Responses to Comments from Reviewer 1

Main suggestions:

Comment 1:

Part 1: Page 11, lines 31-33: The authors state that "for a large number of daily scores in long-term clinical trials, it is computationally challenging to impute the intermittent PROs by MCMC and the monotone missing with MI". My understanding is that in the context of this paper, both the referred "MCMC" and "MI" methods are based on drawing from multivariate normal distributions, where MCMC method is used to draw imputed values, so the distinction in this statement is not clear to me. Does this mean a MAR assumption is used for imputing missing PROs in patients who only have intermittent missing values and a MNAR assumption for those who drop out?

Response: Thank you for the comments to help us to clarify this point. Yes, you are correct that we assumed MAR for the intermittent missing and impute those missing values to make the dataset monotone, then use MNAR to impute monotone missing values (J2R or CIR). We have clarified this point in the first paragraph in Section 3, where we specified "As discussed above, the intermittent missing can be assumed to be MAR and imputed using Markov Chain Monte Carlo (MCMC). Therefore, when handling missing data in practice, we first impute all intermittent missing data by MCMC methods until the missingness becomes monotone. Throughout the following sections, we focus on the setting of drop-out missing for the methodological development, since it requires special imputation model to handle the MNAR situation".

Part 2: When a patient has intermittent missing values earlier in the trial and subsequently drop out completely from the trial, does this imply that a joint distribution for a patient's observed and unobserved values is defined, with a mean vector from their randomised group prior to

dropping out (ie, assuming MAR for intermittent missing values) and a reference-based mean vector post dropping out (assuming MNAR)?

Response: Thank you for your comment. Yes, this is the mechanism for the two types of missingness. Please see the response to Part 1 of this comment.

Comment 2:

Sections 4 and 5: could the authors please include a more detailed description of the procedure for performing MI under MNAR and MAR, at the daily and cycle levels, eg, whether imputation is performed separately by treatment arms, the mean and variance of the treatment arm specific MVN distribution, the priors for the mean and covariance matrix, specification of the joint distributions for the observed and unobserved data for patients who drop out, method for drawing imputations.

Response: Thank you for the comment. For the two-level MI under MNAR, we have added some sentences in the section on Page 10 to clarify the MI process. Please see the added blue texts below.

"Estimates and inferences for J2R and CIR

MI Methods under J2R and CIR: When handling missing data in practice, we first impute all intermittent missing data by MCMC methods until the missingness becomes monotone, as mentioned in Section 3. The estimates and variances by the two-level J2R and CIR can be obtained by multiple imputations (MI) based on the two-level cMMRM model for the daily outcomes (3) with Rubin's rules. Specifically, we first fit the cMMRM model to get means over days within cycles for each treatment group and the covariance matrix, then use formula (7) to impute missing data based on the corresponding J2R and CIR. The imputation is done by treatment group using a Bayesian process with non-informative priors. Each of the completed datasets is then analyzed using the cMMRM model and the results are combined by Rubin's rules."

For the two-level MI under MAR, we have added one sentence in the section on Page 11, please see the added blue texts below.

"4.2 Two-level MAR Methods

MI Methods under MAR: Under MAR, we propose a two-level method to impute the missing daily scores under the defined two-level missing patterns. For the two-level missing pattern, the missing daily scores after dropout in the treatment group will be imputed using the mean profile of the available daily scores from the treatment group on the corresponding days. Specifically, we first fit the cMMRM model to get means over days within cycles for each treatment group and the covariance matrix, then use formula (7) to impute missing data under MAR. The imputation is done by treatment group using a Bayesian process with non-informative priors. Each of the completed datasets is then analyzed using the cMMRM model and the results are combined by Rubin's rules."

For the cycle-level MI under MAR and MNAR, we have added one sentence in the section on Page 13, please see the added blue texts below.

"***Cycle-level MAR Methods***

The estimates and variances by the cycle-level MAR can be obtained by MI and MMRM using the cycle-level model (5). Specifically, we first fit the MMRM model on calculated cycle average scores to get means over cycles for each treatment group and the covariance matrix, then impute missing cycle means under MAR. The imputation is done by treatment group using a Bayesian process with non-informative priors. Each of the completed datasets is then analyzed using the MMRM model and the results are combined by Rubin's rules."

"Cycle-level MNAR Methods: J2R and CIR

MI Methods under J2R and CIR: The MI approaches under MNAR for the missing data at the cycle level are similar to those under the MAR but the imputation will be based on the assumptions of J2R and CIR, and the results are combined by Rubin's rules."

Comment 3:

Sections 4.1, 4.2 and 5: it might be clearer to the readers if further sub-headings are used to distinguish between (i) the derivation of the analytic formulae and (ii) MI (or c/MMRM).

Response: Thank you for your comment. We have addressed your comment by adding underlined sub-headings before the analytic methods and the MI methods, in section 4.1, 4.2. and Section 5.

Comment 4:

Section 6: Generally, I feel that this section could benefit from further details. For a more structured report of the simulation study, the author could consider following the ADEMP framework by Morris et al (DOI: 10.1002/sim.8086), stating more clearly the aims, data-generating mechanisms, estimands, methods, and performance measures.

