Ann Inst Stat Math (2019) 71:837–852
https://doi.org/10.1007/s10463-018-0667-8

Bias reduction using surrogate endpoints as auxiliary variables

Yoshiharu Takagi[1] · **Yutaka Kano**[2]

Received: 31 May 2017 / Revised: 15 March 2018 / Published online: 17 May 2018

Abstract Recently, it is becoming more active to apply appropriate statistical methods dealing with missing data in clinical trials. Under not missing at random missingness, MLE based on direct-likelihood, or observed likelihood, possibly has a serious bias. A solution to the bias problem is to add auxiliary variables such as surrogate endpoints to the model for the purpose of reducing the bias. We theoretically studied the impact of an auxiliary variable on MLE and evaluated the bias reduction or inflation in the case of several typical correlation structures.

Keywords Auxiliary variables · Surrogate endpoints · Direct-likelihood · Not missing at random missingness data

1 Introduction

In clinical trials, it often happens that the endpoints cannot be measured or are missing, due to the subjects' discontinuation from the study (e.g., dropout). It has already been realized that "Missing values represent a potential source of bias in a clinical trial" (International Conference on Harmonisation Guideline E9 1999). Recently, it is becoming more active to apply appropriate statistical methods dealing with missing

✉ Yoshiharu Takagi
Yoshiharu.Takagi@sanofi.com

Yutaka Kano
kano@sigmath.es.osaka-u.ac.jp

[1] Biostatistics and Programming, Sanofi K.K., Nishi-Shinjuku, Tokyo 163-1488, Japan

[2] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan

data, following the recommendations by O'Neill and Temple (2012) and National Research Council (NRC) report (2010).

Missing data are typically classified into three groups according to their causes, which are called missing-data mechanisms. Missing data are referred to as missing completely at random (MCAR) if missingness is independent of all the variables to be collected in the study; missing at random (MAR) if missingness can depend on the components of observed variables only and not on the value of missing variables; and not missing at random (NMAR or MNAR) if missingness is neither MCAR nor MAR (Rubin 1976; Little and Rubin 2002).

When a missing-data mechanism is MCAR, complete-case analysis, omitting all units involving missing values from the data set, will provide consistent estimators for parameters in the model, but the collected information is not fully utilized. In the case of MAR, the direct-likelihood based on observed data, or observed likelihood, will provide consistent estimators (Takai and Kano 2013). However, MCAR and MAR are often not realistic and cannot be expected to hold routinely, because the missingness due to subjects' study discontinuation is possibly affected by the post-study treatment such as lack of efficacy or adverse event.

When a missing-data mechanism is NMAR, we can obtain consistent estimators, using the full-likelihood (Rubin 1976) which includes an appropriate missing-data mechanism in the likelihood. The approach, however, has difficulty with modeling the missing-data mechanism appropriately and computational difficulty in optimizing the likelihood or even to create identification problems. As a result, the estimation based on the direct-likelihood instead of the full-likelihood is often made. In such cases, it has been suggested that auxiliary variables (Ibrahim et al. 2001; O'Neill and Temple 2012) be added to the model to make up for the lost outcomes to be assessed and to supplement the missing information. The inclusion of auxiliary variables seems to reduce the bias of the direct-likelihood estimator. No theory for the bias reduction has been reported, however.

In clinical trial, post-treatment information such as surrogate endpoints can be auxiliary variables to be added to the model for the analysis of the clinical (true) endpoint. Surrogate endpoints are often evaluated instead of the true endpoint so as to reduce the cost and to shrink the study duration. Prentice (1989) proposed some criteria for external variables to be surrogate endpoints: they should be correlated with the clinical endpoint and fully capture the net effect of treatment on the clinical endpoint, that is, conditional independence between an endpoint variable and a treatment given the surrogate variable (Fleming and DeMets 1996). Several studies which utilize surrogate endpoints as auxiliary variables are found (Fleming et al. 1994; Finkelstein and Shoenfeld 1994; Li et al. 2011).

There are some papers which study the impact of adding auxiliary variables on the bias reduction, but those are based on simulation or application of their approach to actual data. In this paper, we theoretically study the bias of a direct-likelihood estimator when a missing-data mechanism is NMAR, and compare the bias between the models with and without an auxiliary variable. We found that there are cases where inclusion of an auxiliary variable can increase the bias in a simple setup, where the clinical endpoint and auxiliary variable are normally distributed. We derive several conditions for an auxiliary variable to reduce or increase the bias.

In our study, we do not use the assumption that given auxiliary variables a treatment group variable is independent of the clinical endpoint, required by Prentice (1989) as a surrogate endpoint. Hence, the results in this paper can also apply to the mixed-effects model repeated measures (MMRM) (Mallinckrodt et al. 2001) including two post-baseline time points in longitudinal data, by considering intermediate time point value as auxiliary variable.

2 Likelihood and estimators

Let X, Y and Y_a be a treatment, clinical endpoint and auxiliary variables, respectively. Suppose that

$$\begin{aligned} X &\sim f_X(x|\psi), \\ Y|X &\sim f_{Y|X}(y|x;\theta), \\ (Y, Y_a)|X &\sim f_{YY_a|X}(y, y_a|x;\theta,\theta_a). \end{aligned} \tag{1}$$

Here, ψ, θ, and θ_a are distinct parameters from one another and θ is a parameter of interest to be estimated. Only the variable Y can be missing. Assume that we have the following sample of size n from a distribution with $f_{XYY_a}(x, y, y_a)$:

$$\begin{bmatrix} X_1 \\ Y_1 \\ Y_{a,1} \end{bmatrix}, \ldots, \begin{bmatrix} X_m \\ Y_m \\ Y_{a,m} \end{bmatrix}, \begin{bmatrix} X_{m+1} \\ missing \\ Y_{a,m+1} \end{bmatrix}, \ldots, \begin{bmatrix} X_n \\ missing \\ Y_{a,n} \end{bmatrix}, \tag{2}$$

where $Y_1, \ldots, Y_m$ are actually observed, and $Y_{m+1}, \ldots, Y_n$ are missing.

