

Mean corrected generalized estimating equations for longitudinal binary outcomes with report bias

Statistical Methods in Medical Research
2022, Vol. 31(2) 315–333
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802211065160
journals.sagepub.com/home/smm



Chao Li¹, Ye Shen¹ , Qian Xiao², Stephen L Rathbun¹,
Hui Huang³, and Yongtao Guan⁴

Abstract

Cocaine addiction is an important public health problem worldwide. Cognitive-behavioral therapy is a counseling intervention for supporting cocaine-dependent individuals through recovery and relapse prevention. It may reduce patients' cocaine uses by improving their motivations and enabling them to recognize risky situations. To study the effect of cognitive behavioral therapy on cocaine dependence, the self-reported cocaine use with urine test data were collected at the Primary Care Center of Yale-New Haven Hospital. Its outcomes are binary, including both the daily self-reported drug uses and weekly urine test results. To date, the generalized estimating equations are widely used to analyze binary data with repeated measures. However, due to the existence of significant self-report bias in the self-reported cocaine use with urine test data, a direct application of the generalized estimating equations approach may not be valid. In this paper, we proposed a novel mean corrected generalized estimating equations approach for analyzing longitudinal binary outcomes subject to reporting bias. The mean corrected generalized estimating equations can provide consistently and asymptotically normally distributed estimators under true contamination probabilities. In the self-reported cocaine use with urine test study, accurate weekly urine test results are used to detect contamination. The superior performances of the proposed method are illustrated by both simulation studies and real data analysis.

Keywords

Cocaine use, generalized estimating equation, drug addiction, self-reported data, survey study, bias correction

1 Introduction

Cocaine use is an important public health problem in the United States and throughout the world. It is associated with many medical consequences and psychosocial characteristics, including increased risks of myocardial infarctions, stroke, infectious diseases, chronic stress, and violence.^{1,2} Although no medications are currently available to treat cocaine addiction effectively, one promising substitute for cocaine is buprenorphine, a partial mu-opioid agonist at the mu-opioid receptor and kappa-opioid antagonist. Its efficacy has been noted in some pharmacotherapy trials.^{3,4,5} Cognitive-behavioral therapy (CBT) is a counseling intervention for drug or alcohol use disorders, which includes learning skills and strategies of affect regulations, changing maladaptive thoughts, and learning new behavioral strategies.^{6,7,8,9}

Our motivating data, the self-reported cocaine use with urine test (SCU) data, is based on a study of the CBT effect on cocaine dependence at the Primary Care Center of Yale-New Haven Hospital. To evaluate the impact of adding the CBT to

¹Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA

²Department of Statistics, Franklin College of Arts and Sciences, University of Georgia, Athens, GA, USA

³School of Mathematics, Sun Yat-Sen University, GuangZhou, China

⁴Department of Management Science, Business School, University of Miami, Coral Gables, FL, USA

Corresponding author:

Ye Shen, Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, USA.
Email: yeshen@uga.com

the physical management (PM) on patients receiving buprenorphine, the subjects were randomly assigned to the treatment and control groups. The control group receives the PM, a 15–20 min session by internal medicine physicians with experience administering buprenorphine. The treatment group receives both the PM and the CBT, which are provided by trained clinicians. Collected outcomes include daily self-reported drug uses and weekly urine test results. A recent publication analyzed the SCU data via a logistic regression model using only the daily self-reported data and found that the CBT effect was not significant.¹⁰ However, their analysis did not consider the report biases of the outcomes which were found to be significant.

The reporting bias exists when a respondent's answer to a survey question differs from the truth.¹¹ It is a major problem in the assessment of the self-reported data, which are commonly collected in clinical researches. When the report bias exists, the conclusion drawn from the self-reported data may not be valid. Neuhaus¹² examined the effect of the reporting bias on the inference of clustered longitudinal binary responses, where consistent parameter estimates have been developed. Chen et al.¹³ proposed the marginal analysis method to analyze longitudinal binary outcomes with report bias, which could yield predominately robust and efficient results.

The generalized estimating equations (GEE) approach¹⁴ has been widely used for analyzing longitudinal data with repeated measurements in drug addiction studies. The GEE extends the generalized linear model to a hierarchical setting with dependent outcomes via specifying a working covariance matrix.¹⁵ Parameter estimates in the GEE are consistent as long as the marginal means of the outcomes are correctly specified, and such consistency is retained even when the covariance structure is misspecified.¹⁴ As the reporting bias may lead to wrong marginal means, estimations from the GEE can be significantly biased when analyzing the self-reported data. In the SCU data, there are roughly 21.4% mismatches when comparing the self-reported cocaine uses and the urine test results, which suggests significant report biases. Throughout this paper, we refer to such mismatches due to the reporting bias as the contamination. Clearly, results from the GEE method analyzing the SCU data are not reliable.

The urine test results are often used as surrogate markers due to their accuracy and reliability. Yet, the collection of such biological tests is more expensive and less convenient compared to self-reports.¹⁶ In practice, urine test results are often collected less frequently than the self-reported data. With the limitations of self-reported data being inaccurate and urine test results being infrequent, it has been suggested to either combine these two measures as joint outcomes or use biological test results to correct the bias in the self-reports to increase the validity.^{17,18,19} In this paper, we propose a novel mean corrected GEE (MCGEE) approach to analyze self-reported longitudinal binary outcomes, which uses the urine test results to correct the marginal means of the self-reports. The MCGEE will lead to consistent and asymptotically normally distributed parameter estimates under true contamination probabilities. In practice, the contamination may come from unintentional mistakes (e.g. forgetfulness or pure recording errors) and intentional mistakes (e.g. hiding true drug use on purpose) and it is subject-specific. We propose a reliable method to estimate such contamination probabilities, which can lead to satisfactory performances in both simulation studies and real data analysis.

The remainder of this paper is organized as follows. Section 2 presents the proposed MCGEE approach and Section 3 illustrates the asymptotic properties of the MCGEE estimators. Section 4 shows the performances of the MCGEE approach on finite samples via simulation studies. Section 5 discusses the analysis of the SCU data using the proposed method. Section 6 concludes and discusses some future works. All proofs and technical details are relegated to the Appendix.

2 Mean corrected generalized estimating equations

Let N and T be the total numbers of subjects and time periods in the study, respectively. Let Y_{it} be the true drug use and X_{it} be the covariate vector for the subject $i \in \{1, \dots, N\}$ at the time $t \in \{1, \dots, T\}$. The binary response $Y_{it} = 1$ if the subject uses the drug; otherwise, $Y_{it} = 0$. The marginal probability mass function of Y_{it} is assumed to be Bernoulli. Denote the marginal mean of Y_{it} as μ_{it} . Let $Y_i = (Y_{i1}, \dots, Y_{iT})'$, $X_i = (X_{i1}, \dots, X_{iT})'$, and $\mu_i = E(Y_i|X_i; \beta) = (\mu_{i1}, \dots, \mu_{iT})'$, where β is the vector of regression parameters. The GEE approach by Liang and Zeger¹⁴ with the logit link has the form:

$$U_\beta(\beta) = \sum_{i=1}^N D_i' V_i^{-1} (Y_i - \mu_i) = 0 \quad (1)$$

where $D_i = \partial \mu_i / \partial \beta = A_i X_i$ and V_i is the working covariance matrix of Y_i . Matrix V_i can be decomposed into the form of $A_i^{1/2} C_i(\gamma) A_i^{1/2}$, where $A_i = \text{diag}(\text{var}(Y_{i1}), \dots, \text{var}(Y_{iT}))$ with $\text{var}(Y_{it}) = \mu_{it} \times (1 - \mu_{it}) = e^{x_{it}\beta} / (1 + e^{x_{it}\beta})^2$ and $C_i(\gamma)$ is the working correlation matrix of Y_i with parameter γ . Some common correlation structures for longitudinal data analysis include: (1) an exchangeable structure where the correlation between two different time points of a subject, that is Y_{id} and Y_{it} ($d \neq t$), is γ , (2) an autoregressive structure where the correlation between Y_{id} and Y_{it} ($d \neq t$) is $\gamma^{|d-t|}$, and (3) an unstructured structure where the correlation between Y_{id} and Y_{it} ($d \neq t$) is γ_{dt} .^{20,21}

In the SCU data, the true drug use Y_{it} cannot be observed directly, and the self-reported outcomes are used to replace Y_{it} . The discrepancy between the two may cause a significant bias in β estimators. As the GEE only requires correct marginal means to achieve consistent estimators of β without the need of specifying correct working correlation structures, we propose the MCGEE approach by correcting the marginal means of self-reported data based on the urine test results.

