Contemporary Clinical Trials Communications 4 (2016) 90–98

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Estimation of response from longitudinal binary data with nonignorable missing values in migraine trials

Fang Fang [a, *], Xiaoyin Fan [b], Ying Zhang [c]

[a] East China Normal University, Shanghai, 200241, China
[b] Novartis Pharmaceuticals, 45 Sidney Street, Cambridge, MA, 02139, USA
[c] Merck Research Laboratories, 351 North Sumneytown Pike, North Wales, PA, 19454, USA

ARTICLE INFO

Article history:
Received 12 September 2015
Received in revised form
10 June 2016
Accepted 17 June 2016
Available online 16 July 2016

Keywords:
Longitudinal binary data
Nonignorable missing
Complete-case analysis
Imputation
Bootstrap

ABSTRACT

In migraine trials pain relief responses from a headache at specific time points and sustained pain relief response over a period of time are important efficacy measures. When there are missing records of individual time point pain scores and/or headache recurrences during a migraine trial, the common approach used in practice to estimate the sustained response is statistically inconsistent even if the data are missing completely at random. Methods dealing with nonignorable longitudinal missing data usually assume certain models for the missing mechanism which can not be checked as they involve unobserved data. Taking advantage of the specific definition of the 'sustained pain relief' response, we propose two estimating methods based on intuitive imputation, which do not require model assumptions on the missing probability or specification of the correlation structure among the longitudinal observations. The consistency of the proposed methods is discussed in theory and their empirical performances are assessed through intensive simulation studies. The simulation results show that the proposed methods perform well in terms of reducing bias and mean square error except in several extreme cases which are unlikely to happen in real trials. The application of the proposed methods is illustrated in a real data analysis.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

This paper was motivated by a dose-finding clinical trial [9], in which different doses of an investigational drug were studied to treat patients with moderate or severe migraine headaches. Patients were required to report their headache severities at baseline and a few specific time points, typically 0.5, 1, 1.5, 2, 3, 4, and 24 h post initial dose. The headache severity was measured on a 4-point scale: 0 = none, 1 = mild, 2 = moderate, 3 = severe. Headache recurrence, defined as moderate or severe headache at any time during the 2–24 h after an initial pain relief (PR, defined as none or mild headache) at 2 h postdose, was also reported. It needs to be noted that the recurrence covers the continuous time period so it is possible for a patient to report PR at all specific time points while still having a recurrence. Patients were allowed to take an optional 2nd dose and/or rescue medication at 2 h postdose or later. One of the secondary endpoint in the study is the 2–24 h sustained PR (SPR), defined as PR at 2 h, no need for the optional 2nd dose or rescue medication and no moderate or severe headache recurrence during the 2–24 h postdose. 2–24 h SPR is generally considered as a more clinically meaningful measurement of the treatment than the 2 h PR. In this paper, we are mainly interested in the estimation of the proportion of patients having SPR for each treatment group and the difference between two groups. If there is no missing data, the 2–24 SPR variable is easily derived and the analysis is straight forward. But if any time point data or the recurrence measure is missing, which is common in real trials, the estimation and analysis become nontrivial.

Using the terminology of Little and Rubin [17]; data are missing completely at random (MCAR) if the missing probability does not depend on both observed and unobserved responses, and data are missing at random (MAR) if the missing probability only depends on the observed responses. These two missing mechanisms are called ignorable in the sense that the likelihood inference may ignore the missing mechanism. A nonignorable missing mechanism (NI) depends on unobserved responses. In our case, the

* Corresponding author. School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China.
E-mail address: ffang@sfs.ecnu.edu.cn (F. Fang).

http://dx.doi.org/10.1016/j.conctc.2016.06.011
2451-8654/© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

missing probability of the recurrence is considered to be related with the values of headache severities at time points which may be missed. So the missing mechanism is typically NI, which is a big challenge to the analysis.

A common approach when there is not a significant amount of missing data is to simply ignore the patients with missed responses, and focus on patients with completely observed data (i.e., complete-case, CC analysis). This simple approach is valid under MCAR. Generalized estimation equation (GEE) proposed by Liang and Zeger [16] is statistically consistent for MCAR longitudinal data. For MAR data, maximum likelihood based approaches [7,13,18]; among others), weighted generalized estimation equation (WGEE) approaches [3,4,20], a complicated imputation method [19], and an alternative multiple imputation approach [15] have been proposed. In general, these methods are biased under nonignorable missing data and some of them also require proper specification of the missing probability model. Much work has also been done to address the nonignorable missing data problems. Available methods include maximum likelihood [2,6,23,27], weighted estimation equations [21,22] and mixed effects models [1,5,8,11,24–26,28,29]. Recent reviews of methods handling nonignorable missing data in longitudinal data can be found in Hogan, Roy, and Korkontzlou [10] and Ibrahim and Molenberghs [12]. The common challenge with these approaches for nonignorable missing data is that they usually depend on certain models for the missing probability, e.g., logistic models, which can not be checked because unobserved data are involved. Their analysis may also be sensitive to the model assumptions on the missing probability.

The migraine data has a special feature in the sense that the patient will have a definitively negative 2–24 h SPR if one of the observed headache severities at 2, 3, 4, and 24 h postdose is moderate or severe. By taking advantage of this feature plus some reasonable assumptions on the missing mechanism, we propose two imputation based methods to deal with nonignorable missing data without model assumptions on the missing probability and the correlation structure among the longitudinal observations. These methods are easy to implement and very efficient as shown in the simulation results.

The rest of this paper is organized as follows. Section 2 introduces the notation and assumptions on the missing mechanism. Two imputation based estimation methods are proposed in Section 3. Results of a simulation study and a real data analysis are represented in Sections 4 and 5 respectively. Several concluding remarks are provided in Section 6. Some proofs are shown in an Appendix.

2. Notation and missing mechanism

For simplicity we start with a single treatment group. Let Y_{it} be the PR indicator of the ith patient at time point t, where $i = 1, \cdots, n$ and $t = 1, \cdots, T$, i.e., $Y_{it} = 1$ if the ith patient has PR at time point t; $Y_{it} = 0$, otherwise. Since the definition of 2–24 h SPR does not involve data prior to the 2 h postdose, for simplicity, in this paper we will ignore data before 2 h postdose and let Y_{i1} be the PR indicator at 2 h postdose, Y_{i2} be the PR indicator at 3 h postdose, etc. Let $\{Y_{it}, t = T+1, \cdots, T+L\}$ denote the additional binary responses other than PR and recurrence, e.g., 'no optional 2nd dose' indicator and 'no rescue medication' indicator. Let X_i be the 'no recurrence' indicator of the ith patient, i.e., $X_i = 1$ if the ith patient reports no headache recurrence; $X_i = 0$ if the ith patient reports recurrence. $X_i = 1$ if and only if the patient does not have moderate or severe headache at any time (including the specific time points $t = 1, \cdots, T$) during the 2–24 h post initial dose. Note that by definition X_i must be 0 if $\prod_{t=1}^{T} Y_{it} = 0$, i.e., the patient definitely has a recurrence if he/she doesn't have PR at any specific time point. Let $Y_i = X_i \prod_{t=T+1}^{T+L} Y_{it}$ be the SPR indicator. A patient has 2–24 h SPR if and only if $Y_i = 1$. Our main goal is to estimate $p = P(Y_i = 1)$ and calculate its confidence interval. The point estimation and variance estimation methods can be directly extended to two treatment groups for comparison purpose. Table 1 shows a hypothetical data set with $T = 4$, $L = 2$ and two treatment groups for illustration.

