stats

MDPI

Article

Improved Small Sample Inference Methods for a Mixed-Effects Model for Repeated Measures Approach in Incomplete Longitudinal Data Analysis

Yoshifumi Ukyo [1,2], Hisashi Noma [3,*], Kazushi Maruo [4] and Masahiko Gosho [4]

1 Department of Biostatistics, Janssen Pharmaceutical K. K., 5-2 Nishi-kanda 3-chome, Chiyoda-ku, Tokyo 101-0065, Japan; yukyo@its.jnj.com
2 Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
3 Department of Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
4 Department of Biostatistics, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan; maruo@md.tsukuba.ac.jp (K.M.); mgosho@md.tsukuba.ac.jp (M.G.)
* Correspondence: noma@ism.ac.jp

Received: 16 January 2019; Accepted: 23 March 2019; Published: 28 March 2019

Abstract: The mixed-effects model for repeated measures (MMRM) approach has been widely applied for longitudinal clinical trials. Many of the standard inference methods of MMRM could possibly lead to the inflation of type I error rates for the tests of treatment effect, when the longitudinal dataset is small and involves missing measurements. We propose two improved inference methods for the MMRM analyses, (1) the Bartlett correction with the adjustment term approximated by bootstrap, and (2) the Monte Carlo test using an estimated null distribution by bootstrap. These methods can be implemented regardless of model complexity and missing patterns via a unified computational framework. Through simulation studies, the proposed methods maintain the type I error rate properly, even for small and incomplete longitudinal clinical trial settings. Applications to a postnatal depression clinical trial are also presented.

Keywords: Bartlett adjustment; MMRM; missing data; longitudinal data analysis; resampling

1. Introduction

Clinical trials for new drug development are often longitudinal trials in which outcome variables are repeatedly measured. In these trials, the primary analyses usually compare the treatment efficacy with a comparator at the end of a follow-up period. However, during the follow-up period, dropouts or missing outcome variables usually occur, and may seriously influence the validity and precision of the statistical inference. In addition, regulatory guidelines for preventing and treating the missing data in clinical trials have been issued [1–3], and adequate practices have been strongly pursued in recent years. Following these discussions, the mixed-effects model for repeated measures (MMRM) [4–6] has been widely applied for primary analyses of clinical trials in drug development. This type of model allows for valid statistical inference under incomplete longitudinal repeated measurements based on the direct likelihood approach.

MMRM is a type of linear mixed model (LMM) [7–9] that directly models the variance-covariance matrix of the longitudinal multivariate outcome variables [5], in which random effects are included as part of the marginal covariance matrix. One of the advantages of MMRM is that it enables flexible modelling of the correlation structure between time points to ensure the validity of inference in treatment efficacy. Further, the ordinary inference methods for the LMM (e.g., restricted maximum

likelihood (REML) method) [10] are based on large sample approximations. Their validities are violated under small or moderate sample settings [11,12]. For the MMRM, Gosho et al. [13] also showed the invalidity of the ordinary inference methods under small or moderate sample settings.

In order to resolve these problems, several related works have been conducted regarding the conventional LMM. One solution is to adopt a higher-order asymptotic theory. Zucker et al. [14] studied the Bartlett correction [15] and the Cox-Reid adjusted likelihood [16] as well as their combination of the likelihood ratio (LR) test. Lyons and Peters [17] and Guolo et al. [18] proposed a higher-order asymptotic approach by adapting Skovgaard's improved modified signed log-likelihood ratio [19] introduced by Barndorff-Nielsen [20]. Stein et al. [21] investigated the modified profile likelihood approach of Barndorff-Nielsen [20] based on the approximation method of Severini [22]. Although the improved methods by Stein et al. [21] performed well in their simulation studies, they compared their methods only with the naïve LR test. In addition, their improved methods require complicated analytical calculations involved in higher-order differentiations of log-likelihood in case-by-case analyses. In particular, when applying the MMRM in longitudinal clinical trials, the marginal covariance structure is usually assumed to be a complicated form to circumvent model misspecifications, and calculations by Stein et al. [21] would not be realistic in practical use. Stein et al. [21] also investigated bootstrap-based approximation inferences, but their discussions and numerical evaluations were also limited within the conventional LMM framework, and the applications to MMRM for incomplete longitudinal studies were not discussed.

In this study, we proposed and investigated two improved inference methods involved in MMRM under small sample size and with incomplete data for longitudinal clinical trials. To circumvent the practical difficulties in implementing the analytical calculations, we adopted numerical approximations using bootstrap inferences [23–25]. The first method was the Bartlett correction [15] with the adjustment term approximated by bootstraps [23]. This approach can effectively circumvent case-by-case complicated analytical calculations and can be generally applied regardless of model complexity and missing patterns via a unified computational framework. In addition, the second involved the Monte Carlo test using empirical distribution constructed by the bootstrap; this was a straightforward approach, and is widely known to be an effective method under these situations. The resampling schemes of both methods allowed outcome variables to be incomplete, and we evaluated their validities and performances under practical situations of longitudinal clinical trials with missing data. In addition, we compared these methods with those of standard methods such as REML using Kenward-Roger's (KR) [11] method and unstructured covariance structure. We also assessed their practical effectiveness, illustrating their application to a postnatal depression clinical trial [26].

This paper is organized as follows: we first review the MMRM for longitudinal data analyses in Section 2. We then provide our approaches to improve the statistical inferences of MMRM in Section 3. We provide simulation evaluations in Section 4, and we apply our methods to the postnatal depression clinical trial data in Section 5. Lastly, we conclude with some discussion in Section 6.

2. Mixed-Effects Model for Repeated Measures (MMRM) for Longitudinal Data Analysis

Suppose subjects were randomized to two treatment groups (e.g., active drug vs. placebo). A continuous outcome was measured repeatedly over n time points. Also, we considered that the total number of subjects in the two groups was N and the number of time points that the outcomes were observed for the ith individual is n_i $(n_i \leq n)$. The primary analysis of interest was to compare the mean difference of the primary endpoint at the final nth time point. In this study, we supposed only monotonic missing data for simplicity, but our discussions can be straightforwardly extended to non-monotonic missingness.

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ was the $n_i \times 1$ random error vector distributed independently as $MVN(0, \boldsymbol{\Sigma}_i)$. $\boldsymbol{\Sigma}_i$ was the $n_i \times n_i$ variance–covariance matrix. The random effect $\boldsymbol{b}_i$ and the error $\boldsymbol{\varepsilon}_i$ were independent, and all data between different subjects were also assumed to be independent. $\boldsymbol{Y}_i$ marginally follows

$MVN(\boldsymbol{X}_i\boldsymbol{\beta},\ \boldsymbol{V}_i)$, where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i$. In the MMRM method, the variations explained by random effects were included as part of the marginal covariance matrix $\boldsymbol{V}_i$ rather than being explicitly modelled as the random effects [5]. An unstructured covariance matrix is often preferred as the structure of $\boldsymbol{V}_i$ because no assumptions are made on the covariance structure [13]. For the statistical inferences of regression parameters, the restricted maximum likelihood (REML) method [10] has been routinely used in practice. In addition, although missing is a common problem in longitudinal clinical trials, validity of the inferences is assured under the missing at random (MAR) mechanism because MMRM adopts the likelihood-based methods [27].

