Statistical Modeling of Subject and Proxy Observations Using Weighted GEE

Mina Hosseini[1*], Nagaraj K. Neerchal[1] & Ann L. Gruber-Baldini[2]

[1]Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, 21250, U.S.A.

[2]Division of Gerontology, Department of Epidemiology & Public Health, University of Maryland School of Medicine, 655 W. Baltimore Street, Baltimore, MD, 21201, U.S.A.

Abstract

In epidemiological studies when the patients become unable to provide responses by themselves due to advancing severity of their conditions, proxy responses by a relative or a caregiver "proxy" are used. The resulting dataset contains a monotonically decreasing missing pattern for subject observations and a monotonically increasing missing pattern for proxy observations. Some statistical models are being investigated that can analyze subject and proxy observations together so that relevant parameters and their standard errors can be estimated in a single framework. The method of weighted Generalized Estimating Equations (GEE), which is commonly used for handling missing data, is applied to the combined proxy and subject dataset.

Key Words: Proxy; GEE; Weighted GEE; WGEE

1. Introduction

Monotone (decreasing) missing data patterns, which are also called dropout(s), occur in epidemiological datasets when the patients become unable to provide responses by themselves due to advancing severity of their conditions. It is common to use proxy responses by a relative or a caregiver in these cases. Proxy observations are substituted for missing subject observations so that the usual statistical methods, which require complete data, can be applied. However, the direct substitution of proxy observations may cause new problems such as biased parameter estimates and incorrect standard errors since as noted by Snow et al. (2005) the direct substitution is like assuming a perfect correlation between subject and proxy observations.

The main hurdle is to postulate a model to describe the joint probability distribution of subject and proxy data. We follow the approaches given in Shardell et al. (2010) and Huang et al. (2005), and propose a methodology of handling subject and proxy data in a single framewrok using the weighted generalized estimating equations (WGEE) method introduced by Robins et al. (1995) for longitudinal data. Performance of the various approaches is investigated via a simulation study.

This paper is structured as follows: In Section 2, missing data and the importance of proxy for older adults in gerontological studies is described. In addition, the structure of subject-proxy dataset that we are using in this paper is explained. In Section 3, a framework for modeling subject and proxy data is illustrated. Section 4 focuses on describing weighted GEE for subject and proxy data. In Section 5, the two WGEE methods for subject and proxy are implemented by using the simulation data.

*Email: mina.hoss.08@gmail.com.

2. Proxy Data in Aging Studies

The monotone dropout pattern of missing observations seems to be frequently encountered in aging studies. As it is described by Gruber-Baldini et al. (2012), sometimes due to physical, cognitive, or other psychological impairment, study participants are unable to provide responses for themselves. Once a subject becomes unable to provide responses, due to advancing age, the condition may never improve. Thus a dropout pattern of missing data is created while the subject is still in the study. Thus, external raters are used to gather information about the participants. The external rater, whose report is substituted for the subject, is called a proxy. A proxy is someone who knows the participant, and can provide information about them. Sometimes, proxy observation can support the information provided by the study participant. It is also possible that the proxy can give more useful information to help the clinicians in understanding the disease.

According to Snow et al. (2005), in some cases, patient and proxy observations are regarded perfectly correlated, and as a result, the two types of observations are analyzed together. The assumption of perfect correlation between proxy and subject observations may not hold in general since the patient and their corresponding proxy are not expected to give the same report due to different characteristics and questions. For example, in a study of older adults by Epstein et al. (1989), it was shown that there was a good agreement between patient and proxy on some factors; however, it was not true about all measures provided by proxies. In general, proxies may over- or under-report and cause biased estimates. As a part of the Baltimore hip fracture studies, Magaziner et al. (1997) compared proxy responses to the subject responses in 5 areas of functioning at one time point. They pointed out that the agreement between proxy and subject varies depending on different functions. They noted that where the patient's emotion was measured, the patient-proxy agreement was different since proxy may not necessarily be aware of the patient's emotion. However, it was noted by Gruber-Baldini et al. (2012) that there are, undoubtedly, some benefits in having proxy observations. Using proxy observations may result in the inclusion of those study subjects who may be illustrative of a selected group. The most important task is to determine how to include proxy observations in the statistical analysis of the study.

As discussed in the introduction, the data structure consists of subjects and proxies involved in a longitudinal study. We will assume that when a subject becomes unable to provide responses on their own, proxy observations are collected. In this stage of our initial work we will assume that either a response from the subject or a proxy is available for all time points. Table 1 schematically shows the data assuming that the longitudinal study consists of 4 time points.

