Impact of per-protocol analysis on treatment effects in randomised controlled trials: a meta-epidemiological study

Mohammod Mostazir,[1] Gordon Taylor,[2,3] William Edward Henley,[3] Edward Robert Watkins,[1] Rod S Taylor[3,4]

[1] School of Psychology, College of Life & Environmental Sciences, University of Exeter, EX4 4QG, United Kingdom; [2]Research Design Service South West (RDS SW), National Institute of Health Research (NIHR) South West, United Kingdom; [3]Institute of Health Research, University of Exeter Medical School, University of Exeter, EX1 2LU; [4]MRC/CSO Social and Public Health Sciences Unit & Robertson Centre for Biostatistics, Institute of Health and Well Being, University of Glasgow, G2 3AX

Declarations of interest: None

Correspondence to:
Mohammod Mostazir
Research Fellow in Statistics
School of Psychology, College of Life and Environmental Sciences (CLES)
University of Exeter, EX4 4QG, Exeter UK
Telephone: +44 1392 726629
Mobile: +44 7432516771
Email: m.mostazir@exeter.ac.uk

Abstract

Objective

To undertake meta-analysis and compare treatment effects estimated by the intention-to-treat (ITT) method and per-protocol (PP) method in randomised controlled trials (RCTs). PP excludes trial participant who are non-adherence to trial protocol in terms of eligibility, interventions, or outcome assessment.

Study design and setting

Five high impact journals were searched for all RCTs published between July 2017 to June 2019. Primary outcome was a pooled estimate that quantified the difference between the treatment effects estimated by the two methods. Results are presented as ratio of odds ratios (ROR). Meta-regression was used to explore the association between level of trial protocol non-adherence and treatment effect. Sensitivity analyses compared results with varying within-study correlations and across various study characteristics.

Results

Random-effects meta-analysis (N = 156) showed that PP estimates were on average 2% greater compared to the ITT estimates (ROR: 1.02, 95% CI: 1.00 to 1.04, p = 0.03). The divergence further increased with higher degree of protocol non-adherence. Sensitivity analyses reassured consistent results with various within-study correlations and across various study characteristics.

Conclusions

There was evidence of larger treatment effect with PP compared to ITT analysis. PP analysis should not be used to assess the impact of protocol non-adherence in RCTs. Instead, in addition to ITT, investigators should consider randomisation based casual method such as Complier Average Causal Effect (CACE).

Key words: Non-adherence, randomised controlled trial, intention-to-treat, per-protocol, meta-epidemiology, complier average causal effect (CACE)

Running title: Impact of per-protocol method on treatment effects in RCTs.

(**Word count:** 193, excluding title, headings, key words and running title)

What is already known

- Per-protocol (PP) analysis is restricted to trial participants who fulfil trial protocol adherence in terms of eligibility, interventions, and outcome assessment.
- PP method of treatment effect estimation is subject to confounding and selection bias

What this study adds

- PP method inflates treatment effect compared to intention-to-treat (ITT) method
- The divergence increases with higher degree of protocol non-adherence
- Complier average causal effect method estimated treatment effect consistent with ITT
- RCTs seeking to explore the impact of protocol non-adherence should avoid PP and instead report CACE to ITT primary analysis

Introduction

Randomised controlled trials (RCTs) are the best source of unbiased estimates of treatment effect. However, RCTs often experience non-adherence to protocol which can take a variety of forms i.e., non-adherence to the treatment protocol (e.g., not starting treatment at all, taking a suboptimal dose at the incorrect time, or discontinuation or switching of treatment) and non-adherence to trial protocol (e.g., inappropriately randomised participants, outcome completion outside the assessment time window, trial discontinuation). The current guidance, including the Consolidated Standards of Reporting Trials (CONSORT),[1] recommend treatment effect from RCTs need to be estimated following intention-to-treat (ITT) strategy i.e. analysing data based on patient randomization regardless whether patients adhered to protocol or not. The fundamentals of ITT strategy are underpinned in Rubin's causal model (RCM) framework[2-5] which evaluates ITT estimate as the difference between average outcome of the treatment group vs. average outcome of the control group without accounting for non-adherence to treatment protocol. Under perfect adherence, ITT thus has the ability to infer 'causal effect' of the intervention as randomisation makes the groups balanced in all aspects except the treatment being offered. The 'causal effect' therefore is a measure of both 'treatment effectiveness' (i.e., a measure of how the adopted treatment policy performs in everyday practice) and 'treatment efficacy' (i.e., a measure of how well the treatment works under perfect adherence and controlled conditions).[6] Since ITT does not account for protocol nonadherence, under suboptimal adherence, ITT underestimates the intervention effect and cannot accurately assess 'treatment efficacy'.[7]

Methods of analyses commonly used to address the issue of non-adherence to protocol that provide an estimate of treatment efficacy include per-protocol (PP) and as-treated (AT) analyses.[8] PP method traditionally includes only trial participants who were defined to adhered to the trial protocol, including treatment. Exclusion of participants from analysis this way can lead to estimating a treatment effect that is prone to selection bias and confounding.[9, 10] As-treated analysis on the other hand introduces bias by simply analysing

participants according to the treatment received regardless of their randomised allocation, and thus destroys the original randomisation.[7] Although in the presence of protocol non-compliance, ITT, PP, and AT all are likely to provide biased estimates, ITT is shielded against selection bias and from known/unknown confounders whilst such claims cannot be established for PP and AT.[7, 10]

Despite the drawbacks of PP analyses, reporting of effect estimates based on PP in RCT reports remains relatively common. Our recent review of RCTs found 52% of the studies reported results using PP analysis[11] and a similar proportion (47%) was reported by another review of a random sample of 100 RCTs.[12] To our knowledge, in a meta-epidemiological setting, no studies have formally quantified the extent to which the PP analysis diverges from the gold-standard ITT analysis. We therefore sought to compare the treatment effects estimated using PP versus ITT analysis in a sample of RCTs.

Methods

We conducted this study and reported findings in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.[13]

Identification of RCTs

We searched the full-text of all RCTs published in a two year period (1st July 2017 to 31st June 2019) in five high impact general medical journals: *The Lancet, New England Journal of Medicine (NEJM), British Medical Journal (BMJ), Journal of American Medical Association (JAMA)*, and *The Annals of Internal Medicine*. The combined search terms included "intention-to-treat" AND "per-protocol".

