Published in final edited form as:

Stat Methods Med Res. 2018 June ; 27(6): 1806–1817. doi:10.1177/0962280216669184.

Inferring marginal association with paired and unpaired clustered data

Douglas J Lorenz¹, Steven Levy², and Somnath Datta³

¹Department of Bioinformatics and Biostatistics, School of Public Health and Information Science, University of Louisville, USA

²Department of Preventive and Community Dentistry, College of Dentistry and Dental Clinics, Department of Epidemiology, College of Public Health, University of Iowa, USA

³Department of Biostatistics, College of Public Health and Health Professions College of Medicine, University of Florida, USA

Abstract

In the marginal analysis of clustered data, where the marginal distribution of interest is that of a typical observation within a typical cluster, analysis by reweighting has been introduced as a useful tool for estimating parameters of these marginal distributions. Such reweighting methods have foundation in within-cluster resampling schemes that marginalize potential informativeness due to cluster size or within-cluster covariate distribution, to which reweighting methods are asymptotically equivalent. In this paper, we introduce a reweighting scheme for the marginal analysis of clustered data that generalizes prior reweighting methods, with a particular application to measuring bivariate correlation in unpaired clustered data, in which observations of two random variables are not naturally paired at the within-cluster level. We develop unpaired clustered data analogs of well-known product moment correlation coefficients (Pearson, Spearman, ϕ), as well as the polyserial coefficient for measuring correlation between one discrete and one continuous variable. We evaluate the performance of these coefficients via a simulation study and demonstrate their use by finding no statistically significant association between dental caries at an early age and dental fluorosis at age 13 using a large dental dataset.

Keywords

Measures of association; correlation; clustered data; marginal analysis; informative cluster size

Reprints and permissions information is available at www.nature.com/reprints Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Douglas J Lorenz, Department of Bioinformatics and Biostatistics University of Louisville, 485 E Gray St, Louisville, KY 40202. djlore01@louisville.edu.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

1 Introduction

The motivating dataset for this paper comes from an observational dental study,¹ with interest in examining the association between dental caries and dental fluorosis, the staining or pitting of enamel during tooth development caused by excess exposure to fluoride. As part of this study, children received dental examinations at ages 5, 9, and 13, and the presence and severity of dental caries were noted for each tooth in each child. Additionally, the presence and severity of dental fluorosis in each tooth was noted at ages 9 and 13. Investigators were interested in identifying associations between the presence of dental caries at age 5 and fluorosis in late-erupting teeth (canines, first and second premolars, second molars) at age 13. The hypothesis was that the protective measures taken by children diagnosed with caries at age 5—when late-erupting teeth are still in their sensitive, formative stages and have yet to erupt—would include the use and potential overuse of fluoride, placing the children at heightened risk of fluorosis in the late-erupting teeth at later ages.

These data provide an example of clustered data, in which observations are collected within units known as clusters. In marginal analyses of clustered data, clusters are typically assumed to be independent and the observations within clusters potentially dependent. Several types of marginal analysis are possible for clustered data, one of which treats the clusters as the primary sampling unit and the marginal distribution of a randomly selected observation from a randomly selected cluster is of interest. For the dental study, such a marginal analysis of the data from the dental study would consider the children as the clusters and the teeth as the observations within clusters.

The data from this study present several challenges in measuring the marginal association between caries at age 5 and fluorosis at age 13 through, say, a correlation analysis. The potential dependence of observations within clusters invalidates the use of standard methods measures based on independent and identically distributed (i.i.d.) theory. A more complex challenge arises from the interest in the marginal distribution of a randomly selected tooth from a randomly selected child. In particular, while it is clear that measurements at age 5 and age 13 should be paired at the cluster (child) level, it is not entirely clear how the observations within clusters (teeth) should be paired. The teeth of interest that will potentially exhibit fluorosis at age 13 have yet to erupt at age 5 and as such could not be measured for dental caries at age 5. Further, virtually all of the deciduous teeth measured for caries at age 5 are no longer present in the mouth at age 13, precluding an assessment of fluorosis.

A possible solution to this pairing problem would be to map deciduous teeth to permanent teeth by location and perform a marginal correlation analysis, but this strategy is biologically problematic. The hypothesized increase in fluoride usage for a child in which caries is detected at age 5 is not tooth-specific, i.e. the stimulus purported to increase the risk of fluorosis is applied to all teeth within a child, particularly those late-erupting teeth that could not be assessed for caries at age 5. By pairing teeth, even teeth present at both ages, information about the association between dental caries and fluorosis might be lost. For example, it is biologically plausible that a child with dental caries on a deciduous canine at age 5 could exhibit fluorosis in the left premolar at age 13, owing to increased fluoride

usage. In short, a marginal correlation analysis with a fixed pairing of observations at the tooth-within-child level does not correspond to the biological process under examination.

An additional problem in the marginal analysis of clustered data that has received recent attention is that of informative cluster size, in which the number of observations within a cluster is correlated with the random variable being measured. When the marginal analysis of interest concerns that of a randomly selected observation from a randomly selected cluster, standard marginal methods for clustered data can be biased under informative cluster size, as observations from large clusters will be overweighted. Hoffman et al.² introduced the within-cluster resampling (WCR) algorithm to mitigate the effects of informative cluster size. In the WCR procedure, an observation is randomly selected from each of the independent clusters to create a pseudo dataset on which methods based on i.i.d. theory can be applied. The process is repeated and the WCR estimator is defined as the average of the estimates produced from the pseudo datasets. Williamson et al.³ noted that, in the limit, the WCR process assigns to each observation a weight equal to the inverse of the cluster size from which the observation came. They introduced cluster-weighted generalized estimating equations (CWGEEs) as an extension of the standard generalized estimating equation (GEE) model. In CWGEEs, standard estimating equations are weighted by the inverse of the cluster size, resulting in estimators asymptotically equivalent to WCR estimators and negating the need for computationally intensive resampling. The cluster-weighting approach has been applied to define clustered data analogs of the Wilcoxon signed rank⁴ and rank sum test,⁵ estimators for survival data,^{6,7} and parametric and nonparametric correlation measures.⁸

An additional type of informativeness in clustered data is the informativeness of covariates at the sub-cluster level, in which the probability that a covariate takes a certain value within a given cluster can be informative for the outcome in question. Huang and Leroux⁹ extended cluster-weighted methods by considering this type of informativeness and introduced doubly weighted generalized estimating equations (DWGEEs). These DWGEEs are based on a two-step WCR procedure—from a given cluster, a particular value of the covariate is randomly chosen from the set of all possible covariate values, followed by the random selection of an observation from the cluster having the randomly selected value. The authors demonstrated that methods correcting for informative cluster size (CWGEEs) and methods correcting for informativeness in covariate distributions, such as inverse probability of treatment weighting,¹⁰ were each biased when both cluster size and sub-cluster covariate informativeness were present, while DWGEE estimators remained unbiased. A particular application of this method of reweighting occurs when the potentially informative covariate is a grouping factor and a comparison of a response among groups is desired, for which Dutta and Datta¹¹ developed a rank sum test.