Table 1: Subject-Proxy Data

Subject	Time			
ID	1	2	3	4
1	S	S	P	P
2	S	P	P	P
⋮	⋮	⋮	⋮	⋮
n	S	S	S	S

The above dataset can be split into 2 parts: One containing only subject responses and the other one containing only proxy responses. Missing values are created when the proxy observations are dropped to form the subject dataset and similarly missing values are created when the subject observations are dropped to form the proxy dataset. Tables 2 and 3 show the subject and proxy datasets. The two datasets exhibit missing patterns which are complementary.

Table 2: Subject Data

Subject	Time			
ID	1	2	3	4
1	S	S	*	*
2	S	*	*	*
⋮	⋮	⋮	⋮	⋮
n	S	S	S	S

Table 3: Proxy Data

Subject	Time			
ID	1	2	3	4
1	*	*	P	P
2	*	P	P	P
⋮	⋮	⋮	⋮	⋮
n	*	*	*	*

Note that the proxy dataset will not have any observations corresponding to subjects who had completed the study without having a need for proxy observations. Likewise, the subject dataset will not have any observations corresponding to subjects whose proxy observations were used starting the first time point. It is now clear from the data structure depicted above that the subject dataset has missing data with a pattern of missing known as dropout. The proxy dataset also has the same structure but in its mirror-image. In this paper, we analyze these datasets separately using the methods used for handling monotone missing (dropout) data. We will also provide a methodology which incorporates both subject and proxy data.

3. A Framework for Modeling Subject and Proxy Data

Here we briefly review a framework provided by Shardell et al. (2010) which considered subject and proxy data simultaneously. This framework will be used in formulating our approach as well as in generating data for our simulation studies. Let $Y_{i(s)}$ and $Y_{i(p)}$ denote the $T \times 1$ vectors of subject and proxy observations, respectively, for the i^{th} subject $(i = 1,2,\ldots,n)$. Also, let $R_{i(s)}$ and $R_{i(p)}$ be the $T \times 1$ indicator vectors such that $R_{it(s)} = 1$ when the i^{th} subject has provided a response (subject data) at time t, and $R_{it(s)} = 0$ when the subject observation is missing, and instead the proxy observation is available. Let $R_{it(p)} = 1 - R_{it(s)}$. Then $Y_{it(obs)}$ can be denoted as

$$Y_{it(obs)} = Y_{it(s)}R_{it(s)} + Y_{it(p)}(1 - R_{it(s)}) \tag{1}$$

or

$$Y_{it(obs)} = Y_{it(s)}R_{it(s)} + Y_{it(p)}R_{it(p)}.$$

For notational simplicity, the model will be stated assuming that there is only one time point per subject. Thus, the subscript "it" is replaced by "i" in equation (1). The notations easily generalize to the case of multiple time points per subject. The approach is first to specify a fairly general joint distribution for $Y_{i(obs)}$ given the value of $R_{i(s)}$ and then to consider a practically special case. To start, it is assumed that subject and proxy responses are jointly normally distributed. Then the following is assumed for $Y_i = (Y_{i(s)}, Y_{i(p)})'$

$$Y_i \mid R_{i(s)} = r, X_i \quad \sim \quad \mathrm{N}_2\left(\begin{pmatrix} \mathrm{X}_i\beta^{(r)} \\ \mathrm{X}_i\theta^{(r)} \end{pmatrix}, \begin{pmatrix} \sigma^{(r)}_{(ss)} & \sigma^{(r)}_{(sp)} \\ \sigma^{(r)}_{(sp)} & \sigma^{(r)}_{(pp)} \end{pmatrix}\right) \tag{2}$$

where X_i is a $1 \times q$ vector of covariates, $\mathrm{R}_{i(s)}$'s are such that

$$\mathrm{R}_{i(s)} \mid X_i \quad \sim \quad \mathrm{Ber}(\pi_s) \quad \text{independent}, \tag{3}$$

and θ and β are $q \times 1$ vectors of parameters. Then, equations (2) and (3) imply that

$$\begin{aligned} Y_{i(s)} \mid R_{i(s)} = r, X_i &\sim \mathrm{N}(X_i\beta^{(r)}, \sigma^{(r)}_{(ss)}) \\ Y_{i(p)} \mid R_{i(s)} = r, X_i &\sim \mathrm{N}(X_i\theta^{(r)}, \sigma^{(r)}_{(pp)}). \end{aligned} \tag{4}$$

The probability distribution of $Y_{i(s)} \mid X_i$ can then be written as

$$\begin{aligned} f(Y_{i(s)} \mid X_i) &= \sum_r \mathrm{P}(R_{i(s)} = r).f(Y_{i(s)} \mid R_{i(s)} = r, X_i) \\ &= \pi_{(s)}.f(Y_{i(s)} \mid R_{i(s)} = 1, X_i) + (1 - \pi_{(s)}).f(Y_{i(s)} \mid R_{i(s)} = 0, X_i). \end{aligned} \tag{5}$$

