Published in final edited form as:

Biometrics. 2016 June ; 72(2): 441–451. doi:10.1111/biom.12456.

Cluster Adjusted Regression for Displaced Subject Data (CARDS): Marginal Inference under Potentially Informative Temporal Cluster Size Profiles

Joe Bible^{1,*}, James D. Beck^{2,**}, and Somnath Datta^{3,***}

¹University of Louisville, Louisville, Kentucky, 40292 U.S.A.

²University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599 U.S.A.

³University of Florida, Gainesville, Florida, 32610 U.S.A.

Summary

Ignorance of the mechanisms responsible for the availability of information presents an unusual problem for analysts. It is often the case that the availability of information is dependent on the outcome. In the analysis of cluster data we say that a condition for *informative cluster size* (ICS) exists when the inference drawn from analysis of hypothetical balanced data varies from that of inference drawn on observed data. Much work has been done in order to address the analysis of clustered data with informative cluster size; examples include Inverse Probability Weighting (IPW), Cluster Weighted Generalized Estimating Equations (CWGEE), and Doubly Weighted Generalized Estimating Equations (DWGEE). When cluster size changes with time, i.e., the data set possess temporally varying cluster sizes (TVCS), these methods may produce biased inference for the underlying marginal distribution of interest. We propose a new marginalization that may be appropriate for addressing clustered longitudinal data with TVCS. The principal motivation for our present work is to analyze the periodontal data collected by Beck et al. (1997, *Journal of Periodontal Research* 6, 497–505). Longitudinal periodontal data often exhibits both ICS and TVCS as the number of teeth possessed by participants at the onset of study is not constant and teeth as well as individuals may be displaced throughout the study.

Keywords

Cluster weighted generalized estimating equations; Informative cluster size; Temporally varying cluster size

1. Introduction

It is no secret that issues involving imbalanced data plague many otherwise well thought out experiments. In the clustered longitudinal data context issues involving data imbalance are

* j bible831@gmail.com. ** Jim_Beck@unc.edu. *** somnath.datta@ufl.edu.

7. Supplementary Materials

Web Appendices referenced in Sections 3.1 and 3.2 as well as Figures 7 and 8 are available with this article at the *Biometrics* website on Wiley Online Library. Also, R code for applying our methods is available at the Online Wiley library.

exacerbated by the temporal availability of subjects where cluster size can change with time. The dental data collected by Beck et al. (1997a) presents a perfect example of the extent to which both baseline cluster size and temporal availability can impact observable information. In the data collected by Beck et al. (1997a) information such as Attachment Loss (AL), pocket depth and number of teeth was collected on teeth belonging to elderly individuals over the course of five years. In this investigation, the primary outcome of interest is Attachment Loss, or more specifically the sum of buccal and mesial attachment losses. AL is a measurement quantifying the loss of a tooth's connective tissue and is used as a primary indicator of periodontal disease. In this study, some of the subjects retain all 32 teeth over the five year period while others have only 1 tooth at baseline. Further complicating the matter, some individuals lose teeth during the study and others are removed from the study entirely due to various causes. We refer to such tooth losses as displacement in this article. Since such displacements are permanent, they can also be considered to be dropouts in a longitudinal study. However, the methods we develop can be applied even when the displacement is not permanent.

Informative cluster size presents itself in a number of cross-sectional settings ranging from litter size in toxicology studies (Hoffman, Sen, and Weinberg, 2001) to the number of retained or available teeth in dental applications where the clustering unit is the patient. The extension to a longitudinal framework further complicates the matter by introducing another layer of nesting such that cluster size may be non-constant with time. Existing methodologies address informative cluster size for cross-sectional applications (Hoffman et al., 2001; Williamson et al., 2003; Benhin et al., 2005; Huang and Leroux, 2011) and for clustered longitudinal data applications when the cluster size is constant with time (Wang et al., 2011), but to our knowledge no existing methodology addresses informative cluster size in clustered longitudinal data with temporally varying cluster sizes.

It is often the case that the number of cluster members available at the onset of a study and an individual cluster members attrition are dependent in some manner on the outcome. Such a dependence is in violation of the underlying assumptions of many methods for analyzing clustered longitudinal data, specifically Generalized Estimating Equations (GEE). Often the motivation for analyzing clustered data is to draw conclusions about the overall effect of various covariates on population means. The utility of GEE methods is that they allow estimation of these (marginal) population parameters while accommodating the within cluster effects, where the within cluster effects are of either auxiliary or no interest to the study question. In contrast to a subject-specific modeling approach where focus is on the effect of covariates on an individual's outcome trajectory. While the periodontal data in this example lends itself to both types of analyses, we motivate the population average (marginal) approach as we are more interested in the effect of environmental factors on the population mean rather than the mean individual trajectory. For further elucidation regarding population average and subject specific models within the GEE context the reader is directed to Zeger, Liang, and Albert (1988).