3. Improved Inference Methods

3.1. Likelihood Ratio (LR) test

We considered the testing problem for individual regression coefficients of MMRM, which corresponded to the primary analysis of longitudinal clinical trials. Without loss of generality, we considered a testing problem of the 1st component of the regression coefficients $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \ldots,\ \beta_p)^T$,

$$\text{Null hypothesis } H_0\text{: } \beta_1 = \beta_1^{null}$$

$$\text{Alternative hypothesis } H_1 : \beta_1 \neq \beta_1^{null}$$

Let $\boldsymbol{\beta}_c = (\beta_2,\ \beta_3,\ \ldots,\ \beta_p)^T$ and let $\boldsymbol{v}$ a parameter vector composed of the components of marginal variance–covariance matrix $\boldsymbol{V}_i$ ($i = 1, 2, \ldots, N$).

To develop the improved inference methods, we first introduced the LR test for MMRM. The LR test statistic for the hypothesis test above was

$$T\left(\beta_1^{null}\right) = -2\left\{l\left(\beta_1^{null},\ \widetilde{\boldsymbol{\beta}}_c,\ \widetilde{\boldsymbol{v}}\right) - l\left(\hat{\beta}_1, \hat{\boldsymbol{\beta}}_c,\ \hat{\boldsymbol{v}}\right)\right\}, \tag{2}$$

where $\left(\hat{\beta}_1,\ \hat{\boldsymbol{\beta}}_c,\ \hat{\boldsymbol{v}}\right)$ was the maximum likelihood (ML) estimates of $(\beta_1,\ \boldsymbol{\beta}_c,\ \boldsymbol{v})$ and $\left(\widetilde{\boldsymbol{\beta}}_c,\ \widetilde{\boldsymbol{v}}\right)$ was the constrained ML estimates under the null hypothesis. Also, $l(\beta_1,\ \boldsymbol{\beta}_c,\ \boldsymbol{v})$ was the log-likelihood function for MMRM,

$$l(\beta_1,\ \boldsymbol{\beta}_c,\ \boldsymbol{v}) = -\frac{1}{2}\sum_{i=1}^{N}\left\{\log|\boldsymbol{V}_i| + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) + n_i\log 2\pi\right\} \tag{3}$$

The ML and the constrained ML estimates were computed by using this log-likelihood function. Asymptotically, the LR test statistic $T\left(\beta_1^{null}\right)$ followed the chi-squared distribution with 1 degree of freedom under the null hypothesis [28].

3.2. Bartlett-Type Adjustment by Bootstrap Resampling

Conventionally, it is widely known that the large sample approximation of the LR test statistic $T\left(\beta_1^{null}\right)$ to the chi-squared distribution is not accurate under small sample settings [23]. To improve the approximations, several higher order approaches have been developed and the Bartlett correction [15] is one of the effective solutions. The Bartlett correction is a correction method for the LR test statistic that aims to improve the approximation to the reference chi-square distribution dividing by a correction term. The adjustment term is an estimate of the first moment of the null distribution of the LR test statistic $\xi = E\left[T\left(\beta_1^{null}\right)\right]$, and the corrected LR test statistic is given by $T^*\left(\beta_1^{null}\right) = T\left(\beta_1^{null}\right)/\hat{\xi}$. Intuitively, if the estimate $\hat{\xi}$ is accurate, the null distribution of the corrected statistic approaches the chi-squared distribution. Theoretically, Barndorff-Nielsen and Hall [29] showed that the Bartlett correction reduces the error of the chi-squared approximation from $O\left(N^{-1}\right)$ to $O\left(N^{-2}\right)$.

In this study, we proposed a practical procedure to apply the Bartlett correction to MMRM effectively for incomplete longitudinal data under small sample size. Many previous studies attempted to obtain analytical forms of the Bartlett correction term by analytical methods [14,30,31]. However, analytical forms of the correction term were not necessarily obtainable when complicated models were assumed and missing data was involved. As an alternative effective approach, Rocke [32] proposed to use a resampling approach, which adopted the parametric bootstrap method to estimate the Bartlett correction term $\hat{\xi}$. Here, we proposed to apply this resampling approach to improve the inferences of MMRM. The resampling approach possibly involved computational burdens, but it had an advantage in that it could be implemented using generic algorithms regardless of the complexity of regression model and covariance structure. The resampling based procedure was formulated as the following Algorithm 1.

Algorithm 1 Bartlett correction using bootstrap resampling technique.

(1) For the MMRM model, compute the constrained ML estimates $\left\{\widetilde{\boldsymbol{\beta}}_c,\ \widetilde{\boldsymbol{v}}\right\}$ under $\beta_1 = \beta_1^{null}$.

(2) Resample $\boldsymbol{Y}_1^{(b)},\ \boldsymbol{Y}_2^{(b)},\ \ldots,\ \boldsymbol{Y}_N^{(b)}$ from the estimated null distribution of the MMRM model with the parameters substituted with $\left\{\beta_1^{null}, \widetilde{\boldsymbol{\beta}}_c,\ \widetilde{\boldsymbol{v}}\right\}$ via parametric bootstrap with reflecting missing patterns of $\boldsymbol{Y}_1,\ \boldsymbol{Y}_2,\ \ldots,\ \boldsymbol{Y}_N$ (i.e., let $\boldsymbol{Y}_i^{(b)}$ have the length of n_i vector), B times ($b = 1,\ 2, \ldots,\ B$).

(3) Compute the ML estimates $\left\{\hat{\beta}_1^{(b)},\ \hat{\boldsymbol{\beta}}_c^{(b)},\ \hat{\boldsymbol{v}}^{(b)}\right\}$ and the constrained ML estimates $\left\{\widetilde{\boldsymbol{\beta}}_c^{(b)},\ \widetilde{\boldsymbol{v}}^{(b)}\right\}$ under the null hypothesis for the bth bootstrap sample $\boldsymbol{Y}_1^{(b)},\ \boldsymbol{Y}_2^{(b)},\ \ldots,\ \boldsymbol{Y}_N^{(b)}$. Replicate it for all B bootstrap samples ($b = 1,\ 2,\ \ldots,\ B$).

(4) Compute the LR test statistics for all B bootstrap estimates,

$$T^{(b)}\left(\beta_1^{null}\right) = -2\left\{l\left(\beta_1^{null}, \widetilde{\boldsymbol{\beta}}_c^{(b)},\ \widetilde{\boldsymbol{v}}^{(b)}\right) - l\left(\hat{\beta}_1^{(b)},\ \hat{\boldsymbol{\beta}}_c^{(b)},\ \hat{\boldsymbol{v}}^{(b)}\right)\right\}, \tag{4}$$

and calculate a bootstrap estimate of ξ,

$$\hat{\xi} = \frac{1}{B}\sum_{b=1}^{B} T^{(b)}\left(\beta_1^{null}\right). \tag{5}$$

(5) We can obtain the corrected LR test statistic,

$$T_{BS}^{*}\left(\beta_1^{null}\right) = T\left(\beta_1^{null}\right)/\hat{\xi}. \tag{6}$$

A statistical test using the corrected LR test statistic $T_{BS}^{*}\left(\beta_1^{null}\right)$ could be performed by using a chi-square distribution with 1 degree of freedom as the reference distribution.

