BMJ Open Sport & Exercise Medicine

Statistical testing of baseline differences in sports medicine RCTs: a systematic evaluation

Ross L Peterson,[1] Matthew Tran,[2] Jonathan Koffel,[3] Steven D Stovitz[4]

To cite: Peterson RL, Tran M, Koffel J, *et al*. Statistical testing of baseline differences in sports medicine RCTs: a systematic evaluation. *BMJ Open Sport Exerc Med* 2017;**3**:e000228. doi:10.1136/bmjsem-2017-000228

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjsem-2017-000228).

Received 22 January 2017
Revised 18 April 2017
Accepted 03 May 2017

[1]Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, Minnesota, USA
[2]College of Biological Sciences, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA
[3]Health Sciences Libraries, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA
[4]Family Medicine and Community Health, University of Minnesota Medical School Twin Cities, Minneapolis, Minnesota, USA

Correspondence to
Ross L Peterson; pet00180@umn.edu

ABSTRACT

Background/Aim The CONSORT (Consolidated Standards of Reporting Trials) statement discourages reporting statistical tests of baseline differences between groups in randomised controlled trials (RCTs). However, this practice is still common in many medical fields. Our aim was to determine the prevalence of this practice in leading sports medicine journals.

Methods We conducted a comprehensive search in Medline through PubMed to identify RCTs published in the years 2005 and 2015 from 10 high-impact sports medicine journals. Two reviewers independently confirmed the trial design and reached consensus on which articles contained statistical tests of baseline differences.

Results Our search strategy identified a total of 324 RCTs, with 85 from the year 2005 and 239 from the year 2015. Overall, 64.8% of studies (95% CI (59.6, 70.0)) reported statistical tests of baseline differences; broken down by year, this percentage was 67.1% in 2005 (95% CI (57.1, 77.1)) and 64.0% in 2015 (95% CI (57.9, 70.1)).

Conclusions Although discouraged by the CONSORT statement, statistical testing of baseline differences remains highly prevalent in sports medicine RCTs. Statistical testing of baseline differences can mislead authors; for example, by failing to identify meaningful baseline differences in small studies. Journals that ask authors to follow the CONSORT statement guidelines should recognise that many manuscripts are ignoring the recommendation against statistical testing of baseline differences.

What are the new findings?

- Across a sample of 324 randomised controlled trials (RCTs) published in leading sports medicine journals in the years 2005 and 2015, about two-thirds reported statistical tests of baseline differences between randomised groups. However, this reporting is discouraged by the CONSORT (Consolidated Standards of Reporting Trials) statement.
- The proportion was not found to differ much between the years 2005 and 2015, despite the release of the CONSORT statement in 2010 that explicitly discourages the practice.
- The CONSORT statement recommends a table describing baseline characteristics of each group. However, about 20% of RCTs failed to present such a table.

INTRODUCTION

For the reporting of randomised controlled trials (RCTs), item 15 of the CONSORT (Consolidated Standards of Reporting Trials) statement recommends that researchers report the baseline characteristics of each group, ideally in a table.[1] However, the same item in the CONSORT statement also discourages statistical testing of differences in baseline covariates between randomised groups. Roughly speaking, standard statistical tests assess the probability that differences were due to chance given that the groups were the same. However, if a study is an RCT, then it is expected that the groups were the same, thus any baseline differences between the groups can be assumed to be due to chance. The CONSORT statement writes, 'Tests of baseline differences are not necessarily wrong, just illogical. Such hypothesis testing is superfluous and can mislead investigators and their readers.'[1] For example, in a study with few participants, there may be large differences between groups that do not reach a level of statistical significance and are thus ignored. Conversely, in a study with lots of participants, even small and meaningless differences may meet statistical significance and thus receive unnecessary attention.

Although discouraged by the CONSORT statement, statistical testing (eg, with the calculation of a p value) of baseline differences in RCTs is still common.[2] Knol *et al* reviewed RCTs published in seven leading medicine journals (eg, *JAMA*, *BMJ* and *Lancet*) from 2008 to 2010 and found that p values were listed in the baseline tables of about 35% of the studies.[3] The primary purpose of this study was to determine the

general proportion of RCTs in the sports medicine literature which included statistical tests of baseline differences. A secondary purpose was to assess the proportion of studies that included a table of baseline characteristics. In order to get a cursory evaluation as to the potential effect of the 2010 CONSORT statement, we chose to study RCTs published in the year 2005 or 2015.

