SciMedCentral

Clinical Research in HIV/ AIDS

Research Article

Extending the Mann-Whitney-Wilcoxon Rank Sum Test for Multiple Treatment Groups and Longitudinal Study Data

Chen R, Wu P, Ma F, Han Y, Chen T, Tu XM*, and Kowalski J

Department of Biostatistics and Computational Biology, University of Rochester, USA

Abstract

Popular models for longitudinal data analysis with continuous outcomes such as linear mixed-effects model and weighted generalized estimating equations lack robustness in the presence of outliers. For example, in a study to evaluate the efficacy of a sexual risk-reduction intervention for sexually active teenage girls in low-income urban settings, some adolescent girls reported very large numbers such as 450 and even 1,000,000 for their unprotected vaginal sex over a three-month period. Although answers like this are clearly not legitimate values of the outcome, they do indicate the extremely high level of sexual activity among these girls and thus should not be completely ignored. However, the mean-based GLMM and WGEE are not capable of dealing with this type of " outliers", due to the sensitivity of the sample mean to values of extremely large magnitude. Rank based methods such as the popular Mann-Whitney-Wilcoxon (MWW) rank sum test are more effective alternatives to address such outliers. Unfortunately, available methods for inference are limited to cross-sectional data and cannot be applied to longitudinal studies, especially in the presence of missing data.

In this paper, we propose to extend the MWW test for comparing multiple groups within a longitudinal data setting, by utilizing the function response models. Inference is based on a class of U-statistics weighted generalized estimating equations, which provides consistent estimates, with asymptotic normal distributions, not only for complete data but also for missing data under MAR, the most popular missing mechanism in real studies. The approach is illustrated with data from both real and simulated studies.

***Corresponding author**

Xin Tu, Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave., Box 630, CTSB 4.239, Rochester, NY 14642, USA, Tel: 585 275-0413; Fax: 585 273-1031;E mail: Xin_Tu@urmc.rochester.edu

Submitted: 19 December 2013

Accepted: 29 January 2014

Published: 31 January 2014

Keywords

- Functional response models
- Missing data
- Outliers
- Sexual health
- U-statistics based weighted generalized estimating equations

INTRODUCTION

Popular models for longitudinal data analysis with continuous outcomes such as linear mixed-effects models (GLMM) and weighted generalized estimating equations (WGEE) lack robustness in the presence of outliers. For example, in a study to evaluate the efficacy of a sexual risk-reduction intervention for sexually active teenage girls in low-income urban settings, a group at elevated risk for HIV, some adolescent girls reported very large numbers such as 450 and even 1,000,000 for their unprotected vaginal sex over a three-month period [1]. Although answers like this are clearly not legitimate values of the outcome, they do indicate the extremely high level of sexual activity among these girls, as compared to the rest of the study sample, and should not be removed for analysis. However, the mean-based GLMM and WGEE are not capable of dealing with this type of "outliers", due to the sensitivity of the sample mean to large values. On the other hand, rank based methods such as the popular Mann-Whitney-Wilcoxon (MWW) rank sum test are more effective to address such outliers. However, available methods for inference are limited to cross-sectional data and cannot be applied to longitudinal data, especially in the presence of missing data. In this paper, we address this issue by extending the MWW test to a longitudinal data and multi-group setting within the framework of the functional response models (FRM). Inference for the FRM-based model is achieved by a class of U-statistics based weighted generalized estimating equations (UWGEE). The approach is illustrated with data from both real and simulated study data. In Section data application in sexual health research as well as simulated data to study the behavior of the estimate for small to moderate sample sizes.

MULTI-SAMPLE MANN-WHITNEY-WILCOXON TESTS

We first briefly review the classic Mann-Whitney-Wilcoxon rank sum test for between-group difference. We then discuss

limitations of existing modeling paradigms to extend it for multi-group comparison within a longitudinal data setting and how the functional response model overcomes such difficulties to achieve the needed generalization.

The mann-whitney-wilcoxon rank sum test

Consider two independent samples with size n_k and let y_{ki} be some continuous outcome from the i th subject within the k th group ($1 \le i \le n_k$, k=1,2). Let R_{ki} denote the rank of y_{ki} in the pooled sample. The Wilcoxon rank sum statistic has the following form [2,3]:

$$\text{Wilcoxon rank sum statistic}: \quad W_n = \sum_{i=1}^{n_1} R_{1i}.$$

Note that the sum of the rank scores $\sum_{j=1}^{n_2} R_{2j}$ from the second group may also be used as a statistic. However, since the two sums add up to $\frac{n(n+1)}{2}$ $(n = n_1 + n_2)$, only one of the sums can be used as a test statistic. An alternative form of this test is the following Mann–Whitney statistic [4,3].:

$$\text{Mann}--\text{Whitney statistic}: \quad U_n = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I_{\{y_{2j} - y_{1i} \le 0\}}, \qquad (1)$$

where $I_{\{u \le 0\}}$ is a set indicator with $I_{\{u \le 0\}} = 1$ if $u \le 0$ and 0 otherwise. Since $W_n = U_n + \frac{n_1(n_1+1)}{2}$ (in the absence of ties), the two tests are equivalent. However, since it is easier to extend the Mann–Whitney version for multi-group comparisons with longitudinal data in the presence of missing values, we focus on (1) and refer to it as the Mann–Whitney–Wilcoxon rank sum test in the remaining discussion unless otherwise stated.

Let $V_n = \binom{n}{2}^{-1} U_n$ be the normalized Mann–Whitney–Wilcoxon statistic in (1) and $\theta = E(V_n)$, where $\binom{n}{2}$ denotes combinations of 2 distinct elements (i,j) from the integer set $\{1,..., n\}$. If y_{ki} have the same distribution, then $\theta = \frac{1}{2}$. Although the reverse is generally not true, $\theta = \frac{1}{2}$ is often considered as the null hypothesis of no between-group difference in practical applications. This connotation is adopted in the following discussion unless otherwise stated. Inference about H_0 has been discussed in the literature [5,3]. Below, we extend this classic test to more than two samples as well as longitudinal data. We start with a brief review of a new class of regression models, which forms the premise for such extensions.

Functional Response Models (FRM)

Existing semi-parametric (distribution-free) regression models are all defined based on a single-subject response. For example, the most popular linear regression model is defined by: $E(y_i \mid \mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\beta$, where y_i (x_i) denotes some response (a vector of predictors or covariates) and β is a vector of parameters. In this model, the response variable is a single-subject response y_i. Although the linear regression model has been extended for modeling more complex types of response variables such as binary, the fact remains that the specification of the model only involves a single subject response. For example, in the generalized linear model defined by: $E(y_i \mid \mathbf{x}_i) = h(\mathbf{x}_i^{\mathrm{T}}\beta)$, the right side is generalized to be a function of the linear predictor, $\mathbf{x}_i^{\mathrm{T}}\beta$, to accommodate the non-linear response y_i, but the left side remains identical to the linear model.

The inherent weakness of such single-subject-response-based regression models is their limited applications to modeling the moments of a response. As a result, many popular statistics that are complex functions of higher-order moments cannot be studied under the regression paradigm. The functional response models (FRM) address this limitation by extending 1) the single-subject y_i to multiple subject outcomes; and 2) the linear response to an arbitrary function:

$$E\left[f\left(y_{i_1},\ldots,y_{i_q}\right) \mid \mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}\right] = h\left(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\beta\right), \qquad (2)$$

where $f(.)$ is some functional, $h(.)$ some smooth functional (with continuous second-order derivatives). By generalizing the response variable in this fashion, this new class of models has been successfully applied to address a range of methodological issues involving second-order moments such as those arising in modeling reliability indices [6-9], modeling population mixtures [10] and structural equation models [11] as well as between-subject attributes such as in modeling social network connectivity [12,13] nonparametric inference for stochastic hypotheses in gene expressions [14] and causal inference for rank-based models [15]. Below, we focus on its application to the current context of extending the MWW test to multiple comparison groups within a longitudinal data setting.

