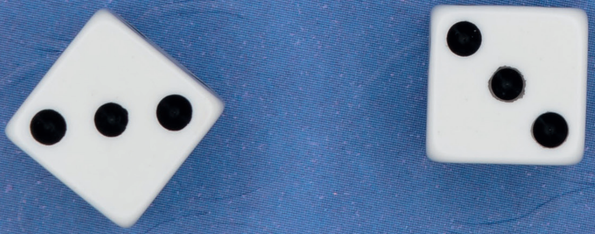
Randomisation can do many things – but it cannot "fail"



Although randomisation has long been seen as crucial to reaching reliable insights from data, it is still falling victim to some peculiar – and troublesome – misconceptions. By **Arthur H. Owora**, **John Dawson**, **Gary Gadbury**, **Luis M. Mestre**, **Greg Pavela**, **Tapan Mehta**, **Colby J. Vorland**, **Pengcheng Xun** and **David B. Allison**

Randomisation is a study design strategy that mitigates the risk of study findings being the result of some form of bias – either known or unknown. In a study of a new medical treatment, say, you would have different groups of people, and each of the people in the study would have different heights, weights and ages, to name just a few variables. By randomising these people to each of our study groups, researchers are able to assume that the distributions of all these variables (and more) will be identical for all groups in the long run. If through repeated or larger experiments, the new treatment given to one of our study groups is seen to perform better than the old treatment given to another study group, the researchers can have confidence that the treatment is more effective overall: that the results are not just because one group consists of all the people who are shorter, or slimmer, or older.

This type of study is called a randomised controlled trial (RCT), and experiments like these are now generally recognised to represent the best method for determining the causal effects of variables.[1–4] They are by no means perfect: as with any other study design, RCTs have inherent limitations and can suffer from suboptimal execution. However, two misconceptions about RCTs seem to be common: one is the belief that properly implemented randomisation can "fail"; the other is the belief that baseline imbalances lock biases into trial findings. Here we explain why these are misconceptions and offer suggestions for responding to some fundamental concerns that may underlie these beliefs.

Misconception 1: Properly implemented randomisation can fail

Researchers often claim that randomisation has "failed" when, after randomising subjects into study groups, they look at the average baseline characteristics of these groups and notice a difference in one or more variables. Table 1 shows an example of a typical report of RCT baseline characteristics. For each group, the mean of age, height and

Table 1: Example of independent *t*-test results for baseline characteristic differences in a two-arm RCT.

Characteristic	Group 1 ($N = 100$) mean (SD)	Group 2 ($N = 100$) mean (SD)	*p*-value
Age (years)	41.0 (3.7)	42.4 (5.3)	0.03
Height (in.)	64.9 (4.6)	64.7 (3.8)	0.74
Weight (kg)	71.9 (10.1)	70.7 (8.9)	0.37

Markus Winkler/Unsplash.com

Arthur H. Owora is assistant professor in the Department of Epidemiology and Biostatistics, Indiana University, School of Public Health-Bloomington.

John Dawson is assistant professor in the Department of Nutritional Sciences, Texas Tech University, College of Human Sciences.

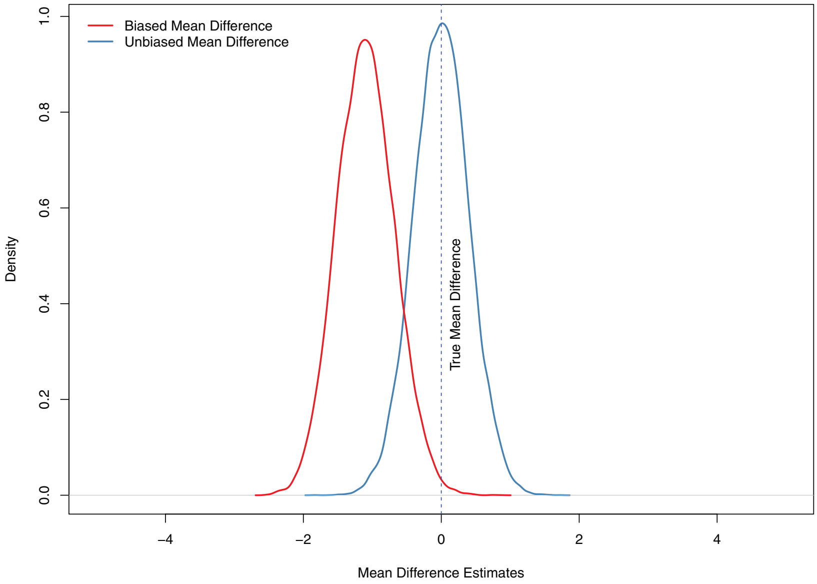


Figure 1: Distribution of biased and unbiased parameter estimates (mean score difference based on 10,000 experiments).

weight is calculated and compared, and a *p*-value is calculated.