These methods for the marginal analysis of clustered data each operate under a reweighting principle, where the weights are tailored to correct the potentially biasing effects of cluster size or covariate informativeness. In this paper, we suggest a reweighting principle that generalizes prior reweighting methods for the marginal analysis of clustered data. In Section 2, we propose our weighted GEEs and illustrate how they generalize the methods of Williamson et al.³ and Huang and Leroux.⁹ We extend the weighted GEEs to data unpaired at the within-cluster level and apply these principles to define several correlation estimators

for such data. Section 3 presents the results of a simulation study testing the validity of these correlation measures and the analysis of data from the dental study.¹ Section 4 presents our concluding remarks.

2 Methods

We seek to estimate the marginal association between two random variables, denoted X and Y , from data collected on M clusters, where the clustering mechanism is defined such that observations from different clusters are independent. Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ denote a vector of parameters derived from the marginal joint distribution of X and Y . In our applications, we will estimate a parameter of marginal association between X and Y that will be a smooth function of this vector of parameters, i.e. $g(\boldsymbol{\theta})$ for smooth function g .

2.1 Notation and preliminaries—paired clustered data

We begin by establishing notation and methods of weighted estimation for paired clustered data; notation and methods for data unpaired at the within-cluster level are detailed in Section 2.2. We observe bivariate data (X, Y) on a set of M clusters, with each cluster consisting of n_i observed pairs. Denote the total number of observations $N = \sum_{i=1}^M n_i$. We assume that the data consist of i.i.d. replicates of observations from each of the M clusters, denoted by the vector $\mathbf{V}$ with i th realization $\mathbf{V}_i = (n_i, (X_{i1}, Y_{i1}), \dots, (X_{in_i}, Y_{in_i}))$. This notation indicates that the size of each cluster, n_i , is a random variable potentially correlated with the bivariate random variable (X, Y) or functionals of the marginal bivariate distribution of (X, Y) .

Let $U_i(Y_{ij}, X_{ij}, \boldsymbol{\theta})$ be an estimating function in cluster i for the marginal parameter $\boldsymbol{\theta}$. To estimate the marginal parameter $\boldsymbol{\theta}$, we suggest the following weighted GEE

$$U(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^M \sum_{j=1}^{n_i} \omega_{ij} U_i(Y_{ij}, X_{ij}, \hat{\boldsymbol{\theta}}) = 0 \quad (1)$$

where the weights ω_{ij} are positive and subject to some natural constraint, e.g. $\sum_{j=1}^{n_i} \omega_{ij} = 1$, $\forall i$. While here we define the constraints per cluster, the weights can easily be normalized over the clusters as well without changing the methodology that follows, e.g.

$\sum_{i=1}^M \sum_{j=1}^{n_i} \omega_{ij} = 1$. In what follows, we demonstrate how equation (1) generalizes cluster-weighted³ and doubly weighted⁹ GEEs. Further, we show the importance of appropriate selection of the weights ω_{ij} in the marginal analysis of clustered data. As will be shown, the choice of weights is intimately related to the type of marginal analysis desired, as well as the nature of the relationship (if any) between the cluster size n_i , the outcome Y , and the marginal association between X and Y .

When the marginal analysis to be conducted is of a typical observation in the population of all observations, a popular choice for estimating the marginal association parameter $\boldsymbol{\theta}$ is the

standard GEE. In this case, θ is a functional of the distribution of a typical observation in the population of all possible observations

$$F_g(x, y) = E_V \left\{ \sum_{i=1}^M \sum_{j=1}^{n_i} N^{-1} I[X_{ij} \leq x, Y_{ij} \leq y] \right\}$$

where E_V indicates the expectation taken with respect to the distribution of V and $I[\cdot]$ is the indicator function. Observations contribute equal information regarding the relationship between X and Y , and thus are weighted equally by setting $\omega_{ij} = 1, \forall i, j$, in which case equation (1) yields the standard GEE with independent working correlation matrix

$$U(\hat{\theta}) = \sum_{i=1}^M \sum_{j=1}^{n_i} U_i(Y_{ij}, X_{ij}, \hat{\theta}) = 0 \quad (2)$$

When cluster size is non-informative, this equation provides an unbiased estimate of θ and any smooth functions thereof in the marginal analysis of a typical observation in the population of all observations.

A second type of marginal analysis for clustered data treats the cluster as the primary sampling unit, so that interest is in a typical observation from a typical cluster. As previously noted,^{2,3,9} for this type of marginal analysis, the standard GEE approach in equation (2) still yields an appropriate estimate of the marginal parameter θ when cluster size is non-informative. However, when cluster size is informative, the standard GEE will yield biased estimates of θ , favoring larger clusters. In equation (2), cluster i receives weight n_i/N and, as such, larger clusters are favored by receiving greater weight. Williamson et al.³ introduced cluster-weighted estimating equations by setting $\omega_{ij} = n_i^{-1}$

$$U(\hat{\theta}) = \sum_{i=1}^M \sum_{j=1}^{n_i} \frac{1}{n_i} U_i(Y_{ij}, X_{ij}, \hat{\theta}) = 0 \quad (3)$$

Under this approach, it is the clusters rather than the observations that receive equal weight ($\sum_j n_i^{-1} = 1, \forall i$), as observations from larger clusters are down-weighted. The marginal parameter θ is a functional of the marginal distribution of X and Y at the cluster level, defined as

$$F_c(x, y) = E_V \left\{ \sum_{i=1}^M \sum_{j=1}^{n_i} n_i^{-1} I[X_{ij} \leq x, Y_{ij} \leq y] \right\}$$

that is, the distribution of a typical observation from a typical cluster. A key distinction between estimation via estimating equations (2) and (3) is that the θ estimated by equation (2) is a marginal parameter in the population of observations, whereas the θ estimated by equation (3) is a marginal parameter in the population of clusters. While these parameters can be closely related or even coincide, they are often distinct, either by definition or through their relationship to the potentially informative cluster size.