FRM-based Multi-sample MWW Tests

Adopting the previous notation, let

$$f\left(y_{1i}, y_{2j}\right) = I_{\left\{y_{1i} \le y_{2j}\right\}}, \quad h(\theta) = \theta, \quad i \in C_1^{n_1}, \quad j \in C_1^{n_2} \qquad (3)$$

where $C_1^{n_k} = \{1,\ldots,n_k\}$. Consider the following FRM:

$$E\left[f\left(y_{1i}, y_{2j}\right)\right] = h(\theta) = \theta, \quad i \in C_1^{n_1}, \quad j \in C_1^{n_2}.$$

Under the null of no between-group difference, $H_0:\theta=\frac{1}{2}$. As will be seen in Section 3.1, this FRM yields the classic MWW test. By framing the MWW test under FRM, we are ready to extend this test to a multi-group setting.

Note that when ties are present, $H_0:\theta=\frac{1}{2}$ does not imply no difference between the groups. The null in this case may be expressed as [16]:

$$H_0: E\left(I_{\{y_{1i}<y_{2j}\}}\right)+\frac{1}{2}E\left(I_{\{y_{1i}=y_{2j}\}}\right)=\frac{1}{2}.$$

Thus, we may redefine the functional response in (3) as

$$f\left(y_{1i},y_{2j}\right)=I_{\{y_{1i}<y_{2j}\}}+\frac{1}{2}I_{\{y_{1i}=y_{2j}\}}.$$

For notational brevity, we assume continuous y_{ki} throughout the discussion unless otherwise noted.

Now consider K groups and let

$$\begin{aligned}&f\left(y_{ki},y_{lj}\right)=I_{\{y_{ki}\le y_{lj}\}},\quad h(\theta)=\theta_{kl},\\&\theta=\left(\theta_{12},\ldots,\theta_{1K},\theta_{23},\ldots,\theta_{2K},\ldots,\theta_{(K-1)K}\right)^{\mathrm{T}},\\&i\in C_1^{n_k},\quad j\in C_1^{n_l},\quad (k,l)\in C_2^K,\end{aligned}\tag{4}$$

where $C_2^K=\{(k,l);k,l\in\{1,\ldots,K\},k<l\}$ denotes all distinct ordered combinations of (k,l) from the integer set {1,2,..., K}. Consider the following FRM:

$$\begin{aligned}&E\left[f\left(y_{ki},y_{lj}\right)\right]=\\&h(\theta)=\theta_{kl},\quad i\in C_1^{n_k},\quad j\in C_1^{n_l},(k,l)\in C_2^K.\end{aligned}\tag{5}$$

If no difference exists across all the K samples, then

$\theta_{kl}=\frac{1}{2}$ for all $(k,l)\in C_2^K$ and vice versa. Thus, we can test the null hypothesis, $H_0:\theta_{kl}=\frac{1}{2}$, $(k,l)\in C_2^K$, to determine if there is any difference across the K samples. If this omnibus test is rejected, we may follow with pair-wise comparisons to identify the sources of differences.

We can also readily extend the FRM-based multiple MWW model to a longitudinal data setting. For convenience, consider a longitudinal study with only two groups and m assessments. Let y_{kit} denote the outcome from the i^{th} subject within the K th group at time t, and let $h_t(\theta_t)=\theta_t$. The FRM for the longitudinal data is defined by:

$$\begin{aligned}&E\left[f\left(y_{1it},y_{2jt}\right)\right]=\\&h_t(\theta),\quad \theta=(\theta_1,\theta_2,...,\theta_m)^{\mathrm{T}},\quad i\in C_1^{n_1},\quad j\in C_1^{n_2}.\end{aligned}\tag{6}$$

If there is no treatment difference over time, we have: $H_0:\theta_t=\frac{1}{2}$ for all t = 1,....,m.

INFERENCE FOR FRM-BASED MANN-WHITNEY-WILCOXON TESTS

We start with cross-sectional data and then extend the results to longitudinal data.

Cross-sectional Data

Consider K groups and let

$$\begin{aligned}&\mathbf{f_i}=\left(f\left(y_{1i_1},y_{2i_2}\right),\ldots,f\left(y_{1i_1},y_{Ki_K}\right),f\left(y_{2i_2},y_{3i_3}\right),\ldots,f\left(y_{(K-1)i_{K-1}},y_{Ki_K}\right)\right)^{\mathrm{T}}\\&\mathbf{h_i}=\left(h_{12},\ldots,h_{1K},h_{23},\ldots,h_{(K-1)K}\right)^{\mathrm{T}},\quad \theta=\left(\theta_{12},\ldots,\theta_{1K},\theta_{23},\ldots,\theta_{(K-1)K}\right)^{\mathrm{T}},\end{aligned}\tag{7}$$

Where $\mathbf{i}=(i_1,\ldots,i_K)\in\ C=C_1^{n_1}\otimes C_1^{n_2}\otimes\cdots\otimes C_1^{n_K}$ ($\otimes$ denotes the Cartesian product of $C_1^{n_k}$ $k=1,\ldots,K$). Define a set of U-statistics based generalized estimating equation (UGEE) as follows:

$$\mathbf{U}_n(\theta)=\sum_{\mathbf{i}\in C}\mathbf{U}_{n,\mathbf{i}}(\theta)=\sum_{\mathbf{i}\in C}G_{\mathbf{i}}S_{\mathbf{i}}=\mathbf{0},\tag{8}$$

Where $S_i = f_i - h_i$ and G_i is some known $\frac{K(K-1)}{2}\times\frac{K(K-1)}{2}$ matrix function of θ. Like the standard GEE [17] the choice of G_i is not unique. In most applications, we set $G_{\mathbf{i}} = D_{\mathbf{i}}V^{-1} = \left(\frac{\partial}{\partial\theta}\mathbf{h}_{\mathbf{i}}\right)V^{-1}$, where V denotes some known $\frac{K(K-1)}{2}\times\frac{K(K-1)}{2}$ such as $V = \mathbf{I}_{\frac{K(K-1)}{2}\times\frac{K(K-1)}{2}}$. Further, $V(\alpha)$ may be parameterized by some vector α. If α is unknown, it must be estimated before the UGEE in (8) can be solved for θ. The UGEE estimate $\hat{\theta}$ obtained as solution to (8) is consistent and asymptotically normal.

Although the consistency of $\hat{\theta}$ is independent of how α is estimated, the asymptotic normality of such estimates is guaranteed only when $\sqrt{n}$ -consistent estimates of α are used [3]. We summarize the asymptotic properties below.

Theorem 1. Let

$$\mathbf{v}_{ki} = E\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{ki}\right),\quad \Sigma_k = Var\left(\mathbf{v}_{ki}\right),\quad B = E\left(G_{\mathbf{i}}D_{\mathbf{i}}^{\mathrm{T}}\right),\quad D_{\mathbf{i}} = \frac{\partial}{\partial\theta}\mathbf{h}_{\mathbf{i}}, \qquad (9)$$

$$n = \sum_{k=1}^{K} n_k,\quad \rho_k^2 = \lim_{n\to\infty}\frac{n}{n_k} < \infty,\quad k = 1,\ldots,K.$$

Then, under mild regularity conditions, we have:

1. $\hat{\theta}$ is consistent.

2. If $\sqrt{n}\left(\hat{\alpha}-\alpha\right) = \mathbf{O}_p(1)$, $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \to_d N\left(\mathbf{0}, \Sigma_\theta = \sum_{k=1}^{K}\rho_k^2 B^{-1}\Sigma_k B^{-\mathrm{T}}\right). \qquad (10)$$

The asymptotic variance Σ_θ generally is not in closed-form, except under some special cases. For example, if $K = 2$, we have

$$\Sigma_1 = E\left[1-F_2\left(y_{1i}\right)\right]^2 - \theta,\quad \Sigma_2 = E\left[F_1\left(y_{2j}\right)\right]^2 - \theta, \qquad (11)$$

where Fk (y) $F_k(y)$ denotes the cumulative distribution function of y_{ki}. Further, under the null

$$H_0 : F_k(y) = F(y),\ \theta = \frac{1}{2}.$$

It follows from (10) and the fact that $F(y_{ki})$ is a uniform ^{U} between 0 and 1 that

$$\Sigma_1 = E\left(1-U\right)^2 - \frac{1}{4} = \frac{1}{12},\quad \Sigma_2 = \frac{1}{12}.$$

Thus in this special case,

$\Sigma_\theta = \frac{1}{12}\left(\rho_1^2 + \rho_2^2\right)$ and a consistent estimate is given by

$\hat{\Sigma}_\theta = \frac{1}{12}\left(\frac{n}{n_1} + \frac{n}{n_2}\right)$ ($n = n_1 + n_2$. These asymptotic results for the classic MWW test has been well documented in the literature [18,5].

