Published in final edited form as:

Stat Med. 2011 November 30; 30(27): 3181–3191. doi:10.1002/sim.4368.

Marginal association measures for clustered data†

Douglas J. Lorenz^{a,*†}, Somnath Datta^a, and Susan J. Harkema^{b,c}

^aDepartment of Bioinformatics and Biostatistics, School of Public Health and Information Science, University of Louisville, Louisville, KY, USA

^bDepartment of Neurological Surgery, University of Louisville, Louisville, KY, USA

^cFrazier Rehab Institute, Louisville, KY, USA

Abstract

The use of correlation coefficients in measuring the association between two continuous variables is common, but regular methods of calculating correlations have not been extended to the clustered data framework. For clustered data in which observations within a cluster may be correlated, regular inferential procedures for calculating marginal association between two variables can be biased. This is particularly true for data in which the number of observations in a given cluster is informative for the association being measured. In this paper, we apply the principle of inverse cluster size reweighting to develop estimators of marginal correlation that remain valid in the clustered data framework when cluster size is informative for the correlation being measured. These correlations are derived as analogs to regular correlation estimators for continuous, independent data, namely, Pearson's ρ and Kendall's τ . We present the results of a simple simulation study demonstrating the appropriateness of our proposed estimators and the inherent bias of other inferential procedures for clustered data. We illustrate their use through an application to data from patients with incomplete spinal cord injury in the USA.

Keywords

measures of correlation; marginal analysis; clustered observations; informative cluster size

1. Introduction

Correlation analyses are widely used in the measurement of association between two random variables. There are many varieties of correlation coefficients, with the Pearson product-moment coefficient likely being the most familiar and the rank-based Spearman coefficient and Kendall concordance coefficient serving as nonparametric alternatives. Inferential calculations for Pearson's ρ , Kendall's τ , and other correlation coefficients rest on the assumption of independence of the observations. In this paper, we consider the measurement of association in the presence of clustered data, for which observations within a given cluster may be dependent.

†The code used for the simulation study and application to real data in this paper is available upon request to the corresponding author, or can be downloaded from <http://www.somnathdatta.org/software/>.

Copyright © 2011 John Wiley & Sons, Ltd.

*Correspondence to: Douglas J. Lorenz, Department of Bioinformatics and Biostatistics, School of Public Health and Information Science, University of Louisville, Louisville, KY 40292, USA.

†djlore01@louisville.edu

Our motivating example is a data set of 323 patients with incomplete spinal cord injury (SCI) participating in a standardized activity-based rehabilitation program [1]. As treatment progresses, patients receive periodic evaluations, during which a battery of functional measurements are taken. Patients are discharged from the program when improvement in the functional outcomes reaches a plateau and further functional gains seem unlikely. This data set provides an example of clustered data—the patients form the clusters, and the repeated functional evaluations form the (possibly dependent) observations per cluster. Clustered data need not arise from studies with repeated measurements on individuals. Additional examples include carcinogen exposure among sets of families or photosensitivity of litter mates in rodent experiments where, respectively, the families and litters are the clusters and the individuals, and rodents are the observations within cluster.

In our motivating example, one may be interested in measuring the marginal association between two measures of functional ability. In this setting, several types of marginal analyses for clustered data are possible. An observation-based approach would consider the association between two functional measures in the population of all cluster members. A cluster-based approach would consider the association between two functional measures for typical members of typical clusters. As has been previously noted [2, 3], these types of marginal analyses reach similar conclusions when the number of observations per cluster is unrelated to the outcome of interest but are characteristically different when the size of the cluster is related to the outcome. This phenomenon is referred to as informative cluster size, and we exemplify it presently in the context of our SCI example. More severely impaired patients with low functional capacity at enrollment tend to remain in the rehabilitation program longer, as there is more room for improvement in the functional outcomes being measured, and discharge from the program is based on functional progress. Hence, patients of lower function at enrollment contribute more observations to the data set and would receive greater weight if correlations were estimated using standard methods. This could be particularly problematic if the association between two functional outcomes of interest is related to the functional status of the patient at enrollment (low or high). Marginal correlation estimates based on standard methods would generally be biased toward the correlation exhibited by these low functioning patients.

A naive approach to estimating marginal correlations in clustered data would be to apply standard methods requiring the independence of observations. Under certain conditions on the relationship among observations within a given cluster, this approach can produce unbiased estimates when cluster size is non-informative. Traditional estimates of standard error for naive correlations would be invalidated by the dependence of observations within a cluster, complicating interval estimation and hypothesis testing. For marginal analyses of univariate data, generalized estimating equations (GEE) are frequently utilized to estimate within-cluster correlation and obtain proper variance estimates for marginal parameters. The GEE approach could conceivably be extended to multivariate data to estimate parameters such as a correlation coefficient and to provide appropriate estimates of standard error. Aside from the conceptual difficulty of such a multivariate GEE, it has been shown that methods based on GEE can be biased when cluster sizes are informative for the outcome of interest [2, 3].

Another approach to estimating a marginal correlation would be to apply within-cluster averaging, which can take two forms in the context of estimating a correlation: (1) averaging observations within each cluster and applying standard correlation methods to the within-cluster averages and (2) calculating standard correlations within each cluster and then averaging the correlations over the clusters. Empirically, these averaging approaches are unsatisfactory in measuring marginal association. In the first averaging approach, the joint distribution of the within-cluster averages may not relate to the marginal joint distribution of

the observations. In particular, the correlation between the pair of within-cluster averages may differ from the marginal correlation of the observations. Because the second approach calculates correlations within each cluster using standard methods, independence within each cluster would still be required for valid inference. Additionally, clusters of small size may produce unreliable correlation estimates that would adversely impact the cluster-averaged correlation estimate.