Definitions

In accordance with the CONSORT guidelines,[14] in this study we defined ITT analysis as "*A strategy for analyzing data in which all participants are included in the group to which they*

were assigned, whether or not they completed the intervention given to the group" and PP analysis as an "*Analytical strategy restricted to only participants who fulfil the protocol in terms of eligibility, interventions, and outcome assessment*".

Some authors reported ITT results labelled as modified-intention-to-treat (mITT) analysis[15] where patients were excluded if a certain minimum dose of intervention was not received.[16-18] These studies also provided separate PP analysis based on complete adherence to study protocol. We included these studies as they were reported i.e., mITT for ITT estimates but we also conducted separate analysis excluding those studies as they may be subject to selection bias.[12] In some other studies authors used the term mITT to indicate that not all patients as randomised were available for analysis i.e., analysis included only those patients for whom an outcome measure was available.[19, 20] In such situations, although authors used the term mITT, we considered those as valid ITT analysis. We included all RCTs irrespective of their design or testing hypothesis i.e., superiority, non-inferiority, and equivalence trials. RCTs that reported only an estimate of treatment effect based on ITT or PP analysis only were excluded. Study selection was undertaken by a single reviewer (MM).

Data extraction and assessing risk of bias

Data on the trial characteristics such as year of publication, sample size, clinical area of investigation, intervention type, trial type, single/multi-centre trial, funding source, blinding, sequencing, allocation concealment, placebo/active control, duration, type of outcome, proportion adhering to trial protocol , and participant characteristics were extracted using a predefined data extraction form. From each included trial, we extracted data on the named primary outcome. Where no primary outcome was defined, the first outcome reported in the abstract was used. To assess small study publication bias, the Harbord test[21] was used for studies with binary outcomes and the Egger test[22] for studies with continuous outcomes. One reviewer (MM) undertook data extraction and risk of bias assessment, and a random sample of 10% of all data were checked for accuracy by a second reviewer (GT).

Data analysis

The sample size for this study was based on findings from our pilot study that we carried out to develop the study protocol for this study (see study protocol, Appendix-A). Since there is no prior estimate of the magnitude of treatment difference between PP and ITT, we powered this study based on a 5% relative difference between PP and ITT estimate an average level of effect observed in our pilot study and considered to be clinically meaningful. We estimated that we required 85 studies to achieve 90% power with alpha = 0.05 and a sampling frame of 304 to obtain the required 85 studies. All main analyses were prespecified. The differences in intervention effects comparing the ITT and PP estimates were quantified using ratios of odds ratios: $ROR=OR_{PP}/OR_{ITT}$. As the primary outcomes of the included trials varied in their nature, we needed to convert them into a common metric for pooled comparison. Both binary and continuous outcomes providing number of event/means/standard deviation (SD) information respectively were converted to log odds ratio, with continuous outcomes first being converted to the standardised mean difference (Hedges' g). We used the reported 95% confidence intervals (CI) to construct the variance for studies those presented treatment effects in ratios but did not provided any information on number of events occurred.[23] Studies that reported treatment effects for multiple groups were combined where possible or discarded to avoid unit of analysis error that rises from double counting samples.[24] All outcome conversion formulae can be found elsewhere.[24, 25] Given that we did not have access individual participant data, our estimates are not adjusted for stratification or minimisation variables applied in the original trials.

The primary interest of our meta-analysis was to test the null hypothesis that reported PP estimates in RCTs are no different than ITT estimates [ROR = log(PP) – log(ITT) = 0]. Since sample size varies for the ITT and PP populations, the within study variance for outcome was estimated combining both ITT and PP populations for each study.[26] Further, because the ITT and PP estimates are sourced from the same study, a moderate to strong correlation

of r = 0.50 was assumed to adjust for within-study correlated outcomes.[25] Details of model derivation and variance estimation can be found in the appendix (Appendix-B).

We present our primary outcome of interest (pooled ROR: difference between PP vs. ITT) using maximum likelihood based random effects model. Pooled ROR is also presented stratified by various study characteristics (i.e., intervention area, superiority/inferiority trial, funding, blinding, randomisation sequencing, allocation concealment, placebo/active control, follow-up duration, outcome type, mITT analysis, and by ITT results where positive trials indicate significant treatment effect reported in the ITT analyses at least at $p<0.05$ and negative trials indicate no-difference between treatment and controls) with p-value for their interaction with the outcome. Multivariate meta-regression analysis was used to explore the effect of protocol non-adherence on the effect size where proportion of non-adherence was calculated as (1–[total number of patients in intervention group in PP population / number randomised to intervention group in ITT population]). Effect of non-adherence was also sought separately for studies where non-adherence was exclusively related to the intervention. Meta-regression models were further adjusted with the study characteristics mentioned above. Separate sensitivity analyses were carried out excluding the studies with mITT estimates and for a range of within-study correlations (i.e., r = 0.10 to 0.90).

As an exploratory analysis, we also applied the 'complier average causal effect' (CACE)[27,28] method for all studies that reported the relevant number of events/mean/SD information and compared the pooled CACE estimate against ITT and PP. To derive CACE estimates, we assumed the principal stratum 'compliers' to be those in the treatment group who were included in the PP population as authors claimed that these participants adhered to the protocol. The rest of the information for other strata were derived following CACE method. The ROR for CACE vs ITT/PP estimates were calculated as explained above.
All analyses were carried out using Stata version-16.

Patient and public involvement

Our study did not involve any patient or member of the public in the design, or conduct, or reporting, or dissemination plans of the research.

