METHOD ARTICLE

Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study [version 1; peer review: 1 approved with reservations]

Yin Mo, Cherry Lim, Mavuto Mukaka, Ben S. Cooper

1Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, 10400, Thailand
2Division of Infectious Diseases, University Medicine Cluster, National University Hospital, Singapore, 119074, Singapore
3Department of Medicine, National University of Singapore, Singapore, 119228, Singapore
4Centre for Tropical Medicine, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, UK

Abstract

Protocol non-adherence is common and poses unique challenges in the interpretation of trial outcomes, especially in non-inferiority trials. We performed simulations of a non-inferiority trial with a time-fixed treatment and a binary endpoint in order to: i) explore the impact of various patterns of non-adherence and analysis methods on treatment effect estimates; ii) quantify the probability of claiming non-inferiority when the experimental treatment effect is actually inferior; and iii) evaluate alternative methods such as inverse probability weighting and instrumental variable estimation. We found that the probability of concluding non-inferiority when the experimental treatment is actually inferior depends on whether non-adherence is due to confounding or non-confounding factors, and the actual treatments received by the non-adherent participants. With non-adherence, intention-to-treat analysis has a higher tendency to conclude non-inferiority when the experimental treatment is actually inferior under most patterns of non-adherence. This probability of concluding non-inferiority can be increased to as high as 0.1 from 0.025 when the adherence is relatively high at 90%. The direction of bias for the per-protocol analysis depends on the directions of influence the confounders have on adherence and probability of outcome. The inverse probability weighting approach can reduce bias but will only eliminate it if all confounders can be measured without error and are appropriately adjusted for. Instrumental variable estimation overcomes this limitation and gives unbiased estimates even when confounders are not known, but typically requires large sample sizes to achieve acceptable power. Investigators need to consider patterns of non-adherence and potential confounders in trial designs. Adjusted analysis of the per-protocol population with sensitivity analyses on confounders and other approaches, such as instrumental variable estimation, should be considered when non-compliance is anticipated. We provide an online power calculator allowing for various patterns of non-adherence using the above methods.
Keywords
Trial methodology, non-inferiority trials, causal inference, non-adherence

This article is included in the Mahidol Oxford Tropical Medicine Research Unit (MORU) gateway.

Corresponding author: Yin Mo (moyin@tropmedres.ac)

Author roles: Mo Y: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Lim C: Formal Analysis, Methodology, Writing – Review & Editing; Mukaka M: Formal Analysis, Writing – Review & Editing; Cooper BS: Conceptualization, Formal Analysis, Methodology, Software, Supervision, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: MY is supported by the Singapore National Medical Research Council Research Fellowship [NMRC/Fellowship/0051/2017]. BSC is supported by UK Medical Research Council / Department for International Development [MR/K006924/1]. CL is supported by the Wellcome Trust [206736/Z/17/Z]. The Mahidol-Oxford Tropical Medicine Research Unit is supported by the Wellcome Trust-Mahidol University-Oxford Tropical Medicine Research Programme [106698]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Mo Y et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Mo Y, Lim C, Mukaka M and Cooper BS. Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study [version 1; peer review: 1 approved with reservations] Wellcome Open Research 2019, 4:207 (https://doi.org/10.12688/wellcomeopenres.15636.1)

First published: 18 Dec 2019, 4:207 (https://doi.org/10.12688/wellcomeopenres.15636.1)
Introduction

Clinical trials designed to determine whether an experimental treatment is no worse than the standard-of-care treatment by a predefined margin are known as non-inferiority trials. Though a widely adopted trial design in the medical literature, the best practices for trial design, analysis and reporting remain debated. These debates often revolve around the appropriateness of the non-inferiority margin, and consistency with historical placebo-controlled trials in the choices of standard-of-care control treatment, study population and outcomes. Non-adherence to allocated treatment, which occurs commonly in all randomized controlled trials, has also been recognized as an important contributor towards making erroneous conclusions in non-inferiority trials1-3.