Let δ_{it} be the non-missing indicator for Y_{it}, i.e., $\delta_{it} = 1$ if Y_{it} is not missing; $\delta_{it} = 0$, otherwise. In the real trial, patients are asked to check a box for whether the 2nd dose or rescue medication was taken. If the boxes are not checked then the answers are recorded as 'no 2nd dose' and 'no rescue medication'. So we can assume that δ_{it} always equals 1 for $t = T+1, \cdots, T+L$, i.e., $\{Y_{it}, t = T+1, \cdots, T+L\}$ are always observed. Let δ_i be the non-missing indicator of X_i. It should be noted that δ_i still could be 0 even if all the $\delta_{it} = 1$ (e.g., the patients 6A, 2B and 10B in Table 1) and vice versa (e.g., the patients 1A and 5B in Table 1). Since $\{Y_{it}, t = T+1, \cdots, T+L\}$ are always observed and $Y_i = X_i \prod_{t=T+1}^{T+L} Y_{it}$, δ_i is also the non-missing indicator of Y_i. Denote $R_i = \{t: \delta_{it} = 1, t \le T+L\}$. A special feature of the data is that when $\prod_{t \in R_i} Y_{it} = 0$, Y_i must be 0 even if it is not directly observed (i.e., X_i is not reported and $\delta_i = 0$). In this case, we still consider Y_i as missing in the data set (e.g., the patients 6A, 14A and 10B in Table 1) but will impute it as 0 later. To address this special feature, we define an alternative non-missing indicator $\tilde{\delta}_i$ for Y_i as: if $\delta_i = 1$,then $\tilde{\delta}_i = 1$ (e.g., patient 1A in Table 1); if $\delta_i = 0$ and $\prod_{t \in R_i} Y_{it} = 0$, then Y_i must be 0 and we define $\tilde{\delta}_i = 1$ (e.g., patients 6A, 14A and 10B); if $\delta_i = 0$ and $\prod_{t \in R_i} Y_{it} = 1$, then we still don't know Y_i should be 0 or 1 and we define $\tilde{\delta}_i = 0$ (e.g., patients 4A, 10A and 2B).

We adopt the following two assumptions on the missing mechanism:

(a1) the missing probability of Y_{it} is independent of the response values and the miss-ing probability of Y_i, i.e., $\{\delta_{it}, t = 1, \cdots, T\}$ are independent of $\{Y_{it}, X_i, Y_i, \delta_i\}$, although $\{\delta_{it}, t = 1, \cdots, T\}$ may be correlated with each other, e.g., in a monotone missing mechanism. This assumes that whether the headache scores are missing at individual time points is not affected by the actual headache scores, recurrence response, 2–24 h SPR or whether recurrence is missing or not.

(a2) the missing probability of Y_i (or X_i) depends on $\{Y_{it}, X_i, Y_i, \delta_{it}, t = 1, \cdots, T+L\}$ only through the values of Y_{it}, $i = 1, \cdots, T+L$, i.e., $P(\delta_i = 1 | Y_{it}, X_i, Y_i, \delta_{it}, t = 1, \cdots, T+L) = P(\delta_i = 1 | Y_{it}, t = 1, \cdots, T+L)$. This assumes that only individual headache scores and whether 2nd dose or rescue medication is taken have direct affects on the recurrence missing. Since Y_{it} may not be observed, the missing mechanism of Y_i is nonignorable. To discuss the applicability of different methods in different situations, we consider the following three mechanisms under (a2):

M1. For all possible values of Y_{it}, the missing probabilities of Y_i are the same, i.e., Y_i is MCAR.

M2. $P(\delta_i = 1 \Big| \prod_{t=1}^{T+L} Y_{it} = 1) = P(\delta_i = 1 \Big| \prod_{t=1}^{T+L} Y_{it} = 0)$ but Y_i is not MCAR.

$$\text{M3.} P(\delta_i = 1 \Big| \prod_{t=1}^{T+L} Y_{it} = 1) \neq P(\delta_i = 1 \Big| \prod_{t=1}^{T+L} Y_{it} = 0)$$

Missing mechanisms M1 and M2 cannot be checked by the data since these assumptions involve unobserved data. So M3 is the most reliable assumption when we analyze the real data.

In practice, complete case analysis is commonly used. The CC method includes only those patients with observed SPR. The estimate of p is given by the observed proportion of SPR:

$$\hat{p}cc = \frac{\sum_{i=1}^{n} \delta_i Y_i}{\sum_{i=1}^{n} \delta_i}.$$

There are several different ways to estimate the variance of $\hat{p}cc$ and construct the 100 $(1-\alpha)$% confidence interval for p. Here we

Table 1
A hypothetical data set with $T = 4$ and $L = 2$ for illustration.

	Treatment group A									Treatment group B								
	Time points									Time points								
Patient	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	X	Y	Patient	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	X	Y
1A	1	.	.	1	1	0	1	0	1B	1	1	1	1	1	1	1	1
2A	1	1	1	1	0	0	0	0	2B	1	1	1	1	1	1	.	.
3A	1	1	1	1	0	0	1	0	3B	1	1	1	1	1	1	1	1
4A	1	.	.	.	1	1	.	.	4B	1	1	1	1	1	1	1	1
5A	1	1	1	1	1	1	1	1	5B	1	.	.	.	1	0	1	0
6A	0	1	1	1	0	0	.	.	6B	0	0	0	1	0	0	0	0
7A	0	1	1	1	1	1	0	0	7B	0	.	.	.	1	1	0	0
8A	0	0	0	.	0	0	0	0	8B	0	0	0	1	0	0	0	0
9A	0	0	0	1	0	0	0	0	9B	0	0	0	1	0	0	0	0
10A	1	.	.	.	1	1	.	.	10B	0	1	1	1	0	0	.	.
11A	0	0	0	1	0	0	0	0	11B	1	.	1	1	1	1	1	1
12A	1	1	1	1	1	1	1	1	12B	0	0	1	1	1	1	0	0
13A	0	0	0	1	0	0	0	0									
14A	0	0	.	1	0	0	.	.									

Note: Y_1, …, Y_4: PR indicator at time point t, Y_5: 'no 2nd dose' indicator, Y_6: 'no rescue medication' indicator, X: 'no recurrence' indicator, Y: SPR, '.': missing value.

use the Wald type as shown below. The confidence interval of p is calculate to be

$$\hat{p}cc \pm z_{\alpha/2}\sqrt{\frac{\hat{p}cc(1-\hat{p}cc)}{\sum_{i=1}^{n}\delta_i}}, \tag{1}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ percentile of the standard normal distribution. CC method is valid in MCAR, i.e., under M1. Moreover, CC method actually is consistent under M2. But it is not consistent under M3. See Appendix A.1 for a detailed discussion.

