating equations, will estimate the risk for birth defects in pups as relatively low, compared to a method that weights each litter equally. With within-cluster resampling, because sampling is cluster-based, larger litters are given the same weight as smaller litters because each resampled-based analysis uses a single observation to represent each cluster. This sampling structure explains why effects of nonignorable cluster sizes are eliminated in a within-cluster resampling analysis.

To explore implications of nonignorable cluster size further, in the context of the toxicological example, consider a particular endpoint, such as a birth defect or birth weight. Many embryos that are abnormal are maternally resorbed and are thus not born to be observed. Litters where pups share a high baseline propensity for a bad outcome will thus tend to be smaller than average. For simplicity, let T denote the event that a pup is toxicant-exposed and assume an additive random effects model for the observed litters, so that $E(Y_{ij}|T) = \alpha_i + \beta$ and $E(Y_{ij}|\bar{T}) = \alpha_i$ for unexposed pups. Suppose that the exposure is via the dam, so that all pups in a given litter share the same exposure status. With pup-based sampling, the marginal $E(Y_{ij}|T)$ will be weighted by the litter size, n, as follows:

$$E(Y_{ij}|T) = E_\alpha\{E(Y_{ij}|T, \alpha)\} = \int \left(\frac{(\alpha + \beta)E(n|\alpha, T)}{\int E(n|\alpha, T)\, d\alpha} \right) d\alpha.$$

This is equivalent to

$$E(Y_{ij}\,|\,T) = E_\alpha\{E(Y_{ij}\,|\,T, \alpha)\} = \int \left(\frac{\alpha E(n\,|\,\alpha, T)}{\int E(n\,|\,\alpha, T)\, d\alpha} \right) d\alpha + \beta.$$

For unexposed litters, the marginal expected outcome is given by

$$E(Y_{ij}\,|\,\bar{T}) = E_\alpha\{E(Y_{ij}\,|\,\bar{T}, \alpha)\} = \int \left(\frac{\alpha E(n\,|\,\alpha, \bar{T})}{\int E(n\,|\,\alpha, \bar{T})\, d\alpha} \right) d\alpha.$$

With nonignorable cluster size, these marginal expected outcomes need not differ by β. Now consider cluster-based sampling. The expected outcome for a randomly sampled pup from a randomly sampled exposed litter will be $E(\alpha_i + \beta)$, where the expectation is now taken across the distribution of α in an unweighted way. Cluster-based sampling thus leads to estimation of β, as in a random-effects analysis. This demonstrates that, even under a linear model, analyses relying on unit-based sampling will not in general converge to the same marginal parameters as will analyses relying on cluster-based sampling.

It may be instructive to compare the interpretation of the cluster-based within-cluster resampling marginal parameter and the observation-based generalised estimating equations marginal parameter. For a logistic model, the traditional interpretation of the odds-ratio is the ratio of odds for the marginal probability of the outcome of interest with and without the exposure. The difference in interpretation between a within-cluster resampling parameter and a generalised estimating equations parameter is subtle: the within-cluster resampling parameter reflects the one-per-cluster sampling scheme of the procedure. The interpretation of the within-cluster resampling parameter estimate based on a logistic model would be the difference in log odds for the marginal probability of the outcome of interest between a randomly sampled exposed observation from a randomly selected cluster and a randomly sampled unexposed observation from a randomly selected cluster. If cluster size is unrelated to the outcome, the two interpretations coincide. If cluster size is related to the outcome, the generalised estimating equations approach is not appropriate, as will be shown. In this sense within-cluster resampling offers improved robustness.

The cluster-based parameter may often be more generalisable than an observation-based parameter. An example may help. Suppose the cluster of interest is a set of pregnancy outcomes for each of a random sample of women. The cluster size will be related to risk, because women at high risk of spontaneous abortion need to have more pregnancies on average to achieve their desired family size. Thus cluster size is nonignorable. In this example, the cluster-based interpretation corresponds to risk for a randomly-sampled woman. These cluster-based parameters should therefore be commensurate across populations or ethnic groups, even if couples in these groups have very different desired family sizes. By contrast, the observation-based risk, if we ignore covariates and nonlinearity of any model, will be systematically higher in a population with larger desired family size, because high-risk women will keep trying and will tend to contribute large numbers of pregnancy outcomes.