Let R_{it} be an indicator variable representing the outcome contamination for the subject $i \in \{1, \dots, N\}$ at the time $t \in \{1, \dots, T\}$. If the self-reported drug use Z_{it} is different from the true drug use Y_{it} , we set $R_{it} = 1$; otherwise, $R_{it} = 0$. We have

$$Z_{it} = Y_{it}(1 - R_{it}) + (1 - Y_{it})R_{it} \quad (2)$$

Define μ_{it}^* as the expected value of Z_{it} , based on equation (2), we have

$$\begin{aligned} \mu_{it}^* &= E(Z_{it}|X_{it}; \beta) \\ &= E(Y_{it}|X_{it}; \beta) - 2E(Y_{it}R_{it}|X_{it}; \beta) + E(R_{it}|X_{it}; \beta) \end{aligned} \quad (3)$$

Generally speaking, the contamination in the self-reported data may come from unintentional mistakes (e.g. forgetfulness or pure recording errors) and intentional mistakes (e.g. hiding true drug use on purpose). In drug dependence studies, since some subjects who actually use the drug may not want others to know about the facts, they may intentionally hide their drug uses. Thus, the contamination probabilities for the cases of $Y = 1$ include both unintentional and intentional mistakes, and they tend to be higher than those for the cases of $Y = 0$, which only include unintentional mistakes. In this paper, we model this scenario as follows. Let the i th subject have a contamination probability $p_{it}^{(0)}$ for unintentional mistakes at the time t , and a probability $p_{it}^{(1)}$ for intentionally hiding the true drug use at the time t . That is, for any time $t = 1, \dots, T$ and given X_{it} and β , we assume $E(R_{it}|Y_{it} = 0) = \text{Prob}(Z_{it} = 1|Y_{it} = 0) = p_{it}^{(0)}$, and $E(R_{it}|Y_{it} = 1) = \text{Prob}(Z_{it} = 0|Y_{it} = 1) = p_{it}^{(0)} + p_{it}^{(1)}$, where $p_{it}^{(0)} \geq 0$, $p_{it}^{(1)} \geq 0$ and $p_{it}^{(0)} + p_{it}^{(1)} \leq 1$. For notation conciseness, we do not write out X_{it} and β as conditions in the expectations and probabilities hereafter.

Based on equation (3) and the facts that $E(Y_{it}R_{it}) = E(Y_{it}R_{it}|Y_{it} = 1)\text{Prob}(Y_{it} = 1) + E(Y_{it}R_{it}|Y_{it} = 0)\text{Prob}(Y_{it} = 0) = (p_{it}^{(0)} + p_{it}^{(1)})\mu_{it}$ and $E(R_{it}) = E(R_{it}|Y_{it} = 1)\text{Prob}(Y_{it} = 1) + E(R_{it}|Y_{it} = 0)\text{Prob}(Y_{it} = 0) = p_{it}^{(1)}\mu_{it} + p_{it}^{(0)}$, we can derive that

$$\mu_{it}^* = \mu_{it} - 2p_{it}^{(0)} \times \mu_{it} - p_{it}^{(1)} \times \mu_{it} + p_{it}^{(0)} \quad (4)$$

Let $\mu_i^* = (\mu_{i1}^*, \dots, \mu_{iT}^*)'$, $p_i^{(0)} = (p_{i1}^{(0)}, \dots, p_{iT}^{(0)})$ and $p_i^{(1)} = (p_{i1}^{(1)}, \dots, p_{iT}^{(1)})$. Then, the MCGEE of the self-reported Z_i can be written as

$$U_\beta^*(\beta) = \sum_{i=1}^N D_i^* V_i^{*-1} (Z_i - \mu_i^*) = 0$$

where $D_i^* = \partial \mu_i^* / \partial \beta = (1 - 2p_i^{(0)} - p_i^{(1)}) \odot \partial \mu_i / \partial \beta = (1 - 2p_i^{(0)} - p_i^{(1)}) \odot A_i X_i$ with $\odot$ being the element-wise product, and all other settings are the same as those for equation (1). As an illustration, $(1 - 2p_i^{(0)} - p_i^{(1)}) \odot A_i X_i = ((1 - 2p_{i1}^{(0)} - p_{i1}^{(1)})a_{i1x_{i1}}, \dots, (1 - 2p_{iT}^{(0)} - p_{iT}^{(1)})a_{iT x_{iT}})'$. V_i^* is the working covariance matrix of Z_i , and can be decomposed into the form of $A_i^{*1/2} C_i^*(\gamma) A_i^{*1/2}$, where $C_i^*(\gamma)$ be the working correlation matrix of Z_i and $A_i^* = \text{diag}(\text{var}(Z_{i1}), \dots, \text{var}(Z_{iT}))$. The MCGEE of the self-reported outcomes Z_i can be written as

$$\begin{aligned} U_\beta^*(\beta) &= \sum_{i=1}^N (1 - 2p_i^{(0)} - p_i^{(1)}) \odot X_i' A_i (A_i^{*1/2} C_i^*(\gamma) A_i^{*1/2})^{-1} \\ &\quad (Z_i - (\mu_i - (2p_i^{(0)} + p_i^{(1)}) \odot \mu_i + p_i^{(0)})) = 0 \end{aligned}$$

The solution $\hat{\beta}$ can be obtained by Fisher's scoring algorithm, which first gives an initial guess for $\hat{\beta}^0$, and then updates $\hat{\beta}^l$ in the l th iteration by taking

$$\begin{aligned} \hat{\beta}^{l+1} &= \hat{\beta}^l - \left(\sum_{i=1}^N \hat{D}_i^* \hat{V}_i^{*-1} \hat{D}_i^* \right)^{-1} \\ &\quad \left(\sum_{i=1}^N \hat{D}_i^* \hat{V}_i^{*-1} (Y_i - \hat{\mu}_i) \right) \end{aligned} \quad (5)$$

where $Y_i = (Y_{i1}, \dots, Y_{iT})'$ and $\hat{\mu}_i = (\hat{\mu}_{i1}, \dots, \hat{\mu}_{iT})'$ are the estimated marginal means by equation (4). The variance of $\hat{\beta}$ can be estimated using the robust variance estimator (sandwich estimator):

$$\left(\sum_{i=1}^N \hat{D}_i^* \hat{V}_i^{*-1} \hat{D}_i^* \right)^{-1} \left(\sum_{i=1}^N \hat{D}_i^* \hat{V}_i^{*-1} \hat{A}_i^* \hat{V}_i^{*-1} \hat{D}_i^* \right) \left(\sum_{i=1}^N \hat{D}_i^* \hat{V}_i^{*-1} \hat{D}_i^* \right)^{-1} \quad (6)$$

Please refer to Liang and Zeger¹⁴ for more details on this standard GEE estimation procedure. Note that equation (6) should be used when the contamination probabilities are known. For unknown contamination probabilities, the variance of $\hat{\beta}$ can be estimated via a standard bootstrap method.

Next, we discuss the estimations of $p_{it}^{(0)}$ and $p_{it}^{(1)}$ in equation (4) when they are unknown in advance. In practice, since the contamination probability is often a subject-specific characteristic, we consider the homogeneous case in this paper; that is, we assume $p_{i1}^{(0)} = \dots = p_{iT}^{(0)} = p_{i*}^{(0)}$ and $p_{i1}^{(1)} = \dots = p_{iT}^{(1)} = p_{i*}^{(1)}$ for the i th subject. In the SCU data, urine samples were collected every $k = 7$ days, and the time period for the drug to be cleared from urine is h days, where h often takes values in $\{1, 2, 3, 4\}$ based on the situations. For the i th subject, let U_{ij} be the urine test result for the subject i at the j th measurement, where $j = 1, \dots, m_i$ and m_i is the total number of urine measurements for the subject i . We divide the whole time period into $m_i k$ -day blocks. In the j th block, if there is any true drug use $Y_{it} = 1$ ($t = kj - h + 1, \dots, kj$) within the clearing time, the urine test result U_{ij} will be 1; otherwise, $U_{ij} = 0$. Clearly, when the clearing time period $h = 1$, we have $U_{ij} = Y_{it}$, where $t = kj$, and it is straightforward to derive the unbiased estimators of contamination probabilities via comparing Z_{it} and Y_{it} given β and X_{it} . We have

$$\begin{aligned} \hat{p}_{i*}^{(0)} &= \text{Prob}(Z_{it} = 1 | Y_{it} = 0) = \frac{\sum_{j=1}^{m_i} Z_{it}(1 - U_{ij})}{\sum_{j=1}^{m_i} (1 - U_{ij})} \\ \hat{p}_{i*}^{(1)} &= \text{Prob}(Z_{it} = 0 | Y_{it} = 1) - \hat{p}_{i*}^{(0)} = \frac{\sum_{j=1}^{m_i} U_{ij}(1 - Z_{it})}{\sum_{j=1}^{m_i} U_{ij}} - \hat{p}_{i*}^{(0)} \end{aligned}$$

where $t = kj$, $\hat{p}_{i*}^{(0)}, \hat{p}_{i*}^{(1)} \geq 0$, and $\hat{p}_{i*}^{(0)} + \hat{p}_{i*}^{(1)} \leq 1$.