For general K, a consistent estimate of Σ_θ is obtained by substituting respective consistent estimates in place of B and Σ_k. A consistent estimate of B is

$$\hat{B} = \frac{1}{\prod_l n_l}\sum\nolimits_{\mathbf{i}\in C} G_{\mathbf{i}}D_{\mathbf{i}}^{\mathrm{T}}.$$

To find a consistent estimate of Σ_k, first note that we can estimate

$$E\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{ki_k}\right)$$

by: $\hat{E}\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{ki_k}\right) = \frac{1}{\prod_{l\neq k} n_l}\sum_{l\neq k, 1\le i_l\le n_l}\mathbf{U}_{n,\left(i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_K\right)}\left(\hat{\theta}\right),\quad k = 1,\ldots,K. \qquad (12)$

Thus, a consistent estimate of Σ_k is given by:

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i_k=1}^{n_k}\hat{E}\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{ki_k}\right)\hat{E}^{\mathrm{T}}\left(\mathbf{U}_{n,\mathbf{i}} \mid y_{ki_k}\right). \qquad (13)$$

Under the null $H_0 : \theta = \frac{1}{2}\mathbf{1}_{\frac{(K-1)K}{2}}$, where 1_K denotes a K X 1 column vector of one's, we can evaluate (12) by substituting $\frac{1}{2}\mathbf{1}_{\frac{(K-1)K}{2}}$ in place of $\hat{\theta}$. We can test H_0 by a Wald-type statistic:

$$W = n\left(\hat{\theta} - \frac{1}{2}\mathbf{1}_{\frac{(K-1)K}{2}}\right)^{\mathrm{T}} \hat{\Sigma}_{\theta}^{-1}\left(\hat{\theta} - \frac{1}{2}\mathbf{1}_{\frac{(K-1)K}{2}}\right) \sim \chi^2_{\frac{(K-1)K}{2}}. \qquad (14)$$

As a special case with $K = 2$, $\theta = \frac{1}{2}$, $\hat{\Sigma}_{\theta}^{-1}$, is a scaler and (14) reduces to $W = n\hat{\Sigma}_{\theta}^{-1}(\hat{\theta} - \frac{1}{2})^2 \sim \chi^2_1$, a widely used asymptotic result for inference about group differences when using the MWW test [18,5].

Longitudinal data

We first consider the case of complete data and then generalize it to address missing values.

Complete data:

Let

$$\mathbf{y}_{ki_k} = \left(y_{ki_k1}, y_{ki_k2}, ..., y_{ki_km}\right)^{\mathrm{T}}, \quad \mathbf{f}_{\mathbf{i},t} = \mathbf{f}\left(y_{1i_1t}, y_{2i_2t}, \ldots, y_{Ki_Kt}\right) \qquad (15)$$

$$\mathbf{f}_{\mathbf{i}} = \left(\mathbf{f}_{\mathbf{i},1}^{\mathrm{T}}, \mathbf{f}_{\mathbf{i},2}^{\mathrm{T}}, ..., \mathbf{f}_{\mathbf{i},m}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \mathbf{h}_{\mathbf{i},t} = \left(h_{12,t}, \ldots, h_{1K,t}, h_{23,t}, \ldots, h_{(K-1)K,t}\right)^{\mathrm{T}},$$

$$\mathbf{h}_{\mathbf{i}} = \left(\mathbf{h}_{\mathbf{i},1}^{\mathrm{T}}, \mathbf{h}_{\mathbf{i},2}^{\mathrm{T}}, ..., \mathbf{h}_{\mathbf{i},m}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \theta_t = \left(\theta_{12,t}, \ldots, \theta_{1K,t}, \theta_{23,t}, \ldots, \theta_{(K-1)K,t}\right)^{\mathrm{T}},$$

$$\theta = \left(\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}}, ..., \theta_m^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Inference about θ does not create any complication and can be made using the same UGEE in (8), with f_i and $h_i(\theta)$ defined in (15). In particular, by setting $G_{\mathbf{i}} = \left(\frac{\partial}{\partial\theta}\mathbf{h}_{\mathbf{i}}\right)V^{-1}$ with $V = \mathbf{I}_{\frac{K(K-1)m}{2} \times \frac{K(K-1)}{2}m}$, we can solve the UGEE in closed form to obtain: $\hat{\theta} = \prod_{k=1}^{K} \frac{1}{n_k} \sum_{\mathbf{i} \in C} \mathbf{f}_{\mathbf{i}}$. As a special case, if $K = 2$, this reduces to the familiar MWW statistic at each assessment t:

$$\hat{\theta} = \frac{1}{n_1 n_2} \sum_{i \in C_1^{n_1}, j \in C_1^{n_2}} \left(I_{\left\{y_{1i1} \le y_{2j1}\right\}}, I_{\left\{y_{1i2} \le y_{2j2}\right\}}, \ldots, I_{\left\{y_{1im} \le y_{2jm}\right\}}\right)^{\mathrm{T}}. \qquad (16)$$

It follows from Theorem 1 that the UGEE estimate $\hat{\theta}$ in (16) is consistent and asymptotically normal. The asymptotic variance can again be estimated by (13). Because of the difference in the definition of f_i and h_i, the null of no difference between the K groups over time has a different expression, $H_0: \theta = \frac{1}{2}\mathbf{1}_{\frac{K(K-1)}{2}m}$, and accordingly the Wald statistic in (14) is given by:

$$W = n\left(\hat{\theta} - \frac{1}{2}\mathbf{1}_{\frac{K(K-1)}{2}m}\right)^{\mathrm{T}} \hat{\Sigma}_{\theta}^{-1}\left(\hat{\theta} - \frac{1}{2}\mathbf{1}_{\frac{K(K-1)}{2}m}\right) \sim \chi^2_{\frac{K(K-1)}{2}m}. \qquad (17)$$

As a special case with $K = 2$, $\theta = \frac{1}{2}\mathbf{1}_m$, (17) reduces to $W \sim \chi^2_m$.

Missing data:

Define a vector of missing (or rather observed) value indicators as follows:

$$r_{ki_kt} = \begin{cases} 1 & \text{if} y_{ki_kt} \text{is observed} \\ 0 & \text{if} y_{ki_kt} \text{is unobserved} \end{cases}, \quad \mathbf{r}_{ki_k} = \left(r_{ki_k1}, r_{ki_k2}, ..., r_{ki_km}\right)^{\mathrm{T}}. \qquad (18)$$

As in the literature, we assume no missing data at baseline $t = 1$ such that $r_{ki_k1} \equiv 1$ for all $i_k \in C_1^{n_k}$. Let

$$\pi_{ki_kt} = \Pr\left(r_{ki_kt} = 1 \mid \mathbf{y}_{ki_k}, \mathbf{x}_{ki_k}\right), \quad \mathbf{x}_{ki_k} = (\mathbf{x}_{ki_k1}^{\mathrm{T}}, ..., \mathbf{x}_{ki_km}^{\mathrm{T}})^{\mathrm{T}}, \qquad (19)$$

$$\varphi_{\mathbf{i},t} = \left(\varphi_{1i_1t}\varphi_{2i_2t}, \ldots, \varphi_{1i_1t}\varphi_{Ki_Kt}, \varphi_{2i_2t}\varphi_{3i_3t}, \ldots, \varphi_{(K-1)i_{K-1}t}\varphi_{Ki_Kt}\right)^{\mathrm{T}},$$

$$\varphi_{\mathbf{i}} = \left(\varphi_{\mathbf{i},1}^{\mathrm{T}}, \varphi_{\mathbf{i},2}^{\mathrm{T}}, \ldots, \varphi_{\mathbf{i},m}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \ddot{\mathbf{A}}_{\mathbf{i}} = diag\left(\varphi_{\mathbf{i}}\right), \quad \varphi_{ki_kt} = \frac{r_{ki_kt}}{\pi_{ki_kt}},$$

Where $diag(\varphi_i)$ denotes a diagonal matrix with φ_i forming the diagonal, $\mathbf{y}_{ki_k}$ is defined in (15), and $\mathbf{x}_{ki_k}$ denotes a vector of other variables collected. Since $r_{ki_k1} \equiv 1$, $\pi_{ki_k1} = 1$ for all $i_k \in C_1^{n_k}$.