The most widely used analysis strategy in all clinical trials, including the non-inferiority design, is the intention-to-treat analysis (ITT). The analysis compares individuals according to their randomly allocated treatment, regardless of what actual treatment an individual receives. Hence, ITT estimates the effect of assigning a treatment instead of the treatment effect itself. The effect of assignment and the effect of treatment will generally differ when there is non-adherence. When non-adherent individuals switch to treatments prescribed in the opposite allocation arm, or take up other treatments with similar efficacy as the standard-of-care, the ITT estimates tend to shift towards zero difference. This property is generally valued in the analysis of superiority trials as the demonstration of superiority becomes more difficult. In non-inferiority trials, however, it may lead to the conclusion of non-inferiority for allocating the experimental treatment when the new treatment is actually inferior in terms of treatment efficacy.

While the effect of allocation may sometimes be reflective of the ‘real world’ practice, the causal effect of treatment on the outcome is often of considerable interest. This is the focus of this paper. To estimate treatment efficacy, a widely used method is the per protocol analysis (PP). This analysis considers only individuals who adhere to the allocated treatment and excludes those who do not. However, because the adherent individuals may have different characteristics compared to the non-adherent individuals in the allocation groups, comparing only the adherent individuals may lead to biased treatment estimates.

The above issues have been highlighted in international guidelines and simulation studies, but consensus on the best way forward has not been reached4-6. Of note, Kim has previously shown that the standard approaches can lead to erroneous conclusions about treatment efficacy in non-inferiority trials with non-adherence and proposed using an instrumental variable estimator as an alternative statistical method7. Sanchez and Chen reached a similar conclusion: depending on the pattern of protocol deviation, both PP and ITT populations may show non-inferiority when the treatment effect is actually inferior8. In the latest CONSORT guideline for non-inferiority trials, it has been suggested that hybrid ITT/PP analyses should be considered9. However, the exact methodology was not specified. Practical guidance is needed when designing trials about how incremental levels of non-adherence affect the chance of reaching different trial conclusions.

It is important to assess the potential patterns of non-adherence that might occur in a non-inferiority trial during the planning stage both to inform power calculations and to allow an appropriate analysis plan to be developed10-12. However, no easily accessible tools are currently available to guide investigators in non-inferiority trial design accounting for these considerations. In this study, we performed simulations of a hypothetical non-inferiority trial with a binary outcome in order to: i) explore the impact of various patterns of non-adherence and analysis methods on trial treatment effect estimates; ii) quantify the probability of claiming non-inferiority when treatment efficacy is actually inferior; iii) compare and evaluate alternative analysis methods such as inverse probability weighting and instrumental variable estimation; and iv) provide a tool for investigators to design non-inferiority trials which anticipate non-adherence.

Methods

We simulated a two arm non-inferiority randomized controlled trial, where treatment, A, and outcome, Y, are binary and time fixed. Randomization, Z, is done in a 1:1 ratio. An example of such a trial is the study on optimising antibiotic treatment duration for community acquired pneumonia13. The experimental treatment is five days of antibiotic treatment (\(A = 1\)), while the control treatment is a duration as decided by the physicians (\(A = 0\)). Outcome is treatment failure as defined by a set of questionnaire scores on day 30 (\(Y = 1\) represents treatment failure, \(Y = 0\) represents treatment success). The effect estimate is the absolute risk difference, calculated as the difference in the proportion of participants with treatment failure between treatment groups.

We calculated the sample size based on the hypothetical assumption that 40% of patients in both experimental and control arms experience treatment failure, with a non-inferiority margin of 10% and tolerable type 1 error of 0.025. This required 505 participants per arm for 90% power14. We explored all simulation scenarios with 60–100% adherence to illustrate the effect of adherence under various patterns of non-adherence and analysis methods. Each simulation was performed with 1000 iterations. All simulation and analyses were performed with R Version 1.1.46315. Simulation code is available on GitHub. (https://github.com/moyinNUHS/NItrlalsimulation.git).

