METHOD ARTICLE

REVISED Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study [version 2; peer review: 2 approved]

Yin Mo [1-4], Cherry Lim [1,4], Mavuto Mukaka [1,4], Ben S. Cooper [1,4]

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

v2 **First published:** 18 Dec 2019, **4**:207
https://doi.org/10.12688/wellcomeopenres.15636.1
Latest published: 24 Apr 2020, **4**:207
https://doi.org/10.12688/wellcomeopenres.15636.2

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.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 24 Apr 2020	✓ report	✓ report
version 1 18 Dec 2019	? report	✓ report

1 **Mimi Y. Kim**, Albert Einstein College of Medicine, Bronx, USA

2 **Matteo Quartagno**, University College London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

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 (Grant ref: MR/K006924/1) and the UK Department of Health and Social Care using UK Aid funding managed by the NIHR (Grant ref: PR-OD-1017-20006). The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health and Social Care or the Department for International Development. 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.

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 2; peer review: 2 approved]** Wellcome Open Research 2020, **4**:207 https://doi.org/10.12688/wellcomeopenres.15636.2

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

REVISED **Amendments from Version 1**

We would like to thank the reviewers for their insightful comments. We have addressed all of the points raised in the new version of the manuscript. We made two main changes here. Firstly, we added a scenario where non-adherent participants receive an alternative inferior treatment compared to both the experimental and control treatments, e.g., declining all treatments, in the methods and results section. This option was also added in the online power calculator application. We now provide the simulation outcome figures for all 18 simulation scenarios (illustrated in Figure 2) in supplementary 2. Secondly, we highlighted that the more influence the confounder has on treatment failure, the more biased per protocol estimates will be, potentially leading to higher type 1 error rates. This is illustrated in Figure 5. We also added the option of choosing the expected magnitude of direct confounder effect on treatment failure in the online power calculator application. The point-by-point replies to the reviewers' comments and specific changes to the manuscript are included in 'Responses to peer review reports'.

Any further responses from the reviewers can be found at the end of the article

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 trials[1–3].

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, 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 arms, 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 reached[3–6]. 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 method[7]. 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 inferior[8]. In the latest CONSORT guideline for non-inferiority trials, it has been suggested that hybrid ITT/PP analyses should be considered[4]. 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 developed[8,9]. 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 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 pneumonia[10]. The experimental treatment is five days of antibiotic treatment ($A = a_1$), while the control treatment is a duration as decided by the physicians ($A = a_0$). 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). With this single end-point, we consider adherence as a binary variable where non-adherent patients in the short arm would receive longer than five days of treatment, and non-adherent patients in the long arm would receive fewer than five days of treatment. The effect estimate is the absolute risk difference, calculated as the difference in the proportion of participants with treatment failure between treatment arms.

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% power[11]. 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 1.1.463[12]. Simulation code is available on GitHub. (https://github.com/moyinNUHS/NItrialsimulation.git).

Notation

In the subsequent paragraphs, $Y^{a=0}$ represents the potential outcome if the control treatment were to be administered ($A = a_0$); $Y^{a=1}$ represents the outcome that would occur if the experimental treatment were to be administered ($A = a_1$); and $Y^{a=2}$ represents the potential outcome if an alternative inferior treatment compared to both the control and experimental treatments were to be administered ($A = a_2$). For an individual, i, $Y_i^{a=a_0}; Y_i^{a=a_1}; Y_i^{a=a_2}$ 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 $Y_i^{a=a_0}; Y_i^{a=a_1}; Y_i^{a=a_2}$ depending on the treatment received, i.e. $Y_i = Y^{A_i}$, where $A_i = a_0$ if the individual received the control treatment, $A_i = a_1$ if the individual received the experimental treatment, and $A_i = a_2$ if the individual received an alternative inferior treatment compared to both the control and experimental treatments[13]. Similarly, the observed outcomes depending on randomization (Z) are represented by $Y_i^{z=1}$, and $Y_i^{z=0}$ 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 arms, regardless of whether the participants had the intended treatment. It estimates the effect of Z on Y, i.e. $Pr[Y^{Z=1} = 1] - Pr[Y^{Z=0} =1]$. The PP analysis only considers participants who received treatment according to their allocation stated in the study protocol, i.e. $Pr[Y^{A=a_1, Z=1} =1] - Pr[Y^{A=a_0, Z=0} =1]$.

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 arm[14].

Lastly, we used instrumental variable estimation in scenarios where non-adherent participants receiving treatment of the opposite arm. 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[15]. 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[16]. 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[17]. 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. When the non-adherence pattern involves switching of treatment to an alternative other than the experimental or control treatment, preference-based analyses using a framework involving 'compliers', 'preferers' and 'insisters' which allows for comparison of treatment effects of two active treatments are available[18]. However, this involves additional assumptions on the treatment effects in the various arms of participants which are often not verifiable. Details of the analysis methods are provided in *Extended data* (Supplementary 1)[18].

Non-inferiority hypothesis testing

The null hypothesis is tested by comparing the upper bound of the two-sided 95% confidence interval of the effect estimate with the non-inferiority margin. Non-inferiority is concluded if the upper bound of the 95% confidence interval for the absolute risk difference between the experimental and control treatments 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[19].

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 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. We considered scenarios where non-adherent participants cross over to the opposite treatment arm, or receive alternative treatments that are inferior to both the control and experimental treatments. 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 arm 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

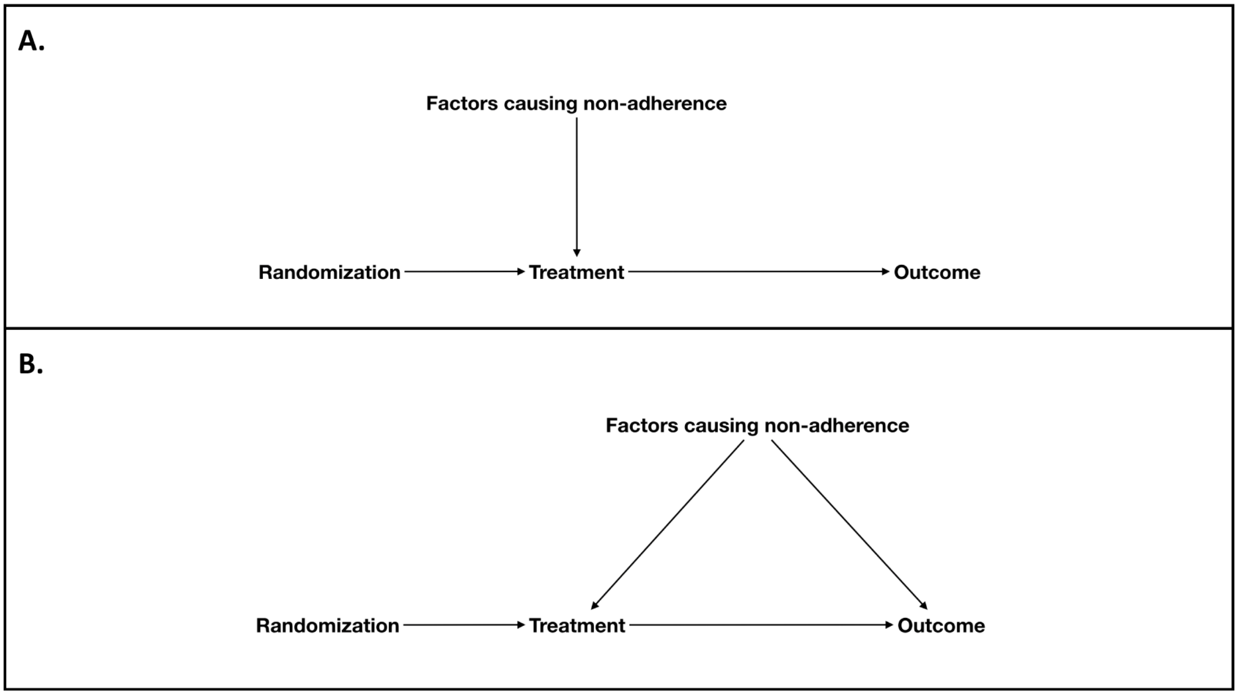

Figure 1. Directed acyclic graphs demonstrating the causal relationships of the variables generated for each study participant. In scenario **A**, factors that cause non-adherence affect the probability of the participant taking up the allocated treatment but do not affect the outcome e.g. minor side 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.