To address informative cluster size in the marginal analysis of clustered data, Hoffman *et al.* [2] introduced within-cluster resampling (WCR). In WCR, a single observation is randomly selected from each cluster to form a resampled data set on which methods requiring independent observations are valid (assuming the mutual independence of the clusters). The resampling procedure is repeated many times with replacement, and the WCR estimate of the marginal quantity of interest is calculated as the average of the estimates calculated on the resampled data sets. The WCR approach was introduced in the context of estimating parameters from generalized linear models. Follman *et al.* [4] illustrated its broad applicability in diverse settings, terming it ‘multiple outputation’. Williamson *et al.* [3] built upon the WCR procedure to introduce cluster-weighted GEEs for marginal quantities, in which estimating equations used for independent data are weighted by the inverse of the cluster size. The clear advantage of this approach over WCR is the reduction in computational burden—a marginal statistic of interest from clustered data can be calculated directly from the data, rather than requiring a large number of Monte Carlo resampling steps as with WCR. This cluster-weighting approach has been applied to define clustered data analogs of several statistics for independent data, including the nonparametric Wilcoxon rank sum [5] and signed rank tests [6], and in the definition of cluster-weighted versions of the proportional hazards model for survival data [7, 8].

To date, research in methods for clustered data has focused on univariate outcomes. In this paper, we consider the measurement of association for multivariate outcomes and define correlation coefficients for clustered data. The rest of the article is arranged as follows. In Section 2, we introduce our notation and further describe WCR and the marginalization principle of Williamson *et al.* [3]. We define two measures of association analogous to the standard Pearson and nonparametric Kendall coefficients for independent data and discuss their asymptotic normality. Section 3 provides the results of a simulation study examining the performance of our cluster-weighted estimator in clustered data with informative cluster size and a comparison with three other potential marginal correlation estimators. An application to SCI data is presented in Section 4, and our concluding remarks are in Section 5.

2. Methods

In this section, we introduce two estimators of marginal association for clustered bivariate data in which cluster size may be informative of the association being measured. We begin by defining our notation and outlining WCR and the marginalization principle upon which our estimators are based. We then derive our measures of association as analogs of the standard Pearson and Kendall correlation coefficients for independent data and describe their asymptotic properties.

2.1. Notation and motivation

We consider paired data (X, Y) collected on a set of M clusters. Let (X_{ij}, Y_{ij}) be the j th bivariate observation for cluster i , where $1 \leq i \leq M$ and $1 \leq j \leq n_i$. Let $\mathbf{B}_i$ denote the random vector $\{(X_{i1}, Y_{i1}), \dots, (X_{in_i}, Y_{in_i})\}$ of observations for cluster i and define $\mathbf{V}_i = (n_i, \mathbf{B}_i)$. We assume that the data consist of M independent and identically distributed (i.i.d.) replicates of $\mathbf{V}_i$.

Our interest is in estimating the marginal correlation of the bivariate pair (X, Y) , in which the clusters form the primary sampling unit. We briefly describe the WCR procedure introduced by Hoffman *et al.* [2] in the context of estimating the marginal correlation $\text{Corr}(X, Y)$ as preliminary motivation for our correlation estimators. From each cluster of data $\mathbf{B}_j$, we randomly select a single paired observation, denoted (X_i^*, Y_i^*) , to create a pseudo data set of M resampled observations. Because we have assumed that the clusters $\mathbf{V}_j$ are i.i.d., we can apply standard methods for estimating association to this resampled data set. The measure of association calculated on this resampled data set, say $\hat{\rho}^*$, however, uses only a fraction of the available data and is artificially random because we randomly choose only one observation per cluster. We can calculate the WCR estimator of the correlation as the average of the standard correlations calculated on all possible realizations of the resampling procedure. In practice, we can generate a suitably large number of pseudo data sets to produce the WCR estimator of the marginal correlation, particularly when the number of clusters and/or cluster sizes are large. Let $\hat{\rho}_q^*$ denote the correlation estimator for q th resampled pseudo data set. We define the WCR estimator of the marginal correlation as the average of the correlation estimators on the resampled data sets over Q iterations of the WCR algorithm, $Q^{-1} \sum_{q=1}^Q \hat{\rho}_q^*$ for large Q . Hoffman *et al.* [2] demonstrated the consistency and asymptotic normality of WCR estimators.

The distribution of the resampled data (X_i^*, Y_i^*) corresponds to the distribution of a randomly chosen pair from a randomly chosen cluster, which we can formally define as follows. Conditional on the cluster size n_j let J_j be a random integer uniformly distributed on $\{1, \dots, n_j\}$. The distribution of (X_{iJ_j}, Y_{iJ_j}) is then

$$F(x, y) = E_V \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} I(X_{ij} \leq x, Y_{ij} \leq y) \right\}, \quad (1)$$

where E_V represents expectation taken with respect to the joint distribution of the vector $\mathbf{V}_j$ and $I[A]$ is the indicator function taking value 1 when event A occurs and value 0 otherwise. We will use the distribution function (1) in subsequent sections to define the marginal correlations in which we are interested.

