RESEARCH

Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study

OPEN ACCESS

Nazmus Saquib *postdoctoral research fellow*, Juliann Saquib *postdoctoral research fellow*, John P A Ioannidis *professor*

Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

Abstract

Objective To assess adjustment practices for primary outcomes of randomized controlled trials and their impact on the results.

Design Meta-epidemiologic study.

Data sources 25 biomedical journals with the highest impact factor according to *Journal Citation Reports* 2009.

Study selection Randomized controlled trials published in print in 2009 that reported primary outcomes. The search yielded 684 eligible papers of randomized controlled trials, of which 200 were randomly selected.

Data extraction Two researchers independently extracted data on study population, intervention, primary outcome, and the adjustment plan for primary outcomes. They also recorded the magnitude and statistical significance of the intervention effect with and without adjustments, and estimated whether adjustment made a difference in the level of nominal significance. They also compared the analysis plan for model adjustment in the published trial versus the trial protocol with information on the protocol collected from registries, design papers, and communication with all corresponding authors.

Results 54% of the trials used stratified randomization, 96% presented baseline characteristics in the compared arms, and 46% also evaluated differences in baseline factors with statistical testing. Half of the trials performed adjusted analyses for the main outcome, as the sole analysis (29%) or along with unadjusted analyses (21%). Adjustment for stratification variables and for baseline variables was performed in 39% (42/108) and 42% (84/199) of the trials, respectively. Among 40 comparisons with both adjusted and unadjusted analyses, 43% had statistically significant effects, 40% had non-significant effects, and 18% had significant effects with only one of the two analyses, but not with the other. Information on analysis plan regarding model adjustment was available in 6% (9/162) of trial registry entries, 78% (21/27) of design papers, and 74% (40/54) of protocols obtained from authors. The analysis plan disagreed between the published trial and the registry, protocol, or design paper in 47% (28/60) of the studies.

Conclusions There is large diversity on whether and how analyses of primary outcomes are adjusted in randomized controlled trials and these choices can sometimes change the nominal significance of the results. Registered protocols should explicitly specify adjustments plans for main outcomes and analysis should follow these plans.

Introduction

The results of primary outcomes in randomized controlled trials are often influenced by factors other than the treatment. Examples include recruiting site characteristics in multicenter trials such as the type and number of sites¹ as well as various participants' characteristics such as age, sex, and body weight.² These variables are often used as stratification factors during randomization.³ Some participants' characteristics are asymmetrically distributed between the study arms despite randomization; and the likelihood of imbalance is higher in trials with a small sample and when randomization procedures are not followed properly. At present there is no consistent practice as to how to handle the imbalance between the study groups in analysis.^{2 4 5} Some studies adjust the outcome model for baseline differences among study groups. Other studies consider baseline differences to be chance findings and therefore not to be adjusted for, and many methodologists argue against even checking for them.⁶

Different choices for adjustment may lead to different estimates of the treatment effect and levels of statistical significance. The variability in treatment effects estimates due to multiple analytical choices has been described as "vibration of effects": each analysis with or without various adjustments may give somewhat different results.⁷ Making adjustments should have little impact on the conclusions, when the effect size is large and its clinical importance is not contestable or even when the effect size is modest but the amount of evidence is large. Most randomized controlled trials are, however, not large⁸ and when

Correspondence to: J P A Ioannidis jioannid@stanford.edu

Extra material supplied by the author (see <http://www.bmj.com/content/347/bmj.f4313?tab=related#webextra>)

Randomized controlled trials selected for study

the presence of an effect is tenuous (for example, when the results hover around the “attractive” P value of 0.05),⁹ decisions regarding adjustment may influence the interpretation of the study outcome. Vibration of effects may lead to biased results, if multiple possible adjustment schemes are performed retrospectively and the most favorable result is then selectively reported or highlighted.

Several measures can be taken to reduce the disparate practices in the analysis of data from randomized controlled trials to increase transparency and overall reproducibility of clinical research. The ideal practice would be that the trial investigators make the protocol, dataset, and analytical code publicly available. Alternately, investigators could provide as much raw data as possible using detailed tables and figures either in the trial publications or in the registry. Minimally, investigators should follow the standardized reporting guidelines that are already in place for randomized controlled trials. For example, the CONSORT statement and ICH E9 (statistical principles for clinical trials) provide explicit instructions on statistical analyses in randomized controlled trials, including model adjustment for baseline differences and stratified randomization factors.^{10 11}

In this analysis we evaluated a sample (n=200) of randomized controlled trials published in high impact journals in 2009. We gathered all relevant information about planned statistical analysis of the primary outcome from trial registries, design papers, and protocols provided by the authors. We assessed differences in the statistical plan for model adjustment between the trial protocol and the trial publication for those trials with available protocols. For the sample of all 200 trials we also assessed the congruency of model adjustment for primary outcomes between the trial methods and results section. We examined which factors were used for adjustment and evaluated the extent to which adjusted and unadjusted treatment effects differ in level of nominal statistical significance.

Methods

We searched PubMed to locate the relevant trials. We categorized the search term according to study type (randomized controlled trial) and journal (*BMJ*, *American Journal of Psychiatry*, *American Journal of Respiratory Critical Care Medicine*, *Annals of Internal Medicine*, *Annals of Neurology*, *Archives of General Psychiatry*, *Archives of Internal Medicine*, *Blood*, *Brain*, *Circulation*, *European Heart Journal*, *Gastroenterology*, *Gut*, *Hepatology*, *Journal of Allergy and Clinical Immunology*, *Journal of the American College of Cardiology*, *Journal of Clinical Oncology*, *Journal of the National Cancer Institute*, *JAMA*, *Lancet*, *Lancet Infectious Diseases*, *Lancet Neurology*, *Lancet Oncology*, *New England Journal of Medicine*, and *PLoS Medicine*). The 25 biomedical journals were selected as having the highest impact factor (per *Journal Citation Reports* 2009 edition) among journals that may publish clinical trials. The “AND” Boolean operator was used to combine search terms between the categories and the “OR” was used within the category for journals. We limited the search to studies that involved human participants and were published in 2009.