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 experimentalexercise 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 regimeand take up nicotine patch or decline all treatments.

Although adherence was considered as a binary variable in the simulations, in the scenario where participants received less effective treatments than the control and experimental treatments may reflect partial adherence to either treatments.

We generated counterfactual outcomes for each individual, one for experimental treatment, one for control treatment and one for alternative treatments inferior to both the control and experimental treatments. The overall average difference between the counterfactual outcomes for experimental and control treatments for all study participants is the pre-defined true treatment effect assumed in the simulations. 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. Graphs illustrating effect estimates and associated type 1 errors for all simulated scenarios are included in the Supplementary 2 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). In the case of non-adherent participants receiving an alternative inferior treatment compared to the experimental and control arms, we assumed the difference in the probability of treatment failure between the control and alternative treatment, and experimental and alternative treatment to be 0.1 and 0.2 respectively. Since the non-inferiority margin is assumed to be 10%, simulation iterations which concluded non-inferiority were considered to have committed type 1 error (*Extended data:* Supplementary 2 Figure 1[18]).

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,

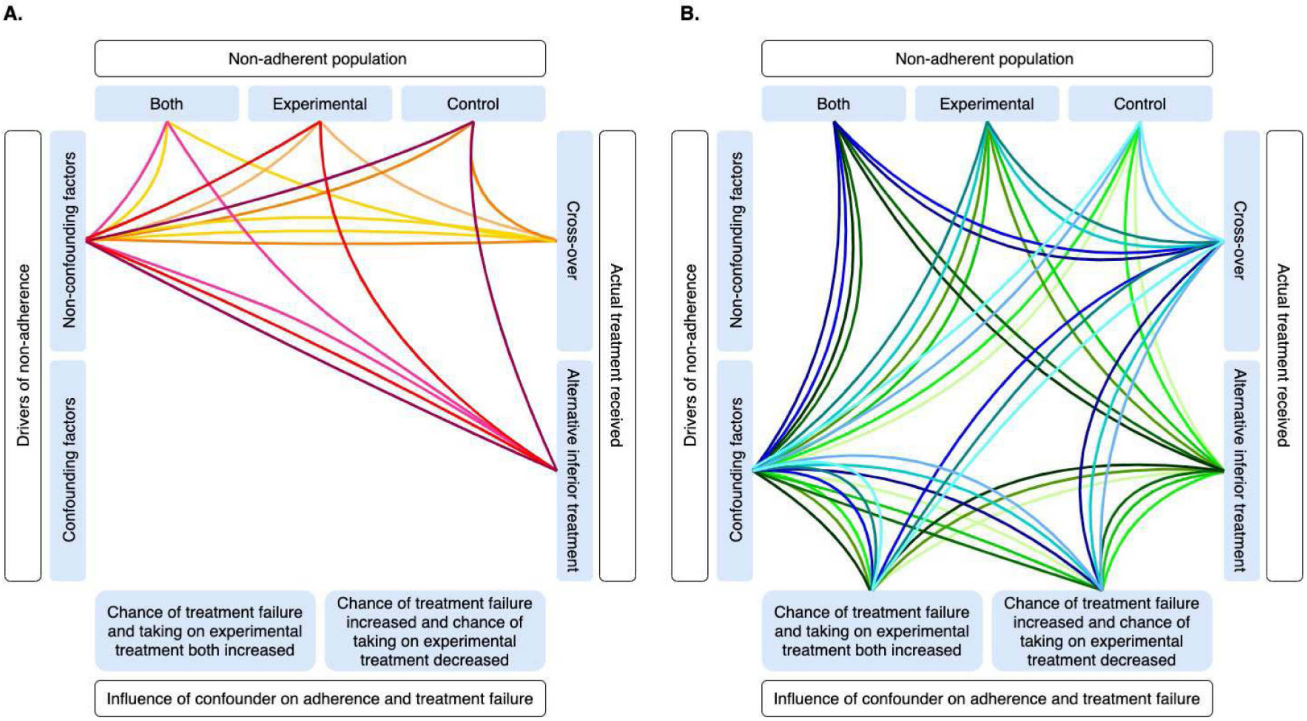
A.
Non-adherent population
Both
Experimental
Control
Drivers of non-adherence
Non-confounding factors
Confounding factors
Cross-over
Alternative inferior treatment
Actual treatment received
Chance of treatment failure and taking on experimental treatment both increased
Chance of treatment failure increased and chance of taking on experimental treatment decreased
Influence of confounder on adherence and treatment failure
B.
Non-adherent population
Both
Experimental
Control
Drivers of non-adherence
Non-confounding factors
Confounding factors
Cross-over
Alternative inferior treatment
Actual treatment received
Chance of treatment failure and taking on experimental treatment both increased
Chance of treatment failure increased and chance of taking on experimental treatment decreased
Influence of confounder on adherence and treatment failure

Figure 2. Simulation scenarios. Simulation scenarios were explored permutations of four factors: i) non-adherent population (both arms, experimental arm, control arm); ii) actual treatment received by the non-adherent population (crossing over to the opposite arm, 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.

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 2 Figure 1[18]).

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-adherent participants receive treatment from the opposite arm

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 arms. 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 in the probability of taking up the control treatment) and increases the probability of treatment failure. This is such that participants with the highest confounder values in the experimental arm cross over to the control arm and participants with the lowest confounder values in the control arm cross over to the experimental arm. 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.