In Shardell et al. (2010), it was further assumed that $\sigma^{(1)}_{(ss)} = \sigma^{(0)}_{(ss)}$, $\sigma^{(1)}_{(sp)} = \sigma^{(0)}_{(sp)}$, and $\sigma^{(1)}_{(pp)} = \sigma^{(0)}_{(pp)}$. They discussed three different classes of models applicable to different scenarios. In this paper, we consider one of these classes, referred to by them as "class of subject-adjusted proxy pattern-mixture models". This model is stated later in section 5.

4. Weighted GEE for Subject and Proxy Data

The method of weighted generalized estimating equations (WGEE) also known as inverse probability weighting (IPW) was proposed by Robins et al. (1995). This method is based on assigning weights to the observed data in datasets with missing observations. According to Fitzmaurice et al. (2011), inverse probability weighting is mostly effective when the multivariate distribution of observations is not known and the likelihood based analysis is not possible. This method is already used to study subject responses.

WGEE is an implementation of the generalized estimating equations (GEE) with weights. The weights are inversely proportional to the probability that a subject response is available. The WGEE estimates are obtained by solving the following equations:

$$U(\beta, \alpha) = \sum_{i=1}^{n} D_i' V_i^{-1} \Delta_i \{Y_i - \mu_i(\beta)\} = 0, \tag{6}$$

where Δ_i is a $T \times T$ diagonal matrix of weights as follows:

$$\Delta_i = \begin{bmatrix} \pi_{i1}(\alpha)^{-1} R_{i1} & 0 & \dots & \dots & 0 \\ 0 & \pi_{i2}(\alpha)^{-1} R_{i2} & \dots & \dots & 0 \\ \vdots & \vdots & \pi_{i3}(\alpha)^{-1} R_{i3} & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \pi_{iT}(\alpha)^{-1} R_{iT} \end{bmatrix}. \tag{7}$$

In equation (7), R_{it} is the t^{th} element $(t = 1, 2, \dots, T)$ of R_i, the $T \times 1$ indicator vector representing the missing pattern for the i^{th} subject $(i = 1, 2, \dots, n)$, and π_{it} is the probability of subject i being observed at time t where

$$\pi_{it} = \mathrm{P}(R_{it} = 1 \mid X_i, Y_i), \tag{8}$$

and under the missing at random (MAR) assumption (See Rubin (1976) and Fitzmaurice et al. (2011)),

$$\pi_{it} = \mathrm{P}(R_{it} = 1 \mid X_i, Y_{i1}, \dots, Y_{it-1}) = \lambda_{it} \,.\, \lambda_{it-1} \,.\, \lambda_{it-2} \,.\, \dots \,.\, \lambda_{i2} \,.\, 1, \tag{9}$$

where

$$\lambda_{it} = \mathrm{P}(R_{it} = 1 \mid R_{it-1} = 1, X_i, Y_{i1}, \ldots, Y_{it-1}). \tag{10}$$

According to Fitzmaurice et al. (2011), a correct model to estimate π_{it} plays a very important role in finding a valid estimate for β. To find this correct model, a logit model is used where the response variable is R_{it}.

$$\mathrm{Logit}(\lambda_{it}) = Z'_{it}\alpha, \tag{11}$$

where Z_{it} is a $q \times 1$ vector of past observations and some of (or all) covariates. The estimated weights are given by

$$\hat{w}_{it} = \hat{\pi}_{it}^{-1}, \tag{12}$$

where

$$\hat{\pi}_{it} = \lambda_{it}(\hat{\alpha}) \,.\, \ldots \,.\, \lambda_{i2}(\hat{\alpha}).1.$$

Based on the theorem by Robins et al. (1995), $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean 0, and the positive definite estimator of variance as

$$\hat{\Gamma}^{-1}\widetilde{C}\hat{\Gamma}^{-1\prime}, \tag{13}$$

where β_0 is the actual solution to the equation in (6) and $\hat{\Gamma}$, $\widetilde{C}$, and A_i are given by

$$\hat{\Gamma} = n^{-\frac{1}{2}}.\frac{\partial U(\hat{\beta},\hat{\alpha})}{\partial \beta'} = n^{-1}.\sum_{i=1}^{n}(\frac{\partial \mu_i(\hat{\beta})}{\partial \beta'})'.V_i^{-1}.\Delta_i(\hat{\alpha}).(\frac{\partial \mu_i(\hat{\beta})}{\partial \beta'}) \tag{14}$$

