International Journal of Clinical Trials
Chen H et al. Int J Clin Trials. 2020 May;7(2):150-153
http://www.ijclinicaltrials.com
pISSN 2349-3240 | eISSN 2349-3259

DOI: http://dx.doi.org/10.18203/2349-3259.ijct20201720

Review Article

Testing for baseline differences in clinical trials

Henian Chen*, Yuanyuan Lu, Nicole Slye

Study Design and Data Analysis Center, College of Public Health, University of South Florida, USA

Received: 29 August 2019
Accepted: 21 January 2020

***Correspondence:**
Dr. Henian Chen,
E-mail: hchen1@usf.edu

ABSTRACT

Reporting statistical tests for baseline measures of clinical trials does not make sense since the statistical significance is dependent on sample size, as a large trial can find significance in the same difference that a small trial did not find to be statistically significant. We use 3 published trials using the same baseline measures to provide the relationship between trial sample size and p value. For trial 1 sequential organ failure assessment (SOFA) score, p=0.01, 10.4±3.4 vs. 9.6±3.2, difference=0.8; p=0.007 for vasopressors, 83.0% vs. 72.6%. Trial 2 has SOFA score 11±3 vs. 12±3, difference=1, p=0.42. Trial 3 has vasopressors 73% vs. 83%, p=0.21. Based on trial 2, supine group has a mean of 12 and an SD of 3 for SOFA score, while prone group has a mean of 11 and an SD of 3 for SOFA score. The p values are 0.29850, 0.09877, 0.01940, 0.00094, 0.00005, and <0.00001 when n (per arm) is 20, 50, 100, 200, 300 and 400, respectively. Based on trial 3 information, the vasopressors percentages are 73.0% in the supine group vs. 83.0% in the prone group. The p values are 0.4452, 0.2274, 0.0878, 0.0158, 0.0031, and 0.0006 when n (per arm) is 20, 50, 100, 200, 300 and 400, respectively. Small trials provide larger p values than big trials for the same baseline differences. We cannot define the imbalance in baseline measures only based on these p values. There is no statistical basis for advocating the baseline difference tests

Keywords: Baseline difference, Statistical significant testing, Randomization, Trial size

INTRODUCTION

Randomization ensures that allocation of patients to different treatment groups is left purely to chance in randomized controlled trials (RCTs). When the groups have similar characteristics at baseline, and only one of the groups receives an intervention, the difference in outcomes between the groups can therefore be attributed to the new intervention. The Consolidated Standards of Reporting Trials (CONSORT) statement advocates a table of descriptive statistics of baseline measures, and discourages the use of significant tests.[1] Statistical significant testing for baseline differences between the groups is inappropriate in well-conducted RCTs, but unfortunately these tests are still common. A review of 300 cluster randomized trials published between 2000-2008 found that 58% reported significance tests of baseline balance.[6] Another review of journals found a decrease in the reporting of p values in baseline tables, from 58% in 1987 and 48% in 1997 to 34.8% of papers published from 2008-2010.[7] A significant decline in baseline comparison reporting would be expected after the publishing of the CONSORT statement; however, this is not the case. A systematic evaluation of sports medicine journals found that 64.0% of studies published in 2015 still reported statistical baseline tests, a slight decrease from 67.1% in 2005.[8]

There are many reasons why testing baseline differences is inappropriate. It is claimed that baseline comparisons show whether randomization was successful, even though there is no cut-off that dictates when differences in baseline measures are in line with proper randomization.[9] Zhao and Berger remark that perfect balanced baseline is

neither feasible nor necessary.[10] Though significance tests can show an imbalance between treatment groups, they do not assess whether these imbalances might have affected the results.[11] Baseline comparisons that show an imbalance may lead to adjustment for covariates that are not strongly related to the outcome or not pre-specified in the statistical analysis plan.[12,13] Differences in baseline characteristics are related to the sample size of the groups. As sample size increases, the baseline differences between groups are expected to decrease. The ability to detect these differences becomes greater as the sample size becomes larger.[14] The same size imbalance will have a greater effect on the statistical tests for larger sample sizes.[15] Below we show the relationship between sample size and p-value of baseline differences using 3 published trials that report two of the same baseline measures.