In most applications, π_{ki_kt} $(t = 2, \ldots, m)$ is unknown and must be estimated. Under MCAR, π_{ki_kt} is independent of $\mathbf{y}_{ki_k}$ and thus $\pi_{ki_kt} = \Pr\left(r_{ki_kt} = 1\right) = \pi_{kt}$. In this case, π_{kt} is a constant independent of $\mathbf{y}_{ki_k}$ and is readily estimated by the sample moment:

$$\pi_{kt} = \frac{1}{n_k}\sum_{i_k=1}^{n_k} r_{ki_kt}, \quad t = 2,\ldots,m; \quad k = 1,\ldots,K.$$

Under MAR, π_{ki_kt} becomes dependent on the observed y_{ki_kt} and $\mathbf{x}_{ki_kt}$, which within the current context contain all y_{ki_ks} and $\mathbf{x}_{ki_ks}$ for s = 1,...,t-1. Denote such a " history" by $\mathbf{z}_{ki_kt^-} = (\mathbf{y}_{ki_ks}, \mathbf{x}_{ki_ks}; s = 1,\ldots,t-1)^{\mathrm{T}}$. Then under MAR,

$$\pi_{ki_kt} = \begin{cases} 1 & \text{if } t = 1, \\ \Pr\left(r_{ki_kt} = 1 \mid \mathbf{z}_{ki_kt^-}\right) & \text{if } t = 2,\ldots,m. \end{cases} \qquad (20)$$

Unlike the definition in (19), r_{ki_kt} does not depend on y_{ki_kt} and $\mathbf{x}_{ki_kt}$, making it possible to model π_{ki_kt} in (20). However, it is still difficult to model and estimate π_{ki_kt} without imposing the monotone missing data pattern (MMDP) assumption, because of the large number of missing data patterns [3,16]. Under MMDP, y_{ki_kt} and $\mathbf{x}_{ki_kt}$ are observed only if all y_{ki_ks} and $\mathbf{x}_{ki_ks}$ prior to time t are all observed. The structured patterns reduce not only the number of missing data patterns, but also the complexity in modeling π_{kit}.

Let $p_{ki_kt} = E\left(r_{ki_kt} = 1 \mid r_{ki_k(t-1)} = 1, \mathbf{z}_{ki_kt^-}\right)$ denote the one-step transition probability from observing the response at t-1 to t. We can readily model p_{ki_kt} using a logistic regression model:

$$\mathrm{logit}\left(p_{ki_kt}\right) = g_{kt}\left(\gamma_{kt}, \mathbf{z}_{ki_kt^-}\right) = \xi_{kt} + \eta_{kt}^{\mathrm{T}}\mathbf{z}_{ki_kt^-}, \quad t = 2,\ldots,m, \qquad (21)$$

where $\gamma_{kt} = \left(\xi_{kt}, \eta_{kt}^{\mathrm{T}}\right)^{\mathrm{T}}$ denotes the model parameters. More complex forms of $g_{kt}\left(\gamma_{kt}, \mathbf{z}_{ki_kt^-}\right)$ such as those involving interactions of the components of $\mathbf{z}_{ki_kt^-}$ are similarly considered. Under MMDP, it is readily checked that

$$\pi_{ki_kt}(\gamma_k) = \Pr\left(r_{ki_k(t-1)} = 1 \mid \mathbf{z}_{ki_kt^-}\right) = \prod_{s=2}^{t} p_{ki_ks}(\gamma_{ks}), \quad t = 2,\ldots,m, \qquad (22)$$

where $\gamma_k = \left(\gamma_{k2}^{\mathrm{T}},\ldots,\gamma_{km}^{\mathrm{T}}\right)^{\mathrm{T}}$.

At this point, we can proceed in one of two ways. We can either estimate π_{ki_kt} from p_{ki_kt} in (21) using the relationship in (22) and incorporate such information into the UGEE in (8) or define a new FRM to model y_{ki_kt} and r_{ki_kt} simultaneously. We discuss both approaches below.

To estimate γ_k, we can use the following estimating equations based on maximum likelihood:

$$\mathbf{Q}_{n,k}(\gamma_k) = \sum_{i_k=1}^{n_k}\mathbf{Q}_{n,ki_k} = \mathbf{0}, \quad \mathbf{Q}_{n,ki_k} = \left(\mathbf{Q}_{n,ki_k2}^{\mathrm{T}},\ldots,\mathbf{Q}_{n,ki_km}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad k = 1,\ldots,K, \qquad (23)$$

where

$$\mathbf{Q}_{n,ki_kt} = \frac{\partial}{\partial\gamma_{kt}}\left\{r_{ki_k(t-1)}\left[r_{ki_kt}\log\left(p_{ki_kt}\right) + \left(1 - r_{ki_kt}\right)\log\left(1 - p_{ki_kt}\right)\right]\right\}, \qquad (24)$$

$t = 2,\ldots,m, \quad i_k \in C_1^{n_k}, \quad k = 1,\ldots,K.$

Let

$$\mathbf{W}_{n,ki_k}(\gamma_k) = \left(\mathbf{0}_{(m-1)\times(k-1)}^{\mathrm{T}}, \mathbf{Q}_{n,ki_k}^{\mathrm{T}}, \mathbf{0}_{(m-1)\times(K-k)}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad k = 1,\ldots,K,$$

Where $\mathbf{0}_J$ denotes a J×1 column vector of 1's. Let $\gamma = \left(\gamma_2^{\mathrm{T}},\ldots,\gamma_m^{\mathrm{T}}\right)^{\mathrm{T}}$. We may express (23) in a compact form:

$$\mathbf{W}_n(\gamma) = \sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\mathbf{W}_{n,ki_k}(\gamma_k) = \left(\sum_{i_1=1}^{n_1}\mathbf{Q}_{n,1i_1},\ldots,\sum_{i_K=1}^{n_K}\mathbf{Q}_{n,Ki_K}\right)^{\mathrm{T}} = \mathbf{0}.$$

To incorporate the estimated π_{ki_kt} into the estimate of θ, we first revise the UGEE to create a set of U-statistics-based weighted generalized estimating equations (UWGEE):

$$\mathbf{U}_n(\theta) = \sum_{\mathbf{i}\in C}\mathbf{U}_{n,\mathbf{i}} = \sum_{\mathbf{i}\in C}G_{\mathbf{i}}\boldsymbol{\Delta}_{\mathbf{i}}S_{\mathbf{i}} = \sum_{\mathbf{i}\in C}G_{\mathbf{i}}\boldsymbol{\Delta}_{\mathbf{i}}\left(\mathbf{f}_{\mathbf{i}} - \mathbf{h}_{\mathbf{i}}\right) = \mathbf{0}, \qquad (25)$$

Where G_i has the same interpretation as in the complete data case. For example, for K=2,

$$\pi_{ki_k t} = \Pr\left(r_{ki_k t} = 1 \mid \mathbf{z}_{ki_k t^-}\right), \quad \varphi_{ki_k t} = \frac{r_{ki_k t}}{\pi_{ki_k t}}, \quad \varphi_{\mathbf{i},t} = \varphi_{1i_1 t}\varphi_{2i_2 t}, \tag{26}$$

$$\varphi_{\mathbf{i}} = \left(\varphi_{\mathbf{i},1}, \varphi_{\mathbf{i},2}, \ldots, \varphi_{\mathbf{i},m}\right)^{\mathrm{T}}, \quad \boldsymbol{\Delta}_{\mathbf{i}} = diag\left(\varphi_{\mathbf{i}}\right),$$

$$\mathbf{W}_n(\gamma) = \sum_{i=1}^{n_1} \mathbf{W}_{n,1i}(\gamma_1) + \sum_{j=1}^{n_2} \mathbf{W}_{n,2j}(\gamma_2),$$

$$\mathbf{W}_{n,1i}(\gamma_1) = \left(\mathbf{Q}_{n,1i}^{\mathrm{T}}, \mathbf{0}_{m-1}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \mathbf{W}_{n,2j}(\gamma_2) = \left(\mathbf{0}_{m-1}^{\mathrm{T}}, \mathbf{Q}_{n,2j}^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Like UGEE, UWGEE is a generalization of the weighted generalized estimating equations (WGEE) for inference about distribution-free regression models [3] Since $\pi_{ki_k t}$ are estimated, we need to account for sampling variability when estimating θ, from (25). The theorem below not only shows how to accomplish this task, but guarantee the consistent and asymptotic normality properties of the UWGEE estimates as well.

Theorem 2. Let $\hat{\theta}$ denote the UWGEE estimate from solving the estimating equations in (25). Let

$$\mathbf{v}_{ki_k} = E\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{y}_{ki_k}, \mathbf{r}_{ki_k}\right), \quad \Sigma_k = Var\left(\mathbf{v}_{ki_k}\right), \quad B = E\left(G_{\mathbf{i}}\boldsymbol{\Delta}_{\mathbf{i}} D_{\mathbf{i}}^{\mathrm{T}}\right), \tag{27}$$

$$C = E\left[\frac{\partial}{\partial\gamma}\left(G_{\mathbf{i}}\boldsymbol{\Delta}_{\mathbf{i}} S_{\mathbf{i}}\right)\right]^{\mathrm{T}}, \quad H = \sum_{k=1}^{K} \rho_k^2 E\left(\frac{\partial}{\partial\gamma}\mathbf{W}_{n,ki_k}(\gamma_k)\right)^{\mathrm{T}},$$

$$\Phi_k = CH^{-1}Var\left(\mathbf{W}_{n,ki_k}\right)H^{-\mathrm{T}}C^{\mathrm{T}} - E\left(\mathbf{v}_{ki_k}\mathbf{W}_{ki_k}^{\mathrm{T}}H^{-\mathrm{T}}C^{\mathrm{T}}\right) - \left[E\left(\mathbf{v}_{ki_k}\mathbf{W}_{ki_k}^{\mathrm{T}}H^{-\mathrm{T}}C^{\mathrm{T}}\right)\right]^{\mathrm{T}},$$

$$n = \sum_{k=1}^{K} n_k, \quad \rho_k^2 = \lim_{n\to\infty}\frac{n}{n_k} < \infty, \quad k = 1, \ldots, K.$$

Then, under mild regularity conditions and a $\sqrt{n}$ -consistent estimates of $\hat{\theta}$,

1. $\hat{\theta}$ is consistent.

2. If $\sqrt{n}\left(\hat{\alpha} - \alpha\right) = \mathbf{O}_p(1)$, $\hat{\theta}$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N\left(\mathbf{0}, \Sigma_\theta = \sum_{k=1}^{K} \rho_k^2 B^{-1}\left(\Sigma_k + \Phi_k\right)B^{-\mathrm{T}}\right). \tag{28}$$

The asymptotic variance above is almost identical to its counterpart in (10), except for an added term Φ_k. This extra Φ_k accounts for additional variability due to estimation of γ.

A consistent estimate of Σ_θ is obtained by substituting consistent estimates in place of the respective quantities in (28). For example, the following are consistent estimates of the components in (28):

$$\hat{B} = \left[\prod_{k=1}^{K}\binom{n_k}{1}\right]^{-1}\sum_{\mathbf{i}\in C}\hat{G}_{\mathbf{i}}\hat{\Delta}_{\mathbf{i}}\hat{D}^{\mathrm{T}}, \quad \hat{C} = \left[\prod_{k=1}^{K}\binom{n_k}{1}\right]^{-1}\sum_{\mathbf{i}\in C}\left[\frac{\partial}{\partial\gamma}\left(\hat{G}_{\mathbf{i}}\hat{\Delta}_{\mathbf{i}}\hat{S}_{\mathbf{i}}\right)\right]^{\mathrm{T}},$$

$$\hat{H} = \sum_{k=1}^{K}\frac{n}{n_k}\left(\frac{1}{n_k}\sum_{i_k=1}^{n_k}\frac{\partial}{\partial\gamma}\mathbf{W}_{n,ki_k}(\gamma_k)\right), \quad \hat{v}_{ki_k} = \hat{E}\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{y}_{ki_k}, \mathbf{r}_{ki_k}\right),$$

$$\hat{\Phi}_k = \hat{C}\hat{H}^{-1}\hat{Var}\left(\mathbf{W}_{n,ki_k}\right)\hat{H}^{-\mathrm{T}}\hat{C}^{\mathrm{T}} - \frac{1}{n_k}\sum_{i_k=1}^{n_k}\hat{v}_{ki_k}\hat{W}_{ki_k}^{\mathrm{T}}\hat{H}^{-\mathrm{T}}\hat{C}^{\mathrm{T}} - \frac{1}{n_k}\sum_{i_k=1}^{n_k}\left(\hat{v}_{ki_k}\hat{W}_{ki_k}^{\mathrm{T}}\hat{H}^{-\mathrm{T}}\hat{C}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i_k=1}^{n_k}\hat{v}_{ki_k}\hat{v}_{ki_k}^{\mathrm{T}},$$

where $\hat{A}$ denotes the quantity by substituting $\hat{\beta}$, $\hat{\alpha}$, and $\hat{\gamma}$ in the respective parameters in A.

Alternatively, with the flexibility of FRM, we can readily define a new FRM to concurrently model $y_{ki_k t}$ and $r_{ki_k t}$. For notational brevity, consider only two groups and let

$$E\left[f_{1it}\left(y_{1it},y_{2jt}\right)\right]=h_{1it}\left(\theta,\gamma\right),\quad E\left[f_{2it}\left(r_{1it},r_{2jt}\right)|\mathbf{z}_{1it^-},\mathbf{z}_{1jt^-}\right]=h_{2it}\left(\theta,\gamma\right),\qquad(29)$$

$$E\left[f_{3it}\left(r_{1it},r_{2jt}\right)|\mathbf{z}_{2it^-},\mathbf{z}_{2jt^-}\right]=h_{3it}\left(\theta,\gamma\right),\quad f_{1it}\left(y_{1it},y_{2jt}\right)=I_{\left\{y_{1it}\le y_{2jt}\right\}},$$

$$f_{2it}\left(y_{1it},y_{2jt}\right)=r_{1it},\quad f_{3it}\left(y_{1it},y_{2jt}\right)=r_{2jt},\quad h_{1it}\left(\theta,\gamma\right)=\theta_t,$$

$$\text{logit}\left(h_{2it}\left(\theta,\gamma\right)\right)=\xi_{1t}+\eta_{1t}^{\mathrm{T}}\mathbf{z}_{1it^-},\quad \text{logit}\left(h_{3it}\left(\theta,\gamma\right)\right)=\xi_{2t}+\eta_{2t}^{\mathrm{T}}\mathbf{z}_{2jt^-}.$$

Unlike the preceding approach, θ and γ are both the parameters of the FRM in (29). With the redefined G_i, Δ_i, f_i and h_i below, the UWGEE in (25) can again be used to provide simultaneously inference about $\zeta=\left(\theta^{\mathrm{T}},\gamma^{\mathrm{T}}\right)^{\mathrm{T}}$:

$$G_{\mathbf{i}}=D_{\mathbf{i}}V_{\mathbf{i}}^{-1},\quad \mathbf{f_i}=\left(\mathbf{f}_{1\mathbf{i}}^{\mathrm{T}},\mathbf{f}_{2\mathbf{i}}^{\mathrm{T}},\mathbf{f}_{3\mathbf{i}}^{\mathrm{T}}\right)^{\mathrm{T}},\quad \mathbf{h_i}\left(\theta\right)=\left(\mathbf{h}_{1\mathbf{i}}^{\mathrm{T}},\mathbf{h}_{2\mathbf{i}}^{\mathrm{T}},\mathbf{h}_{3\mathbf{i}}^{\mathrm{T}}\right)^{\mathrm{T}},\quad \mathbf{f}_{l\mathbf{i}}=\left(f_{l\mathbf{i}1},\ldots,f_{l\mathbf{i}m}\right)^{\mathrm{T}},\qquad(30)$$

$$\mathbf{h}_{l\mathbf{i}}=\left(h_{l\mathbf{i}1},\ldots,h_{l\mathbf{i}m}\right)^{\mathrm{T}},\quad l=1,2,3,$$

$$\pi_{1it}=\begin{cases}1 & \text{if}t=1\\ \prod_{s=2}^{t}h_{2s\mathbf{i}}\left(\theta\right) & \text{if}t=2,\ldots,m'\end{cases},\quad \pi_{2jt}=\begin{cases}1 & \text{if}t=1\\ \prod_{s=2}^{t}h_{3s\mathbf{i}}\left(\theta\right) & \text{if}t=2,\ldots,m'\end{cases}$$