When sub-cluster covariate informativeness is present, not only is the value of the covariate X related to the outcome Y but so is its within-cluster distribution. In this setting, estimation through standard GEEs or even cluster-weighted GEEs can lead to biased estimation of θ . For a simple example, consider the scenario in which Y is positively related to a binary variable X and that clusters with large values of Y tend to be large and are more likely to have $X=1$. Equation (2) will result in a biased estimate of θ in this scenario, given the overabundance of observations with large values of Y and $X=1$. Equation (3) will also produce a biased estimate of θ . Even though observations from larger clusters are down-weighted by the inverse cluster size, observations with large values of Y remain more likely to come from clusters in which observations with $X=1$ predominate. The DWGEEs of Huang and Leroux⁹ are obtained by defining the weights in equation (1) in terms of the sub-cluster covariate distribution

$$U(\hat{\theta}) = \sum_{i=1}^M \sum_{j=1}^{n_i} \frac{1}{n_{iX_{ij}}} U_i(Y_{ij}, X_{ij}, \hat{\theta}) = 0 \quad (4)$$

where $n_{iX_{ij}}$ represents the number of observations in cluster i taking value X_{ij} . For example, if X is categorical with K levels, then the i th cluster will be composed of n_{i1} observations from level 1, n_{i2} observations from level 2, ..., and n_{iK} observations from level K , so that $n_i = \sum_{k=1}^K n_{ik}$.

The distinct values of the covariate X define “sub-clusters,” which are equally weighted by estimating equation (4). In the K -level categorical X example, each level of X receives weight 1, so that each cluster receives weight K . When θ is estimated using the CWGEE (3), we note that the marginal analysis of interest is for a typical observation from a typical cluster. The same is true of the DWGEE (4). The distinction between CWGEEs and DWGEEs lies in the interpretation of “typical.” For CWGEEs, a typical X observation is defined by the observed distribution of X values. For DWGEEs, a typical X observation is defined so that each possible value for the covariate X is equally likely, i.e. the observed within-cluster distribution of X is marginalized. In this way, any bias in estimating θ that results from a relationship between Y and the within-cluster distribution of X is marginalized. The marginal distribution under consideration when estimating by DWGEEs is

$$F_d(x, y) = E_V \left\{ \sum_{i=1}^M \sum_{j=1}^{n_i} n_{iX_{ij}}^{-1} I[X_{ij} \leq x, Y_{ij} \leq y] \right\}$$

The variance of the estimator $\hat{\theta}$ obtained by solving equation (1) can be estimated in sandwich form by $\hat{\Sigma} = \mathcal{M}^{-1} \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1}$, where

$$\begin{aligned} \hat{\mathbf{A}} &= \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^{n_i} \omega_{ij} \frac{\partial \mathbf{U}_i(Y_{ij}, X_{ij}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \\ \hat{\mathbf{B}} &= \frac{1}{M} \sum_{i=1}^M \left\{ \sum_{j=1}^{n_i} \omega_{ij} \mathbf{U}_i(Y_{ij}, X_{ij}, \hat{\boldsymbol{\theta}}) \right\} \times \left\{ \sum_{j=1}^{n_i} \omega_{ij} \mathbf{U}_i(Y_{ij}, X_{ij}, \hat{\boldsymbol{\theta}}) \right\}^T \end{aligned}$$

This variance estimator corresponds with estimators established for both CWGEEs³ and DWGEEs.⁹ Next, we define marginal correlation coefficient estimators that are smooth functions, g , of the marginal parameter $\boldsymbol{\theta}$. The correlation estimators are defined simply as $g(\hat{\boldsymbol{\theta}})$, with variance estimators defined via the delta method as $\hat{\Sigma} = \mathcal{M}^{-1} \hat{\mathbf{I}}^T \hat{\Sigma} \hat{\mathbf{I}}$, where $\hat{\mathbf{I}}$ is the vector of first partial derivatives of g with respect to $\boldsymbol{\theta}$ evaluated at the estimated value of $\boldsymbol{\theta}$

$$\hat{\mathbf{I}} = \left(\frac{\partial g}{\partial \theta_1} \bigg|_{\theta_1 = \hat{\theta}_1}, \dots, \frac{\partial g}{\partial \theta_K} \bigg|_{\theta_K = \hat{\theta}_K} \right)$$

2.2 Unpaired clustered data

When data are unpaired at the within-cluster level, as in the dental data described in the introduction,¹ unconventional notation is required. We continue to denote the random variables in whose correlation we are interested as X and Y and the data in the i th cluster as $\mathbf{V}_i$. Now, we define

$$\mathbf{V}_i = (n_{iX}, n_{iY}, X_{i1}, \dots, X_{in_{iX}}, Y_{i1}, \dots, Y_{in_{iY}})$$

Several characteristics of data unpaired at the within-cluster level are implicit in this notation—that the number of X observations n_{iX} and the number of Y observations n_{iY} in each cluster are random, potentially unequal, and potentially correlated with, and thus informative for, X , Y , and any association between X and Y . Our interest remains in the marginal association between X and Y , defined as some function g of the marginal parameter $\boldsymbol{\theta}$ and estimating function $\mathbf{U}_i(X_{ij}, Y_{ij}, \boldsymbol{\theta})$, which require no changes from before. In what follows, we will refer to this type of data as unpaired clustered data.