When the clearing time $h \geq 2$, since all true drug uses Y_{it} ($t = kj - h + 1, \dots, kj$) must be 0 during the h -day clearing period if its corresponding urine test result $U_{ij} = 0$, we can derive the unbiased estimators of $p_{i*}^{(0)}$ using only those self-reported data Z_{it} whose corresponding urine test results $U_{ij} = 0$. Then, we have

$$\hat{p}_{i*}^{(0)} = \frac{\sum_{j=1}^{m_i} \left(\sum_{t=kj-h+1}^{kj} Z_{it} \right) \times (1 - U_{ij})}{\sum_{j=1}^{m_i} h \times (1 - U_{ij})} \quad (7)$$

The estimation of $p_{i*}^{(1)}$ is more challenging for cases of $h \geq 2$. When having a positive urine test $U_{ij} = 1$, we will not know which Y_{it} ($t = kj - h + 1, \dots, kj$) is 1 and how many 1s are there. Let $S_{ij} = 1$ if any of the Z_{it} ($t = kj - h + 1, \dots, kj$) in the clearing time of the j th urine test is 1; otherwise, $S_{ij} = 0$. Using the self-reported data Z_{it} whose corresponding urine test results $U_{ij} = 1$, we can derive the lower bound of $\hat{p}_{i*}^{(1)}$:

$$\hat{p}_{i,\text{LOW}}^{(1)} = \frac{\sum_{j=1}^{m_i} U_{ij}(1 - S_{ij})}{\sum_{j=1}^{m_i} U_{ij}} - \hat{p}_{i*}^{(0)}$$

This lower bound is generally not achievable for $h \geq 2$, since it assumes only one true drug use when $U_{ij} = 1$ and does not consider which Y_{it} ($t = kj - h + 1, \dots, kj$) is 1 in the clearing period. Clearly, the true $\hat{p}_{i*}^{(1)}$ should be no less than $\hat{p}_{i,LOW}^{(1)}$ when sufficient data are available. Another extreme situation is that when $U_{ij} = 1$, all true drug uses Y_{it} ($t = kj - h + 1, \dots, kj$) are 1 in the clearing period, which leads to the upper bound of $\hat{p}_{i*}^{(1)}$:

$$\hat{p}_{i,UP}^{(1)} = \frac{\sum_{j=1}^{m_i} U_{ij} \left(\sum_{t=kj-h+1}^{kj} (1 - Z_{it}) \right)}{h \sum_{j=1}^{m_i} U_{ij}} - \hat{p}_{i*}^{(0)}$$

It is straightforward that the true $\hat{p}_{i*}^{(1)}$ should be no more than $\hat{p}_{i,UP}^{(1)}$ when sufficient data are available. In theory, any value of $\hat{p}_{i*}^{(1)}$ within the range $(\hat{p}_{i,LOW}^{(1)}, \hat{p}_{i,UP}^{(1)})$ is possible, and further assumptions are required to uniquely identify $\hat{p}_{i*}^{(1)}$. For practical use, we develop an estimator of $p_{i*}^{(1)}$ via assuming time-independence of Y_{it} in the clearing period. It works very well in both simulation studies and real data analysis even when this assumption is not true; see Sections 4 and 5 for details.

Under the time-independence assumption, we have the unbiased estimator of $\text{Prob}(Y_i = 0) = ((m_i - \sum_{j=1}^{m_i} U_{ij})/m_i)^{1/h}$. In

addition, it is straightforward that $\text{Prob}(Z_i = 0) = 1 - \sum_{t=1}^n Z_{it}/n$. Based on the Bayes rule, we have

$$\begin{aligned} \hat{p}_{i*}^{(0)} + \hat{p}_{i*}^{(1)} &= \text{Prob}(Z_i = 0 | Y_i = 1) \\ &= 1 - \text{Prob}(Z_i = 1 | Y_i = 1) \\ &= 1 - \frac{\text{Prob}(Z_i = 1)(1 - \text{Prob}(Y_i = 0 | Z_i = 1))}{\text{Prob}(Y_i = 1)} \\ &= 1 - \frac{1 - \text{Prob}(Z_i = 0)}{1 - \text{Prob}(Y_i = 0)} \\ &= \left(1 - \frac{\text{Prob}(Y_i = 0)\text{Prob}(Z_i = 1 | Y_i = 0)}{1 - \text{Prob}(Z_i = 0)} \right) \end{aligned}$$

Thus, we have

$$\begin{aligned} \hat{p}_{i*}^{(1)} &= 1 - \frac{\sum_{t=1}^n Z_{it}/n}{1 - ((m_i - \sum_{j=1}^{m_i} U_{ij})/m_i)^{1/h}} \\ &\quad \left(1 - \frac{\hat{p}_{i*}^{(0)}((m_i - \sum_{j=1}^{m_i} U_{ij})/m_i)^{1/h}}{\sum_{t=1}^n Z_{it}/n} \right) - \hat{p}_{i*}^{(0)} \end{aligned} \quad (8)$$

where the unbiased estimator $\hat{p}_{i*}^{(0)}$ is from equation (7). Note that when the data size is small, the calculated $\hat{p}_{i*}^{(1)}$ in equation (8) may be outside the range $(\hat{p}_{i,LOW}^{(1)}, \hat{p}_{i,UP}^{(1)})$. In such situations, we should use the bound to correct $\hat{p}_{i*}^{(1)}$. When analyzing real data, to avoid zero values in the denominators of equation (8), we can set the ranges of $\text{Prob}(Y_i = 0)$ and $\text{Prob}(Z_i = 0)$ to be $[\tau, 1 - \tau]$ where τ is a small enough positive value.

3 Asymptotic properties of the estimators

In practice, a large number of subjects (N) may be studied for a certain time period T . In this part, we study the asymptotic properties of the MCGEE estimator by proving its existence, consistency, and asymptotic normality as the sample size

$N \rightarrow \infty$, where the time period T is bounded and the true contamination probabilities are found for all subjects. Based on the seminal work of Liang and Zeger¹⁴ and Yuan and Jennrich,²² we aim to show that the solution $\hat{\beta}_N$ to the MCGEE $U_N(\beta) = 0$ is consistent, that is $\hat{\beta}_N \rightarrow \beta_0$ almost surely for the true value β_0 , and $\hat{\beta}_N$ is approximately normally distributed as $N \rightarrow \infty$. Similar to Liang and Zeger¹⁴ and Yuan and Jennrich,²² some standard assumptions are needed.

Assumption A: the subjects are independently sampled and there exists an upper bound $M < \infty$ such that $m_i < M$ for all subjects, where m_i is the number of urine measurements for subject i .

Assumption B: there exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects.

Assumption C: the matrix $(1/N) \sum_{i=1}^N X_i X_i' \rightarrow B$ as $N \rightarrow \infty$, where B is a positive definite matrix.

Assumption D: $\hat{\gamma}$ is $\sqrt{N}$ consistent given β .

Here, we provide some brief and intuitive understandings of these assumptions. Assumption A ensures that data from a finite number of subjects do not dominate the parameter estimation. Assumption B ensures that the estimating functions $(1/N)U_N(\beta)$ have bounded first- and higher-order derivatives with respect to β . Assumption C means that for sufficiently large N , $(1/N)E(\frac{\partial}{\partial \beta'} U_N(\beta))$ will be positive definite and there is no redundancy in the predictors. Assumption D ensures that the asymptotic properties of $\hat{\beta}$ do not depend on the choice of the γ estimator if it is $\sqrt{N}$ consistent. More detailed justifications on these assumptions can be found in Liang and Zeger¹⁴ and Yuan and Jennrich.²²

Let the N matrices $I_0^*(\beta) = \lim_{N \rightarrow \infty} (1/N) \frac{\partial}{\partial \beta'} U_N(\beta) = \lim_{N \rightarrow \infty} (1/N) \sum_{i=1}^N (D_i^* V_i^{*-1} D_i^*)$, and $I_1^*(\beta) = \lim_{N \rightarrow \infty} (1/N) \sum_{i=1}^N (D_i^* V_i^{*-1} A_i^* V_i^{*-1} D_i^*)$. It is straightforward that Assumption B ensures the existence of these limits and Assumption C ensures $I_0^*(\beta)$ and $I_1^*(\beta)$ being positive definite. The following theorem shows that the MCGEE is strongly consistent for β .

Theorem 1 Under Assumptions A–D, with probability one there exist zeros $\hat{\beta}_N$ of $U_N(\beta) = 0$ such that $\hat{\beta}_N \rightarrow \beta_0$ as $N \rightarrow \infty$.

Theorem 1 can be proved by showing $(1/N)U_N(\beta_0) \rightarrow 0$ a.s., as $N \rightarrow \infty$ and $(1/N)(\partial/\partial \beta')U_N(\beta)$ converges uniformly to a non-stochastic limit which is nonsingular at β_0 . The results then follow the Theorem 2 of Yuan and Jennrich.²² All detailed proofs in this section are included in the Appendix.

Next, we show the MCGEE estimator is approximately normally distributed for large N .

Theorem 2 Under assumptions A–D, $\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{L} N(0, I_0^{*-1}(\beta_0)I_1^*(\beta_0)I_0^{*-1}(\beta_0))$, as $N \rightarrow \infty$.

The asymptotic properties of $\hat{\beta}$ hold under $\sqrt{N}$ consistent estimates of the correlation parameters $\hat{\gamma}$ and the true values of the contamination probabilities $p_{i*}^{(0)}$ and $p_{i*}^{(1)}$. However, if contamination probabilities have not been correctly estimated, there may exist bias of $\hat{\beta}$

$$E_{\beta_0}(U_{\beta}^*(\beta)) = \sum_{i=1}^N D_i^* V_i^{*-1} (E(Z_i) - \mu_i^*) \neq 0$$

Instead, we can only have $E_{\beta_0}(U_{\beta}^*(\beta^*)) = 0$ $\hat{\beta}_N \rightarrow \beta^*$, as $N \rightarrow \infty$. In the next section, we conduct a simulation study to assess the bias of the MCGEE estimators on finite samples.