Also, the corresponding confidence interval of β_1 could be constructed by a set of β_1^{null} that fulfill the following inequality,

$$T_{BS}^{*}\left(\beta_1^{null}\right) \leq \chi_{1,\,\alpha}^2, \tag{7}$$

where $\chi_{1,\,\alpha}^2$ is the upper αth quantile of the chi-square distribution with 1 degree of freedom. Note that while it was technically possible to apply a nonparametric bootstrap method, the parametric bootstrap method would be preferred to estimate the Bartlett correction term ξ under small sample settings, because the bootstrap distributions might have been too discrete [33,34].

3.3. Monte Carlo Test Using an Estimated Null Distribution by Bootstrap

Using the parametric bootstrap, we could also estimate the null distribution of the test statistic directly via the Monte Carlo technique [35]. We could construct an estimate of the null distribution of $T\left(\beta_1^{null}\right)$ if we resampled a large number of LR test statistics under the null hypothesis using

a parametric bootstrap technique. This method used the Monte Carlo estimate of the null distribution as the reference distribution of LR test, instead of the chi-squared distribution. This approach would be an alternative to the former proposed method that had the same advantages for the inferences of small sample settings.

With processes 1–4 of Algorithm 1, we had the bootstrap LR test statistics $T^{(b)}\left(\beta_1^{null}\right)$ $(b = 1, 2, \ldots, B)$. The Monte Carlo estimate of the null distribution was obtained as the empirical distribution of $T\left(\beta_1^{null}\right)$. Also, the bootstrap-based critical value of the nominal α level $(0 < \alpha < 1)$ corresponded to the upper αth quantile of the empirical distribution function. The Monte Carlo test can be constructed by the following Algorithm 2.

Algorithm 2 Bootstrap-based adjustment of LR test.

(1) Conduct processes 1–4 of Algorithm 1.

(2) Calculate the p-value by the following formula [24].

$$p = \frac{1}{B+1}\left\{1 + \sum_{b=1}^{B} I\left[T^{(b)}\left(\beta_1^{null}\right) > T\left(\beta_1^{null}\right)\right]\right\} \tag{8}$$

Here, $I(x)$ is an indicator function, and it returns 1 if x is true and 0 otherwise.

Also, $100 \times (1-\alpha)\%$ confidence intervals can be constructed with a set of β_1^{null} that fulfill [36],

$$T\left(\beta_1^{null}\right) \leq \hat{q}_{bs,\,(1-\alpha)}.$$

where $\hat{q}_{bs,(1-\alpha)}$ for the upper αth quantile of the estimated null distribution. According to Rocke [32], more than 1000 resamplings were recommended when estimating the tail of a distribution, such as the upper αth quantile of the null distribution.

4. Simulation Studies

4.1. Design and Setting

We conducted a series of simulation studies to assess the performances of the two methods, the Bartlett-type correction for LR test statistic-based test (LR_{Bart}) and the bootstrap adjustment test for LR test statistic based test (LR_{Boot}) under practical situations of longitudinal small clinical trials. We compared the effectiveness of these methods with the conventional ordinary LR test and Kenward-Roger (KR) method [11], which is the current standard inference method in MMRM analyses. We considered the same scenarios of the simulation studies of Gosho et al. [13], which conducted extensive simulations to evaluate the performances of MMRM for longitudinal clinical trials. We supposed two group comparative longitudinal clinical trials (e.g., active drug group vs. placebo group) and the number of post baseline visits (n) to be 7. The total number of subjects was determined as $N = 20$ (i.e., 10 subjects per group, respectively). The outcome variables Y_{it} (t = 1., 2., ... 7) were generated from the following model,

$$Y_{it} = mean_{it} + subject_i + error_{it} \tag{10}$$

where $mean_{it}$ was a fixed effect, $subject_i$ was a subject effect, and $error_{it}$ was a random error (i = 1, 2, ... , N; t = 1, 2, ... , 7). The mean values of Y_{it} assumed the four scenarios illustrated in Figure 1. Here, we were interested in evaluating the mean difference between the two groups at the final (7th) time point. Scenarios 1 and 2 corresponded to the null hypothesis that the mean values of the outcome variables were the same between the two groups at the final time point. By contrast, in scenarios 3 and

4, the mean value of the treatment efficacy at the final time point differs between the two groups and corresponded to the alternative hypothesis.

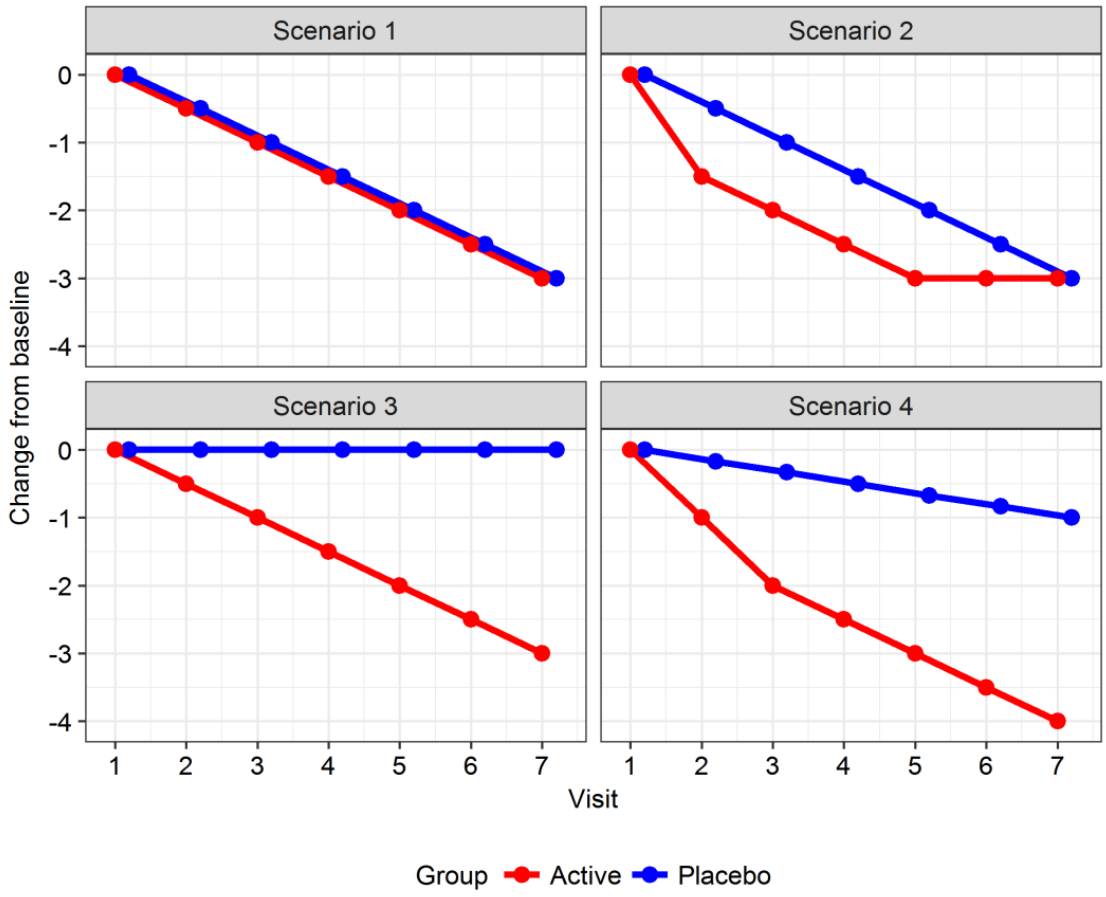

Figure 1. Mean parameter settings under the four scenarios in the simulation studies.