In migraine trials, an intuitive way to extend the CC method is to impute the Y_i by 0 if $\delta_i = 0$ and $\tilde{\delta}_i = 1$, and estimate p by sample mean:

$$\hat{p}_{\tilde{C}C} = \frac{\sum_{i=1}^{n}\tilde{\delta}_i\tilde{Y}_i}{\sum_{i=1}^{n}\tilde{\delta}_i},$$

where $\tilde{Y}_i = Y_i$ when $\delta_i = 1$; $\tilde{Y}_i = 0$ when $\delta_i = 0$ and $\tilde{\delta}_i = 1$; $\tilde{Y}_i = missing$, if otherwise. The confidence interval is given with analogous expression of Equation (1). We denote this method as $\tilde{CC}$. This is a commonly used approach in practice. But unfortunately, this estimator may have large bias unless the proportion of nonresponse is small. This is because only value of 0 can be imputed, so this estimator usually underestimates even under MCAR.

3. Proposed estimation methods based on imputation

3.1. Proposed imputation method 1 (IM_1)

An intuitive approach to impute the missing Y_i is to use the proportion of SPR in the patients that have non-missing SPR responses and similar $\{Y_t\}$ pattern with $\{Y_{it}, t \in R_i\}$. We denote this method as IM_1, which is described as follows.

Let $\tilde{Y}_i$ be the imputed value of Y_i. $\tilde{Y}_i = Y_i$ when $\delta_i = 1$. $\tilde{Y}_i = 0$ when $\delta_i = 0$ and $\tilde{\delta}_i = 1$. This is similar to method $\tilde{CC}$. But instead of ignoring the remaining patients with $\tilde{\delta}_i = 0$, we impute them by

$$\tilde{Y}_i = \frac{\sum_{j=1}^{n} I\{\delta_j Y_j = 1, \prod_{t\in R_i}\delta_{jt}Y_{jt} = 1\}}{\sum_{j=1}^{n} I\{\delta_j = 1, \prod_{t\in R_i}\delta_{jt}Y_{jt} = 1\}}. \tag{2}$$

Finally, we estimate p by $\sum_{i=1}^{n}\tilde{Y}_i/n$. This method has quite intuitive explanations. Table 2 further explains how the imputations are done using an illustrative example with some hypothetical data. In this example, we have a single treatment group with 7 patients. Similar to a real trial, we let $T = 4$ and $L = 2$. The 1st and 2nd patient have missing SPR.

Step 1 For the 1st patient, since he/she doesn't have PR at time point 2 ($Y_{12} = 0$), we know for sure he/she doesn't have SPR (i.e., $\tilde{\delta}_1 = 1$). Therefore we impute Y_1 by 0.

Step 2 For the 2nd patient, since all the non-missing $\{Y_{2t}, t = 1, \cdots, 6\}$ are 1 (i.e., $\tilde{\delta}_2 = 0$), he/she may or may not have SPR. Notice that $R_2 = \{1, 2, 4, 5, 6\}$. Take all the patients with non-missing SPR and have exactly the same $\{Y_t\}$ values as $\{Y_{2t}, t \in R_2\}$ and no matter $\{Y_t, t \in /R_2\}$ are missing or not, i.e., patients 3 to 7. Calculate the proportion of SPR of these patients, which is 2/5 in this case. Therefore we impute Y_2 by 2/5. Finally $\hat{p} = \sum_{i=1}^{n}\tilde{Y}_i/n = (2/5 + 1 + 1)/7 = 12/35$.

We can show that (2) is a consistent estimator of $P(Y = 1|\delta = 1, \prod_{t\in R}Y_t = 1)$. The basic idea of this method is to use $P(Y = 1|\delta = 1, \prod_{t\in R}Y_t = 1)$ to estimate $P(Y = 1|\delta = 0, \prod_{t\in R}\delta_t Y_t = 1, \delta_{t,t\notin R} = 0)$, which equals to $P(Y = 1|\delta = 0, \prod_{t\in R}Y_t = 1)$ when the assumption (a1) on the missing mechanism holds.

Theoretically, this imputation estimator is not consistent under M2 and M3. How ever, as discussed in AppendixA.2, pragmatically positive and negative biases of imputed responses with different patterns of R_i often offset much against each other. Hence the estimator performs reasonably well except in several extreme cases as shown in the simulation study.

3.2. Proposed imputation method 2 (IM_2)

The proposed imputation method 1 is not consistent theoretically. In this subsection, we propose a consistent estimation method. The key point is to find a consistent estimator for $P(Y = 1|\delta = 0, \prod_{t\in R}\delta_t Y_t = 1, \delta_{t,t\notin R} = 0)$, which equals $P(Y = 1|\ \delta = 0, \prod_{t\in R}Y_t = 1)$ und-er assumption (a1). The proposed imputation method 1 actually estimates $P(Y = 1|\delta = 0, \prod_{t\in R}Y_t = 1)$ by $P(Y = 1|\delta = 1, \prod_{t\in R}Y_t = 1) = P(Y = 1|\delta = 1, \prod_{t\in R}\delta_t Y_t = 1)$ which is usually in-consistent. We need to figure out an adjusting factor to make it consistent. By Bayes' formula, we have

Table 2
Illustration of proposed imputation method 1.

Patient	Time points								Step 1	Step 2
	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	Y_{i6}	X_i	Y_i	Impute Y_i with $\tilde{\delta}_i = 1$.	Impute Y_i with $\tilde{\delta}_i = 0$ by observed responses with similar $\{Y_t, t \in R_i\}$ pattern.
1	1	0	1	1	1	1	.	.		
2	1	1	.	1	1	1	.	.	Impute $\tilde{Y}_1 = 0$	$R_2 = \{1, 2, 4, 5, 6\}$
3	1	1	.	1	1	1	0	0	since $\tilde{\delta}_1 = 1$	
4	1	1	.	1	1	1	1	1		Take patients 3 to 7
5	1	1	0	1	1	1	0	0		Impute $Y_2 = \frac{2}{5}$
6	1	1	1	1	1	1	1	1		
7	1	1	1	1	1	1	0	0		$\hat{p} = \left(\frac{2}{5} + 1 + 1\right)/7 = \frac{12}{35}$

Note: $Y_{i1}, \ldots, Y_{i4}$: PR indicator at time point t, Y_{i5}: 'no 2nd dose' indicator, Y_{i6}: 'no rescue medication' indicator, X_i: 'no recurrence' indicator, Y_i: SPR, '.': missing value.