GEE's enjoy the luxury of a wide range of applications and popularity afforded by their ease of use. Their behavior in the presence of dependence between the availability of data and the outcome within the context of non-longitudinal settings has received considerable attention

over the past decade. In Section 2 we will discuss *informative cluster size* and *temporally varying cluster size*, the situations where there is a dependence between the outcome and the availability of information. In Section 3 we discuss various marginal analyses and motivate the implementation of each according to contextual appraisal of the data. In particular, we introduce a novel reweighted generalized estimating equation that could be used to properly treat the TVCS situation.

In Section 4, we provide the results of a simulation comparing marginal observed cluster inference drawn from an applicable design to that of the standard GEE. Our work was motivated principally by the need to analyze periodontal data. In such a setting the clustering unit is the subject under study and the teeth are the individual cluster members.

Heuristically, one can understand the relationship between the number of teeth an individual has and that individual's overall oral health. Similarly, not all teeth are retained throughout study and the number of teeth retained at a given point in study cannot, in good conscience, be considered independent of the number of teeth present at baseline. In Section 5, we analyze the data from the Piedmont study using an appropriate marginalization and compare our results with standard approaches.

2. Informative Cluster Size

2.1. Notation

Let (Y_{ijk}, X_{ijk}) be the outcome and covariate vector of the j th unit belonging to the i th cluster at observation k and Q the number of clusters. In our example $i \in \{1, 2, \dots, Q\}$ would index the Q individuals under study, $j \in \{1, 2, \dots, 32\}$ would index the 32 possibly observable teeth and $k \in \{1, 2, 3, 4\}$ would denote each of the four observation occasions. For the sake of simplicity throughout the rest of the article we concern ourselves with fixed observation times such that $t_{ijk} = t_{i'j'k}$ where t_{ijk} denotes the time (in units) corresponding to the observation occasion k . We assume that units under study do not return once they miss an observation occasion. The assumption of permanent dropout is not necessary but does simplify notation and calculation. Let, N_i^k be the number of units belonging to the i th cluster which were observed at the k th observation time, N_i is the cluster size at baseline and, k_{ij} be the number of observations made on the j th unit (k_{ij} , i.e., the number of temporal observations made on a subunit will also be referred to as the subunit's temporal profile).

2.2. Informative Cluster Size

We can regard ICS as dependence between outcome and cluster size. Consider a hypothetical example simpler than the periodontal data where we define Y_{ij} as a response measured on Q individuals indexed by $i \in \{1, 2, \dots, Q\}$ with N_i repeated measures indexed by $j \in \{1, 2, \dots, N_i\}$ made on each individual. Say $Y_{ij} = a_i + \epsilon_{ij}$, where a_i 's are i.i.d. and $a_i \sim N(\mu, \sigma^2)$, ϵ_{ij} 's are i.i.d. and $\epsilon_{ij} \sim N(0, \tau^2)$. Further assume that f is an increasing function taking values in $\{1, 2, \dots\}$ so that $\text{corr}(f(a_i), a_i) > 0$, and $N_i = f(a_i)$. We consider two ways of calculating the expectation of Y_{ij} :

$$\hat{\mu}_1 = \frac{\frac{1}{Q} \sum_i \sum_j Y_{ij}}{\frac{1}{Q} \sum_i N_i}, \quad (1)$$

and

$$\hat{\mu}_2 = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}. \quad (2)$$

Note that

$$\begin{aligned} E(\hat{\mu}_1) &= \frac{E\left(\frac{1}{Q} \sum_i \sum_j Y_{ij}\right)}{E\left(\frac{1}{Q} \sum_i N_i\right)} + o(1), \\ &= \frac{E(\alpha_1 f(\alpha_1))}{E(f(\alpha_1))} + o(1) \\ &= \frac{\text{cov}(\alpha_1, f(\alpha_1))}{E(f(\alpha_1))} + \mu + o(1) > \mu, \quad \text{for large } Q. \end{aligned}$$

However,

$$E(\hat{\mu}_2) = E\left(\frac{1}{f(\alpha_1)} E\left(\sum_j Y_{1j} \mid \alpha_1\right)\right) = E(\alpha_1) = \mu.$$

The discrepancy exhibited by equations 1 and 2 is an example of the effects of informative cluster size and the marginalization in equation 2 is analogous to that of *Cluster Weighted Generalized Estimating Equations* (CWGEE), Williamson, Datta, and Satten (2003); also see Benhin, Rao, and Scott (2005) and Hoffman, Sen, and Weinberg (2001).

One popular approach for addressing bias in the presence of ICS is the implementation of *Weighted Generalized Estimating Equations* (WGEE's). WGEE's exploit the flexibility of the underlying standard (i.e., unweighted) GEE framework by reweighting individual components of the corresponding estimating equation. To this point a great deal of the literature concerned with ICS considers cross sectional or non-temporal repeated measures type data. Wang et al. (2011) extended the CWGEE framework developed by Williamson et al. intended to address non-temporal clustered data, in order to accommodate longitudinal data where the number of temporal measures made on each unit is constant. It was shown that under certain conditions CWGEE could, through marginalization, produce unbiased estimators of the corresponding complete cluster regression parameters.