4.2. Correlation Structures

For the variance–covariance structure of the error terms, a first order heterogeneous autoregressive (ARH (1)) structure was adopted, of which $(t,\ t')$ element is defined as $\sigma_t\sigma_{t'}\rho^{|t'-t|}$ where σ_t and $\sigma_{t'}$ are the standard deviances of tth and t'th time points and ρ is the correlation coefficient between the two points. Following Gosho et al. [13], the diagonal elements σ_t^2 was set to $9\{1+3(t-1)/6\}$ and the correlation coefficient ρ was set to 0.7. Also, the subject effect was generated by $N(0,\ 3^2)$.

4.3. Missing-data Mechanism

In this simulation, we considered two missing-data mechanisms, missing completely at random (MCAR) and missing at random (MAR). Only the monotone missing was assumed, i.e., once missingness occurred, all outcome values after the time point were missing for the corresponding individual. We denoted the probability of missingness of Y_{it} as λ_{it}. The missingness probability λ_{it} was assumed to follow the logistic regression model,

$$\text{logit}(\lambda_{it}) = \gamma_0 + \gamma_1 Y_{i,\,t-1},\ t = 2,\ 3, \ldots,\ 7. \tag{11}$$

The regression coefficients of the logistic regression model for the missingness probability were defined as $\gamma_1 = 0$ for MCAR and $\gamma_1 = -1$ for MAR. Table 1 summarizes the missing-data mechanisms and the coefficients of the logistic regression model that had a defined dropout rate for each treatment

group at the final time point. The total dropout rates for the two groups were set to 0%, 20% or 40% for the four scenarios.

4.4. Analysis Methods

The simulated data were analysed using four methods as mentioned above. We adopted the standard MMRM that included a group variable and time variables as dummy variables and the group-by-time interactions in the regression function. An unstructured covariance structure was adopted for the covariance structure model for the outcome variables. Parametric bootstraps for the proposed two methods were performed via 3000 resamplings. The results concerning convergence of MMRM analyses are reported in the Appendix A. The numbers of simulations were 1000 for all scenarios. All computations were performed by SAS ver. 9.4. Also, the significance levels were set to be 0.05.

Table 1. Dropout rates at the final time point and parameter settings.

Scenario	Missing Mechanism	Overall Dropout (%)	Dropout (%)		γ_0	
			Placebo	Active	Placebo	Active
1	MCAR	20	20	20	3.2	3.2
		40	40	40	2.4	2.4
	MAR	20	20	20	7.1	7.1
		40	40	40	4.2	4.2
2	MCAR	20	22	18	3.1	3.4
		40	44	36	2.3	2.6
	MAR	20	22	18	6.6	6.6
		40	44	36	3.7	3.7
3	MCAR	20	24	16	3.0	3.5
		40	46	34	2.2	2.6
	MAR	20	24	16	7.8	7.8
		40	46	34	4.8	4.8
4	MCAR	20	24	16	3.0	3.5
		40	46	34	2.2	2.6
	MAR	20	24	16	7.2	7.2
		40	46	34	4.2	4.2

4.5. Results

Figure 2 shows the type I error rates for scenarios 1 and 2 under $N = 20$ (i.e., 10 subjects per group). The blue dashed lines correspond to the 95% intervals of the Monte Carlo errors. At first, the type I error rates of LR test increased greatly from 5% as the dropout rate increased. In scenario 1 with a 40% dropout rate, the type I error rates of LR was 11.3% under MCAR, and 9.8% under MAR, respectively. Besides, the type I error rates of LR_{Bart} and LR_{Boot} were maintained at 5% irrespective of the missing-data mechanism and dropout rate. For example, the type I error rates under scenario 1 with a 40% dropout rate under MAR were 5.5%, 5.6% for LR_{Bart} and LR_{Boot}, respectively. On the other hand, the type I error rates of the KR method were not maintained at around 5% under MAR and were too conservative. Under MAR with a 40% dropout rate, the type I error rates of the KR method were 3.6% and 3.5% for scenarios 1 and 2, respectively. Besides, under MCAR scenarios, the type I error rates of the KR method were maintained at around 5%. The convergence rates of these methods were not significantly different (see Appendix A). Note that the type I errors for the LR, LR_{Bart} and LR_{Boot} were inflated under scenario 1 with dropout rate 40%, but they fell within the ranges of Monte Carlo errors. The results of the convergence for scenarios 1 and 2 appears in the appendix section as Figure A1.

Figure 3 shows the powers in scenarios 3 and 4 for $N = 20$ (i.e., 10 subjects per group). At first, the powers of LR was higher than those of other methods, ranging from approximately 14% to 20%

depending on the dropout rate. However, since the type I error rates of LR were not maintained at 5% under scenarios 1 and 2, it should be considered to have liberal properties in general. In all three methods other than LR test, the powers decreased as the dropout rate increased, due to the reduction of available statistical information. In scenario 3, with 10 subjects per group and a 40% dropout rate under MAR, the powers of LR_{Bart}, LR_{Boot} and KR were 7.1%, 7.4% and 6.5%, respectively. The overall trends concerning powers of the four methods agreed with the rejection rates under scenario 1 and 2, although they depended on the sample size and effect sizes.

The results of the convergence for scenarios 3 and 4 appears in the appendix section as Figure A2.

