

Some Main Problems Eroding the Credibility and Relevance of Randomized Trials

John P. A. Ioannidis, M.D., Ph.D.

Abstract

Randomized trials are an excellent research design with major advantages. However, randomized trials are not immune to biases, and inferences from them may be sometimes flawed or irrelevant. The present review addresses, in brief, some of the major threats to the credibility and relevance of the results of clinical trials: power problems, biases affecting internal validity (poor design, conduct, and analysis), biases affecting the total randomized evidence on a specific topic (publication bias and selective outcome and analysis reporting bias), lack of relevance, poor generalizability, and biases in the interpretation of the results.

The development of the randomized trial design has been one of the most important, major advances in medicine. Randomization has proven to be the best method to alleviate concerns about confounding variables and leads to the most appropriate creation of two or more groups to be compared. Nevertheless, as any research design in real life, randomized trials are not perfect or immune to biases, and their inferences may sometimes be flawed or even irrelevant.

Empirical evidence suggests that discrepancies over time can occur in randomized trials on the same topic.¹ In particular, diminishing effects are common in clinical medicine. For example, across 100 meta-analyses of mental health related

interventions, effect sizes associated with pharmacotherapies were far more likely to diminish rather than increase with the appearance of new trials.² Even for the most frequently cited randomized trials, discrepancies and lack of replication have been documented in some investigations.³ Some of these discrepancies may reflect genuine differences in the treatment effects of various trials (legitimate effect modification with different settings, patients, co-interventions, and so forth); however, many of them may simply reflect biases and lack of credibility.

There are many problems that are able to erode the credibility and relevance of randomized trials, but few efforts have been made to collect all the empirical evidence about these problems and see the total magnitude of the challenges at hand.⁴ The importance of all these problems may vary from topic to topic, and there is also some remaining controversy, even among experts, about which problems are the most critical and how frequently they are encountered. While sorting out these controversies is beyond the scope of this article, the following review will present a list of probable, common problems that are useful to consider when evaluating a randomized trial and its results (Table 1). I acknowledge up-front that the emphasis that I give to different items is unavoidably subjective (and possibly even biased).

Power Problems

Most randomized trials are underpowered to study the hard clinical outcomes in which most clinicians and patients are interested. The current, average sample size is only 80 patients.⁵ It can be shown⁶ that if a trial finds a formal statistically significant result, then the probability that the effect is not null indeed is given by $JR/(JR+\alpha)$, where J is the study power, R is the pre-study odds of an effect being present, and α is the type I error. If a trial finds a nonstatistically significant result, then the probability that there is no effect

John P. A. Ioannidis, M.D., Ph.D., is Professor and Chairman, Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine and Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece, and Adjunct Professor of Medicine, Tufts University School of Medicine, Boston, Massachusetts.

Correspondence: John P. A. Ioannidis, M.D., Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece 45110; jioannid@cc.uoi.gr.

Table 1 Some Main Questions to Ask on the Credibility and Relevance of Randomized Trials

-
1. Is the trial adequately powered for a clinically relevant outcome?
 2. Is there bias in the specific trial based on its design, conduct, and reporting?
 3. Even if this trial seems unbiased, is there bias in the total randomized evidence that has been accumulated from similar trials on similar questions?
 4. Is the randomized evidence relevant, asking the right question(s)?
 5. What is the generalizability of the results?
 6. Is the interpretation of the results appropriate?
-

is given by $(1-\alpha)/[(R-JR + 1-\alpha)]$. For example, if we have an underpowered trial with $J = 0.20$, $R = 0.2$, and $\alpha = 0.04$, we get $0.20 \times 0.2 / (0.20 \times 0.2 + 0.04) = 50\%$ probability for a “positive” result to be correct and $(1-0.04)/[0.2-0.20 \times 0.2 + 1-0.04] = 0.96/1.12 = 86\%$ probability for a “negative” result to be correct. With much higher pre-study odds, such as $R = 1$, the probabilities become 83% and 55%, respectively, i.e., what is gained in positive predictive value is lost in negative predictive value. These calculations are under ideal circumstances where there is absolutely no bias operating, clearly a very optimistic, if not unrealistic, scenario.

We should note also that power depends on the postulated magnitude of the effect size under study. Few medical interventions have very large effect sizes (e.g., almost all patients die if they do not take the treatment, while almost all patients live if they do take the treatment). In the vast majority of cases, even optimistic anticipations aim at modest incremental benefits. This means that even sizeable trials have low power to demonstrate these small benefits for hard clinical outcomes.

Bias in a Single Study

As has been shown,^{6,7} a research finding cannot reach credibility over 50% unless u is less than R , i.e., bias must be less than the pre-study odds, where bias is generally defined as any reason (other than random error) that creates a formally statistically significant finding (“positive”) while this should not be so. Thus, if $R = 0.2$, and if 20% of the nonstatistically significant (“negative”) results can become “positive,” a trial claiming a formally statistically significant effect would always be more likely to be false than true.

The poor quality of randomized trials is a research topic that has drawn intense attention in the literature.⁸ Early empirical evaluations suggested that effect sizes may depend on aggregate quality scores. This has been dismissed since there are so many quality scores, with the result that inferences are widely different.⁹ Other empirical evaluations suggested that specific quality items, such as problems in the generation of the randomization sequence, lack of blinding, and lack of allocation concealment in randomized controlled trials, in particular, may inflate treatment effects.^{8,10} However, it now seems more likely that such quality deficits may be associated either with inflated or with deflated treatment effects.¹¹ Also, it is difficult to tell up-front what quality deficits do and how they operate on the magnitude or even

presence of a treatment effect. “Averaging” quality is probably wrong.¹² A randomized trial with one major flaw may obtain the wrong answer. A randomized trial with two major flaws may produce a result that is even more wrong or may paradoxically produce a somehow more correct answer if the two flaws pull the treatment effect in opposite directions.

The appraisal of quality in randomized trials is further hampered by the suboptimal reporting of these studies. The CONSORT (Consolidated Standards of Reporting Trials) statement¹³ is a valiant effort toward improving the reporting of trials, and empirical studies suggest that reporting has improved over time. It is less clear whether improvements in reporting reflect genuine improvements also in the design of the trials and their protection from bias. They simply may be an effort to fit accepted norms of reporting, with dissociation from the actual design and conduct of the trial, which may still be biased. One should understand the difference between reported quality and actual quality and that they do not always coincide. A trial that seems to be of high quality at reporting may have had serious problems while it was being conducted that are not transparent in the reported paper. Conversely, a trial that does not report information on some important design issues cannot be assumed to have failed these aspects of the design.¹⁴ Finally, even with careful reporting, some major design terms are used with a different connotation by different investigators. For example, the terms “double-blind” or “masking” can have subtle but important differences in their use across trials.^{15,16}

In broader terms, one could consider two kinds of bad quality. First, the quality may be bad on purpose, because a conflict exists and there is pressure to show a specific result. Then, one has to expect that the effect sizes are almost always inflated. Second, quality may be bad because of “stupidity”: poor design, conduct, and analysis without conflicted purpose. Subsequently, the effect sizes may be affected in either direction; it may be even more likely that they are deflated compared with the truth (a consequence of accumulated, nondifferential misclassification errors).

Biases in the Total Randomized Evidence

Every randomized trial is not a single statistical analysis in isolation. Usually, there are many outcomes analyzed and many analyses performed per outcome in the same trial. Moreover, there are often several other similar trials on the same or similar themes. Due to a conglomerate of forces,

the presented “visible” evidence may be different from the real total evidence.

Potential conflicts of interest (financial, corporate, academic, scientific, etc.) may lead to publication bias,¹⁷ time lag bias,¹⁸ selective outcome and analysis reporting bias,¹⁹ or even fraudulent results.²⁰ While the latter category is probably uncommon, the other biases are likely to have a major common impact on the credibility of the available, visible randomized evidence.²¹

The problem with all these biases is that unless one pursues them with very elaborate detective work, they largely remain invisible. Efforts have been undertaken to reduce the hiding places. Trial registration²² is a very important step in the direction of diminishing publication bias and raising awareness about the existence of unpublished data. However, registration alone does not address time lag bias, and it also does not currently protect from selective analysis and outcome reporting. Protocols are registered often with very minimal information on their analysis plans. This leaves a great deal of room for selection. Evaluation of large domains of randomized research suggests that, overall, there are more significant trials compared to what one would expect, even if the described effects were true.²¹ Selective reporting biases are probably greater in observational research,²³ but its impact in randomized trials should not be underestimated. With prevailing conflicts of interest, biases may be very strong on some occasions. The problem is that it is very difficult to pinpoint where exactly these biases have operated more heavily and which particular interventions would be affected. With increasing awareness of the importance of detailed registration,²⁴ one hopes that these biases will diminish in the future.

Poor Relevance

Under the issue of poor relevance, one may consider all the reasons why a randomized trial may not be truly relevant, no matter how well it is conducted, analyzed, and reported. Strictly speaking, the credibility of the result remains high from a quantitative perspective, but qualitatively this is no longer the key issue. In brief, the trial has answered (probably correctly) the wrong question.

There are many biases that belong in this category. Setting the research agenda may be influenced by a virtually uncountable number of factors, not all of which pertain to helping people and promoting science.^{25,26} Truly clinically useful, pragmatic trials seem to be in the minority.²⁷ Trials may neglect prior evidence²⁸ or may be informed by biased prior evidence. New interventions may continue to be compared against placebo or no treatment (often because of regulatory dicta), while the true question is how well they perform against other available treatments that have been documented to be effective.²⁶ Active comparators may be selected that have suboptimal efficacy or even cause harm, so as to show a superior benefit from the new intervention (strawman design).²⁹ Head-to-head comparisons of effective

agents may be avoided sometimes, while in other situations head-to-head comparisons may be very popular for promoting drugs that have never been shown to be conclusively effective against placebo or no treatment. At the opposite of low power (described above), some trials may be overpowered to detect significant benefits for surrogate end points that bear no clinical importance.³⁰ Finally, trials may be designed and reported aiming at demonstrating effectiveness, while measurement and reporting of harms may be neglected,³¹⁻³³ thus causing an imbalance in the risk-benefit information flowsheet.

I will only briefly mention here also the issue of the genuine generalizability of results,^{34,35} i.e., whether the trial results can be extrapolated to different populations than those where the trial was conducted. Extrapolation of applicability is always a leap of faith but not necessarily an extravagant one. It is very difficult or even impossible to represent all of the different types of patients and settings in a trial population. Even with several trials conducted, a few important subpopulations may be underrepresented in the accumulated evidence. However, sometimes the underrepresentation is systematic, i.e., patients at high risk, those with important (and common) comorbidities, and those taking other drugs may have been systematically understudied. Deciding whether a seemingly beneficial intervention would be effective in these additional settings becomes often a difficult judgment call.

Interpretation Biases

Interpretation of any research finding adds another level of complexity and leaves more room for bias. Often there is some unavoidable subjectivity about interpretation and varying respected views can possibly be defended with appropriate arguments. However, one should take into account the following issues.

First, effects with similar credibility may look different, depending on how they are presented. Some classic tricks are that odds ratios may be larger than risk ratios and relative risks in general are more impressive to read, compared with absolute risk reductions. For example, a relative risk reduction of 30% may correspond to an absolute risk reduction of 0.03%, if the risk in the control group is 0.1%. Even worse, is the focus on p values rather than effect sizes. P values alone say nothing about how effective or harmful an intervention is.³⁶

Selective discussion and dissemination of results is an area that needs more study. However, it is possible that several studies may suffer from selective invocation of external evidence (mentioning only the data from other studies that are favorable), silencing or downplaying of limitations,³⁷ and inappropriate generalization.³⁴ Conflicts of interest again need to be considered. For example, it has been demonstrated that for a similar profile of treatment effects and adverse events, trials sponsored by the industry are more likely to be presented as showing favorable results compared with

nonindustry sponsored studies.³⁸

Going one step further, to the network of scientific information, scientific citation bias³⁹ is well documented, favoring the citation of studies with “positive” and the most favorable results and avoiding citations of studies with “negative” results. In the current research environment, the majority of high profile clinical trials are funded by the industry alone, and very few such trials do not have at least some co-funding by the industry.⁴⁰ Thus, projects that are highly cited become citation “poles” that cannot be ignored, even by the most independent and nonconflicted scientists.

There is also evidence for “ghost management” of the literature,⁴¹ where companies hire professional ghostwriters, who write up trial results, as well as comments, editorials, and so forth. These papers are then published with the names of prestigious clinician investigators in the article’s byline. Skewed public dissemination with direct patient advertising⁴² and a huge, multifarious marketing effort^{43,44} can compound the picture further.

Finally, given the prestige of the randomized experimental design and the regulatory aura surrounding such, it is unavoidable that at times “positive” trials may create a barrier to the conduct of additional randomized trials. If the results are correct, then ethical considerations would suggest that this is a proper choice.⁴⁵ However, if the results are not correct, we run into the situation where a false finding stifles further efforts to overturn it. Discerning the false from the true, based on the above, is clearly not an easy thing to do.

Conclusions

Randomized controlled trials are a brilliant, simple design with a solid history of successful utilization in clinical research. They can offer extremely useful evidence, and they are a must for documenting the effectiveness of proposed interventions. Nevertheless, this does not mean that they cannot suffer from important major biases that can erode their credibility and relevance. I have attempted to cover some of the main problems but more may be present.⁴⁶ Moreover, I have limited the discussion to clinical trials of traditional, mainstream medical interventions. Most likely, these problems get worse when we enter alternative or complementary medicine or other non-mainstream domains that, nevertheless, also have a flurry of randomized trials being conducted, often with exclusively favorable results in numerous settings.⁴⁷ For example, it is probably fair to say that clinical trials on homeopathy interventions simply measure bias (or placebo effects) rather than actual, specific treatment effects.⁴⁸ This is not the case with mainstream medical interventions. Overall, we have learned a lot from well conducted randomized trials. These designs will most likely continue to be a strong bastion of clinical research in the foreseeable future. The accompanying uncertainty and healthy skepticism do not mean that decisions should not be made based on their outcomes,⁴⁹ but that it is

important always to acknowledge the wisdom of *caveat lector*—let the reader beware.

Disclosure Statement

The author has no financial or proprietary interest in the subject matter or materials discussed, including, but not limited to, employment, consultancies, stock ownership, honoraria, and paid expert testimony.

References

- Ioannidis JPA, Lau J. Evolution of treatment effects over time: empirical evidence from recursive cumulative meta-analyses. *Proc Natl Acad Sciences, USA.* 2001;98:831-6.
- Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol.* 2004;57:1124-30.
- Ioannidis JPA. Contradicted and initial stronger effects in highly-cited clinical research. *JAMA.* 2005;294:218-28.
- Gluud LL. Bias in clinical intervention research. *Am J Epidemiol.* 2006;163:493-501.
- 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.
- Ioannidis JP. Why most published research findings are false: author’s reply to Goodman and Greenland. *PLoS Med.* 2007;4:e215.
- Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. *BMJ.* 2001;323:42-6.
- Jüni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282:1054-60.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408-12.
- Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA.* 2002;287:2973-82.
- Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol.* 1994;140:290-6.
- Altman DG, Schulz KF, Moher D, et al: The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134:663-94.
- Soares HP, Daniels S, Kumar A, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *BMJ.* 2004;328:22-4.
- Montori VM, Bhandari M, Devereaux PJ, et al. In the dark: the reporting of blinding status in randomized controlled trials. *J Clin Epidemiol.* 2002;55:787-90.
- Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med.* 2002;136:254-9.
- Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet.* 1991;337:867-72.

18. Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279:281-6.
19. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*. 2005;330:753.
20. Welch HG, Mogielnicki J. Presumed benefit: lessons from the American experience with marrow transplantation for breast cancer. *BMJ*. 2002;324:1088-92.
21. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4:245-53.
22. Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration: looking back and moving ahead. *Lancet*. 2007;369:1909-11.
23. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. *PLoS Med* 2007;4:e79.
24. Reveiz L, Krleza-Jeric K, Chan AW, De Aguiar S. Do trialists endorse clinical trial registration? Survey of a PubMed sample. *Trials* 2007;8:30.
25. Ioannidis JP. Indirect comparisons: the mesh and mess of clinical trials. *Lancet*. 2006;368:1470-2.
26. Salanti G, Kavvoura FK, Ioannidis JPA. Exploring the geometry of treatment networks. *Ann Intern Med*. 2008;148:544-53.
27. Zwarenstein M, Oxman A. Pragmatic Trials in Health Care Systems (PRACTIHC). Why are so few randomized trials useful, and what can we do about it? *J Clin Epidemiol*. 2006;59:1125-6.
28. Bero L, Oostvogel F, Bacchetti P, et al. Factors associated with findings of published trials of drug-drug comparisons: why some statins appear more efficacious than others. *PLoS Med*. 2007;4:e184.
29. Chalmers TC, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. *Stat Med*. 1996;15:1263-8.
30. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125:605-13.
31. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA*. 2001;285:437-43.
32. Ioannidis JP, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. *Ann Intern Med*. 2006;144:298-300.
33. Ioannidis JP, Evans SJ, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141:781-8.
34. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet*. 2005;365:82-93.
35. Travers J, Marsh S, Williams M, et al. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*. 2007;62:219-23.
36. Goodman SN. Toward evidence-based medical statistics. 1: The p value fallacy. *Ann Intern Med*. 1999;130:995-1004.
37. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol*. 2007;60:324-9.
38. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? *JAMA*. 2003;290:921-8.
39. Kjaergard LL, Gluud C. Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol*. 2002;55:407-10.
40. Patsopoulos NA, Ioannidis JP, Analatos AA. Origin and funding of the most frequently cited papers in medicine: database analysis. *BMJ*. 2006;332:1061-4.
41. Sismondo S. Ghost management: how much of the medical literature is shaped behind the scenes by the pharmaceutical industry? *PLoS Med*. 2007;4:e286.
42. Donohue JM, Cevalco M, Rosenthal MB. A decade of direct-to-consumer advertising of prescription drugs. *N Engl J Med*. 2007;357:673-81.
43. Gagnon MA, Lexchin J. The cost of pushing pills: a new estimate of pharmaceutical promotion expenditures in the United States. *PLoS Med*. 2008;5:e1.
44. Giannakakis I, Ioannidis JPA. Arabian nights: 1001 tales of how pharmaceutical companies cater to the material needs of doctors. *BMJ*. 2000;321:1563-4.
45. Mann H, Djulbegovic B. Choosing a control intervention for a randomised clinical trial. *BMC Med Res Methodol*. 2003;3:7.
46. Yazici H. Use and abuse of the controlled clinical trial. *Bull NYU Hosp Jt Dis*. 2007;65:132-4.
47. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials*. 1998;19:159-66.
48. Shang A, Huwiler-Müntener K, Nartey L, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet* 2005;366:726-32.
49. Djulbegovic B, Hozo I. When should potentially false research findings be considered acceptable? *PLoS Med*. 2007;4:e26.