First, let us estimate θ based on (X, Y). The direct-likelihood (Rubin 1976) or the observed likelihood, without any missing-data mechanism, is defined as

$$DL(\theta, \psi|X, Y) = \prod_{i=1}^{m} f_{XY}(X_i, Y_i|\psi, \theta) \prod_{i=m+1}^{n} f_X(X_i|\psi). \tag{3}$$

Let $\tilde{\theta}$ denote the MLE which maximizes the likelihood in (3). Next, let us estimate θ based on (X, Y, Y_a). The direct-likelihood then becomes

$$\begin{aligned} DL_+(\theta, \theta_a, \psi|X, Y, Y_a) = &\prod_{i=1}^{m} f_{XYY_a}(X_i, Y_i, Y_{a,i}|\theta, \theta_a, \psi) \\ &\times \prod_{i=m+1}^{n} f_{XY_a}(X_i, Y_{a,i}|\theta_a, \psi). \end{aligned} \tag{4}$$

Let $\hat{\theta}$ denote the MLE which maximizes the likelihood in (4).

Assume further multivariate normality for $(X, Y, Y_a)'$ as

$$\begin{bmatrix} X \\ Y \\ Y_a \end{bmatrix} \sim N_3(\mu, \Sigma) \text{ with } \mu = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_{y_a} \end{bmatrix}, \ \Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xy_a} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yy_a} \\ \sigma_{y_a x} & \sigma_{y_a y} & \sigma_{y_a y_a} \end{bmatrix}. \tag{5}$$

In the case, the marginal distribution of X and that of (Y, Y_a) given X are obtained as follows:

$$\begin{aligned} X &\sim N(\psi_1, \psi_2), \\ \begin{bmatrix} Y \\ Y_a \end{bmatrix} | X &\sim N_2\left(\begin{bmatrix} \theta_1 + \theta_2 X \\ \theta_{a1} + \theta_{a2} X \end{bmatrix}, \begin{bmatrix} \theta_3 & \theta_{a4} \\ \theta_{a4} & \theta_{a3} \end{bmatrix}\right). \end{aligned} \tag{6}$$

The parameter vector $(\psi_1, \psi_2, \theta_1, \theta_2, \theta_3, \theta_{a1}, \theta_{a2}, \theta_{a3}, \theta_{a4})'$ in (6) can be expressed by those of μ and Σ in (5), and the correspondence is one to one. Let $\psi = (\psi_1, \psi_2)'$, $\theta = (\theta_1, \theta_2, \theta_3)'$, $\theta_a = (\theta_{a1}, \theta_{a2}, \theta_{a3}, \theta_{a4})'$, and $\theta_+ = (\theta', \theta_a')'$. We write the sample means and (co)variances based on complete cases as:

$$\begin{aligned} \bar{X}_{(m)} &= \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \\ S_{xx_{(m)}} &= \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X}_{(m)})^2, \quad S_{yy_{(m)}} = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_{(m)})^2, \\ S_{xy_{(m)}} &= \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X}_{(m)})(Y_i - \bar{Y}_{(m)}). \end{aligned}$$

The MLE $\tilde{\theta}$ for θ based on the likelihood in (3) is known to have the form:

$$\begin{aligned} \tilde{\theta}_1 &= \bar{Y}_{(m)} - S_{yx_{(m)}} S_{xx_{(m)}}^{-1} \bar{X}_{(m)}, \\ \tilde{\theta}_2 &= S_{xx_{(m)}}^{-1} S_{xy_{(m)}}, \\ \tilde{\theta}_3 &= S_{yy\cdot x_{(m)}} = S_{yy_{(m)}} - S_{yx_{(m)}}^2 S_{xx_{(m)}}^{-1}. \end{aligned} \tag{7}$$

Proposition 1 *Let $\tilde{\theta}_i$'s be defined in (7). The MLE $\hat{\theta}_i$'s based on (X, Y, Y_a) defined in (4) can be expressed as follows:*

$$\begin{aligned} \hat{\theta}_1 &= \tilde{\theta}_1 + \frac{S_{yy_a\cdot x_{(m)}}}{S_{y_a y_a\cdot x_{(m)}}} \left\{ \left(\bar{Y}_{a_{(n)}} - \frac{S_{xy_{a_{(n)}}}}{S_{xx_{(n)}}} \bar{X}_{(n)} \right) - \left(\bar{Y}_{a_{(m)}} - \frac{S_{xy_{a_{(m)}}}}{S_{xx_{(m)}}} \bar{X}_{(m)} \right) \right\}, \\ \hat{\theta}_2 &= \tilde{\theta}_2 + \frac{S_{y_a y\cdot x_{(m)}}}{S_{y_a y_a\cdot x_{(m)}}} \left(\frac{S_{xy_{a_{(n)}}}}{S_{xx_{(n)}}} - \frac{S_{xy_{a_{(m)}}}}{S_{xx_{(m)}}} \right), \\ \hat{\theta}_3 &= \tilde{\theta}_3 + \left(\frac{S_{yy_a\cdot x_{(m)}}}{S_{y_a y_a\cdot x_{(m)}}} \right)^2 \left(S_{y_a y_a\cdot x_{(n)}} - S_{y_a y_a\cdot x_{(m)}} \right), \end{aligned}$$

$$\hat{\theta}_{a1} = \bar{Y}_{a_{(n)}} - S_{y_a x_{(n)}} S_{x x_{(n)}}^{-1} \bar{X}_{(n)},$$
$$\hat{\theta}_{a2} = S_{x x_{(n)}}^{-1} S_{x y_{a_{(n)}}},$$
$$\hat{\theta}_{a3} = S_{y_a y_a \cdot x_{(n)}},$$
$$\hat{\theta}_{a4} = S_{y_a y \cdot x_{(m)}} \frac{S_{y_a y_a \cdot x_{(n)}}}{S_{y_a y_a \cdot x_{(m)}}},$$

where the sample means and (co)variances such as $\bar{X}_{(n)}$, $\bar{Y}_{a_{(n)}}$, $S_{x y_{a_{(n)}}}$, $\bar{Y}_{a_{(m)}}$, $S_{x y_{a\,(m)}}$, $S_{y y_a \cdot x_{(m)}}$, $S_{y_a y_a \cdot x_{(m)}}$ will be defined in "Appendix A."

Derivations of the results in Proposition 1 will be given in "Appendix A." It should be noted that every $\hat{\theta}_i$ is an addition of certain terms to $\tilde{\theta}_i$.