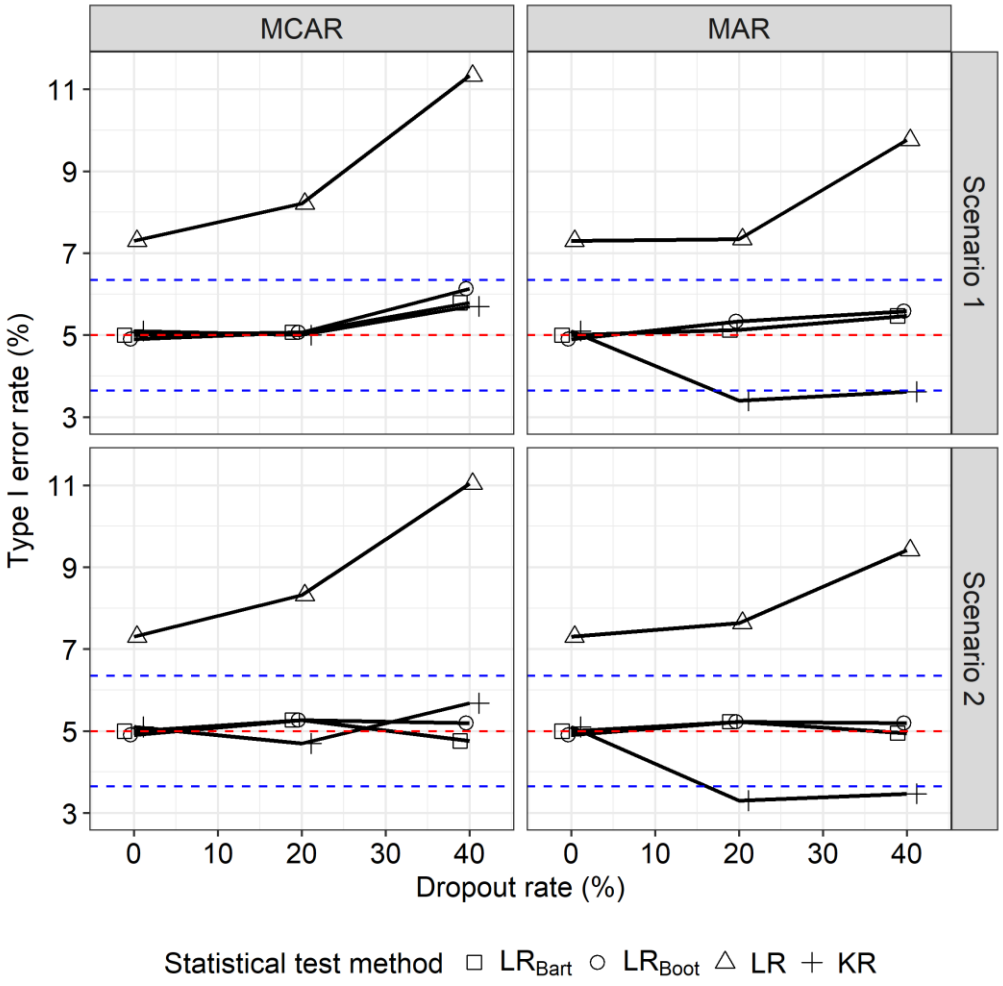

Figure 2. Type I error rates under scenarios 1 and 2 in the simulation study. Red dashed line, 5%; Blue dashed line, 95% intervals of Monte Carlo error with 1000 iterations.

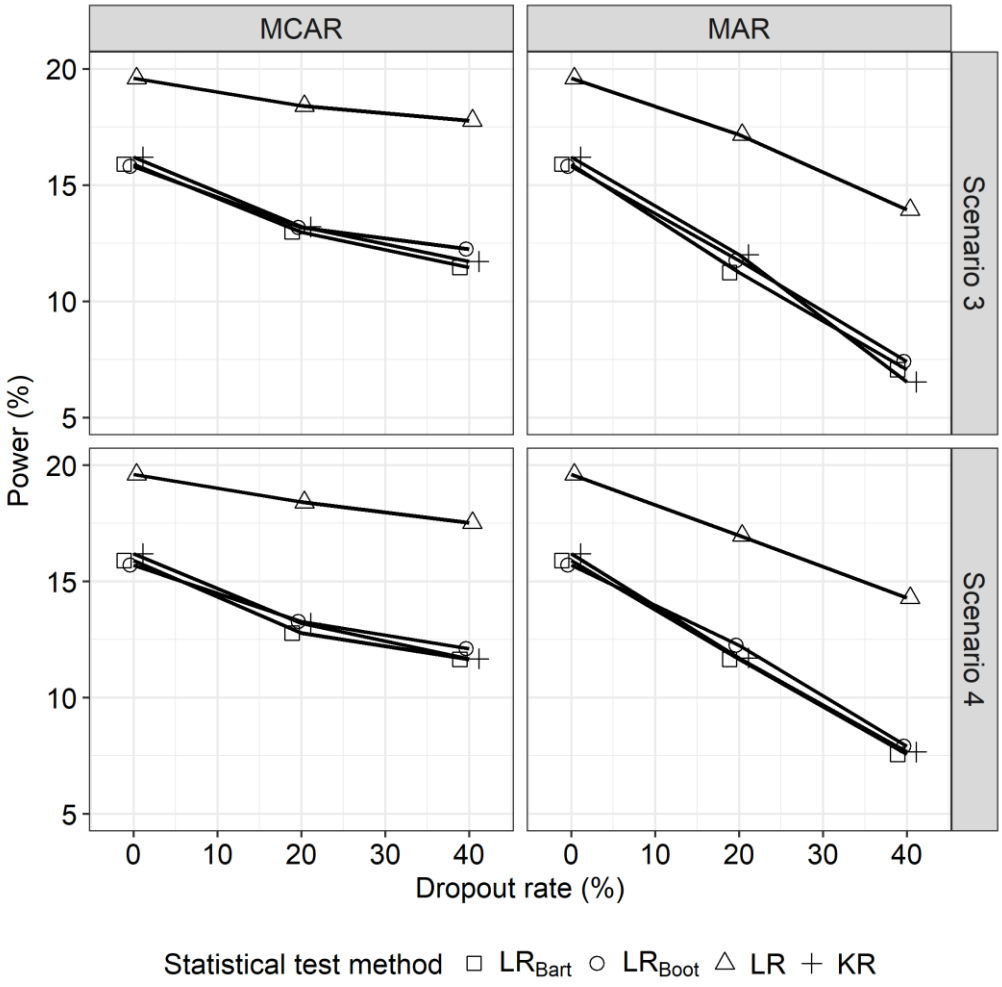

Figure 3. Power under scenarios 3 and 4 in the simulation study.

5. Application: Postnatal Depression Trial

Postnatal depression is commonly treated with antidepressants and counselling. Transdermal administration of estrogen has also been shown to be effective, and Gregoire et al [26] conducted a double-blind, placebo-controlled study in 61 women within 3 months of giving birth [26,37]. Although the study planned to enroll 100 subjects, eventually 61 women were randomly assigned to the placebo group (27 subjects) or the estrogen group (34 subjects). The women were assessed twice prior to treatment and then monthly for 6 months after treatment using the Edinburgh postnatal depression scale (EPDS), with higher scores indicating more severe depression. Approximately 37.0% (10/27) of subjects in the placebo group and 17.6% (6/34) of subjects in the estrogen group had missing EPDS scores at the final time point. All data had monotone missing patterns.

The baseline EPDS score was defined as the average of the scores at the 1st and 2nd months in this study. The outcome variables were measured on the visits between the 3rd and 8th months. We considered analysing this longitudinal dataset using MMRM and the following regression function model,

$$E[Y_{it}|G_i, t_{it}] = \beta_0 + \beta_1 G_i + \beta_{2t} t_{it} + \beta_{3t} G_i \times t_{it} \tag{12}$$

where Y_{it} denotes the EPDS score for the participant i $(i = 1, 2, \ldots, 61)$ on the tth occasion $(t = 1, 2, \ldots, 8)$. G_i was a dummy variable that equals 1 if the participant i belongs to the estrogen group and equals 0, otherwise. For the covariance structure of the outcome variables, we assumed the unstructured structure. Here, our primary subject of interest was the evaluation of the mean

difference of outcome variables on the final time point. In addition, we considered a subgroup analysis for a group of participants with clinically severe depressive symptoms that was defined as baseline EPDS > 21 [38]. There were 30 participants in subgroup (15 participants in both placebo and estrogen groups). At the final month, the proportions of dropout were 40.0% (6/15) and 20.0% (3/15) for placebo and estrogen group, respectively.