4 A simulation study

In this section, we conduct numerical studies simulating the SCU data with different sample sizes (N), time periods (T), contamination probabilities ($p^{(0)}$ and $p^{(1)}$) and cocaine clearing time (h). In both of the simulated treatment and control groups, there are $N/2$ subjects whose outcomes are repeatedly measured at T time points. The true drug uses are generated by Bernoulli trials with $\text{Prob}(Y_{it} = 1|X_i; \beta) = \mu_{it}$, and we assume $\log(\mu_{it}/(1 - \mu_{it})) = \beta_0 + \beta_1 X_i + \epsilon_i$ where $i = 1, \dots, N$ and $t = 1, \dots, T$. The treatment indicator $X_i = 1$ when the subject is in the treatment group; otherwise $X_i = 0$. The random effect ϵ_i follows the normal distribution with mean zero and constant variance $v = 0.01$. We assume that the true intercept β_0 and treatment effect β_1 are -0.5 and -1.5 , respectively. The correlation structures $C_i(\gamma)$ ($i = 1, \dots, N$) for the logit of the marginal mean of Y_{it} 's are assumed to be AR(1) with $\gamma = 0.1$. The AR(1) process is a commonly used correlation structure in longitudinal data analysis. It assumes that the current value in the series depends only on the value just prior to it. The parameters of the MCGEE approach are estimated using methods discussed in Section 2 and their standard errors are estimated using the standard bootstrap method. Note that we do not assume time-independence of Y_{it} 's in this simulation, and we will show the estimation of $p^{(1)}$ in equation (8) is still satisfactory. The urine samples are collected every $k = 7$ days, and their outcomes are the corresponding true drug uses. The self-reported outcomes Z_{it} are generated by flipping (i.e. reversing the binary outcomes) the true drug uses Y_{it} according to the contamination probabilities: $p_{it}^{(0)} = \text{Prob}(Z_{it} = 1|Y_{it} = 0)$ (for the cases of $Y = 0$) and $p_{it}^{(0)} + p_{it}^{(1)} = \text{Prob}(Z_{it} = 0|Y_{it} = 1)$ (for the cases of $Y = 1$).

Table 1. Parameters' estimations and SEs of the GEE and the MCGEE approaches when the clearing time $h = 1$.

Effect	N	T	$p^{(0)}$	$p^{(1)}$	$\hat{p}^{(0)}$	$\hat{p}^{(1)}$	GEE	MCGEE
β_0 (SE) True value = -0.5	100	140	0.05	0.2	0.04	0.22	-0.82(0.03)	-0.50(0.04)
				0.4	0.03	0.41	-1.23(0.03)	-0.56(0.05)
			0.1	0.2	0.07	0.24	-0.81(0.03)	-0.49(0.05)
				0.4	0.05	0.44	-1.26(0.03)	-0.55(0.04)
		280	0.05	0.2	0.04	0.22	-0.81(0.02)	-0.51(0.04)
				0.4	0.03	0.42	-1.24(0.03)	-0.53(0.04)
		280	0.1	0.2	0.07	0.24	-0.81(0.02)	-0.50(0.04)
				0.4	0.05	0.45	-1.26(0.03)	-0.51(0.03)
	200	140	0.05	0.2	0.04	0.22	-0.81(0.02)	-0.50(0.03)
				0.4	0.03	0.42	-1.23(0.02)	-0.56(0.04)
			0.1	0.2	0.07	0.25	-0.81(0.02)	-0.49(0.03)
				0.4	0.05	0.45	-1.26(0.02)	-0.55(0.04)
		280	0.05	0.2	0.04	0.22	-0.81(0.02)	-0.50(0.03)
				0.4	0.03	0.42	-1.23(0.02)	-0.52(0.03)
β_1 (SE) True value = -1.5	100	140	0.05	0.2	0.04	0.22	-1.14(0.05)	-1.51(0.07)
				0.4	0.03	0.41	-1.07(0.05)	-1.63(0.08)
			0.1	0.2	0.07	0.24	-0.95(0.04)	-1.46(0.07)
				0.4	0.05	0.44	-0.88(0.05)	-1.60(0.08)
		280	0.05	0.2	0.04	0.22	-1.14(0.03)	-1.46(0.05)
				0.4	0.03	0.42	-1.06(0.04)	-1.53(0.06)
				0.1	0.2	0.07	-0.96(0.03)	-1.43(0.06)
	200	140	0.05	0.4	0.05	0.45	-0.89(0.04)	-1.51(0.06)
				0.2	0.04	0.22	-1.14(0.03)	-1.50(0.05)
			0.1	0.4	0.03	0.42	-1.07(0.04)	-1.60(0.05)
				0.2	0.07	0.25	-0.96(0.03)	-1.44(0.06)
		280	0.05	0.4	0.05	0.45	-0.89(0.04)	-1.56(0.06)
				0.2	0.04	0.22	-1.14(0.03)	-1.46(0.04)
				0.4	0.03	0.42	-1.08(0.03)	-1.53(0.05)
		280	0.1	0.2	0.07	0.24	-0.96(0.02)	-1.43(0.04)
				0.4	0.05	0.45	-0.89(0.03)	-1.50(0.04)

SE: standard error; GEE: generalized estimating equation; MCGEE: mean corrected GEE.

Specifically, we consider several combinations of different sample sizes ($N = 100$ or 200), time periods ($T = 140$ or 280), contamination probabilities ($p^{(0)} = 0.05$ or 0.1 and $p^{(1)} = 0.2$ or 0.4), and the cocaine clearing time ($h = 1$ or 4). For each scenario, the contamination probabilities are assumed to be the same for different subjects. Thousand replications are performed for all methods and we show the median results. All analyses are implemented using *R*.

Table 1 reports the parameters' estimations and their standard errors for both the GEE and the MCGEE approaches under different scenarios when the cocaine clearing time h is one day. In the top two plots of Figure 1, we further compare the biases of the GEE and MCGEE estimators for β_1 under different scenarios. From Table 1 and Figure 1, it is seen that the biases of the MCGEE estimators are much lower than those of the GEE estimators for all cases. We can also see that the difference of biases between the GEE and MCGEE approaches increases as the contamination probability $p^{(0)}$ or $p^{(1)}$ increases. Clearly, the proposed MCGEE can significantly improve the traditional GEE for analyzing self-reported data subject to report bias. In addition, from Table 1, it is seen that the estimated contamination probabilities are very similar to their true values. In this simulation, although the time-independence assumption is not satisfied, our proposed method for estimating $p^{(1)}$ is still accurate. The standard errors of parameters' estimations in both the GEE and MCGEE approaches decrease as the sample sizes and time periods increase.

In Table 2, we report the parameters' estimations and their standard errors for both the GEE and the MCGEE when assuming the cocaine clearing time $h = 4$ days. In the bottom two plots of Figure 1, we compare the biases of the GEE and MCGEE estimators of β_1 for the $h = 4$ cases. From Table 2 and Figure 1, it is clear that the biases of the MCGEE estimators are much lower compared to those from the GEE. As $p^{(0)}$ or $p^{(1)}$ increases, the difference between the GEE and MCGEE estimators also increases, since it becomes easier for the MCGEE to detect the contamination and correct