3 Bias of direct-likelihood-based MLE

In this section, we shall derive the bias of $\tilde{\theta}$ and $\hat{\theta}$ under NMAR missingness, and then compare between those biases theoretically. For the purpose, we need to specify a missing-data mechanism. Let R_Y be a response indicator taking 0 or 1 for Y being observed or missing, respectively. We use a shared-parameter model (Follmann and Wu 1995; Albert and Follmann 2009) as a distribution of R_Y. We introduce a latent variable Z which connects between R_Y and Y, and suppose that:

$$Z|[X, Y_a, Y] \sim f_{Z|XY_aY}(z|x, y_a, y; \varphi), \tag{8}$$
$$R_Y|Z \sim P(R_Y = 1|z; \tau), \tag{9}$$
$$[X, Y_a, Y] \perp\!\!\!\perp R_Y|Z, \tag{10}$$

where $A \perp\!\!\!\perp B|C$ stands for conditional independence between A and B given C, in which $\perp\!\!\!\perp$ is called Dawid's symbol (Lauritzen 1996, p. 28; Dawid 1979). The missing-data mechanism can then be expressed in the form

$$\begin{aligned} &P(R_Y = 1|x, y_a, y; \psi, \theta_+, \varphi, \tau) \\ &= \int P(R_Y = 1|x, y_a, y, z; \psi, \theta_+, \varphi, \tau) f_{Z|XY_aY}(z|x, y_a, y; \varphi) dz \\ &= \int P(R_Y = 1|z; \tau) f_{Z|XY_aY}(z|x, y_a, y; \varphi) dz \end{aligned}$$

in view of the assumption (10). Note that the mechanism is unrelated with ψ and θ_+, and we can then write the missing-data mechanism as

$$P(R_Y = 1|x, y_a, y; \varphi, \tau) = \int P(R_Y = 1|z; \tau) f_{Z|XY_aY}(z|x, y_a, y; \varphi) dz. \tag{11}$$

It is seen that the conditional probability in (11) generally depends on y, and thus, the missing-data mechanism is not MAR.

Take the following conditional distribution of R_y given Z as an example:

$$P(R_Y = 1|z; \tau) = \begin{cases} 1, & \text{if } z \leq \tau, \\ 0, & \text{if } z > \tau \end{cases}, \tag{12}$$

and assume the following distribution:

$$Z|(X, Y_a, Y) \sim N(\varphi_0 + \varphi_1 X + \varphi_2 Y_a + \varphi_3 Y, \varphi_4). \tag{13}$$

We then have

$$\begin{aligned} P(R_Y = 1|x, y_a, y; \varphi, \tau) &= \int_{-\infty}^{\tau} f_{Z|XY_aY}(z|x, y_a, y; \varphi)dz \\ &= \int_{-\infty}^{\tau} N(z|\varphi_0 + \varphi_1 x + \varphi_2 y_a + \varphi_3 y, \varphi_4)dz \\ &= \Phi\left((\tau - \varphi_0 - \varphi_1 x - \varphi_2 y_a - \varphi_3 y)/\sqrt{\varphi_4}\right), \end{aligned}$$

where $N(z|\mu, \sigma^2)$ means the probability density function of the normal distribution with mean μ and variance σ^2, and $\Phi(z)$ means the cumulative distribution function of the standard normal distribution. Note that the missing-data mechanism is a probit regression model of $(\tau - \varphi_0 - \varphi_1 x - \varphi_2 y_a - \varphi_3 y)/\sqrt{\varphi_4}$. If we specify the functional form in (12) appropriately, we can express a wide range of missing-data mechanisms.

By the weak law of large numbers we easily see that $\bar{X}_{(n)} S_{xx_{(n)}} S_{xy_{a_{(n)}}}$ converge in probability to μ_x, σ_{xx} and σ_{xy_a}, respectively. Under the above setup, we have that the statistics, $\bar{X}_{(m)} S_{xx_{(m)}}$ and $S_{xy_{a_{(m)}}}$ using complete cases only, converge when $n \to \infty$ as follows:

$$\begin{aligned} \bar{X}_{(m)} &\xrightarrow{P} \mu_x + \frac{\sigma_{xz}}{\sigma_{zz}} E[Z - \mu_z|R_Y = 1], \\ S_{xx_{(m)}} &\xrightarrow{P} \sigma_{xx} + \frac{\sigma_{xz}^2}{\sigma_{zz}^2} \left(Var[Z|R_Y = 1] - \sigma_{zz}\right), \\ S_{xy_{a_{(m)}}} &\xrightarrow{P} \sigma_{xy_a} + \frac{\sigma_{xz}\sigma_{zy_a}}{\sigma_{zz}^2} \left(Var[Z|R_Y = 1] - \sigma_{zz}\right). \end{aligned} \tag{14}$$

The other statistics such as $\bar{Y}_{(m)}$, $\bar{Y}_{a_{(m)}}$, $S_{yy_{(m)}}$, $S_{y_a y_{a_{(m)}}}$, $S_{xy_{(m)}}$, $S_{yy_{a_{(m)}}}$ have similar convergence properties. Proofs will be provided in "Appendix B."

Using those results we obtain the asymptotic bias of $\tilde{\theta}$ as follows.

Proposition 2 *Under the model without auxiliary variable Y_a, we have:*

$$\tilde{\theta}_1 - \theta_1 \xrightarrow{P} \frac{\sigma_{zy \cdot x}}{\sigma_{zz} + \rho_{zx}^2 A} \left(E[Z - \mu_z|R_Y = 1] - \frac{\sigma_{xz} A \mu_x}{\sigma_{xx}\sigma_{zz}}\right), \tag{15}$$

$$\tilde{\theta}_2 - \theta_2 \xrightarrow{P} \frac{\sigma_{xz}(\sigma_{xx}\sigma_{zz})^{-1} A \sigma_{zy \cdot x}}{\sigma_{zz} + \rho_{xz}^2 A}, \tag{16}$$

$$\tilde{\theta}_3 - \theta_3 \xrightarrow{P} \frac{\sigma_{zz}^{-1} A \sigma_{zy\cdot x}^2}{\sigma_{zz} + \rho_{zx}^2 A}, \tag{17}$$

where $A = Var[Z|R_Y = 1] - \sigma_{zz},\ \sigma_{zy\cdot x} = \sigma_{zy} - \sigma_{zx}\sigma_{xx}^{-1}\sigma_{xy}$.
Note that $\sigma_{zz} + \rho_{xz}^2 A$ is always nonnegative and positive if $\sigma_{xx\cdot z} > 0$. Under the model without Y_a, we can easily see that the MLEs are consistent if $Z \perp\!\!\!\perp Y|X$, implying that the missing is MAR, since each asymptotic bias has the factor $\sigma_{zy\cdot x}$.