$$
\begin{aligned}
&P\left(Y=1|\delta=0, \prod_{t\in R} Y_t = 1\right) \\
&= \frac{P\left(\delta=0|Y=1, \prod_{t\in R} Y_t = 1\right) P\left(Y=1, \prod_{t\in R} Y_t = 1\right)}{P\left(\delta=0|\prod_{t\in R} Y_t = 1\right) P\left(\prod_{t\in R} Y_t = 1\right)} \\
&= \frac{\frac{P\left(\delta=0|Y=1, \prod_{t\in R} Y_t = 1\right)}{P(\delta=1|Y=1, \prod_{t\in R} Y_t = 1)} P\left(\delta=1, Y=1, \prod_{t\in R} Y_t = 1\right)}{\frac{P(\delta=0|\prod_{t\in R} Y_t = 1)}{P(\delta=1|\prod_{t\in R} Y_t = 1)} P\left(\delta=1, \prod_{t\in R} Y_t = 1\right)} \\
&= P\left(Y=1 \middle| \delta=1, \prod_{t\in R} Y_t = 1\right) \frac{O_1}{O_2},
\end{aligned}
\tag{3}
$$

where

$$
\begin{aligned}
O_1 &= \frac{P(\delta=0|Y=1, \prod_{t\in R} Y_t = 1)}{P(\delta=1|Y=1, \prod_{t\in R} Y_t = 1)} \text{ and } O_2 \\
&= \frac{P(\delta=0|\prod_{t\in R} Y_t = 1)}{P(\delta=1|\prod_{t\in R} Y_t = 1)}
\end{aligned}
\tag{4}
$$

are conditional odds of missing probability, and $\frac{O_1}{O_2}$ is a bias adjusting factor similar to Kim and Yu [14]. Since $\prod_{t\in R} Y_t$ and $\prod_{t=1}^{T+L} Y_t$ must be 1 when $Y = 1$,

$$
O_1 = \frac{P\left(\delta=0 \middle| Y=1, \prod_{t=1}^{T+L} Y_t = 1\right)}{P\left(\delta=1 \middle| Y=1, \prod_{t=1}^{T+L} Y_t = 1\right)} = \frac{P\left(\delta=0 \middle| \prod_{t=1}^{T+L} Y_t = 1\right)}{P\left(\delta=1 \middle| \prod_{t=1}^{T+L} Y_t = 1\right)},
\tag{5}
$$

where the second equation holds because of assumption (a2). Equations (3)–(5) and assumption (a1) together give us

$$
\begin{aligned}
P\left(Y=1 \middle| \delta=0, \prod_{t\in R} \delta_t Y_t = 1, \delta_{t, t\notin R} = 0\right) &= P\left(Y=1 \middle| \delta=1, \prod_{t\in R} \delta_t Y_t = 1\right) \frac{P\left(\delta=0 \middle| \prod_{t=1}^{T+L} \delta_t Y_t = 1\right)}{P\left(\delta=1 \middle| \prod_{t=1}^{T+L} \delta_t Y_t = 1\right)} \Big/ \frac{P(\delta=0|\prod_{t\in R} \delta_t Y_t = 1)}{P(\delta=1|\prod_{t\in R} \delta_t Y_t = 1)} \\
&= \frac{P(\delta=1, Y=1, \prod_{t\in R} \delta_t Y_t = 1) P\left(\delta=0, \prod_{t=1}^{T+L} \delta_t Y_t = 1\right)}{P\left(\delta=1, \prod_{t=1}^{T+L} \delta_t Y_t = 1\right) P(\delta=0, \prod_{t\in R} \delta_t Y_t = 1)}.
\end{aligned}
\tag{6}
$$

So we can impute Yi with $\tilde{\delta}_i = 0$ by

$$
\tilde{Y}_i = \frac{\sum_{j=1}^{n} I\left\{\delta_j Y_j = 1, \prod_{t\in R_i} \delta_{jt} Y_{jt} = 1\right\} \sum_{j=1}^{n} I\left\{\delta_j = 0, \prod_{t=1}^{T+L} \delta_{jt} Y_{jt} = 1\right\}}{\sum_{j=1}^{n} I\left\{\delta_j = 1, \prod_{t=1}^{T+L} \delta_{jt} Y_{jt} = 1\right\} \sum_{j=1}^{n} I\left\{\delta_j = 0, \prod_{t\in R_i} \delta_{jt} Y_{jt} = 1\right\}}
\tag{7}
$$

to obtain a consistent estimator of p. This method is consistent under M1, M2 and M3.

As for the efficiency, if the denominator of (6) is too small, the variance of the estimator may be large. We do not need to worry too much about $P(\delta = 0, \prod_{t\in R} \delta_t Y_t = 1)$ since its small value also means only few Y_i needs to be imputed by the formula. So the main concern is small $P(\delta = 1, \prod_{t=1}^{T+L} \delta_t Y_t = 1)$. Based on our simulation experience, as long as $P(\delta = 1, \prod_{t=1}^{T+L} \delta_t Y_t = 1)$ is greater than 5%, the final estimator of p will be quite efficient. It means that we only need to have more than 5% of the patients to have PR at all time points, do not take the second dose, do not take the rescue medicine, and have no missing data at all, which is common in real trials, to make sure the estimator is efficient.

3.3. Variance estimation and confidence interval

Since the proposed imputation estimators above do not have explicit variance expression, we apply the bootstrap method to estimate their variances, which is conducted in the following steps.

(1) From each data set, draw a simple random sample of size n with replacement from the set of patients (respondents or nonrespondents). For each patient in the bootstrap sample, the bootstrap data consist of the Y_t and Y values. If the Y_t or Y is missing, the bootstrap datum is also treated as missing.
(2) Apply the same imputation methods as we described in the previous two subsections to the bootstrap sample generated in step (1), and get the estimator $\hat{p}^*$.
(3) Repeat the previous steps in dependently for B times and obtain $\hat{p}^{*1}, \ldots, \hat{p}^{*B}$. Esti-mate the variance of $\hat{p}$ by the sample variance of $\hat{p}^{*1}, \ldots, \hat{p}^{*B}$.

Table 3
The PR rates and SPR rate in the simulation.

Time points								No 2nd dose	No rescue medication	SPR
0	0.5	1	1.5	2	3	4	24			
0	0.23	0.42	0.54	0.61	0.64	0.66	0.66	0.59	0.57	0.348

The 100(1 − α)% confidence interval is given by

$$\text{point estimate} \pm z_{\alpha/2}\sqrt{\textit{bootstrap variance estimator}},$$

where $z_{\alpha/2}$ is the 100(1−α/2)% percentile of the standard normal distribution.