3. Asymptotic normality and consistent variance of the within-cluster resampling estimator

We assume a generalised linear model for data arising under cluster-based sampling. Thus $h\{E(Y|X = x)\} = \beta^{\mathrm{T}}x$, where h is a monotone and differentiable link function and x

is a vector of covariates. For each $q = 1, \ldots, Q$, $\hat\beta(R; q)$ is a realisation of the standard maximum likelihood estimator for a generalised linear model based on I independent observations, under the qth iteration of the within-cluster resampling sampling scheme. We assume sufficient regularity conditions (Sen & Singer 1993, pp. 111–8) to ensure that $I^{\frac{1}{2}}\{\hat\beta(R; q) - \beta\} \sim N\{0, J^{-1}(\beta; q)\}$ in distribution as $I \to \infty$, where $J(\beta; q)$ is the expected information matrix (McCullagh & Nelder, 1989, pp. 324–8; Fahrmeir & Kaufmann, 1985). Further details are given in E. B. Hoffman's unpublished 1998 Ph.D. thesis from the University of North Carolina.

The within-cluster resampling estimator is the average of identically-distributed but dependent maximum likelihood estimates and the Central Limit Theorem is not directly applicable. Asymptotic normality of the within-cluster resampling estimator, $\bar\beta$, is established by rewriting the average of the Q resample-based score statistics as the sum of independent cluster-specific pieces so that a Central Limit Theorem can be applied.

THEOREM 1. *Let the within-cluster resampling estimator be defined as*

$$\bar\beta = Q^{-1} \sum_{q=1}^{Q} \hat\beta(R; q)$$

for Q very large. As the number of clusters $I \to \infty$, under the usual regularity conditions, $I^{\frac{1}{2}}(\bar\beta - \beta) \to N(0, \Sigma)$ in distribution, where Σ is finite and positive-definite.

The proof of Theorem 1 can be found in Appendix 1.

A consistent variance estimator is provided in Theorem 2.

THEOREM 2. *Let I and $\bar\beta$ be defined as in Theorem 1. Define*

$$\hat\Sigma = \hat{\text{var}}\{I^{\frac{1}{2}}(\bar\beta - \beta)\} \doteq I\left\{Q^{-1} \sum_{q=1}^{Q} \hat\Sigma(R; q) - (Q-1)Q^{-1}S_{\hat\beta}^2\right\},$$

where $\hat\Sigma(R; q)$ is the estimated covariance matrix from the qth analysis and

$$S_{\hat\beta}^2 = (Q-1)^{-1} \sum_{q=1}^{Q} \{\hat\beta(R; q) - \bar\beta\}\{\hat\beta(R; q) - \bar\beta\}'$$

is the estimated covariance matrix among the Q resample-based estimates $\hat\beta(R; q)$. Then $\hat\Sigma$ is consistent for $\Sigma = \text{var}\{I^{\frac{1}{2}}(\bar\beta - \beta)\}$.

The proof of Theorem 2 can be found in Appendix 2.

4. SIMULATIONS

To characterise the behaviour of within-cluster resampling, we simulated clustered binary data. For each simulated cluster, a baseline risk was sampled from a $\text{Be}(a, b)$ distribution. Individuals were independently designated as 'exposed' or 'unexposed', each with probability $\frac{1}{2}$.

The number of resamples, Q, required depends on the scenario simulated. As the number of possible distinct resampled datasets increases, more resamples may be required to achieve stable parameter and variance estimates. To ensure that each analysis used enough resamples a large value, 10 000 or 50 000, was used. Fewer resamples may be needed with a particular dataset, where one can apply empirical criteria for stability of the estimates.

The simulated datasets all had varying cluster sizes, obtained from a Bi(8, 0·75) distribution, truncated by discarding clusters of assigned size zero.

Scenarios with nonignorable cluster size were simulated by imposing a negative relationship between the size of the cluster and the baseline risk as follows: clusters with a sampled baseline risk lower than the Be(a, b) mean had cluster size taken from a truncated Bi$(9, 0{\cdot}75)$, excluding the 0's and 9's; clusters with a sampled baseline risk higher than the Be(a, b) mean had their cluster size taken from a truncated Bi$(9, 0{\cdot}25)$, excluding the 0's and 9's. Under such a scenario a randomly selected unexposed observation, unit-based sampling, has risk lower than the Be(a, b) mean of $a/(a+b)$, while under cluster-based sampling the corresponding risk is equal to the beta mean. By contrast, if cluster size had been ignorable, a randomly selected unexposed observation would have risk equal to the Be(a, b) mean, whether the sampling is unit-based, as in generalised estimating equations, or cluster-based, as in within-cluster resampling.

We additionally assumed that exposed observations have an increased risk for the outcome of interest, corresponding to the addition of a fixed effect on the logit scale. Under our simulation set-up, each resampled dataset yields a 2×2 table. Particularly with small I, there is a non-negligible probability of a zero cell, producing infinite parameter and variance estimates. To avoid this problem, only the finite portion of the resampling distribution is used for parameter estimation and variance calculations. The usual maximum-likelihood-based large-sample approximations were used to estimate the parameters and variances from the 2×2 tables (Stokes et al., 1995).