Notation

In the subsequent paragraphs, \(Y_{\text{ctrl}}\) represents the outcome that would occur if the experimental treatment were to be administered (\(A = 1\)); and \(Y_{\text{exp}}\) represents the potential outcome if the control treatment were to be administered (\(A = 0\)). For an individual, \(i\), \(Y_{\text{ctrl}}\) and \(Y_{\text{exp}}\) are therefore counterfactual outcomes. Because only one of the outcomes is observed in the real world, the actual observed outcome, \(Y_i\), is either equal to \(Y_{\text{ctrl}}\) or \(Y_{\text{exp}}\) depending on the treatment received, i.e. \(Y_i = Y_{\text{ctrl}}\) if the individual received the experimental treatment and \(A_i = 1\) if the individual received the control treatment16. Similarly, the observed outcomes depending on randomization...
(Z) are represented by $Y_{iz=1}$ and $Y_{iz=0}$ respectively. C refers to the confounding factors that may increase or decrease the probabilities of adhering to the allocated treatment and outcome.

**Analysis methods**

The ITT analysis considers all randomized participants according to their assigned groups, regardless of whether the participants had the intended treatment. It estimates the effect of Z on Y, i.e. $Pr[Y_{z=1}] - Pr[Y_{z=0}]$]. The PP analysis only considers participants who received treatment according to their allocation stated in the study protocol, i.e. $Pr[Y_{A=1}, Z=1] - Pr[Y_{A=1}, Z=0]$. In addition, we used an inverse probability weighting approach to estimate the causal effect of treatment on the outcome. This approach applies a logistic regression model incorporating the confounder as an explanatory variable to estimate an individual’s probability of adhering to a particular allocation arm. The inverse of these predicted probabilities are used as weights to inflate or deflate the individual’s influence on the overall treatment effect in the group.

Lastly, we used instrumental variable estimation. This approach analyzes all participants by quantifying first, the degree to which allocated treatment predicts actual treatment and, second, the degree to which treatment predicts outcome. We adopted the structural mean model, first proposed by Robins and Rotnitzky for estimation of the received treatment effect on a dichotomous outcome in randomized trials. The main assumptions in using instrumental variable estimation are that: i) the instrument, Z, is associated with the actual treatment received, A; ii) Z does not affect the outcome, Y, except through its potential effect on A; and iii) Z and Y do not share causes. Out of these conditions, only the first is verifiable. In the context of a randomized controlled trial, randomization is an appropriate instrument. When done correctly, randomization satisfies the first and third conditions as it randomly allocates treatment to the participants, independent of the final outcomes. The second condition is satisfied in a successfully double blinded study. Details of the analysis methods are provided in Extended data.

**Non-inferiority hypothesis testing**

The null hypothesis is tested by comparing the upper bound of the two-sided or one-sided 95% confidence interval of the effect estimate with the non-inferiority margin. Non-inferiority is concluded if the upper confidence interval bound is less than the non-inferiority margin.

**Simulation mechanism**

We generated individual level data which included the following variables: treatment allocation, participant characteristics, which may affect adherence and outcome, actual treatment received, counterfactual outcomes and observed outcomes. Allocation is a binary variable with each individual having a 50% probability of being allocated to the experimental treatment. Participant characteristics were represented by a single continuous variable on the interval [0, 1] drawn from a Beta distribution. This can be thought of as a disease risk score.

We considered two common reasons for non-adherence. The first is when non-adherence is due to factors which affect the probability of taking up the allocated treatment but do not affect the study outcome through any other pathway (Figure 1A). The

![Figure 1](image-url)

**Figure 1.** Directed acyclic graphs demonstrating the causal relationships of the variables generated for each study participant. In scenario A, factors that cause non-adherence affect the probability of the participant taking up the allocated treatment but do not affect the outcome e.g. minor side effects of the treatment drug. In scenario B, factors causing non-adherence affect both the probability of the participant taking up the allocated treatment as well as the outcome e.g. disease severity.
second is driven by confounders, defined as the study participants’ prognostic factors that affect both the probability for taking up the allocated treatment and the outcome (Figure 1B).