EXAMPLES

When statistical tests are conducted to assess baseline comparability, they can be reported in several ways: report p values in a column of the table, make a footnote under the table, or have a statement in the main text about which baseline measures were statistically significantly different.[16-18] To illustrate how sample size can affect the significance of similar baseline differences, we selected two baseline measures: sequential organ failure assessment (SOFA) score and vasopressors (%) from three published randomized controlled trials (RCTs) related to the effect of prone positioning in severe acute respiratory distress syndrome (Table 1).[16-18] Trial 1 evaluated the effect of early application of prone positioning on outcomes in patients with severe acute respiratory distress syndrome (ARDS), trial 2 compared the effect of prone positioning vs. supine position on the duration of mechanical ventilation, and trial 3 evaluated the effect of prone ventilation being implemented early in the course of ARDS, applied for most of the day, and maintained for a prolong period of time.[16-18] Trial 1 has 229 and 237 patients in the supine group and the prone group, respectively, and provides descriptive statistics for both SOFA score and vasopressors in the baseline table. The authors of this paper make a footnote under the baseline table and state in the results section that the mean SOFA score and percentage of vasopressors are significantly different between the supine group and the prone group without presenting the exact p-value. We calculated the p-value using a t-test and a Chi-square test for SOFA score and vasopressors (%) based on their baseline table: p=0.01 for SOFA score, 10.4±3.4 vs. 9.6±3.2, difference=0.8; p=0.007 for vasopressors, 83.0% vs. 72.6%. The authors added these two covariates in their data analysis models to adjust for the imbalance of them between the groups. Trial 2 has 19 and 21 patients in the supine and the prone group, and provides descriptive statistics and p-value for SOFA score in the baseline table: 11±3 vs. 12±3, difference=1, p=0.42. Trial 3 has 60 and 76 patients in the supine and the prone group, and provides descriptive statistics and p-value for vasopressors in the baseline table: 73% vs. 83%, p=0.21. When we compare trial 1 and trial 2 in baseline SOFA score, we find that trial 1 has a smaller difference (0.8) than trial 2 (1.0). However, trial 1 had p<0.05, while trial 2 had p>0.05. Trial 1 and trial 3 reported a similar difference in percentages of vasopressors (83% vs. 73%), but p=0.21 for trial 3 while p=0.007 for trial 1.

Table 1: Characteristics of the participants from three published clinical trials.

	Trial 1[18]			Trial 2[16]			Trial 3[17]		
	Supine group (n=229)	**Prone group (n=237)**	**P value**	**Supine group (n=19)**	**Prone group (n=21)**	**P value**	**Supine group (n=60)**	**Prone group (n=76)**	**P value**
SOFA score mean (SD)	10.4 (3.4)	9.6 (3.2)	0.011	12 (3)	11 (3)	0.42			
Vasopressors (%)	83.0	72.6	0.007				73	83	0.21

Table 2: P values with corresponding sample size based on the same difference from trial 2 and trial 3.*

Sample size (per group)	Trial 2 p value (unpaired T test)	Trial 3 p value (Chi-square test)
20	0.29850	0.4452
50	0.09877	0.2274
100	0.01940	0.0878
200	0.00094	0.0158
300	0.00005	0.0031
400	0.00000	0.0006
500	0.00000	0.0001

*Trial 2[16]; Trial 3[17]

SMALL TRIALS VERSUS BIG TRIALS

Below we provide p values for the same difference for different trial sample sizes. Based on trial 2 information, supine group has a mean of 12 and an SD of 3 for SOFA score, while prone group has a mean of 11 and an SD of 3 for SOFA score. The p values are 0.29850, 0.09877, 0.01940, 0.00094, 0.00005, and <0.00001 when n (per arm) is 20, 50, 100, 200, 300 and 400, respectively (Table 2). Based on trial 3 information, the vasopressors percentages are 73.0% in the supine group vs. 83.0% in the prone group. The p values are 0.4452, 0.2274, 0.0878, 0.0158, 0.0031, and 0.0006 when n (per arm) is 20, 50, 100, 200, 300 and 400, respectively (Table 2). This huge spread in p values is only due to differences in the trial sample size, as the two groups have the exact same difference.

HOW TO USE THE BASELINE TABLE?

What is the baseline table for if the p values cannot be used to evaluate the quality of randomization and detect the imbalance? There are four uses of the baseline table that are appropriate.

First, to describe participants: it is important to know the characteristics of the participants who were actually recruited and how comparable the groups were. The baseline table allows readers, especially clinicians, to judge how relevant the results of a trial might be to a particular patient.

Next, to establish compliance with protocol: the inconsistencies between trial protocols and the final reports have been frequently documented. The baseline table is useful for an assessment of consistency between the final report and the trial protocol, for example, to assess the participant eligibility, trial sample size, and missing data.

Third, to replicate the trial and compare to similar RCTs. The baseline information is necessary for replicating a trial and can be used to compare other similar RCTs. Meta-analysis of clinical trials should be conducted for similar RCTs (for example, similar patient characteristics) based on their baseline tables. It is inappropriate to just put all published trials together from the same outcome measure and the same intervention for a meta-analysis, and ignore the similarity of these trials.