- To be more specific in the text, please refer to models (3) and (4) under which the PROs data are generated.

  Response: Thank you for the comment. We have added details in the first paragraph in Section 6 about how the daily score are generated according to the 4X7 dimensional multivariate normal distribution (4) for the simulated trial with 4 postbaseline cycles with 7 days in each cycle. Please see the added text in the first paragraph in Section 6 in the revised manuscript.

  In Addition:

- We have added the aims of the simulation in the first paragraph of Section 6. Please see the blue test below:

  "The aims of the simulations are 1) to evaluate the statistical properties of the proposed two-level approaches as compared to the conventional cycle-level analysis; 2) to evaluate the performance of the proposed analytic methods to the MI approach."

- We have added the estimands, methods, and performance measures before the subsection "Results", please see the blue texts:

  "We consider two hypothetical estimands in the simulations. The first is a per-protocol like estimand which assess a hypothetical effect under the assumption that patients who discontinued would stay in the study on their assigned therapy. This estimand is

evaluated using the cMMRM or MMRM under the MAR assumption. The second estimand is to assess the attributable effect from the randomized treatment under the assumption that patients who discontinued would stay in the study without continuing their assigned therapy or taking any other medication (such as rescue medication). This estimand is assessed using the RBI methods such as J2R and CIR.

In this simulation study, we want to measure the treatment difference between the treatment group and the control group using all the MAR and MNAR two-level methods and cycle-level methods, by evaluating the estimates of treatment difference, bias, percent relative absolute bias (|estimate−true value|/true value), standard error (SE) which is the model-based SE averaged over 1000 repetitions, empirical SE, and power. Additionally, the Type I error will be evaluated under the null hypothesis in which the data will be generated with the same cycle means in the treatment and control groups."

- Are the missingness patterns generated with similar % completers in the two arms?

  Response: Thank you for the comment. We assume the percents of completers are the same for each group, and the details about how to analyze the data by the cMMRM and MMRM are also added. The blue texts have been added to the paragraph before the 4 missing scenarios (MAR1, MAR2, MNAR1, MNAR2):

  "We assume the percents of completers are the same for both groups. For the two-level methods, the daily scores are analyzed by the daily cMMRM model (3). For the cycle-level methods, firstly we need to calculate the cycle means based on the rule: the cycle mean is assumed to be missing if the number of days with available scores in the corresponding cycle is <3, and calculated by the average of the available scores in that cycle otherwise; secondly the cycle means are analyzed by the cycle-level MMRM model (5)."

- The selection parameters were the same for MAR1 and MAR2, as well as for MNAR1 and MNAR2 – typos?

  Response: Thank you or the comment. The selection parameters were the same for MAR1 and MAR2, and the same for MNAR1 and MNAR2. The difference between MAR1

and MAR2 is that the proportion of missing in MAR2 is higher; the same is true for the difference between MNAR1 and MNAR2.

- Page 15, lines 23-28: it might be clearer to the readers if this is described as MAR and MNAR methods performed at either the daily level or the cycle level. Again this could be made clearer in sections 4 and 5.

  Response: Thank you for the helpful comment. The details of the MAR methods and MNAR methods have been added to the paragraph at cycle-level and daily level. The details of the MAR methods and MNAR methods at the daily level and cycle-level have also been added to Section 4 and 5, as suggested by the comment 2 above.

- Could the authors please give further details on MI, eg, how many imputations are created, which software package (and if possible, command syntax) is used to perform MI, which method is used for drawing the imputations.

  Response: Thank you for the helpful comment. The details of the MI methods have been added to Paragraph 1 on Page 16, such as the terms of the MMRM/cMMRM model used for MI, and the number of imputations. We implemented these methods in R-codes because there is no existing software package for these methods.

- Please consider using further sub-headings to separate the method from the results.

  Response: Thank you for the good suggestion. Sub-headings of methods and results have been added to the section of simulations.

- Since bias is one of the performance measures (alongside SE, power, type 1 error), please could the authors report the bias or percentage bias instead of the mean. I assume the SE is model-based SE averaged over 1000 simulation repetitions?
  Please also consider reporting a comparison between the empirical and model-based SEs, as well as Monte Carlo errors for the performance measure estimates (see Morris et al).

  Response: Thank you for your comment. We have added reports of the percent relative absolute bias and empirical SEs in Table 2. Two bullets have been added to the paragraph of results:

- The two-level J2R and CIR methods produce unbiased estimates for the true values of β^{J2R} and β^{CIR}, and cycle-level J2R and CIR methods produce unbiased estimates for the true values of $\tilde{\beta}^{J2R}$ and $\tilde{\beta}^{CIR}$, which can be seen from the small percent relative absolute bias reported in Table 2. This is also expected by the theorems.
- The model-based SEs from the MNAR-MI methods are slighly larger than the empirical SEs, while the model-based SEs from the analytic methods are very close to the empirical SEs.