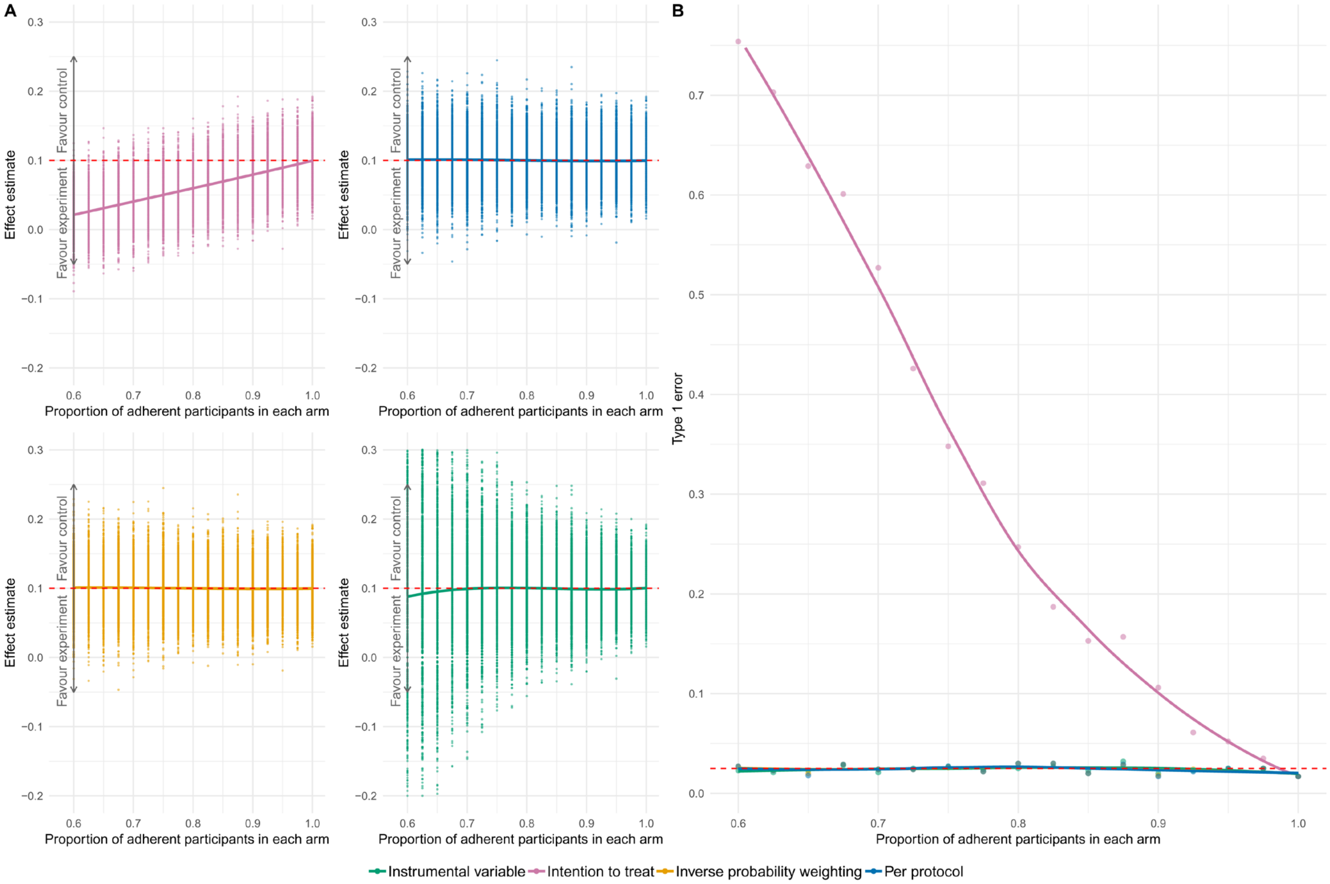

Figure 3. Non-adherence caused by non-confounding factors. A: Dots represent trial estimates calculated from each iteration. Coloured lines present the Locally Weighted Scatterplot Smoothing (LOESS) lines through mean trial estimates from all iterations. Because our outcome in the simulated trial refers 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 tolerable type I error is set at 0.025 at full adherence.

The more influence the confounder has on treatment failure, the more biased PP estimates will be, leading to higher type 1 error rates (Figure 5). When the confounder increases 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 6).

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 7). 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.

Non-adherent participants receive an alternative inferior treatment compared to both the experimental and control treatments

If non-adherent participants do not cross over to the opposite arm, they may receive an alternative inferior treatment or default care. The effect of this on the ITT treatment estimates depends on the allocation arm that is predominantly non-adherent. When most of the non-adherent participants are from the control arm, the control treatment will appear worse compared to the experimental treatment using the ITT analysis, thereby favouring the experimental treatment (Supplementary 2 Figures 2F, 2L, 2R). However, when most of the non-adherent participants are from the experimental arm, the experimental treatment will appear worse compared to the experimental treatment using the ITT analysis, thereby favouring the control treatment (Supplementary 2 Figures 2D, 2J, 2P).

Where non-adherence is caused by confounding factors, PP estimates become biased. The direction of bias is determined by the difference in the underlying prognostic characteristics of the

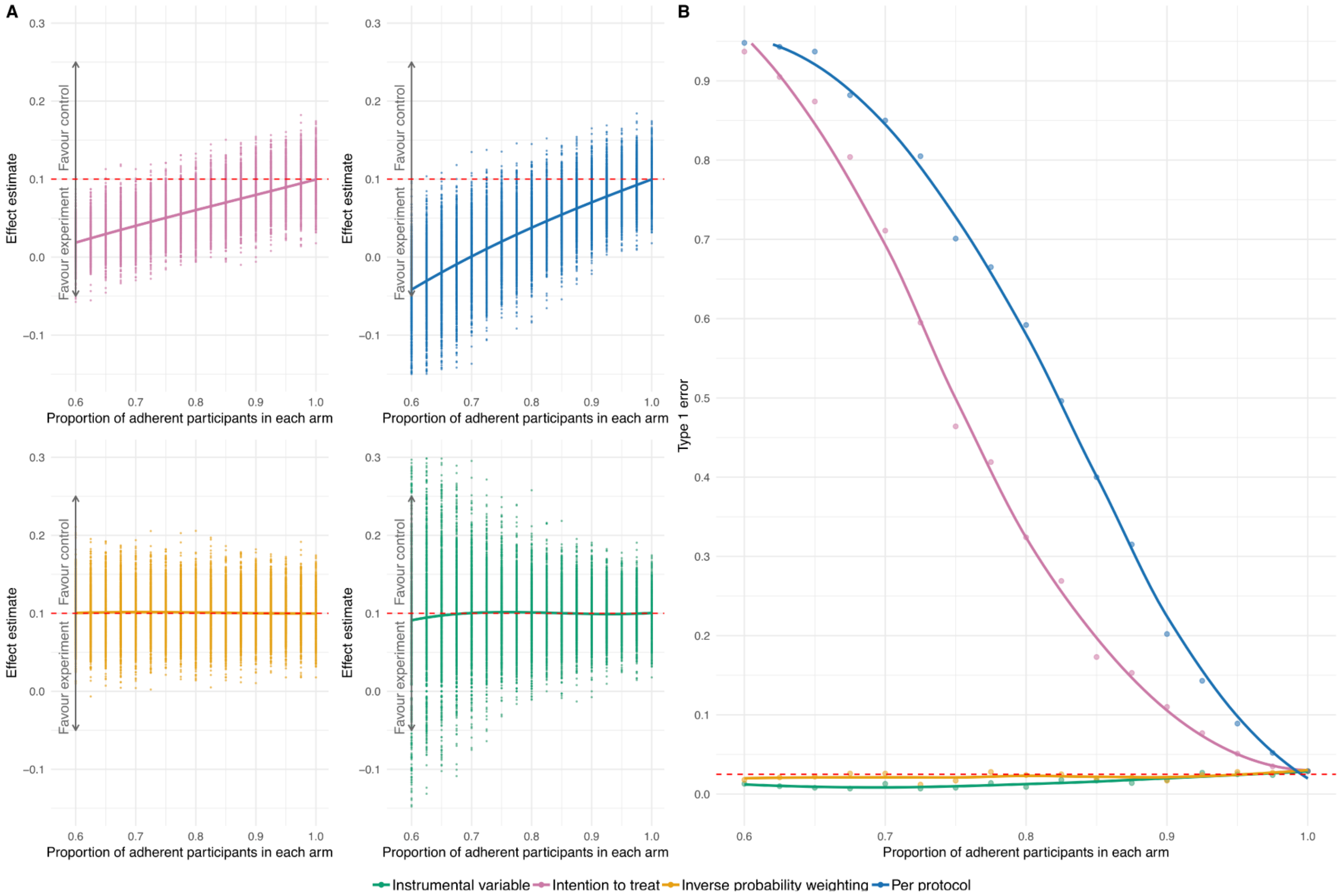

Figure 4. Non-adherence caused by confounding factors I. Non-adherence caused by confounding factors where participants with higher confounder values have lower probability of taking up the allocated treatment regardless of the allocation, and increases the probability of treatment failure.