Similarly, we also obtain the asymptotic bias of $\hat{\theta}$ as follows.

Proposition 3 *Under the model with auxiliary variable Y_a, we have:*

$$\hat{\theta}_1 - \theta_1 \xrightarrow{P} \frac{\sigma_{zy\cdot xy_a}}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a\cdot x}^2 (1 - \rho_{xz}^2) A} \left(E[Z - \mu_z | R_Y = 1] - \frac{\sigma_{xz} A \mu_x}{\sigma_{xx}\sigma_{zz}} \right), \tag{18}$$

$$\hat{\theta}_2 - \theta_2 \xrightarrow{P} \frac{\sigma_{xz}(\sigma_{xx}\sigma_{zz})^{-1} A \sigma_{zy\cdot xy_a}}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a\cdot x}^2 (1 - \rho_{xz}^2) A}, \tag{19}$$

$$\hat{\theta}_3 - \theta_3 \xrightarrow{P} \frac{\sigma_{zz}^{-1} A \sigma_{zy\cdot xy_a}}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a\cdot x}^2 (1 - \rho_{xz}^2) A} \left(2\sigma_{zy\cdot x} - \frac{(\sigma_{zz} + \rho_{xz}^2 A)\sigma_{zy\cdot xy_a}}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a\cdot x}^2 (1 - \rho_{xz}^2) A} \right), \tag{20}$$

where $A = Var[Z|R_Y = 1] - \sigma_{zz},\ \sigma_{zy\cdot xy_a} = \sigma_{zy\cdot y_a} - \sigma_{zx\cdot y_a}\sigma_{xx\cdot y_a}^{-1}\sigma_{xy\cdot y_a}$.
Note that $\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a\cdot x}^2 (1 - \rho_{xz}^2) A$ is always nonnegative and positive if $\sigma_{xx\cdot z} > 0$ and $\rho_{y_a z\cdot x}^2 < 1$.

We can also easily see that the MLEs are consistent if $Z \perp\!\!\!\perp Y|(X, Y_a)$, implying MAR, since each asymptotic bias has the factor $\sigma_{zy\cdot xy_a}$.

Here, we use ratio to evaluate the bias of MLE with Y_a against that without Y_a. We have the following proposition.

Proposition 4 *If $\sigma_{zy\cdot x} \neq 0$, we obtain the bias of each MLE with auxiliary variable (Y_a) divided by that without Y_a as follows:*

$$\frac{\hat{\theta}_1 - \theta_1}{\tilde{\theta}_1 - \theta_1} \xrightarrow{P} B, \tag{21}$$

$$\frac{\hat{\theta}_2 - \theta_2}{\tilde{\theta}_2 - \theta_2} \xrightarrow{P} B, \tag{22}$$

$$\frac{\hat{\theta}_3 - \theta_3}{\tilde{\theta}_3 - \theta_3} \xrightarrow{P} B(2 - B), \tag{23}$$

where

$$B = \left(1 - \frac{\rho_{zy_a \cdot x}\rho_{y_a y \cdot x}}{\rho_{zy \cdot x}}\right) \frac{\sigma_{zz} + \rho_{xz}^2 A}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a \cdot x}^2 (1 - \rho_{xz}^2) A}. \tag{24}$$

An interesting point of the proposition is that it holds true irrespective of the form of the missing-data mechanism in terms of Z given in (9).

4 Several investigations on bias reduction

We shall consider the bias ratio under several conditions to study effects on bias reduction due to inclusion of an auxiliary variable. It follows from Proposition 4 that the biases of MLEs for θ_1 and θ_2 and for θ_3 will be reduced by adding the auxiliary variable Y_a if $|B| < 1$ for θ_1 and θ_2 and $|B-1| < \sqrt{2}$ for θ_3, respectively. In particular, the biases of the MLEs for all the parameters reduce if $0 \leq B < 1$.

Next, we explore some situations on the constant B to meet the conditions above. First, consider the case where $\rho_{zy \cdot xy_a} = 0$, and then the expression in (24) shows the bias is zero. The case is, however, that the missing-data mechanism under the model with Y_a is MAR, and thus, the bias of the direct MLE is known to be zero (see, e.g., Little and Rubin 2002; Takai and Kano 2013).

In the case where $\rho_{zy_a \cdot xy} = 0$, meaning no direct effect of Y_a on Z, it follows that $\rho_{zy_a \cdot x} = \rho_{zy \cdot x}\rho_{yy_a \cdot x}$. Hence, we obtain:

$$B = \frac{(1 - \rho_{yy_a \cdot x}^2)(\sigma_{zz} + \rho_{xz}^2 A)}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy \cdot x}^2 \rho_{yy_a \cdot x}^2 (1 - \rho_{xz}^2) A}, \tag{25}$$

from which we have that $0 \leq B < 1$. Thus, the biases will reduce due to the Y_a.

In the case where $\rho_{yy_a \cdot xz} = 0$, meaning no direct effect of Y_a on Y, we can consider the case as an extreme case with small correlation between Y and Y_a. Notice that Y_a is no longer said to be a surrogate endpoint. In the situation, it follows that $\rho_{yy_a \cdot x} = \rho_{yz \cdot x}\rho_{zy_a \cdot x}$. Hence, we obtain

$$B = \frac{(1 - \rho_{zy_a \cdot x}^2)(\sigma_{zz} + \rho_{xz}^2 A)}{\sigma_{zz} + \rho_{zx}^2 A + \rho_{zy_a \cdot x}^2 (1 - \rho_{xz}^2) A} \tag{26}$$

from which we also have that $0 \leq B < 1$, and the biases will reduce due to the Y_a.

Summarizing these results, we obtain the following result.