Since the estimated variance is based on subtraction, it is theoretically possible for the matrix to fail to be positive-definite. In our experience, however, this was exceedingly rare, occurring once in 61 000 simulated datasets. Such an aberration should suggest to the investigator that the number of resamples was insufficient or that the sample size I was insufficient for the asymptotic approximation to hold.

The generalised estimating equations approach was also applied to the simulated datasets for comparison. In that approach, reflecting the actual data generation, the exchangeable correlation structure was used for all scenarios. In addition, the independence correlation structure was used for comparison under the nonignorable cluster-size scenario. Within-cluster resampling does not require the user to specify a correlation structure.

We will show by a simulated scenario that when cluster size is nonignorable the behaviour of the generalised estimating equations approach breaks down. Since it is therefore not a valid altemative to within-cluster resampling, the generalised estimating equations approach was not applied in subsequent simulations with nonignorable cluster size.

Tables 1–4 report the following results from the simulations: the average across simulations of the parameter estimates,

$$\bar{b}_0 = \frac{\sum_{s=1}^{S} \hat{b}_0(s)}{S}, \quad \bar{b}_1 = \frac{\sum_{s=1}^{S} \hat{b}_1(s)}{S},$$

where $s = 1, \ldots, S = 1000$ indexes the simulation; empirical standard errors,

$$\text{ESE}(\bar{b}_0) = \left[\frac{\sum_{s=1}^{S} \{\hat{b}_0(s) - \bar{b}_0\}^2}{S(S-1)}\right]^{\frac{1}{2}}, \quad \text{ESE}(\bar{b}_1) = \left[\frac{\sum_{s=1}^{S} \{\hat{b}_1(s) - \bar{b}_1\}^2}{S(S-1)}\right]^{\frac{1}{2}};$$

average estimated standard errors for the parameter estimates,

$$\text{avg}\{\text{SE}(\hat{b}_0)\} = \frac{\sum_{s=1}^{S} [\text{var}\{\hat{b}_0(s)\}]^{\frac{1}{2}}}{S}, \quad \text{avg}\{\text{SE}(\hat{b}_1)\} = \frac{\sum_{s=1}^{S} [\text{var}\{\hat{b}_1(s)\}]^{\frac{1}{2}}}{S};$$

empirical coverage of nominal 95% confidence intervals for the exposure parameter; the empirical power or, under the null, the empirical alpha; a paired power comparison for within-cluster resampling versus generalised estimating equations; and the number of zero-celled resampled datasets across the simulated datasets. The paired power comparison reports the number of discordant pairs in favour of within-cluster resampling or generalised estimating equations, where we count the number of rejections based on a two-sided 0·05-level test of $b_1 = 0$, and apply McNemar's test (Stokes et al., 1995). We used C programs to simulate the data and to perform the within-cluster resampling analyses. The robust variance estimates were used in the generalised estimating equations approach.

We first considered a null exposure effect scenario with cluster size $n_i \leqslant 8$ for all i ($i = 1, \ldots, I = 250$) and $S = 1000$. We varied the within-cluster correlation, using $\rho = 0{\cdot}0$ and $\rho = 0{\cdot}2$. Subsequent simulated datasets had a positive exposure effect with $\rho = 0{\cdot}2$, and included both cluster-specific and observation-specific exposures with ignorable cluster size data. The number of clusters varied from 50 to 250 for the scenarios with cluster-specific exposures, and was 250 for the scenarios with observation-specific exposure. The number of resamples was increased from 10 000 in the cluster-specific exposure simulations to 50 000 for the scenarios with observation-specific exposure.

Table 1. *Results for* 1000 *simulations per scenario of a null exposure effect scenario, with cluster-specific exposures,* $n_i \leqslant 8$, $Q = 10\,000$, $I = 250$, $b_0 = -1{\cdot}0986$, $b_1 = 0{\cdot}0$ *and* $\rho = 0$ *or* 0·2

	WCR	GEE	WCR	GEE
ρ	0·0	0·0	0·2	0·2
$\bar{b}_0$	−1·1066	−1·1007	−1·1051	−1·0988
ESE$(\bar{b}_0)$	0·0036	0·0033	0·0044	0·0043
$\bar{b}_1$	−0·0068	−0·0042	−0·0056	−0·0026
ESE$(\bar{b}_1)$	0·0051	0·0046	0·0062	0·0060
avg{SE$(\hat{b}_0)$}	0·1112	0·1028	0·1363	0·1343
avg{SE$(\hat{b}_1)$}	0·1575	0·1459	0·1929	0·1900
Coverage	0·945	0·952	0·945	0·942
Empirical alpha	0·055	0·048	0·055	0·058
Discordant pairs favouring the above method	20	13	19	22

WCR, within-cluster resampling; GEE, generalised estimating equations; ESE, empirical standard error, i.e. standard error among estimates derived from simulations; avg(SE), average among estimated standard errors.