Although intuitively appealing, the WCR procedure is computationally intensive. The marginalization principle introduced by Williamson *et al.* [3] based upon WCR provides estimators for clustered data based on weighting standard estimating equations by the inverse of the cluster size. Such estimators were shown to be asymptotically equivalent to WCR estimators and subsequently consistent for the marginal parameters of interest and asymptotically normal. The concept rests largely on noting that we can define the WCR estimator of a given marginal population quantity (such as a correlation) as the conditional expectation of the sample quantity calculated on one realization of the WCR procedure given the original data $\mathbf{V}_j$. As an example, consider the estimation of the marginal mean of X in the present bivariate framework, $E_F(X) = \iint x dF(x, y)$, where F is the distribution function in (1). One realization of the WCR procedure produces data $\{X_i^*, 1 \leq i \leq M\}$ and the estimator $M^{-1} \sum_{i=1}^M X_i^*$. Following the marginalization principle and taking the conditional expectation of this estimator given the data, we get

$$E_F(\bar{X}) = E \left[\left(\frac{1}{M} \sum_{i=1}^M X_i^* \right) \middle| \{\mathbf{V}_1, \dots, \mathbf{V}_M\} \right] = \frac{1}{M} \sum_{i=1}^M E[X_i^* | \{\mathbf{V}_1, \dots, \mathbf{V}_M\}] = \frac{1}{M} \sum_{i=1}^M \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}. \quad (2)$$

Note that in the above, the expectation E is taken with respect to the distribution (1). This simple moment calculation provides the foundation for other moment-based estimators for clustered data, such as the Pearson correlation discussed in the following paragraphs.

2.2. Marginal Pearson correlation

The first marginal measure of association for clustered data we define is an analog of the standard Pearson correlation. The Pearson correlation coefficient for independent data has a convenient representation as a function of the first and second moments of the bivariate distribution function of the observations. We adapt this representation to the clustered data setting to define our marginal Pearson-type correlation estimator for clustered data. We begin by defining the population quantity of interest in terms of the marginal bivariate distribution $F(x, y)$ in (1). Denote the first and second moments of $F(x, y)$ as follows: $\omega_1 = \iint x dF(x, y)$, $\omega_2 = \iint y dF(x, y)$, $\omega_3 = \iint xy dF(x, y)$, $\omega_4 = \iint x^2 dF(x, y)$, and $\omega_5 = \iint y^2 dF(x, y)$. We define the marginal population Pearson-type correlation between X and Y , denoted ρ_m ('m' for marginal), in terms of these quantities as follows:

$$\rho_m = \frac{\omega_3 - \omega_1 \omega_2}{\sqrt{(\omega_4 - \omega_1^2)(\omega_5 - \omega_2^2)}}. \quad (3)$$

We define the standard Pearson correlation estimator by replacing the population first and second moments with their sample equivalents. We proceed in the same fashion to develop our estimator for ρ_m . We can easily define marginal estimators of the ω_k following the simple moment calculation in (2). For convenience of notation, let

$$W_{1i} = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}, W_{2i} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}, W_{3i} = n_i^{-1} \sum_{j=1}^{n_i} X_{ij} Y_{ij}, W_{4i} = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}^2, \text{ and } W_{5i} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}^2.$$

Note that the W_{ki} define the cluster-specific conditional first and second moments of $F(x, y)$ given V_i . Following (2), we define the sample moments $\bar{W}_k = M^{-1} \sum_{i=1}^M W_{ki}$ for $k = 1, \dots, 5$. Our Pearson correlation estimator, $\hat{\rho}_m$, for clustered data is then defined by replacing the population moments in (3) with their corresponding sample moments:

$$\hat{\rho}_m = \frac{\bar{W}_3 - \bar{W}_1 \bar{W}_2}{\sqrt{(\bar{W}_4 - \bar{W}_1^2)(\bar{W}_5 - \bar{W}_2^2)}}. \quad (4)$$

Let $\hat{\Sigma}$ be the empirical variance-covariance matrix of $\mathbf{W}_i = (W_{1i}, W_{2i}, W_{3i}, W_{4i}, W_{5i})^T$,

$$\hat{\Sigma} = \frac{1}{M-1} \sum_{i=1}^M (\mathbf{W}_i - \bar{\mathbf{W}})(\mathbf{W}_i - \bar{\mathbf{W}})^T, \quad (5)$$

where $\bar{\mathbf{W}} = M^{-1} \sum_{i=1}^M \mathbf{W}_i$. Define the function $g(\mathbf{z}) = (z_3 - z_1 z_2) / \sqrt{(z_4 - z_1^2)(z_5 - z_2^2)}$ for a vector $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5)^T$ and note that $\rho_m = g(\boldsymbol{\omega})$ for $\boldsymbol{\omega} = E(\mathbf{W}_i)$ and $\hat{\rho}_m = g(\bar{\mathbf{W}})$. Because we have assumed that the clusters V_i are independent, the vectors $\mathbf{W}_i$ are independent as well. Therefore, the central limit theorem applied to the vector $\mathbf{W}_i$ and the delta method give us that $\hat{\rho}_m$ is asymptotically normal with mean ρ_m and variance estimated by $M^{-1} \boldsymbol{\ell}^T \hat{\Sigma} \boldsymbol{\ell}$, where $\boldsymbol{\ell}$

is the vector of first partial derivatives of g evaluated at $\bar{\mathbf{W}}$, $\boldsymbol{\ell}^T = \left(\frac{\partial g}{\partial z_1} \Big|_{z_1=\bar{W}_1}, \dots, \frac{\partial g}{\partial z_5} \Big|_{z_5=\bar{W}_5} \right)$. We omit the details of the calculation of $\boldsymbol{\ell}$ as they are rudimentary and algebraically tedious.

We will illustrate the asymptotic properties of $\hat{\rho}_m$ via the simulation study and application to SCI data in Sections 3 and 4.