The *p*-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.[5] Here we see that for age, the reported *p*-value is 0.03, meaning that the probability of observing a difference at least this extreme given the assumption of no difference in mean age between the groups is only 3%. This *p*-value is below the conventional threshold of "statistical significance" ($p = 0.05$, with the significance level also being known by the term alpha, or α), so a researcher might look at this table and conclude that "randomisation has failed". But they would be wrong to do so, for two reasons.

Firstly, randomisation does not guarantee that on any given "roll of the dice" all variables will have identical empirical distributions across all groups in any finite sample.[6] Some deviation from perfect equality is expected – and, indeed, if we did not observe at least some deviation from perfect equality, that would raise suspicions as to whether investigators were truly randomly assigning their subjects. Carlisle found precisely this in an influential 2017 study of retracted RCTs: "the distribution of means for baseline variables in randomized, controlled trials was inconsistent with random sampling, due to an excess of very similar means and an excess of very dissimilar means".[7] Similar findings had been seen years earlier by Schulz *et al*.: "In reports of trials that had, apparently, used unrestricted randomisation, the differences in sample sizes between treatment and control groups were much smaller than would be expected due to chance".[8] Such studies support the suspicion that investigators can be so uncomfortable with deviations from the equality expected under randomisation that they may deliberately attempt to reduce them through (presumably non-random) assignment methods.

It is important to keep in mind that random assignments, when faithfully and properly implemented, may produce surprisingly different distributions for the various variables affecting experimental outcomes – and those differences may even reach statistical significance. But this brings us to our second issue, which is that calculating the *p*-value of such differences is illogical.

p-values are based on the assumption that randomness alone is responsible for the observed difference. As such, they cannot also be used to test that assumption. A small *p*-value, suggesting that the observed difference is improbable if randomisation was properly implemented, either means randomisation was not properly implemented – either because of error or fraud – or that the improbable has happened. If the researcher knows that randomisation *was* properly implemented, then this is simply a chance finding. Thus, the common practice of testing for the statistical significance of baseline values in RCTs makes no sense except as a detector for fraud or errors, and in no way indicates a failure of randomisation. Simply put, randomisation is a *process*, not an *outcome*. Even when the process is implemented correctly, sometimes outcomes may appear to be non-random but, in fact, are the result of a random process.

Misconception 2: Baseline imbalances lock biases into trial findings

A researcher who looks at Table 1 and (wrongly) concludes that randomisation has failed might then make a second mistake, by deciding that the difference that exists in the

Gary Gadbury is professor emeritus in the Department of Statistics, Kansas State University, College of Arts and Sciences.

Luis M. Mestre is a research assistant and PhD student in epidemiology with a minor in data science at Indiana University, School of Public Health-Bloomington.

Greg Pavela is associate professor in the Department of Health Behavior, The University of Alabama at Birmingham, School of Public Health.

FAQs for practitioners

Having made the case that baseline imbalances in the samples of randomised experiments neither demonstrate failed randomisation nor create bias, does this mean that no further consideration of baseline imbalances is warranted? Not at all.

Is it always pointless to test for baseline imbalances?
No. Testing for significant differences in baseline variables is entirely reasonable if the goal is to detect selection bias (e.g., in cluster randomised trials where subject recruitment is performed after random allocation of clusters to treatment).[9] Testing is also reasonable if the goal is to detect potential errors or fraud in implementing the randomisation.[10,11] However, testing for significant differences in baseline variables in RCTs, unless one has specific reasons to look for errors or fraud in data production, is pointless. If one knows one has randomised properly, statistical tests of significance offer no meaningful information.

Should we ignore methods for improving balance in baseline covariates?
No. Multiple methods are available to minimise baseline imbalances, and benefits exist to doing so.[11] This may be especially valuable in studies with small numbers of randomised units, especially in cluster randomised controlled trials.[12] It is important to note that implementing such procedures, which may include various forms of blocked, adaptive, minimisation, matched or stratified randomisation, may bring with it additional analytic nuances or requirements to be implemented when the data are analysed after collection.[13] These different random allocation techniques reduce the prevalence of imbalances but do not completely rule out the possibility of imbalances, whose expected magnitudes could decrease with increasing sample size.[14,15] But none of this implies that failing to address these nuances creates a bias in truly randomised experiments.

Should we forget about statistically adjusting for pre-randomisation covariates on which imbalances may occur?
No. Controlling for pre-randomisation covariates, most obviously baseline values of variables in RCTs when those variables are risk factors (or correlates) but not longitudinal mediators of the outcome variables, can be sensible and, in almost all cases, will improve statistical power in truly randomised experiments.[16] In doing so, pre-planning and pre-specification is advised.[17–20] For example, controlling for baseline age when examining risk reduction of a dichotomous outcome (e.g., type 2 diabetes) between two weight-loss interventions in an RCT is wise (i.e., age is a prognostic factor for type 2 diabetes) and can improve statistical power.[20] However, again, while doing this is prudent, not planning to statistically adjust and not statistically adjusting for perceived imbalances in baseline values or potential imbalances in baseline values does not create bias in true and properly executed RCTs.