2.3. Temporally Varying Cluster Size

In general, *Temporally Varying Cluster Size* (TVCS) introduces another dimension to the situation which cannot be disregarded for a proper analysis of the data. For the purposes of this investigation we concern ourselves with permanent subject displacement, that is we

assume that when a subject is unobservable at a given time they remain unobservable at all subsequent time points. This is a natural assumption when data are collected on a mortal cohort. This is also meaningful in dental studies since a lost tooth will remain unobservable for subsequent occasions. Under this pattern of dropout the temporally varying cluster sizes will be a decreasing function of time.

The following simple example illustrates why TVCS may lead to a discrepancy between the observed (available) and complete cluster relationships. Consider a hypothetical cluster with baseline cluster size $N_1 = 100$. Assume, that the individuals ($j \in 1, 2, \dots, 100$) experience an outcome Y_{1jk} temporally over the course of 20 years. Suppose the outcome follows the model,

$$Y_{1jk} = k \times \gamma_{1j} + \epsilon_{1jk},$$

with $\gamma_{1j} \sim N(0, 1)$ and $\epsilon_{1jk} \sim N(0, 1)$. This defines the distribution of a hypothetical complete cluster. Now, let us consider that each individual has the possibility of being displaced with a constant hazard proportional to $.2 \times (1 - \Phi(\gamma_{1j}))$, where $\Phi(\cdot)$ is the standard normal CDF. The implication of the dependence of the constant hazard on the cluster member specific slope (i.e., γ_{1j}) is that cluster members with larger positive slopes will be retained longer, on average, than those with larger negative slopes. Figure 1 plots the observed distribution given the example design and highlights how in the longitudinal setting temporally varying cluster size cannot be ignored. Looking at each of the four panels it is easy to see how as time passes the distribution of observed information becomes more obscured from that of the underlying complete cluster. While in the first panel the observed distribution is similar to that of the complete cluster the distribution of the subjects remaining in the fourth panel cannot be considered a representative sample of the corresponding complete cluster.

On a related note, a data analyst may need to decide which inference is of primary interest. Suppose that during a hypothetical study we are able to observe only panel 4 of Figure 1, i.e., 16 to 20 years. Given the mechanism responsible for subject displacement, it might be possible to make inference about the complete cluster distribution. However, given that more than 60% of the original subjects are displaced by the 16th year it could be argued that the resulting distribution of retained 16 to 20 year olds would be of considerably more interest. This would be an example of a marginal conditional analysis.

3. The Case for Scientifically Motivated Marginalization and Statistical Inference

In the marginal analysis of clustered data we assume that each cluster is drawn from some common distribution, similarly, subjects belonging to said clusters are drawn from another common distribution. The observed cluster distributions may vary but the variations are assumed to be the result of some underlying common probabilistic function. As such, when we wish to estimate the distribution of a typical cluster or a typical subject belonging to a typical cluster, the contribution of individual subjects or clusters becomes of paramount importance. If we are interested in estimating the distribution of a typical cluster, it would be

reasonable to conclude that each observed cluster should contribute proportional information in the estimation process, thus mitigating the ability of atypically sized clusters to dominate the resulting analysis. Similarly, in the presence of TVCS if we were concerned with estimating the distribution of a typical subject belonging to a typical cluster who was observable at some point over a given interval, it would be reasonable to conclude that each subject within a given cluster should be allowed to contribute proportionately to that clusters contribution to the analysis and that the sum of the marginalized subject contributions should be reweighted as to allow proportionate contribution of each cluster.

Given the often interrelated nature of ICS at baseline and TVCS, estimation of an underlying complete cluster distribution is not necessarily possible without making considerable assumptions about the form of the underlying distribution and the mechanism responsible for TVCS; for examples, see Seaman, Pavlou, and Copas (2014). We proceed now by justifying a number of relevant marginal analyses and motivate their implementation based on contextual relevance of the marginal distributions that they seek to describe.

3.1. Weighted Generalized Estimating Equations

Generalized Estimating Equations as described by Liang and Zeger (1986) provide a flexible platform for the analysis of clustered data. Through the use of a general link function GEE's are capable of modeling various non-linear outcomes as a function of linear combinations of covariates. Furthermore, GEE's can accommodate the within cluster variance structure which addresses the mutually correlated relationship of cluster specific outcomes without estimating a cluster or subject specific effect. Let, $X\beta$ define the linear predictor where β is a p -dimensional vector of parameters, $g(\beta) = \mu = E[Y|X]$ define the link relating the linear predictor to Y such that $g^{-1} X\beta$ and corresponding variance covariance relation as V .