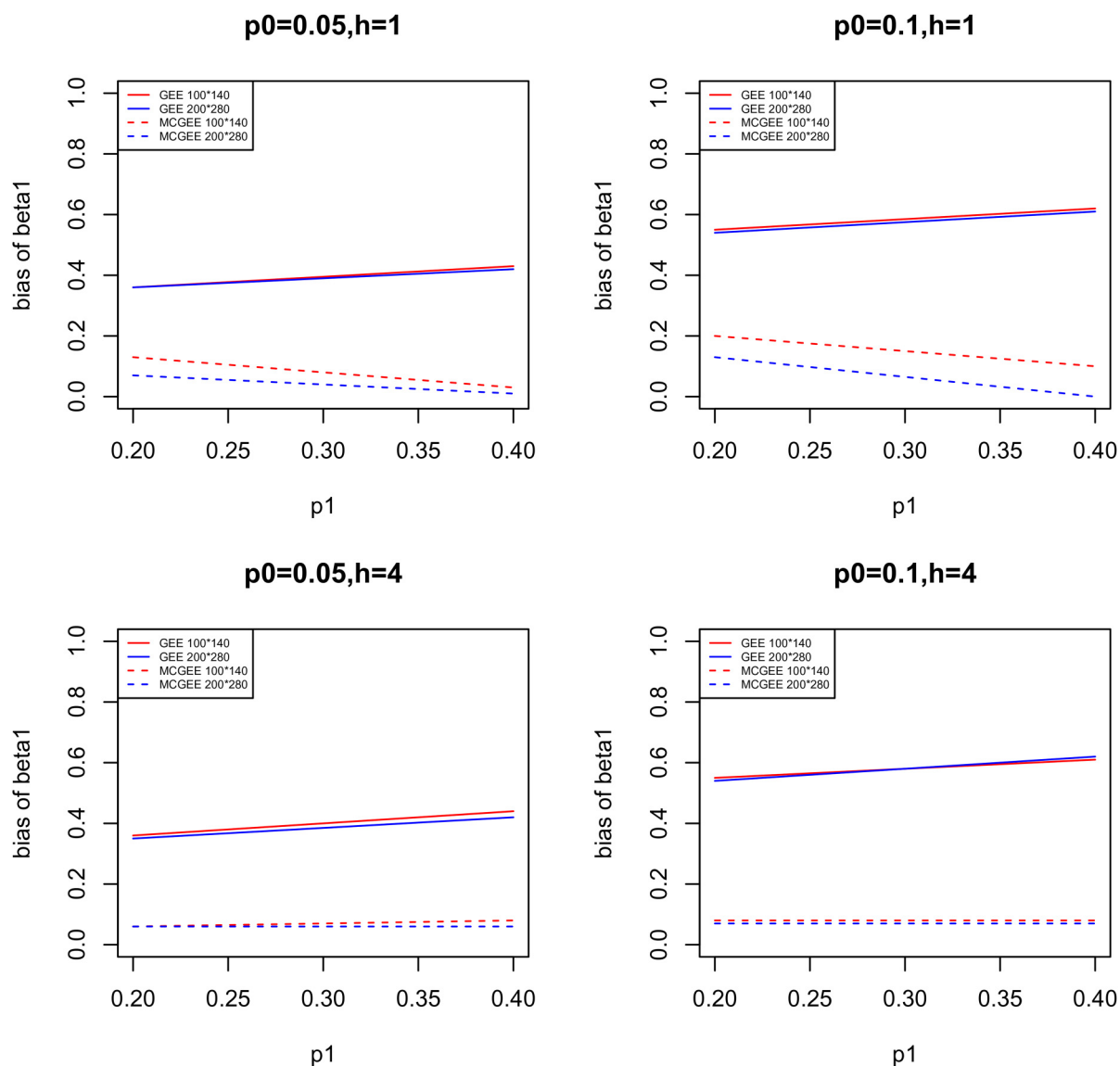


Figure 1. Biases of β_1 estimators in the GEE and MCGEE approaches. GEE: generalized estimating equation; MCGEE: mean corrected GEE.

the marginal means of the GEE. The estimated contamination probabilities in Table 2 are also very similar to their true values. Compared to the $h = 1$ cases in Table 1, the MCGEE estimators for the $h = 4$ cases in Table 2 are more stable but slightly less accurate. One potential reason is that we use the derived upper and lower bounds of $\hat{p}^{(1)}$ to further restrict the estimators, which may stabilize the estimation given limited data. From Table 2, we can also see the standard errors of parameters' estimates in both the GEE and MCGEE approaches decrease as the sample sizes and time periods increase. Note that we have also performed various simulations under similar circumstances with the correlation structure parameter γ up to 0.5. The resulting MCGEE parameters' estimations had minimal changes.

5 Analysis of the SCU data

In this section, we consider the SCU data collected at the Primary Care Center of Yale-New Haven Hospital, where 140 patients were followed for 1 to 6 months. All enrolled patients had met the current Diagnostic and Statistical Manual of Mental Disorders (4th edition) (DSM-IV) diagnostic criteria for cocaine dependence,²³ and they were randomly assigned to the treatment and control groups after a 2-week induction and stabilization period. Both groups received buprenorphine once per day, which is a substitute for cocaine. The control group received the physical management (PM), a 15–20 min

Table 2. Parameters' estimations and SEs of the GEE and the MCGEE approaches when the clearing time $h = 4$.

Effect	N	T	$p^{(0)}$	$p^{(1)}$	$\hat{p}^{(0)}$	$\hat{p}^{(1)}$	GEE	MCGEE
β_0 (SE) True value = -0.5	100	140	0.05	0.2	0.04	0.27	$-0.81(0.03)$	$-0.42(0.04)$
				0.4	0.03	0.46	$-1.23(0.04)$	$-0.45(0.04)$
			0.1	0.2	0.07	0.28	$-0.81(0.03)$	$-0.42(0.05)$
				0.4	0.05	0.48	$-1.26(0.03)$	$-0.43(0.04)$
		280	0.05	0.2	0.04	0.24	$-0.81(0.02)$	$-0.48(0.03)$
				0.4	0.03	0.44	$-1.24(0.03)$	$-0.49(0.04)$
			0.1	0.2	0.07	0.25	$-0.81(0.02)$	$-0.48(0.04)$
				0.4	0.05	0.46	$-1.26(0.03)$	$-0.48(0.04)$
	200	140	0.05	0.2	0.04	0.27	$-0.82(0.02)$	$-0.42(0.03)$
				0.4	0.03	0.46	$-1.23(0.02)$	$-0.45(0.04)$
			0.1	0.2	0.07	0.28	$-0.81(0.02)$	$-0.41(0.03)$
				0.4	0.05	0.48	$-1.26(0.02)$	$-0.44(0.03)$
		280	0.05	0.2	0.04	0.24	$-0.81(0.02)$	$-0.47(0.03)$
				0.4	0.03	0.44	$-1.23(0.02)$	$-0.49(0.03)$
			0.1	0.2	0.07	0.25	$-0.80(0.02)$	$-0.48(0.02)$
				0.4	0.05	0.46	$-1.27(0.02)$	$-0.49(0.02)$
β_1 (SE) True value = -1.5	100	140	0.05	0.2	0.04	0.27	$-1.14(0.05)$	$-1.44(0.07)$
				0.4	0.03	0.46	$-1.06(0.05)$	$-1.44(0.08)$
			0.1	0.2	0.07	0.28	$-0.95(0.04)$	$-1.42(0.07)$
				0.4	0.05	0.48	$-0.89(0.05)$	$-1.43(0.08)$
		280	0.05	0.2	0.04	0.24	$-1.15(0.04)$	$-1.44(0.05)$
				0.4	0.03	0.44	$-1.06(0.04)$	$-1.44(0.06)$
			0.1	0.2	0.07	0.25	$-0.95(0.03)$	$-1.42(0.06)$
				0.4	0.05	0.46	$-0.89(0.04)$	$-1.44(0.06)$
	200	140	0.05	0.2	0.04	0.27	$-1.15(0.03)$	$-1.45(0.05)$
				0.4	0.03	0.46	$-1.08(0.04)$	$-1.44(0.06)$
			0.1	0.2	0.07	0.28	$-0.95(0.03)$	$-1.43(0.05)$
				0.4	0.05	0.48	$-0.89(0.04)$	$-1.43(0.06)$
		280	0.05	0.2	0.04	0.24	$-1.15(0.03)$	$-1.44(0.04)$
				0.4	0.03	0.44	$-1.08(0.03)$	$-1.44(0.05)$
			0.1	0.2	0.07	0.25	$-0.96(0.02)$	$-1.43(0.05)$
				0.4	0.05	0.46	$-0.88(0.03)$	$-1.43(0.04)$

SE: standard error; GEE: generalized estimating equation; MCGEE: mean corrected GEE.

session by internal medicine physicians with experiences as buprenorphine providers. The treatment group received both the PM and a CBT. The CBT is a counseling intervention that has demonstrated the potential for treating a variety of psychiatric conditions and drug dependence. It was provided by master- or doctoral-level clinicians who were trained with a manual adapted from a guidance of using CBT for cocaine dependence.²⁴ Its main components focused on performing a functional analysis of behavior, promoting behavioral activation, identifying and coping with drug cravings, enhancing drug refusal and decision-making skills about high-risk situations, and improving problem-solving skills.¹⁰ The complete SCU data include results on cocaine use, marijuana use, alcohol use, bup use, and others. In this study, we focus on the analysis of cocaine use, and the proposed method can be easily adapted for analyzing other substances' uses. The outcomes on cocaine use include the daily self-reported cocaine use and the weekly urine test results on cocaine.

Table 3. Analysis of cocaine use via different approaches in the case study.

Model	$\hat{p}^{(0)}$	$\hat{p}^{(1)}$	Parameter	Estimate	SE	P-value
GEE	—	—	β_0	-3.31	0.26	<0.0001
			β_1	0.37	0.34	0.28
MCGEE ($h = 3$)	0.02	0.82	β_0	-1.02	0.09	<0.0001
			β_1	-0.09	0.12	0.45

SE: standard error; GEE: generalized estimating equation; MCGEE: mean corrected GEE.

The currently popular way to analyze this data is to apply the GEE approach with a logistic link and an AR(1) correlation structure using only the self-reported outcomes. We show its results in Table 3, where $\hat{\beta}_1 = 0.37$ with a p -value of 0.28. It indicates that there is no significant treatment effect. Yet, when we compare the self-reported drug uses and the urine test results, we find that there are roughly 21.4% mismatches, which may suggest significant contamination. To avoid the biases of GEE estimators, we need to identify the contamination probabilities and correct the marginal means of the self-reported data.

Given that the time period for cocaine to be cleared from urine is roughly 3 days,²⁵ we set $h = 3$ in the proposed MCGEE approach for this case study. We show its results in Table 3, where the MCGEE estimator $\hat{\beta}_1 = -0.09$ with a p -value of 0.45. Again, it indicates that the treatment effect is not significant. Compared to the GEE estimator $\hat{\beta}_1 = 0.37$, we think the MCGEE estimator $\hat{\beta}_1 = -0.09$ makes more practical sense, since the CBT is generally considered to be beneficial or at least not harmful,²⁴ which would correspond to a negative treatment effect (though not significant). While the GEE and MCGEE methods reached the same conclusion with insignificant estimates of treatment effects, we have more confidence in the effect size reported from the proposed MCGEE method based on the background study on contamination. In the SCU data, there are a substantial amount of missing records on urine tests, resulting in a significant drop in the effective sample size. For some patients, the total time periods T are as small as 8 after excluding the missing records. To draw a more deterministic conclusion on the CBT effect, follow-up studies are needed to increase the effective sample size that can be used in our proposed MCGEE model.