Results

Study selection and study characteristics

A total 334 studies were initially identified by our searches. Of these, 156 (47%) were included. One study had four independent groups with '2 X2' factorial design which provided an additional record for analysis forming the analyses sample to 157. Figure 1 shows the flow of data through the study, from screening to the final dataset. The major reason for exclusion at full text review was the PP analysis estimates not being reported (N = 83). Table 1 presents the study level characteristics (N = 156). Most studies were multi-centred (96%) non-blinded (58%) drug intervention (57%) superiority trials (70%) with a follow-up duration of 6-months or more (54%). A total of 123 (79%) studies presented all the necessary data required for meta-analysis using both ITT and PP analysis (binary outcomes: 105/133; Continuous outcomes: 18/23). The median non-adherence rate to the trial protocol was 9% (range 1 to 77%). Authors of 3 studies did not provide any definition of their per-protocol population while for N = 80 (51%) studies, the PP population excluded patients who were non-adherent to the treatment protocol other than any other protocol violations. N = 74 studies (47%) excluded patients from the PP population for multiple reasons inclusive of treatment non-adherence where the other reasons being discontinuation/withdrawal of treatment, switching treatment, withdrawal for adverse effects, developing exclusion criteria, accidental blinding, lost to follow-up etc.

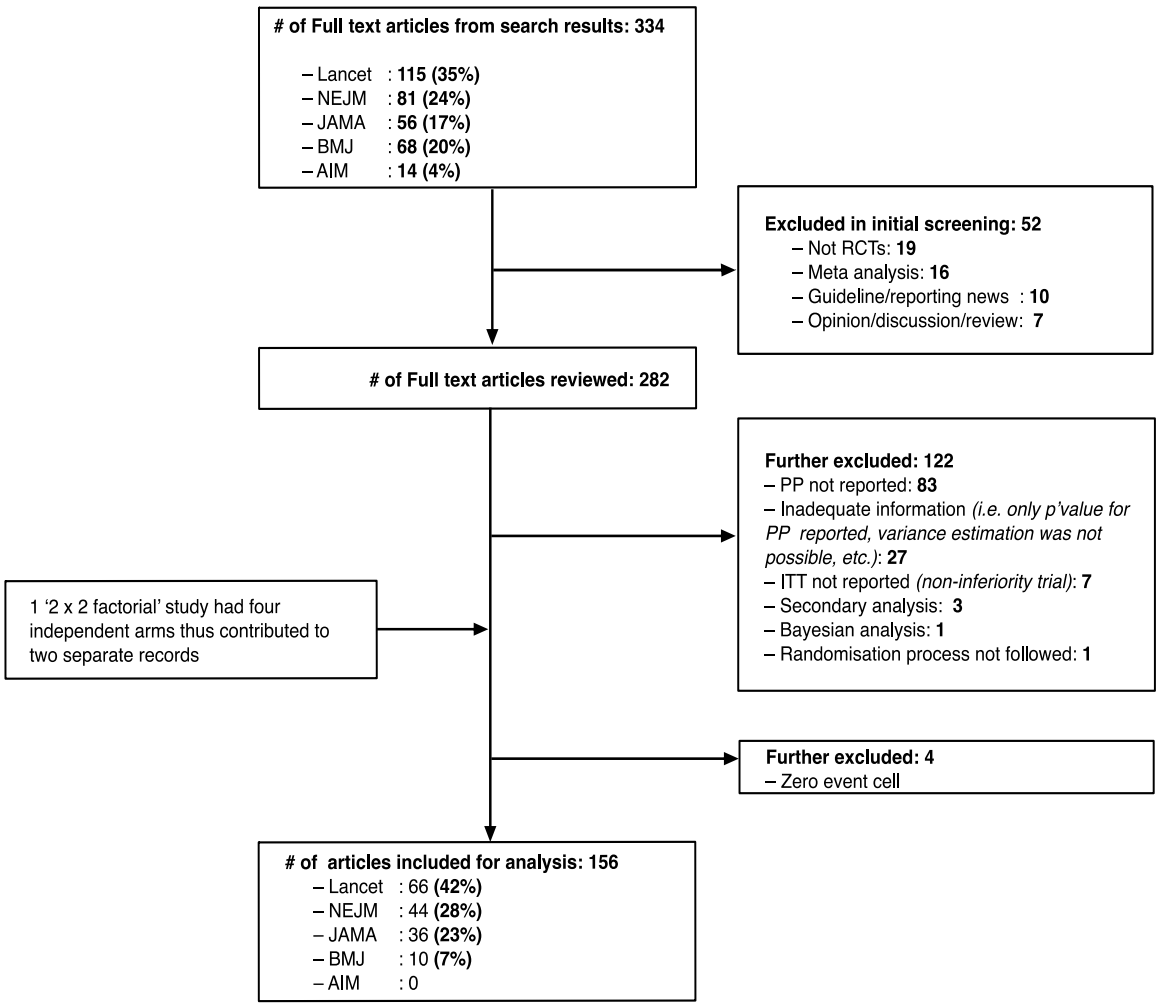


Fig 1: Flow diagram of study inclusion and exclusion process

Characteristics	# of Studies (%)
Number of studies	156 (100)
Year of publication	–
2017	21 (13)
2018	82 (53)
2019	53 (34)
Journals	–
Lancet	66 (42)
New England Journal of Medicine	44 (28)
Journal of the American Medical Association	36 (23)
British Medical Journal	10 (6)
Annals of Internal Medicine	0 (0)
Intervention clinical area	–
Cardiology	39 (25)
Oncology	18 (12)
Gynaecology/Obstetrics/Paediatrics	18 (12)
Osteoarthritis/Inflammatory	11 (7)
Virology/Immunology/Infectious disease	12 (8)
Neurology	10 (6)
Pulmonology/Respiratory	10 (6)
Nephrology/Hepatology/Gastroenterology	9 (6)
Others	29 (19)
Intervention type	–
Drug	89 (57)
Surgical/Medical device/Others	67 (43)
Study type	–
Superiority trial	109 (70)
Non-inferiority trial	45 (29)
Equivalence trial	2 (1)
Single/multi-centre	–
Multi-centre trial	150 (96)
Single-centre trial	6 (4)
Funding type	–
Govt./charity/public institution funded	98 (63)
Pharmaceutical/Bio-technological company funded	58 (37)
Blinding type	–
Unblinded/Open label	90 (58)
Double blinding	49 (31)
Single blinded	17 (11)
Randomization sequence generation	–
Adequate	106 (68)
Inadequate	50 (32)
Allocation concealment	–
Concealed	90 (58)
Not concealed	66 (42)
Placebo/Active control	–
Active control/Usual treatment	121 (78)
Placebo	35 (22)
Follow up duration	–
> 6 months	85 (54)
6 months or less	71 (46)
Type of primary outcome	–
Binary/Ratio	133 (85)
Continuous	23 (15)
Non-adherence	–
%Non-adherence in treatment group: Median%, range	9%, 1-77
Non-adherence type	
Treatment protocol non-adherence only	80 (51%)
Treatment or trial protocol non-adherence	74 (47%)
Undefined	3 (2%)
Participant characteristics	–
Participant recruited: Mean (SD)	1752 (3532)
Age: Mean (SD)	50 (20)
%Male	156 (54)