Should we only statistically adjust for imbalances of key pre-randomisation covariates if this was pre-planned?
No. It is reasonable to include as covariates variables on which baseline imbalances have occurred in the sample, even if doing so was not pre-planned or pre-specified.

However, such analyses should be labelled as secondary or sensitivity analyses, and the fact that they were not pre-planned should be disclosed. The choice of including a covariate should not be based on the statistical significance of baseline treatment group differences.[17] Extensive discussions on the value of, and methods for, adjusting for baseline covariates on which imbalances have occurred in randomised experiments exist.[18] Yet, to reiterate, failing to adjust for any such baseline variables on which imbalances have occurred in the sample of a truly and properly randomised experiment does not create bias. On the other hand, adjusting for a baseline imbalance due to chance could bias a previously unbiased result (e.g., adjusting for mediating factors).[19]

It should be emphasised that the importance of baseline imbalances depends on the nature of the effect being investigated. For example, suppose you are working on an RCT for preventing cardiometabolic outcomes. If, at baseline, 40% of patients with obesity are in one arm and only 5% of patients with obesity are in the other arm, regardless of the *p*-value associated with the baseline obesity proportion difference between the two RCT arms, obesity is a critical prognostic/risk factor for cardiometabolic outcomes, and therefore this should be considered in the interpretation of treatment efficacy. But, to reiterate, the observed imbalance does not mean that randomisation failed. Rather, it is a chance finding – a result that is possible even with randomisation.

baseline age characteristic will bias the final outcome. This is not true, and to understand why, it is important to be clear about what bias is and what *p*-values are actually telling us.

Bias is most familiar as a systematic shift in the expected observed value of a sample estimate of a parameter – say, the mean value – from its true population value. These deviations have many potential causes, such as the source of patients included in an RCT, or hand-picking which ones get treated and which do not. These sorts of biases are precisely what randomisation is used to eliminate. The mistake researchers make when they worry about baseline imbalances is to think that any *single* RCT will be free of differences that might look at first glance to be systematic bias. As we have already seen, randomisation is quite capable of producing differences that look like bias – but, unlike true bias, these "anomalies" will cancel each other out as more RCTs are conducted.

Bias can also appear in the context of statistical tests that generate *p*-values. This time the bias affects the probability that a test rejects a hypothesis, given the assumptions made by the test. In the case of *p*-values, the null hypothesis is that there is no real difference between (say) the survival of patients in both the treatment and placebo arm, with any observed difference being due simply to chance. The definition of a *p*-value then implies that if we find a *p*-value of, say, 0.012, we can be sure the probability of observing a *p*-value lower than this is also 0.012. If the test is biased, however, we would not have that assurance.

Put more formally, a biased test may be defined as one that produces a distribution

Tapan Mehta is associate professor and director of research in the Department of Health Services Administration, The University of Alabama at Birmingham, School of Health Professions.

Colby J. Vorland is a post-doctoral fellow in the Department of Applied Health Science, Indiana University, School of Public Health-Bloomington.

Pengcheng Xun is an associate director of epidemiology and outcomes research at Atara Biotherapeutics, United States.

David B. Allison is dean, distinguished professor, and provost professor at Indiana University, School of Public Health-Bloomington.

of parameter estimates and/or *p*-values that do not conform to the rule that Prob($p < \alpha$) = α for all α, even though the null hypothesis is true. Thus, just as with a biased estimate of a parameter, the bias of a test refers to long-range expectations over (usually hypothetical) repeated runs of the experiment.

To explain it another way, bias – whether in the estimate of parameters characterising treatment effects or in the production of *p*-values in significance testing procedures – is a characteristic of the sampling distribution of sample statistics (e.g., mean score difference). In either case, bias has no meaning when referring to a single realisation of a sample statistic extracted from a randomised sample because the bias is known to be zero. Hence, we can speak of biased estimation procedures and biased testing procedures, but we cannot speak of a single realisation of a testing procedure as biased except in two instances: first, when the expected value of the estimate has been theoretically or empirically shown to have a mean unequal to the true parameter being estimated (Figure 1, page 21); and/or second, when we are speaking colloquially to mean that it is the result of a biased testing procedure.