In the introduction, we noted that both CWGEEs and DWGEEs for paired clustered data arise from limiting calculations applied to within-cluster resampling² schemes. For

CWGEEs, one bivariate observation per cluster is randomly selected to create the pseudo datasets on which the marginal parameter θ is estimated and averaged.³ Resampling and estimation proceed in the same way for DWGEEs, except that the random selection of an observation within each cluster is preceded by the random selection of a sub-cluster based on the discrete values of the informative covariate X .⁹ A within-cluster resampling scheme for unpaired clustered data involves randomly and independently selecting one X and one Y observation from each cluster to build the pseudo datasets on which θ is estimated and averaged. Such a sampling scheme is in accordance with the characteristics of unpaired clustered data, in which X and Y observations are not naturally paired and all possible pairings of X and Y observations from a given cluster are equally valid. A limiting calculation applied to this resampling scheme produces the following estimating equation for unpaired clustered data

$$\mathbf{U}(\hat{\theta}) = \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} \mathbf{U}_i(Y_{ij_y}, X_{ij_x}, \hat{\theta}) = 0 \quad (5)$$

In equation (5), each possible pairing of X and Y observations is represented and assigned a weight equal to the inverse of the number of possible X – Y pairings in each cluster ($n_{ix}n_{iy}$). Thus, estimators for unpaired clustered data arise in a similar way as from the weighted estimating equation (1), although the notation there does not accommodate the notation for unpaired clustered data. Further, if either or both of the sizes of the X and Y data in a given cluster are informative for θ , then estimation via equation (5) will remain valid. Like CWGEEs and DWGEEs, the marginal analysis of interest here still concerns a typical observation from a typical cluster. Because the data are unpaired at the within-cluster level, a typical observation is characterized as any pairing of a single X and a single Y observation from a particular cluster, which has the marginal distribution

$$F_u(x, y) = E_{\mathbf{V}} \left\{ \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} I[X_{ij_x} \leq x, Y_{ij_y} \leq y] \right\}$$

The variance estimator $\hat{\Sigma}$ takes the same form previously shown, with

$$\begin{aligned} \hat{\mathbf{A}} &= \frac{1}{M} \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} \left. \frac{\partial \mathbf{U}_i(Y_{ij_y}, X_{ij_x}, \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}} \\ \hat{\mathbf{B}} &= \frac{1}{M} \sum_{i=1}^M \left\{ \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} \mathbf{U}_i(Y_{ij_y}, X_{ij_x}, \hat{\theta}) \right\} \times \left\{ \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} \mathbf{U}_i(Y_{ij_y}, X_{ij_x}, \hat{\theta}) \right\}^T \end{aligned}$$

The delta method calculations for estimating smooth functions of θ noted previously remain unchanged in the unpaired clustered data framework.

As an example of estimation for unpaired clustered data, consider estimation of the marginal moment $m_{kl} = E[X^k Y^l]$. The estimating function for $\theta = m_{kl}$ is $U_i(Y_{ij}, X_{ij}, \theta) = X_{ij}^k Y_{ij}^l - m_{kl}$.

By solving equation (5), we obtain the estimator

$$\hat{m}_{kl} = \frac{1}{M} \sum_{i=1}^M \frac{1}{n_{ix} n_{iy}} \sum_{j_y=1}^{n_{iy}} \sum_{j_x=1}^{n_{ix}} X_{ij_x}^k Y_{ij_x}^l \quad (6)$$

When either k or l is 0 and thus a marginal moment of X ($l=0$) or a marginal moment of Y ($k=0$) is being estimated, the interior sum in equation (6) corresponding to the zero exponent drops out, and the moment estimator corresponds to the CWGEE estimator. For example, in estimating $m_{20} = E[X^2]$ for unpaired clustered data, we find that

$$\hat{m}_{20} = M^{-1} \sum_{i=1}^M n_{ix}^{-1} \sum_{j_x=1}^{n_{ix}} X_{ij_x}^2, \text{ since } \sum_{j_y=1}^{n_{iy}} n_{iy}^{-1} = 1. \text{ We will now use these moment}$$

estimators to motivate the estimation of product moment correlations detailed in Section 2.3.

2.3 Applications to correlation analysis for unpaired data

Product moment correlations are functions of five moments of varying type from the bivariate distribution of X and Y . Letting the vector $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5)$ represent the five moments, the familiar product moment formula is

$$g(\mathbf{z}) = \frac{z_3 - z_1 z_2}{\sqrt{(z_4 - z_1^2)(z_5 - z_2^2)}} \quad (7)$$

The Pearson correlation is the most well-known product moment correlation, for which g in equation (7) is evaluated at the vector of the first five raw moments of the bivariate distribution of X and Y : $\rho = g(m_{10}, m_{01}, m_{11}, m_{20}, m_{02})$. The estimator is formed by applying the product moment formula to the first five raw sample moments, $\hat{\rho} = g(\hat{m}_{10}, \hat{m}_{01}, \hat{m}_{11}, \hat{m}_{20}, \hat{m}_{02})$, where $\hat{m}_{kl}$ is as defined in Section 2.2. We briefly note that the correlation coefficient ρ is not directly estimated as the solution of the weighted estimating equation (1). However, ρ is a smooth function of the first five raw moments of X and Y , estimators of which can be obtained as solutions of equation (1).

The Spearman correlation coefficient is a popular alternative to the Pearson coefficient, and is typically employed to detect nonlinear association or relationships between discrete or skewed continuous random variables. The Spearman coefficient is also a product moment correlation, but is calculated on the first five “rank moments” of the bivariate distribution of X and Y . These rank moments are not defined as solutions to the weighted estimating equation (1), in contrast with the Pearson coefficient. Therefore, we must define a cluster-

weighted rank function to define the rank moments that will comprise a cluster-weighted Spearman coefficient. As previously noted,^{4,5} informative cluster size invalidates standard formulas for rank functions such as $\text{rank}(X_{i'j'_x}) = \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} I[X_{ij_x} \leq X_{i'j'_x}]$. We define the rank of X_{ij_x} among all X observations as $R_{X_{ij_x}} = \frac{1}{2}(\hat{F}_X(X_{ij_x}) + \hat{F}_X(X_{ij_x} -))$, where $\hat{F}_X(x) = \frac{1}{M} \sum_{i=1}^M F_{X_i}(x)$, and $F_{X_i}(x) = \sum_{j=1}^{n_{ix}} \frac{1}{n_{ix}} I[X_{ij} \leq x]$. Ranks for the Y data, $R_{Y_{ij_y}}$, are defined analogously, with corresponding functions $\hat{F}_Y(y)$, and $\hat{F}_Y(y)$. These cluster-weighted rank functions have been previously used to define rank sum⁵ and signed rank⁴ tests for clustered data, and we note that tied observations are accounted through a midrank calculation implied by the formulas. The cluster-weighted rank moments for unpaired clustered data are then defined as

$$\hat{r}_{kl} = \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} R_{X_{ij_x}}^k R_{Y_{ij_y}}^l$$

and the marginal Spearman correlation for unpaired clustered data can be calculated as $\hat{\rho}_s = g(\hat{F}_{10}, \hat{F}_{01}, \hat{F}_{11}, \hat{F}_{20}, \hat{F}_{02})$.