4. A simulation study

We conducted a simulation study to compare the empirical performances of the *CC*, $\widetilde{CC}$ and the proposed methods in a wide range of scenarios. We focused on a single treatment group scenario in this Section. In Section 5.2 we will show some simulation results for two treatment groups. We simulated data sets with similar pattern to the data in the motivating trial. Let X_t, $t = 1, \ldots, 8$ denote the headache severity at baseline and 0.5, 1, 1.5, 2, 3, 4, 24 h postdose, respectively. We generate X_t as follows: $X_1 = 2$ with probability 2/3, $X_1 = 3$ with probability 1/3. X_{t+1} is generated from a multinomial distribution conditional on the value of X_t. The conditional probabilities of $P(X_{t+1}|X_t)$ are similar to what we observed from the motivating trial. Then we define $Y_1 = I\{X_5 < 2\}$, $Y_2 = I\{X_6<2\}$, $Y_3 = I\{X_7<2\}$, and $Y_4 = I\{X_8<2\}$. 'No recurrence' indicator $X = 0$ if $\prod_{t=1}^{4} Y_t = 0$; $X = 1$ with probability 0.95 if $\prod_{t=1}^{4} Y_t = 1$ and $X_5 = 0$; $X = 1$ with probability 0.91 if $\prod_{t=1}^{4} Y_t = 1$ and $X_5 = 1$. 'No 2nd dose' indicator Y_5 and 'no rescue medication' indicator Y_6 are generated from binomial distributions conditional on X_5. Values of P(no 2nd dose|X_5) and P(no rescue medication|X_5) are also similar to the motivating trial. Then SPR indicator $Y = XY_5Y_6$. We consider the scenario with the following PR rate profile: (see Table 3).

The non-missing probabilities are determined by the following parameters. $r_t = P(\delta_t = 1)$, $t = 1, \ldots, 6$, $r = P(\delta = 1 \mid \prod_{t=1}^{6} Y_t = 1)$, $s = P(\delta = 1 \mid \prod_{t=1}^{6} Y_t = 0)$. Notice that $r_5 = r_6 = 1$ and s is determined by $s_1 = P(\delta = 1|Y_1Y_2Y_3 = 1, Y_4 = 0)$, $s_2 = P(\delta = 1|Y_1Y_2Y_3 = 0, Y_4 = 1)$, $s_3 = P(\delta = 1|Y_1Y_2Y_3 = 0, Y_4 = 0)$ and $s_4 = P(\delta = 1|Y_1Y_2Y_3Y_4 = 1, Y_5Y_6 = 0)$. We con-duct simulations under all the three missing mechanisms M1, M2, and M3.

M1: $r = s = s_1 = s_2 = s_3 = s_4$.

M2: $r = s$, but s_1, s_2, s_3 and s_4 are not all the same.

M3: r ≠s.

For each missing mechanism, we consider several cases with different non-missing parameters which are given in Table 4. In the simulation study, we set the missing probabilities much higher than which in the reality in order to distinguish the performances of different methods.

For each missing mechanism and each case, we run the simulation 1000 times. The sample size $n = 400$. The bootstrap round B is 200. The relative bias, mean square error (MSE), and the coverage probability (CP) of the 95% confidence interval are reported in Table 5.

The simulation results in Table 5 can be summarized as follows:

(1) The *CC* method performs well under M1 and M2 in terms of relative bias and MSE. But it has large bias under M3. The $\widetilde{CC}$ method only works well when the missing probability is small (e.g., Case 1 of M1). Otherwise it underestimates noticeably as expected.
(2) The proposed imputation method 1 performs well in most of the cases except Case 5/6 of M2 and M3. In these 4 cases, we set $r_1 = r_2 = r_3 = 1$ and $r_4 = 0.3$, thus the only possible R_i is {1, 2, 3, 5, 6} or {1, 2, 3, 4, 5, 6}. As discussed in Appendix A.2, when the pattern of R_i is not diversified, the proposed method 1 may have large bias, which is verified by the simulation results.
(3) The proposed imputation method 2 performs best in terms of the robustness. The only concern of this method is when $P(\delta = 1, \prod_{t=1}^{T+L}\delta_t Y_t = 1)$ is extremely small. For example, in Case 6 of M3, $P(\delta = 1, \prod_{t=1}^{T+L}\delta_t Y_t = 1) = 0.028$. The relative bias and MSE are relatively large compared to other cases. But it is still better than all the other methods. Since in reality the probability $P(\delta = 1, \prod_{t=1}^{T+L}\delta_t Y_t = 1)$ usually is not that small, the proposed imputation method 2 is an ideal method to deal with missing data problem in migraine trials and in other similar situations.

Table 4
The non-missing parameters in each missing mechanism and different cases.

	Case	r_1	r_2	r_3	r_4	r	s	s_1	s_2	s_3	s_4	$P(\delta = 1)$	$P(\tilde{\delta} = 1)$	$P(\delta = 1, \prod \delta_t Y_t = 1)$
M1	1	0.9	0.9	0.9	0.9	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.24
	2	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.96	0.22
	3	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.92	0.19
	4	0.9	0.9	0.9	0.9	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.74	0.07
M2	1	0.9	0.9	0.9	0.9	0.26	0.26	1	1	0	0	0.26	0.72	0.06
	2	0.9	0.9	0.9	0.9	0.74	0.74	0	0	1	1	0.74	0.90	0.18
	3	0.9	0.9	0.9	0.9	0.53	0.53	1	0	1	0	0.53	0.82	0.13
	4	0.9	0.9	0.9	0.9	0.47	0.47	0	1	0	1	0.47	0.80	0.11
	5	1	1	1	0.3	0.53	0.53	1	0	1	0	0.53	0.82	0.06
	6	1	1	1	0.3	0.74	0.74	0	0	1	1	0.74	0.89	0.08
M3	1	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.8	0.84	0.96	0.22
	2	0.9	0.9	0.9	0.9	0.9	0.5	0.5	0.5	0.5	0.5	0.65	0.96	0.22
	3	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.86	0.92	0.19
	4	0.9	0.9	0.9	0.9	0.5	0.9	0.9	0.9	0.9	0.9	0.74	0.81	0.12
	5	1	1	1	0.3	0.3	0.9	0.9	0.9	0.9	0.9	0.67	0.73	0.034
	6	1	1	1	0.3	0.25	0.9	0.9	0.9	0.9	0.9	0.65	0.72	0.028

Table 5
The simulation results for one treatment group.

Case		Relative bias%				MSE*1000				CP%			
		CC	$\tilde{CC}$	IM_1	IM_2	*CC*	$\tilde{CC}$	IM_1	IM_2	*CC*	$\tilde{CC}$	IM_1	IM_2
M1	1	−0.0	−1.3	0.0	−0.2	0.570	0.585	0.557	0.561	95.5	94.5	94.6	94.5
	2	−0.8	−7.4	−1.0	−1.4	0.581	1.179	0.538	0.546	94.5	83.0	94.5	95.3
	3	−1.1	−14.6	−0.8	−1.1	0.738	3.178	0.566	0.577	93.5	41.0	94.5	95.0
	4	0.3	−59.3	0.3	0.1	1.970	43.050	0.781	0.812	95.5	0.0	93.0	93.5
M2	1	0.9	−64.3	−1.5	−0.3	2.507	50.358	0.806	0.786	93.5	0.0	94.5	94.0
	2	−0.1	−18.2	0.4	−0.5	0.748	4.561	0.554	0.558	95.5	25.5	94.8	96.4
	3	0.3	−35.9	−0.5	−0.6	1.114	16.106	0.674	0.681	95.5	0.0	95.0	95.0
	4	−0.7	−41.4	0.5	−0.1	1.144	21.995	0.574	0.594	96.0	0.0	95.0	94.5
	5	0.5	−35.7	−2.3	0.6	1.082	15.928	0.601	1.446	96.0	0.0	96.5	95.5
	6	−0.1	−17.3	3.1	0.0	0.778	4.233	0.726	0.967	94.5	33.0	90.3	94.0
M3	1	7.5	−6.5	0.3	−0.0	1.473	1.175	0.662	0.660	77.5	79.0	93.5	93.5
	2	39.7	−6.1	0.7	0.1	19.874	0.931	0.503	0.497	0.0	88.0	96.0	96.5
	3	−7.8	−14.0	−0.4	−0.6	1.309	2.895	0.547	0.551	81.5	44.5	95.1	96.0
	4	−34.1	−39.3	−0.6	−0.4	14.696	19.206	0.653	0.655	1.0	0.0	96.0	95.4
	5	−55.9	−59.5	−5.8	0.4	38.336	43.333	1.242	3.453	0.0	0.0	87.5	95.5
	6	−61.7	−65.0	−6.9	3.6	46.524	51.176	1.434	5.420	0.0	0.0	86.2	95.1