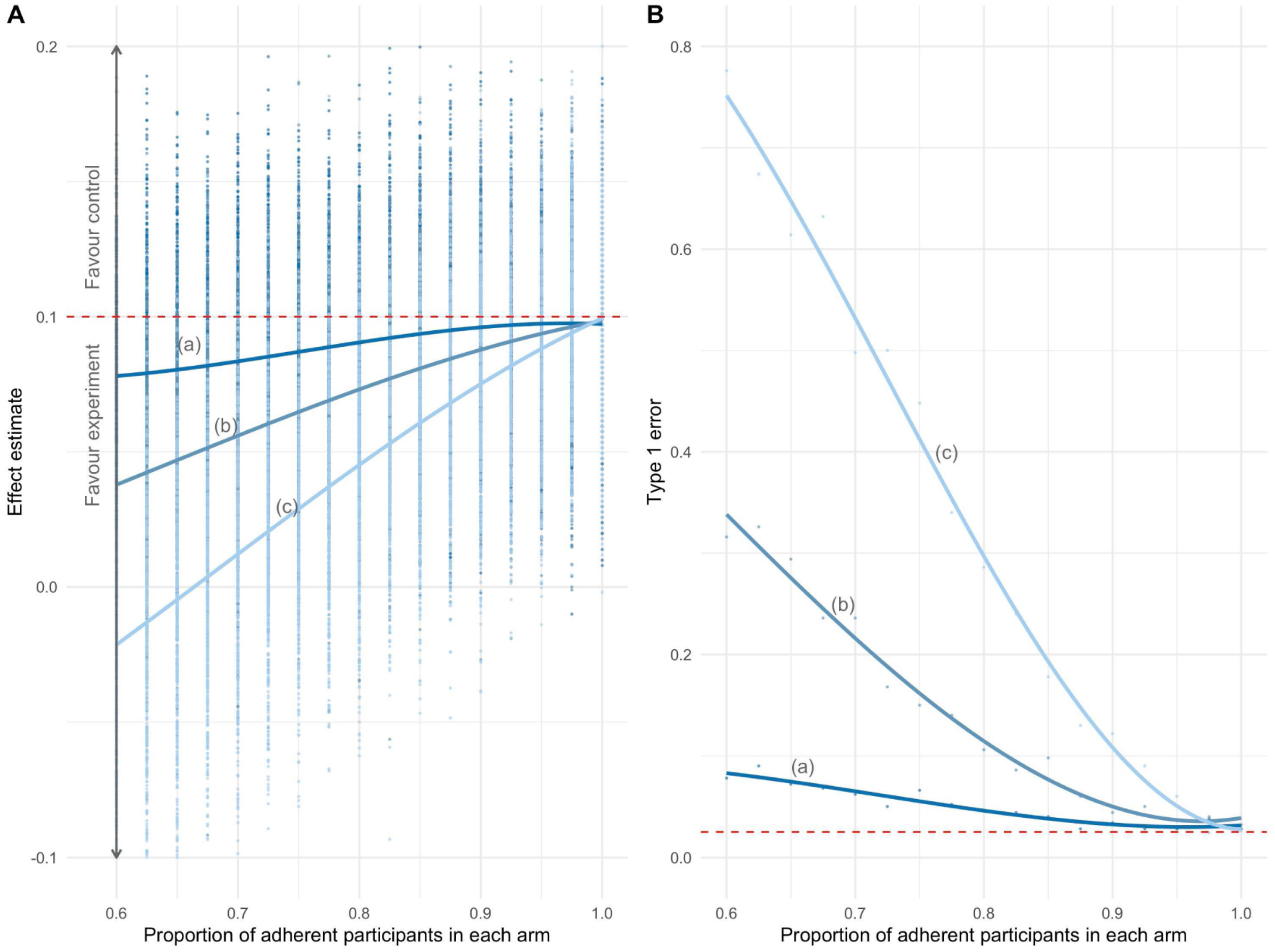

Figure 5. Non-adherence caused by confounding factors II. Non-adherence caused by confounding factors where participants with higher confounder values have lower probability of taking up the experimental treatment regardless of the allocation, and increases the probability of treatment failure. Per protocol analysis is shown (in various shades of blue) to illustrate the impact of increasing direct confounder effect on treatment failure, in terms of the treatment estimates and associated type 1 errors. The magnitude of direct confounder effect on treatment failure is calculated with treatment failure as the dependent variable, and confounder as the independent variable, in a linear regression. (a) magnitude of direct confounder effect on treatment failure = 1; (b) magnitude of direct confounder effect on treatment failure = 5; (c) magnitude of direct confounder effect on treatment failure = 9.

non-adherent participants, similar to the cases where non-adherent participants cross over to the opposite arms.

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 8). In the presence of non-adherence, the more influence the confounder has on treatment failure, the lower the power.

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 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