2.3. Marginal Kendall correlation

The Kendall correlation coefficient is a frequently used nonparametric alternative to the Pearson correlation. In the following paragraphs, we define a nonparametric marginal correlation estimator analogous to the standard Kendall correlation for independent data. The Kendall coefficient has several formulations as a function of the number of concordant pairs of observations in a set. We adopt the U -statistic formulation [9], for which the order 2 kernel of $\tau + 1$ is given by $h((x_1, y_1), (x_2, y_2)) = 2I[(y_2 - y_1)(x_2 - x_1) > 0]$. Applying this formulation to one realization of the WCR procedure, that is, for resampled data

$\{(X_i^*, Y_i^*), 1 \leq i \leq M\}$, we get $\tau^* = 2 \binom{M}{2}^{-1} \sum_{i=1}^{M-1} \sum_{i'=i+1}^M I[(Y_{i'}^* - Y_i^*)(X_{i'}^* - X_i^*) > 0] - 1$. As before, we apply the marginalization principle and take the conditional expectation of τ^* given the data to arrive at the marginal Kendall-type correlation estimator:

$$\hat{\tau}_m = 2 \binom{M}{2}^{-1} \sum_{i=1}^{M-1} \sum_{i'=i+1}^M \frac{1}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} I[(Y_{i'j'} - Y_{ij})(X_{i'j'} - X_{ij}) > 0] - 1. \quad (6)$$

We note that in Equation 6, observations from the same cluster are not compared in the argument of the indicator function as is appropriate in a marginal analysis where clusters form the primary sampling unit.

It is not difficult to see that we can express $\hat{\tau}_m$ as a U -statistic based on the $\mathbf{V}_i$. Define the kernel $h(\mathbf{v}_1, \mathbf{v}_2) = 2n_1^{-1}n_2^{-1} \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} I[(y_{1j_1} - y_{2j_2})(x_{1j_1} - x_{2j_2}) > 0]$ for vectors $\mathbf{v}_i = (n_i, (x_{i1}, y_{i1}), \dots, (x_{in_i}, y_{in_i}))$, $i = 1, 2$, and note that $\hat{\tau}_m + 1 = \binom{M}{2}^{-1} \sum_{i_1=1}^{M-1} \sum_{i_2=i_1+1}^M h(\mathbf{V}_{i_1}, \mathbf{V}_{i_2})$. Standard asymptotic theory for U -statistics is not applicable in the current framework because the sampling units, the vectors $\mathbf{V}_i$ are of different, random lengths. However, basic projection arguments still apply, and we can demonstrate the asymptotic normality of $\hat{\tau}_m$ along similar lines as standard U -statistic theory.

Consider the Hájek projection of $\hat{\tau}_m$, which, up to a constant, has k th summand $S_k = E[\hat{\tau}_m + 1 | \mathbf{V}_k]$. To calculate S_k , define the function $h_1(\mathbf{v}) = E[h(\mathbf{v}, \mathbf{V}_2)]$ for the vector $\mathbf{v} = (n, (x_1, y_1), \dots, (x_n, y_n))$, where h is the kernel function associated with $\hat{\tau}_m$ defined in the preceding paragraphs. After some algebra, we get

$$h_1(\mathbf{v}) = \frac{2}{n} \sum_{j=1}^n (F^l(x_j, y_j) + F^r(x_j, y_j)), \quad (7)$$

where $F^l(x, y) = P(X < x, Y < y)$ and $F^r(x, y) = P(X > x, Y > y)$, and P is the probability measure corresponding to the distribution function (1). Therefore,

$$S_k = \frac{2}{M} h_1(\mathbf{V}_k) = \frac{4}{M n_k} \sum_{j=1}^{n_k} (F^l(X_{kj}, Y_{kj}) + F^r(X_{kj}, Y_{kj})). \quad (8)$$

The quantity S_k is of little value in practice, as F^l and F^r are unknown. Define $\hat{S}_k$ by replacing $F^l(x, y)$ and $F^r(x, y)$ with

$$\hat{F}^l(x, y) = \frac{1}{M} \sum_{i=1}^M \hat{F}_i^l(x, y), \quad \hat{F}^r(x, y) = \frac{1}{M} \sum_{i=1}^M \hat{F}_i^r(x, y), \quad (9)$$

where $\hat{F}_i^l(x, y) = n_i^{-1} \sum_{j=1}^{n_i} I[X_{ij} < x, Y_{ij} < y]$ and $\hat{F}_i^r(x, y) = n_i^{-1} \sum_{j=1}^{n_i} I[X_{ij} > x, Y_{ij} > y]$. We can then estimate the asymptotic variance of $\hat{\tau}_m$, denoted $\hat{\sigma}_\tau^2$, by M times the empirical variance of the summands of the Hájek projection: $\hat{\sigma}_\tau^2 = M(M-1)^{-1} \sum_{i=1}^M (\hat{S}_i - \bar{S})^2$, where $\bar{S} = M^{-1} \sum_{i=1}^M \hat{S}_i$. Following the arguments of Datta and Satten [5], we can show that $\hat{\tau}_m$ is asymptotically normal, with mean

$$\tau_m = 2 P \left[(Y_{i'j'} - Y_{ij})(X_{i'j'} - X_{ij}) > 0 \right] - 1 \quad (10)$$

and variance estimated by $\hat{\sigma}_\tau^2$. The sampling properties of $\hat{\tau}_m$ are further explored in Sections 3 and 4.