In addition, we have added the reference by Morris et al (2019) and evaluated the Monte Carlo SEs. The evaluations have been added to the last paragraph of Section 6.

- The MAR methods seem to perform relatively well in terms of bias under the posited MNAR mechanisms – is this because of the small values of the selection parameters?

  Response: Thank you for your comment. The MAR methods seem to perform relatively well in terms of bias under the posited MNAR mechanisms, this may be because of the small values of the selection parameters. However, the MAR methods provide slightly smaller biases under the posited MAR mechanisms.

Comment 5:

Section 7: again, please consider adding more information on how MI is performed.

Response: Thank you for the comment. We added one paragraph (the second last paragraph) in section 7 for more information on how MI is performed.

Minor suggestions

- The term “Rubin’s rules” (in plural form) is typically used.

  Response: Thank you for the comment. All have been changed to Rubin’s rules per your comment.

- Is the *j* term before $\pi_{K,j}$ and $\pi_{K,M-j+1}$ redundant in equations (8) and (9) (and also in the appendices), respectively?

  Response: Thank you for your comment. I have validated the derivations again, the term j is not redundant in equations (8) and (9). In addition, we can examine this by the magnitude of the quantity from equation (8). If there is no j, the quantity of β_K^{J2R} will be less than 1/M of β_K ; with j, the quantity of of β_K^{J2R} will be proportional to β_K .
- Please consider adding more descriptions in the figure and table captions.

  Response: Thank you for your comment. We have added more descriptions for the Table 2. In addition, the following details of Table 2 have been added to the first paragraph in the section of Results:

  "The estimate, bias, percent relative absolute bias (|estimate−true value|/true value), standard error(SE) which is the model-based SE averaged over 1000 repetitions, empirical SE, power, and Type I error for each of the methods are reported from the simulated data in Table 2."
- MI is abbreviated multiple times throughout the paper – please check.

  Response: Thank you for your comment. We have made edits for consistence of the use of abbreviated MI.
- Abstract, lines 29-, 30: "For patients who discontinue from the trial, the proposed methods impute the missing daily outcomes based on the proposed methods" – I feel that this sentence is not so clear; please consider rewriting this?

  Response: Thank you for your helpful comment. After reading it again, I think that paragraph in the abstract flows better without this sentence, so we have deleted it.

Responses to Comments from Reviewer 2

Comment 1:

The motivating example came from an insomnia trial where the patient report was sleep time. While somewhat variable for insomnia sufferers, many PRO measures, especially those based on single items, can show much larger variability. Please comment on the generalizability of the methods presented for scores with high variability.

Response: In clinical trials with daily PROs, large variability is one of the reasons that cycle means are usually used as analysis endpoints. Therefore, we are focusing on handling missing data in trials with daily PROs and cycle means as endpoints. The two-level approaches provide alternative methods to the conventional cycle-level methods. The methods considered in this manuscript are general for this type of trials with daily outcomes, since they work well for endpoints based on cycle means. We have added some discussion in the first paragraph in Section 8.

Comment 2:

Often PRO data show an increasing attrition rate over time and although this pattern was observed in your example, available data rates were still quite high. How might higher rates of attrition impact the methodology?

Response: The MNAR assumptions such as J2R and CIR are trying to use conservative models for imputing missing data in sensitivity analyses. In practice, the unobserved values can never be truly recovered from assumed models. Therefore, in clinical trials we should make all efforts

to minimize the rate of missing data. When the rate is very high, sensitivity analyses may show different results than the primary analyses. Thus, high rate of missing data may put the validity of trials in question. For this reason, we ran simulation under mild and moderate missing scenarios that are commonly seen in many clinical trials. Some discussion about this has been added to the second last paragraph in Section 8.

Comment 3:

Your use of multilevel structures to handle missing data are refreshing, but the analysis methods are either marginalized by the MMRM framework or too large to feasibly be fit. How do your imputation methods compare to other methodologies such as latent variable models that can incorporate measurement and structural features into estimation?

Response: Our methods use MMRM which is an established model in clinical trials with repeated measures. The proposed cMMRM/MMRM are very general with unstructured covariance matrix which is often required to accommodate the multilevel variability and avoid mis-specification of the correlation. We have followed the MMRM model in the literature for daily outcomes (Thomas, Harel, Little, 2016) but provided more theoretical developments and simulation studies.

We agree with you that latent variable model may be another method for handling the daily data which allows for more specific correlation structure of the multilevel data. This is an interesting topic for future research. We have added some discussion in the last paragraph in Section 8.

Comment 4:

Measurement error is inherent to patient reports. Do you foresee needing to account for measurement error within the multilevel structure you have designed and if so, how might one go about doing so?

Response: We have accounted for the variation in the covariance structure using cMMRM/MMRM, as used in the type of trials with daily PROs in the literature. The model

accounts for both sources of variation including daily variation and measurement errors. We recommend unstructured covariance for capturing the overall variation. We don't differentiate the measurement errors from the total variation but we acknowledge this is an issue for PROs and is another interesting topic for future research.