Proposition 5 *Suppose that $\rho_{zy_a \cdot xy} = 0$ or $\rho_{yy_a \cdot xz} = 0$ holds true. Then, we have $0 \leq B < 1$, and the biases of the MLEs for all parameters are reduced by adding the auxiliary variable Y_a. The results hold true for any functional form of $P(R_Y = 1|z; \tau)$ in (9).*

Next, consider general unstructured covariances, where the assumption in Proposition 5 does not necessarily hold. Let us write $B = B_1 B_2$, where

$$B_1 = 1 - \frac{\rho_{zy_a \cdot x} \rho_{y_a y \cdot x}}{\rho_{zy \cdot x}},$$
$$B_2 = \frac{\sigma_{zz} + \rho_{xz}^2 A}{\sigma_{zz} + \rho_{xz}^2 A + \rho_{zy_a \cdot x}^2 (1 - \rho_{xz}^2) A}.$$

In case of $A \geq 0$, we can easily see that $0 < B_2 \leq 1$, and that a sufficient condition for reducing the biases by adding Y_a is $|B_1| < 1$, that is,

$$0 < \frac{\rho_{zy_a \cdot x} \rho_{y_a y \cdot x}}{\rho_{zy \cdot x}} < 2. \tag{27}$$

Since B_2 is monotonically decreasing in A on $A > -\sigma_{zz}$, we have

$$\frac{\rho_{xz}^2}{\rho_{xz}^2 + \rho_{zy_a \cdot x}^2 (1 - \rho_{xz}^2)} < B_2 < \frac{1}{1 - \rho_{zy_a \cdot x}^2}. \tag{28}$$

Thus, a sufficient condition to reduce the biases of MLE by adding Y_a is

$$|B_1| \frac{1}{1 - \rho_{zy_a \cdot x}^2} < 1, \tag{29}$$

which is equivalent to

$$\rho_{zy_a \cdot x}^2 < \frac{\rho_{zy_a \cdot x} \rho_{y_a y \cdot x}}{\rho_{zy \cdot x}} < 2 - \rho_{zy_a \cdot x}^2. \tag{30}$$

Now we obtain the following result.

Proposition 6 *Under the inequality assumption in* (30)*, the bias of MLE by adding* Y_a *is reduced. If* $A \geq 0$*, the assumption in* (30) *can be relaxed as that in* (27)*.*

In general, the condition $A \geq 0$ does not hold, as shown below. Suppose Z is normally distributed and the conditional distribution of R_Y given Z is expressed in (12), and then Z follows a truncated normal distribution. It is well known that the variance of a truncated normal distribution is smaller than that of the untruncated normal distribution. In the case, we have that $A < 0$.

5 Numerical evaluations of bias ratio

Here, we shall numerically calculate and plot the bias ratios under several conditions. We put in (24) $\sigma_{xx} = \sigma_{zz} = 1$, and take $A = -0.95, 0, 1$. Note that $A = V[Z|R_Y = 1] - 1 > -1$.

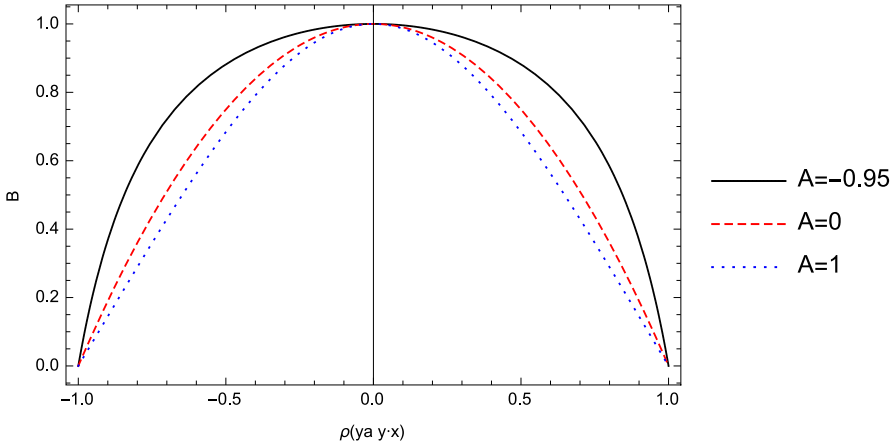

Fig. 1 Graph of B against $\rho_{y_a y \cdot x}$ in the case where $\rho_{z y_a \cdot xy} = 0$

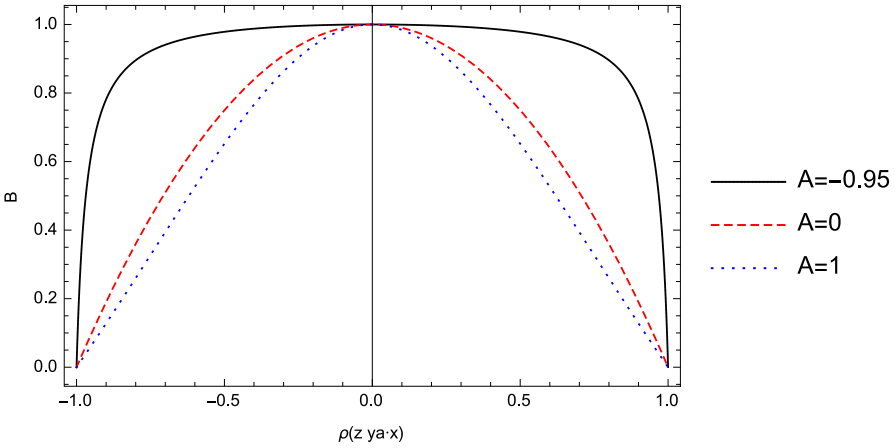

Fig. 2 Graph of B against $\rho_{z y_a \cdot x}$ in the case where $\rho_{y_a y \cdot xz} = 0$

In the case where $\rho_{z y_a \cdot xy} = 0$, the graph of B against $\rho_{y_a y \cdot x}$ is shown as in Fig. 1. It can be seen that the bias is always reduced when adding Y_a with $\rho_{y_a y \cdot x} \neq 0$, and that the larger $|\rho_{y_a y \cdot x}|$ is, the more reduced the bias of estimator with Y_a is. It can be interpreted that the bias reduction is due to the supplement of missing information of endpoint values by adding the auxiliary variable Y_a. The smaller A is, the more gradual the reduction of the bias ratio is.