3. Simulation study

We conducted a small simulation study to explore the performance of our marginal correlation estimators on clustered data with informative cluster sizes. We also compared the performance of our estimators with the three candidate marginal correlation estimators mentioned in the introduction: (1) the naive Pearson and Kendall correlation coefficients for independent data calculated on all observations; (2) the cluster-averaged Pearson and Kendall coefficients, defined as the standard Pearson and Kendall coefficients for independent data calculated within each cluster and then averaged over the clusters; and (3) observation-averaged Pearson and Kendall coefficients, in which the standard Pearson and Kendall coefficients are applied to within-cluster averages of the observations.

We simulated M clusters of bivariate data (X, Y) as follows. As before, let i index the clusters and j the observations within each cluster ($1 \leq i \leq M, 1 \leq j \leq n_i$). Our model for creating the simulated bivariate data was

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} = \begin{pmatrix} u_i \\ v_i \end{pmatrix} + \begin{pmatrix} \mu_{ij} \\ \nu_{ij} \end{pmatrix},$$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix} \right), \quad \begin{pmatrix} \mu_{ij} \\ \nu_{ij} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad (11)$$

effectively, a multivariate random effects model, with the (u_i, v_i) serving as random effects for the clusters and the (μ_{ij}, ν_{ij}) as the model errors. We generated both the random effects and model errors i.i.d. and independently of each other. Under this design, the marginal distribution of (X_{ij}, Y_{ij}) was the same for all pairs with $E(X_{ij}) = E(Y_{ij}) = 0$, $\text{Var}(X_{ij}) = \text{Var}(Y_{ij}) = 2$, and $\text{Cov}(X_{ij}, Y_{ij}) = \rho + \gamma$; hence, the marginal correlation ρ_m between X and Y was $(\rho + \gamma)/2$.

We fixed the number of clusters at $M = 100$ and ran simulations for three values of the model errors correlation parameter ρ , -0.5 , 0.2 , and 0.8 , and two values for the random effects correlation parameter γ , 0 and 0.4 . We generated cluster sizes as the following function of the random effects (u_i, v_i) :

$$n_i = \begin{cases} n_1 & \text{if } \text{sign}(u_i * v_i) > 0, \\ n_2 & \text{otherwise.} \end{cases} \quad (12)$$

In one set of simulations, we defined $(n_1, n_2) = (20, 10)$ so that clusters with concordant random effects ($\text{sign}(u_i * v_i) > 0$) had larger cluster sizes. In the second set of simulations, we set $(n_1, n_2) = (10, 20)$ so that clusters with discordant random effects had larger cluster sizes. For each configuration of the design settings, we calculated the average of each correlation estimator over 10,000 Monte Carlo replicates.

Our cluster-weighted approach exhibited unbiasedness under all configurations of the design parameters for both the Pearson-type and Kendall-type coefficients (Table I). In particular, we noted that the cluster-weighted approach was resistant to changes in the cluster size structure, that is, whether concordant or discordant observations were more heavily weighted.

The cluster-averaged coefficients were also resistant to changes in our informative cluster size configuration but did not estimate the true marginal correlation. Because the cluster-averaged coefficients calculate correlations within each cluster and then average over the clusters, the variance and covariance added by the random effects in simulation model (11) are effectively ignored. Hence, the cluster-averaged Pearson coefficient estimated the model errors correlation ρ rather than the marginal correlation $\rho_m = (\rho + \gamma)/2$. Similarly, the cluster-averaged Kendall coefficient did not estimate τ_m , rather τ from a model without the added unit of random effects variance. We briefly note that τ from such a model are approximately $-0.33, 0.13$, and 0.59 for the three values of ρ , and that the cluster-averaged coefficients closely estimated these quantities.

The observation-averaged coefficients also did not estimate the true marginal correlation. The averaging of observations within each cluster heavily diluted the correlation of the model errors (ρ), leaving only the correlation of the random effects (γ). We can see this in the differences between simulations in which the random effects covariance parameter γ was equal to zero (first block, Table I) and when γ was equal to 0.4 (second block, Table I). When there was no covariation in the random effects ($\gamma = 0$), the observation-averaged correlations were of substantially reduced magnitude relative to the marginal correlations being estimated. When random effects covariation was introduced ($\gamma = 0.4$), the observation-averaged correlations picked this up. Under both scenarios, the observation-averaged correlations increased as the model-errors covariance parameter ρ increased, although to a much lesser degree as mentioned.

The results for the naive Pearson and Kendall correlation coefficients illustrated the biasing effect of informative cluster size. Clusters with concordant random effects provide more evidence of a positive association between X and Y . When these clusters systematically have larger cluster sizes (i.e., when $n_1 = 20$), the effect on measures of association that do not adjust for informative cluster size would be to overestimate the true marginal association between X and Y . This was illustrated by our results. When $(n_1, n_2) = (20, 10)$ in our simulations, both the Pearson and Kendall coefficients overestimated the true marginal correlation. Conversely, when clusters with discordant random effects had larger sizes (i.e., when $(n_1, n_2) = (10, 20)$), the naive correlations substantially underestimated the true marginal association measures.

To evaluate the normal approximation for our cluster-weighted coefficients, we constructed asymptotic 95% confidence intervals for our cluster-weighted Pearson and Kendall coefficients according to the variance formulas given in Section 2. The normal confidence

intervals appeared to work well for both coefficients (Table II), although coverage probabilities tended to be slightly less than the nominal confidence level of 0.95. This slight under-coverage corresponded with known results for asymptotic confidence intervals for the Pearson and Kendall coefficients for independent data (see, e.g., [9]). Given the substantial bias for the true marginal association exhibited by the other coefficients (Table I), we did not calculate coverage probabilities for them.