$$\widetilde{C} = n^{-1}\sum_i A_i A_i' \tag{15}$$

$$A_i = \left[U_i(\hat{\beta},\hat{\alpha}) - \left(\sum_i U_i(\hat{\beta},\hat{\alpha})S_i'(\hat{\alpha})\right)\left(\sum_i S_i(\hat{\alpha})S_i'(\hat{\alpha})\right)^{-1} S_i(\hat{\alpha})\right], \tag{16}$$

and also

$$S(\alpha) = \sum_i S_i(\alpha) = \sum_i \sum_t R_{it-1}.Z_{it}.(R_{it} - \lambda_{it}) \tag{17}$$

for $i = 1, 2 \ldots, n$ and $t = 2, 3 \ldots, T$.

Since

$$\mathrm{L}(\alpha) = \prod_i \mathrm{L}_i(\alpha) = \prod_i \prod_t [\lambda_{it}(\alpha)^{R_{it}}.(1 - \lambda_{it}(\alpha))^{1-R_{it}})]^{R_{it-1}}, \tag{18}$$

$S_i(\alpha)$ in equation (17) is obtained as follows:

$$S_i(\alpha) = \left\{\frac{\partial \mathrm{Log}(\mathrm{L}_i(\alpha))}{\partial \alpha}\right\}. \tag{19}$$

When α is fixed, $\left(\sum_i U_i(\hat{\beta},\hat{\alpha})S_i'(\hat{\alpha})\right)\left(\sum_i S_i(\hat{\alpha})S_i'(\hat{\alpha})\right)^{-1} S_i(\hat{\alpha})$ is equal to zero since $S_i(\alpha) = 0$. In this case the covariance is the simple sandwich estimator of covariance. However, in reality, α is estimated and covariance of $\hat{\beta}$ is adjusted for the estimated α.

There are two approaches to implement the WGEE method described above. The first approach is to explicitly code the two steps involved (a 2-step code). The first step is to use a PROC GENMOD run to obtain predicted values of the λ_{it}'s. The predicted values are then used

to obtain weights, w_{it}'s. The second step is another PROC GENMOD run with a WEIGHT statement to estimate β and its covariance, given in equations (6) and (13) respectively. In the second approach, PROC GEE in SAS can be used, where the mentioned two steps are integrated. See Lin & Rodriguez (2014) for more details. While this approach is convenient, PROC GEE does not provide access to all intermediate results and datasets created internally. In particular, the procedure does not provide access to the values of the weights used in producing the WGEE. We coded the two steps explicitly and confirmed that the two approaches provide identical estimates for the same datasets. Furthermore, our code can also be conveniently translated to other packages such as R (or STATA) because the individual steps rely only on widely available generalized model estimation methods.

As noted in Table 1, subject dataset already has a dropout missing pattern. Thus, implementation of WGEE to the subject dataset is straight forward. On the other hand, as shown in Figure 1, the data structure of the proxy dataset is the mirror-image of the subject dataset. Therefore, a suitable restructuring of this dataset is necessary before obtaining the weights given in equation (12). The pattern of missingness for the proxy dataset is monotone increasing. In addition, one of the assumptions in applying WGEE to subject observations is the availability of observations at the first time point. This assumption is not satisfied when WGEE for proxy observations is of interest.

	$t=1$	$t=2$	$t=3$	$t=4$
$i=1$	$Y_{11(p)}$	$Y_{12(p)}$	$Y_{13(p)}$	$Y_{14(p)}$
$i=2$	$Y_{21(s)}$	$Y_{22(p)}$	$Y_{23(p)}$	$Y_{24(p)}$
$i=3$	$Y_{31(s)}$	$Y_{32(s)}$	$Y_{33(p)}$	$Y_{34(p)}$
$i=4$	$Y_{41(s)}$	$Y_{42(s)}$	$Y_{43(s)}$	$Y_{44(p)}$
$i=5$	$Y_{51(s)}$	$Y_{52(s)}$	$Y_{53(s)}$	$Y_{54(s)}$

	$t'=1$ $t=4$	$t'=2$ $t=3$	$t'=3$ $t=2$	$t'=4$ $t=1$
$i=1$	$Y_{14(p)}$	$Y_{13(p)}$	$Y_{12(p)}$	$Y_{11(p)}$
$i=2$	$Y_{24(p)}$	$Y_{23(p)}$	$Y_{22(p)}$	$Y_{21(s)}$
$i=3$	$Y_{34(p)}$	$Y_{33(p)}$	$Y_{32(s)}$	$Y_{31(s)}$
$i=4$	$Y_{44(p)}$	$Y_{43(s)}$	$Y_{42(s)}$	$Y_{41(s)}$
$i=5$	$Y_{54(s)}$	$Y_{53(s)}$	$Y_{52(s)}$	$Y_{51(s)}$

Figure 1: Subject-Proxy Data

In order to change the non-dropout pattern of the proxy portion of the data, the order of the observations is reversed so that WGEE method can be applied. This idea is illustarted in Figure 1: the table on the right shows a mirror-image of the data presented on the left. According to Robins et al. (1995), WGEE is applied under the assumption of MAR. Under MAR, the probability of being observed at time t depends on all available observed responses and the covariates. Thus, MAR assumption is also applicable to the proxy portion of the data, and the WGEE method can be used to analyze flipped proxy observations.