When X and Y are both binary random variables taking values 0 or 1, equation (7) yields the phi coefficient, a commonly used measure of association for binary data. The phi coefficient also has an alternative formulation in terms of cell counts from a 2×2 contingency table. Much like the rank functions, the cell counts from 2×2 tables calculated on clustered data with potentially informative cluster size require weighting. The cluster-weighted cell counts for unpaired clustered data are

$$\hat{n}_{kl} = \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \frac{1}{n_{ix}n_{iy}} I[X_{ij_x} = k, Y_{ij_y} = l]$$

where k and l take values 0 or 1. The cluster-weighted marginal totals for the 2×2 table can be calculated as $\hat{n}_{1.} = \hat{n}_{11} + \hat{n}_{10}$, $\hat{n}_{0.} = \hat{n}_{01} + \hat{n}_{00}$, $\hat{n}_{.1} = \hat{n}_{11} + \hat{n}_{01}$, and $\hat{n}_{.0} = \hat{n}_{10} + \hat{n}_{00}$. In terms of these weighted cell counts, the alternative formulation for the cluster-weighted phi coefficient is

$$\hat{\phi} = \frac{\hat{n}_{11}\hat{n}_{00} - \hat{n}_{10}\hat{n}_{01}}{\sqrt{\hat{n}_{1.}\hat{n}_{.1}\hat{n}_{0.}\hat{n}_{.0}}}$$

The delta method calculation described previously can be employed to derive variance estimators for each of the three product moment correlation coefficients. Because the five cluster-weighted moments used in the product moment correlation formula are

asymptotically normal^{2,3} and the product moment correlations are a smooth function of the cluster-weighted moments, the asymptotic distributions of the Pearson, Spearman, and phi coefficients are normal. We use this fact to generate confidence intervals for each of the product moment coefficients, the performance of which will be evaluated in Section 3.

The polyserial correlation¹² is commonly used for bivariate data in which one variable is continuous and one discrete. We now assume X to be continuous and Y discrete taking values $1, \dots, K$. It is assumed that values of Y arise from a latent random variable $U \sim N(0, 1)$ according to the formula

$$P[Y = k] = P[\xi_{k-1} \leq U < \xi_k], \quad k = 1, \dots, K-1$$

where ξ_k are partition points of the real line, $\xi_0 = -\infty$, and $\xi_K = \infty$. The pair (X, U) is assumed to be bivariate normal and, for brevity of presentation, we assume X to be standard normal as well. When X follows a non-standard normal distribution, its mean and variance can be estimated using the cluster-weighted moment estimators described in Section 2.2.

In developing the biserial coefficient, Cox¹² defined the estimating equations for the latent correlation ρ and the partition points ξ_k . Cluster-weighted versions of these estimating equations for unpaired clustered data are

$$\begin{aligned} U_i(\rho) &= \frac{1}{(1-\rho^2)^{3/2}} \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \sum_{k=1}^K \frac{I[Y_{ij_y} = k]}{n_{ix}n_{iy}} \times \left\{ \rho \frac{\xi_k \phi_k(X_{ij_x}) - \xi_{k-1} \phi_{k-1}(X_{ij_x})}{\Phi_k(X_{ij_x}) - \Phi_{k-1}(X_{ij_x})} - \frac{\phi_k(X_{ij_x}) - \phi_{k-1}(X_{ij_x})}{\Phi_k(X_{ij_x}) - \Phi_{k-1}(X_{ij_x})} \right\} \\ U_i(\xi_r) &= \frac{1}{(1-\rho^2)^{1/2}} \sum_{i=1}^M \sum_{j_x=1}^{n_{ix}} \sum_{j_y=1}^{n_{iy}} \sum_{k=1}^K \frac{I[Y_{ij_y} = k]}{n_{ix}n_{iy}} \times \left\{ \frac{\phi_k(X_{ij_x}) I[k=r]}{\Phi_k(X_{ij_x}) - \Phi_{k-1}(X_{ij_x})} - \frac{\phi_{k-1}(X_{ij_x}) I[k-1=r]}{\Phi_k(X_{ij_x}) - \Phi_{k-1}(X_{ij_x})} \right\} \end{aligned} \quad (8)$$

where $\Phi_y(x) = \Phi((\xi_y - \rho x)/\sqrt{1-\rho^2})$, $\phi_y(x) = \phi((\xi_y - \rho x)/\sqrt{1-\rho^2})$, ϕ and Φ denote the standard normal density and distribution functions, $\phi_0(x) = \phi_K(x) = 0$, $\Phi_0(x) = 0$, and $\Phi_K = 1$.

The estimates $\hat{\rho}$ and $\hat{\xi}_r$ are obtained by solving equation (8) numerically. Starting values for the vector $(\rho, \xi_1, \dots, \xi_{K-1})^T$ are obtained as cluster-weighted analogs of the starting estimates provided by Cox¹² for i.i.d. data. Let $\pi_k = M^{-1} \sum_{i=1}^M n_{iy}^{-1} \sum_{j_y=1}^{n_{iy}} I[Y_{ij_y} = k]$ be the cluster-weighted proportion of Y observations taking value k . Then the starting values are given by

$$\xi_r^* = \Phi^{-1} \left(\sum_{k=1}^r \pi_k \right)$$

$$\rho^* = \frac{s_y^r xy}{\sum_{k=1}^{K-1} \phi(\Phi^{-1}(\pi_k))}$$

in which we have that

$$s_y^2 = M^{-1} \sum_{i=1}^M n_{iy}^{-1} \sum_{j_y=1}^{n_{iy}} Y_{ij_y}^2 - \left(M^{-1} \sum_{i=1}^M n_{iy}^{-1} \sum_{j_y=1}^{n_{iy}} Y_{ij_y} \right)^2$$

a cluster-weighted variance estimate for Y , and r_{xy} is the cluster-weighted product moment correlation of (X, Y) defined earlier in this section. Finally, the variance-covariance matrix of $(\hat{\rho}, \hat{\theta}_1, \dots, \hat{\theta}_{K-1})^T$ can be estimated using the sandwich formula noted in Sections 2.1 and 2.2.