The actual treatment received by an individual differs from the allocated treatment when there is non-adherence. In the case where the participant characteristics cause an individual to switch to an experimental treatment, their probability for crossing over to the experimental treatment when randomized to the control group is increased. An example is a trial studying an experimental treatment for a terminal disease which has few effective treatment options. An individual with more severe disease may be more likely to switch to the experimental treatment even when they are randomized to the control treatment. In another case where the factor causing non-adherence discourages an individual to take up an experimental treatment, their probability for adhering to the experimental treatment after being randomized to the experimental arm is decreased. The individual might take up the control treatment or refuse treatment altogether. An example is a trial comparing an experimental exercise regime to nicotine patches for smoking cessation. An individual with chronic obstructive lung disease may be more likely to be non-adherent to the experimental exercise regime and take up nicotine patch or decline all treatments.

We generated counterfactual outcomes for each individual, one for experimental treatment and one for control treatment. The overall averaged difference between the counterfactual outcomes for all study participants is the pre-defined true treatment effect assumed in the simulation. The participant characteristics may cause an increase or decrease in the probability of having the outcome depending on the direction of influence the confounder has on the outcome. The observed outcome is then chosen from one of the counterfactuals depending on the actual treatment that the individual received. Detailed descriptions of the simulations are included in the supplementary material.

We simulated 18 different patterns of non-adherence. The conditions of these non-adherent patterns are shown in Figure 2.

Comparing the analysis methods
To examine type 1 error, i.e. concluding non-inferiority when the experimental treatment is actually inferior, we assumed a difference in the probability of treatment failure between the control and experimental arms of 0.1 (i.e. the experimental treatment is inferior and its true treatment effect is -0.1 on an absolute scale). Since the non-inferiority margin is assumed to be 10%, simulation iterations which concluded non-inferiority

![Figure 2. Simulation scenarios.](image-url) Simulation scenarios were explored permutations of four factors: i) non-adherent population (both groups, experimental group, control group); ii) actual treatment received by the non-adherent population (crossing over to the opposite group, another treatment inferior to both the experimental and control treatment); iii) reason for non-adherence (due to confounding factors or non-confounding factors); iv) if non-adherence is due to non-confounding factors, direction of influence of the confounders on the probability of taking up the experimental treatment and outcome (both probabilities may increase or decrease, or the two probabilities are in opposite directions). Left Panel shows the six possible scenarios when non-adherence is due to non-confounding factors. Right Panel shows the 12 possible scenarios when non-adherence is due to confounding factors. Each coloured line represents one scenario.
were considered to have committed type 1 error (Extended data: Supplementary Figure 1).

Power, given by one minus the type 2 error, is the proportion of non-inferiority trials which conclude non-inferiority correctly. Here, we assumed the true treatment effect to be zero. Thus, the experimental treatment arm has the same probability of having treatment failure i.e. non-inferior to the control treatment. Simulation iterations which concluded inferiority were considered to have committed a type 2 error (Extended data: Supplementary Figure 1).

The above assumptions on treatment effects used in calculating type 1 error and power for the scenarios below are arbitrary and intended for illustrative purposes. Other assumptions can be explored with the Shiny app (https://moru.shinyapps.io/samplesize_nonadherence/).

Results
Non-adherence due to non-confounding factors
In most patterns of non-adherence, ITT estimates tend to shift towards zero difference between the control and experimental groups. The only exceptions are when study participants allocated to the experimental arm actually received no treatment or a treatment inferior to both treatments offered in the trial. Compared to treatment efficacy estimates, ITT analysis has a higher tendency of claiming non-inferiority when the experimental treatment is actually inferior when there is non-adherence. Figure 3 illustrates the case where non-adherent study participants cross over to the opposite arm. Even at a relatively high adherence of 90%, the type 1 error of the ITT estimate can be as high as 10%. All other analysis methods are unbiased in this case where non-adherence is due only to non-confounding factors. Note the different scale for the instrumental variable estimates, and the high variance at low adherence.