Both within-cluster resampling and the generalised estimating equations method performed well in the null case, see Table 1, with good coverage and Type I error rates consistent with 0·05. The presence of correlated outcomes increased variances slightly but did not otherwise impair the performance of either procedure.

For scenarios with a positive exposure effect, see Table 2, within-cluster resampling was slightly biased when the number of clusters was small, $I = 50$ and $I = 100$. As the number of clusters increased to 250, the bias disappeared for both the scenarios with cluster-specific and with observation-specific exposure. Within-cluster resampling coverage and power were comparable to that for the generalised estimating equations method, except for the within-cluster resampling simulation case of $I = 50$ in Table 2, where the power was significantly better for within-cluster resampling. Another exception was the observation-specific case, where the generalised estimating equations approach outperformed

Table 2. *Results from* 1000 *simulations per scenario, with a positive exposure effect,* $n_i \leqslant 8$, $\rho = 0{\cdot}2$, $b_0 = -1{\cdot}0986$ *and* $b_1 = 0{\cdot}4796$

	WCR	GEE	WCR	GEE	WCR	GEE	WCR	GEE
No. of resamples (Q)	10 000		10 000		10 000		50 000	
No. of clusters (I)	50	50	100	100	250	250	250	250
Covariate type	CS	CS	CS	CS	CS	CS	OS	OS
$\bar{b}_0$	−1·1642	−1·11711	−1·1256	−1·1122	−1·1105	−1·1058	−1·1106	−1·1031
ESE$(\bar{b}_0)$	0·0101	0·0101	0·0073	0·0071	0·0045	0·0044	0·0041	0·0039
$\bar{b}_1$	0·5158	0·4849	0·5018	0·4944	0·4807	0·4776	0·4829	0·4807
ESE$(\bar{b}_1)$	0·0147	0·0139	0·0093	0·0091	0·0059	0·0059	0·0049	0·0044
avg{SE$(\hat{b}_0)$}	0·3055	0·2988	0·2140	0·2117	0·1362	0·1342	0·1244	0·1168
avg{SE$(\hat{b}_1)$}	0·4193	0·4125	0·2942	0·2907	0·1865	0·1843	0·1511	0·1327
Coverage	0·933†	0·940	0·947	0·954	0·949	0·950	0·944	0·942
Power	0·259	0·239	0·387	0·389	0·732	0·741	0·882	0·944
Discordant pairs favouring above method	62*	42	60	62	49	58	10	72*
No. of zero cells out of 1000Q	13 847		12		0		0	

†More than 2 standard errors from nominal value, 0·95. *Significant McNemar's test at 0·05.

WCR, within-cluster resampling; GEE, generalised estimating equations; ESE, empirical standard error, i.e. standard error among estimates derived from simulations; avg(SE), average among estimated standard errors; CS, cluster-specific exposures; OS, observation-specific exposures.

within-cluster resampling. The total number of resampled datasets with zero cells is recorded in Table 2, the denominator being $1000Q$.

The current literature on marginal methods for the analysis of clustered binary data has not addressed issues related to nonignorable cluster size. For example, Liang & Zeger (1986) assume a common cluster size and comment that missing data, yielding smaller clusters, must be missing completely at random, i.e. ignorable in our sense. Therefore, the presence of nonignorable cluster size violates an implicit assumption of the generalised estimating equations approach.

Table 3 demonstrates a nonignorable cluster size scenario with observation-specific covariates, $\rho = 0{\cdot}2$, $n_i \leqslant 8$, $I = 500$ and $S = 2000$. The generalised estimating equations approach, with two different correlation structures specified, produces statistically incompatible estimates, despite the large sample size. This is in contrast to the theoretical robustness of the method to choice of the correlation structure, a choice which should at worst affect the efficiency of generalised estimating equations estimation (Liang & Zeger, 1986; Zeger & Liang, 1986). The intercept parameter estimates are dramatically inconsistent with each other. The two exposure coefficients are also statistically different from each other. The variance estimation also misbehaved: with an exchangeable working structure, the empirical variances exceeded the average estimated variances by more than 10%. Under within-cluster resampling, the empirical variances were nearly identical to the average estimated variance.