We screened the abstract using the inclusion criteria of randomized controlled trial, published in print in 2009, and main trial paper with primary outcome. We randomly selected 200 from the eligible articles.

Protocol assessment of included randomized controlled trials

Registration entries—We searched the trial papers to determine whether the study was registered. If the registry number was available, we checked the appropriate trial registry online and extracted information on the use of model adjustment for the primary outcome from general registration information, study results, or full study protocols that might be available in the registry for each trial.

Original protocol through personal communication—We identified the corresponding author and his or her contact information for each trial. We emailed each author with a brief description of our study and requested the full original study protocol. We waited two weeks for a response and sent a reminder email to those who had not responded.

Design paper search—We searched the registry as well as the reference list of the trial publication to see whether a design paper had been published. We obtained full text of the design papers through PubMed and extracted information on the use of model adjustment for the primary outcome. Additionally, some authors sent us a design paper during our email communication with them.

Data extraction from publications of randomized controlled trials

From each trial we extracted the following information: study arms, primary outcomes, sample size, number of sites, whether randomization was stratified, and whether arms were compared for baseline characteristics. We considered baseline differences to be statistically significant when P values were given or the text included the word “significant” for the comparison.

We recorded whether any adjustments were planned for the primary outcome in the Methods section and whether these analyses were respectively reported in the Results section. We defined a model as adjusted if explicit statements regarding adjustment were made, covariates were listed, or a statistical test that by definition is multivariate (for example, multiple logistic regression, multivariate analysis of variance) was reported. Whenever adjustments were made, we recorded whether they used variables that had been used in the stratification of the randomization or the baseline characteristics, and which variables were involved. To avoid double counting, if a variable had been used for the stratified randomization then it was not counted as a baseline characteristic; otherwise, baseline variables include all those mentioned in the text or table as measured at baseline. Finally, we recorded the unadjusted estimate and 95% confidence interval of the treatment effect of the primary outcome and any adjusted estimates for the same treatment comparison. For trials with multiple primary outcomes, we selected the outcome on which the sample size calculation was based.

Two researchers (NS and JS) independently extracted data. The pilot phase involved data extraction from 30 articles; the model adjustment data were in full agreement in two thirds of the cases. At the end of the pilot phase, the two researchers met with the senior investigator (JPAI) to resolve the discrepancies, to finalize the definition, and to streamline the protocol. At the end of the full data extraction phase, the two researchers discussed discrepancies in model adjustment data and were able to reach a consensus in all but four cases, which were arbitrated by the senior investigator.

Statistical analysis

We compared the model adjustment plan provided in the main trial publication with the corresponding information extracted from the registry, protocol, or design paper, to see if these did or did not agree. We also calculated the proportion of trials that reported statistically significant effects in each of the categories of agreement. For trials where the protocol and the publication had different analysis plans, we also recorded what the results were for analyses that had not been specified in the protocol but had been added in the publication.

From the sample of all 200 trials we generated frequency tables for categorical variables and measures of central tendency (that is, mean, median) and range for continuous variables pertaining to general trial characteristics and adjustment procedures. We identified the most common variables adjusted for and calculated the percentage of trials that reported them.

To achieve consistency across trials, we presented effect sizes in such a manner that relative risk metrics <1.00 and risk difference and mean difference metrics <0 indicate that the experimental intervention is better than the control intervention. Whenever both adjusted and unadjusted treatment effects were reported, we examined the concordance of the level of nominal statistical significance, based on 95% confidence intervals being entirely on one side of the null, P values <0.05 , or a statement in the text. Whenever nominal significance was reached with only one analysis but not the other, we also noted whether the authors of the trial focused on the significant or non-significant result primarily in interpreting their findings.

Statistical analyses were conducted in SAS version 9.2 (Cary, NC). All P values are two tailed.

Results

The search resulted in 1123 articles. We excluded a total of 439 articles; 163 papers were not randomized controlled trials, 94 papers were published in print in 2010 (published electronically in 2009, but appeared in print in 2010), and 182 were secondary studies using the population of the original randomized controlled trial (that is, extended follow-up or subset analyses). Then, of the 200 randomly selected articles, two papers each reported the results from two separate trials and these were considered separately; conversely, we excluded three trials that had not analyzed primary outcomes between the study arms. Analysis included 199 trials (see supplementary file).

Comparison between trial protocols and main paper

Most of the 199 trials (81%, $n=162$) were registered; the most common registries included clinicaltrials.gov ($n=109$), International Standard Randomized Controlled Trial Number (ISRCTN) ($n=41$), and Australia New Zealand Clinical Trial Registry (ANZCTR) ($n=6$). General information on registration was available for the majority of the trials; study results were available for only 12% (20/162) of the trials. Statistical procedures such as model adjustment were found only in 6% (9/162) of the trials. Of the trials, 15% (30/199) had published design papers and 70% (21/30) of them provided information on model adjustment. We evaluated 54 original protocols, sent to us by the authors, and found that 74% (40/54) included an analysis section with a model adjustment plan. In total, 31% (61/199) of the trials had available information on adjustment from the registry, design paper, or protocol (figure 1).