METHODS

We sought to identify RCTs published in the year 2005 or 2015 in sports medicine journals. We included the top 10 highest-impact factor sports medicine journals according to the 2014 Journal Citation Reports that published RCTs. Our biomedical librarian, JK, conducted a search in Medline through PubMed on 22 June 2016 using the search string *clinical trial[pt] OR randomly OR randomized OR randomised*. The journals were as follows: *American Journal of Sports Medicine*; *British Journal of Sports Medicine*; *Gait & Posture*; *Journal of Applied Physiology*; *Journal of Orthopaedic and Sports Physical Therapy*; *Journal of Science and Medicine in Sport*; *Knee Surgery, Sports Traumatology, Arthroscopy*; *Medicine and Science in Sports and Exercise*; *Scandinavian Journal of Medicine & Science in Sports*; *Sports Medicine*. The full search strategy is presented in online supplementary appendix 1.

As shown in figure 1, our PubMed search of RCTs retrieved 1109 articles that were then filtered using keywords to remove obvious systematic reviews, other non-RCTs and crossover trials to develop a set of 598 potential articles to evaluate. Then, two reviewers (RLP and MT) independently examined each article and came to a consensus on which articles were RCTs, which contained tables of baseline differences between randomised groups, and which had significance testing of baseline differences. Ultimately, 324 articles were included. For the outcome of significance testing of baseline differences, we counted articles that included various forms of significance testing in either the baseline table or in the text, such as p values, t-tests and analyses of variance. Inter-rater reliability was around

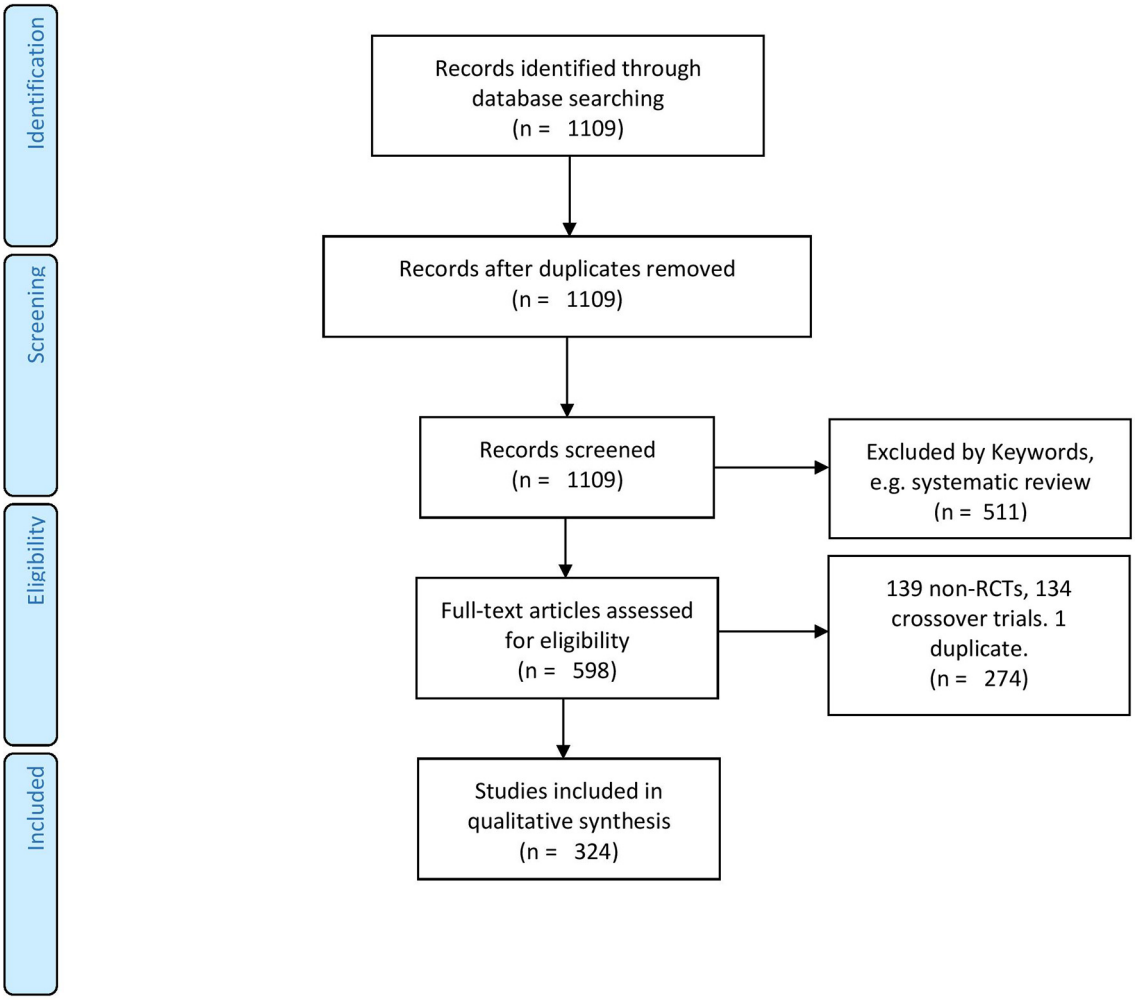

Figure 1 PRISMA flow diagram for selection of RCTs from leading sports medicine journals for the publication years 2005 and 2015. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT, randomised controlled trial.