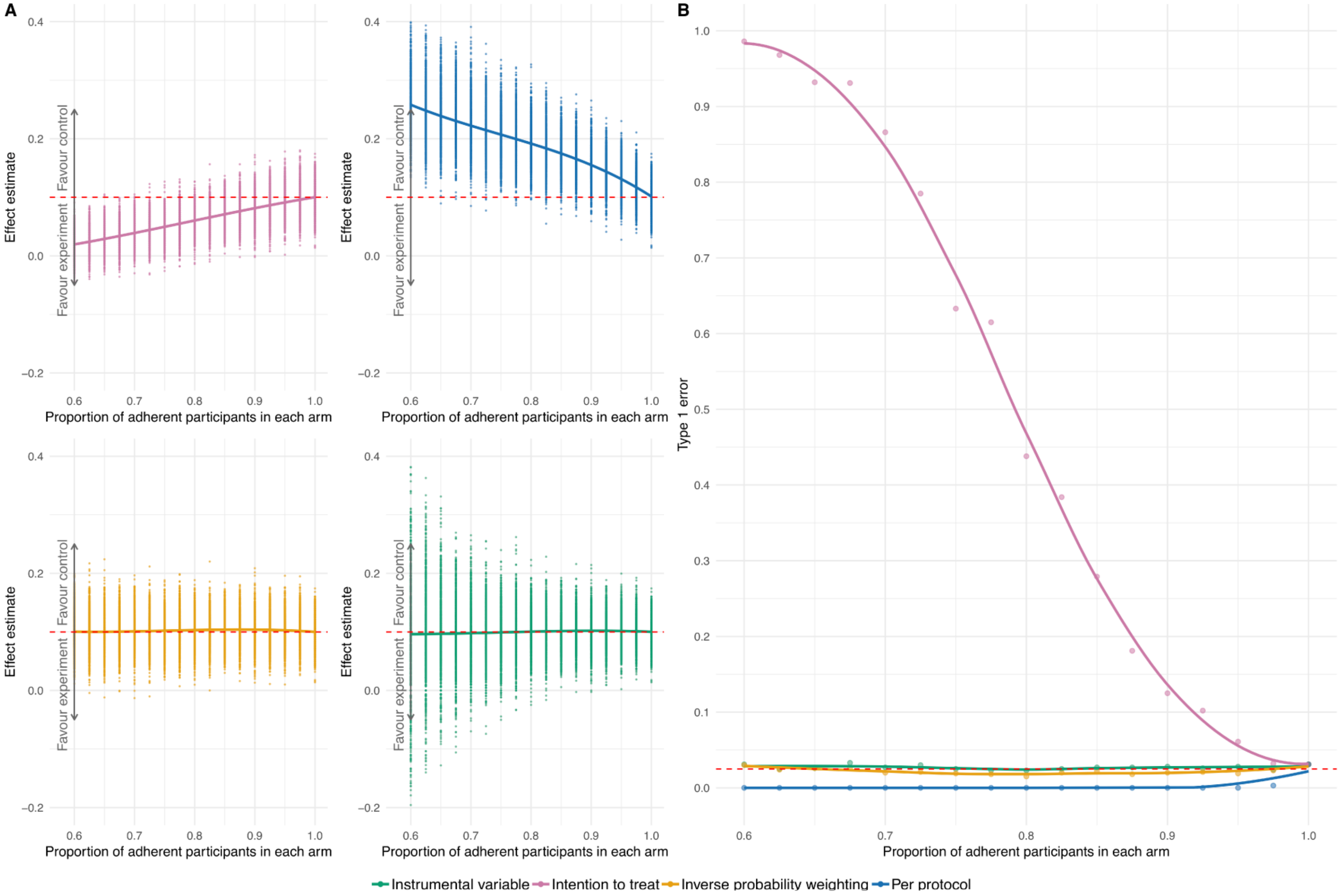

Figure 6. Non-adherence caused by confounding factors III. Non-adherence caused by confounding factors where participants with higher confounder values have higher probabilities of taking up the experimental treatment regardless of the allocation, and treatment failure.

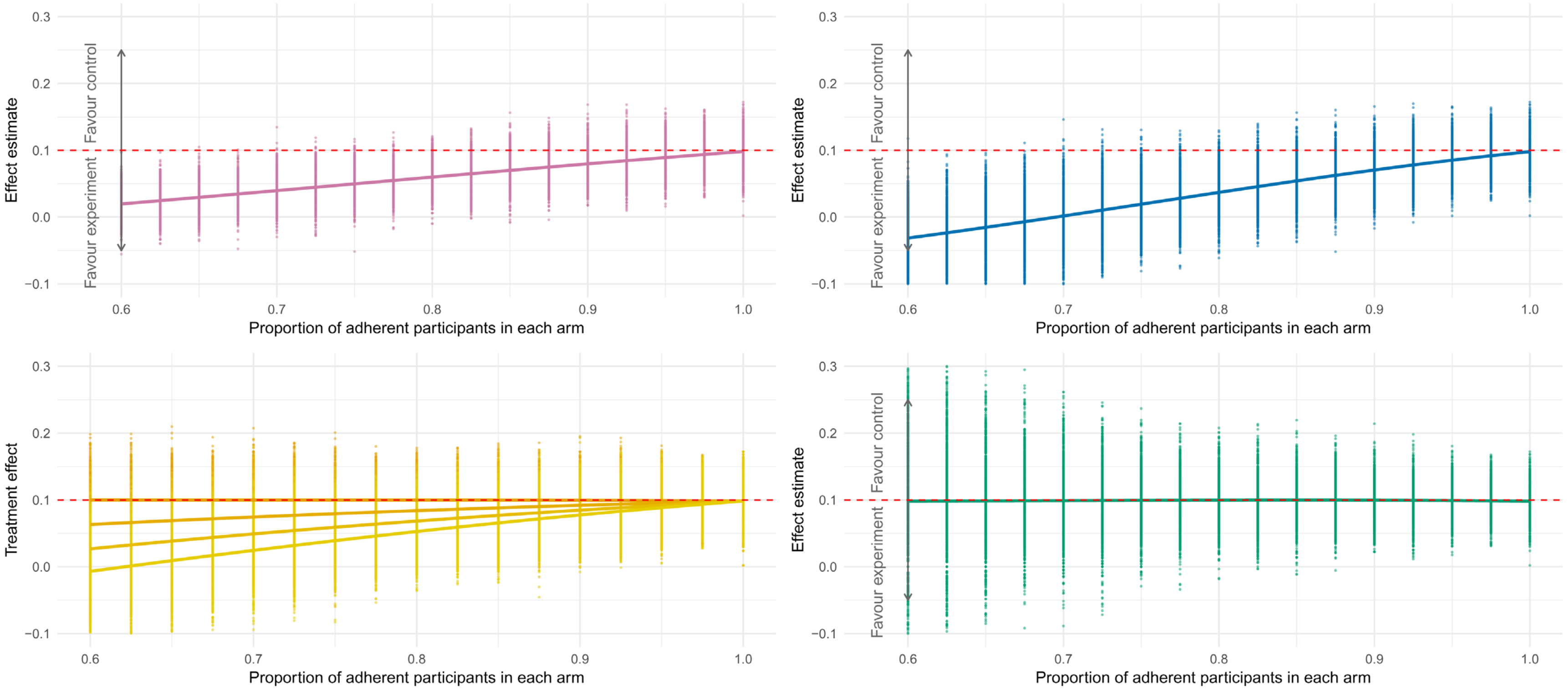

Figure 7. Non-adherence caused by both known and unknown confounders. Four confounders were added in the simulation. ITT, PP and instrumental variable analyses did not adjust for any confounders. For the inverse probability weighting analysis, the four lines represent situations when one, two, three and all of confounders were adjusted for. With more confounders accounted for, treatment estimates become less biased.