Non-adherence due to confounders and no unobserved confounding
In the case where confounders influence non-adherence behavior, PP analysis is biased in estimating the causal effect of treatment. Figure 4 illustrates an example where increasing confounder value decreases the probability of taking up the experimental treatment (with a corresponding increase the probability of taking up the control treatment) and increases the

Figure 3. Non-adherence caused by non-confounding factors. A: Dots represent trial estimates calculated from each iteration. Coloured lines present the Locally Weighted Scatterplot Smoothing (LOESS) lines through mean trial estimates from all iterations. Because our outcome in the simulated trial refers to treatment failure, higher effect estimate values favour control treatment. The red dotted line is the true effect size estimate assumed in the simulations. B: Dots represent type 1 error calculated from all iterations at various degrees of adherence. The tangible type I error is set at 0.025 at full adherence.
probability of treatment failure. This will lead to an inflated type 1 error rate. In this case, inverse probability weighting and instrumental variable estimation give unbiased estimates with conservative type I error rates.

When the confounder decreases both the probability of taking up the experimental treatment and of treatment failure, the treatment effect estimated with the PP analysis will be higher than the true value. (Figure 5)

Non-adherence due to confounders with unobserved confounding
In practice, not all confounders will be observed, and those which can be observed may not be measured perfectly so that it will only be possible to partially adjust for confounding. In such cases, inverse probability weighting can become biased (Figure 6). Adjusting for more confounders can reduce but not eliminate bias in treatment estimates. Instrumental variable estimation, on the other hand, remains unbiased even with unobserved confounders, as it does not depend on the knowledge of the confounders to compute treatment effect estimates when all the above-mentioned assumptions are met.

Effect of non-adherence on power
In addition to affecting treatment estimates, non-adherence decreases the power to detect truly non-inferior experimental treatments. We consider the effect of non-adherence on inverse probability weighting and instrumental variable effect estimates as these methods can potentially give unbiased treatment efficacy estimates despite non-adherence.

To maintain power, the sample size required for instrumental variable estimation increases drastically when adherence falls below 95%. In contrast, sample size for inverse probability weighting changes linearly with the decrease in adherence (Figure 7).

Different patterns of non-adherence and choice of analysis methods affect power to differing degrees. To aid investigators in planning for clinical trials anticipating non-adherence, a power calculator is available online based on the simulation mechanisms shown here (https://moru.shinyapps.io/samplesize_nonadherence/). Using the same simulation mechanism as above, the calculator caters for a two-arm non-inferiority trial with a binary outcome and time-fixed treatment. The application
Figure 5. Non-adherence caused by confounding factors II. Non-adherence caused by confounding factors where higher confounder value decreases the probability of taking up the allocated treatment and treatment failure.

Figure 6. Non-adherence caused by both known and unknown confounders. Four confounders were added in the simulation. ITT, PP and instrumental variable analyses did not adjust for any confounders. For the inverse probability weighting analysis, the four lines represent situations when one, two, three and all of confounders were adjusted for. With more confounders accounted for, treatment estimates become less biased.
is an interactive platform that calculates power using user inputs for the following: whether non-adherence is mainly caused by non-confounding and confounding factors; number of participants who are anticipated to cross-over to the opposite arm; and the various directions of influence the confounders have on adherence and probability of outcomes.