In practice, the true correlation structure is, generally, unknown and some appropriate correlation structure is assumed. We will consider a working variance covariance matrix V that is block diagonal with blocks defined by $k_{ij} \times k_{ij}$ matrices V_{ij} . The matrix V_{ij} represents the working variance covariance matrix for the temporal observations on unit j of cluster i . Some common choices for the corresponding correlation matrix include the identity matrix, the compound symmetric matrix, and autoregressive correlation structures. In simulation, we will demonstrate that ignoring the between subject correlation has little effect on the observed bias but that assuming some reasonable subject specific correlation structure has a notable impact on the power; such results are in-line with the general GEE methodology.

Given the specified variance covariance relation we define β^{GEE} as the solution to

$$\sum_{i=1}^Q \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}) = 0, \quad (3)$$

where we note that $\frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij})$ is a p -dimensional vector. This will provide a marginal inference based on each subject irrespective of their cluster membership. If we

wished to make marginal inference on the distribution of a typical cluster we could consider the following marginalization:

$$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}) = 0. \quad (4)$$

This marginalization has the interpretation of describing the distribution of a typical subject belonging to a typical cluster. This is the CWGEE marginalization proposed by Williamson et al. in 2003 which was shown to be asymptotically equivalent to the Within Cluster Resampling (WCR) introduced by Hoffman et al. (2001). This will be appropriate provided each subject experiences the same number of temporal observations, or it can reasonably be assumed that the mechanism responsible for TVCS was such that subject specific temporal profiles are not informative of their responses.

A shortcoming of this approach as pointed out by Huang and Leroux (2011) is that it does not properly marginalize with respect to the distribution of cluster member covariates. In other words, it does not accommodate potentially informative covariate distributions within a cluster. To that end, they proposed a novel marginalization called DWGEE and its variant DWGEE2.

In the treatment of TVCS, we propose a modification to the reweighting scheme for the marginalization proposed by equation 4. Marginalizing over all the temporal observations made on each subject has the effect that each individual, regardless of how long they were observed, would contribute the same amount of information to the marginal distribution within the cluster it belongs. Furthermore, the inverse cluster size weighting as in equation 4 ensures that each cluster contributes equally irrespective of its size to the overall distribution. This leads to the estimating equation:

$$\sum_{i=1}^Q \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{1}{k_{ij}} \right) \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} (Y_{ij} - \mu_{ij}) = 0. \quad (5)$$

The marginalization in equation 5 has the interpretation of estimating the distribution of a typical temporal observation on a typical subject belonging to a typical cluster. Note that equation 5 makes no assumptions about the mechanism responsible for TVCS.

An augmentation to the WCR resampling scheme could be proposed in order to accommodate the marginalization of equation 5 for the special case of the identity correlation matrix. In order to accomplish this, one would need to produce the resampled data sets $\{Y_i^*, X_i^*\}$ by first resampling one primary unit (e.g., tooth) from each cluster (e.g., patient), followed by resampling one temporal record (e.g., attachment loss at a given visit) at random from the set of all its temporal records. The resampled data sets should then be analyzed by an estimating equation of the form

$$\sum_{i=1}^Q \frac{\partial \mu_i^*}{\partial \beta} (Y_i^* - \mu_i^*) = 0, \quad (6)$$

and the estimates aggregated. For utilizing a general working variance covariance matrix V_{ij} , the entire vector of temporal records corresponding to a selected subunit from each cluster needs to be sampled with probability proportional to the inverse of the subunits' temporal profiles. The resulting data sets need to be analyzed using a GEE and the results aggregated. This leads to a normalized form of estimation equation 5; see Web-Appendix A.

3.2. Variance of WGEE's

The sandwich estimator of the variance of β^{GEE} for clustered longitudinal data can be obtained by,

$$\Sigma^{GEE} = \Omega^{-1} M \Omega^{-1}, \quad (7)$$

with,

$$M = \sum_{i=1}^Q \left(\left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \epsilon_{ij} \right\} \left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \epsilon_{ij} \right\}^T \right), \quad (8)$$

where $\Omega = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} V_{ij}^{-1} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right)$. We propose a generalized sandwich estimator applicable to WGEE's for drawing inference about parameters obtained via marginal analysis. Let ω_{ij} be the diagonal matrix of marginal weights applied to the j th subject

belonging to the i th cluster. For example, in equation 5, $\omega_{ij} = \text{diag} \left\{ \frac{1}{N_i k_{ij}}, \dots, \frac{1}{N_i k_{ij}} \right\}$. We can then express the variance of the general sandwich estimator for WGEE's as,

$$\Sigma^{WGEE} = \tilde{\Omega}^{-1} \tilde{M} \tilde{\Omega}^{-1}, \quad (9)$$

with

$$\tilde{M} = \sum_{i=1}^Q \left(\left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\} \left\{ \sum_{j=1}^{N_i} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \epsilon_{ij} \right\}^T \right), \quad (10)$$

where $\tilde{\Omega} = \sum_{i=1}^Q \sum_{j=1}^{N_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} V_{ij}^{-1}) \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right)$. The asymptotic normality of β^{WGEE} follows as a consequence of the independence of the block sums of the WGEE estimating

equation. We defer the algorithmic details of parameter estimation to the Web Appendices B and C.