Under $\rho_{y y_a \cdot xz} = 0$, we obtained the graph of B against $\rho_{z y_a \cdot x}$ shown as Fig. 2. It can be seen that the bias is always reduced when adding Y_a with $\rho_{z y_a \cdot x} \neq 0$, as before. The larger $|\rho_{z y_a \cdot x}|$ is, the more reduced the bias of estimator with Y_a is. It can be interpreted that the bias is reduced as the variation of Z is reduced by adding Y_a, so that the connection between Y and R_Y is weaken. The smaller A is, the more gradual the reduction of the bias ratio is.

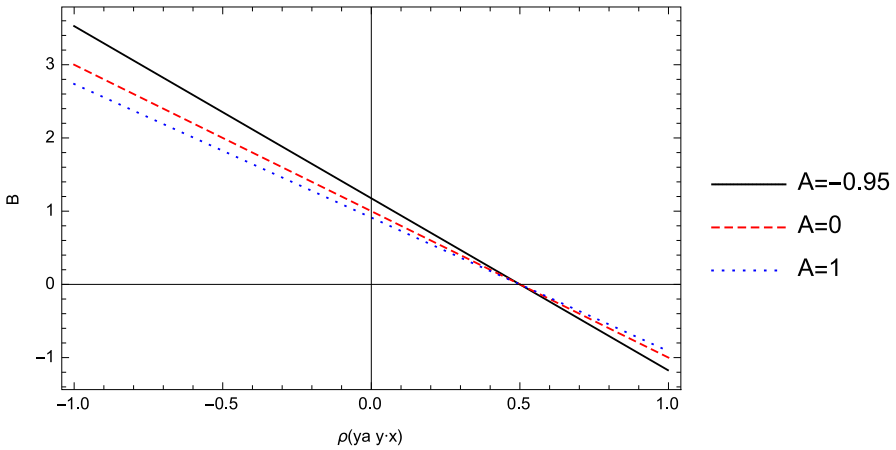

Fig. 3 Graph of B against $\rho_{y_a y \cdot x}$ in the case of unstructured correlations

In the case of unstructured correlations unlike the above situations, we presented the graph of B against $\rho_{y_a y \cdot x}$ as Fig. 3 by putting in B $\rho_{zy \cdot x} = 0.2$, $\rho_{zy_a \cdot x} = 0.4$, $\rho_{xz} = 0.5$. The graph shows that the bias ratio is 0 when $\rho_{yy_a \cdot x} = 0.5$, in which $\rho_{zy \cdot xy_a} = 0$, implying MAR under the model with Y_a. When $\rho_{y_a y \cdot x} < 0$, the bias increases by adding Y_a regardless of A. When $A \geq 0$, the bias is reduced by adding Y_a as long as $\rho_{y_a y \cdot x} > 0$.

6 Application to age-related macular degeneration study data

In this section, we take an actual data set of a clinical study of age-related macular degeneration (ARMD) to see how our theoretical results in Sect. 2 agree to an analysis of real data. For details about the study and results, see "Pharmacological Therapy for Macular Degeneration Study Group (1997)." ARMD causes impairment of visual acuity and vision loss at worst. The study is double-blinded, placebo-control, parallel-group study to evaluate the efficacy and safety in patients with ARMD. Four hundred and eighty-one patients are randomized to assign one of four dose groups (α-2a 1.5 MIU, 3 MIU, 6 MIU, and placebo). For efficacy, the visual acuity test was performed using Early Treatment of Diabetic Retinopathy Study chart. The actual data set of visual acuity scores assessed at baseline, Week 4, 12, 24, and 52 from the study, is available as an example data of the nlmeU package on R language.

We consider change from the baseline at Week 52 as clinical endpoint (Y) and change from the baseline at Week 24 as a surrogate, auxiliary variable (Y_a), and compare the two estimates with and without Y_a for the treatment difference in change at Week 52 between 6MIU and Placebo (6MIU − Placebo).

The treatment difference MLE without auxiliary variable Y_a is calculated as -4.122, whereas that with Y_a is -4.619. These estimates are consistent with those from MMRM (Mallinckrodt et al. 2001). The observation is reasonable because the struc-

ture of the direct-likelihood is similar to that of MMRM. Note that the estimate from MMRM using all time point variables available is -4.862.

The primary purpose of the propositions given in Sect. 5 and the numerical evaluations here is to give theoretical evidence to reduction or inflation of the biases of the direct-likelihood MLE, and to give cautions practical users who blindly add auxiliary variables.

For practical use, obviously more easy-to-check sufficient conditions need to be developed. That is our future issue.

Appendix A: MLE based on direct-likelihood with Y_a

We shall obtain MLE based on direct-likelihood with Y_a according to Anderson (1957) in which MLE is more easily derived by using characteristics of normal distribution and reparametrization.

Assuming that (X, Y, Y_a) have a normal distribution in (5), the conditional distribution of Y given (X, Y_a) follows a normal distribution with mean $\beta_0 + \beta_x X + \beta_{y_a} Y_a$ and variance σ_e^2, where:

$$\begin{aligned}
&\beta_0 = \mu_y - \beta_x \mu_x - \beta_{y_a} \mu_{y_a}, \quad \beta_x = \sigma_{xx\cdot y_a}^{-1} \sigma_{xy\cdot y_a}, \quad \beta_{y_a} = \sigma_{y_a y_a \cdot x}^{-1} \sigma_{y_a y\cdot x}, \\
&\sigma_e^2 = \sigma_{yy\cdot x y_a} = \sigma_{yy\cdot x} - \sigma_{yy_a\cdot x}^2 \sigma_{y_a y_a\cdot x}^{-1}.
\end{aligned} \tag{31}$$

Hence, the direct-likelihood DL_+ can be rewritten by the reparametrization as follows:

$$\begin{aligned}
DL_+ = &\prod_{i=1}^{m} f_{Y|XY_a}(Y_i|X_i, Y_{a,i}; \beta_0, \beta_x, \beta_{y_a}, \sigma_e^2) \\
&\times \prod_{i=1}^{n} f_{XY_a}(X_i, Y_{a,i}|\mu_x, \mu_{y_a}, \sigma_{xx}, \sigma_{y_a y_a}, \sigma_{x y_a}).
\end{aligned} \tag{32}$$