We compared the analysis plans from the main paper with the plan reported in the registry, protocol, or design paper for 60

trials (the plan was unclear for one published trial). The adjustment plan matched in 53% (32/60) of the trials; of them, 62% (20/32) reported a significant effect. Among the trials where the adjustment plan did not match, 53% (15/28) reported a significant effect. Of the 28 identified trials where the plan in the protocol did not match what was reported eventually in the publication, 75% ($n=21$) reported analyses that had not been specified in the protocol and, of them, 62% (13/21) were statistically significant; whereas 25% ($n=7$) did not report analyses that had been specified in the protocol and 29% (2/7) of the reported analyses were statistically significant.

Characteristics of selected trials

Three quarters (77%) of the 199 reviewed trials had only two study arms, and a similar percentage (76%) of the trials assessed a single primary outcome. Most trials (69%) enrolled participants from multiple study sites (median 20, range 2-862 sites). More than half (54%) used stratified randomization. Almost all trials (96%) presented baseline characteristics per arm. About half (46%) tested statistically for differences in baseline characteristics, and 22% claimed statistically significant differences in baseline characteristics (table 1).

Adjustment practices

Among the 199 trials, the most common approach specified in the Methods section was to use an unadjusted analysis either alone (48%) or in conjunction with adjusted analysis (21%), whereas 30% of the trials planned adjusted analysis only. A similar picture was seen in the Results section. There were six trials where the plan in the Methods section was incongruent with what was shown in the Results section: three trials where adjusted analysis were promised in the Methods section but not shown in the Results section and three trials where adjusted analyses appeared in the Results section without being stated in the Methods section. The use or not of adjustment was unclear in both the Methods and the Results sections for two trials (table 2).

Adjustment types

Ninety nine trials reported adjusted analysis in the Results section (adjusted only $n=57$, in conjunction with unadjusted $n=42$). In 92% (91/99) of the trials that used adjustments, the authors presented clearly the exact adjusting variables, whereas in 8% (8/99) of the trials the list of adjusting variables remained unclear. The median number of variables used in the adjustment was 3 (range 1-13, $n=91$). Ninety two per cent (84/91) adjusted for baseline variables and 46% (42/91) adjusted for stratification variables. The most common variables used for adjustment were age (31%), study center/site (27%), sex (23%), socioeconomic status (9%), smoking (6%), race (5%), body mass index or body weight (5%), and self rated health (5%). Adjustment for baseline variables was substantially more common when significant baseline differences had been found (60% v 38%, $P=0.02$).

Comparison of nominal significance between unadjusted and adjusted analyses

Of the 42 trials that performed both unadjusted and adjusted analyses for the selected primary outcomes, the determination of nominal statistical significance was possible for 38 (range 12-49) trials. Twenty eight of them provided specific effect estimates for both types of analyses (table 3); 10 trials provided nominal significance for both types of analyses but were missing one or both effect estimates (table 4). Two trials^{19 36} had two comparisons each (two different active interventions versus

control), making a total of 40 comparisons: 43% (17/40) had a statistically significant effect with both analyses, 40% (16/40) had non-significant effects with both analyses, and 18% (7/40) had significant effect with only one of the two analyses (three unadjusted only, four adjusted only) (table 3).

Among the seven comparisons (six trials)^{15 19 20 37 42 45} with discrepant levels of nominal significance in unadjusted versus adjusted analyses, the authors interpreted the trial focusing primarily on the statistically significant result in four trials (five comparisons)^{19 37 42 45} where the experimental treatment tended to be better. Conversely, they focused primarily on the non-statistically significant result in one trial¹⁵ where the experimental treatment was worse. Finally, the authors of one trial carefully balanced between the significant and non-significant results, admitting that according to the primary analysis plan the results were non-significant.²⁰

Discussion

This empirical evaluation shows that there is wide diversity in adjustment practices for the analysis of primary outcomes in randomized controlled trials. The use or not of adjustments can make a substantial difference in the statistical inferences for several trials. In 18% of the comparisons that provided both adjusted and unadjusted analysis results, nominal statistical significance was attained with only one of the two approaches; in six of these seven cases, authors focused primarily on the result that was more favorable or less unfavorable for the experimental treatment. The lack of standardization in adjustment practices offers an opportunity for subjectively steering the results of randomized controlled trials towards specific conclusions. Registries generally did not include information on this aspect of the analysis, few trials had published design papers, and protocols could be obtained from authors for the minority of the trials. Even then, analysis plans regarding model adjustment disagreed between the available information in the protocol and the published trial in about half of the cases.

Selective reporting

In the current practice, it is not possible to evaluate whether the analysis plan changed between the original protocol and main trial paper using the trial registries. Although most trials that we evaluated were registered, the information regarding statistical analyses, specifically model adjustment, was rarely provided. Only 12% of registry records provided study results and in only half of those cases was information on adjustment given. Trial investigators should provide full protocols with detailed analysis plans to the registry prior to conducting the study. Analysis plans in the registry, protocol, or design paper disagreed with the plans in the published paper in approximately half of the studies. It is unlikely that the disagreement rate would be lower in the trials where the protocol was neither publicly available nor retrievable from the authors. Our findings are consistent with a growing literature on selective reporting of outcomes and analyses in diverse study designs, including randomized controlled trials. Other empirical studies⁵⁰⁻⁵² have shown selective reporting based on comparisons of protocols and published results; many outcomes specified in the protocols were missing in the published reports, while new outcomes were introduced.

We observed large variability in terms of the practices of adjustment for stratification or baseline variables. Less than half of the trials that used a stratified randomization technique adjusted the model for the stratification variables, even though

this is suggested by both the CONSORT and the ICH E9 guidelines. Although we do not advocate testing for differences in baseline variables, our results showed that more trials adjusted for baseline variables when significant differences had been noted between the compared arms in these variables; however, there were many exceptions to this pattern. Many studies with significant differences at baseline did not adjust for them and many studies without significant differences still adjusted for baseline factors. This inconsistency is not surprising since a consensus does not exist on testing and adjusting for multicenter trials. The studies that adjusted for center typically did not report the exact model that they used to adjust for site. Multiple approaches exist to account for site in the analysis, and results may differ among them.¹ Another empirical study found that only 25% among a sample of trials published in 2000 and 2006 reported adjusted analyses and only 5% and 10% of trials in these two years reported both adjusted and unadjusted analyses. The respective proportions were higher in our survey, but they were still low, and the difference may reflect the more select nature of our sample of trials (trials published in major journals) rather than improvement in reporting of adjustments over time.⁵³

The adjusted model occasionally lacked information regarding which covariates it included. Sometimes the text specified the list of covariates that were tested in univariate analyses against the outcome, but did not report which of them qualified for the final adjusted model⁵⁴; whereas in other trials the text reported covariates that were significantly associated with the outcome without mentioning whether the adjusted model also contained any non-significant covariates.^{55 56} Flexibility in the use of different combinations of covariates provides an additional mechanism for “vibration of effects.” Different adjustment choices may yield different results and opens a window for potential selective reporting.

Limitations of this study

Our analysis has limitations. Inaccurate reporting of the methods and results cannot be excluded. However, the Methods and Results sections were almost always congruent in specifying the types of analyses. Moreover, some data were missing in the published reports. The estimates for the intervention effect and the associated significance levels were not mentioned in several trials that performed both unadjusted and adjusted analysis. Thus one would be unable to see the magnitude of the differences, if any, between the two analytical approaches. Finally, despite our intense efforts we could not retrieve the protocols for more than two thirds of the trials. It does not necessarily follow that these trials suffer from selective reporting, but lack of transparency creates opportunities for such bias.

Conclusions

We have shown the potential of selective reporting on the primary outcomes of trials published in the most high profile clinical journals, due to the alternative choices on how to adjust or not the results of these outcomes. Adjustments may affect the interpretation of the study findings and eventually their impact on medical practice and policy. Other authors have also identified that non-significant results in randomized controlled trials are subject to “spin” in their interpretation to make them seem more favorable.⁵⁷ Overall, the plethora of analytical and interpretation options may infuse subjectivity in the evidence procured by randomized controlled trials. One possibility to help minimize these problems is to request that registered protocols should present in meticulous detail any adjustments plans for the main outcomes. Significant advances have been

made in trial registration^{58 59} and there is impetus to improve also the quality of registered protocols. For example, the SPIRIT initiative recommends that protocols explicitly specify whether an adjusted analysis will be undertaken; and if so, to list the adjustment variables as well as techniques to handle them in the model.⁶⁰ This is in line with efforts in other biomedical fields, where the movement of reproducible research has led to the request to deposit detailed protocols, analytical codes, and the raw data of published studies.⁶¹⁻⁶⁵ Increased transparency regarding the choices of adjusted and unadjusted analyses may enhance the reliability of inferences from randomized trials.

We thank the corresponding authors who responded to our email request and contributed to this work by providing us with the trial protocols or design papers.

Contributors: JPAI had the original idea and all three authors conceived and designed the study. NS and JS identified the eligible trials and extracted the relevant data. All three authors performed the statistical analyses, wrote the manuscript, and approved the final version of the manuscript. JPAI is the guarantor. All authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding: This study received no funding.

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 support from any organization for the submitted work; no financial relationships with any organization that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: Datasets are available from the corresponding author at jioannid@stanford.edu.

- Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 2001;135:112-23.
- Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917-30.
- Piantadosi S. Clinical trials: a methodological perspective. Wiley, 1997.
- Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials* 1998;19:249-56.
- Matts JP, McHugh RB. Analysis of accrual randomized clinical trials with balanced groups in strata. *J Chronic Dis* 1978;31:725-40.
- Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.
- Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008;19:640-8.
- Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;365:1159-62.
- Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
- Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Stat Med* 1999;18:1903-42.
- Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med* 2010;7:e1000251.
- Siebers AG, Klinkhamer PJ, Grefte JM, Massuger LF, Vedder JE, Beijers-Broos A, et al. Comparison of liquid-based cytology with conventional cytology for detection of cervical cancer precursors: a randomized controlled trial. *JAMA* 2009;302:1757-64.
- Lee SM, Woll PJ, Rudd R, Ferry D, O'Brien M, Middleton G, et al. Anti-angiogenic therapy using thalidomide combined with chemotherapy in small cell lung cancer: a randomized, double-blind, placebo-controlled trial. *J Natl Cancer Inst* 2009;101:1049-57.
- Piepers S, Veldink JH, de Jong SW, van der Tweel I, van der Pol WL, Uijtendaal EV, et al. Randomized sequential trial of valproic acid in amyotrophic lateral sclerosis. *Ann Neurol* 2009;66:227-34.
- Eddleston M, Eyer P, Worek F, Juszczak E, Alder N, Mohamed F, et al. Pralidoxime in acute organophosphorus insecticide poisoning—a randomised controlled trial. *PLoS Med* 2009;6:e1000104.
- Achan J, Tibenderana JK, Kyabayinze D, Wabwire-Mangen F, Kamya MR, Dorsey G, et al. Effectiveness of quinine versus artemether-lumefantrine for treating uncomplicated falciparum malaria in Ugandan children: randomised trial. *BMJ* 2009;339:b2763.
- ASTEC Study Group, Kitchener H, Swart AM, Qian Q, Parmar MK. Efficacy of systematic pelvic lymphadenectomy in endometrial cancer (MRC ASTEC trial): a randomised study. *Lancet* 2009;373:125-36.
- Bensimon G, Ludolph A, Agid Y, Vidaliht M, Payan C, Leigh PN, et al. Riluzole treatment, survival and diagnostic criteria in Parkinson plus disorders: the NNIPPS study. *Brain* 2009;132(Pt 1):156-71.

- De Smet AM, Kluytmans JA, Cooper BS, Mascini EM, Benus RF, van der Werf TS, et al. Decontamination of the digestive tract and oropharynx in ICU patients. *New Engl J Med* 2009;360:20-31.
- O'Connor CM, Whellan DJ, Lee KL, Keteyian SJ, Cooper LS, Ellis SJ, et al. Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA* 2009;301:1439-50.
- GISSI AF Investigators, Disertori M, Latini R, Barlera S, Franzosi MG, Staszewsky L, et al. Valsartan for prevention of recurrent atrial fibrillation. *New Engl J Med* 2009;360:1606-17.
- Sax PE, Tierney C, Collier AC, Fischl MA, Mollan K, Peoples L, et al. Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. *New Engl J Med* 2009;361:2230-40.
- Colon Cancer Laparoscopic or Open Resection Study Group, Buunen M, Veldkamp R, Hop WC, Kuhry E, Jeekel J, et al. Survival after laparoscopic surgery versus open surgery for colon cancer: long-term outcome of a randomised clinical trial. *Lancet Oncol* 2009;10:44-52.
- Von Minckwitz G, du Bois A, Schmidt M, Maass N, Cufer T, de Jongh FE, et al. Trastuzumab beyond progression in human epidermal growth factor receptor 2-positive advanced breast cancer: a german breast group 26/breast international group 03-05 study. *J Clin Oncol* 2009;27:1999-2006.
- Wiegel T, Bottke D, Steiner U, Siegmann A, Golz R, Storkel S, et al. Phase III postoperative adjuvant radiotherapy after radical prostatectomy compared with radical prostatectomy alone in pT3 prostate cancer with postoperative undetectable prostate-specific antigen: ARO 96-02/AUO AP 09/95. *J Clin Oncol* 2009;27:2924-30.
- Staedke SG, Mwebaza N, Kamya MR, Clark TD, Dorsey G, Rosenthal PJ, et al. Home management of malaria with artemether-lumefantrine compared with standard care in urban Ugandan children: a randomised controlled trial. *Lancet* 2009;373:1623-31.
- Bhardwaj P, Garg PK, Maulik SK, Saraya A, Tandon RK, Acharya SK. A randomized controlled trial of antioxidant supplementation for pain relief in patients with chronic pancreatitis. *Gastroenterology* 2009;136:149-59 e2.
- Ekeberg OM, Bautz-Holter E, Tveita EK, Juel NG, Kvalheim S, Brox JI. Subacromial ultrasound guided or systemic steroid injection for rotator cuff disease: randomised double blind study. *BMJ* 2009;338:a3112.
- Hassiotis A, Robotham D, Canagasabay A, Romeo R, Langridge D, Blizard R, et al. Randomized, single-blind, controlled trial of a specialist behavior therapy team for challenging behavior in adults with intellectual disabilities. *Am J Psychiatry* 2009;166:1278-85.
- Sezer M, Cimen A, Aslanger E, Elitok A, Ummam B, Bugra Z, et al. Effect of intracoronary streptokinase administered immediately after primary percutaneous coronary intervention on long-term left ventricular infarct size, volumes, and function. *J Am Coll Cardiol* 2009;54:1065-71.
- Wake M, Baur LA, Gerner B, Gibbons K, Gold L, Gunn J, et al. Outcomes and costs of primary care surveillance and intervention for overweight or obese children: the LEAP 2 randomised controlled trial. *BMJ* 2009;339:b3308.
- Lo GH, Chen WC, Wang HM, Lin CK, Chan HH, Tsai WL, et al. Low-dose terlipressin plus banding ligation versus low-dose terlipressin alone in the prevention of very early rebleeding of oesophageal varices. *Gut* 2009;58:1275-80.
- CLOTS Trials Collaboration, Dennis M, Sandercock PA, Reid J, Graham C, Murray G, et al. Effectiveness of thigh-length graduated compression stockings to reduce the risk of deep vein thrombosis after stroke (CLOTS trial 1): a multicentre, randomised controlled trial. *Lancet* 2009;373:1958-65.
- Dennis CL, Hodnett E, Kenton L, Weston J, Zupancic J, Stewart DE, et al. Effect of peer support on prevention of postnatal depression among high risk women: multisite randomised controlled trial. *BMJ* 2009;338:a3064.
- Steinberg MB, Greenhaus S, Schmelzer AC, Bover MT, Foulds J, Hoover DR, et al. Triple-combination pharmacotherapy for medically ill smokers: a randomized trial. *Ann Intern Med* 2009;150:447-54.
- Cowling BJ, Chan KH, Fang VJ, Cheng CK, Fung RO, Wai W, et al. Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial. *Ann Intern Med* 2009;151:437-46.
- Ahmad F, Hogg-Johnson S, Stewart DE, Skinner HA, Glazier RH, Levinson W. Computer-assisted screening for intimate partner violence and control: a randomized trial. *Ann Intern Med* 2009;151:93-102.
- Schnipper JL, Hamann C, Ndumele CD, Liang CL, Carty MG, Karson AS, et al. Effect of an electronic medication reconciliation application and process redesign on potential adverse drug events: a cluster-randomized trial. *Arch Intern Med* 2009;169:771-80.
- Sorensen PS, Mellgren SI, Svenningsson A, Elovaara I, Frederiksen JL, Beiske AG, et al. NORdic trial of oral Methylprednisolone as add-on therapy to Interferon beta-1a for treatment of relapsing-remitting Multiple Sclerosis (NORMIMS study): a randomised, placebo-controlled trial. *Lancet Neurol* 2009;8:519-29.
- Adler A, Aschenbeck J, Yenerim T, Mayr M, Aminalai A, Drossel R, et al. Narrow-band versus white-light high definition television endoscopic imaging for screening colonoscopy: a prospective randomized trial. *Gastroenterology* 2009;136:410-6 e1; quiz 715.
- Dumville JC, Worthy G, Bland JM, Cullum N, Dowson C, Iglesias C, et al. Larval therapy for leg ulcers (VenUS II): randomised controlled trial. *BMJ* 2009;338:b773.
- Lainchbury JG, Troughton RW, Strangman KM, Frampton CM, Pilbrow A, Yandle TG, et al. N-terminal pro-B-type natriuretic peptide-guided treatment for chronic heart failure: results from the BATTLESCARRED (NT-proBNP-Assisted Treatment To Lessen Serial Cardiac Readmissions and Death) trial. *J Am Coll Cardiol* 2009;55:53-60.
- Muss HB, Berry DA, Cirincione CT, Theodoulou M, Mauer AM, Kornblith AB, et al. Adjuvant chemotherapy in older women with early-stage breast cancer. *New Engl J Med* 2009;360:2055-65.
- Piper ME, Smith SS, Schlam TR, Fiore MC, Jorenby DE, Fraser D, et al. A randomized placebo-controlled clinical trial of 5 smoking cessation pharmacotherapies. *Arch Gen Psychiatry* 2009;66:1253-62.
- Plint AC, Johnson DW, Patel H, Wiebe N, Correll R, Brant R, et al. Epinephrine and dexamethasone in children with bronchiolitis. *New Engl J Med* 2009;360:2079-89.
- Hofmeijer J, Kappelle LJ, Algra A, Amelink GJ, van Gijn J, van der Worp HB, et al. Surgical decompression for space-occupying cerebral infarction (the Hemispherectomy After Middle Cerebral Artery infarction with Life-threatening Edema Trial [HAMLET]): a multicentre, open, randomised trial. *Lancet Neurol* 2009;8:326-33.
- Spencer A, Prince HM, Roberts AW, Prosser JW, Bradstock KF, Coyle L, et al. Consolidation therapy with low-dose thalidomide and prednisolone prolongs the survival of multiple myeloma patients undergoing a single autologous stem-cell transplantation procedure. *J Clin Oncol* 2009;27:1788-93.

What is already known on this topic

At present the imbalance between study groups in randomized controlled trials is inconsistently handled

Some studies adjust the outcome model for baseline differences among study groups, whereas others consider them chance findings or not even worth consideration

Different choices for adjustment may lead to different estimates of the treatment effect and levels of statistical significance

What this study adds

Covariates are handled in diverse ways in the analysis of primary outcomes of randomized controlled trials

It is common for analysis plans on adjustments to differ between protocols and published papers, but this can be discerned only when protocols are obtained from authors, since detailed information on the analysis plan is rarely available

Moreover, unadjusted versus adjusted results sometimes differ in the level of nominal significance, and investigators usually select to report the more favorable results

- 48 Kesecioglu J, Beale R, Stewart TE, Findlay GP, Roubey JJ, Holzapfel L, et al. Exogenous natural surfactant for treatment of acute lung injury and the acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2009;180:989-94.
- 49 Vandemheen KL, O'Connor A, Bell SC, Freitag A, Bye P, Jeanneret A, et al. Randomized trial of a decision aid for patients with cystic fibrosis considering lung transplantation. *Am J Respir Crit Care Med* 2009;180:761-8.
- 50 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
- 51 Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
- 52 Chan AW, Kroleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171:735-40.
- 53 Yu LM, Chan AW, Hopewell S, Deeks JJ, Altman DG. Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. *Trials* 2010;11:59.
- 54 Adamsen L, Quist M, Andersen C, Moller T, Herrstedt J, Kronborg D, et al. Effect of a multimodal high intensity exercise intervention in cancer patients undergoing chemotherapy: randomised controlled trial. *BMJ* 2009;339:b3410.
- 55 Ades PA, Savage PD, Toth MJ, Harvey-Berino J, Schneider DJ, Bunn JY, et al. High-calorie-expenditure exercise: a new approach to cardiac rehabilitation for overweight coronary patients. *Circulation* 2009;119:2671-8.
- 56 Kim GP, Sargent DJ, Mahoney MR, Rowland KM, Jr, Philip PA, Mitchell E, et al. Phase III noninferiority trial comparing irinotecan with oxaliplatin, fluorouracil, and leucovorin in patients with advanced colorectal carcinoma previously treated with fluorouracil: N9841. *J Clin Oncol* 2009;27:2848-54.
- 57 Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058-64.
- 58 Zarin DA, Ide NC, Tse T, Harlan WR, West JC, Lindberg DA. Issues in the registration of clinical trials. *JAMA* 2007;297:2112-20.
- 59 Zarin DA, Tse T, Ide NC. Trial registration at ClinicalTrials.gov between May and October 2005. *New Engl J Med* 2005;353:2779-87.
- 60 Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013;346:e7586.
- 61 Peng RD. Reproducible research in computational science. *Science* 2011;334:1226-7.
- 62 Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. *Science* 2011;334:1230-2.
- 63 Diggle PJ, Zeger SL. Embracing the concept of reproducible research. *Biostatistics* 2010;11:375.
- 64 Keiding N. Reproducible research and the substantive context. *Biostatistics* 2010;11:376-8.
- 65 Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. *Ann Intern Med* 2007;146:450-3.

Accepted: 27 June 2013

Cite this as: *BMJ* 2013;347:f4313

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>.

Tables

Table 1 | Description of trials in analysis. Values are numbers (percentages) unless stated otherwise.

Variables	Trials (n=199)
No of study arms:	
2	154 (77)
3	28 (14)
≥4	17 (9)
No of primary outcomes:	
1	152 (76)
2	27 (14)
≥3	20 (10)
Setting:	
Single site	58 (29)
Multisite	137 (69)
Missing information	4 (2)
Randomization:	
Not stratified	91 (46)
Stratified	108 (54)
Baseline comparison between study groups:	
No comparison	8 (4)
Comparison but no statistical testing	99 (50)
Comparison and statistical testing	92 (46)
Baseline difference between study groups*:	
No difference	55 (60)
Difference	35 (38)
Unclear about difference	2 (2)
Sample size:	199
Median (range) sample size	286 (14-243 000)
No of sites:	124
Median (range) No of sites†	20 (2-862)

*Denominator is 92 studies with statistical testing of baseline differences.

†Information on number of sites was not provided in 13 multisite trials.

Table 2| Adjustment practices of trials in analysis. Values are numbers (percentages) unless stated otherwise

Variables	Trials (n=199)
Analysis plan in Methods section	
Unadjusted only	96 (48)
Adjusted only	59 (30)
Both unadjusted and adjusted	42 (21)
Unclear regarding adjustment	2 (1)
Analysis reported in Results section	
Unadjusted only	96 (48)
Adjusted only	57 (29)
Both unadjusted and adjusted	42 (21)
Unclear regarding adjustment	4 (2)
Adjustment for baseline factors	
Groups were similar:	
Not adjusted	74 (59)
Adjusted	46 (37)
Unclear	5 (4)
Groups were different:	
Not adjusted	18 (41)
Adjusted	25 (57)
Unclear	1 (2)
Group difference was unclear:	
Not adjusted	17 (57)
Adjusted	13 (43)
Adjustment for stratification variables (n=108)*	
Not adjusted	60 (56)
Adjusted	42 (39)
Unclear	6 (6)
Any adjustment (stratification or baseline)	
None	95 (48)
Adjusted	99 (50)
Unclear	5 (3)
No of adjustment variables	99
Median (range) No of adjustment variables	3 (1-13)

*See supplementary file.

Table 3| Adjusted and unadjusted estimates of intervention effect (n=28)

Intervention comparison	Primary outcome	Metric	Unadjusted estimate (95% CI)	Adjusted estimate (95% CI)
Liquid based cytology versus conventional PAP test in screened women ¹²	Detection of CIN grade II+*	Detection rate ratio/incidence rate	0.99 (0.85 to 1.16)	1.00 (0.83 to 1.19)
Thalidomide versus placebo in lung cancer ¹³	Death	Hazard ratio	1.09 (0.93 to 1.27)	1.11 (0.94 to 1.32)
Valproic acid versus placebo in amyotrophic lateral sclerosis ¹⁴	Death	Hazard ratio	1.67 (0.83 to 3.33)	1.54 (0.75 to 3.22)
Pralidoxime chloride versus saline placebo in organophosphorus poisoning ¹⁵	Death	Hazard ratio	1.82 (1.01 to 3.28)†‡	1.69 (0.88 to 3.26)
Artemether-lumefantrine versus oral quinine in malaria ¹⁶	Treatment failure	Hazard ratio	0.09 (0.03 to 0.30)†	0.09 (0.03 to 0.30)†
Standard surgery+pelvic lymphadenectomy versus standard surgery in endometrial cancer ¹⁷	Death	Hazard ratio	1.16 (0.87 to 1.54)	1.04 (0.74 to 1.45)
Riluzole versus placebo in Parkinson's disease ¹⁸	Death	Hazard ratio	1.09 (0.88 to 1.35)	1.09 (0.88 to 1.35)
Selective digestive tract decontamination versus standard care in intensive care unit patients ¹⁹	Death	Hazard ratio	0.94 (0.82 to 1.08)	0.83 (0.72 to 0.97)†‡
Selective oropharyngeal tract decontamination versus standard care in intensive care unit patients ¹⁹	Death	Hazard ratio	0.95 (0.82 to 1.10)	0.86 (0.74 to 0.99)†‡
Usual care+aerobic exercise training versus usual care alone in heart failure ²⁰	All cause mortality or admission to hospital	Hazard ratio	0.93 (0.84 to 1.02)	0.89 (0.81 to 0.99)†‡
Valsartan versus placebo in atrial fibrillation ²¹	Recurrence of atrial fibrillation	Hazard ratio	0.98 (0.85 to 1.13)	0.97 (0.83 to 1.14)
Abacavir-lamivudine versus tenofovir-emtricitabine in patients infected with HIV ²²	Virologic failure	Hazard ratio	2.33 (1.46 to 3.72)†	2.08 (1.28 to 3.37)†
Laparoscopic versus open resection in colon cancer ²³	Disease progression	Hazard ratio	1.09 (0.87 to 1.35)	1.08 (0.87 to 1.35)
Capecitabine+trastuzumab versus capecitabine alone in breast cancer ²⁴	Progression	Hazard ratio	0.69 (0.48 to 0.97)†	0.66 (0.43 to 0.99)†
Radiotherapy+radical prostatectomy versus radical prostatectomy alone in prostate cancer ²⁵	Death or biochemical progression	Hazard ratio /risk ratio	0.53 (0.37 to 0.79)†	0.56 (0.37 to 0.83)†
Home management versus usual care in patients with malaria ²⁶	Treatment incidence density per person-year*	Incidence rate ratio	0.57 (0.47 to 0.68)†	0.58 (0.49 to 0.70)†
Antioxidant supplement versus placebo in chronic pancreatitis ²⁷	Reduction in painful days per month*	Mean difference	-4.15 (-6.23 to -2.07)†	-2.33 (-3.51 to -1.15)†
Local versus systemic steroids in rotator cuff disease ²⁸	Pain and disability index score	Mean difference	-5.2 (-13.9 to 3.5)	-4.1 (-12.3 to 4.1)
Behavioral analysis+standard treatment versus standard treatment in adults with intellectual disabilities ²⁹	Aberrant behavior checklist	Mean difference	-1.38 (-2.54 to -0.22)†	-0.89 (-1.74 to -0.04)†
Intracoronary streptokinase versus no additional therapy in patients requiring percutaneous coronary intervention ³⁰	Infarct size at six months	Mean difference	-10.5 (-16.9 to -4.1)†	-10.2 (ND)
Lifestyle intervention versus control in obese people ³¹	Body mass index at 12 months	Mean difference	-0.11 (-0.77 to 0.55)	-0.11 (-0.45 to 0.22)
Terlipressin infusion+banding ligation versus terlipressin infusion alone in acute oesophageal variceal bleeding ³²	Treatment failure, early bleeding	Odds ratio	0.10 (0.013 to 0.87)†	0.08 (0.010 to 0.63)†
Routine care+thigh length graduated compression stockings versus routine care in stroke ³³	Deep vein thrombosis	Odds ratio	0.97 (0.75 to 1.26)	0.98 (0.76 to 1.27)
Peer support versus standard care in women at risk for postpartum depression ³⁴	Depression scale (Edinburgh)	Odds ratio	0.47 (0.30 to 0.72)†	0.44 (0.28 to 0.68)†
Triple therapy versus nicotine patch alone in smokers ³⁵	Abstinence at 26 weeks*	Odds ratio	0.43 (0.19 to 0.97)†	0.40 (0.16 to 0.95)†
Hand hygiene only versus control in households ³⁶	Secondary attack ratio	Odds ratio	0.52 (0.27 to 1.00)	0.57 (0.26 to 1.22)
Hand hygiene+facemask versus control in households ³⁶	Secondary attack ratio	Odds ratio	0.67 (0.36 to 1.25)	0.77 (0.38 to 1.55)
Computer based multirisk assessment report versus standard care in women ³⁷	Interpersonal violence detection*	Risk ratio	0.48 (0.22 to 1.00)†‡	0.50 (0.24 to 1.11)

Table 3 (continued)

Intervention comparison	Primary outcome	Metric	Unadjusted estimate (95% CI)	Adjusted estimate (95% CI)
Computer based medication reconciliation versus standard care in hospital patients ³⁸	Potential adverse drug events	Risk ratio	0.74 (0.60 to 0.89)†	0.72 (0.52 to 0.99)†
Methylprednisolone versus placebo in multiple sclerosis ³⁹	Relapse rate	Risk ratio	0.37 (0.23 to 0.59)†	0.33 (0.19 to 0.56)†

CIN=cervical intraepithelial neoplasia; NA=no data.

Two trials^{19 36} had two comparison categories in each (for example, treatment A versus control and treatment B versus control), making a total of 30 comparisons in 28 trials. Across all trials, effect sizes are presented such that relative risk metrics <1.00 and mean difference metrics <0 mean that experimental intervention is better than control intervention.

*Favorable outcome. Estimates and 95% confidence intervals recalculated for consistency across table: effect size was inverted (1/effect size) for relative risk metrics and was changed sign for mean difference metrics.

†Significant effect based on 95% confidence interval.

‡Discordant results between unadjusted and adjusted effect.

Table 4| Reasons intervention estimates were missing for either unadjusted or adjusted model (n=10)

Study	Quotes
Adler et al ⁴⁰	Adjustment for age and sex had no effect on the results either
Dumville et al ⁴¹	Time to ulcer healing did not differ between the groups (log rank test 1.00, df=2, P=0.62)
Lainchbury et al ⁴²	Cox regression (allowing for time to event, age ≤75 or >75 years, and time by age interactions), an overall treatment effect spanned the 3 years (P=0.033)
Muss et al ⁴³	The rates of relapse in the Capecitabine group were nearly twice those in the standard chemotherapy group (unadjusted). Adjusted estimate was given HR=2.09 . . .
Piper et al ⁴⁴	The same effects (group) were obtained when logistic regression analyses controlled for race, sex, and site
Plint et al ⁴⁵	The relative risk of admission by day 7 in group 1 as compared with group 4 was 0.65 (95% confidence interval, 0.45 to 0.95; P=0.02 and P=0.07 for the unadjusted and adjusted analyses, respectively)
Hofmeijer et al ⁴⁶	These results were the same after adjustment for age . . .
Spencer et al ⁴⁷	PFS adjusted for either . . . or . . . , or both, remained significant
Kesecioglu et al ⁴⁸	The estimated odds ratio for death at day 28 in the usual care group versus the HL 10 group was 0.75 (95% CI, 0.48–1.8; P=0.22). A Kaplan-Meier survival plot for the 28-day mortality is shown in Figure 2 (P=0.34 by log-rank test)
Vandemheen et al ⁴⁹	Adjustment of the results for sex had no effect on the results

Figure

Flow diagram for 199 evaluated trials: availability of information on model adjustment for primary outcomes from trial registry, design paper, or full protocol provided by trial authors