To implement the WGEE for proxy data, we use the 2-step code. In the first step of the code, the probability that subject i is not observed at time $t' = T - t + 1$ $(t' = 1, 2, \ldots, T)$ is estimated using a logit model with $R_{it'(p)} = 1 - R_{it'(s)}$ as a response and the observed proxy responses at the previous time points (which are future responses in the original dataset) as covariates. Then the estimated probabilities are used to calculate weights. Finally, weights are incorporated into equation (20) in the second step of the code, similarly to what is done in WGEE for subject data.

$\theta^{(0)}$ in equation (20) is the parameter when proxy observations are available,

$$g(\mathrm{E}(Y_{(p)} \mid X)) = X\theta^{(0)}. \tag{20}$$

5. A WGEE Small Simulation Study

To compare parameter estimates obtained using WGEE for subject and the proposed WGEE for proxy datasets, 1000 studies per parameter combination were simulated. In each study, two sets of longitudinal data were generated based on the "class of subject-adjusted proxy pattern-mixture model" by Shardell et al. (2010). The model is stated here for convenient reference.

In this model, $\beta^{(r)} = (\theta^{(r)} - \psi)/\gamma$ for $r \in \{0,1\}$, or in other words $\theta^{(r)} = \gamma\beta^{(r)} + \psi$. Using this equality along with the variance covariance assumptions, $\sigma^{(0)}_{(ss)} = \sigma^{(1)}_{(ss)}$, $\sigma^{(1)}_{(pp)} = \sigma^{(0)}_{(pp)}$, and $\sigma^{(1)}_{(sp)} = \sigma^{(0)}_{(sp)}$, the bivariate distribution of $Y_{i(s)}$ and $Y_{i(p)}$ conditional on $R_{i(s)}$ and X_i is as follows:

$$\begin{aligned}
((\mathrm{Y}_{i(s)}, \mathrm{Y}_{i(p)})' \mid \mathrm{R}_{i(s)} = 1, \mathrm{X}_i) &\sim \mathrm{N}_2\left(\begin{pmatrix} \mathrm{X}_i\beta^{(1)} \\ \mathrm{X}_i(\gamma\beta^{(1)} + \psi) \end{pmatrix}, \begin{pmatrix} \sigma^{(1)}_{(ss)} & \sigma^{(0)}_{(sp)} \\ \sigma^{(0)}_{(sp)} & \sigma^{(0)}_{(pp)} \end{pmatrix}\right) \\
((\mathrm{Y}_{i(s)}, \mathrm{Y}_{i(p)})' \mid \mathrm{R}_{i(s)} = 0, \mathrm{X}_i) &\sim \mathrm{N}_2\left(\begin{pmatrix} \mathrm{X}_i\beta^{(0)} \\ \mathrm{X}_i(\gamma\beta^{(0)} + \psi) \end{pmatrix}, \begin{pmatrix} \sigma^{(1)}_{(ss)} & \sigma^{(0)}_{(sp)} \\ \sigma^{(0)}_{(sp)} & \sigma^{(0)}_{(pp)} \end{pmatrix}\right).
\end{aligned}$$

In this model, we have that $\sigma^{(0)}_{(ss)} = \sigma^{(1)}_{(ss)}$, and $\sigma^{(1)}_{(sp)} = \sigma^{(0)}_{(sp)}$, and that the mean of $Y_{i(p)}$ conditional on $Y_{i(s)}$ and X_i is independent from $R_{i(s)}$. Therefore, it follows that, $\sigma^{(1)}_{(sp)}/\sigma^{(1)}_{(ss)} = \sigma^{(0)}_{(sp)}/\sigma^{(0)}_{(ss)} = \gamma$. As a result

$$\begin{aligned}
\mathrm{E}(Y_{i(p)} \mid Y_{i(s)}, R_{i(s)} = 0, X_i) &= \mathrm{E}(Y_{i(p)} \mid Y_{i(s)}, R_{i(s)} = 1, X_i) \\
X\theta^{(0)} + \gamma\left(Y_{i(s)} - X\beta^{(0)}\right) &= X\theta^{(1)} + \gamma\left(Y_{i(s)} - X\beta^{(1)}\right) \\
\theta^{(0)} - \gamma\beta^{(0)} &= \theta^{(1)} - \gamma\beta^{(1)},
\end{aligned}$$

where it is assumed that

$$\psi = \theta^{(0)} - \gamma\beta^{(0)} = \theta^{(1)} - \gamma\beta^{(1)}. \tag{21}$$