3 Results

3.1 Simulation study

We evaluated the correlation estimators for unpaired data from Section 2.3 with a simple simulation study. We simulated data under an unpaired “multivariate random effects” structure, the details of which follow. For each cluster, we first simulated the random effects for X and Y as the bivariate normal pair

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix} \right)$$

We then independently simulated the cluster sizes n_{ix} and n_{iy} in each cluster as

$$n_{ix}, n_{iy} \sim \begin{cases} \text{BIN}(n_1, 9) & \text{if } u_i v_i > 0 \\ \text{BIN}(n_2, 9) & \text{otherwise} \end{cases}$$

where $(n_1, n_2) = (20, 10)$ in one set of simulations and $(n_1, n_2) = (10, 20)$ in another. Using these cluster sizes, we then simulated the “model errors” for each cluster as the vector

$$\left(\mu_{i1}, \dots, \mu_{in_{ix}}, \nu_{i1}, \dots, \nu_{in_{iy}} \right) \sim N \left(\mathbf{0}, \begin{pmatrix} \Delta_x & \mathbf{P}^T \\ \mathbf{P} & \Delta_y \end{pmatrix} \right)$$

where $\mathbf{0}$ denotes a vector of $n_{ix}+n_{iy}$ zeros, Σ_x the $n_{ix} \times n_{ix}$ compound symmetry matrix with correlation parameter δ_x , Σ_y the $n_{iy} \times n_{iy}$ compound symmetry matrix with correlation parameter δ_y , and $\mathbf{P}$ the $n_{iy} \times n_{ix}$ matrix with every element equal to ρ . The data in cluster i were then generated as the sum of the random effects and model errors

$$\mathbf{V}_i = \left\{ n_{ix}, n_{iy}, X_{i1}, \dots, X_{in_{ix}}, Y_{i1}, \dots, Y_{in_{iy}} \right\}$$

$$= \left\{ n_{ix}, n_{iy}, (u_i + \mu_{i1}), \dots, (u_i + \mu_{in_{ix}}), (v_i + \nu_{i1}), \dots, (v_i + \nu_{in_{iy}}) \right\}$$

Data simulated under this model have several important features. Under the compound symmetry structure, $\text{Corr}(X_{ij1}, X_{ij2}) = \delta_x$ and $\text{Corr}(Y_{ij1}, Y_{ij2}) = \delta_y$. Since the random effects u_i and v_i were generated independently of the model errors μ_{ijx} and ν_{ijy} , we have that $\text{Var}(X) = \text{Var}(Y) = 2$ and $\text{Cov}(X, Y) = \rho + \gamma$ and thus $\text{Corr}(X_{ij1}, Y_{ij2}) = (\rho + \gamma)/2$. Of particular importance to note is that the cluster sizes n_{ix} and n_{iy} were informative for the correlation between X and Y . Under the setting in which $(n_1, n_2) = (20, 10)$, clusters with concordant (discordant) random effects u_i and v_i were simulated from a binomial distribution with size 20 (10). Thus, clusters with concordant random effects—clusters with evidence of a positive correlation—tended to be larger in size than clusters with discordant random effects. The reverse was true when $(n_1, n_2) = (10, 20)$ —clusters with discordant random effects tended to be larger in size.

We evaluated each of our correlation estimators for $\rho = -0.5, 0.2, 0.7$ and $\gamma = -0.4, 0, 0.4$. The compound symmetry correlation parameters for the X and Y data were set as $\delta_x = \delta_y = 0.7$. To create binary data for evaluating the phi coefficient, we dichotomized X as $\mathbb{I}[X > 0.5]$ and Y as $\mathbb{I}[Y > 0.5]$. To create discrete data for testing the biserial correlation, we discretized Y into a four-level categorical variable using $-1.2, -0.1$, and 1.3 as cut points. For each of the design settings, we calculated the average of our estimators over 10,000 Monte Carlo loops and defined the number of clusters to be $M=100$.

Our proposed correlation estimators for unpaired clustered data were unbiased and the asymptotic confidence intervals exhibited close to nominal coverage probabilities (Table 1). The estimators were resistant to the potentially biasing effect of informative cluster size—the estimators were approximately unbiased when clusters with concordant random effects were larger and when clusters with discordant random effects were larger. Coverage probabilities were slightly lower but generally in correspondence with the nominal 95% level. This phenomenon corresponded with the same well-known phenomenon for asymptotic confidence intervals for product moment correlations for i.i.d. data.

3.2 Application to dental data

To illustrate the use of the unpaired correlation estimators, we analyzed the dental dataset described in the introduction.¹ At ages 5 and 13, each subject was evaluated for caries and fluorosis, respectively. An ordinal fluorosis score was assigned for each tooth of each enrolled subject using the Fluorosis Risk Index,¹³ with 0 indicating none, 1 indicating questionable fluorosis, 2 indicating definitive white striations, and 3 indicating staining or

pitting of the tooth. Additionally, an alphabetized, ordinal dental caries score was assigned for each tooth of each enrolled subject, with “S” indicating a sound zone (no caries), “D0”, “D1”, and “D2” indicating increasing severity of caries, and “F” indicating a tooth with filled caries. We briefly note that both fluorosis and caries were scored on several zones for each tooth; to simplify our analyses, we considered the maximum score across all zones to be the fluorosis or caries score for a given tooth.