All characteristics are expressed in frequency and percent unless expressed otherwise

Table 1: Study-level characteristics of included RCTs

Pooled comparison between PP and ITT estimates

Table 2 presents the results for primary and exploratory analyses. A total of 157 comparisons across 156 trials contributed data for meta-analysis. The pooled ROR showed that trials' reported PP estimates were on an average 2% larger compared to their ITT estimates, rejecting the null hypothesis that the PP estimates are not significantly different to the ITT estimates (ROR: 1.02, 95% CI: 1.00 to 1.04, p = 0.03). When analysis was restricted the studies those reported PP estimates due to the non-adherence to treatment protocol, the effect was slightly larger (N = 80; ROR: 1.04, 95% CI: 1.01 to 1.06, p = 0.01).

Analyses type	**N**	**Ratio of odds Ratio (ROR) (95% CI)**	**Study heterogeneity τ^2 (I^2%)**
Primary outcome: Comparing PP vs. ITT	**–**	**–**	–
PP vs. ITT (All studies)	157	1.02 (1.00 to 1.04)*	0.00 (0.00)
Meta regression (PP vs. ITT)[a]	154	–	–
%Nonadherence	–	1.02 (1.00 to 1.04)**	–
ITT results (Positive vs. Negative trials)[b]	–	1.06 (1.01 to 1.11)*	–

[a]Meta-regression: Effects of other study characteristics were not significant. Regression coefficient corresponds to 10% increase in non-adherence
[b]'Positive trials': statistical superiority of treatment versus control group. 'Negative trials': no statistical superiority of treatment versus control group or statistical superiority of control versus treatment group
*denotes p<0.05, **p<0.01; τ^2: Between study variance; I^2: Heterogeneity statistic

Table 2: Primary outcome and meta regression

Impact of the level of non-adherence

Meta-regression results in Table 2 shows that the protocol non-adherence rate was significantly associated with greater PP estimates (ROR: 1.02, 95% CI: 1.00 to 1.04, p = 0.02). Since non-adherence rate is in proportion scale, the coefficient was rescaled dividing by 10 so that one-unit increase corresponds to 10% non-adherence. The divergence in ROR estimates compared to the ITT increased with increasing levels of protocol non-adherence. Figure 2 shows that at non-adherence rate 10%, PP estimates were significantly different from ITT (p = 0.04) and diverging further from the ITT estimates significantly as non-adherence rate increases.

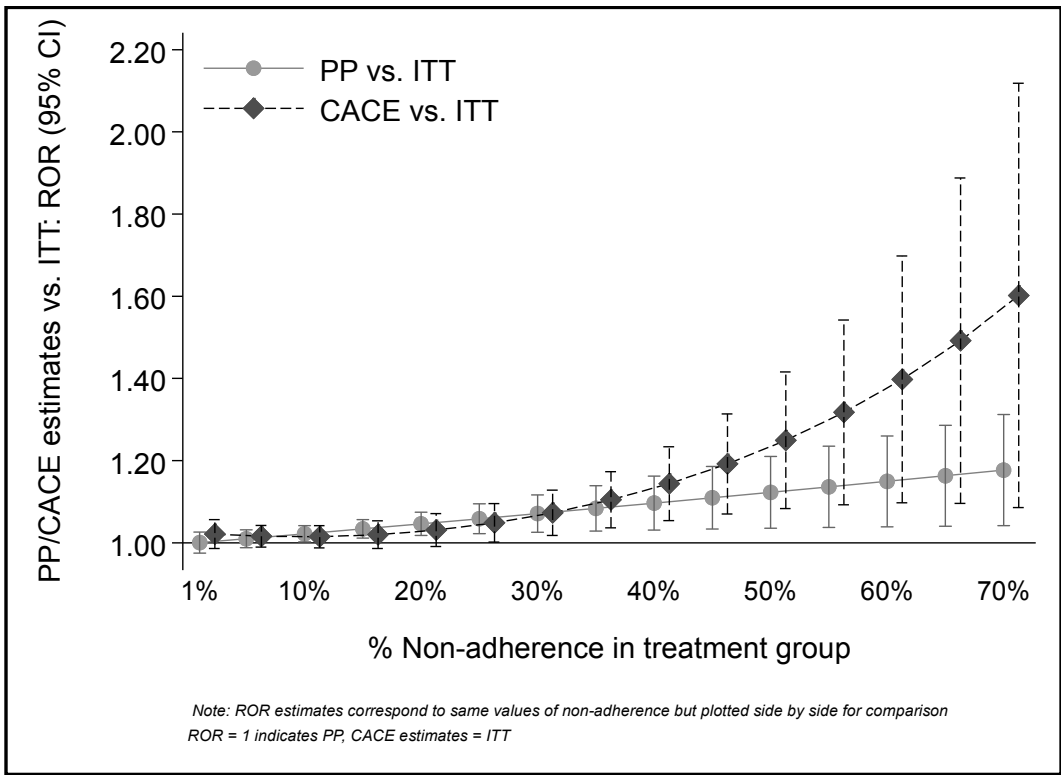


Fig 2: Comparison of PP vs ITT and CACE vs ITT estimates against level of non-adherence in treatment group

Subgroup comparisons and sensitivity analyses

24 trials employed some forms of mITT analyses and excluding these trials did not change the overall effect size estimated earlier (ITT vs PP - ROR: 1.03, CI: 1.00 to 1.05, p = 0.03). Figure 3 presents the overall pooled estimate with heterogeneity statistics besides presenting effect sizes by study characteristics. None of the differences between the study characteristics showed any statistical significance apart from the studies reporting significant ITT treatment effect which also appear to have presented larger PP estimate (p = 0.03). Sensitivity analyses using wide range of within-study bivariate correlations (r = 0 ~ 0.90) between the ITT and PP showed no change in ROR estimates and p-values remained significant with correlation as strong as r = 0.90 (Supplementary Figure 1, panel-a).