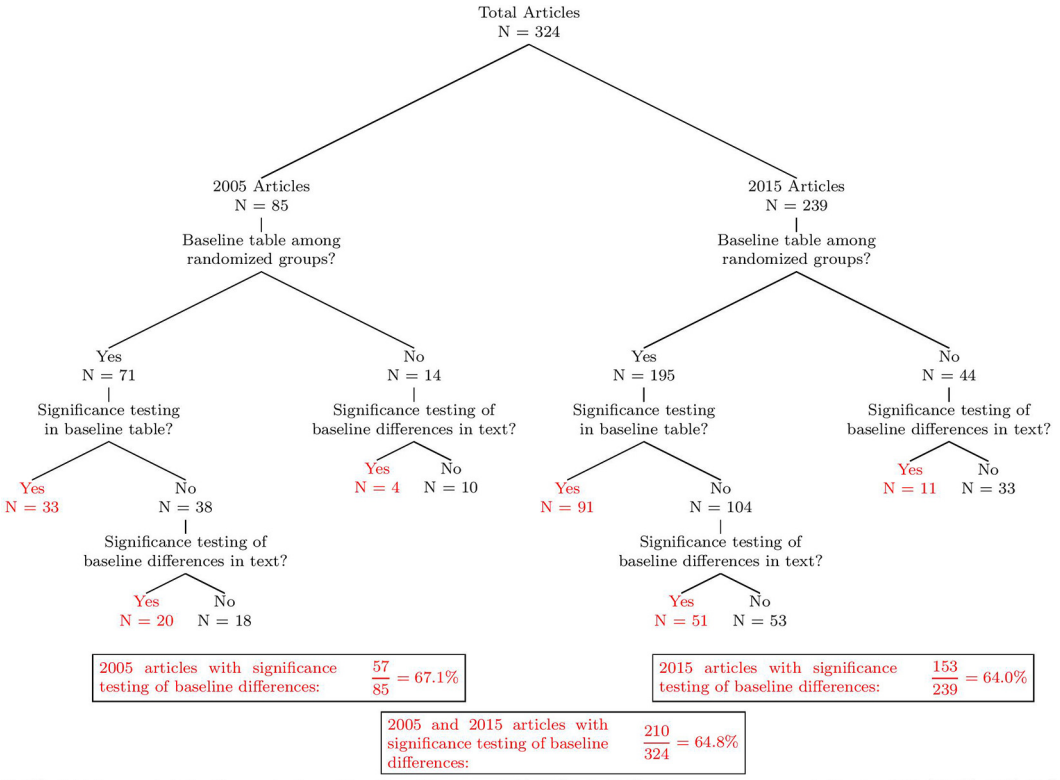

Figure 2 Significance testing of baseline differences in sports medicine RCTs from 2005 to 2015. RCT, randomised controlled trial.

88% and the two reviewers met in person to resolve disagreements.

RESULTS

As shown in figure 2, our search identified 85 and 239 RCTs published in the years 2005 and 2015, respectively. Overall, 64.8% of studies reported statistical testing of baseline differences (95% CI (59.6, 70.0)). Broken down by year, this percentage was 67.1% in 2005 (95% CI (57.1, 77.1)) and 64.0% in 2015 (95% CI (57.9, 70.1)). For both years, about 20% of studies did not contain a table displaying baseline covariates. Among those studies with a baseline table, significance testing (generally in the form of a p value) was reported in 46.5% (33/71) of baseline tables in 2005 and 46.7% (91/195) of baseline tables in 2015. The dataset and code for the statistical analysis are available as a GitHub repository (https://github.com/RPeterson4/Baseline_Covariate_Files).

DISCUSSION

The CONSORT group is a long-standing international collaboration of medical professionals from a variety of areas related to research conduct and publication including 'trialists, methodologists and medicine journal editors.'[4] Their primary aim has been to improve the reporting of RCTs, and their recommendations are endorsed by nearly 600 journals. Regarding RCTs, item 15 of the CONSORT statement advises presenting a table with baseline characteristics of the randomised groups, but discourages statistical testing of baseline differences.[1] Our systematic evaluation of RCTs published in the sports medicine literature in 2005 and 2015 found that about two-thirds reported statistical testing of baseline differences. Our results suggest that the practice of statistical testing for baseline differences is more common in sports medicine journals than in the highest-impact medical journals such as *JAMA*, *BMJ* and *Lancet*.[3] We also found that about 20% of studies across both years failed to include a baseline table.