Lastly, to study trial generalizability, knowing the baseline characteristics of the trial participants allows us to assess how generalizable the results of the trial will be. The baseline demographic and clinical characteristics of a trial is useful for studying this trial's generalizability.[19] The trial sample may potentially differ from the general patient population from which they were drawn and it may not be representative of the target patient population. We can compare the baseline demographic and clinical characteristics between the trial sample and the general patient population, measure the affinity, and quantify the similarity between the trial and the patient population on the basis of the propensity score based on the information from the baseline table.

DISCUSSION

Theoretically, randomization distributes both known and unknown covariates equally between the treatment groups. Statistical significant testing for baseline differences cannot help to address whether the randomization was actually done correctly, and non-significant results cannot prove that patients were allocated randomly. As we can see, small trials provide larger p-values than big trials for the same baseline differences due to the impact of sample size on p value. We cannot define the imbalance in baseline measures only based on these p-values. Adjustment for variables because they differ significantly at baseline is likely to bias the estimated effect. It is unfair to declare that the randomization is successful based on p>0.05 for baseline measures. In other words, small trials will always do better on randomization than big trials. From a statistical standpoint, it is misleading and wrong to declare that the randomization was actually done correctly or patients were allocated randomly based on the non-significant p-values from the baseline difference tests; however, our own experience in RCT data analysis is that co-authors, reviewers, and even editors are still persistent in their demand for these significance tests and p values. There is no statistical basis for advocating the baseline difference tests. Authors of a report of an RCT should follow the CONSORT statement by including a table of baseline characteristics of patients in each group, without statistical significant tests and p values.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Moher D, Schulz S, Gotzsche P, Egger M. CONSORT 2010 explanation and elaboration: Updated guideline for reporting parallel group randomized trials. BMJ. 2010;340:c869.
2. Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994;13(17):1715-26.
3. Altman D, Doré C. Randomisation and baseline comparisons in clinical trials. Lancet. 1990;335(8682):149-53.
4. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355(9209):1064-9.
5. Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. J Clin Epidemiol. 2015;68(6):603-9.

6. Schulz KF. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA. 1994;272(2):125-8.
7. Knol M, Groenwold R, Grobbee D. P-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. Eur J Prev Cardiol. 2012;19(2):231-2.
8. Peterson RL, Tran M, Koffel J, Stovitz SD. Statistical testing of baseline differences in sports medicine RCTs: a systematic evaluation. BMJ Open Sport Exerc Med. 2017;3(1):e000228.
9. Boer MRD, Waterlander WE, Kuijper LD, Steenhuis IH, Twisk JW. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. Int J Behav Nutr Phys Act. 2015;12:4.
10. Zhao W, Berger V. Imbalance control in clinical trial subject randomization—from philosophy to strategy. J Clin Epidemiol. 2018;101:116-8.
11. Altman DG. Comparability of Randomised Groups. Statistician. 1985;34(1):125.
12. 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(19):2917-30.
13. Mutz DC, Pemantle R, Pham P. The perils of balance testing in experimental design: Messy analyses of clean data. Am Stat. 2018;73(1):32-42.
14. Sedgwick P. Randomised controlled trials: balance in baseline characteristics. BMJ. 2014;349:g5721.
15. Roberts C, Torgerson DJ. Understanding controlled trials: Baseline imbalance in randomised controlled trials. BMJ. 1999;319:185.
16. Voggenreiter G, Aufmkolk M, Stiletto RJ, Baacke MG, Waydhas C, Ose C, et al. Prone positioning improves oxygenation in post-traumatic lung injury - A prospective randomized trial. J Trauma. 2005;59(2):333-43.
17. Mancebo J, Fernández R, Blanch L, Rialp G, Gordo F, Ferrer M, et al. A multicenter trial of prolonged prone ventilation in severe acute respiratory distress syndrome. Am J Respir Crit Care Med. 2006;173(11):1233-9.
18. Guerin C, Reignier J, Richard J, Beuret P, Gacouin A, Boulain T, et al. Prone positioning in severe acute respiratory distress syndrome. N Engl J Med. 2013;368(23):2159-68.
19. Wang W, Ma Y, Huang Y, Chen H. Generalizability analysis for clinical trials: a simulation study. Stat Med. 2017;36:1523-31.

Cite this article as: Chen H, Lu Y, Slye N. Testing for baseline differences in clinical trials. Int J Clin Trials 2020;7(2):150-3.