6 Discussion

In this paper, we propose an MCGEE approach for analyzing the longitudinal binary outcomes with report bias. When the contamination probabilities can be correctly specified, the MCGEE approach will yield consistent parameters' estimates, and such consistency is retained even when the working correlation matrix is misspecified. We propose to use the weekly urine test results to estimate the contamination probabilities and correct the marginal means of the daily self-reported results. Compared to the traditional GEE method, the MCGEE approach is found to be more accurate in parameters' estimations in both simulation studies and real data analysis, when the self-reported data are subject to significant report bias.

When the time period for cocaine to be cleared from urine is $h = 1$ day, the derived contamination probabilities $p_{i*}^{(0)}$ and $p_{i*}^{(1)}$ in Section 2 are unbiased, and thus the developed MCGEE estimators are consistent. Yet, when the clearing time $h \geq 2$ days, we will only have unbiased estimators of $p_{i*}^{(0)}$ and the upper and lower bounds for the estimators of $p_{i*}^{(1)}$ without further assumptions. In practice, we find that the estimators of $p_{i*}^{(1)}$ based on a time-independence assumption works very well. It provides reasonable performances even when this assumption is not satisfied. An interesting future research question is how to develop estimators $\hat{p}_{i*}^{(1)}$ for the $h \geq 2$ cases to further improve the performances. One possible way is to assume known working correlation structures, for example, exchangeable or AR(1) correlation, and then apply the Monte Carlo (MC) method. We can start from the current estimators of $p_{i*}^{(0)}$ and $p_{i*}^{(1)}$, and then estimate all parameters in the MCGEE approach under the assumed correlation structures. Subsequently, we simulate random paths of the true responses, update the contamination probabilities based on the simulated true outcomes, and repeat these steps until convergence. The MC method can be computationally expensive and unstable, which requires additional studies from both the theoretical and practical aspects.

Declaration of conflicting interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ye Shen  <https://orcid.org/0000-0002-6662-0048>

Supplemental Material

Supplemental material for this article is available online.

References

1. Macdonald S, Erickson P, Wells S, et al. Predicting violence among cocaine, cannabis, and alcohol treatment clients. *Addict Behavior* 2008; **33**: 201–205.

2. Qureshi AI, Suri MF, Guterman LR, et al. Cocaine use and the likelihood of nonfatal myocardial infarction and stroke: data from the third national health and nutrition examination survey. *Circulation* 2001; **103**: 502–506.
3. Brown EE, Finlay JM, Wong JT, et al. Behavioral and neurochemical interactions between cocaine and buprenorphine: implications for the pharmacotherapy of cocaine abuse. *J Pharmacol Exp Ther* 1991; **256**: 119–125.
4. Kamien JHB and Spealman RD. Modulation of the discriminative-stimulus effects of cocaine by buprenorphine. *Behav Pharmacol* 1991; **2**: 517–520.
5. Mooney LJ, Nielsen S, Saxon A, et al. Cocaine use reduction with buprenorphine (CURB): rationale, design, and methodology. *Contemp Clin Trials* 2013; **34**: 196–204.
6. Magill M and Ray LA. Cognitive-behavioral treatment with adult alcohol and illicit drug users: a meta-analysis of randomized controlled trials. *J Stud Alcohol Drugs* 2009; **70**: 516–527.
7. Beck AT, Wright F, Newman C, et al. *Cognitive therapy of substance abuse*. New York: Guilford Press, 1993.
8. Maude-Griffin PM, Hohenstein JM, Humfleet GL, et al. Superior efficacy of cognitive-behavioral therapy for urban crack cocaine abusers: Main and matching effects. *J Consult Clin Psychol* 1998; **66**SEP: 832–837.
9. Gonzalez VM, Schmitz JM and DeLaune KA. The role of homework in cognitive-behavioral therapy for cocaine dependence. *J Consult Clin Psychol* 2006; **74**: 633–637.
10. Fiellin DA, Barry DT, Sullivan LE, et al. A randomized trial of cognitive behavioral therapy in primary care-based buprenorphine. *Am J Med* 2013; **126**: 74.
11. Del Boca FK and Noll JA. Truth or consequences: The validity of self-report data in health services research on addictions. *Addiction* 2000; **95**: 347–360.
12. Neuhaus JM. Analysis of clustered and longitudinal binary data subjects to response misclassification. *Biometrics* 2002; **58**: 675–683.
13. Chen Z, Yi GY and Wu C. Marginal methods for correlated binary data with misclassified responses. *Biometrika* 2011; **98**: 647–662.
14. Liang K and Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
15. Fitzmaurice GM, Davidian M, Verbeke G, et al. *Longitudinal data analysis*. Boca Raton, FL: Chapman Hall/CRC, 2009.
16. Magura S, Goldsmith D, Casriel C, et al. The validity of methadone clients' self-reported drug use. *Int J Addict* 1987; **22**: 727–749.
17. Babor TF, Steinberg K, Anton R, et al. Talk is cheap: Measuring drinking outcomes in clinical trials. *J Stud Alcohol Drugs* 2000; **61**: 55–63.
18. Blattman C, Jamisonb J, Koroknay-Paliczc T, et al. Measuring the measurement error: A method to qualitatively validate survey data. *J Dev Econ* 2016; **120**: 99–112.
19. Wilcox CE, Bogenschutz MP, Nakazawa M, et al. Concordance between self-report and urine drug screen data in adolescent opioid-dependent clinical trial participants. *Addict Behav* 2013; **38**: 2568–2574.
20. Munoz A, Carey V, Schouten JP, et al. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* 1992; **48**: 733–742.
21. Westgate PM. Improving the correlation structure selection approach for generalized estimating equations and balanced longitudinal data. *Stat Med* 2014; **33**: 2222–2237.
22. Yuan KH and Jennrich RI. Asymptotics of estimating equations under natural conditions. *J Multivar Anal* 1998; **65**: 245–260.
23. Argoff CE and McCleane G. *Pain Management Secrets* (3rd ed.). Philadelphia, PA: Mosby/Elsevier, 2009.
24. Carroll KM. *A cognitive behavioral approach: Treating cocaine addiction*. Rockville, MD: National Institute on Drug Abuse, 1998.
25. Cone EJ, Tsadik A, Oyler J, et al. Cocaine metabolism and urinary excretion after different routes of administration. *Ther Drug Monit* 1998; **20**: 556–560.

Appendix I

From Section 3, $U_N(\beta) = \sum_{i=1}^N D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)$, define $\psi(Z_i; \beta) = D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)$. The assumptions we need to prove the consistency and asymptotic normality of $\hat{\beta}_N$ are:

Assumption A. The subjects are independently sampled and there exists an upper bound $M < \infty$ such that the number of replicates $m_i < M$ for all subjects $i = 1, 2, \dots$

Assumption B. There exists an upper bound $b < \infty$ such that $|X_i| < b$ for all subjects $i = 1, 2, \dots$

Assumption C. It is assumed that $\frac{1}{N} \sum X_i X_i' \rightarrow B$ as $N \rightarrow \infty$, where B is a positive definite matrix.

Assumption D. $\hat{\gamma}$ is $\sqrt{N}$ consistent given β .

Assumption E. $\hat{\rho}_i$ is $\sqrt{N}$ consistent given β .

Define the matrices

$$\begin{aligned}
 I_0^*(\beta) &= \lim_{N \rightarrow \infty} \frac{1}{N} \frac{\partial}{\partial \beta'} U_N(\beta) \\
 &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (D_i^{*'} V_i^{*-1} D_i^*)
 \end{aligned}$$

and

$$I_1^*(\beta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^*)$$

The existence of these limits is ensured by Assumption B. Moreover, Assumption C ensures that $I_0^*(\beta)$ and $I_1^*(\beta)$ are positive definite.

To prove the solution $\hat{\beta}_N$ of $U_N(\beta) = 0$ is consistent and asymptotic normally distributed for large N , we need to show that

1. $(1/N)U_N(\beta_0) \rightarrow 0$ a.s., as $N \rightarrow \infty$.
2. $(1/N)\frac{\partial}{\partial \beta^T} U_N(\beta)$ converge uniformly to a nonstochastic limit, which is nonsingular at β_0 .
3. With probability one, $\psi(Z_i; \beta)$ are twice continuously differentiable with respect to $\beta \in B$, and $|\frac{\partial^2}{\partial \beta_j \partial \beta_k} \psi(Z_i; \beta)| < \infty$.
4. $|\psi(Z_i; \beta)| < \infty$, and $\frac{1}{\sqrt{N}} U_N(\beta_0) LN(0, I_1^*(\beta_0))$.

Since our proof is based on some conditions and theorems from Yuan and Jennrich,²² we verify their conditions in our case in the following section, consistency and the asymptotic normality of MCGEE estimator will be proved when the conditions are satisfied.