MLEs of $\mu_x, \mu_{y_a}, \sigma_{xx}, \sigma_{xy_a}, \sigma_{y_a y_a}$ are easily obtained from the second factor of the DL_+ as follows:

$$\begin{aligned}
&\hat{\mu}_x = \bar{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_{y_a} = \bar{Y}_{a_{(n)}} = \frac{1}{n}\sum_{i=1}^{n} Y_{a,i}, \\
&\hat{\sigma}_{xx} = S_{xx_{(n)}} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}_{(n)}\right)^2, \quad \hat{\sigma}_{y_a y_a} = S_{y_a y_{a_{(n)}}} = \frac{1}{n}\sum_{i=1}^{n} \left(Y_{a,i} - \bar{Y}_{a_{(n)}}\right)^2, \\
&\hat{\sigma}_{x y_a} = S_{x y_{a_{(n)}}} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_{(n)})(Y_{a,i} - \bar{Y}_{a_{(n)}}).
\end{aligned}$$

The MLE of θ_{a1}, θ_{a2}, and θ_{a3} will easily be found from the results.

MLEs of the remaining parameters of $\beta_0, \beta_x, \beta_{y_a}, \sigma_e^2$ are obtained from the first factor of DL_+ from the standard results on the linear regression as follows:

$$\begin{aligned}
\hat{\beta}_x &= S_{xx\cdot y_{a(m)}}^{-1} S_{xy\cdot y_{a(m)}}, \quad \hat{\beta}_{y_a} = S_{y_a y_a \cdot x_{(m)}}^{-1} S_{y_a y\cdot x_{(m)}}, \\
\hat{\beta}_0 &= \bar{Y}_{(m)} - \hat{\beta}_x \bar{X}_{(m)} - \hat{\beta}_{y_a} \bar{Y}_{a_{(m)}}, \\
\hat{\sigma}_e^2 &= \hat{\sigma}_{yy\cdot xy_a} = S_{yy\cdot xy_{a(m)}} = S_{yy\cdot x_{(m)}} - S_{yy_a\cdot x_{(m)}}^2 S_{y_a y_a\cdot x_{(m)}}^{-1},
\end{aligned}$$

where

$$\begin{aligned}
\bar{X}_{(m)} &= \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y}_{a_{(m)}} = \frac{1}{m}\sum_{i=1}^{m} Y_{a,i}, \quad \bar{Y}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \\
S_{xx_{(m)}} &= \frac{1}{m}\sum_{k=1}^{m}(X_i - \bar{X}_{(m)})^2, \quad S_{y_a y_{a(m)}} = \frac{1}{m}\sum_{k=1}^{m}\left(Y_{a,i} - \bar{Y}_{a_{(m)}}\right)^2, \\
S_{xy_{a(m)}} &= \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X}_{(m)})(Y_{a,i} - \bar{Y}_{a_{(m)}}), \\
S_{uv\cdot w_{(m)}} &= S_{uv_{(m)}} - S_{uw_{(m)}} S_{ww_{(m)}}^{-1} S_{wv_{(m)}}.
\end{aligned}$$

It follows from (31) and the relationship in parameters between (5) and (6) that

$$\begin{aligned}
\theta_1 &= \mu_y - \sigma_{yx}\sigma_{xx}^{-1}\mu_x = (\beta_0 + \beta_x\mu_x + \beta_{y_a}\mu_{y_a}) - \theta_2\mu_x, \\
\theta_2 &= \sigma_{xx}^{-1}\sigma_{xy} = \sigma_{xx}^{-1}(\sigma_{xx}\beta_x + \sigma_{y_a x}\beta_{y_a}) = \beta_x + \sigma_{xx}^{-1}\sigma_{xy_a}\beta_{y_a}, \\
\theta_3 &= \sigma_{yy\cdot x} = \sigma_e^2 + \beta_{y_a}^2\sigma_{y_a y_a\cdot x}, \\
\theta_{a4} &= \sigma_{y_a y\cdot x} = \sigma_{y_a y_a\cdot x}\beta_{y_a}.
\end{aligned}$$

Hence, we obtain MLE of these parameters as follows, which means the results in Proposition 2:

$$\begin{aligned}
\hat{\theta}_1 &= (\hat{\beta}_0 + \hat{\beta}_x\hat{\mu}_x + \hat{\beta}_{y_a}\hat{\mu}_{y_a}) - \hat{\theta}_2\hat{\mu}_x, \\
\hat{\theta}_2 &= \hat{\beta}_x + \hat{\sigma}_{xx}^{-1}\hat{\sigma}_{xy_a}\hat{\beta}_{y_a}, \\
\hat{\theta}_3 &= \hat{\sigma}_e^2 + \hat{\beta}_{y_a}^2\hat{\sigma}_{y_a y_a\cdot x}, \\
\hat{\theta}_{a4} &= \hat{\sigma}_{y_a y_a\cdot x}\hat{\beta}_{y_a}.
\end{aligned}$$

Appendix B: Limit of statistics using complete cases only

Here, we shall show the following convergences for limits of $\bar{X}_{(m)}$ and $S_{xx_{(m)}}$ as n tends to infinity. The limits of the other statistics also have the same properties.

$$\bar{X}_{(m)} \xrightarrow{P} \mu_x + \frac{\sigma_{xz}}{\sigma_{zz}} E[Z - \mu_z | R_Y = 1], \tag{33}$$

$$S_{xx_{(m)}} \xrightarrow{P} \sigma_{xx} + \frac{\sigma_{xz}^2}{\sigma_{zz}^2} (Var[Z|R_Y = 1] - \sigma_{zz}). \tag{34}$$

Assuming that (X, Y, Y_a, Z) have a normal distribution in addition to (5), where the mean and variance of Z are μ_z and σ_{zz}, respectively, and covariance between Z and (X, Y, Y_a) is $(\sigma_{zx}, \sigma_{zy}, \sigma_{zy_a})$.