There were 525 subjects in the study with assessments of dental caries at age 5 and assessments of fluorosis at age 13, representing 10,363 teeth at age 5 and 14,001 teeth at age 13. Here, we restrict our analysis to the measurements taken on 7,732 late-erupting teeth at age 13 (canines, premolars, and second molars). Table 2 provides some descriptive statistics for these 525 individuals. The majority of subjects had full complements of teeth at age 5 (465 of 525, 89%) and full complements of late-erupting teeth at age 13 (317 of 525, 60%), but variation in the number of teeth per individual was present.

Many subjects (36%) exhibited caries on at least one tooth at age 5, defined here as a caries score greater than “S”. There was some evidence of potentially informative cluster size, as both the proportion of individuals with at least one tooth with caries and the mean proportion of teeth with caries per subject appeared to vary by the number of teeth. Some subjects (26%) exhibited fluorosis on at least one tooth at age 13, defined here (and generally) as a fluorosis score greater than 1. The proportion of individuals with at least one tooth exhibiting fluorosis varied over subjects with different tooth counts, as did proportion of teeth with fluorosis. Two distinct, simplified, subject-level marginal analyses indicated that caries at age 5 and fluorosis at age 13 are not related. The proportion of subjects with fluorosis at age 13 was fairly consistent among those with caries at age 5 (45 of 189, 23%) and without caries at age 5 (92 of 336, 27%). Further, the Pearson correlation between subject-averaged ordinal caries scores at age 5 and subject-averaged Fluorosis Risk Index measurements at age 13 was low (-0.08 , 95% CI $[-0.17, 0.00]$).

We examined the caries–fluorosis relationship at the tooth-within-subject level by calculating the unpaired Spearman coefficient for the caries at age 5 score and the fluorosis at age 13 score. We additionally calculated phi coefficients based on various dichotomizations of the caries and fluorosis scores. The use of unpaired coefficients was in part justified by the presence of different teeth at ages 5 and 13, as noted in the introduction—no permanent teeth were present in any subject at age 5 and only late-erupting permanent teeth were of interest at age 13. Our marginal analysis showed no caries–fluorosis association (Table 3), as all estimated coefficients were close to zero for all dichotomizations of the fluorosis and caries scores.

4 Discussion

In this paper, we have suggested a generalized weighted estimating equation for the marginal analysis of clustered data in which the cluster is the primary sampling unit and the marginal distribution of a typical observation from a typical cluster is of interest. The estimating equations that we suggest here are straightforward extensions of standard GEEs, and generalize previously defined extended GEEs designed to adjust for the potentially biasing

effects of informative cluster size and sub-cluster covariate informativeness.^{3,9} We used these estimating equations to develop correlation estimators for unpaired clustered data, in which observations of two random variables to be associated are paired at the cluster level but unpaired at the observation-within-cluster level. The simulation study in Section 3.1 demonstrated the unbiasedness of our correlation estimators for unpaired clustered data and the approximate correct coverage of the associated asymptotic 95% confidence intervals. We were able to use these correlation estimators to show that dental caries at a younger age and fluorosis at an older age were not associated in a large sample of children enrolled in a dental study.

A potential alternative option in the marginal analysis of clustered data is a model-based approach in which all sources of variability are specified.¹⁴ However, a similar approach for bivariate data under informative cluster size has yet to be developed. In particular, the joint model approach must also include a component for the cluster size, and the performance of the association measures under mis-specification of this component needs to be studied.

We have shown examples of weights ω_{ij} to be used in estimating equation (1), and each was associated with a resampling scheme designed to adjust for potentially problematic features of clustered data. Inverse cluster size weights ($\omega_{ij} = n_i^{-1}$) were employed to adjust for informative cluster size.³ Inverse sub-cluster covariate weights ($\omega_{ij} = n_{ix_{ij}}^{-1}$) were employed to adjust for sub-cluster covariate informativeness.⁹ We introduced variable-specific inverse cluster size weights ($\omega_{ijxjy} = (n_{ix}n_{iy})^{-1}$) for unpaired clustered data with potentially informative cluster size. We note that the sub-cluster covariate weighting approach can be integrated into this by defining $\omega_{ijxjy} = (n_{ix_{ij}}n_{iy_{ij}})^{-1}$. We did not explore this approach here, as sub-cluster covariate informativeness did not appear to be a feature of the dental dataset. An important feature of estimating equation (1) is that it is not limited to such resampling-based weights. For example, in our dental data application, it is possible that proximity plays a role, in that fluorosis may occur by age 13 in teeth nearby a tooth with caries at age 5. In this scenario, the weights ω_{ijxjy} can be defined as a function of a proximity mapping to accommodate association of teeth by proximity, e.g., ω_{ijxjy} can be defined as a function of $I[X_{ijx} \equiv Y_{ijy}]$, where $\equiv$ indicates that X_{ijx} and Y_{ijy} come from teeth defined to be in proximity to each other.

While the customization of the weights in equation (1) provides flexibility in the marginal analysis of clustered data, care needs to be taken in selecting appropriate weights and in noting what equation (1) estimates for a given weight. For example, consider a version of the doubly weighted GEEs proposed by Huang and Leroux⁹ for scenarios in which not all possible values of the potentially informative covariate X are observed in each cluster, termed by the authors DWGEE2. Seaman et al.¹⁵ showed that DWGEE2 estimates a parameter in a population in which all clusters exhibit all possible values of X . In other words, it is assumed that the observed data come from a population in which each value of X is represented in each cluster. Among other criticisms, Seaman et al.¹⁵ point out that this can be philosophically problematic, in that such a population might be purely hypothetical. For example, in the context of dental studies, one could find oneself modeling the dental hygiene

of individuals with no teeth. As such, careful consideration of the weights is important, particularly with regard to what is being estimated for a particularly specified weight.

Acknowledgments

We thank an anonymous reviewer for many useful comments that improved this manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Institutes of Health (grant numbers 1R03DE020839, 1R03DE022538, R01-DE09551, R01-DE12101, M01-RR00059, UL1-RR024979).