**Discussion**

Our simulations illustrate the complexities in interpreting non-inferiority trials with non-adherence, taking both qualitative and quantitative perspectives. Intention-to-treat effect estimates, due to the ‘dilution’ from the participants who received other treatments different from the allocated treatment, tend to be lower than true treatment effects at low adherence under most non-adherence patterns. As non-adherence increases, the chance that ITT analysis will conclude non-inferiority increases. The probability of concluding non-inferiority when the treatment is actually inferior can be increased to as high as 0.1 from the acceptable 0.025 when non-adherence in 90%. The direction of bias in PP analysis is dependent on whether the confounders increase or decrease the probability of taking up the allocated treatment and the probability of the outcome occurring.

Inverse probability weighting accounts for the difference in confounders between the allocation groups to ensure that the reweighted groups are similar and comparable. It eliminates bias if all confounders can be appropriately adjusted for, but in general this will not be possible. Sensitivity analysis methods are available to address unobserved confounding and covariate measurement errors. In contrast, instrumental variable estimation can account for unknown confounders but requires the “exclusion restriction” to be fulfilled (i.e. treatment allocation only influences the outcome through the treatment and not through any other pathways). This assumption is unverifiable and we are only likely to be confident that it holds in a double blinded study. The other drawback of using an instrumental variable is the need for large sample sizes when adherence is low as the method relies heavily on the strength of the instrument (i.e. randomization) in predicting the treatment. Recent methods using doubly robust procedures have been developed to boost power when using instrumental variable estimation.

Though our simulation mainly illustrates the analysis of time fixed treatments and outcomes, time varying treatments and outcomes can be analyzed with inverse probability weighting and g-estimation methods. These methods are also used to address missing data and censoring.

Some degree of non-adherence is near ubiquitous in clinical trials. Though ITT will, under some circumstances, represent the ‘real-world’ effectiveness of treatment allocation, the effects of treatment itself are relevant estimates generalizable to other situations with different adherence patterns. They are also likely to be of particular interest for those with agency in their adherence. When the interest is in the actual treatment effects, as we have shown, the conservative nature of ITT in a conventional superiority trial (i.e. lower probability of concluding superiority in the presence of non-adherence) is compromised under many patterns of non-adherence in a non-inferiority trial.
In conclusion, given the potential inflation in the probability of concluding non-inferiority with non-adherence, we propose that during the planning stage of clinical trials, especially those that expect non-adherence to be more than 0.05, investigators should anticipate the likely patterns and magnitude of non-adherence and devise ways to reduce it. Ideally, power calculations should account for such anticipated non-adherence. Potential confounders should be carefully measured and recorded for subsequent analysis. Adjusted analysis of the PP population using inverse probability weighting or g-estimation can reduce bias in treatment effect estimates introduced by non-adherence. In the case of double blinded trials with large sample sizes, instrumental variable estimation may also be appropriate.

Data availability
Underlying data
Simulation code is available on GitHub: https://github.com/moyinNUHS/NItrialsimulation.git.


License: Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

References


The authors have conducted extensive simulations to evaluate the impact of various patterns of non-adherence on the treatment effect estimates, Type 1 error rate and power of different methods for analyzing non-inferiority trials. Many of the key findings, particularly those pertaining to the ITT, per-protocol, and instrumental variables methods, have been reported by others. What is novel in this paper is that potential confounders of non-adherence are considered, and a user-friendly app was developed that allows one to evaluate the impact on power of different degrees of non-adherence. The paper could be improved, however, by resolving several inconsistencies and errors in notation, and by providing additional details about how the simulations were performed. Below are some specific comments:

- It should be explicitly stated in the paper that in the simulation studies, it was assumed that non-adherence is all or nothing (i.e., no partial non-adherence). Also, when non-adherence does occur, it is assumed to be either always cross-over, or always switching to an alternative that is inferior to both experimental and control treatments. The authors may want to comment on whether this second set of assumptions is very realistic since in a non-inferiority trial with an active control arm, one would expect that both cross-over and switching to an inferior alternative would occur in the same study. The smoking cessation trial described in the paper is actually an example where both could happen since an individual assigned to the experimental exercise regimen may not adhere by either crossing over and taking up the nicotine patch (control) or switching to the inferior alternative of declining all treatments.