Using the response indicator R_Y, $\bar{X}_{(m)}$ is expressed in the form:

$$\bar{X}_{(m)} = \frac{1}{\sum_{i=1}^n R_{Y_i}} \sum_{i=1}^n R_{Y_i} X_i.$$

By the weak law of large numbers, we obtain

$$\bar{X}_{(m)} \xrightarrow{P} \frac{E[R_Y X]}{E[R_Y]} = \frac{E[X|R_Y = 1]P(R_Y = 1)}{P(R_Y = 1)} = E[X|R_Y = 1].$$

By using the condition $X \perp\!\!\!\perp R_Y | Z$, we obtain (33) shown as follows:

$$E[X|R_Y = 1] = E[E[X|Z]|R_Y = 1] = \mu_x + \frac{\sigma_{xz}}{\sigma_{zz}} E[Z - \mu_z | R_Y = 1]. \tag{35}$$

For $S_{xx_{(m)}}$, we can rewrite using response indicator R_Y as follows:

$$S_{xx_{(m)}} = \frac{1}{\sum_{i=1}^n R_{Y_i}} \sum_{i=1}^n R_{Y_i} \left(X_i - \bar{X}_{(m)}\right)^2.$$

By applying the weak law of large numbers,

$$\begin{aligned}
S_{xx_{(m)}} =& \frac{1}{\sum_{i=1}^n R_{Y_i}} \sum_{i=1}^n R_{Y_i} X_i^2 - \left(\bar{X}_{(m)}\right)^2 \\
\xrightarrow{P}& \frac{E\left[R_Y X^2\right]}{E[R_Y]} - \left(E\left[X|R_Y = 1\right]\right)^2 \\
=& E\left[X^2|R_Y = 1\right] - \left(E\left[X|R_Y = 1\right]\right)^2 \\
=& E\left[\{X - E[X|R_Y = 1]\}^2 |R_Y = 1\right] = Var[X|R_Y = 1] \\
=& E\left[\{(X - E[X|Z]) + (E[X|Z] - E[X|R_Y = 1])\}^2 |R_Y = 1\right] \\
=& E\left[\{X - E[X|Z]\}^2 |R_Y = 1\right] \\
&+ E\left[\{E[X|Z] - E[X|R_Y = 1]\}^2 |R_Y = 1\right] \\
&+ 2E\left[\{X - E[X|Z]\}\{E[X|Z] - E[X|R_Y = 1]\}|R_Y = 1\right].
\end{aligned} \tag{36}$$

By noting that $X \perp\!\!\!\perp R_Y|Z$, we can evaluate the third term as follows:

$$\begin{aligned}
&2E\left[\{X - E[X|Z]\}\{E[X|Z] - E[X|R_Y = 1]\}|R_Y = 1\right] \\
&\quad = 2E\left[E\left[\{X - E[X|Z]\}\{E[X|Z] - E[X|R_Y = 1]\}|Z, R_Y = 1\right]|R_Y = 1\right] \\
&\quad = 2E\left[E\left[X - E[X|Z]|Z, R_Y = 1\right]\{E[X|Z] - E[X|R_Y = 1]\}|R_Y = 1\right] \\
&\quad = 2E\left[E\left[X - E[X|Z]|Z\right]\{E[X|Z] - E[X|R_Y = 1]\}|R_Y = 1\right] \\
&\quad = 0.
\end{aligned}$$

The first term is written as follows:

$$\begin{aligned}
E\left[\{X - E[X|Z]\}^2|R_Y = 1\right] &= E\left[E\left[\{X - E[X|Z]\}^2|Z\right]|R_Y = 1\right] \\
&= \sigma_{xx\cdot z}. \qquad (37)
\end{aligned}$$

The second term is written by using (35) as follows:

$$\begin{aligned}
&E\left[\{E[X|Z] - E[X|R_Y = 1]\}^2|R_Y = 1\right] \\
&\quad = E\left[\frac{\sigma_{xz}^2}{\sigma_{zz}^2}(Z - E[Z|R_Y = 1])^2|R_Y = 1\right] \\
&\quad = \frac{\sigma_{xz}^2}{\sigma_{zz}^2}Var[Z|R_Y = 1]. \qquad (38)
\end{aligned}$$

Hence, we finally obtain:

$$S_{xx_{(m)}} \xrightarrow{P} \sigma_{xx\cdot z} + \frac{\sigma_{xz}^2}{\sigma_{zz}^2}Var[Z|R_Y = 1] = \sigma_{xx} + \frac{\sigma_{xz}^2}{\sigma_{zz}^2}(Var[Z|R_Y = 1] - \sigma_{zz}).$$

Similar derivations have been used in Kano (2015).

References

Albert, P. S., Follmann, D. A. (2009). Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 433–452). Boca Raton, FL: Chapman & Hall/CRC Press.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of American Statistical Association*, *52*(278), 200–203.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B (Methodological)*, *41*(1), 1–31.

Finkelstein, D., Shoenfeld, D. (1994). Analysing survival in the presense of an auxiliary variable. *Statistics in Medicine*, *13*, 1747–1754.

Fleming, T. R., DeMets, D. L. (1996). Surrogate end points in clinical trials: Are we being misled? *Annals of Internal Medicine*, *125*, 605–613.

Fleming, T. R., Prentice, R. L., Pepe, M. S., Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, *13*, 955–968.

Follmann, D., Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, *51*, 151–168.

Ibrahim, J. G., Lipsitz, S. R., Horton, N. (2001). Using auxiliary data for parameter estimation with non-ignorability missing outcomes. *Applied Statistics*, *50*(3), 361–373.

International Conference on Harmonisation E9 Expert Working Group. (1999). Statistical principles for clinical trials: ICH Harmonised Tripartite Guideline. *Statistics in Medicine, 18*, 955–968.

Kano, Y. (2015). *Developments in multivariate missing data analysis*. A paper presented at International Meeting of the Psychometric Society (IMPS2015). Peking, China.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Li, Y., Taylor, J. M. G., Little, R. J. A. (2011). A shrinkage approach for estimating a treatment effect using intermediate biomarker data in clinical trials. *Biometrics*, *67*, 1434–1441.

Little, R. J. A., Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Mallinckrodt, C. H., Clark, W. S., David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, *11*(1&2), 9–21.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials (Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education)*. Washington, DC: The National Academies Press.

O'Neill, R. T., Temple, R. (2012). The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clinical Pharmacology and Therapeutics*, *91*(3), 550–554.

Pharmacological Therapy for Macular Degeneration Study Group. (1997). Interferon alpha-2a is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. *Archives of Ophthalmology*, *115*(7), 865–872.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, *8*, 431–440.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Takai, K., Kano, Y. (2013). Asymptotic inference with incomplete data. *Communications in Statistics—Theory and Methods*, *42*(17), 3174–3190.