We conducted an additional simulation to evaluate the performance of our cluster-weighted coefficients for large correlations near the boundary of 1.0, in correspondence with results from our analysis of SCI data in the following section. For these simulations, we set ρ and γ from (11) to be 0.85 and 0.95, respectively, so that $\rho_m = 0.9$ and $\tau_m = 0.71$. The cluster-weighted coefficients remained unbiased under both cluster size configurations (empirical averages = 0.90, 0.71 for $\hat{\rho}_m, \hat{\tau}_m$), and the asymptotic 95% confidence intervals exhibited good coverage (between 0.9377 and 0.956).

4. Application to NeuroRecovery Network data

To illustrate the application of our marginal association measures, we examined functional outcome data for individuals with incomplete SCI. The data set comes from the Christopher and Dana Reeve Foundation NeuroRecovery Network (NRN), a specialized network of treatment centers in the USA that provide a standardized activity-based therapy and locomotor training for individuals with incomplete SCI. Patients eligible for enrollment in the NRN have incomplete SCI with lesions at spinal level T10 or above, are not participating in inpatient rehabilitative programs, and meet other previously detailed eligibility criteria [1]. Participants in the NRN undergo periodic evaluations of several cardiovascular, health, functional, and quality-of-life outcome measures until discharge from the program. The data set analyzed here consisted of 1519 evaluations of the outcome measures taken on 323 NRN participants. In the context of clustered data, the 323 participants served as the clusters, and the 1519 evaluations served as the observations within clusters. The median number of observations per patient/cluster was three with a minimum of one and maximum of 18. Demographic and clinical characteristics and the recovery of functional capability by NRN participants have been previously reported and explored [10]. Pertaining to the present analysis, we have shown that functional capability prior to enrollment was related to the length of time enrolled in the NRN, and hence the number of observations contributed per patient [10]. Therefore, cluster size was informative for the outcome measures described in the following paragraphs, as patients of poor function tended to contribute a greater number of observations.

The focus of our analysis was on marginal relationships among two measures of walking function, the Six Minute Walk (6MW) and Ten Meter Walk (10MW) Tests [11], and two measures of balance capability, the Berg Balance Scale (BBS) [12,13] and Modified Functional Reach (MFR) [14]. We calculated the cluster-weighted Pearson and Kendall coefficients for each pairing of these four outcome measures as well as the naive, cluster-averaged, and observation-averaged Pearson and Kendall coefficients. We additionally calculated asymptotic 95% confidence intervals for our cluster-weighted coefficients.

We observed that the association between the two measures of walking capability was quite strong (Figure 1; Table III). The BBS corresponded well with the walking measures along a curve (up to the ceiling of 56 for BBS), and the MFR was poorly related to the walking measures and moderately related to the BBS. Clinically, these results indicated that walking endurance and gait speed, ostensibly measured by the 6MW and 10MW tests, respectively, were closely related in incomplete SCI patients. Further, balance capability as measured by the BBS was strongly tied to walking function, indicating the importance of proper balance

in ambulatory capability. The absence of association between the MFR and walking tests was not unexpected as the MFR is largely an assessment of sitting balance capacity, a function largely unrelated to walking capability. The BBS is largely composed of items that assess aspects of standing balance with only a few items challenging sitting balance, explaining the moderate relationship between the BBS and MFR. We briefly note that the Kendall coefficients are preferable measures of association for these data. The 6MW, 10MW, and MFR are characteristically right-skewed for the incomplete SCI population (and were so for this sample), and the Berg Balance Scale is bounded on both sides. Hence, the assumption of pairwise bivariate normality required of the Pearson coefficient was likely violated by these data.

The cluster-weighted, naive, and observation-averaged coefficients were in general correspondence for all pairings of outcome variables, where the observation-averaged coefficients were consistently higher and the naive coefficients consistently lower than the cluster-weighted coefficients. The disparity among the coefficients was inversely related to their magnitude—pairs of functional measures exhibiting weaker association exhibited greater disparity among these three coefficients. The cluster-averaged correlations were substantially lower than the other three coefficients for all pairings of outcome measures. We note that the cluster-averaged methodology necessarily excluded clusters with two or fewer observations, as it is not possible to calculate a standard correlation on a single observation and a standard correlation calculated for two observations will have absolute magnitude 1. Hence, the 134 (41%) NRN patients with two or fewer evaluations were excluded from the calculation of the cluster-averaged coefficients, a possible (partial) explanation of the disparity between this coefficient and the other three. These 134 patients accounted for 235 (15%) of the observations in the data set.

Figure 1 in part depicts an explanation for the reduction in the naive coefficients relative to the cluster-weighted coefficients. The disparities between the two coefficients were lowest among the BBS, 6MW, and 10MW, and the lower triangle of Figure 1 shows that the standard Pearson correlation was fairly stable as a function of the cluster size. Correlations involving the MFR were less stable as a function of cluster size (first column, Figure 1), and larger, more heavily weighted clusters tended to exhibit lower correlations. Hence, the naive correlations involving the MFR tended to be lower than the cluster-weighted correlations to a greater degree than those not involving the MFR.

In our simulation study, the observation-averaged coefficients picked up correlation in the random effects and diluted the correlation of the model errors. The cluster-averaged coefficients picked up correlation in the model errors and ignored the random effects. This phenomenon serves as a potential explanation for the estimated values for these coefficients in the NRN data set, in which the observation-averaged correlations were greater than and cluster-averaged correlations less than the cluster-weighted correlations. One possible conclusion is that much of the correlation among these outcome measures was due to correlated random effects rather than correlated observation errors.