At baseline, the mean EPDS scores of the placebo and estrogen groups were 21.26 (3.11) and 21.59 (3.06), respectively. Table 2 summarizes the mean difference estimates of EPDS scores at the final month, as well as their 95% confidence intervals and p-values by the conventional and proposed methods. We added the t-test on the single point analysis at the final month in these analyses as a reference. The numbers of resampling for the proposed methods were set to be 3000.

Table 2. Results of the analyses of the postnatal depression trial: Inferences of the mean difference at the final month.

	Whole population (N = 61)		**Subgroup (N = 30)**	
	Estimate [95% CI]	***p*-Value**	**Estimate [95% CI]**	***p*-Value**
LR_{Bart}	4.34 [1.67, 7.66]	0.0031	3.93 [−0.19, 10.24]	0.0586
LR_{Boot}	4.34 [1.70, 7.64]	0.0050	3.93 [−0.16, 10.21]	0.0583
LR	4.34 [1.81, 7.52]	0.0019	3.93 [0.29, 9.74]	0.0383
KR	4.34 [1.45, 7.23]	0.0040	3.93 [−1.23, 9.09]	0.1288
t-test	4.36 [1.43, 7.29]	0.0045	3.17 [−2.24, 8.57]	0.2349

In the whole population analysis, all of the five methods showed significant differences and provided similar estimates. However, the p-value of LR test was a bit smaller than the proposed methods, and that of KR was a bit larger. These trends might corresponds to the liberal and conservative properties of these methods. These trends became clearer for the subgroup analysis for the participants with severe symptoms. Only the LR test showed significant difference, and the other four methods provided non-significant results. The p-values of LR_{Bart} and LR_{Boot} were 0.0586 and 0.0583, but that of KR was 0.1288. These results might reflect the conservative property of KR, and it was possibly improved by the proposed two methods. In addition, the t-test for the subgroup analysis provided a larger p-value (0.2349) with a considerably smaller estimate and larger standard error. Previous numerical evidence (e.g., Ashbeck and Bell [39]) showed possible bias and information reduction of the single time point analysis by t-test, and this result might correspond to this evidence. With LR_{Bart} and LR_{Boot}, the computational times were 55 and 38 minutes for whole population and subgroup, respectively (we used a general laptop computer with an Intel (R) Core (TM) i7-6500U and SAS 9.4). The computational times would be dramatically improved by applying parallel computation techniques.

Figure 4 shows the histogram of the empirical distribution of the LR test statistics resampled by the parametric bootstrap method under the null effect hypothesis. The mean values of the empirical distribution were designated by the vertical blue dashed line in each histogram and were 1.09 and 1.20 in the whole population and the subgroup, respectively. In addition, the 95th percentiles of the empirical distribution were 4.14 and 4.55 for the whole population and the subgroup, respectively. If the chi-squared approximations are accurate, the means and 95th percentiles of the null distribution were expected to be 1.0 and 3.84. These results would show that the distribution of the LR test statistic in incomplete longitudinal data with a small sample size shifted and adequate corrections were needed. The proposed methods would improve the approximations and enable improvements of the inferences as shown in the simulations.

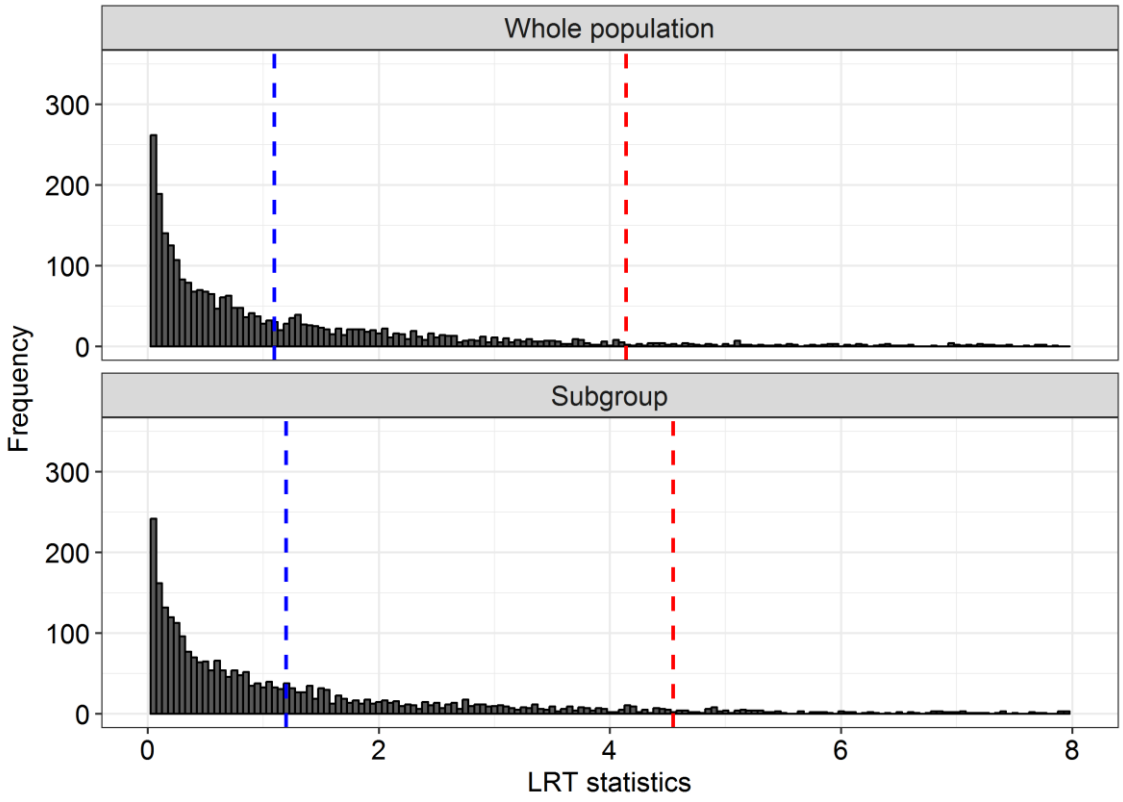

Figure 4. Empirical distribution by parametric bootstrap in postnatal depression trial. Blue dashed line, mean of empirical distribution (1.09 and 1.20 for whole population and subgroup, respectively); Red dashed line, 95th percentile of empirical distribution (4.14 and 4.55 for whole population and subgroup, respectively).

6. Discussion

MMRM with the KR method has been widely applied as a standard analysis method for longitudinal clinical trials. If a sufficient number of samples are available, there are no problems using statistical tests and confidence intervals based on large sample theory. However, the asymptotic approximations cannot be appropriate in cases with small sample sizes. In addition, most clinical trials involve missing data. As methods to improve validity of the statistical inferences, we proposed resampling-based approaches. Throughout the simulations and real data applications, we demonstrated the effectiveness of the proposed methods compared with existing standard methods.