Two drug dose levels and 4 time points were considered. The number of subjects were chosen to be 100 in each of the 2 drug dose levels. Conditional on $R_{i(s)} = 1$ and X_i for all i's ($i = 1, 2, \ldots, 200 \times 4$), $(Y_{i(s)}, Y_{i(p)})'$ was generated from a bivariate normal distribution with mean $(X_i\beta^{(1)}, X_i\theta^{(1)})'$. Two different $\beta^{(1)}$'s were considered, $(-1.5, 0.1061, 0.54)'$ and $(1.5, 0.9, -0.54)'$. ψ was assumed to be $(0.21, 0.21, 0.21)'$. Also, two different covariance matrices were considered. The elements of the covariance matrix in one of these matrices were assumed to be $\sigma^{(1)}_{(ss)} = 1$, $\sigma^{(1)}_{(pp)} = 1.5$, and $\sigma^{(0)}_{(sp)} = 0.85$, and the elements of the second one were assumed to be $\sigma^{(1)}_{(ss)} = 0.2$, $\sigma^{(1)}_{(pp)} = 0.9$, and $\sigma^{(0)}_{(sp)} = 0.21$. Similarly, conditional on $R_{i(s)} = 0$ and X_i, the second set of $(Y_{i(s)}, Y_{i(p)})'$ was generated from a bivariate normal distribution with mean $(X_i\beta^{(0)}, X_i\theta^{(0)})'$, where $\beta^{(0)}$ and $\theta^{(0)}$ were chosen to be equal to $\beta^{(1)}$ and $\theta^{(1)}$ respectively, and the covariance matrix the same as what was chosen for the first set of the points. In generating both sets the correlation between subject and proxy was assumed to be 0.7 and 0.5 when using the first and the second covariance matrices respectively which were effective in choosing the value of $\sigma^{(0)}_{(sp)}$.

The dropout pattern $R_{(s)}$ was generated separately to specify if $Y_{i(obs)}$ $(i = 1,2,\ldots,n)$, at each time point, is a subject observation or a proxy observation. The data generating process used for the dropout pattern is consistent with MAR assumption. If $R_{i(s)} = 1$ then $Y_{i(s)}$ from the first simulated set was picked, but if $R_{i(s)} = 0$, $Y_{i(p)}$ from the second simulated set was picked. $Y_{i(s)}$'s and $Y_{i(p)}$'s were saved into two different datasets as subject and proxy datasets.

Using simulation data, WGEE for subject and proxy data was implemented. At first, the weights were estimated for both subject and proxy observations. The estimated weights, in separate analyses of subject and proxy data, were incorporated into the models to estimate the subject and proxy parameters. Tables 4 and 6 show the results for different initial $\theta^{(0)}$'s and $\beta^{(1)}$'s. On the top portion of both tables the results of analyzing subject observations is shown where they were analyzed separately from the proxy observations, using the WGEE method. The mean bias was calculated by subtracting the true subject parameter (when subject observations were available), $\beta^{(1)}$, from the mean estimate of subject parameter. The middle parts of Tables 4 and 6 similarly show the results of analyzing proxy observations in separate analyses where the mean bias was calculated by subtracting the true proxy parameter (when subject observations are not available), $\theta^{(0)}$, from the mean estimate of proxy parameter. The bottom parts of Tables 4 and 6 show the results of analyzing data when proxy observations were substituted for missing subject observations and the GEE method was used while the missing mechanism was assumed to be MCAR. The mean bias was calculated by subtracting the true subject parameter (when proxy observations were substituted for missing subject observations), $\beta^{(1)}$, from the mean estimate of subject/proxy parameter.