4. Simulation

In the simulation portion we focus our attention on a design for which equation 5 is applicable. We report the mean, bias and variance of the marginal parameter estimates obtained from a Monte Carlo simulation in which we calculate both the standard GEE and the marginal WGEE estimates obtained using equation 5.

Let $Y_{ijk} \sim N(X_{ijk}\beta + \alpha_i + \gamma_{ij}, 1^2)$, where X is a $4 \times N$ design matrix, $\alpha_i \sim N(0, 1)$ and $\gamma_{ij} \sim N(0, .5^2)$. The components of $\beta = (.5, 1.5, .75, .8)$ correspond to the intercept, time effect, cluster constant exposure, and exposure time interaction with half of the Q clusters having time constant exposure $X_{2,i} = 1$. We denote $X_{2,ijk}$ by $X_{2,i}$ since it is constant with respect to j and k . The cluster size at baseline $N_i^1 \sim \text{Pois}(\exp\{\lambda_i\}) + 1$ where $\lambda_i = 1 + .25\alpha_i + X_{2,i}$. The number of temporal observations made on each subject varies between 2 and 4 with the individual number of observations determined by the ordinal logit model with probabilities

$$\begin{aligned}\pi_{k_{ij}=2} &= \frac{e^{-.01 \times N_i^1}}{1 + e^{-.01 \times N_i^1}}, \\ \pi_{k_{ij}=3} &= 1 - \frac{e^{-.01 \times N_i^1}}{1 + e^{-.01 \times N_i^1}} - \frac{1}{1 + e^{1.5 - .01 \times N_i^1}}, \quad \text{and} \\ \pi_{k_{ij}=4} &= \frac{1}{1 + e^{1.5 - .01 \times N_i^1}}.\end{aligned}$$

For the indicated design we conduct two Monte Carlo simulations generating 500 data sets from the corresponding distributions, one with $Q = 50$ and another with $Q = 200$. For each iteration we calculate the standard GEE estimate as well as the marginal WGEE estimate of equation 5. We report the summary results of the simulation in Table 1 and Figures 2 and 7 (Web Appendix).

4.1. Simulation Results

Analysis of data generated from the indicated design using the standard GEE results in bias being introduced to $\hat{\beta}_0$ and $\hat{\beta}_2$ in both the $M = 50$ and the $M = 200$ settings. In fact the bias of $\hat{\beta}_2$ actually increases in the larger sample setting. In the small sample case ($M = 50$), we see relative biases of around .75 and .06 associated with $\hat{\beta}_0$ and $\hat{\beta}_2$ respectively. These values for the larger sample ($M = 200$) setting were around .72 and .16, respectively.

The marginal WGEE provides consistent bias reduction w.r.t. the standard GEE method in the estimation of the intercept term in both the small and the large sample settings. The absolute bias corresponding to the exposure variable, i.e., that of $\hat{\beta}_2$, is not reduced in the small sample setting, but is corrected in the larger sample setting. It should be noted that regardless of the choice of correlation structure marginal WGEE estimates are in reasonable agreement with one another. The power plots of Figures 2 and 7 (Web Appendix) point to the true benefit of specifying a reasonable working correlation structure. As can be seen from Figures 2 and 7 (Web Appendix) in all cases the marginal WGEE appears to maintain

appropriate size but there is consistent improvement to the power. The improvements to power are considerably less w.r.t. β_0 and β_2 , i.e., the variables for which marginal WGEE mitigates the induced bias of a standard GEE analysis. However, making a reasonable accommodation for the within subject correlation provides consistent improvement w.r.t to power which is consistent with the general GEE methodology.

5. Periodontal Application

We illustrate our method by analyzing the data collected by Beck et al. (1997a) for the Piedmont 65 + Dental Study. About 800 dentulous elderly patients residing in the Piedmont Region of North Carolina participated in the study. For each of the participants various measures of dental health were collected on available teeth including Attachment Loss (AL), Gingival Recession, and Pocket Depth. Similarly, demographic and lifestyle information was collected on each individual at baseline which included sex, race (Black and White), tobacco use, level of education obtained, length of time since last visit to dentist, marital status, socioeconomic information, and whether an individual resided in an urban community; the reader is directed toward Beck et al. (1997a) for a more detailed accounting of the study design. Beck et al. (1997a), Beck et al. (1997b) and Wang et al. (2011), among others, have analyzed these data in an effort to describe the dependence of AL on various demographic and lifestyle factors.

We retain 452 individuals in our analysis with 7823 teeth for a total of 25,183 records. The 452 retained individuals correspond to those who were observed at both baseline, the 18 month period and whose subsequent observations followed a monotonic displacement pattern; that is to say that we removed those subjects that missed an intermediate visit and then returned to study. In this particular investigation the choice to retain only the subjects with monotonic displacement was made to simplify estimation of the correlation parameters. This is in contrast to the investigation conducted by Wang et al. (2011) where only 74 individuals, which were observed at all 4 observation occasions without loss of teeth, were retained for analysis.