A.1 Verifying the conditions

Yuan and Jennrich²² proved consistency and asymptotic normality of M estimators based on the following conditions, which are:

1. $(1/N)U_N(\beta_0) \rightarrow 0$ a.s., as $N \rightarrow \infty$.
2. There exists a neighborhood M of β_0 on which with probability one, all $\frac{1}{N} U_N(\beta)$ are continuously differentiable and $(1/N)(\partial/\partial \beta^T)U_N(\beta)$ converge uniformly to a nonstochastic limit which is nonsingular at β_0 .
- 3.

$$\frac{1}{\sqrt{N}} U_N(\beta_0) LN(0, I_1^*(\beta_0))$$

as $N \rightarrow \infty$. To prove that the MCGEE estimator is consistent and asymptotically normally distributed, we will demonstrate the conditions above are satisfied.

By Theorem 5 of Yuan and Jennrich,²² to verify condition 1 of Yuan and Jennrich²² it suffices to verify the condition 4 of Yuan and Jennrich.²²

4. For each i , $\psi(Z_i; \beta_0)$ has mean zero and variance–covariance matrix K_i , such that

$$\frac{1}{N} \sum_{i=1}^N K_i \rightarrow K$$

for some positive-definite matrix K .

For self-reported data Z_i , since $E(Z_i) = \mu_i^*$, then $E(\psi(Z_i; \beta_0)) = 0$, and

$$\text{var}(\psi(Z_i; \beta_0)) = D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^*$$

$$D_i^* = (1 - 2p_i) \odot A_i X_i$$

$$A_i^* = \text{diag}(\text{var}(Z_{i1}), \dots, \text{var}(Z_{iT}))$$

$$\text{var}(Z_{it}) = (1 - 2p_{it})^2 \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} + p_{it}(1 - p_{it})$$

then

$$A_i^* = \text{diag}((1 - 2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1 + e^{\beta' X_{i1}})^2} + p_{i1}(1 - p_{i1}), \dots, (1 - 2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1 + e^{\beta' X_{iT}})^2} + p_{iT}(1 - p_{iT}))$$

Since $|X_i| < b < \infty$ for all $i = 1, 2, \dots$ by assumption, and the contamination probability, $0 \leq p_{it} \leq 1$, then

$$0 \leq (1 - 2p_i)^2 \leq 1$$

$$0 \leq p_{it}(1 - p_{it}) \leq \frac{1}{4}$$

and

$$0 \leq \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} \leq \frac{1}{4}$$

Therefore

$$D_i^* = (1 - 2p_i) \odot A_i X_i < \infty$$

$$0 \leq \text{var}(Z_{it}) = (1 - 2p_{it})^2 \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} + p_{it}(1 - p_{it}) \leq \frac{1}{2}$$

and

$$V_i = A_i^{*1/2} C_i^*(\gamma) A_i^{*1/2} < \infty$$

Hence, the variance-covariance matrix K_i of $\psi(\beta_0)$ satisfies

$$\frac{1}{N} \sum_{i=1}^N K_i \rightarrow K$$

for some positive-definite matrix K . Condition 1 of Yuan and Jennrich²² has been verified.

To verify condition 2 of Yuan and Jennrich²² it suffices to verify the following assumptions of Yuan and Jennrich²²:

5. With probability one, $\psi(Z_i; \beta)$ are twice continuously differentiable with respect to $\beta \in B$.
6. For each $\beta \in B$

$$\frac{1}{N} \sum_{i=1}^N E\left(\frac{\partial}{\partial \beta^T} \psi(Z_i; \beta)\right) \rightarrow I_0^*(\beta)$$

where $I_0^*(\beta)$ is nonsingular and with probability one

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial \beta_t} \psi(Z_i; \beta) \rightarrow I_0^*(\beta)$$

as $N \rightarrow \infty$.

7. For each i

$$\left| \frac{\partial^2}{\partial \beta_j \partial \beta_k} \psi(Z_i; \beta) \right| \leq S$$

for some upper bound $S < \infty$.

Yuan and Jennrich²² proved that under conditions 6, 7, and 8, condition 2 is satisfied. To verify condition 6, we have

$$\begin{aligned}\frac{\partial}{\partial \beta} \psi(Z_i; \beta) &= \frac{\partial}{\partial \beta} (D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)) \\ &= D_i^{*'} V_i^{*-1} D_i^* + \left(\frac{\partial}{\partial \beta} D_i^{*'} \right)' V_i^{*-1} (Z_i - \mu_i^*) \\ &\quad + D_i^{*'} \left(\frac{\partial}{\partial \beta} V_i^{*-1} \right) (Z_i - \mu_i^*)\end{aligned}$$

Since $E(Z_i) = \mu_i^*$, the last two terms in the expression above have expectation zero, so

$$E\left(\frac{\partial}{\partial \beta} \psi(Z_i; \beta)\right) = D_i^{*'} V_i^{*-1} D_i^*$$

Moreover

$$\frac{\partial}{\partial \beta} D_i^* = (1 - 2p_i) \odot \left(\frac{\partial}{\partial \beta} A_i \right) X_i$$

where $A_i = \text{diag}(\text{var}(Y_{i1}), \dots, \text{var}(Y_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial \beta} (\text{var}(Y_{it})) &= \frac{\partial}{\partial \beta} \frac{e^{\beta' X_{it}}}{(1 + e^{\beta' X_{it}})^2} \\ &= \frac{X_{it} e^{\beta' X_{it}} (1 - e^{\beta' X_{it}})}{(1 + e^{\beta' X_{it}})^3} \\ \frac{\partial}{\partial \beta} A_i &= \text{diag}\left(\frac{X_{i1} e^{\beta' X_{i1}} (1 - e^{\beta' X_{i1}})}{(1 + e^{\beta' X_{i1}})^3}, \dots, \right. \\ &\quad \left. \frac{X_{iT} e^{\beta' X_{iT}} (1 - e^{\beta' X_{iT}})}{(1 + e^{\beta' X_{iT}})^3}\right)\end{aligned}$$

Since $|X_i| < b < \infty$, $0 \leq (1 - 2p_i) \leq 1$, $0 \leq e^{\beta' X_{it}} / (1 + e^{\beta' X_{it}})^2 \leq 1/4$, and $0 \leq 1/(1 + e^{\beta' X_{it}}) \leq 1$

$$\frac{\partial}{\partial \beta} D_i^* < \infty$$

And

$$\begin{aligned}\frac{\partial}{\partial \beta} V_i^{*-1} &= -V_i^{*-1} \left(\frac{\partial}{\partial \beta} V_i^* \right) V_i^{*-1} \\ \frac{\partial}{\partial \beta} V_i^* &= \left(\frac{\partial}{\partial \beta} A_i^{*1/2} \right) C_i(\gamma) A_i^{*1/2} + A_i^{*1/2} C_i(\gamma) \left(\frac{\partial}{\partial \beta} A_i^{*1/2} \right)\end{aligned}$$

$A_i^* = \text{diag}(\text{var}(Z_{i1}), \dots, \text{var}(Z_{iT}))$, and

$$\begin{aligned}\frac{\partial}{\partial \beta} (\text{var}(Z_{it})) &= \frac{\partial}{\partial \beta} \left((1 - 2p_{it})^2 \frac{e^{X_{it}\beta}}{(1 + e^{X_{it}\beta})^2} + p_{it}(1 - p_{it}) \right) \\ &= (1 - 2p_{it})^2 \frac{X_{it} e^{X_{it}\beta} (1 - e^{X_{it}\beta})}{(1 + e^{X_{it}\beta})^3}\end{aligned}$$

Then

$$\begin{aligned}
 \frac{\partial}{\partial \beta} A_i^{*1/2} &= \text{diag} \left[\frac{\partial}{\partial \beta} \sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}, \dots, \right. \\
 &\quad \left. \frac{\partial}{\partial \beta} \sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})} \right] \\
 &= \frac{1}{2} \text{diag} \left[\frac{(1-2p_{i1})^2 \frac{X_{i1} e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} - 2 \frac{X_{i1} e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, \dots, \right. \\
 &\quad \left. \frac{(1-2p_{iT})^2 \frac{X_{iT} e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} - 2 \frac{X_{iT} e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}} \right] \\
 &= \frac{1}{2} \text{diag} \left[\frac{(1-2p_{i1})^2 \mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}}, \dots, \right. \\
 &\quad \left. \frac{(1-2p_{iT})^2 \mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}} \right]
 \end{aligned}$$

Therefore, $(\partial/\partial \beta) A_i^{*1/2} < \infty$, $((\partial/\partial \beta) V_i^* < \infty$, $(\partial/\partial \beta) D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*) + D_i^{*'} (\frac{\partial}{\partial \beta} V_i^{*-1}) (Z_i - \mu_i^*) < \infty$, and

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial \beta} \psi(Z_i; \beta) = \frac{1}{N} \sum_{i=1}^N D_i^* V_i^{*-1} D_i^*$$