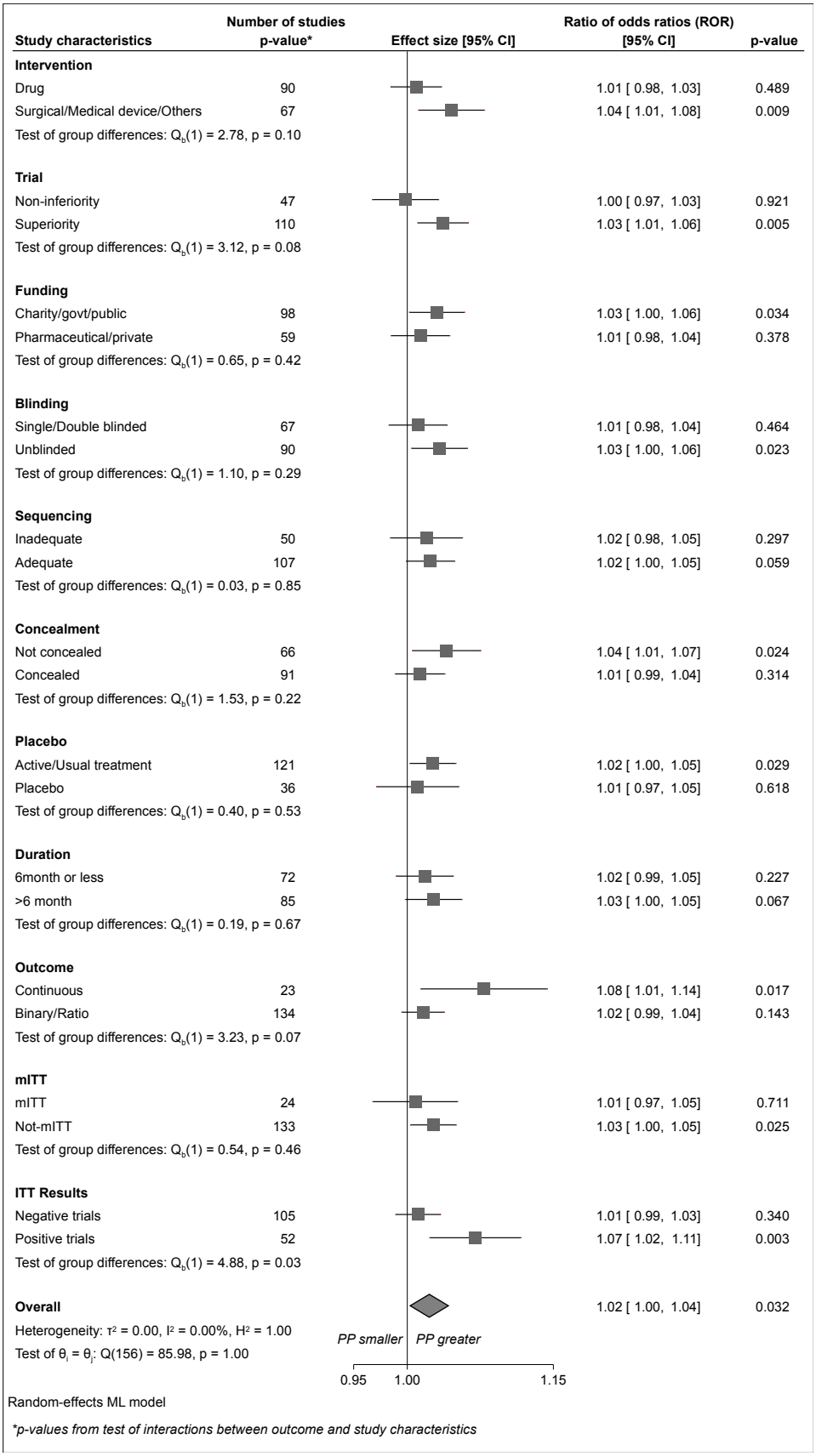


Fig 3: Ratio of odds ratios comparing PP vs ITT treatment effect in trials stratified by key trial characteristics

Publication bias

The test of small study publication bias for both binary outcome and continuous outcome suggested that there was no small study effects as both tests were statistically non-significant (Harbord test: p = 0.12, Egger test: p = 0.63).

Exploratory analyses: Comparison of ITT and PP with CACE

When comparison was made between the ITT and CACE (Supplementary Table 1, Supplementary Figure 2), pooled CACE estimate was 3% larger compared to the ITT estimates (ROR: 1.03, 95% CI: 1.00 to 1.05, p = 0.02). In meta-regression (Supplementary Table 1), higher non-adherence rate showed association with greater CACE estimates (ROR: 1.03, 95% CI: 1.00 to 1.05, p = 0.03). However, association of CACE estimates with non-adherence was not linear. We found CACE estimates were not significantly different to ITT estimates (p>0.05) when non-adherence rate was below <25% (Figure 2). Sensitivity analyses showed that ROR remains unchanged at wide range of within study correlations (Supplementary Figure 1b). We found no difference when comparison was made between CACE and PP estimates (ROR: 1.00, 95% CI: 0.98 to 1.03, p = 0.72) (Supplementary Table 1).

Discussion

Our meta-epidemiology study primarily sought to test the null hypothesis of no difference in RCT treatment effects when estimated by the PP analysis compared to the ITT analysis. We found evidence that PP analysis results in larger treatment effects than ITT analysis of RCTs thus rejecting the null hypothesis. Although the average increase in treatment effect size was relatively small (2%), the magnitude of this difference increased with higher levels of non-adherence to the trial protocol i.e., eligibility, interventions, and outcome assessment. For example, we estimate that a trial with 20% trial protocol non-adherence will result in a PP estimate 5% larger than ITT (ROR: 1.05, 95% CI: 1.02 to 1.07, p = 0.01).