5. Discussion

The marginal correlations defined in this paper represent one type of marginal analysis of clustered data; there are several possibilities for a marginal analysis. Indeed, the cluster-averaged, observation-averaged, and standard Pearson and Kendall coefficients investigated in our simulation study themselves are forms of marginal analyses, although the marginal quantities being estimated by these coefficients are complex functionals of the distribution of the vector V_i ; and in particular of the random (and possibly informative) cluster size n_i . In contrast, our formulas indeed estimate the corresponding correlations in the true marginal

model of individual pairs under the natural assumption of identical marginal distribution of all pairs within a cluster. This holds regardless of whether the cluster size is informative or not.

In the SCI data set analyzed in Section 4, cluster size was potentially informative because patients with greater impairment and lower function remained in the rehabilitation program over longer periods, contributing more observations. The marginal correlations we calculated mitigate the effect of informative cluster size and represent general measures of association among the four selected functional outcome measures for individuals with incomplete SCI. If there is interest in examining how the correlations among these four outcomes vary according to functional status at enrollment, then our marginal analysis would not be appropriate—our correlations marginalize the effect of cluster size, which in this case was a proxy for functional status. One could conduct a conditional analysis of correlation by stratifying the study sample on some (external) classifier of functional status.

For clustered data, focus has traditionally been placed on handling correlation among observations within a cluster. Familiar examples for univariate data include extensions of the general linear model to clustered data, such as GEEs and mixed effects models. However, as noted by Datta and Satten [6], when cluster sizes are informative, these methods generally estimate quantities under a distribution under which individual observations, rather than

individuals, are equally weighted, say $F'(x) = N^{-1} \sum_{i=1}^M \sum_{j=1}^{n_i} I(X_{ij} \leq x)$, where $N = \sum_{i=1}^M n_i$. For non-random or non-informative n_i , this generally does not pose a problem, but when n_i is informative, estimators based on $F'(x)$ can be biased for certain marginal parameters of interest, as demonstrated by the simulation results in Section 3. In addition, from a theoretical perspective, this distribution does not correspond to a proper empirical distribution when the cluster size is informative, even if the vectors V_i are i.i.d.

The marginalization technique of Williamson *et al.* [3] enjoys the benefits of producing estimators that are asymptotically equivalent to WCR estimators but are calculable functions of the data not requiring computationally taxing randomization techniques. The method has wide applicability for clustered data problems. Developments based upon this technique have focused on univariate data—cluster-weighted GEEs [3], Wilcoxon tests [5,6], and survival data [7,8]. The paper of Follman *et al.* [4] suggests the applicability of WCR to multivariate data, but to our knowledge the marginalization principle has yet to be explored for multivariate data. The correlation coefficients introduced in this paper represent a first step into the marginal analysis of clustered multivariate data with informative cluster size. Extensions of the marginalization principle to additional multivariate methodologies, including the calculation of alternative correlation coefficients such as Spearman's ρ and coefficients for binary data, are certainly possible.

Acknowledgments

The authors are grateful for the use of data from the NeuroRecovery Network and thank the directors of centers participating in the NRN: Steve Ahr (Frazier Rehab Institute, Louisville, KY); Mary Schmidt Read, DPT (Magee Rehabilitation, Philadelphia, PA); Steve Williams, MD (Boston Medical Center, Boston, MA); Daniel Graves, PhD (Memorial Hermann/The Institute of Rehabilitation and Research, Houston, TX); Keith Tansey, MD, PhD (Shepherd Center, Atlanta, GA); Gail Forrest, PhD (Kessler Medical Rehabilitation Research and Education Corporation, West Orange, NJ); and D. Michele Basso PT, EdD (The Ohio State University Medical Center, Columbus, OH). We also thank the two anonymous referees for their helpful critical commentary on an earlier version of the manuscript. This research was supported in part by a grant (1R03DE020839-01A1) from the National Institutes of Health.

Contract/grant sponsor: Christopher and Dana Reeve Foundation (CDRF) and Centers for Disease Control and Prevention (CDC). Grant/Cooperative Agreement Number U10/CCU220379.

References

1. Harkema SJ, Schmidt-Read M, Behrman A, Bratta A, Sisto SA, Edgerton VR. Establishing the NeuroRecovery Network: multi-site rehabilitation centers that provide activity based therapies and assessments for neurologic disorders. *Archives of Physical Medicine and Rehabilitation*. 2011
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. Follman D, Proschan M, Liefer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics*. 2003; 59:420–429. [PubMed: 12926727]
5. Datta S, Satten GA. Rank-sum tests for clustered data. *Journal of the American Statistical Association*. 2005; 100(471):908–915.
6. Datta S, Satten GA. A signed-rank test for clustered data. *Biometrics*. 2008; 65:501–507. [PubMed: 17970820]
7. Cong XJ, Yin G, Shen Y. Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*. 2007; 63:663–672. [PubMed: 17825000]
8. Williamson JM, Kim HY, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. *Statistics in Medicine*. 2008; 27:543–555. [PubMed: 17640035]
9. van der Vaart, AW. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press; 1998.
10. Harkema SJ, Schmidt-Read M, Lorenz DJ, Edgerton VR, Behrman A. Balance and ambulation improvements in individuals with chronic incomplete spinal cord injury using locomotor training-based rehabilitation. *Archives of Physical Medicine and Rehabilitation*. 2011
11. van Hedel HJ, Wirz M, Dietz V. Assessing walking ability in subjects with spinal cord injury: validity and reliability of 3 walking tests. *Archives of Physical Medicine and Rehabilitation*. 2005; 86(2):190–196. [PubMed: 15706542]
12. Berg K, Wood-Dauphinee S, Williams JI, Gayton D. Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada*. 1989; 41:304–311.
13. Berg KO, Wood-Dauphinee S, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. *Canadian Journal of Public Health*. 1992; 83:S7–S11.
14. Adegoke BO, Ogwumike OO, Olatemiju A. Dynamic balance and level of lesion in spinal cord injured patients. *African Journal of Medicine and Medical Sciences*. 2002; 31:357–360. [PubMed: 15027780]

Figure 1.