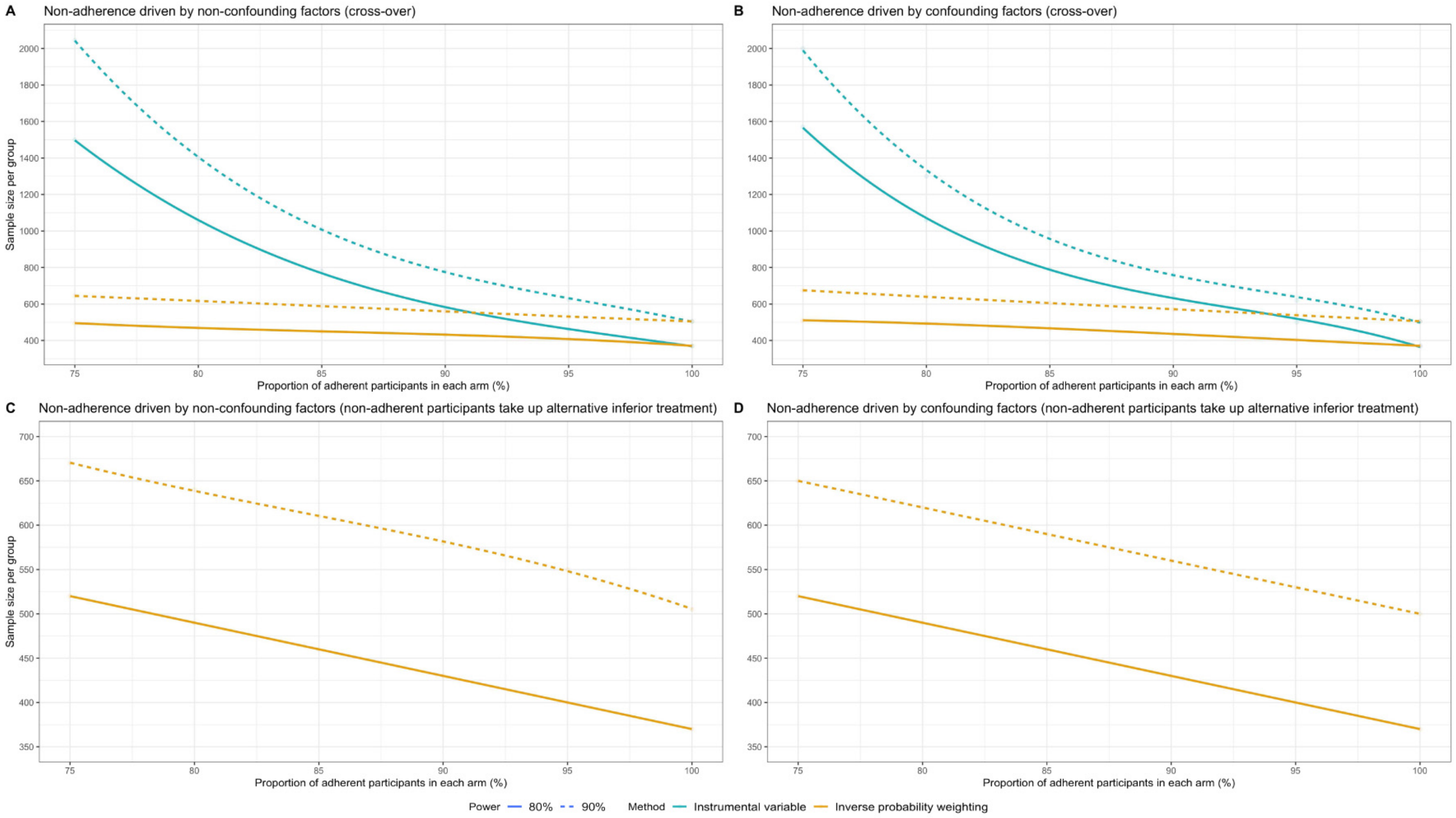

Figure 8. Decrease in adherence requires inflated sample sizes to maintain power. Panel **A** and **C** show the output the scenario where non-adherence is driven by non-confounding factors. Panel **B** and **D** show the output from simulating the scenario where non-adherence is driven by confounding factors. Panels **A** and **B** show the impact of cross-over type of non-adherence on power. The assumed true proportion of treatment failure in the experimental treatment was set to be the same as the control treatment at 40%. Panels **C** and **D** show the impact on power when non-adherent participants receive an alternative treatment inferior to the control and experimental treatments by 10%. In all the four scenarios, the non-inferiority margin was set at 10%.

participants who are anticipated to be non-adherent; the expected influence the confounder is likely to have on treatment failure; 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 is 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. This bias is increased when the confounder is more influential on the outcome.

Inverse probability weighting accounts for the difference in confounders between the allocation arms to ensure that the reweighted arms 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[20,21]. 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[22].

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[23] and g-estimation methods[24]. These methods are also used to address missing data and censoring[25,26]. Another limitation in our study is that non-adherence is either due to cross-over or switching to a treatment that is inferior to both the control and experimental treatments. In practice, both types of non-adherence may occur within the same trial. However, our simulations use these extreme examples to clarify the impacts of non-adherence on trial analyses and outcomes.

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 even in cases where expected non-adherence is as low as 5%, we propose that during the planning stage of clinical trials, 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.

Archived code as at time of publication: https://doi.org/10.5281/zenodo.3746705[27].

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

Extended data

Zenodo: Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study (Supplementary material), http://doi.org/10.5281/zenodo.3746706[18].

This project contains the following extended data:

- Supplementary material 1: Simulation models and analysis methods.
- Supplementary material 2: Supplementary figures.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Software availability

Power calculator accounting for non-adherence in a non-inferiority trial: https://moru.shinyapps.io/samplesize_nonadherence/

Acknowledgements

We thank Thomas Crellen and Jacob Armstrong for providing technical support in the programming of the simulations.

References

1. Wu Y, Zhao L, Hou Y, *et al.*: **Correcting for non-compliance in randomized non-inferiority trials with active and placebo control using structural models.** *Stat Med.* 2015; **34**(6): 950–65.
PubMed Abstract | Publisher Full Text

2. Parpia S, Julian JA, Thabane L, *et al.*: **Treatment crossovers in time-to-event non-inferiority randomised trials of radiotherapy in patients with breast cancer.** *BMJ Open.* 2014; **4**(10): e006531.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Sheng D, Kim MY: **The effects of non-compliance on intent-to-treat analysis of equivalence trials.** *Stat Med.* 2006; **25**(7): 1183–99.
PubMed Abstract | Publisher Full Text

4. Piaggio G, Elbourne DR, Pocock SJ, *et al.*: **Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement.** *JAMA.* 2012; **308**(24): 2594–604.
PubMed Abstract | Publisher Full Text

5. Committee for Proprietary Medicinal Products: **Points to consider on switching between superiority and non-inferiority.** *Br J Clin Pharmacol.* 2001; **52**(3): 223–8.
PubMed Abstract | Free Full Text

6. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), *et al.*: **Non-inferiority clinical trials to establish effectiveness: guidance for industry**. 2016; 6, Accessed Oct 2017.
Reference Source

7. Kim MY: **Using the instrumental variables estimator to analyze noninferiority trials with noncompliance.** *J Biopharm Stat.* 2010; **20**(4): 745–58.
PubMed Abstract | Publisher Full Text

8. Matilde Sanchez M, Chen X: **Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat.** *Stat Med.* 2006; **25**(7): 1169–81.
PubMed Abstract | Publisher Full Text

9. Hernán MA, Robins JM: **Per-Protocol Analyses of Pragmatic Trials.** *N Engl J Med.* 2017; **377**(14): 1391–8.
PubMed Abstract | Publisher Full Text

10. Uranga A, España PP, Bilbao A, *et al.*: **Duration of Antibiotic Treatment in Community-Acquired Pneumonia: A Multicenter Randomized Clinical Trial.** *JAMA Intern Med.* 2016; **176**(9): 1257–65.
PubMed Abstract | Publisher Full Text

11. Blackwelder WC: **"Proving the null hypothesis" in clinical trials.** *Control Clin Trials.* 1982; **3**(4): 345–53.
PubMed Abstract | Publisher Full Text

12. RStudio: **RStudio.** 2014; (accessed 7 May 2019).
Reference Source

13. Hernan MA, Robins JM: **Causal Inference.** Taylor & Francis. 2019.
Reference Source

14. Mansournia MA, Altman DG: **Inverse probability weighting.** *BMJ.* 2016; **352**: i189.
PubMed Abstract | Publisher Full Text

15. Angrist JD, Imbens GW, Rubin DB: **Identification of Causal Effects Using Instrumental Variables.** *J Am Stat Assoc.* 1996; **91**(434): 444–55.
Publisher Full Text