In the simulation experiments, the KR method and our proposed methods maintained almost the same type I error rate under MCAR, which was close to 5%. However, under MAR scenarios with large dropout rates, the KR method had conservative type I error rate compared with our proposed methods. Our proposed methods might have an advantage even if the missing-data mechanism is MAR compared with KR. In addition, it should be noted that Algorithm 1 uses bootstrap samples to estimate the mean of the null distribution, whereas Algorithm 2 uses them to estimate a quantile of the null distribution. In general, the latter is a more unstable quantity for Monte Carlo inferences and thus requires a larger number of resamplings in general [23]. In our simulation studies, 3000 resamplings were performed for both the LR_{Bart} and LR_{Boot} methods, and we obtained similar results. The number 3000 was determined considering Monte Carlo errors, and they would be sufficient. Although these might require large computation burdens, they would not be so problematic under a modern computational environment, in which parallel computations are available for standard statistical software.

In addition, another possible effective approach to be considered in future research might be the Bayesian approach. The Bayesian method might also accommodate small sample sizes, if the choices

of the prior distributions are appropriate. The advantages and potential drawbacks are discussed in Van De Schoot et al. [40]. Also, another concern is extensions to multi-parameter inferences. However, the proposed methods are quite general methods and could be straightforwardly extended to the multi-parameter inferences.

The effectiveness of our proposed two resampling approaches for MMRM were clearly shown through simulation studies and real data applications. To assure scientific validity in developments of new drug and health technology, accurate statistical inference methods are essential tools. The proposed methods can be applied as effective options in statistical analyses for small and incomplete longitudinal clinical trials.

Author Contributions: Y.U. and H.N. conceived and designed this study. Y.U., H.N., K.M. and M.G. conducted developments of the methods. Y.U. conducted simulation and real data analyses. Y.U. and H.N. interpreted the results, and drafted the manuscript. All authors approved its final version.

Funding: This study was partly supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant numbers: JP15K15954, JP15H03390).

Conflicts of Interest: Y.U. is employee of Janssen Pharmaceutical K.K. H.N., K.M., and M.G. declare that they have no competing interests.

Appendix A

The results of convergence proportion with $N = 20$ are presented as figures below. Figure A1 displays the results in scenario 1 and 2 and Figure A2 displays the results in scenario 3 and 4, respectively.

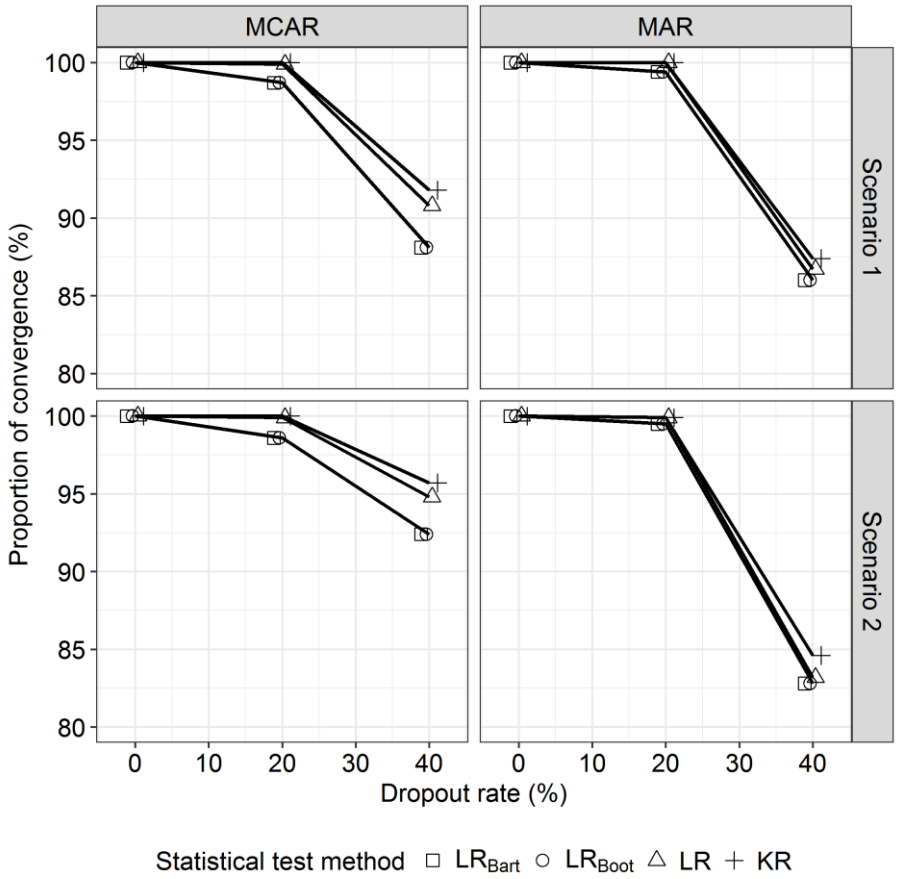

Figure A1. Convergence proportion (%) in scenarios 1 and 2 under MCAR and MAR assuming 10 subjects per group in the simulation study.

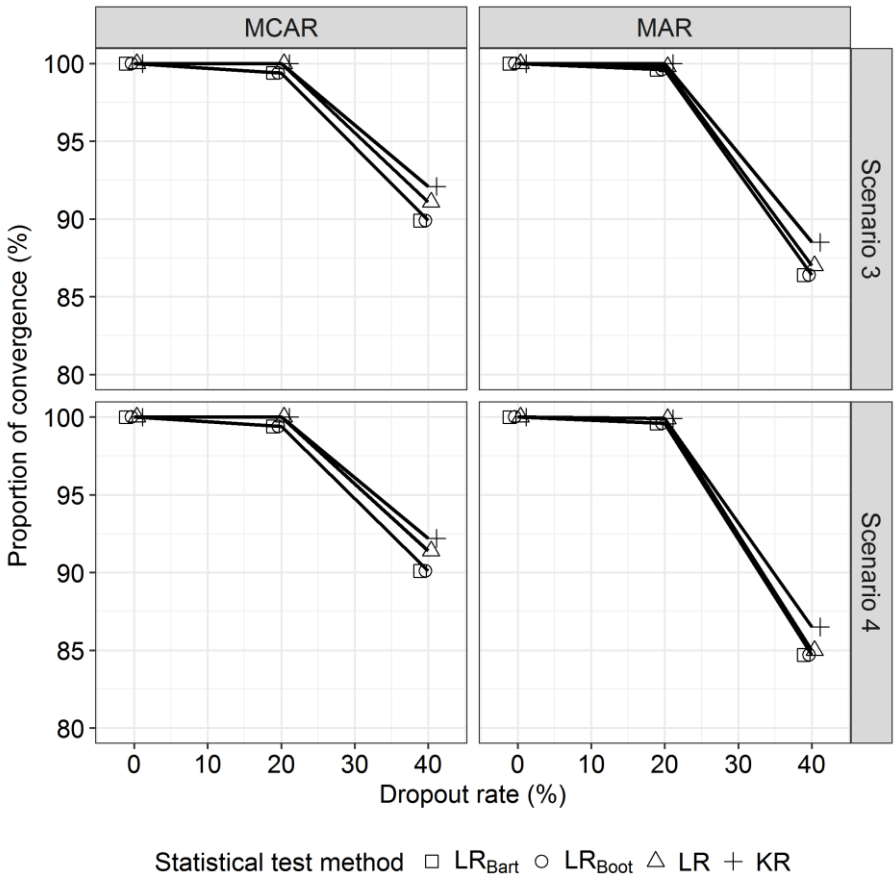

Figure A2. Convergence proportion (%) in scenarios 3 and 4 under MCAR and MAR assuming 10 subjects per group in the simulation study.