As such, the number of temporal observations made on each individual and each tooth varies. For further clarification, Figure 3 is included showing the distributions of teeth at baseline, the number of teeth summarized by total number of observations contributed, and the number of records contributed by each individual. In particular, Panel (c) of Figure 3 illustrates the contrast between the number of records each individual contributes to the analysis.

To further motivate our marginal analysis, Figure 8 (Web Appendix) is included to underline the relationship between number of teeth at baseline, tooth retention, and AL. The discrepancies between Panels (b) and (c) of Figure 8 (Web Appendix) convey the complexity of the relationship between AL and the availability of information.

In our investigation we consider three analyses, the standard GEE, the CWGEE, and the marginal WGEE described in equation 5; in each we employ a racially stratified analysis. Rationalization for investigating the Black and White participants separately comes from

disparities between the distributions of demographic, socio-economic and educational information of the two groups. For example, if we compare the mean years of education attained for the 215 Blacks which were retained in our analysis to that of the 237 retained Whites, the mean of the Black group is 3.66 years higher than that of the White group. Similarly the retained Black participants, on average, visited the dentist more frequently, had higher socio-economic indicators, and were less likely to use tobacco than the retained White participants.

We model AL using a weighted generalized estimating equation where we assume that the identity link is appropriate; we accommodate within tooth temporal correlation assuming an AR-1 correlation structure. An independence working correlation structure between teeth belonging to the same individual is assumed in the estimation of $\hat{\beta}^{WGEE}$. The weightings employed correspond to the CWGEE (equation 4) and that as described in equation 5, respectively. We motivate the use of the marginalization described in equation 5 upon consideration of the relationship between the number of teeth at baseline, the number of temporal observations made on a given tooth and AL described in Figure 8 (Web Appendix). Use of the CWGEE marginalization is difficult to motivate as there is not a one to one relation between the number of records each individual contributed and the number of teeth present at baseline; however, it is included for a comparison.

For our racially stratified analysis we investigate the following linear model:

$$\begin{aligned}
 E[AL|covariates] &= \beta_0 + \beta_1 T + \beta_2 SEX \\
 &+ \beta_3 EDUC \\
 &+ \beta_4 WHENDDS \\
 &+ \beta_5 MARRIED \\
 &+ \beta_6 SES \\
 &+ \beta_7 TOBACCO \\
 &+ \beta_8 URBAN \\
 &+ \beta_9 INCOME \\
 &+ \beta_{10} (SEX \times T) + \beta_{11} (EDUC \times T) + \beta_{12} (WHENDDS \times T) \\
 &+ \beta_{13} (MARRIED \times T) \\
 &+ \beta_{14} (SES \times T) + \beta_{15} (TOBACCO \times T) \\
 &+ \beta_{16} (URBAN \times T) \\
 &+ \beta_{17} (INCOME \times T).
 \end{aligned} \tag{11}$$

The factor effects SEX, MARRIED, TOBACCO, and URBAN in equation 11 are coded 0 and 1 such that SEX = 1 implies the subject is male, URBAN = 1 implies that a subject resided in an urban community, TOBACCO = 0 corresponds to the tobacco users. EDUCATION represents to the number of years of education received. While insignificant, MARRIED, WHENDDS, SES, and INCOME represent marital status, years since dentist visit, socioeconomic status and income, respectively. We report only the effect sizes of the significant effects, where significance is determined by a Wald test with corresponding

variance estimates obtained according to equation 7 in the case of the GEE analysis and equation 9 in the case of CWGEE and WGEE analyses. The resulting marginal WGEE model describes the conditional expectation of AL of a typical record on a typical tooth belonging to a typical individual sampled from the corresponding population. Thus, for example, an estimated TOBACCO use effect size of -1.489 corresponding to the White population would indicate that in a typical examination on a typical tooth belonging to a typical individual in the White population who does not use tobacco would, on the average, have an AL value 1.489 lower than a similar tooth belonging to tobacco users (everything else being equal).

Table 2 presents the results of the racially stratified GEE, CWGEE, and marginal WGEE analyses.