The PP analysis compares outcomes of the intervention according to initial random allocation but excluding participants who do not adhere to the trial protocol in terms of eligibility, interventions, or outcome assessment.[14] The method therefore not only is exposed to selection bias but also relies on implausible assumptions that known/unknown confounders for prognosis factors are same between those excluded versus those retained. While majority of these studies reported PP analysis as a form of sensitivity analysis alongside ITT, 37% of the studies (N = 57) discussed PP results in the discussion section and emphasised their findings. Almost all studies, used PP estimates to confirm the robustness of the trial findings by statements such as "*Analysis of the primary outcome in the per-protocol population confirmed this result*" or "*We undertook a per protocol analysis for the primary outcome to check the robustness of conclusions*", which in our view may critically cloud the clinical judgement. The possibility of selection bias and our evidence of inflated PP estimate thus suggest that PP method in RCTs should be avoided.

Several statistical methods[11] have been developed for estimating the causal treatment effects that take account of the intervention nonadherence without introducing the biases inherent to PP analyses. CACE and Instrumental Variable (IV)[3] methods are two such unbiased alternatives to the ITT when adherence to the treatment is suboptimal and the estimated treatment effect has the ability to infer 'causal' effect. In this review, our exploratory analysis showed that the CACE method is more likely to provide estimates closer to ITT when non-adherence rate is below 25% i.e., lower bound of 95% confidence interval overlaps with ROR = 1.00 with statistical significance $p>0.05$ (Figure 2). This probably suggests that the CACE is as good an alternative method as the ITT under low non-adherence. Our finding conforms with past simulation study that showed ITT and CACE to have equal power when number of patients in 'never-takers' stratum is low.[29] Figure 2 also shows the exponential departure of CACE estimates from the ITT as non-adherence rate further increases. At this level of non-adherence, perhaps none of the estimates are

reliable and our finding conforms with previous study finding where unreliably large CACE estimates have been reported in the presence of high non-adherence.[30] CACE method is a randomisation-based efficacy estimator and in our evidence, its conformity with the ITT under certain levels of non-adherence (<25%) makes it a better supplement to ITT. However, it is also important to note that the estimation procedure of CACE may be complex when multiple arms are involved or switching between arms is not restricted.[29, 31-33]

Limitations and strengths

In a meta-epidemiological setting, we believe this to be the first study to empirically quantify the potential divergence associated with the PP analysis relative to the ITT analysis in estimating treatment effect under non-adherence in RCTs. The majority of our selected articles (79%) provided necessary information required therefore indirect variance estimation was minimal, which provided strong reliability of our estimates. However, we recognise our study had limitations. First, we limited our study to the RCTs published in selected high impact general medical journals and our sample may not be representative of all published RCTs. Second, as we selected only those RCTs that reported both ITT and PP, it is possible that we retrieved studies that only had insignificant treatment effects from ITT and investigators therefore additionally reported PP estimates with the anticipation of a greater treatment effect from the PP estimates. However, if this was the case, we would expect studies with insignificant ITT results to be positively associated with studies with larger PP estimates. We did not find any such association: 31% (49/157) of our studies reported treatment effects using both ITT and PP methods when both were statistically significant at least at $p<0.05$ and 64% (101/157) reported results using both methods when both were statistically non-significant at $p>0.05$. The χ^2 test statistic of association whether studies with significant/insignificant ITT estimates reported smaller/greater PP estimates was not statistically significant ($\chi^2 = 0.81$, $p=0.37$). Third, to undertake this study we had to assume a within-study bivariate correlation between ITT and PP. Reassuringly, our findings remained robust to the sensitivity analyses we carried out using a wide of range correlations. Fourth,

our derived CACE estimates are moment based rather model based and the assumption of 'compliers' stratum may not be as precise as it would have been if they were reported by the study investigators. Fifth, since the AT method destroys randomisation, we have not focused on this method in this study. Finally, because of the large number of studies, we could not provide a legible forest plot, but we provided a bubble plot with study specific weights that contributed towards the estimation of effect sizes (Supplementary Figure 3).

Policy implications

The CONSORT guideline requires RCTs to present results using the ITT method and use of PP is discouraged.[1] High rates of protocol non-adherence make all forms of analyses including the ITT unreliable. However, both the ITT and CACE involve the maintenance of randomisation, enabling an efficacy estimation that is shielded against selection bias, whereas PP is not shielded against selection bias. In addition, CACE estimates were similar to ITT estimates at higher levels of treatment non-adherence (compared to PP) making it an attractive supplement to ITT.

Conclusions

This study confirmed that there is a significant level of treatment effect overestimation with PP analysis relative to ITT analysis. Whilst at an absolute level this overall overestimation was relatively small, we also found an increasing overestimation of treatment effect at higher rates of protocol non-adherence in terms of eligibility, interventions, and outcome assessment. Given the potential for selection bias, this argues that PP estimates should be avoided in the context of RCTs testing superiority. Further research including meta-epidemiological studies using individual patient data and simulation methods are required to confirm these findings.

Contributors: MM and RT conceived and designed the study. RT also provided overall supervision for the study. MM has undertaken literature search, carried out data extraction,

statistical analyses, and drafted the manuscript. GT carried out data accuracy check. WH advised on methodological and statistical aspects of the study. EW contributed towards constructing arguments, article structure, formats, and presentation of results. All authors had their final approval of the submitted version.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare no financial relationships with any organisations that might have an interest in the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

References

1. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010-03-24 00:06:15 2010;340

2. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986;81(396):945-960. doi:10.2307/2289064

3. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*. 1996;91(434):444-455. doi:10.2307/2291629

4. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974;66(5):688-701. doi:10.1037/h0037350

5. Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*. 1978;6(1):34-58.

6. Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. *Clinical trials (London, England)*. 09/23 2012;9(1):48-55. doi:10.1177/1740774511420743

7. White IR. Uses and limitations of randomization-based efficacy estimators. Review. *Statistical Methods in Medical Research*. August 2005;14(4):327-347. doi:http://dx.doi.org/10.1191/0962280205sm406oa

8. Ten Have TR, Normand S-LT, Marcus SM, Brown C, Lavori P, Duan N. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatric Annals*. Dec 2008;38(12):772-783. doi:http://dx.doi.org/10.3928/00485713-20081201-10

9. Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health*. 2000/05/01 2000;21(1):121-145. doi:10.1146/annurev.publhealth.21.1.121

10. Shrier I, Steele RJ, Verhagen E, Herbert R, Riddell CA, Kaufman JS. Beyond intention to treat: What is the right question? *Clinical Trials*. February 2014;11(1):28-37. doi:http://dx.doi.org/10.1177/1740774513504151

11. Mostazir M, Taylor RS, Henley W, Watkins E. An overview of statistical methods for handling nonadherence to intervention protocol in randomized control trials: a methodological review. (1878-5921 (Electronic))

12. Dodd S, White I, Williamson P. Departure from treatment protocol in published randomised controlled trials: A review. Conference Abstract. *Trials*. 13 Dec 2011;12doi:http://dx.doi.org/10.1186/1745-6215-12-S1-A129

13. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. Jul 21 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097

14. Moher D, Hopewell S Fau - Schulz KF, Schulz Kf Fau - Montori V, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. (1743-9159 (Electronic))

15. Gupta SK. Intention-to-treat concept: A review. *Perspectives in Clinical Research*. Jul-Sep 2011;2(3):109-112. doi:10.4103/2229-3485.83221

16. Raskob GE, van Es N, Verhamme P, et al. Edoxaban for the Treatment of Cancer-Associated Venous Thromboembolism. *N Engl J Med*. Feb 15 2018;378(7):615-624. doi:10.1056/NEJMoa1711948

17. Gallant J, Lazzarin A, Mills A, et al. Bictegravir, emtricitabine, and tenofovir alafenamide versus dolutegravir, abacavir, and lamivudine for initial treatment of HIV-1 infection (GS-US-380-1489): a double-blind, multicentre, phase 3, randomised controlled non-inferiority trial. *The Lancet*. 2017;390(10107):2063-2072. doi:10.1016/S0140-6736(17)32299-7

18. Bornstein NM, Saver JL, Diener HC, et al. An injectable implant to stimulate the sphenopalatine ganglion for treatment of acute ischaemic stroke up to 24 h from onset (ImpACT-24B): an international, randomised, double-blind, sham-controlled, pivotal trial. *The Lancet*. 2019;394(10194):219-229. doi:10.1016/S0140-6736(19)31192-4

19. Caraceni P, Riggio O, Angeli P, et al. Long-term albumin administration in decompensated cirrhosis (ANSWER): an open-label randomised trial. *Lancet*. May 31 2018;doi:10.1016/s0140-6736(18)30840-7

20. Kaufmann R, Halm JA, Eker HH, et al. Mesh versus suture repair of umbilical hernia in adults: a randomised, double-blind, controlled, multicentre trial. *Lancet*. Mar 3 2018;391(10123):860-869. doi:10.1016/s0140-6736(18)30298-8

21. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. Oct 2006;25(20):3443-57. doi:10.1002/sim.2380

22. Egger M, Smith Gd Fau - Phillips AN, Phillips AN. Meta-analysis: principles and procedures. (0959-8138 (Print))

23. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials*. 2007/06/07 2007;8(1):16. doi:10.1186/1745-6215-8-16

24. Higgins JPT. Green S, ed. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, 2011; 2011. http://handbook.cochrane.org

25. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd; 2009.

26. Bucher HC, Guyatt Gh Fau - Griffith LE, Griffith Le Fau - Walter SD, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. (0895-4356 (Print))

27. Frangakis CE, Rubin DB. Principal Stratification in Causal Inference. *Biometrics*. 2002;58(1):21-29.

28. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. Review. *Statistical Methods in Medical Research*. August 2005;14(4):369-395. doi:http://dx.doi.org/10.1191/0962280205sm403oa

29. Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine*. 15 Nov 2002;21(21):3161-3181. doi:http://dx.doi.org/10.1002/sim.1267

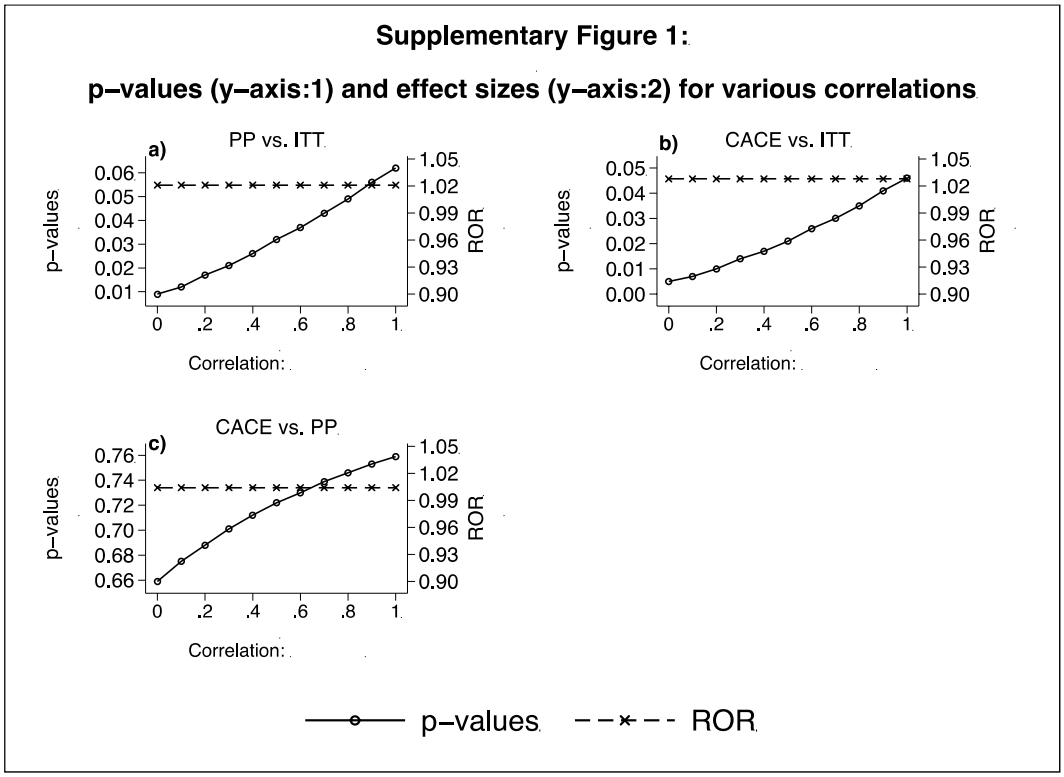
30. Matsui S. Stratified analysis in randomized trials with noncompliance. *Biometrics*. September 2005;61(3):816-823+895. doi:http://dx.doi.org/10.1111/j.1541-0420.2005.00339.x

31. Jo B. Estimation of intervention effects with noncompliance: Alternative model specifications. Empirical Study. *Journal of Educational and Behavioral Statistics*. Win 2002;27(4):385-409. doi:http://dx.doi.org/10.3102/10769986027004385

32. Yau L, Little RJ. Inference for the complier-average causal effect for longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*. 2001;96(456):1232-1244.