Table 4: Comparing Results from Subject and Proxy Data Separately to Current GEE Approach

$\beta^{(1)}_{\text{True}} = (0.1061\ ,\ 0.54)'$, $\theta^{(0)}_{\text{True}} = (0.1061\gamma + 0.21\ ,\ 0.54\gamma + 0.21)'$								
Proxy(%)	$(\sigma_{ss}, \sigma_{sp} = \sigma_{ps}, \sigma_{pp})$	ρ_{sp}	γ	Parm	Mean Est.	Mean Bias	MCSE	Mean SE
Subject Parameters (Only Subject Data, WGEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.1052	-0.0009	0.0762	0.0747
				Time	0.5413	0.0013	0.0341	0.0336
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.1057	-0.0004	0.0340	0.0332
				Time	0.5406	0.0006	0.0152	0.0150
Proxy Parameters (Only Proxy Data, WGEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.3094	0.0093	0.3436	0.3043
				Time	0.6581	-0.0109	0.2154	0.1931
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.3298	0.0084	0.2736	0.2416
				Time	0.7687	-0.0084	0.1710	0.1537
Subject/Proxy Parameters (Combined Subject-Proxy Data, GEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.1301	0.0240	0.0826	0.0803
				Time	0.6167	0.0767	0.0333	0.0336
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.1316	0.0255	0.0496	0.0483
				Time	0.6221	0.0821	0.0195	0.0198

Directly substituting proxy observations for the missing observations and using the GEE method results in biased estimates. Instead of applying the GEE method to the combined data directly, we defined indicator variables that control the use of covariates for subject and proxy depending on the response that is from either of subject or proxy. Inclusion of dummy variables allows us to obtain separate set of parameter estimates for subject and proxy observations. Thus, our approach essentially amounts to including interaction between the dose and time effects and the factor of whether an observation is a subject response or a proxy response. In addition, the WGEE method was used to analyze data. The estimated weights were the ones from separate

Table 5: Comparing Results from Combined Subject and Proxy Data to Current GEE Approach

$\beta^{(1)}_{\text{True}} = (0.1061\ ,\ 0.54)'$, $\theta^{(0)}_{\text{True}} = (0.1061\gamma+0.21\ ,\ 0.54\gamma+0.21)'$								
Proxy(%)	$(\sigma_{ss}, \sigma_{sp}=\sigma_{ps}, \sigma_{pp})$	ρ_{sp}	γ	Parm	Mean Est.	Mean Bias	MCSE	Mean SE
Subject Parameters (Combined Subject-Proxy Data, WGEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.1054	-0.0007	0.0762	0.0747
				Time	0.5413	0.0013	0.0341	0.0337
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.1058	-0.0003	0.0341	0.0333
				Time	0.5406	0.0006	0.0152	0.0150
Proxy Parameters (Combined Subject-Proxy Data, WGEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.3123	0.0121	0.3397	0.3049
				Time	0.6565	-0.0125	0.2137	0.1950
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.3303	0.0089	0.2715	0.2423
				Time	0.7676	-0.0094	0.1695	0.1551
Subject/Proxy Parameters (Combined Subject-Proxy Data, GEE Analysis)								
10.40	(1, 0.85, 1.5)	0.7	0.85	Dose	0.1301	0.0240	0.0825	0.0803
				Time	0.6167	0.0767	0.0333	0.0336
9.81	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.1316	0.0255	0.0496	0.0483
				Time	0.6221	0.0821	0.0195	0.0198

analyses of subject and proxy data. Using this method, $\beta^{(1)}$'s and $\theta^{(0)}$'s were estimated. The top and the middle parts of Tables 5 and 7 show the results. For the convenience, the results of simple GEE for the combined data is presented in the bottom of Tables 5 and 7 one more time. The results in Tables 5 and 7 show that the mean bias for both subject and proxy parameters is very small compared to the bias of parameters from GEE analysis. It is also important that all available data is used in WGEE analysis for combined subject and proxy observations. The standard errors from WGEE analysis; however, are large for proxy parameters.

Table 6: Comparing Results from Subject and Proxy Data Separately to Current GEE Approach

$\beta^{(1)}_{\text{True}} = (0.9\ ,\ -0.54)'$, $\theta^{(0)}_{\text{True}} = (0.9\gamma+0.21\ ,\ -0.54\gamma+0.21)'$								
Proxy(%)	$(\sigma_{ss}, \sigma_{sp}=\sigma_{ps}, \sigma_{pp})$	ρ_{sp}	γ	Parm	Mean Est.	Mean Bias	MCSE	Mean SE
Subject Parameters (Only Subject Data, WGEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.8984	-0.0016	0.0819	0.0797
				Time	-0.5388	0.0012	0.0363	0.0356
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.8992	-0.0009	0.0358	0.0353
				Time	-0.5394	0.0006	0.0161	0.0157
Proxy Parameters (Only Proxy Data, WGEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.9806	0.0056	0.2284	0.2180
				Time	-0.2546	-0.0056	0.1416	0.1347
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	1.1649	0.0099	0.1787	0.1718
				Time	-0.3615	-0.0045	0.1114	0.1069
Subject/Proxy Parameters (Combined Subject-Proxy Data, GEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.9702	0.0703	0.0890	0.0865
				Time	-0.4032	0.1369	0.0344	0.0346
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	1.0065	0.1065	0.0570	0.0580
				Time	-0.4126	0.1274	0.0216	0.0217