Table 3. *Results from* 2000 *simulations showing inconsistency of generalised estimating equations with nonignorable cluster sizes; observation-specific exposures,* $n_i \leqslant 8$, $Q = 50\,000$, $I = 500$, $\rho = 0{\cdot}2$, $b_0 = -1{\cdot}0986$ *and* $b_1 = 0{\cdot}4796$

	WCR	GEE(i)	GEE(e)
$\bar{b}_0$	−1·1052	−1·5047	−1·3634
ESE$(\bar{b}_0)$	0·0022	0·0019	0·0020
$\bar{b}_1$	0·4848	0·4991	0·4796
ESE$(\bar{b}_1)$	0·0027	0·0022	0·0021

ESE, empirical standard error, i.e. standard error among estimates derived from simulations; WCR, within-cluster resampling; GEE, generalised estimating equations; GEE(i) indicates independence; GEE(e) indicates exhangeable.

Since the remaining simulation results presented will be for scenarios with nonignorable cluster size, only results for within-cluster resampling are given. In Table 4, all intercept estimates are slightly but significantly biased whereas the estimates for β_1 are not significantly biased. The within-cluster resampling coverages for β_1 are consistent with the nominal 0·95 and the power increases nicely with the sample size.

5. Example

We analysed data from the Intergenerational Epidemiologic Study of Periodontitis (Gansky et al., 1998, 1999). Multiple family members spanning three generations were included in the study. Each tooth of each participant was examined by several dentists

Table 4. *Results from* 1000 *simulations per scenario, analysed with within-cluster resampling, positive exposure effect with nonignorable cluster size,* $n_i \leqslant 8$, $\rho = 0{\cdot}2$, $b_0 = -1{\cdot}0986$ *and* $b_1 = 0{\cdot}4796$

No. of resamples (Q)	10 000	10 000	10 000	50 000
No. of clusters (I)	50	100	250	250
Covariate type	cs	cs	cs	os
$\bar{b}_0$	−1·1497	−1·1344	−1·1126	−1·1101
ESE($\bar{b}_0$)	0·0118	0·0076	0·0048	0·0046
$\bar{b}_1$	0·4811	0·4996	0·4922	0·4907
ESE($\bar{b}_1$)	0·0154	0·0103	0·0066	0·0057
avg{SE($\hat{b}_0$)}	0·3387	0·2349	0·1488	0·1402
avg{SE($\hat{b}_1$)}	0·4554	0·3178	0·2000	0·1735
Coverage	0·943	0·944	0·938	0·936
Power	0·211	0·356	0·701	0·814
No. of zero cells out of 1000Q	13 059	17	0	0

ESE, empirical standard error, i.e. standard error among estimates derived from simulations; avg(SE), average among estimated standard errors; CS, cluster-specific exposures; OS, observation-specific exposures.

for periodontal disease. The tooth number, 1–32, and type of tooth, molar, premolar, etc., were recorded as was whether or not the tooth met a priori criteria for periodontal disease, i.e. whether or not the mean clinical attachment level exceeded 3 mm. Demographic and dental hygiene information were also included. In this analysis, the study subject is the ‘cluster’ and each tooth is an ‘observation’. The eldest family member from the second generation of each represented family was selected for illustrative analysis. Additionally, only premolars and molars were selected for analysis, so the maximum cluster size was 16. Altogether 198 subjects were included, with an average of 13 teeth per subject.

There is reason to think that cluster size may be nonignorable in this dataset. People who are more susceptible to periodontal disease may already have lost some teeth from the disease, suggesting the plausibility of a negative relationship between the number of teeth and periodontal disease. We accordingly stratified the data by whether or not the number of teeth exceeded 12, and computed the proportion of teeth with and without periodontal disease. If cluster size is ignorable, then the two groups, few/more teeth, should have about the same proportion of teeth with periodontal disease. The proportion of teeth with periodontal disease in the few-teeth group was 0·52, according to an observation-based estimate, and the proportion of teeth with periodontal disease in the more-teeth group was 0·23. Therefore, not only is the validity of the generalised estimating equations approach for these data uncertain, but we can expect different results from within-cluster resampling and generalised estimating equations. For comparison we applied within-cluster resampling, and the generalised estimating equations method with independent and exchangeable covariances.