$$D_{\mathbf{i}}=\frac{\partial}{\partial\theta}\mathbf{h_i},\quad \mathbf{\Delta_i}=\begin{pmatrix}diag\left(\varphi_{\mathbf{i}}\right) & 0 & 0\\ 0 & \mathbf{I}_{m-1} & 0\\ 0 & 0 & \mathbf{I}_{m-1}\end{pmatrix},\quad V_{\mathbf{i}}^{-1}=\begin{pmatrix}\mathbf{I}_m & 0 & 0\\ 0 & V_{2\mathbf{i}}^{-1} & 0\\ 0 & 0 & V_{3\mathbf{i}}^{-1}\end{pmatrix},$$

$$\tilde{V}_{2\mathbf{i}}=\begin{pmatrix}h_{22\mathbf{i}}\left(1-h_{22\mathbf{i}}\right) & \cdots & 0\\ \cdots & \ddots & \cdots\\ 0 & \cdots & h_{2m_{1\mathbf{i}}\mathbf{i}}\left(1-h_{2m_{1\mathbf{i}}\mathbf{i}}\right)\end{pmatrix},\quad V_{2\mathbf{i}}=\begin{pmatrix}\tilde{V}_{2\mathbf{i}} & 0\\ 0 & \mathbf{I}_{m-m_{1\mathbf{i}}}\end{pmatrix},$$

$$\tilde{V}_{3\mathbf{i}}=\begin{pmatrix}h_{32\mathbf{i}}\left(1-h_{32\mathbf{i}}\right) & \cdots & 0\\ \cdots & \ddots & \cdots\\ 0 & \cdots & h_{3m_{2\mathbf{i}}\mathbf{i}}\left(1-h_{3m_{2\mathbf{i}}\mathbf{i}}\right)\end{pmatrix},\quad V_{3\mathbf{i}}=\begin{pmatrix}\tilde{V}_{3\mathbf{i}} & 0\\ 0 & \mathbf{I}_{m-m_{2\mathbf{i}}}\end{pmatrix},$$

where $m_{k\mathbf{i}}=\max\left\{t;r_{kit}=1,t=1,\ldots,m\right\}+1$ and φ_i is defined in (26).

We can again apply Theorem 2 to characterize the asymptotic behavior of the estimate $\hat{\zeta}$, except that we no longer need to adjust for the variability of estimated $\hat{\gamma}$, since the latter is estimated together with θ. Under mild regularity conditions, $\hat{\zeta}$ is asymptotically normal:

$$\sqrt{n}\left(\hat{\zeta}-\zeta\right)\to_d N\left(\mathbf{0},\Sigma_\zeta=B^{-1}\left(\sum_{k=1}^{K}\rho_k^2\Sigma_k\right)B^{-\mathrm{T}}\right),$$

where ρ_k^2, Σ_k and B are defined in (27). To find Σ_ζ, note that

$$\Sigma_k=Var\left(\mathbf{v}_{ki_k}\right)=E\left(\mathbf{v}_{ki_k}\mathbf{v}_{ki_k}^{\mathrm{T}}\right),\quad k=1,2.$$

Thus,

$$\Sigma_1=\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{v}_{1i}\hat{v}_{1i}^{\mathrm{T}},\quad \hat{v}_{1i}=\frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{U}_{n,ij},\quad \hat{\Sigma}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}\hat{v}_{2j}\hat{v}_{2j}^{\mathrm{T}},\quad \hat{v}_{2j}=\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{U}_{n,ij}.$$

Also, we estimate B by:

$$\hat{B}=\binom{n_1}{1}^{-1}\binom{n_2}{1}^{-1}\sum_{\mathbf{i}\in C}G_{\mathbf{i}}\ddot{\mathbf{A}}_{\mathbf{i}}D_{\mathbf{i}}^{\mathrm{T}}.$$

Thus, a consistent estimate of Σ_θ is given by:

$$\hat{\Sigma}_\zeta = \hat{B}^{-1}\left(\frac{n}{n_1}\hat{\Sigma}_1 + \frac{n}{n_2}\hat{\Sigma}_2\right)\hat{B}^{-\mathrm{T}}.$$

Alternatively, since

$\Sigma_1 = E\left[E\left(\mathbf{U}_{n,\mathbf{i}} \mid \mathbf{y}_{1i}, \mathbf{r}_{1i}\right) E\left(\mathbf{U}_{n,\mathbf{i}}^{\mathrm{T}} \mid \mathbf{y}_{1i}, \mathbf{r}_{1i}\right)\right] = E\left(\mathbf{U}_{n,ij}\mathbf{U}_{n,ik}^{\mathrm{T}}\right)$, a consistent estimate of Σ_1 is given by the U-statistic:

$$\hat{\Sigma}_1 = \binom{n_1}{1}^{-1}\binom{n_2}{2}^{-1}\sum_{i\in C_1^{n_1}}\sum_{(j,k)\in C_2^{n_2}}\mathbf{V}_{ijk},$$ where $\mathbf{V}_{ijk}$ is a symmetric version $\mathbf{U}_{n,ij}\mathbf{U}_{n,ik}^{\mathrm{T}}$ with respect to permutations of (*j*,*k*), i.e., $\tilde{V}_{ijk} = \tilde{V}_{ij'k}$, [3]. A similar estimate is obtained for $\hat{\Sigma}_2$.

APPLICATIONS

We demonstrate our considerations with both simulated and real data. We first investigate the performance of the proposed approach by simulation and then present an application to a real study on sexual health for a group of teenage girls in low-income urban settings who were at elevated risk for HIV, sexually transmitted infections (STIs), and unintended pregnancies. In all the examples, we applied the second approach for inference as discussed in Section 3.2 and set the statistical significance at = 0.05. All analyses were carried out using codes developed by the authors for implementing the models considered using the Matlab software [17].

Simulation study

We conducted a simulation study to examine the performance of the proposed FRM-based multi-sample Mann-Whitney-Wilcoxon Model for longitudinal data analysis. The data were simulated from a longitudinal study with two groups and three assessments under both complete and missing data. For space consideration, we only report results for three sample sizes, $n_1(=n_2)$=50, 100, and 300, representing small, moderate and large sample sizes, respectively. All simulations were performed with a Monte Carlo sample of 1,000.

We first simulated $\mathbf{y}_{ki} = \left(y_{ki1}, y_{ki2}, y_{ki3}\right)^{\mathrm{T}}$ K = (1, 2) from a trivariate normal, $N\left(\mathbf{0}, C_3(0.5)\right)$, with $C_3(0.5)$ denoting a compound symmetry correlation matrix (Kowalski and Tu, 2007). We modeled the data using the FRM in (6) with K = 2 and m = 3. For complete data, by applying UGEE, we obtain from (16) the following estimate of $\theta = \left(\theta_{121}, \theta_{122}, \theta_{123}\right)^{\mathrm{T}}$:

$$\hat{\theta} = \frac{1}{n_1 n_2}\sum_{(i,j)\in C_1^{n_1}\otimes C_1^{n_2}}\mathbf{f}_{ij} = \left(I_{\left\{y_{1i1}\le y_{2j1}\right\}}, I_{\left\{y_{1i2}\le y_{2j2}\right\}}, I_{\left\{y_{1i3}\le y_{2j3}\right\}}\right)^{\mathrm{T}}.$$

The asymptotic distribution of $\hat{\theta}$ is given by:

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \to_d N\left(\mathbf{0}, \Sigma_\theta = \frac{1}{2}\left(\Sigma_1 + \Sigma_2\right)\right),\quad \Sigma_k = E\left[E\left(\mathbf{U}_{n,(i,j)} \mid \mathbf{y}_{ki}\right)E^{\mathrm{T}}\left(\mathbf{U}_{n,(i,j)} \mid \mathbf{y}_{ki}\right)\right].$$

Under the null of no between-group difference over time, H_0, $\theta = \theta_0 = \frac{1}{2}\mathbf{1}_3$ (a 3×1 vector of 1's), a consistent estimate of Σ_k is given by:

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i_k=1}^{n_k}\hat{E}\left(\mathbf{U}_{n,(i,j)} \mid \mathbf{y}_{ki}\right)\hat{E}^{\mathrm{T}}\left(\mathbf{U}_{n,(i,j)} \mid \mathbf{y}_{ki}\right),$$

$$\hat{E}\left(\mathbf{U}_{n,(i,j)} \mid \mathbf{y}_{ki}\right) = \begin{cases}\dfrac{1}{n_2}\displaystyle\sum_{1\le j\le n_2}\mathbf{U}_{n,(1i,2j)}\left(\theta_0\right) & \text{if } k=1\\ \dfrac{1}{n_1}\displaystyle\sum_{1\le j\le n_1}\mathbf{U}_{n,(1j,2i)}\left(\theta_0\right) & \text{if } k=2\end{cases}.$$

Thus, it follows from Theorem 1 that the Wald statistic,

$$W = n\left(\hat{\theta}-\theta\right)^{\mathrm{T}}\hat{\Sigma}_\theta^{-1}\left(\hat{\theta}-\theta\right) \sim \chi_3^2.$$

For the missing data case, we assumed no missing value at baseline t = 1and simulated the missing response $r_{ki,t}$ at post-baseline under MAR according to (21) with the transition probability $p_{ki_k t}$ modeled by the logistic regression under a one-step Markov condition below:

$$\text{logit}\left(p_{ki_k t}\right) = \xi_{kt} + \eta_{kt}y_{ki_k(t-1)},\quad t = 2,3,\quad k = 1,2. \qquad (31)$$

Under (31), missingness only depends on the most recently observed response $y_{ki_k(t-1)}$ prior to time *t*. Although this assumption was used only for convenience purposes, it provides a reasonable model for most real studies.

We set $\eta_{kt}=3$ *t = 2,3* and solved the following equations for ξ_{kt} to create about 15% and 25% missing responses y_{ki_kt} at time *t = 2,3*, respectively,

$$\sum_{i_k=1}^{n_k} p_{ki_k2}=0.85n_k,\quad \sum_{i_k=1}^{n_k} p_{ki_k2}p_{ki_k3}=0.75n_k. \qquad (32)$$

To ensure MMDP, we first simulated the missing data indicator r_{ki_k2} from the Bernoulli distribution, $Bern\left(p_{ki_k2}\right)$ (k = 1,2). Then, we simulate r_{ki_k3} by conditioning on r_{ki_k2}, i.e., setting $r_{ki_k3}=0$ if $r_{ki_k2}=0$ and simulating $r_{ki_k3}\sim Bern\left(p_{ki_k3}\right)$ otherwise.

Shown in Table 1 are the UGEE and UWGEE estimates of θ, along with standard errors and type I errors for the complete and missing data cases based on 1,000 MC replications. For missing data under MAR, we used (a) the FRM in (15) with inference based on the UWGEE in (25) and Theorem 2, and (b) the FRM in (29) for jointly modeling $y_{ki_k(t-1)}$ and $r_{ki_k,t}$ with inference based on UWGEE in (25), but redefined G_i, Δ_i, f_i and h_i in (30). Since the results were quite similar, only the ones from the latter approach were reported. As well, only estimates of θ were shown in the table, as they are of primary interest. The results from the logistic regression in (31) for the missing data were quite close to the true values set for the simulation.

As seen, both the UGEE and UWGEE estimates of $\hat{\theta}$ were quite accurate, even for the small sample size n_k=50. The standard errors showed a stead decrease as n_k increased. Also, the corresponding standard errors were slightly larger for the UWGEE estimates because of the loss of information due to missing data. The type I error rates based on the Wald statistic showed a small upward for the small sample size n_k=50, which is typical of the anti-conservative behavior of this statistic, [18-22,9] but the bias disappeared at the larger sample size n_k = 100 and 300.

Real study

Teenage girls in low-income urban settings are at elevated risk for HIV, sexually transmitted infections (STIs), and unintended pregnancies. A randomized controlled trial was recently conducted to evaluate the efficacy of a sexual risk-reduction intervention, supplemented with post-intervention booster sessions, targeting low-income, urban, sexually active teenage girls [1]. The study recruited sexually-active urban adolescent girls aged 15-19 from the Rochester, New York, a mid-size, northeastern U. S. city, and randomized them to a theory-based, sexual risk reduction intervention or to a structurally-equivalent health promotion control group. Assessments and behavioral data were collected at baseline, and again at 3 and 6 months post-intervention. The primary interest of the study is to compare frequency of unprotected vaginal sex between the intervention and controlled condition. More details about the demographic characteristics of the study population, the treatment conditions and the assessment battery can be found in [1].

As mentioned in Section 1, a difficult problem with the data are the extremely large values some subjects reported with respect to their sexual activities. For example, seven subjects reported over 100 episodes of unprotected vaginal sex over the past 3 months at the 3 month follow-up, with the largest one being 1,000,000. A common approach to this issue in psychosocial research is to trim such outliers using some ad-hoc rules such as the one based on trimming large values by setting such outliers at 3 times the standard deviation of the outcome [19,1]. However, these methods induce artifacts, because of their dependence on the specific rules used and subjective criteria used in each method. Rank-based approaches such as the proposed FRM model address this issue in a much more objective fashion.

Table 1: UGEE (for complete data) and UWGEE (for missing data) estimates, standard errors and type I errors for testing no effect of time for a simulated longitudinal stuty with 2 groups and 3 assessments under complete data and missing data with MAR.

Estimates of θ (standard error) and type I errors from simulated data				
	Complete data (UGEE)			
Sample Size (per group)	θ_{121}	θ_{122}	θ_{123}	**Type I error**
				$H_0:\theta_{12}=\frac{1}{2}\mathbf{1}_3$
50	0.497 (0.057)	0.502 (0.058)	0.501 (0.059)	0.054
100	0.499 (0.042)	0.502 (0.045)	0.502 (0.048)	0.051
300	0.499 (0.023)	0.500 (0.024)	0.499 (0.025)	0.054
	Missing data under MAR (UWGEE)			
50	0.500 (0.061)	0.504 (0.071)	0.499 (0.069)	0.069
100	0.502 (0.043)	0.501 (0.057)	0.501 (0.049)	0.051
300	0.500 (0.024)	0.499 (0.037)	0.500 (0.029)	0.051

Our analysis was based on the 639 subjects who completed at least one of the three assessments (n_1=310,49% in the control and n_2=329,51% in the intervention). We were interested in comparing the two treatment groups for the unprotected vaginal sex. Let $y_{ki_k t}$ $\left(r_{ki_k,t}\right)$ denote such an outcome (indicator for missing data) at time t for the kth treatment, with t=1, 2 and 3 denoting the baseline, 3 and 6 months post-intervention and k=1 for the control and 2 for the intervention. As in the simulation study, the UWGEE estimates of θ and γ were obtained from the FRM in (29) by jointly modeling $y_{ki_k t}$ and $r_{ki_k,t}$. However, since some of $y_{ki_k t}$ were extremely large, we used the rankings $R_{ki_i(t-1)}$, rather than the actual values of $y_{ki_k(t-1)}$, as the predictor in the logistic regression for missing data:

$$\text{logit}\left(h_{2it}(\theta,\gamma)\right)=\xi_{1t}+\eta_{1t}R_{1i(t-1)},\quad \text{logit}\left(h_{3it}(\theta,\gamma)\right)=\xi_{2t}+\eta_{2t}R_{2j(t-1)},\quad t=2,3. \qquad (33)$$

Also, as ties are inevitable for the intrinsically discrete $y_{ki_k t}$ within our context, we used $f_{1it}\left(y_{1it},y_{2jt}\right)=I_{\left\{y_{1it}<y_{2jt}\right\}}+\frac{1}{2}I_{\left\{y_{1it}=y_{2jt}\right\}}$ as the functional response in the FRM to account for their presence.

Shown in Table 2 are the UWGEE estimates of the intercept ($_{1t}$) and slope (η_{1t}) from the fitted logistic regression component in (29) for modeling missingness at 3 and 6 months post-intervention. The occurrence of missing data did not depend on the observed (ranking of the) outcome at the prior visit for either treatment condition, suggesting no evidence for rejecting MCAR. Note that the one-step Markov condition was again adopted in (33). This assumption appeared to be sufficient, since we also tried to include $R_{1i(t-1)}$ in the second logistic model in (33), but $R_{1i(t-1)}$ was not a significant predictor.

Shown in Table 3 are the estimated θ, standard errors and p-values for testing the null of no between-group difference at each assessment time, along with the test statistic and p-value for testing the null of no temporal trend over post-intervention. The estimated θ_{12t} showed a steady decrease, implying that the likelihood for those in the intervention condition to engage in unprotected vaginal sex over time from baseline to 3 and 6 months post-intervention. The decline was significant at 6 months, but was a trend at 3 months (p-value is close to 0.10). The continued decline from 3 to 6 months was confirmed by the near significant p-value for testing the null of no temporal trend over post-intervention $H_0: \theta_{122} = \theta_{123}$. The increased gain of the intervention effect at 6 months was likely due to two booster sessions the study subjects received at 3 months [1]. The booster sessions, 90 minute long (as opposed to four 120 minute regular sessions delivered during the intervention), address behavioral patterns of girls that are expected to occur as they age and can promote maintenance of gains observed with health-behavior interventions [24].

Table 2: UWGEE estimates of parameters of logistic regression for modeling missingness at 3 and 6 months post-intervention for the randomized controlled trial on sexual health.

Parameter estimates of logistic regression for occurrence of missing data from the RCT on sexual health				
	Month 3		**Month 6**	
	Intercept	Prior response	Intercept	Prior response
	Control/Treat	Control/Treat	Control/Treat	Control/Treat
	Unprotected vaginal sex			
estimate	1.386/20.3 6	0.001/ -0.0006	2.246/2.42 4	0.002/0.00 3
standard error	0.349/0.37 0	0.001/0.001	0.488/0.57 1	0.002/0.00 3
p-value	< 0.001/ < 0.001	0.157/0.57 1	< 0.001/ <0.001	0.436/0.28 0

Table 3: UWGEE estimates of parameters for comparing the number of unprotected vaginal sex from baseline to 3 to 6 months post-intervention for the randomized controlled trial on sexual health.

Estimates, standard errors and p-values from the RCT on sexual health			
Baseline	**Estimate**	**Standard error**	**p-value** $\left(H_0:\theta_{12t}=\frac{1}{2}\right)$
Baseline	0.508	0.566	0.733
month	0.462	0.600	0.112
month	0.441	0.616	0.017
Test statistic (p-value) for testing no differential treatment effect			
between 3 and 6 months: $H_0:\theta_{122}=\theta_{123}$			
.769 (0.055)			

DISCUSSION

In this paper, we extended the classic Mann-Whitney-Wilcoxon (MWW) for multi-group comparison within a longitudinal data setting. We achieved this generalization by utilizing the functional response models (FRM), which is uniquely positioned to model rank-based outcomes as in the MWW rank sum test within our context. Inference is based on the U-statistics weighted generalized estimating equations. Which provides consistent and asymptotically normal estimates not only for complete data but also for missing data under MAR, the most popular missing mechanism in real studies [3,25,26].

We examined the performance of the proposed approach through both simulated and real study data. Results from the simulation study show that the proposed approach performed really well, with good parameter and type I estimates even for a sample as small as 50 per group. The proposed approach applies to both continuos and discrete outcomes. As demonstrated by the real study on sexual health, it handled ties well as the number of unprotected vaginal sex is an intrinsically discrete outcome.

In addition to the MWW test, median regression may also be used to address the outlier issue arising from the sexual health study [27,28]. However, these methods may not work well, since they either do not address missing data in longitudinal outcomes or require a unique median. Given that discrete outcomes typically do not have a unique median and MAR is popular in most real studies, applications of such methods in practice are very limited.

We performed all the simulation and real data analyses using a program we developed in Matlab. Readers interested in applying the methods can download this program from "CTSpedia.org", a popular reference and resource website as well as a repository of statistical and utility macros to facilitate and promote multidisciplinary interactions and collaborations involving biostatisticians.

The proposed approach has also limitations. For example, it cannot control for any covariate, which is particularly important for observational studies. Current work is underway to further extend the Mann-Whitney-Wilcoxon to a regression setting.

ACKNOWLEDGEMENT

This research was supported in part by grant R33 DA027521 from the National Institutes of Health, and by the University of Rochester CTSA award UL1TR000042 from the National Center for Advancing Translational Sciences of the National Institutes of Health.

REFERENCES

1. Morrison-Beedy D, Jones SH, Xia Y, Tu X, Crean HF, Carey MP. Reducing sexual risk behavior in adolescent girls: results from a randomized controlled trial. J Adolesc Health. 2013; 52: 314-321.
2. Wilcoxin F. Probability tables for individual comparisons by ranking methods. Biometrics. 1947; 3: 119-122.
3. Kowalski J, Tu, XM. Modern Applied U Statistics. Wiley: New York. 2007; 1-378.
4. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 1947; 18: 50-60.
5. Serfling RJ. Approximation Theorems of Mathematical Statistics. Wiley: New York. 1980.
6. Tu XM, Feng C, Kowalski J, Tang W, Wang H, Wan C, et al. Correlation analysis for longitudinal data: applications to HIV and psychosocial research. Stat Med. 2007; 26: 4116-4138.
7. Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. Biometrics. 2008; 64: 781-789.
8. Ma Y, Tang W, Tu XM. Modeling Concordance Correlation Coefficient for longitudinal study data. Psychometrika. 2010; 75: 99-119.
9. Ma Y, Tang W, Tu XM. Modeling Cronbach Coefficient Alpha for longitudinal study data. Statistics in Medicine. 2011; 29: 659-670.
10. Yu Q, Tang W, Kowalski J, Tu XM. Multivariate U-Statistics: A tutorial with applications. WIREs Computational Statistics. 2011; 3: 457-471.
11. Gunzler D, Tang W, Lu N, Wu P, Tu XM. A Class of Distribution-Free Models for Longitudinal Mediation Analysis. Psychometrika. 2013 .
12. Yu Q, Chen R, Tang W, He H, Gallop R, Crits-Christoph P, et al. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. Stat Med. 2013; 32: 2390-2405.
13. Lu N, White AM, Wu P, He H, Hu J, Feng C, Tu XM. Social network endogeneity and its implications for statistical and causal inferences. In Social Networking: Recent Trends, Emerging Issues and Future Outlook, Lu N, White AM, Tu XM, editors. Nova Science. New York. 2013.
14. Kowalski J, Powell J. Nonparametric inference for stochastic linear hypotheses: Application to high-dimensional data. Biometrika. 2004; 91: 393-408.
15. Wu P, Han Y, Chen T, Tu XM. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. Stat Med. 2013; .
16. Tang W, He H, Tu XM. Applied Categorical Data Analysis. Chapman & Hall/CRC. 2012.
17. MathWorks Inc. MatLab version 7.12.
18. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73: 13-22.
19. Randles RH, Wolfe DA. Introduction to the Theory of Nonparametric Statistics. Wiley: New York. 1979.

20. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990; 77: 485-497.

21. Boos DD, Brownie C. A rank-based mixed model approach to multisite clinical trials. Biometrics. 1992; 48: 61-72.

22. Guo X, Pan W, Connett JE, Hannan PJ, French SA. Small-sample performance of the robust score test and its modifications in generalized estimating equations. Stat Med. 2005; 24: 3479-3495.

23. Schroder KE, Carey MP, Vanable PA. Methodological challenges in research on sexual risk behavior: I. Item content, scaling, and data analytical options. Ann Behav Med. 2003; 26: 76-103.

24. Greenberg J, Hennessy M, MacGowan R, Celentano D, Gonzales V, Van Devanter N, et al. Modeling intervention efficacy for high-risk women. The WINGS Project. Eval Health Prof. 2000; 23: 123-148.

25. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd Edn. Wiley: New York. 1987.

26. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. JASA. 1995; 90: 106-121.

27. Yi GY, He W. Median regression models for longitudinal data with dropouts. Biometrics. 2009; 65: 618-625.

28. Yan M, Alejandro GD, Hui Z, Tu XM. A U-statistics-based approach for modeling Cronbach coefficient alpha within a longitudinal data setting. Stat Med. 2010; 29: 659-670.

Cite this article

Chen R, Wu P, Ma F, Han Y, Chen T, et al. (2014) Extending the Mann-Whitney-Wilcoxon Rank Sum Test for Multiple Treatment Groups and Longitudinal Study Data. Clin Res HIV/AIDS 1(1): 1005.