References

1. Levy SM, Warren JJ, Broffitt BA, et al. Fluoride, beverages and dental caries in the primary dentition. *Caries Res.* 2003; 37:157–165. [PubMed: 12740537]
2. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika.* 2001; 88:1121–1134.
3. Williamson JM, Datta S, Satten GA. Marginal analysis of clustered data when cluster size is informative. *Biometrics.* 2003; 59:36–42. [PubMed: 12762439]
4. Datta S, Satten GA. A signed-rank test for clustered data. *Biometrics.* 2008; 65:501–507.
5. Datta S, Satten GA. Rank-sum tests for clustered data. *J Am Stat Assoc.* 2005; 100:908–915.
6. Cong XJ, Yin G, Shen Y. Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics.* 2007; 63:663–672. [PubMed: 17825000]
7. Williamson JM, Kim HY, Manatunga A, et al. Modeling survival data with informative cluster size. *Stat Med.* 2008; 27:543–555. [PubMed: 17640035]
8. Lorenz DJ, Datta S, Harkema SJ. Marginal association measures for clustered data. *Stat Med.* 2011; 30:3181–3191. [PubMed: 21953204]
9. Huang Y, Leroux B. Informative cluster sizes for sub-cluster level covariates and weighted generalized estimating equations. *Biometrics.* 2011; 67:843–851. [PubMed: 21281273]
10. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000; 11:550–560. [PubMed: 10955408]
11. Dutta S, Datta S. A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics.* 2016; 72:432–440. [PubMed: 26575695]
12. Cox NR. Estimation of the correlation between a continuous and a discrete variable. *Biometrics.* 1972; 30:171–178.
13. Pendrys DG. The Fluorosis Risk Index: a method for investigating risk factors. *J Public Health Dent.* 1990; 50:291–298. [PubMed: 2231522]
14. Chen Z, Zhang B, Albert PS. A joint modeling approach to data with informative cluster size: robustness to the cluster size model. *Stat Med.* 2011; 30:1825–1836. [PubMed: 21495060]
15. 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]

Table 1

Simulation results. For each setting of the design parameters, “True” provides the true value of the coefficient, “Est.” lists the average of the 10,000 Monte Carlo estimates, and “Cov.” lists the empirical coverage probability of the asymptotic 95% confidence interval.

Design parameters		Pearson			Spearman			Phi			Biserial		
(n_1, n_2)	ρ	γ	True	Est.	Cov.	True	Est.	Cov.	True	Est.	True	Est.	Cov.
(20, 10)	-0.5	-0.4	-0.45	-0.45	0.9322	-0.43	-0.43	0.9309	-0.27	-0.27	0.9466	-0.45	0.9385
		0.0	-0.25	-0.25	0.9381	-0.24	-0.24	0.9442	-0.14	-0.15	0.9399	-0.25	0.9438
		0.4	-0.05	-0.05	0.9362	-0.05	-0.05	0.9408	-0.03	-0.03	0.9452	-0.05	0.9408
		0.2	-0.4	-0.10	0.9298	-0.10	-0.10	0.9364	-0.06	-0.06	0.9412	-0.10	0.9398
	0.0	0.0	0.10	0.10	0.9358	0.10	0.10	0.9392	0.06	0.06	0.9441	0.10	0.9369
		0.4	0.30	0.30	0.9357	0.29	0.29	0.9424	0.19	0.19	0.9449	0.30	0.9455
		-0.4	0.15	0.15	0.9344	0.14	0.14	0.9382	0.09	0.09	0.9396	0.15	0.9501
		0.0	0.35	0.35	0.9362	0.34	0.33	0.9403	0.22	0.22	0.9448	0.35	0.9389
	0.7	0.4	0.55	0.55	0.9391	0.53	0.53	0.9361	0.36	0.36	0.9436	0.55	0.9402
		-0.4	-0.45	-0.45	0.9376	-0.43	-0.43	0.9385	-0.27	-0.27	0.9448	-0.45	0.9421
		0.0	-0.25	-0.25	0.9336	-0.24	-0.24	0.9395	-0.14	-0.15	0.9358	-0.25	0.9439
		0.4	-0.05	-0.05	0.9355	-0.05	-0.05	0.9431	-0.03	-0.03	0.9450	-0.05	0.9388
(10, 20)	-0.5	-0.4	-0.10	-0.10	0.9364	-0.10	-0.10	0.9432	-0.06	-0.06	0.9469	-0.10	0.9452
		0.0	0.10	0.10	0.9381	0.10	0.09	0.9431	0.06	0.06	0.9428	0.10	0.9411
		0.4	0.30	0.30	0.9339	0.29	0.29	0.9375	0.19	0.19	0.9451	0.30	0.9397
		-0.4	0.15	0.15	0.9346	0.14	0.14	0.9393	0.09	0.09	0.9442	0.15	0.9423
	0.0	0.0	0.35	0.35	0.9377	0.34	0.33	0.9417	0.22	0.22	0.9435	0.35	0.9452
		0.4	0.55	0.55	0.9417	0.53	0.53	0.9382	0.36	0.36	0.9476	0.55	0.9361

Table 2
Descriptive statistics for the $M=525$ subjects with assessments of caries at age 5 and fluorosis at age 13.

Age 5						Age 13			
No. of teeth	No. of subjects	At least 1 with caries	Mean proportion of teeth with caries	No. of teeth		No. of subjects	At least 1 with fluorosis	Mean proportion of teeth with fluorosis	
17	13	0.54	0.16	11	41	0.29	0.15		
18	31	0.45	0.14	12	36	0.25	0.09		
19	16	0.38	0.08	13	22	0.23	0.14		
20	465	0.35	0.06	14	54	0.20	0.07		
				15	55	0.27	0.08		
				16	317	0.27	0.12		
Total	525	0.36	0.07	Total	525	0.26	0.11		

Table 3

Estimated unpaired correlation coefficients (95% confidence interval) examining the relationship between dental caries at age 5 and fluorosis at age 13. Phi coefficients were calculated under the indicated dichotomizations of the dental caries and fluorosis scores.

Coefficient	Age 5 caries	Age 13 fluorosis	Coefficient (95% CI)
Spearman			−0.04 (−0.06, −0.01)
Phi	D0, D1, D2, F	1, 2, 3	−0.03 (−0.06, 0.00)
	D0, D1, D2, F	2, 3	−0.03 (−0.05, −0.01)
	D1, D2, F	1, 2, 3	−0.03 (−0.06, 0.00)
	D1, D2, F	2, 3	−0.03 (−0.05, −0.01)