(4) As for the efficiency, the two proposed imputation methods have comparable MSEs, and they both outperform the *CC* and $\tilde{CC}$ methods.
(5) The bootstrap variance estimators work well for both imputation methods. The coverage probabilities of the confidence intervals are all around 95% except in the cases when the proposed imputation method 1 has relatively large bias in the extreme cases.

5. A real data analysis

5.1. Data analysis results

In this section we illustrate the application of our proposed methods in a real migraine trial data. The trial is a phase 3 confirmatory study to test the efficacy and safety of a calcitonin-gene related peptide (CGRP) antagonist. There are 3 active doses plus a placebo control arm in the study. For simplicity and demonstration purpose, we only choose the one active dose and placebo here. The sample sizes of the two groups are 333 and 348 respectively. The completer case (*CC*) proportions, i.e., $P(\delta = 1)$, are 82.9% and 75.6% respectively. Using $\tilde{CC}$ method the non-missing rates of SPR, i.e.,$P(\tilde{\delta} = 1)$, are 96.1% and 98.6% in the two groups.

The analysis results based on *CC*, $\tilde{CC}$ and the two proposed imputation methods (IM_1 and IM_2) are reported in Table 6. As expected and also demonstrated in simulation (case 1 from Table 9 in Section 5.2), the *CC* approach overestimates the individual treatment effect and the $\tilde{CC}$ method underestimates. Both IM_1 and IM_2 make adjustment to those estimates in the correct directions. For the treatment effect $(p_A - p_B)$ estimate, both *CC* and $\tilde{CC}$ approaches underestimate the treatment difference while IM_1 and IM_2 also appear to have made adjustment in the correct direction. Note that because the $\tilde{CC}$ missing data proportion is low, the adjustment is thus also minor. But potentially the adjustment may be more significant when the missing data rate is high (as seen in the case 2 from Table 9).

5.2. Simulation based on the real data

We conducted an additional simulation based on the real data to evaluate the empirical performances of our proposed methods for two treatment groups. For each treatment group (sample size of 333 and 348 respectively), we generate X_t, Y_t, X and Y in a similar way to the simulation study in Section 4. The conditional probabilities needed for data generation are calculated from the real data with all available samples. The PR rate profile for the two treatment groups is given in Table 7. Due to the effects of missing data when calculating the conditional probabilities for data generation, the calculated PR rates and SPR rates in the simulation are different from what have been observed in the real trial.

We consider two cases for the non-missing probabilities. In the first case, the missing probability parameters r_t, r and s_t are calculated from the real trial based on all available samples. In the second case, we adjust the parameters to make the missing probabilities a little bit larger in order to distinguish the

Table 6
Analysis results of the real migraine trial data using different methods.

Statistics	Method of analysis			
	CC	$\tilde{CC}$	IM_1	IM_2
$\hat{p}A$	0.3406	0.2938	0.3175	0.3061
se($\hat{p}A$)	0.0285	0.0255	0.0263	0.0262
95% CI	(0.2846,0.3965)	(0.2438,0.3436)	(0.2658,0.3691)	(0.2546,0.3575)
$\hat{p}B$	0.2053	0.1574	0.1667	0.1597
se($\hat{p}B$)	0.0249	0.0197	0.0205	0.0202
95% CI	(0.1565,0.2541)	(0.1189,0.1960)	(0.1265,0.2069)	(0.1202,0.1992)
$\hat{p}A-\hat{p}B$	0.1353	0.1363	0.1508	0.1464
se($\hat{p}A-\hat{p}B$)	0.0379	0.0322	0.0334	0.0331
95% CI	(0.0610,0.2095)	(0.07326,0.1994)	(0.0853,0.2162)	(0.0815,0.2112)

Table 7
The PR rates and SPR rates in the simulation.

	Time points								No 2nd dose	No rescue med.	SPR
	0	0.5	1	1.5	2	3	4	24			
Trt A	0	0.11	0.24	0.42	0.49	0.62	0.73	0.85	0.75	0.39	0.221
Trt B	0	0.11	0.22	0.24	0.28	0.44	0.48	0.87	0.64	0.23	0.124

Table 8
The non-missing parameters in each case and each treatment group.

Case	Trt	r_1	r_2	r_3	r_4	r	s	s_1	s_2	s_3	s_4	$P(\delta=1)$	$P(\tilde{\delta}=1)$	$P(\delta=1,\prod\delta_t Y_t=1)$
1	A	0.96	0.85	0.84	0.92	0.94	0.83	1	0.77	0.78	0.96	0.86	0.98	0.15
	B	0.97	0.86	0.85	0.9	0.97	0.73	1	0.68	0.74	1	0.76	0.99	0.09
2	A	0.96	0.85	0.84	0.92	0.8	0.53	1	0.5	0.5	0.5	0.60	0.95	0.13
	B	0.97	0.86	0.85	0.9	0.8	0.51	1	0.5	0.5	0.5	0.55	0.97	0.07

performances of different methods. The parameters in the two cases are listed in Table 8.

For each case, we run the simulation 1000 times. The bootstrap round B is 200. The relative bias (RB), standard deviation (SD), mean square error (MSE), standard error (the estimated standard deviation, SE), and the coverage probability (CP) of the 95% confidence interval are reported in Table 9.

The simulation results are summarized as follows. First, the CC method has large bias as expected since the two cases considered here are both M3. The $\tilde{CC}$ method underestimates especially when the missing probability is relatively large in Case 2. Second, our proposed methods IM_1 and IM_2 work quite well in terms of negligible relative biases and comparable or smaller mean square errors compared with CC and $\tilde{CC}$. Third, the bootstrap method produces nearly unbiased estimator for standard deviation. The coverage probabilities of the confidence intervals based on our proposed methods and bootstrap variance estimators are close to 95%. These results are also consistent with the observations from the M3 cases in the previous more general simulation study.