Taking the second derivative

$$\begin{aligned}
 \frac{\partial}{\partial \beta} (D_i^* V_i^{*-1} D_i^*) &= \left(\frac{\partial}{\partial \beta} D_i^* \right) V_i^{*-1} D_i^* \\
 &\quad + D_i^{*'} \left(\frac{\partial}{\partial \beta} V_i^{*-1} \right) D_i^* \\
 &\quad + D_i^{*'} V_i^{*-1} \left(\frac{\partial}{\partial \beta} D_i^* \right)
 \end{aligned}$$

We also have

$$\begin{aligned}
 \frac{\partial}{\partial \beta} D_i^* &= (1-2p_i) \odot \left(\frac{\partial}{\partial \beta} A_i \right) X_i \\
 \frac{\partial}{\partial \beta} A_i &= \text{diag} \left(\frac{X_{i1} e^{\beta' X_{i1}} (1 - e^{\beta' X_{i1}})}{(1 + e^{\beta' X_{i1}})^3}, \dots, \right. \\
 &\quad \left. \frac{X_{iT} e^{\beta' X_{iT}} (1 - e^{\beta' X_{iT}})}{(1 + e^{\beta' X_{iT}})^3} \right).
 \end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \beta} V_i^{*-1} &= -V_i^{*-1} \left(\frac{\partial}{\partial \beta} V_i^* \right) V_i^{*-1} \\ \frac{\partial}{\partial \beta} V_i^* &= \left(\frac{\partial}{\partial \beta} A_i^{*1/2} \right) C_i(\gamma) A_i^{*1/2} + A_i^{*1/2} C_i(\gamma) \left(\frac{\partial}{\partial \beta} A_i^{*1/2} \right)\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial}{\partial \beta} A_i^{*1/2} &= \text{diag} \left[\frac{\partial}{\partial \beta} \sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})} \right. \\ &\quad , \dots , \\ &\quad \left. \frac{\partial}{\partial \beta} \sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})} \right] \\ &= \frac{1}{2} \text{diag} \left[\frac{(1-2p_{i1})^2 \frac{X_{i1} e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} - 2 \frac{X_{i1} e^{2\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^3}}{\sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}} \right. \\ &\quad , \dots , \\ &\quad \left. \frac{(1-2p_{iT})^2 \frac{X_{iT} e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} - 2 \frac{X_{iT} e^{2\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^3}}{\sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}} \right] \\ &= \frac{1}{2} \text{diag} \left[\frac{(1-2p_{i1})^2 \mu_{i1}(1-\mu_{i1})(1-2\mu_{i1})X_{i1}}{\sqrt{(1-2p_{i1})^2 \frac{e^{\beta' X_{i1}}}{(1+e^{\beta' X_{i1}})^2} + p_{i1}(1-p_{i1})}} \right. \\ &\quad , \dots , \\ &\quad \left. \frac{(1-2p_{iT})^2 \mu_{iT}(1-\mu_{iT})(1-2\mu_{iT})X_{iT}}{\sqrt{(1-2p_{iT})^2 \frac{e^{\beta' X_{iT}}}{(1+e^{\beta' X_{iT}})^2} + p_{iT}(1-p_{iT})}} \right]\end{aligned}$$

We have already shown that $(\partial/\partial \beta) D_i^* < \infty$, $(\partial/\partial \beta) A_i^{*1/2} < \infty$, and $(\partial/\partial \beta) V_i^* < \infty$. Condition 6 is verified.

To verify condition 7 of Yuan and Jennrich, Yuan and Jennrich²² the derivative of $\psi(Z_i; \beta)$ with respect to β is

$$\begin{aligned}\frac{\partial}{\partial \beta} \psi(Z_i; \beta) &= \frac{\partial}{\partial \beta_t} (D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)) \\ &= D_i^* V_i^{*-1} D_i^* + \left(\frac{\partial}{\partial \beta} D_i^* \right)' V_i^{*-1} (Z_i - \mu_i^*) \\ &\quad + D_i^{*'} \left(\frac{\partial}{\partial \beta} V_i^{*-1} \right) (Z_i - \mu_i^*)\end{aligned}$$

We have already shown the following equations when we verifying condition 6:

$$E\left(\frac{\partial}{\partial\beta}\psi(Z_i; \beta)\right) = D_i^* V_i^{*-1} D_i^*$$

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial\beta}\psi(Z_i; \beta) = \frac{1}{N} \sum_{i=1}^N D_i^* V_i^{*-1} D_i^*$$

To complete verifying condition 7, we need to show that

$$\frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial\beta}\psi(Z_i; \beta) \rightarrow I_0^*(\beta)$$

almost surely as $N \rightarrow \infty$.

Since

$$\frac{1}{N} \sum_{i=1}^N D_i^* V_i^{*-1} D_i^* = \frac{1}{N} \sum_{i=1}^N ((1 - 2p_i) \odot A_i X_i')' V_i^{*-1} (1 - 2p_i) \odot A_i X_i'$$

and $(1 - 2p_i)$, V_i , A_i are all bounded from the previous proof. Then $((1 - 2p_i) \odot A_i)' V_i^{*-1} (1 - 2p_i) \odot A_i$ is bounded below by a positive constant b_i .

Let a denote any $T \times 1$ vector, then

$$\begin{aligned} & \frac{1}{N} a' \sum_{i=1}^N X_i ((1 - 2p_i) \odot A_i)' V_i^{*-1} (1 - 2p_i) \odot A_i X_i' a \\ & \geq \frac{1}{N} b_i a' \sum_{i=1}^N X_i X_i' > 0 \end{aligned}$$

by Assumption C, which is

$$\frac{1}{N} \sum_{i=1}^N X_i X_i' \rightarrow B$$

as $N \rightarrow \infty$, where B is a positive definite matrix.

Then

$$\begin{aligned} & \frac{\partial}{\partial\beta}\psi(Z_i; \beta) < \infty \\ & \frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial\beta}\psi(Z_i; \beta) \rightarrow I_0^*(\beta) \end{aligned}$$

almost surely as $N \rightarrow \infty$.

To verify condition 8 of Juan and Jennrich,²² we have already shown that each term of the second derivatives of $\psi(Z_i; \beta)$ with respect to β is bounded when we verify condition 6 $((\partial/\partial\beta)D_i^* < \infty, (\partial/\partial\beta)V_i^{*-1} < \infty)$.

Hence

$$\frac{\partial^2}{\partial\beta\partial\beta}\psi(Z_i; \beta) < \infty$$

In conclusion, condition 2 of Yuan and Jennrich²² has been verified.

Liapounov's theorem and Cramer–Wald theorem are used to verify condition 3 of Yuan and Jennrich²²

$$\frac{1}{\sqrt{N}} U_N(\beta_0) L N(0, I_1^*(\beta_0))$$

as $N \rightarrow \infty$.

As defined earlier

$$U_N(\beta_0) = \sum_{i=1}^N \psi(Z_i; \beta_0) = \sum_{i=1}^N D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)$$

Let a denote any $T \times 1$ vector, to apply Liapounov's theorem, take

$$r_i = a' D_i^{*'} V_i^{*-1} Z_i$$

Then the mean of r_i is

$$m_i = E(r_i) = a' D_i^{*'} V_i^{*-1} \mu_i^*$$

and the variance of r_i is

$$\text{Var}(r_i) = a' D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a$$

Define

$$\begin{aligned} c_n^2 &= \sum_{i=1}^N \text{Var}(r_i) \\ &= \sum_{i=1}^N a' D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a \\ &= O(N) \end{aligned}$$

since

$$\frac{1}{N} \sum_{i=1}^N D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^* \rightarrow I_1^*$$

under condition 4.

Assume $E(|Z_i - \mu_i^*|^3) = \mu_{3i}^* < \infty$. Taking $\delta = 1$, the third central moment is

$$\begin{aligned} E(|r_i - m_i|^3) &= E(|a' D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)|^3) \\ &\leq (a' D_i^{*'} V_i^{*-1})^3 E(|Z_i - \mu_i^*|^3) \\ &= (a' D_i^{*'} V_i^{*-1})^3 \mu_{3i}^* \end{aligned}$$

So

$$\sum_{i=1}^N E(|r_i - m_i|^3) = O(N)$$

since D_i^* and V_i^* are bounded, which have been shown when verifying condition 4 of Yuan and Jennrich.²²

Then

$$\frac{1}{c_n^3} \sum_{i=1}^N E(|r_i - m_i|^3) = \frac{O(N)}{O(N^{3/2})} = O(N^{-1/2})$$

which converges to zero as $N \rightarrow \infty$. Therefore, the conditions of Liapounov's theorem are satisfied, and

$$\begin{aligned} T_N &= \frac{\sum_{i=1}^N (r_i - m_i)}{c_n} \\ &= \frac{\sum_{i=1}^N a' D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*)}{\sqrt{\sum_{i=1}^N a' D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^* a}} \\ &\rightarrow LN(0, 1) \end{aligned}$$

as $N \rightarrow \infty$.

By Slutsky's theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N a' D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*) LN(0, a' I_1^*(\beta) a)$$

By the Cramer-Wold theorem

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N D_i^{*'} V_i^{*-1} (Z_i - \mu_i^*) LN(0, I_1^*(\beta))$$

where $I_1^*(\beta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (D_i^{*'} V_i^{*-1} A_i^* V_i^{*-1} D_i^*)$.

Thus, condition 3 has been verified.

Therefore, the consistency and asymptotic normality of the MCGEE estimator have been proved.