The outcome for the analysis is tooth-specific periodontal disease. One dichotomous covariate, defined by whether or not the subject brushed at least twice a day, is included in the logistic models to assess the effects of dental hygiene. We used $Q = 10\,000$ resamples in the within-cluster resampling analysis. The use of an exchangeable correlation structure for generalised estimating equations is consistent with the usual assumption of equal correlation between teeth within a mouth. However, the independent working correlation

Table 5. *Dental data example: modelled risk of periodontal disease in relation to dental hygiene, $\hat{\beta}$, with estimated standard errors in brackets, computed by three methods*

	WCR	GEE(i)	GEE(e)
Intercept	0·0685 (0·1725)	−0·3893 (0·1820)	−0·0510 (0·1710)
Brush $\geqslant 2$	−0·7862 (0·2188)	−0·4988 (0·2252)	−0·6893 (0·2164)
$\pi_{\bar{B}}$	0·52	0·40	0·49
π_B	0·33	0·29	0·32

WCR, within-cluster resampling; GEE, generalised estimating equations; GEE(i) indicates independence; GEE(e) indicates exhangeable; $\pi_{\bar{B}}$ (π_B) is the apparent rate of disease for those who do not (do) brush at least twice a day, computed according to cluster-based sampling (WCR) versus unit-based (GEE) sampling.

structure was also used, to illustrate the discrepancy between parameter estimates within the generalised estimating equations framework.

Results are shown in Table 5, revealing major discrepancies among the parameter estimates. From all the analyses, people who brushed at least twice a day had significantly less periodontal disease. However, in the within-cluster resampling analysis we found a stronger relationship between brushing and periodontal disease, and the evidence for nonignorable cluster size suggests that inference based on within-cluster resampling is more trustworthy than that based on generalised estimating equations.

Since there is reason to suspect nonignorability of cluster size, within-cluster resampling is appropriately a mouth-based analysis. By contrast, the generalised estimating equations approach weights the larger/lower risk clusters more, resulting in a deflated and invalid estimate of the risk of periodontal disease. The resulting difference can be seen by comparing the estimated probabilities of periodontal disease for those who did not brush at least twice per day, given as $\pi_{\bar{B}}$ in Table 5. The corresponding estimated probabilities of periodontal disease for a person who does brush at least twice per day are also given, as π_B in Table 5.

6. Discussion

The within-cluster resampling method is highly intuitive, but computationally intensive. However, it is simple to implement the within-cluster resampling algorithm in most computer languages or statistical software packages, such as C or SAS (SAS Institute Inc., 1989). As technology improves, the time required to perform a large number of resamples for a within-cluster resampling analysis will diminish.

Although this paper only presents simulations for correlated binary outcomes, the asymptotics presented in § 3 are valid for any clustered data under a generalised linear model. Other simulation studies are needed to characterise the finite sample operating characteristics of within-cluster resampling for non-binary clustered data, such as ordinal or continuous outcomes.

Acknowledgement

The research of Dr Elaine Borland Hoffman was supported by a pre-doctoral Intramural Research Training Award and a Environmental Training Grant, both from the National

Institute of Environmental Health Sciences. The authors would like to thank Dr Jane Weintraub and the National Institute of Health/National Institute of Dental Research for the use of the periodontal data, Dr Stuart for helpful comments, and Dr Lawrence L. Kupper, Dr Marie Davidian and Dr Irva Hertz-Picciotto for their guidance and encouragement of this research. The authors would also like to thank the reviewers of this paper for their helpful comments and suggestions.

Appendix

Proof of Theorem 1

We begin by taking the Taylor series expansion of the loglikelihood, $\log L_I(\beta)$, around the true parameter β. Let $\|u\| < K$, where $u \in \mathbb{R}^p$ such that $\|u\|$ is less than K for positive K. Define the following function of u for each q:

$$\begin{aligned}\lambda_I^{[q]}(u) &= I^{-\frac{1}{2}} u' U_I(\beta;\, q) + I^{-\frac{1}{2}} u' \frac{\partial^2}{\partial\beta\,\partial\beta'} \log L_I(\beta)|_{\beta^*} u \\ &= I^{-\frac{1}{2}} u' U_I(\beta;\, q) - \frac{1}{2I} u' J_I(\beta;\, q) u + Z_I^{[q]}(u), \qquad \text{(A1·1)}\end{aligned}$$

where $U_I(\beta;\, q)$ is the score vector for the qth resampled dataset, β^* belongs to the line segment joining β and $\beta + I^{-\frac{1}{2}} u$, and each $Z_I^{[q]}(u)$ is $o(1)$ in the first mean; see E. B. Hoffman's thesis for more details. Averaging (A1·1) across q for $q = 1, \ldots, Q$, we have

$$Q^{-1} \sum_{q=1}^{Q} \lambda_I^{[q]}(u) = Q^{-1} \sum_{q=1}^{Q} \left\{ I^{-\frac{1}{2}} u' U_I(\beta;\, q) - \frac{1}{2I} u' J_I(\beta;\, q) u + Z_I^{[q]}(u) \right\} \qquad \text{(A1·2)}$$

and, maximising (A1·2) with respect to u, we get

$$\begin{aligned}\hat{u} &= I^{\frac{1}{2}} Q^{-1} \sum_{q=1}^{Q} J_I^{-1}(\beta;\, q) U_I(\beta;\, q) + o(1) \text{ in the first mean} \\ &= I^{-\frac{1}{2}} Q^{-1} J^{-1}(\beta) \sum_{q=1}^{Q} U_I(\beta;\, q) + o(1) \text{ in the first mean}, \qquad \text{(A1·3)}\end{aligned}$$

since

$$Q^{-1} \sum_{q=1}^{Q} o(1) \text{ in the first mean} = o(1) \text{ in the first mean}.$$