6. Concluding remarks

We proposed two imputation based estimation methods to deal with the nonignorable missing data in migraine trials with longitudinal binary responses by leveraging the special data feature of the sustained response. We illustrated the application of our proposed methods by analyzing data from a real migraine clinical trial, and compared their performances to the complete-case method (CC) and the current method used in real trials ($\tilde{CC}$) in comprehensive simulation studies. The $\tilde{CC}$ method has large bias unless the missing probability is small. The CC method is consistent only in some special missing mechanisms. The proposed methods generally perform very well even in nonignorable missingness except in several extreme cases which are unlikely to happen in real trials. Also they are more efficient than the CC method in terms of smaller SMEs. The proposed methods do not need any specific model assumptions on the missing probabilities (e.g., logistic models) or the correlation structure among the longitudinal observations. They are direct estimation methods in the sense that the nuisance longitudinal missing data do not need to be estimated first.

The proposed imputation methods can be easily extended when stratification is needed. For example, we could split the patients into several strata by their headache severities at baseline and conduct the imputation within each stratum. The overall estimation will be a weighted average.

The bootstrap was applied to obtain the variance estimation and to conduct confidence interval. It worked quite well in the simulation study. However, we also realize that it may have some difficulty when $n^*P(\delta=1,\prod_{t=1}^{T+L}\delta_t Y_t=1)$ is very small. In this situation, when we draw a bootstrap sample from the original data,

Table 9
The simulation results for two treatment groups.

Method	Case 1					Case 2				
	RB%	SD	MSE*1000	SE	CP%	RB%	SD	MSE*1000	SE	CP%
Treatment A										
CC	9.3	0.0247	1.039	0.0254	88.6	33.8	0.0316	6.576	0.0322	37.0
$\tilde{CC}$	−5.1	0.0220	0.612	0.0225	91.0	−15.0	0.0214	1.567	0.0220	65.0
IM_1	−0.5	0.0226	0.512	0.0229	95.2	0.6	0.0227	0.518	0.0234	96.0
IM_2	−1.0	0.0226	0.515	0.0228	95.0	−0.1	0.0227	0.517	0.0235	96.1
Treatment B										
CC	27.1	0.0222	1.630	0.0224	70.8	44.6	0.0261	3.752	0.0277	49.2
$\tilde{CC}$	−2.6	0.0174	0.312	0.0175	92.9	−18.2	0.0152	0.744	0.0164	70.2
IM_1	0.5	0.0177	0.315	0.0178	95.0	0.0	0.0167	0.279	0.0180	95.8
IM_2	−0.5	0.0178	0.315	0.0177	94.5	−1.2	0.0169	0.287	0.0180	95.4
Treatment A - Treatment B										
CC	−13.5	0.0328	1.248	0.0338	94.4	19.9	0.0404	2.005	0.0425	93.8
$\tilde{CC}$	−8.3	0.0280	0.847	0.0285	94.4	−11.0	0.0262	0.799	0.0274	94.2
IM_1	−1.7	0.0290	0.843	0.0289	94.6	1.2	0.0285	0.812	0.0295	95.8
IM_2	−1.7	0.0289	0.837	0.0289	94.4	1.3	0.0286	0.817	0.0296	96.0

Note: RB: relative bias, SD: standard deviation, MSE: mean square error, SE: standard error (estimated standard deviation), CP: coverage probability.

it may happen that $\sum_{i^*=1}^{n} I\{\delta_{i^*}=1, \prod_{t=1}^{T+L}\delta_{i^*t}Y_{i^*t}=1\}=0$. Then our proposed methods are not applicable in the bootstrap sample. So the variance estimation in small sample needs to be further addressed. If we allow the modeling on the missing probability and take the correlation structure among the binary responses into account, the comparison among our methods and some other methods such as GEE, maximum likelihood and weighted GEE is an interesting research topic, although we conjecture that our proposed methods should be the most robust approaches since they are free of many model assumptions. How to address all these issues will remain as our future research topics.

Acknowledgements

The authors would like to thank the editor and two anonymous reviewers for their valuable comments and suggestions which improved the presentation of this paper. Fang Fang's research was partially supported by Shanghai Nature Science Foundation 15ZR1410300, Shanghai Rising Star Program (16QA1401700), Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019).

Appendix A

A.1. Discussion of the consistency of CC method

The $p\hat{}_{CC}$ is a consistent estimator of $P(Y=1|\delta=1)$. When assumption (a2) holds, we have

$$P(Y=1|\delta=1)=\frac{P\left(X=1,\prod_{t=1}^{T+L}Y_t=1,\delta=1\right)}{P(\delta=1)}$$

$$=\frac{P\left(X=1,\delta=1\middle|\prod_{t=1}^{T+L}Y_t=1\right)P\left(\prod_{t=1}^{T+L}Y_t=1\right)}{P(\delta=1)}$$

$$=\frac{P\left(\delta=1\middle|\prod_{t=1}^{T+L}Y_t=1\right)P\left(X=1\middle|\prod_{t=1}^{T+L}Y_t=1\right)P\left(\prod_{t=1}^{T+L}Y_t=1\right)}{P(\delta=1)}$$

$$=\frac{P\left(\delta=1\middle|\prod_{t=1}^{T+L}Y_t=1\right)}{P(\delta=1)}P(Y=1).$$

Hence $p\hat{}_{CC}$ is consistent for $P(Y=1)$ if and only if $P(Y=1|\delta=1,\prod_{t\in R}Y_t=1)=P(\delta=1)$, which is equivalent to M2.

A.2. Discussion of the consistency of proposed imputation method 1

The basic idea of proposed method 1 is using $P(Y=1|\delta=1,\prod_{t\in R}Y_t=1)$ to estimate $P(Y=1|\delta=0,\prod_{t\in R}\delta_t Y_t=1,\delta_{t,t\notin R}=0)$, which equals to $P(Y=1|\delta=0,\prod_{t\in R}Y_t=1)$ when the assumption (a1) on the missing mechanism holds, where R could be any fixed subs-et of $\{1,\ldots,T+L\}$. When the assumption (a2) also holds, these two probabilities are equal if and only if $A=B$, where

$$A=\frac{P\left(\delta=1\middle|\prod_{t=1}^{T+L}Y_t=1\right)}{P(\delta=1|\prod_{t\in R}Y_t=1)}\quad\text{and}\quad B=\frac{P\left(\delta=0\middle|\prod_{t=1}^{T+L}Y_t=1\right)}{P(\delta=0|\prod_{t\in R}Y_t=1)}.$$

Under M1, $A=B=1$, so proposed method 1 is consistent. But when M1 doesn't hold, A and B are not necessarily equal. Here is a simple example:

For simplicity, take $T=2, L=0$. Then R could be $\{1,2\}$, $\{1\}$, $\{2\}$, or $\{\phi\}$. We assume Y_1 and Y_2 are independent. Denote $p_t=P(Y_t=1)$, $r_t=P(\delta_t=1)$, $r=P(\delta=1|Y_1Y_2=1)$, $s=P(\delta=1|Y_1Y_2=0)$, $s_1=P(\delta=1|Y_1=1,Y_2=0)$, $s_2=P(\delta=1|Y_1=0,Y_2=1)$, $s_3=P(\delta=1|Y_1=0,Y_2=0)$. Then s, s_1, s_2, s_3 should satisfy

$$s=\frac{s_1p_1(1-p_2)+s_2p_2(1-p_1)+s_3(1-p_1)(1-p_2)}{1-p_1p_2}$$

When $R=\{1,2\}$, $A=B=1$, then there is no problem with the imputation method. When $R=\{\phi\}$, $A=B$ if and only if M2 holds. The A and B values in the other two cases are listed in the following table. The last column lists the probability of the corresponding R will occur when we conduct the imputation.