16. Robins J, Rotnitzky A: **Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models.** *Biometrika.* 2004; **91**(4): 763–83.
Publisher Full Text

17. Clarke PS, Windmeijer F: **Instrumental Variable Estimators for Binary Outcomes.** *J Am Stat Assoc.* 2012; **107**(500): 1638–52.
Publisher Full Text

18. Yin M, Cherry L, Mavuto M, *et al.*: **Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study (Supplementary material).** 2019.
http://www.doi.org/10.5281/zenodo.3746706

19. Miettinen OS: **Stratification by a multivariate confounder score.** *Am J Epidemiol.* 1976; **104**(6): 609–20.
PubMed Abstract | Publisher Full Text

20. Rudolph KE, Stuart EA: **Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods.** *Am J Epidemiol.* 2018; **187**(3): 604–13.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Stürmer T, Schneeweiss S, Avorn J, *et al.*: **Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration.** *Am J Epidemiol.* 2005; **162**(3): 279–89.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Vansteelandt S, Didelez V: **Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators.** *Scand Stat Theory Appl.* 2018; **45**(4): 941–61.
Publisher Full Text

23. Hernán MA, Lanoy E, Costagliola D, *et al.*: **Comparison of dynamic treatment regimes via inverse probability weighting.** *Basic Clin Pharmacol Toxicol.* 2006; **98**(3): 237–42.
PubMed Abstract | Publisher Full Text

24. Robins J, Hernan M: **Estimation of the causal effects of time-varying exposures.** Chapman & Hall/CRC Handbooks of Modern Statistical Methods. 2008; 553–99.
Reference Source

25. Seaman SR, White IR: **Review of inverse probability weighting for dealing with missing data.** *Stat Methods Med Res.* 2013; **22**(3): 278–95.
PubMed Abstract | Publisher Full Text

26. Bang H, Robins JM: **Doubly robust estimation in missing data and causal inference models.** *Biometrics.* 2005; **61**(4): 962–73.
PubMed Abstract | Publisher Full Text

27. moyinNUHS: **moyinNUHS/NItrialsimulation: Statistical considerations in the design and analysis of non-inferiority trials with binary endpoints in the presence of non-adherence: a simulation study (Version v1.1.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3746705

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 06 May 2020

https://doi.org/10.21956/wellcomeopenres.17386.r38525

✓ **Mimi Y. Kim**

Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

I very much appreciate the authors' thorough responses to my comments and the revisions to the paper appear appropriate.

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.

Reviewer Report 04 May 2020

https://doi.org/10.21956/wellcomeopenres.17386.r38526

✓ **Matteo Quartagno**

MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK

Thanks for addressing my previous comments. I have no additional comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical Trials, Statistics.

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.

Version 1

Reviewer Report 26 February 2020

https://doi.org/10.21956/wellcomeopenres.17132.r37977

✓ **Matteo Quartagno**
MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK

Summary: This is a well-planned and well-written paper, explaining some known, yet poorly handled, issues in non-inferiority trials affected by non-adherence. The paper provides results of simulations, and I particularly enjoyed how simple and effective was the presentation of results in graphs. Conclusions of the paper are also reasonable. I only have a few minor comments that I believe might improve the paper:

Comments:

1. I think one important point that you should try to stress more, is that you only considered non-adherence to one of the two arms. Potentially, in non-inferiority trials both arms could suffer from non-adherence, as patients could also stop getting active control treatment. This is important to note particularly because non-adherence in both arms could complicate the implementation of Instrumental Variables methods. It would be good to mention the general problem somewhere at the beginning of the paper and the specific issue with IV in the discussion.

2. Similarly, here you only considered binary non-adherence. In reality, patients might get treatment only for some time, then switch, then maybe switch back, etc etc. Of course, I agree it was a good idea to start from a simpler example, but I still think it would be good to both (i) clarify that this is the case and possibly (ii) justify your example a bit better. Your example is a duration reduction trial, so isn't it the typical situation where it is difficult to think of adherence in binary terms? Could you explain in words exactly what would be adherence or non-adherence in that hypothetical trial?

3. Please clarify across the paper, e.g. when you say that significance level was 0.025 in the methods section, that you used one-sided tests. In section "Non-inferiority hypothesis testing", it should be 97.5% one-sided or 95% two-sided (in your examples).

4. In Figure 4, it seems like with no non-adherence, type 1 error is well below the expected 2.5%. With 1000 simulations, this seems well outside the Monte-Carlo confidence interval. Could you please (i) check that there is nothing wrong with these simulations and (ii) add to the plot the two lines for the Monte-Carlo confidence interval, so that one can immediately check whether results are outside regular random variation. This is partly the case also for Figure 5.

5. In the final recommendations, you say that researchers should consider consequences of non-adherence if they expect at least 5% non adherence. I personally believe it is always better not to give specific numbers, as these could be successively used to justify choices without much critical thinking. I would rather prefer a more general sentence like "especially those who expect some degree of non-adherence".

Typos / formatting / plots:

1. Reference for the FDA guidelines: FDA considered like a given name, please fix.

2. Sometimes you put citations at the end of a sentence, rather than next to the name of the author. See, e.g., citation 7 and 8, page 3, left column. I'd personally prefer this latter citation format, but feel free to ignore, particularly if this is journal requirement.

3. Page 4, last paragraph of left column, you wrote: *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.*
Please fix punctuation.

4. I like a lot how you visually presented results of simulations. When it comes to methods, though, I would personally prefer to have a simple table of all the simulation scenarios you explored. To me, Figure 2 looks much more convoluted than a simple Table where different dimensions were columns and different scenarios rows. Could you provide that sort of table at least as additional online material?

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

Yes

Is the description of the method technically sound?

Yes

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

Yes

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: Clinical Trials, Statistics.

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.

Author Response 09 Apr 2020

Yin Mo, Mahidol University, Bangkok, Thailand

Dear Dr Quartagno,

Thank you very much for your detailed review and very thoughtful comments. We hope to have addressed all of your comments in version 2 of the manuscript.

1. I think one important point that you should try to stress more, is that you only considered non-adherence to one of the two arms. Potentially, in non-inferiority trials both arms could suffer from non-adherence, as patients could also stop getting active control treatment. This is important to note particularly because non-adherence in both arms could complicate the implementation of Instrumental Variables methods. It would be good to mention the general problem somewhere at the beginning of the paper and the specific issue with IV in the discussion.

Apologies for not being clear in the manuscript that our simulations do address non-adherence in one arm and both arms. We have clarified this in all the graphs by relabelling the x-axes: “Proportion of adherent participants in each arm”. In the new supplementary 2, we include figures from all simulated scenarios, including those where participants from one arm are non-adherent and those where participants from both arms are not adherent.

2. Similarly, here you only considered binary non-adherence. In reality, patients might get treatment only for some time, then switch, then maybe switch back, etc etc. Of course, I agree it was a good idea to start from a simpler example, but I still think it would be good to both (i) clarify that this is the case and possibly (ii) justify your example a bit better. Your example is a duration reduction trial, so isn’t it the typical situation where it is difficult to think of adherence in binary terms? Could you explain in words exactly what would be adherence or non-adherence in that hypothetical trial?

We agree that this may be a limitation in this simulation study as it is only applied to time-fixed interventions. We have added in the methodology: “Although adherence was considered as a binary variable in the simulations, in the scenario where participants received less effective treatments than the control and experimental treatments may reflect partial adherence to either treatments.” To further explain why in an antibiotic duration trial, adherence is often considered binary, we have added: “With this single end-point, we consider adherence as a binary variable where non-adherent patients in the short arm would receive longer than five days of treatment, and non-adherent patients in the long arm would receive than less than five days of treatment.”