If an RCT is conducted properly and the statistical analysis is chosen and performed properly, then an unbiased estimator (e.g., a difference in the means of two Gaussian random variables) cannot produce a biased estimate or a biased test of statistical significance. That is, if appropriate estimation and statistical significance testing procedures are being used, then unbiased estimates and test results are produced, on average, in the long run. That is what unbiased means. There is no guarantee that in any one "roll of the dice" the number that comes up will be exactly the population parameter when providing a sample estimate. Similarly, there is no guarantee that one will never make a Type I error (a mistaken rejection of the null hypothesis) or obtain a *p*-value less than α when conducting a significance test, even if the testing procedure is correct and the null hypothesis is true. The tests allow for some errors to occur under the null hypothesis; this is not an indication of bias.

If, say, the experimental design of a study is such that it is at risk of producing deviations from what one expects from the theory of sampling – such as the absence of randomisation – then any deviations observed could indeed be due to bias. But to look at data from what we presume to be an appropriately obtained randomised sample and say "these data don't look like what I would expect them to look like" and then insist that this can only be the result of bias makes no sense. Yet that is exactly what is done when one focuses on those samples in which baseline differences appear large and then claims that, under that subset of realisations, there is bias. In randomised samples, there is no bias in the long run, and it is only in the long run that we can make claims about the unbiasedness of our estimation and testing procedures.

In conclusion…

Randomisation is a strategy for addressing biases that might otherwise undermine the reliability of insights from a study. It ensures the unbiasedness of parameter estimates and *p*-values testing the causal effects of independent variables, regardless of whether any real or perceived imbalance of baseline values occurs. Crucially, randomisation can only be expected to work its "magic" in the long run and not necessarily in any one single study. And if done properly, free from error and fraud, randomisation *cannot* "fail". ■

Acknowledgements
The authors thank Matthew Loop, Keith Baggerly, Carol Etzel, Rebecca Lipscomb, and Jennifer Holmes for helpful comments on this paper.

Disclosure statement
The authors declare no competing interests in relation to this article.

References

1. Greenland, S. (1990) Randomisation, statistics, and causal inference. *Epidemiology*, **1**(6), 421–429.

2. Cartwright, N. (2009) What are randomised controlled trials good for? *Philosophical Studies*, **147**(1), 59.

3. Backmann, M. (2017) What's in a gold standard? In defence of randomised controlled trials. *Medicine, Health Care and Philosophy*, **20**(4), 513–523.

4. Attia, P. (2018) Studying studies: Part IV – randomisation and confounding. peterattiamd.com/ns004/

5. Wasserstein, R. L. and Lazar, N. A. (2016) The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, **70**(2), 129–133.

6. Altman, D. G. (1985) Comparability of randomised groups. *The Statistician*, **34**(1), 125–136.

7. Carlisle, J. B. (2017) Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, **72**(8), 944–952.

8. Schulz, K. F., Chalmers, I., Grimes, D. A. and Altman, D. G. (1994) Assessing the quality of randomisation from reports of controlled trials published in obstetrics and gynecology journals. *Journal of the American Medical Association*, **272**(2), 125–128.

9. Bolzern, J. E., Mitchell, A. and Torgerson, D. J. (2019) Baseline testing in cluster randomised controlled trials: Should this be done? *BMC Medical Research Methodology*, **19**(1), 106.

10. Bolland, M. J., Avenell, A., Gamble, G. D. and Grey, A. (2016) Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology*, **87**(23), 2391.

11. Carlisle, J. B. (2012) The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, **67**(5), 521–537.

12. Ivers, N. M., Halperin, I. J., Barnsley, J., *et al*. (2012) Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials*, **13**(1), 120.

13. Hewitt, C. E. and Torgerson, D. J. (2006) Research methods: Is restricted randomisation necessary? *British Medical Journal*, **332**(7556), 1506–1508.

14. Xiao, L., Lavori, P. W., Wilson, S. R. and Ma, J. (2011) Comparison of dynamic block randomization and minimization in randomized trials: A simulation study. *Clinical Trials*, **8**(1), 59–69.

15. Carter, B. R. and Hood, K. (2008) Balance algorithm for cluster randomized trials. *BMC Medical Research Methodology*, **8**, 65.

16. Allison, D. B. (1995) When is it worth measuring a covariate in a randomized clinical trial? *Journal of Consulting and Clinical Psychology*, **63**(3), 339–343.

17. Moher, D., Hopewell, S., Schulz, K. F., *et al*. (2010) CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal*, **340**, c869.

18. Wei, L. and Zhang, J. (2001) Analysis of data with imbalance in the baseline outcome variable for randomized clinical trials. *Drug Information Journal*, **35**(4), 1201–1214.

19. Rothman, K. J. (1977) Epidemiologic methods in clinical trials. *Cancer*, **39**(4 Suppl.), 1771–1775.

20. Hernández, A. V., Steyerberg, E. W. and Habbema, J. D. F. (2004) Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, **57**(5), 454–460.