We recognise that there are compelling reasons for why authors would argue in favour of reporting statistical tests of baseline differences. One may be to check if the randomisation was successful.[2] If there are covariate imbalances that far exceed what one would expect, then there is reason to question the randomisation process. In addition, p values provide a uniform measure of baseline differences, combining the magnitude of the differences and the sample size into a single number. Statistical tests are also easy to perform and provide the reader with more quantitative information. However, there are sound reasons to not present statistical tests of baseline differences.[5] The aim of statistical testing is to find the probability that the baseline differences would be due to chance if the groups were the same. Yet, as described by the CONSORT statement, if the participants were truly

randomised, then it is known that any baseline differences were due to chance. Furthermore, the main concern of the CONSORT group is that statistical testing can 'mislead investigators and their readers'[1] (p 21).

How can statistical tests mislead investigators? Consider a study with a small number of participants. If investigators use baseline differences as a measure to assess which covariates were different and which were roughly equal between randomised groups, then they may not consider potential confounding from a covariate if the p value falls above the cut-off for statistical significance, generally 0.05. This may be a substantial problem in the sports medicine literature where many studies have small sample sizes. For example, in a 2007 RCT studying operative versus non-operative management as treatments for mid-shaft clavicle fractures,[6] the outcomes (eg, strength, fracture non-unions) were known to correlate with the sex of the patient, a baseline covariate subject to randomisation. The operative group comprised 85% men and the non-operative group comprised 69% men, an absolute difference of 16%. Since there were only 111 participants in the trial, the p value for difference in sex was 0.06 and the authors did not adjust their analyses, stating, "there were no demographic differences between the operative and non-operative groups"[6] (p 6). However, there was a large difference in the percentage of men and women in the groups, although the difference did not reach a level of statistical significance using a p value cut-off of 0.05. Since men tend to be stronger and have fewer non-unions, this difference would be expected to affect the outcomes of the treatment groups. Our evaluation of studies published in sports medicine journals found that about 80% sampled fewer than 100 participants. Conversely, in large studies, small differences may meet statistical significance yet not be meaningful. Authors may then adjust their analyses for these differences, adding extraneous covariates to their models that may have no consequences for the results.

Our study had limitations. While we aimed to identify all RCTs in the included journals, it is possible that some were missed in our PubMed search due to ambiguous language in the title and abstracts. We attempted to minimise this risk by using a publication type term (applied as part of formal indexing in PubMed) and several keywords related to randomisation. Although we had each article independently reviewed by two authors, there is a chance that some articles were misclassified. Finally, out of convenience we selected only the years 2005 and 2015. There is a small possibility that these two years were outliers. We welcome other researchers applying similar methods to RCTs published in other past and future years.

In summary, we found that 65% of RCTs in the sports medicine literature reported statistical testing of baseline differences between randomised groups, a value that changed little when comparing articles from 2005 and 2015. Reporting statistical tests of baseline differences contrasts with recommendations from the 2010 CONSORT statement.[1] Authors should understand the rationale for and against statistical testing of baseline differences. Ideally, prior to the analysis, authors should select baseline covariates for adjustment (ie, those known to affect the outcome) and incorporate these covariates into their models. Journals that ask authors to follow the CONSORT statement guidelines should beware that many manuscripts are ignoring the recommendation against statistical testing of baseline differences.

Acknowledgements The authors would like to acknowledge Alan M Batterham, PhD, and Ian Shrier, MD, PhD, for their helpful comments on earlier versions of this manuscript.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Moher D, Hopewell S, Schulz KF, *et al*. 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Int J Surg* 2012;10:28–55 http://www.bmj.com/content/340/bmj.c869
2. de Boer MR, Waterlander WE, Kuijper LD, *et al*. 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 http://www.ijbnpa.org/content/12/1/4
3. Knol MJ, Groenwold R, Grobbee DE. P-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. *Eur J Prev Cardiol* 2012;19:231–2.
4. Consort—The CONSORT Group. http://www.consort-statement.org/about-consort/the-consort-group
5. Senn S. Testing for baseline balance in clinical trials. *Stat Med*1994;13:1715–26 https://www.ncbi.nlm.nih.gov/pubmed/7997705
6. The Canadian Orthopaedic Trauma Society. Nonoperative treatment compared with plate fixation of displaced midshaft clavicular fractures. *J Bone Jt Surg 2007* 2007;89:1–10.