- It is unclear from the description of the simulations how the impact of switching to an alternative treatment that is inferior to both the experimental or control treatment was evaluated. Step 4 in the supplementary material addresses only the cross-over non-adherence pattern (A = 0, 1). Also, the assumed probability of failure on this inferior alternative was not stated and would presumably have a major impact on the ITT results.

- More detailed results should be provided for the non-adherence pattern where subjects could receive no treatment or one that is inferior to the trial treatments. Figure 3 seems to address only
Page 5: “…when the experimental treatment is actually inferior, we assumed a difference in the probability of treatment failure between the control and experimental arms of 0.1 (i.e. the experimental treatment is inferior and its true treatment effect is -0.1 on an absolute scale).” The outcome was defined on Page 3 as treatment failure (Y=1 represents treatment failure and Y=0 represents treatment success). Since outcome rates that are higher in the experimental compared to the control arm are in the direction of inferiority, wouldn’t it be more accurate to re-state the above sentence as “…we assumed a difference in the probability of treatment failure between the experimental and controls arms of 0.1 (i.e. the experimental treatment is inferior and its true treatment effect is 0.1 on an absolute scale).”

- It is difficult to identify from the colored lines in Figure 2 which six scenarios in the left panel and which twelve scenarios in the right panel were considered in the simulation studies. It would help if different line patterns in addition to colors were used to denote the various scenarios.

- Figure 2 suggests that the simulations considered scenarios where non-adherence occurs in both groups, experimental group only or control group only, so it is not clear in all the X-axes that are labeled as “proportion of adherent participants”, which group(s) this proportion refers to. Is the proportion of adherent participants assumed to be the same in the both treatment groups?

- Page 6 states that “Figure 4 illustrates an example where increasing confounder value decreases the probability of taking up the experimental treatment”. Is this confounding relationship assumed only in the experimental arm? What non-compliance pattern/rate is assumed in the control arm? Or did the authors mean they assumed increasing confounding value decreases the probability of adherence/taking up the allocated treatment in both arms as suggested by the Figure 4 title? The right and left panels of Figure 2 also state “Chance of treatment failure and taking up experimental treatment both increased or decreased.” Please clarify if this is supposed to be experimental treatment (in experimental arm only?) or allocated treatment (in both arms?).

- The app which allows one to explore the impact of non-adherence in a non-inferiority trial is potentially useful, but it seems it is only applicable to the case where cross-over is the only form of non-compliance. This should be clearly stated. Also, in the case of confounding factors, one needs to specify the “effect of confounder on taking up the experimental treatment”. Again, it should be clarified whether the authors indeed mean the effect of confounder on taking the experimental treatment in the experimental group only or taking the assigned treatment (whichever group the participant was assigned to). The supplementary material (Step 2) and the titles for Figures 5 and 7 suggest the latter (Figure 5: “…higher confounder value decreases the probability of taking up the allocated treatment”).

- With the authors’ approach for evaluating the impact of potential confounders, one does not have a clear sense for exactly how much confounding is occurring since only the direction of the effect of the confounder on the adherence and treatment failure probabilities is specified and not the magnitude of the confounding effect. It would be informative if, for a few of the confounding scenarios, the mean estimated failure rates (from all simulation iterations) in the non-adherers could be contrasted with the corresponding mean failure rate among adherers (separately in the experimental and control arms) so readers have a more intuitive idea of the degree of selection bias that is occurring due to the specific way that confounding was generated in the simulations.
- It is not clear why the Type 1 error rates for the IV and IPP methods are below the nominal 0.025 level in Figures 4 and 5.

- Page 4, line 10: there is an error in the second term of the expression for the PP effect.

- Page 4, Non-inferiority hypothesis testing: it should be clarified that the upper bound of the 95% CI for the absolute risk difference (experimental – control) needs to be less than the NI margin.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** biostatistics, clinical trials

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.