While the effect of tobacco use is consistently significant across all six analyses the effect size varies from one analysis to another. In the White sample the same effects remain significant across all three analyses; however, there are marked discrepancies between the effect sizes in each. We see an increase in the estimated intercept from 7.686 to 8.857 , the estimated effect size of URBAN dwelling decreases from $-.989$ to -1.610 , the effect size of SEX decreases from -1.138 to -1.439 , and TOBACCO use decreases from -1.351 to -1.489 when comparing the GEE analysis to the WGEE analysis. Such a result is consistent with our expectation given the interrelated nature of number of teeth at baseline, number of teeth present at a given observation occasion, AL, and various demographic information. The Black sample tells a slightly different story. While there are still differences in the estimated effect sizes, the TIME and TIME $\times$ EDUCATION interaction terms are significant in the Black WGEE analysis whereas they are not in the standard GEE (and neither are significant in either the standard GEE or WGEE analysis of the White population). In a similar vein, the difference in the estimated effect size of TOBACCO use from the standard GEE to the WGEE analysis in the Black population is in the opposite direction from that of the White population. Figure 4 is included to outline a number of discrepancies between the distribution of demographic information between the two samples. Note in Figure 4 we can see that the Black population on average visited the dentist more frequently (b), had more education (a), higher socioeconomic status (c), and had higher incomes (d) than their White counterparts. It is in light of these discrepancies that we defend disparities between estimated effect sizes and significant effects for each population.

6. Discussion

The GEE framework is well established and its use is widely supported in the marginal analysis of clustered data. We have provided a novel WGEE formulation as well as a sandwich estimator of the corresponding variance of $\hat{\beta}^{WGEE}$ for clustered longitudinal data. We have demonstrated through simulation and the analysis of data from an observational study how various marginal distributions and the resulting inference can differ. We advocate the thoughtful application of marginal WGEE analyses in situations where there is reasonable concern about ICS or TVCS and their effects on population averaged effects of interest.

In our present investigations we have considered only data with units experiencing monotonic temporal displacement, though such a condition is not necessary for the application of a marginal WGEE investigation. Depending on the desired marginalization, the unrestricted retention of units experiencing any of a number of combinatorial arrangements of observations occasions can lead to a number of problems in the estimation of marginal WGEE parameters. The first issue worth considering is the estimation of the correlation parameters. When all retained subjects experience continuous observation a first order approximation is appropriate. In situations where retained units are allowed to miss observations, more complicated estimation methods must be employed to acquire appropriate estimates. The second being that sparsity issues can arise if there exist a number of units experiencing unique observation profiles.

We understand that coding individual analysis specific algorithms to implement marginal WGEE analyses is a major limiting factor to its wide spread acceptance. As such in various settings we have repeated a number of the analyses from this investigation in both SAS and R. There does not appear to be an existing function/package in R for appropriately addressing marginal WGEE analyses, however, use of the `< weights= >` option in Proc Genmod provides parameter estimates remarkably similar to those obtained using our formulation. It is also very important to note that often the working correlation structure is misspecified (as was the case in our applied analyses and a number of our simulations). Such a misspecification (ignoring dependencies between units) will result in liberal variance estimates of the marginal parameters if the incorrect dependencies are extended to the variance covariance calculation. One way to mitigate this issue is to write a case specific function, another is to bootstrap the variance estimator sampling the largest unit with dependencies. In our periodontal example this means sampling with replacement from each racially stratified population, obtaining estimates in this manner provides a reasonable approximation to the correct variances.

The present marginal WGEE approach is recommended as an alternative to a fully likelihood-based approach. Joint modeling while effective requires that the analyst make a number of additional assumptions regarding the distributions giving rise to the observed information. The quality of joint model analyses is dependent on the tenability of these extra assumptions, see, for example, Chen, Zhang, and Albert (2011) and Dunson, Chen, and Harry (2003). In the present context such models will be more complicated and would require specification of all (correlated) sources of randomness including the response, the baseline cluster size, and the temporally varying cluster size (due to successive tooth loss). In future works, we may explore the joint model approach including efficient computation of the likelihood function, robustness with respect to model misspecification for the ICS and TVCS and a side-by-side comparison with the marginal model approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the editor, the associate editor, and the reviewers for many constructive comments during the reviewing of this manuscript. This research was supported by NIH grants R03DE020839 and R03DE022538. We thank Kevin Moss for helpful discussions regarding the periodontal data.

References

- Beck JD, Cusmano L, Greene-Helms W, Koch GG, Offenbacher S. A 5-year study of attachment loss in community-dwelling older adults: Incidence density. *Journal of Periodontal Research*. 1997b; 32:506–515. [PubMed: 9379318]
- Beck JD, Sharp T, Koch GG, Offenbacher S. A study of attachment loss patterns in survivor teeth at 18 months, 36 months and 5 years in community-dwelling older adults. *Journal of Periodontal Research*. 1997a; 6:497–505. [PubMed: 9379317]
- Benhin E, Rao J, Scott A. Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*. 2005; 92:435–450.
- Chen Z, Zhang B, Albert PS. A joint modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine*. 2011; 30:1825–1836. [PubMed: 21495060]
- Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit specific outcomes. *Biometrics*. 2003; 59:521–530. [PubMed: 14601753]
- Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001; 88:1121–1134.
- Huang Y, Leroux B. Informative cluster sizes for sub-cluster level covariates and weighted generalized estimating equations. *Biometrics*. 2011; 67:843–851. [PubMed: 21281273]
- Liang KY, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Seaman SR, Pavlou M, Copas AJ. Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics*. 2014; 70:449–456. [PubMed: 24479899]
- Wang M, Kong M, Datta S. Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*. 2011; 20:347–367. [PubMed: 20223781]
- Williamson JM, Datta S, Satten G. Marginal analysis of clustered data when cluster size is informative. *Bio-metrics*. 2003; 59:36–42.
- Zeger SL, Liang KY, Albert PS. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*. 1988; 44:1049–1060. [PubMed: 3233245]

Figure 1. Snapshots of various intervals over a 20 year period. Recall that $E[Y] = 0$ and note that as time progresses the estimated mean (thick gray line), based on observed information gets further from that for the hypothetical complete distribution.