R	A(imputed)	B(true)	$P(R,\delta\delta=1)$
{1}	$\frac{r}{s_1(1-p_2)+rp_2}$	$\frac{1-r}{1-s_1(1-p_2)+rp_2}$	$r_1(1-r_2)p_1[1-s_1(1-p_2)-rp_2]$
{2}	$\frac{r}{s_2(1-p_1)+rp_1}$	$\frac{1-r}{1-s_2(1-p_1)+rp_1}$	$r_2(1-r_1)p_2[1-s_2(1-p_1)-rp_1]$

As we can see from the table, the imputed value and the true value are not always the same. To be more specific, let $p_1=0.8$, $p_2=0.5$, $r=s=0.3$, $s_1=0.4$, $s_2=0$, $s_3=0.2$. Then when $R=\{1\}$,$A=0.857$,$B=1.076$, so the imputed value is smaller than the true value. When $R=\{2\}$, $A=1.25$, $B=0.92$, so the imputed value is larger than the true value. If we take $r_1=1$, $r_2=0$, then only possible $R=\{1\}$. Then the imputed value will always be smaller than the true value. The imputation method underestimates. On the other hand, if we take $r_1=0$, $r_2=1$ then the imputed value will always be larger than the true value. The imputation method overestimates. Notice that in this example, M2 holds. So it illustrates that even in M2, this imputation method may have bias.

But we should notice that the imputation overestimates missing values for some R, and underestimates for some other R. The biases may be cancelled each other if the pattern of R is diversified. So overall speaking, the bias of this imputation method may not be a problem. Actually this is verified in the simulation study in Section 4.

References

[1] P.S. Albert, D.A. Follmann, Modeling repeated count data subject to informative dropout, Biometrics 56 (2000) 667–677.
[2] S.G. Baker, Marginal regression for repeated binary data with outcome subject to non-ignorable non-response, Biometrics 51 (1995) 1042–1052.
[3] B. Chen, G.Y. Yi, R.J. Cook, Weighted generalized estimating functions for longitudinal response and covariate data which are missing at random, J. Am. Stat. Assoc. 105 (2010) 336–353.
[4] B. Chen, X. Zhou, Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates, Biometrics 67 (2011) 830–842.
[5] M.R. Conaway, The analysis of repeated categorical measurements subject to nonig- norable nonresponse, J. Am. Stat. Assoc. 87 (1992) 817–824.
[6] G.M. Fitzmaurice, N.M. Laird, G.E.P. Zahner, Multivariate logistic models for incomplete binary responses, J. Am. Stat. Assoc. 91 (1996) 99–108.
[7] G.M. Fitzmaurice, G. Molenberghs, S.R. Lipsitz, Regression models for longitudinal binary responses with informative drop-outs, J. R. Stat. Soc. Ser. B 57 (1995) 691–704.
[8] D. Follmann, M. Wu, An approximate generalized linear model with random effects for informative missing data, Biometrics 51 (1995) 151–168.
[9] T.W. Ho, L.K. Mannix, X. Fan, et al., Randomized controlled trial of an oral CGRP receptor antagonist,MK-0974,inacutetreatmentofmigraine, Neurology

70 (16) (2008) 1304–1312.
[10] J.W. Hogan, J. Roy, C. Korkontzelou, Handling dropout in longitudinal studies, Stat. Med. 23 (2004) 1455–1497.
[11] J.G. Ibrahim, M. Chen, S.R. Lipsitz, Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable, Biometrika 88 (2001) 551–564.
[12] J.G. Ibrahim, G. Molenberghs, Missing data methods in longitudinal studies: a review, TEST 18 (2009) 1–43.
[13] R.I. Jennrich, M.D. Schluchter, Unbalanced repeated-measures models with structured covariance matrices, Biometrics 42 (1986) 805–820.
[14] J.K. Kim, C.L. Yu, A semiparametric estimation of mean functionals with nonignorable missing data, J. Am. Stat. Assoc. 106 (2011) 157–165.
[15] X. Li, D. Mehrotra, J. Barnard, Analysis of incomplete longitudinal binary data using multiple imputation, Stat. Med. 25 (2006) 2107–2124.
[16] K. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[17] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data (2nd Edition), Wiley, New York, 2002.
[18] G. Molenberghs, M.G. Kenward, E. Lesaffre, The analysis of longitudinal ordinal data with nonrandom drop-out, Biometrika 84 (1997) 33–44.
[19] M.C. Paik, The generalized estimating equation approach when data are not missing completely at random, J. Am. Stat. Assoc. 92 (1997) 1320–1329.
[20] J.M. Robins, A. Rotnitzky, L.P. Zhao, Analysis of semiparametric regression models for repeated outcome in the presence of missing data, J. Am. Stat. Assoc. 90 (1995) 106–121.
[21] A. Rotnitzky, J.M. Robins, D.O. Scharfstein, Semiparametric regression for repeated outcomes with nonignorable nonresponse, J. Am. Stat. Assoc. 93 (1998) 1321–1339.
[22] D.O. Scharfstein, A. Rotnitzky, J.M. Robins, Adjusting for nonignorable drop-out using semiparametric nonresponse models, J. Am. Stat. Assoc. 94 (1999) 1096–1120.
[23] S.K. Sinha, Robust analysis of longitudinal data with nonignorable missing responses, Metrika 75 (2012) 913–938.
[24] A.L. Stubbendick, J.G. Ibrahim, Maximum likelihood methods for nonignorable missing responses and covariates in random effects models, Biometrics 59 (2003) 1140–1150.
[25] A.L. Stubbendick, J.G. Ibrahim, Likelihood based inference with nonignorable missing responses and covariates in models for discrete longitudinal data, Stat. Sin. 16 (2006) 1143–1167.
[26] T.R. Ten Have, A.R. Kunselman, E.P. Pulkstenis, et al., Mixed effects logistic regression models for longitudinal binary response data with informative drop-out, Biometrics 54 (1998) 367–383.
[27] A.B. Troxel, S.R. Lipsitz, D.P. Harrington, Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data, Biometrika 85 (1998) 661–672.
[28] M.C. Wu, R.J. Carrol, Estimating and comparison of changes in the presence of informative right censoring by modeling the censoring process, Biometrics 44 (1988) 175–188.
[29] M. Yuan, R.J.A. Little, Mixed-effect hybrid models for longitudinal data with nonig- norable dropout, Biometrics 65 (2009) 478–486.