In (A1·3), it is necessary to assume weak consistency (Sen & Singer, 1993, pp. 39–40), namely that

$$\lim_{I \to \infty} I^{-1} J_I(\beta) = J(\beta)$$

is a finite and positive definite matrix. We are able to take the derivative inside the summation in (A1·2) because of the uniform continuity established under the usual regularity conditions, which can be found in Sen & Singer (1993, pp. 337–9) and Casella & Berger (1990, pp. 213–20).

We can rewrite the average on the right-hand side of (A1·3), the average of Q score vectors, as the sum of independent and identically distributed random variables:

$$\begin{aligned}\bar{U}(\beta) &= Q^{-1} \sum_{q=1}^{Q} U_I(\beta;\, q) = Q^{-1} \sum_{q=1}^{Q} \sum_{i=1}^{I} U_i(\beta;\, q) \\ &= Q^{-1}\{(x'_{11} d_{11}^{-1} r_{11} + \ldots + x'_{I1} d_{I1}^{-1} r_{I1}) + \ldots + (x'_{1Q} d_{1Q}^{-1} r_{1Q} + \ldots + x'_{IQ} d_{IQ}^{-1} r_{IQ})\} \\ &= Q^{-1}\{(x'_{11} d_{11}^{-1} r_{11} + \ldots + x'_{1Q} d_{1Q}^{-1} r_{1Q}) + \ldots + (x'_{I1} d_{I1}^{-1} r_{I1} + \ldots + x'_{IQ} d_{IQ}^{-1} r_{IQ})\} \\ &= Q^{-1} \sum_{i=1}^{I} X_{Ii}'^{*} D_{Ii}^{*-1}(\beta) S_{Ii}^{*}(\beta), \qquad \text{(A1·4)}\end{aligned}$$

where d_{iq} is a partial derivative, r_{iq} is the residual, x_{iq} is a vector of covariates,

$$X_{Ii}'^{*} = [x_{i1}, \dots, x_{iQ}], \quad D^{*}_{Ii}(\beta) = \mathrm{diag}(d_{i1}, \dots, d_{iQ}), \quad S^{*}_{Ii}(\beta) = [r_{i1}, \dots, r_{iQ}]'.$$

Since (A1·4) is the sum of independent and identically random variables, we may appeal to a Central Limit Theorem (Sen & Singer, 1993, pp. 111–8) to establish the asymptotic normality of the standardised average of the score vectors across resamples.

By the definition of u, we can write

$$I^{\frac{1}{2}}(\bar{\beta} - \beta) = I^{-\frac{1}{2}} Q^{-1} J^{-1}(\beta) \sum_{q=1}^{Q} U_I(\beta; q) + o_p(1).$$

By Slutsky's Theorem, the desired result follows as $I \to \infty$:

$$I^{\frac{1}{2}}(\bar{\beta} - \beta) \to N_p(0, \Sigma),$$

in distribution, as $I \to \infty$, where Σ is an unspecified but finite and positive-definite matrix. It can be verified that

$$\Sigma = \mathrm{var}\left\{ I^{-\frac{1}{2}} Q^{-1} \sum_{q=1}^{Q} \sum_{i=1}^{I} U_i(\beta; q) \right\}$$

goes to a nonzero finite limit.

Appendix 2

Proof of Theorem 2

To prove Theorem 2, we apply an iterated variance formula to a single resample-based estimator, $\hat{\beta}(R)$, giving

$$\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\} = E[\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R) \mid \mathrm{data}\}] + \mathrm{var}[E\{I^{\frac{1}{2}}\hat{\beta}(R) \mid \mathrm{data}\}], \qquad (\mathrm{A}2{\cdot}1)$$

where the arguments on the right-hand side of (A2·1) are the respective expectation and variance over the resampling distribution for $\hat{\beta}(R)$, given the data. This variance representation can be simplified by noticing that the last term on the right-hand side of (A2·1) is the desired variance since $E\{\hat{\beta}(R) \mid \mathrm{data}\} = \bar{\beta}$, for very large Q. Thus

$$\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\} = E[\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R) \mid \mathrm{data}\}] + \mathrm{var}(I^{\frac{1}{2}}\bar{\beta}). \qquad (\mathrm{A}2{\cdot}2)$$