Modified scatterplot matrix of four NRN outcome measures. Upper triangle depicts all unclustered observations. Lower triangle depicts standard Pearson correlations calculated as a function of cluster size. Dashed lines represent cluster-weighted correlation estimates. MFR, Modified Functional Reach; BBS, Berg Balance Scale; 6MW, Six Minute Walk; 10MW, Ten Meter Walk.

Empirical averages of correlation estimators over 10,000 Monte Carlo replications from simulation study of $M = 100$ clusters.

Table I

Design parameters			Pearson coefficients					Kendall coefficients				
ρ	γ	(n_1, n_2)	True ρ_m	CW(*)	N	CA	OA	True τ_m	CW(*)	N	CA	OA
-0.5	0	(20, 10)	-0.25	-0.25	-0.15	-0.48	-0.03	-0.16	-0.16	-0.09	-0.33	-0.02
		(10, 20)		-0.25	-0.35	-0.48	-0.03		-0.16	-0.24	-0.33	-0.03
0.2		(20, 10)	0.10	0.10	0.20	0.19	0.01	0.06	0.06	0.14	0.13	0.01
		(10, 20)		0.10	0.00	0.19	0.01		0.06	-0.01	0.13	0.00
0.8		(20, 10)	0.40	0.40	0.51	0.79	0.06	0.26	0.26	0.35	0.59	0.04
		(10, 20)		0.40	0.30	0.79	0.06		0.26	0.19	0.59	0.03
-0.5	0.4	(20, 10)	-0.05	-0.05	0.03	-0.49	0.34	-0.03	-0.03	0.03	-0.33	0.23
		(10, 20)		-0.05	-0.16	-0.48	0.33		-0.03	-0.11	-0.33	0.21
0.2		(20, 10)	0.30	0.30	0.37	0.19	0.39	0.19	0.19	0.25	0.13	0.26
		(10, 20)		0.30	0.20	0.19	0.38		0.19	0.12	0.13	0.25
0.8		(20, 10)	0.60	0.60	0.66	0.79	0.42	0.41	0.41	0.47	0.59	0.29
		(10, 20)		0.60	0.52	0.79	0.43		0.41	0.34	0.59	0.28

The estimators we advocate are indicated by asterisks. CW, cluster weighted; N, naive; CA, cluster averaged; OA, observation averaged. True τ_m were calculated using Equation (10) under the parametric assumptions of simulation model (11).

Table II

Empirically estimated coverage probabilities of the asymptotic normal 95% confidence intervals calculated using the cluster-weighted Pearson and Kendall coefficients.

Design parameters			Coverage Probability	
ρ	γ	(n_1, n_2)	Pearson	Kendall
-0.5	0	(20, 10)	0.9359	0.9355
		(10, 20)	0.9398	0.9477
0.2		(20, 10)	0.9387	0.9417
		(10, 20)	0.9412	0.9408
0.8		(20, 10)	0.9337	0.9521
		(10, 20)	0.9319	0.9240
-0.5	0.4	(20, 10)	0.9380	0.9277
		(10, 20)	0.9403	0.9604
0.2		(20, 10)	0.9420	0.9306
		(10, 20)	0.9419	0.9492
0.8		(20, 10)	0.9362	0.9416
		(10, 20)	0.9364	0.9403

The simulation setup was the same as in Table I with $M = 100$ clusters.

Table III

Measures of association for four NRN outcome measures with 95% normal confidence intervals for the cluster-weighted correlations.

Measure	Coefficient	MFR	BBS	6MW	10MW
MFR	Naïve	–	0.40	0.17	0.19
	Cluster-averaged	–	0.22	0.14	0.14
	Observation-averaged	–	0.48	0.25	0.25
	Cluster-weighted (CW*)	–	0.44	0.23	0.22
	—95% CI (using CW)	–	(0.36, 0.51)	(0.12, 0.34)	(0.12, 0.32)
BBS	Naïve	0.31	–	0.77	0.77
	Cluster-averaged	0.18	–	0.61	0.58
	Observation-averaged	0.36	–	0.80	0.78
	Cluster-weighted (*)	0.32	–	0.79	0.77
	—95% CI	(0.26, 0.39)	–	(0.75, 0.82)	(0.73, 0.81)
6MW	Naïve	0.13	0.65	–	0.94
	Cluster-averaged	0.11	0.53	–	0.74
	Observation-averaged	0.17	0.69	–	0.96
	Cluster-weighted (*)	0.17	0.67	–	0.94
	—95% CI	(0.08, 0.25)	(0.63, 0.71)	–	(0.91, 0.95)
10MW	Naïve	0.14	0.68	0.81	–
	Cluster-averaged	0.11	0.49	0.62	–
	Observation-averaged	0.16	0.71	0.86	–
	Cluster-weighted (*)	0.16	0.69	0.83	–
	—95% CI	(0.08, 0.24)	(0.66, 0.72)	(0.81, 0.85)	–

The estimators we advocate are indicated by asterisks. Kendall coefficients are on lower triangle, Pearson coefficients on the upper triangle.