References

1. European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials. 2010. Available online: https://www.ema.europa.eu/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf (accessed on 11 January 2019).
2. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*; National Academies Press: Washington, DC, USA, 2010.
3. Ratitch, B.; O'Kelly, M.; Tosiello, R. Missing data in clinical trials: From clinical assumptions to statistical analysis using pattern mixture models. *Pharm. Stat.* **2013**, *12*, 337–347. [CrossRef]
4. Mallinckrodt, C.H.; Sanger, T.M.; Dubé, S.; DeBrota, D.J.; Molenberghs, G.; Carroll, R.J.; Potter, W.Z.; Tollefson, G.D. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol. Psychiatry* **2003**, *53*, 754–760. [CrossRef]
5. Mallinckrodt, C.H.; Lane, P.W.; Schnell, D.; Peng, Y.; Mancuso, J.P. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inf. J.* **2008**, *42*, 303–319. [CrossRef]
6. Mallinckrodt, C.H.; Watkin, J.G.; Molenberghs, G.; Carroll, R.J. Choice of the primary analysis in longitudinal clinical trials. *Pharm. Stat.* **2004**, *3*, 161–169. [CrossRef]
7. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974. [CrossRef]
8. Fitzmaurice, G.M.; Laird, N.M.; Ware, J.H. *Applied Longitudinal Analysis*; John Wiley & Sons: New York, NY, USA, 2012.
9. Diggle, P.J.; Heagerty, P.J.; Liang, K.-Y.; Zeger, S. *Analysis of Longitudinal Data*; Oxford University Press: Oxford, UK, 2002.
10. Patterson, H.D.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [CrossRef]

11. Kenward, M.G.; Roger, J.H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **1997**, *53*, 983–997. [CrossRef]
12. Schluchter, M.D.; Elashoff, J.T. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *J. Stat. Comput. Simul.* **1990**, *37*, 69–87. [CrossRef]
13. Gosho, M.; Hirakawa, A.; Noma, H.; Maruo, K.; Sato, Y. Comparison of bias-corrected covariance estimators for MMRM analysis in longitudinal data with dropouts. *Stat. Methods Med. Res.* **2017**, *26*, 2389–2406. [CrossRef] [PubMed]
14. Zucker, D.M.; Lieberman, O.; Manor, O. Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *J. R. Stat. Soc. B* **2000**, *62*, 827–838. [CrossRef]
15. Bartlett, M.S. Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A* **1937**, *160*, 268–282.
16. Cox, D.R.; Reid, N. Parameter orthogonality and approximate conditional Inference. *J. R. Stat. Soc. B* **1987**, *49*, 1–39. [CrossRef]
17. Lyons, B.; Peters, D. Applying skovgaard's modified directed likelihood statistic to mixed linear models. *J. Stat. Comput. Simul.* **2000**, *65*, 225–242. [CrossRef]
18. Guolo, A.; Brazzale, A.R.; Salvan, A. Improved inference on a scalar fixed effect of interest in nonlinear mixed-effects models. *Comput. Stat. Data Anal.* **2006**, *51*, 1602–1613. [CrossRef]
19. Skovgaard, I.M. An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **1996**, *2*, 145–165. [CrossRef]
20. Barndorff-Nielsen, O. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **1983**, *70*, 343–365. [CrossRef]
21. Stein, M.C.; da Silva, M.F.; Duczmal, L.H. Alternatives to the usual likelihood ratio test in mixed linear models. *Comput. Stat. Data Anal.* **2014**, *69*, 184–197. [CrossRef]
22. Severini, T.A. An approximation to the modified profile likelihood function. *Biometrika* **1998**, *85*, 403–411. [CrossRef]
23. Cordeiro, G.M.; Cribari-Neto, F. *An Introduction to Bartlett Correction and Bias Reduction*; Springer: New York, NY, USA, 2014.
24. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997.
25. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: New York, NY, USA, 1994.
26. Gregoire, A.J.; Kumar, R.; Everitt, B.; Henderson, A.F.; Studd, J.W. Transdermal oestrogen for treatment of severe postnatal depression. *Lancet* **1996**, *347*, 930–933. [CrossRef]
27. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, NY, USA, 2000.
28. Cox, D.R.; Hinkley, D.V. *Theoretical Statistics*; Chapman & Hall: London, UK, 1974.
29. Barndorff-Nielsen, O.E.; Hall, P. On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **1988**, *75*, 374–378. [CrossRef]
30. Lawley, D.N. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **1956**, *43*, 295–303. [CrossRef]
31. Melo, T.F.N.; Ferrari, S.L.P.; Cribari-Neto, F. Improved testing inference in mixed linear models. *Comput. Stat. Data Anal.* **2009**, *53*, 2573–2582. [CrossRef]
32. Rocke, D.M. Bootstrap Bartlett adjustment in seemingly unrelated regression. *J. Am. Stat. Assoc.* **1989**, *84*, 598–601. [CrossRef]
33. Noma, H.; Nagashima, K.; Maruo, K.; Gosho, M.; Furukawa, T.A. Bartlett-type corrections and bootstrap adjustments of likelihood-based inference methods for network meta-analysis. *Stat. Med.* **2018**, *37*, 1178–1190. [CrossRef] [PubMed]
34. Wehrens, R.; Putter, H.; Buydens, L.M.C. The bootstrap: A tutorial. *Chemometr. Intell. Lab.* **2000**, *54*, 35–52. [CrossRef]
35. Efron, B. *Bootstrap Methods: Another Look at the Jackknife*; Springer: New York, NY, USA, 1992.
36. Zeng, Q.; Davidian, M. Bootstrap-adjusted calibration confidence intervals for immunoassay. *J. Am. Stat. Assoc.* **1997**, *92*, 278–290. [CrossRef]
37. Everitt, B.S. *A Handbook of Statistical Analyses Using S-Plus*; CRC Press: New York, NY, USA, 2001.
38. McCabe-Beane, J.E.; Segre, L.S.; Perkhounkova, Y.; Stuart, S.; O'Hara, M.W. The identification of severity ranges for the Edinburgh Postnatal Depression Scale. *J. Reprod. Infant Psychol.* **2016**, *34*, 293–303. [CrossRef]

39. Ashbeck, E.L.; Bell, M.L. Single time point comparisons in longitudinal randomized controlled trials: Power and bias in the presence of missing data. *BMC Med. Res. Methodol.* **2016**, *16*, 43. [CrossRef] [PubMed]
40. Van De Schoot, R.; Broere, J.J.; Perryck, K.H.; Zondervan-Zwijnenburg, M.; Van Loey, N.E. Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *Eur. J. Psychotraumatol.* **2015**, *6*, 25216. [CrossRef] [PubMed]