Table 7: Comparing Results from Combined Subject and Proxy Data to Current GEE Approach

$\beta^{(1)}_{\text{True}} = (0.9\ ,\ -0.54)'$, $\theta^{(0)}_{\text{True}} = (0.9\gamma+0.21\ ,\ -0.54\gamma+0.21)'$								
Proxy(%)	$(\sigma_{ss}, \sigma_{sp} = \sigma_{ps}, \sigma_{pp})$	ρ_{sp}	γ	Parm	Mean Est.	Mean Bias	MCSE	Mean SE
Subject Parameters (Combined Subject-Proxy Data, WGEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.8984	-0.0016	0.0819	0.0797
				Time	-0.5388	0.0012	0.0362	0.0356
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	0.8992	-0.0008	0.0360	0.0353
				Time	-0.5395	0.0005	0.0161	0.0158
Proxy Parameters (Combined Subject-Proxy Data, WGEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.9802	0.0052	0.2288	0.2181
				Time	-0.2542	-0.0052	0.1410	0.1350
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	1.1646	0.0096	0.1789	0.1718
				Time	-0.3611	-0.0041	0.1115	0.1071
Subject/Proxy Parameters (Combined Subject-Proxy Data, GEE Analysis)								
19.11	(1, 0.85, 1.5)	0.7	0.85	Dose	0.9702	0.0703	0.0890	0.0865
				Time	-0.4032	0.1369	0.0344	0.0346
18.29	(0.2, 0.21, 0.9)	0.5	1.05	Dose	1.0065	0.1065	0.0570	0.0580
				Time	-0.4126	0.1274	0.0216	0.0217

6. Discussion

The focus of this paper was to find a way to analyze subject and proxy observations together in gerontological studies. We considered the situation when the subject drops out of the study and only proxy observations become available subsequently. Current approach is to use the generalized estimating equations (GEE) method to analyze data where proxy observations are used in the place of missing subject responses. Gerontology literature makes note of possible biases resulting from using proxy observation in the place of missing subject response. Such substitutions completes the data structure needed to apply standard statistical methodology but the model parameter estimates and standard errors may be incorrect. In this paper we proposed a way to analyze subject and proxy observations together so that the relevant parameters and their standard errors can be estimated in a single framework. A simulation study was condcuted to study the properties of the proposed approach and compare its properties to the current GEE approach.

References

Epstein, A. M., Hall, J. A., Tognetti, J., Son, L. H., & Conant, L. (1989). Using proxies to evaluate quality of life: Can they provide valid information about patients' health status and satisfaction with medical care. *Medical Care*, *27*(3), S91–S98. Retrieved from http://www.jstor.org.proxy-bc.researchport.umd.edu/stable/3765656

Fitzmaurice, G., Laird, N., & Ware, J. (2011). *Applied Longitudinal Analysis*. Hoboken, New Jersey: John Wiley and Sons.

Gruber-Baldini, A. L., Shardell, M., Lloyd, K. D., & Magaziner, J. (2012). Use of proxies and informants. In A. B. Newman & J. A. Cauley (Eds.), *The epidemiology of aging* (pp. 81–90). Springer. doi: 10.1007/978-94-007-5061-6

Huang, R., Liang, Y., & Carriere, K. C. (2005). The role of proxy information in missing data analysis. *Statistical Methods in Medical Research*, *14*(5), 457–471. doi: 10.1191/0962280205sm411oa

Lin, G., & Rodriguez, R. N. (2014). Weighted methods for analyzing missing data with the GEE procedure. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS166-2014.pdf

Magaziner, J., Zimmerman, S. I., Gruber-Baldini, A. L., Hebel, J. R., & Fox, K. M. (1997). Proxy reporting in five areas of functional status. comparison with self-reports and observations of performance. *American Journal of Epidemiology*, *146*(5), 418-428. Retrieved from http://aje.oxfordjournals.org/content/146/5/418.abstract

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *American Statistical Association*, *90*(429), 106-121. doi: 10.2307/2291134

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581

Shardell, M., Hicks, G. E., Miller, R. R., Langenberg, P., & Magaziner, J. (2010). Pattern-mixture models for analyzing normal outcome data with proxy respondents. *Statistics in Medicine*, *29*(14), 1522–1538. doi: 10.1002/sim.3902

Snow, A. L., Cook, K. F., Lin, P., Morgan, R. O., & Magaziner, J. (2005). Proxies and other external raters: Methodological considerations. *Health Services Research*, *40*, 1676–1693. doi: 10.1111/j.1475-6773.2005.00447.x