3. Please clarify across the paper, e.g. when you say that significance level was 0.025 in the methods section, that you used one-sided tests. In section “Non-inferiority hypothesis testing”, it should be 97.5% one-sided or 95% two-sided (in your examples).

We agree with this recommendation and have clarified in the methodology section that “The null hypothesis is tested by comparing the upper bound of the two-sided 95% confidence interval of the effect estimate with the non-inferiority margin.”

4. In Figure 4, it seems like with no non-adherence, type 1 error is well below the expected 2.5%. With 1000 simulations, this seems well outside the Monte-Carlo confidence interval.

Could you please (i) check that there is nothing wrong with these simulations and (ii) add to the plot the two lines for the Monte-Carlo confidence interval, so that one can immediately check whether results are outside regular random variation. This is partly the case also for Figure 5.

We have changed the way we derive the standard errors of the simulated treatment estimates from using standard formulae to empirically from the simulated distributions. We have updated these graphs showing type 1 error rates close to the nominal 0.025 level. We did not include the Monte-Carlo confidence intervals as the type 1 error rates now fall nicely on the expected value of 0.025, and adding further shadings and lines for confidence intervals may reduce the prominence of the lines formed by estimates.

5. In the final recommendations, you say that researchers should consider consequences of non-adherence if they expect at least 5% non adherence. I personally believe it is always better not to give specific numbers, as these could be successively used to justify choices without much critical thinking. I would rather prefer a more general sentence like "especially those who expect some degree of non-adherence".

We agree with this concern, and have changed this sentence to “In conclusion, given the potential inflation in the probability of concluding non-inferiority with non-adherence even in cases where expected non-adherence is as low as 5%, we propose that during the planning stage of clinical trials investigators should anticipate the likely patterns and magnitude of non-adherence and devise ways to reduce it.”

6. Reference for the FDA guidelines: FDA considered like a given name, please fix.

We have amended the reference to “Center for Drug Evaluation and Research (Food and Drug Administration). Non-inferiority clinical trials to establish effectiveness: guidance for industry 2016. Accessed Oct 2017;6.” in version 2 of the manuscript.

7. Sometimes you put citations at the end of a sentence, rather than next to the name of the author. See, e.g., citation 7 and 8, page 3, left column. I’d personally prefer this latter citation format, but feel free to ignore, particularly if this is journal requirement.

We will liaise with the journal editors to ensure the format of the references are correct.

8. Page 4, last paragraph of left column, you wrote: 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. Please fix punctuation.

We have removed the comma after ‘second’.

9. I like a lot how you visually presented results of simulations. When it comes to methods, though, I would personally prefer to have a simple table of all the simulation scenarios you explored. To me, Figure 2 looks much more convoluted than a simple Table where different dimensions were columns and different scenarios rows. Could you provide that sort of table at least as additional online material?

We have now included this table in supplementary 2 with all the simulation scenarios included.

Please do not hesitate to contact us again for further comments. Once again, we would like to express our gratitude towards your insightful comments!

Yours sincerely,
Mo Yin

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 January 2020

https://doi.org/10.21956/wellcomeopenres.17132.r37440

?

Mimi Y. Kim

Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

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 the cross-over case.

- 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 in 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.

Author Response 09 Apr 2020

Yin Mo, Mahidol University, Bangkok, Thailand

Dear Professor Kim,

Thank you very much for your detailed review and very thoughtful comments. We hope to have addressed all of your comments in version 2 of the manuscript.

Below we describe the specific improvements and corrections we made according to your recommendations:

1. 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.

We agree that a limitation of the simulations is that non-adherence is either due to cross-over or switching to an alternative treatment that is inferior to both the control and experimental treatments. We have acknowledged this in the discussions by adding “Another limitation in our study is that non-adherence is either due to cross-over or switching to a treatment that is inferior to both the control and experimental treatments. In practice, both types of non-adherence may occur within the same trial. However, our simulations use these extreme examples to clarify the impacts of non-adherence on trial analyses and outcomes.” We have also acknowledged in the manuscript that non-adherence is a binary variable in the simulations.

2. 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.

Thank you for this valuable feedback. We have updated the methodology sections in both the manuscript and the supplementary material to include the case where non-adherent participants receive an alternative treatment inferior to both the control and experimental treatments. The presumed probability of failure for the inferior alternative treatment is also included in the methodology section in version 2 of the main manuscript.

3. 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 the cross-over case.

We have added descriptions of our key findings from the scenarios where subjects could receive no treatment or one that is inferior to the trial treatments in the results section. We have also included all 18 simulation scenarios in version 2 of the supplementary material 2.

4. 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)."

We have amended this error i.e. changed -0.1 to 0.1.

5. 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.
All the scenarios are explored in the simulations. This includes 6 on the left panel (non-adherence driven by non-confounding factors) and 12 on the right panel (non-adherence driven by confounding factors.

To improve clarity, we have included graphs produced from all the scenarios in the version 2 of the supplementary material 2.

6. 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?

We have changed the x-axis label from 'proportion of adherent participants' to 'proportion of adherent participants in each arm' such that the proportion of adherent participants is the same in both groups.

7. 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 effect of the confounder (increasing confounder value decreases the probability of taking up the experimental treatment) has a corresponding increase in the probability of taking up the control treatment. We have added 'This is such that participants with the highest confounder values in the experimental arm cross over to the control arm, and participants with the lowest confounder values in the control arm cross over to the experimental arm.' to further explain this point. We have also changed the captions of Figure 4 and 5 to "participants with higher confounder values have lower/higher probability of taking up the experimental treatment regardless of the allocation, and lower/higher probability of treatment failure" improve clarity.

8. 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").

We have added to our shiny app an additional input that includes the scenario where non-adherent participants take up an alternative treatment. We have also clarified the effect of the confounder on the shiny app by elaborating on the tabs: "Participants with high confounder values from the standard-of-care arm tend to be non-adherent; participants with low confounder values from from the experimental arm tend to be non-adherent" and "Participants with low confounder values from the standard-of-care arm tend to be non-adherent; participants with high confounder values from from the experimental arm tend to be non-adherent".

9. 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.

We have added an additional figure (figure 5) to highlight that the more influence the confounder has on treatment failure, the more biased per protocol estimates will be, leading to higher type 1 error rates.

10. 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.

We have changed the way we derive the standard errors of the simulated treatment estimates from using standard formulae to empirically from the simulated distributions. We have updated these graphs showing type 1 error rates close to the nominal 0.025 level.

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

Thank you for pointing out his typo. This notation error has been corrected from $\Pr[Y^{A=1,\ Z=1} = 1]$ - $\Pr[Y^{A=1,\ Z=1} = 1]$ to $\Pr[Y^{A=1,\ Z=1} = 1]$ - $\Pr[Y^{A=0,\ Z=0} = 1]$.

12. 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.

We have included "Non-inferiority is concluded if the upper bound of the 95% confidence interval for the absolute risk difference between the experimental and control treatments is less than the non-inferiority margin.” in the non-inferiority hypothesis testing.

Once again, we thank you for your comments and please do not hesitate to raise further queries!

Your sincerely,
Mo Yin

Competing Interests: No competing interests were disclosed.