Figure 2.

Power plots obtained using the Wald statistic on point estimates obtained from Marginal WGEE analysis and the 500 Monte Carlo data generations from the indicated design with $M = 50$. The solid curve indicates the power curve associate with the marginal WGEE estimates obtained assuming independence working correlation, the dotted lines correspond to the AR-1 and exchangeable correlation structures.

Figure 3. Graphical representation of the available data. (a) Baseline cluster size of each of the 452 individuals retained. (b) Number of observations made on each of the 7823 retained teeth. (c) Total number of observations contributed by each individual.

Figure 4. Distribution of various demographic information stratified by race. All values have been centered according to the respective overall sample means.

Table 1

Summary of simulation results from the indicated design with $M = 500$ *Monte Carlo replicates* $Q = 50$ and 200. The mean of the parameters estimates, bias and variance are reported for the standard GEE with independent working correlation and for the marginal WGEE with various choices of subject specific working correlation structures.

	$Q = 50$				$Q = 200$			
	Standard GEE				Standard GEE			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Mean	0.873	1.503	0.782	0.804	0.862	1.502	0.833	0.807
Bias	0.373	0.003	0.032	0.004	0.362	0.002	0.083	0.007
Var	0.235	0.011	0.437	0.015	0.179	0.013	0.357	0.018
	MWGEE (Ind.)				MWGEE (Ind.)			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Mean	0.513	1.506	0.715	0.802	0.500	1.503	0.741	0.806
Bias	0.013	0.006	-0.035	0.002	0.000	0.003	-0.009	0.006
Var	0.449	0.118	0.615	0.141	0.233	0.064	0.317	0.076
	MWGEE (AR-1)				MWGEE (AR-1)			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Mean	0.520	1.500	0.718	0.800	0.504	1.500	0.748	0.800
Bias	0.020	<0.001	-0.032	<0.001	0.004	<0.001	-0.002	<0.001
Var	0.398	0.011	0.562	0.013	0.203	0.006	0.286	0.007
	MWGEE (Exch.)				MWGEE (Exch.)			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Mean	0.520	1.500	0.718	0.800	0.504	1.500	0.748	0.799
Bias	0.020	<0.001	-0.032	<0.001	0.004	<0.001	-0.002	<0.001
Var	0.398	0.009	0.562	0.011	0.203	0.005	0.288	0.005

Table 2

Aggregate results of racially stratified GEE, CWGEE and marginal WGEE analyses of AL. In each analysis an AR-1 working correlation was assumed. The effects reported are those significant according to a Wald test (covariance estimates obtained via sandwich and weighted sandwich) at the $\alpha = .05$ level.

White					Black				
Standard GEE					Standard GEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	7.686	0.560	188.2	<0.001	Intercept	6.018	0.393	234.4	<0.001
SEX	-1.138	0.379	9.0	0.003	SEX	-0.635	0.289	4.8	0.028
SES	-0.003	0.001	5.3	0.020	TOBACCO	-1.044	0.297	12.3	<0.001
TOBACCO	-1.351	0.380	12.6	<0.001					
URBAN	-0.982	0.360	7.4	0.006					

CWGEE					CWGEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	8.055	0.592	184.8	<0.001	Intercept	5.933	0.397	223.4	<0.001
SEX	-1.369	0.407	11.3	<0.001	TIME	0.152	0.077	3.9	0.048
SES	-0.003	0.001	5.7	0.016	TOBACCO	-0.969	0.294	10.9	0.001
TOBACCO	-1.571	0.399	15.5	<0.001	TIME $\times$ EDUCATION	-0.020	0.010	4.0	0.045
URBAN	-1.111	0.380	8.5	0.003					

Marginal WGEE					Marginal WGEE				
Effect	$\hat{\beta}$	std. err	Wald	p	Effect	$\hat{\beta}$	std. err	Wald	p
Intercept	8.857	0.665	177.5	<0.001	Intercept	6.188	0.476	169.0	<0.001
SEX	-1.439	0.531	7.3	0.007	TIME	0.211	0.082	6.6	0.010
SES	-0.003	0.002	3.9	0.047	TOBACCO	-0.863	0.383	5.0	0.024
TOBACCO	-1.490	0.512	8.4	0.004	TIME $\times$ EDUCATION	-0.023	0.011	4.8	0.027
URBAN	-1.610	0.531	9.1	0.002					