Rearranging (A2·2), we get

$$\mathrm{var}(I^{\frac{1}{2}}\bar{\beta}) = \mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\} - E[\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R) \mid \mathrm{data}\}]. \qquad (\mathrm{A}2{\cdot}3)$$

The first term on the right-hand side of (A2·3) is the true unconditional variance of a single resample-based estimator. Each $I\hat{\Sigma}(R)$, where $\hat{\Sigma}(R)$ is the estimated variance matrix for a random resampled dataset, is a consistent estimator of $\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\}$. The average of Q consistent estimators, $Q^{-1}I\sum\hat{\Sigma}(R; q)$, is consistent for $\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\}$.

The second term on the right-hand side of (A2·3) is

$$E[\mathrm{var}\{I^{\frac{1}{2}}\hat{\beta}(R)\} \mid \mathrm{data}] = Q^{-1} I E\left[\sum_{q=1}^{Q} \{\hat{\beta}(R; q) - \bar{\beta}\}\{\hat{\beta}(R; q) - \bar{\beta}\}' \right] = E(IS^2_{\beta}).$$

To show that $I\{S^2_{\beta} - E(S^2_{\beta})\} \to 0$ in probability, it suffices to verify that $\mathrm{var}(IS^2_{\beta}) \to 0$. By Markov's Theorem (Gnedenko, 1962, pp. 234–5),

$$\begin{aligned} \mathrm{var}(IS^2_{\beta}) \simeq{} & Q^{-1}\, \mathrm{var}[I^{\frac{1}{2}}\{\hat{\beta}(R; q) - \bar{\beta}\} I^{\frac{1}{2}}\{\hat{\beta}(R; q) - \bar{\beta}\}'] \\ & + \mathrm{cov}([I^{\frac{1}{2}}\{\hat{\beta}(R; q) - \bar{\beta}\} I^{\frac{1}{2}}\{\hat{\beta}(R; q) - \bar{\beta}\}'], [I^{\frac{1}{2}}\{\hat{\beta}(R; q^*) - \bar{\beta}\} I^{\frac{1}{2}}\{\hat{\beta}(R; q^*) - \bar{\beta}\}']), \end{aligned}$$

which for large Q approximately equals

$$I^2\, \mathrm{cov}[\{\hat{\beta}(R; q) - \bar{\beta}\}\{\hat{\beta}(R; q) - \bar{\beta}\}', \{\hat{\beta}(R; q^*) - \bar{\beta}\}\{\hat{\beta}(R; q^*) - \bar{\beta}\}']. \qquad (\mathrm{A}2{\cdot}4)$$

It is shown in E. B. Hoffman's thesis that the expression in (A2·4) goes to zero as $I \to \infty$. Consequently, a consistent estimator of Σ is

$$\hat{\Sigma} \simeq I \left\{ Q^{-1} \sum_{q=1}^{Q} \hat{\Sigma}(R;\, q) - (Q-1) Q^{-1} S_{\beta}^{2} \right\},$$

where Q is very large, $\hat{\Sigma}(R;\, q)$ is the estimated covariance matrix from the qth analysis, and

$$S_{\beta}^{2} = (Q-1)^{-1} \sum_{q=1}^{Q} \{\hat{\beta}(R;\, q) - \bar{\beta}\}\{\hat{\beta}(R;\, q) - \bar{\beta}\}'$$

is the sample covariance matrix among the $\hat{\beta}(R;\, q)$'s.

References

Casella, G. & Berger, R. L. (1990). *Statistical Inference.* Belmont, CA: Duxbury.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalised linear models. *Ann. Statist.* **13**, 342–68.

Gansky, S. A., Weintraub, J. A. & Shain, S. (1998). Parental periodontal predictors of oral health in adult children of a community cohort. *J. Dental Res.* **77(Spec Iss B)**, 707.

Gansky, S. A., Weintraub, J. A. & Shain, S. (1999). Family aggregation of periodontal status in a two generation cohort. *J. Dental Res.* **78(Spec Iss)**, 123.

Gnedenko, B. V. (1962). *Theory of Probability.* New York: Chelsea.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.

Mancl, L. A. & Leroux, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500–11.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. & Fisher, M. R. (1996). A survey of methods for analysing clustered binary response data. *Int. Statist. Rev.* **64**, 89–118.

SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6.* Cary, NC: SAS Institute.

Sen, P. K. & Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* New York: Chapman and Hall.

Stokes, M. E., Davis, C. S. & Koch, G. G. (1995). *Categorical Data Analysis using the SAS System.* Cary, NC: SAS Institute.

Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–30.

[*Received May* 1999. *Revised May* 2001]