33. Mealli F, Imbens GW, Ferro S, Biggeri A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics (Oxford, England)*. Apr 2004;5(2):207-222.

Supplementary Figure 1:



Supplementary Table 1: Comparison of CACE estimates with ITT and PP

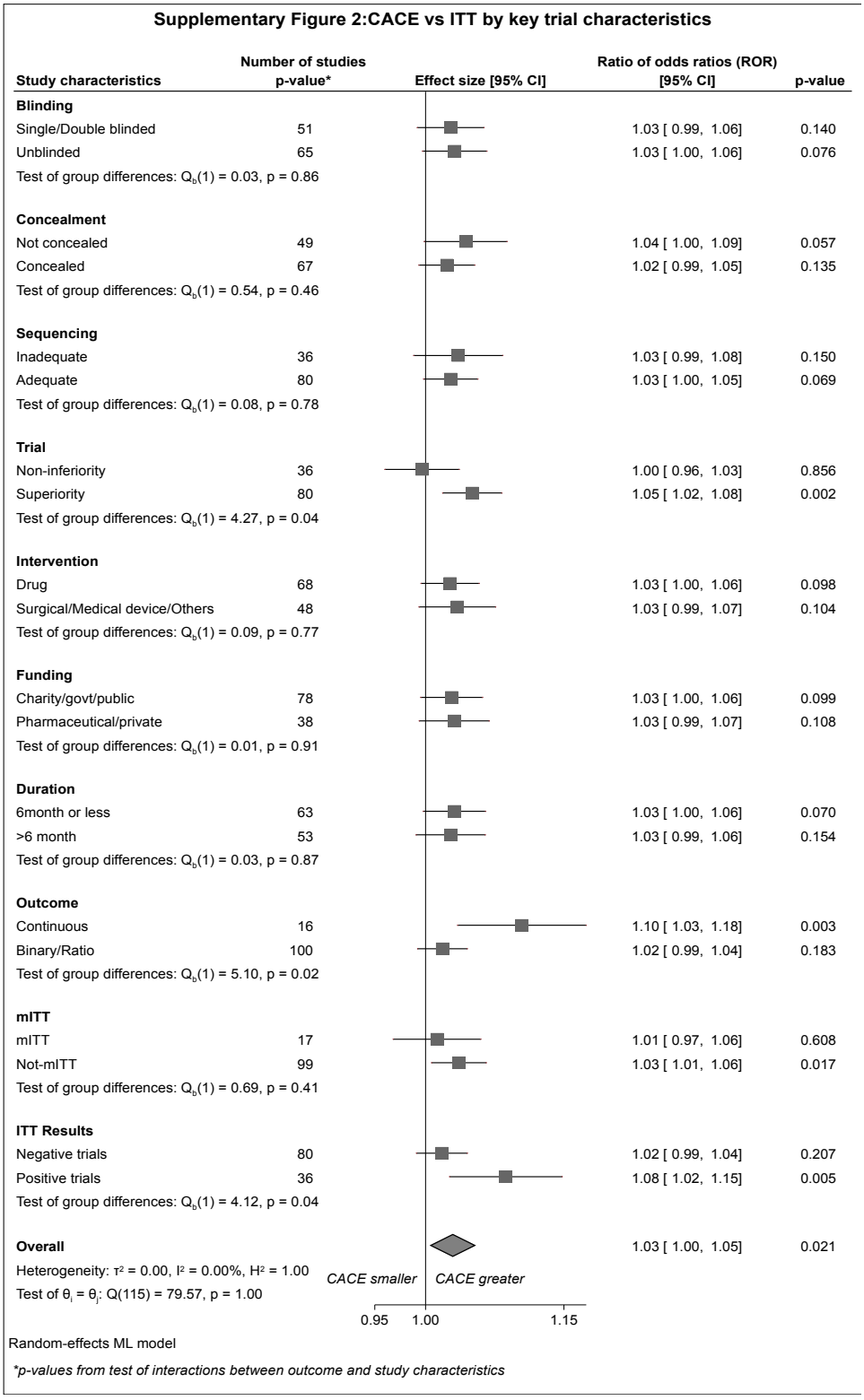
Analyses type	N	Ratio of odds Ratio (ROR) (95% CI)	Study heterogeneity τ^2 (I^2%)
Exploratory analysis: Comparison with CACE	**–**	**–**	–
CACE vs. ITT (All studies)	116	1.03 (1.00 to 1.05)*	0.00 (0.00)
CACE vs. PP (All studies)	116	1.00 (0.98 to 1.03)	0.00 (0.00)
Meta regression[a]	–	–	–
CACE vs. ITT	116	–	–
%Nonadherence	–	1.03 (1.00 to 1.05)*	–
ITT results (Positive vs. Negative trials)[b]	–	1.07 (1.00 to 1.13)*	–
CACE vs. PP	116	–	–
%Nonadherence	–	0.99 (0.97 to 1.02)	–

[a]Meta regression: Effects of other study characteristics were not significant. Regression coefficients correspond to 10% increase in non-adherence

[b]'Positive trials': statistical superiority of treatment versus control group. 'Negative trials': no statistical superiority of treatment versus control group or statistical superiority of control versus treatment group

*denotes $p<0.05$; τ^2: Between study variance; I^2: Heterogeneity statistic

Supplementary Figure 2:



Supplementary Figure 3: