Most ankle sprain research is either false or clinically unimportant: A 30-year audit of Randomized Controlled Trials

Bleakley CM PhD [1]

Matthews M PhD [2]

Smoliga JM PhD [3]

[1]School of Health Science, Ulster University, County Antrim, Northern Ireland

[2]School of Sport, Ulster University, County Antrim, Northern Ireland

[3]Department of Physical Therapy, Congdon School of Health Science, High Point University, North Carolina, USA

Corresponding author

Dr Chris Bleakley

Room 1D117

School of Health Science,

Ulster University,

County Antrim,

Northern Ireland

Tel: +442890366025

c.bleakley@ulster.ac.uk

Abstract

Lateral ankle sprain (LAS) is the most common musculoskeletal injury. Although clinical research in this field is growing, there is a broader concern that clinical trial outcomes are often false and fail to translate into patient benefits. The aim of this review was to audit 30 years of experimental research related to LAS management (n=74 RCT) and to determine if reports of treatment effectiveness could be validated beyond statistical certainty. Seventy-seven percent of trials reported positive treatment effects but there was a high risk of false discovery. Most trials were unregistered and relied solely on statistical significance, or lack of statistical significance, rather than interpreting key measures of minimum clinical importance (eg. minimal detectable change, minimal clinically important difference). Future clinical trials must adopt higher standards of reporting and data interpretation. This includes consideration of the ethical responsibility to preregister their research; and interpretation of clinical outcomes beyond statistical significance.

Background

Lateral ankle sprain (LAS) is the most prevalent musculoskeletal injury in physically active populations.[1] Although often considered innocuous, LAS has the highest re-injury rate across all lower limb musculoskeletal injuries,[2] and the annual costs associated with sports-related ankle sprain in the Netherlands is estimated at €187,200,000.[3] LAS also occurs frequently in the general population, with large cohorts suffering chronic problems;[4] indeed, 30[5]-75%[6] develop a clinical condition known as chronic ankle instability (CAI), characterized by recurrent injury and self-reported instability.[5] The long-term costs associated with LAS and CAI are significant[7 8] and relate to lower quality of life,[9] physically inactivity[4] and an increased risk of post-traumatic ankle osteoarthritis.[5 10-12,13]

Randomized controlled trials (RCTs) are currently considered to be the gold standard methodology for determining treatment superiority.[14] The first RCT involving acute LAS was published in 1972.[15] The Physiotherapy Evidence Database (PEDro) now archives over 150 RCTs involving patients with LAS or CAI, and a 2017 meta-evaluation[16] in this field included 46 systematic reviews. Having access to high volumes of experimental research should improve the quality of healthcare, but there is much concern that many clinical trial outcomes are either false[17,18] or they fail to translate into clinical benefits for patients.[19] False discovery in science (eg. erroneously claiming a treatment is effective) often occurs due to over reliance on frequentist reasoning and *p*-value thresholds;[20] a problem further compounded by unplanned multiple testing, selected reporting, and confirmation bias.[21]

Recently we introduced a four-point checklist (FAIR), which aims to validate experimental research beyond statistical certainty.[22] The checklist assesses the following criteria: ***False Positive Risk (FPR)***, which is 'the probability of observing a statistically significant *p*-value and declaring that an effect is real, when it is not'.[23] ***A priori registration***, which is essential for controlling the 'degrees of freedom' researchers have during data analysis and reporting,[21] thereby reducing the risk of false positive findings. ***Clinical Importance***, whereby the magnitude of treatment effect is compared to relevant minimal detectable change (MDC) and minimal clinically important difference (MCID)[24] data. And finally, ***Replication***, which should underpin all scientific discovery.

Evidence based health care relies on the production of valid experimental data that translates into clinical benefits. This review examines the validity of conclusions from 30 years of clinical trials into one of the most common musculoskeletal injuries - LAS and CAI. Our primary objective was to examine the extent to which reports of treatment effectiveness in this field, could be validated beyond statistical certainty. The FAIR checklist[22] was applied, with higher validity placed on trials presenting with: low false positive risk; pre-registration; treatment effect magnitudes which exceeded relevant MDC and MCID values; and the corroboration of treatment effectiveness through independent replication.

Methods

Trial selection

Review methods aligned with PRISMA.[25] Electronic searching was undertaken independently by two authors (CB, MM) on MEDLINE, and the Physiotherapy Evidence

Database (PEDro).[26 27] In MEDLINE we undertook a broad search strategy based on MeSH terms (ankle AND randomized controlled trial) and we used the PEDro search interface to run three separate searches for clinical trials using the terms 'ankle sprain', 'chronic ankle instability', and 'CAI'. Citation tracking was also undertaken using a recent meta-evaluation.[16] To be eligible for inclusion, trials must have met the following criteria: a randomized controlled design; participants with LAS and/or CAI managed with at least one conservative treatment intervention; assessment of at least one clinically relevant outcome measure (eg. pain, function, range of motion, strength, balance). Trials were excluded if they involved any surgical intervention. No restrictions were placed on injury severity, participant demographics or follow-up duration. We did not include RCTs using: >2 treatment arms, equivalency or non-inferiority trials, pilot trials or trials published prior to 1990. Any disagreements in trial selection were resolved through consensus with a third reviewer (JS).

Data extraction and analysis

PICO (population, intervention, comparison, outcome) characteristics were extracted from the full text of all eligible trials, in addition to aims and hypothesis, n participants, follow-up time points, and the total number of between-group statistical comparisons undertaken. Included trials were then classified as being either statistically significant or null. A statistically significant trial was defined as a trial having a *p*-value less than 0.05 in the trial results tab for any clinical outcome.[28] We also calculated the proportion of between-group comparisons that resulted in statistically significant findings within each individual trial, and whether they were recorded in primary or secondary outcome

measures. When trials included multiple outcome measures but did not clearly specify a 'primary' outcome, the primary outcome was determined by the authors based on the nature of the research question and the following definition of a primary outcome 'a specific key measurement(s) or observation(s) used to measure the effect of experimental variables in a trial.'[29] The FAIR checklist[22] was applied as follows:

False Positive Risk

Calculation of FPR followed methods used in a previous research audit in this field.[30] FPR calculation is a special case of Bayesian analysis. It allows the p-value to be supplemented by a single number that gives a much better idea of the strength of the evidence than a p-value alone.[23] We calculated FPR for all trials reporting a statistically significant finding from their primary outcome. All FPR calculations were performed using the False Positive Risk Web Calculator (version 1.5) using the following data: the *n* of participants in each group; a relevant *p*-value; and the corresponding effect size (Hedges g).[31] Further details of the analysis script and simulated examples of FPR calculations can be found in Colquhoun's recent articles.[20 23] If a trial reported a *p*-value threshold such as p<0.05, rather than an exact *p*-value, we assumed that the *p*-value was one decimal place below the threshold value (e.g. p<0.05 was inputted as 0.049). The calculation of FPR also requires an estimation of the prior probability that there is a real effect [P(H1)] for a given treatment. In all trials, we initially assumed that P(H1) was 0.5 – ie. treatment interventions had a 50:50 chance of a (positive) real effect before the experiment was done.[18 20] In all cases FPR estimations were calculated using the p-equals method, as our aim was to interpret a single *p*-value from a single experiment

(rather than trying to estimate the long term error rate).[31] Descriptive statistics were used to determine the median FPR and the number (%) of statistically significant *p*-values associated with FPR less than 5%.

A Priori trial registration

We determined the number (%) of eligible trials reporting preregistration; defined as the trial protocol being publicly available within a trial registry (*e.g.*ClinicalTrials.gov) prior to the initiation of participant recruitment. In a secondary analysis, we used odds ratios (ORs) and 95% confidence intervals (95% CI) to determine whether the likelihood of reporting a statistically significant outcome was influenced by *a priori* trial registration.

Clinical Importance

Initially, we determined the number (%) of trials that referenced or reported MDC and/or MCID values within the full text manuscript. When enough data were available, we calculated the mean differences (MD) and 95% confidence intervals (CI) for each clinical outcome, where MD = mean $_{experimental}$ – mean $_{control}$. MD (95% CI) data were then compared to corresponding MDC and MCID data. If a trial did not report MDC or MCID data for a particular outcome, we searched the literature for relevant figures and inputted them. MDC was set at confidence levels of 95% and considered to be 'the amount of change that must be observed before it is considered above the bounds of measurement error'.[32] MCID was considered to be 'the smallest change that would be important to patients', and could have been quantified by externally referenced (anchor) or internally referenced (distribution) methods.[33]

Replication

PICO criteria were compared across trials. If possible, homogeneous trials were sub grouped and their trial effects (magnitude and direction) were compared to screen for successful replication.

Results

We screened 1098 titles and abstracts (937 from Medline and 161 from PEDro), with 169 selected for full-text review. n=74 RCTs were eligible for inclusion (Supplemental data 1), with the remainder (n=95) excluded (>2 treatment arms (n=45); no clinical outcomes (n=9); non RCT (n=8); non English language (n=8) surgical intervention (n=7); non inferiority / equivalency (n=5), non-ankle sprain/CAI (n=5); other (n=8) (Figure 1). Trials included participants with either LAS (n=53 trials) or CAI (n=21 trials). In most trials, the primary intervention involved external supports (n=30), exercise intervention (n=18), pharmacotherapy (n=14) manual therapy or electro-physical agents (n=11). The mean sample size was n=85.1 (SD=96.8; range 13-522) and 50% (37/74) reported using *a priori* sample size calculation. Most sample size estimations included alpha (Type 1 error) and beta (Type 2 error) levels of 5% and 20% respectively, with the average effect size estimated at 0.7 (SD=0.45) a priori.

Insert Figure 1 here.

Twenty-three percent (17/74) of RCTs were classed as null (no treatment effects reported). The remaining 77% (57/74) reported statistically significant findings from at least one outcome measure. We extracted an aggregate of 966 *p*-values relating to between-group statistical comparisons involving primary or secondary outcomes, of which 35.4% (342/966) were statistically significant ($p<0.05$) (Figure 2A). Most statistically significant findings were derived from secondary outcomes, with just 17% (58/342) derived from primary outcome measures (Figure 2B). Out of the 966 *p*-values reported in the literature, only 11 (1%) represented statistically significant findings in a primary outcome measure reported from a pre-registered trial (Figure 2C). (Supplemental data 2)

Insert Figure 2 here

False positive risk

Enough data were available to calculate effect sizes and FPR in 68% of trials (39/57) reporting significant effects ($p<0.05$) in their primary outcome. FPR is summarized in Figure 3; the median FPR was 14% (range 0.6 to 100%) and 28% of trials (11/39) had FPR less than 5%. (also see Supplemental data 3)

Insert Figure 3

A Priori trial registration

Only 19% (14/74) of trials were preregistered. The average number of between-group comparisons reported across registered and unregistered trials was similar [12.8 (SD 9.0) vs 13.3 (SD 10.9) respectively], however unregistered trials were more likely to report *p*-values less than 0.05 (OR=1.7 Cis: 1.2 to 2.4; p=.004).

Clinical importance

Of the 57 trials reporting statistical significance, only 9% (5/57) made any reference to either MDC or MCID values. In a further 16 trials, we were able to extract relevant MDC and/or MCID values extracted from the existing literature, for the following outcomes measures: Foot and ankle outcome measure (FAAM);[34 35] Cumberland ankle instability tool (CAIT);[36] Lower extremity functional scale (LEFS);[37] isometric / isokinetic ankle strength;[38 39] limb circumference / swelling;[40 41] range of motion;[38 42] postural control;[27] pain[43]. Effect magnitudes (MD) exceeded the respective MDC or MCID values in 12 and 7 trials respectively. Effect magnitudes exceeded both MDC and MCID in just 3 trials (also see Supplemental data 3)

Replication

Figure 4 summarizes the number of trials meeting more than one of the FAIR criteria. Three trials were both pre-registered and reported a low FPR (<5%), and one of the pre-registered trials also reported a clinically important effect. No trial met all the following conditions: preregistered; low false positive risk (<5%); clear evidence that the magnitude of treatment effect exceeded both MDC and MCID values. There were no instances when a positive treatment effect was independently replicated.

Insert Figure 4 here

Discussion

There is concern that a large proportion of scientific research is based on false positive, non-replicable conclusions.[17] Strategies known to reduce the risk of false discovery include: mandatory trial registration;[21] false positive risk calculation,[20] and use of MDC and MCID values to determine if reported treatment magnitudes are clinically meaningful.[22 24] There is a dearth of empirical meta-research investigating the credibility of research practices in SEM research. Recent audits have highlighted a high propensity for questionable research practices (eg. HARKing, cherry picking, p-hacking) in high impact SEM journals;[44] and we have previously found a high risk of false positive claims in the sports physiotherapy literature.[30] This is the first piece of meta-research using a saturation of RCTs from a single field of musculoskeletal medicine. n=74 trials met our inclusion criteria, with 77% reporting statistically significant findings from at least one outcome measure. However, in most trials, data interpretation was limited to all or nothing Null Hypothesis Significance Testing, and most positive conclusions could not be validated beyond statistical certainty.

Only 19% of trials in the LAS/CAI research literature were preregistered. Trial registration is now required as a condition of ethical approval,[45] and audits of clinical trials undertaken in other fields of medicine (cardiology, rheumatology, and gastroenterology), show better adherence to current guidelines.[46] One of our key findings was that unregistered trials were 70% more likely to report statistical significance (OR=1.7 Cis: 1.2-2.4) compared to those that were registered *a priori*. Unregistered trials typically carry a higher risk of false discovery due to: significance seeking, selective reporting of outcomes,[47] or HARKing (hypothesizing after the results are known).[21] In contrast, preregistration helps to control the 'degrees of freedom' a researcher has during data analysis and reporting,[21] reducing

such risks. A related finding was that out of the 342 statistically significant *p*-values (<0.05) reported across trials, only 11 were generated from primary outcomes within pre-registered trials. Consequently, the vast majority of statistically significant findings within the LAS/CAI evidence base, are derived from secondary outcomes in unregistered trials, and should therefore be considered exploratory or hypothesis generating.[21]

Measures of minimum clinical importance, (MDC and MCID) are increasingly recognized as important thresholds for evaluating the efficacy of an intervention. However, the reporting of clinical significance is poor in RCTs involving patients with LAS or CAI, with just 9% of trials, referring to MDC or MCID data. After extracting MDC and MCID for clinical outcomes relating to pain, function, instability, strength and swelling, we were able to examine clinical efficacy in 21 trials; however, the results were disappointing with 50% of trials recording treatment effects which could not be differentiated from measurement error. Furthermore, in most trials, the treatment effects did not exceed relevant MCID figures, and are therefore unlikely to be considered important by patients with LAS and CAI. An initial audit[48] of interventional research in the sports medicine literature, found that MDC or MCID was considered in 53% and 40% of trials respectively. However, a much larger audit of orthopaedic literature, found that only 7.5% of clinical science articles made reference to MCID,[24]

It is expected that musculoskeletal injuries are managed from an evidence-based perspective, whereby the best available evidence is integrated with patient preference, clinical expertise, and the clinical context. As RCTs represent the gold standard

methodology for determining treatment superiority, they have a considerable influence on the relevance of adopting an evidence-based framework when treating patients with LAS or CAI. Our results raise fundamental questions about the current value of evidence-based practice in this field and clarify that future clinical trials must adopt higher standards of reporting and data interpretation. Interestingly, there is a lack of robust clinical interpretation in other fields of medicine,[49] and continuing to rely solely on NHST, not only wastes research funding, but erodes credibility and slows down scientific progress.[50] Although NHST remains an important step for determining treatment effectiveness, it is most efficient in the context of long-run repeated testing.[50] We support the idea that *p*-values are supplemented with a formal estimation of the false positive risk[18 31] which represents "the probability, in the light of the *p*-value that you observe, you declare that an effect is real, when in fact, it isn't."[23] Although it is often assumed that the FPR is equal to the reported *p*-value, they are different constructs and often vary considerably. Indeed, our audits shows that the median FPR associated with statistically significantly findings ($p<0.05$) was 14% (range 0.6-100%), and only 27% of trials had a FPR lower than 5%. These figures suggest that statistical significance alone is not a solid foundation for determining treatment effect, particularly when it is based on binary thresholds ($p<0.05$).

Limitations

Higher validity was assumed under the following conditions: derived from registered trials; low false positive risk; treatment effects exceeding MDC and MCID values. This is not an exhaustive list and we did not fully consider false discoveries relating to multiple treatment arms, the analysis of multiple outcomes, or multiple analyses of the same outcome at

different times.[51] We acknowledge although preregistration increases the transparency and validity of trial conclusions, it is not a cure-all for efficient and accurate dissemination. Audits of clinicaltrials.gov show that approximately 20% of registered trials disseminate their results within 1 year of completion,[52] with others highlighting quite a high risk of discordance between the original registry data and the published data.[53]

We must also consider that our FPR calculations were based on assumptions that the prior probability of effect was 50%, but it is likely that some trials were underpinned by more extreme hypotheses. In previous data simulations,[28] we have shown that a positive conclusion from an optimistic research question (*i.e*. a higher prior probability) is likely to be correct; whereas an unlikely hypotheses (where researchers are driven by pursuit of novelty) will have a much higher risk of false-positive reporting. Alternatives to FPR have been discussed by Colquhoun.[23] Perhaps the most clinically intuitive option is use of a reverse Bayesian approach,[54] where the observed p-value is used to calculate the prior probability required to achieve a specific or minimal false positive risk (eg. 5%). This then allows the researcher to determine whether the calculated prior is plausible or not.[30] Finally, many latent constructs influence false discovery; this includes a scientific culture which places most value on statistically significant findings or novel discoveries.[21]

Conclusion

There is a high risk of false positive discovery in a core field of musculoskeletal research. A key concern is that most of the research in this field remains unregistered, and relies solely on statistical significance, or lack of statistical significance, rather than interpreting the magnitude of change. Researchers must consider the ethical responsibility to

preregister their research; and their interpretation of clinical outcomes must evolve beyond statistical significance.

Author Contributions

CB and JS conceived of the presented idea. CB and MM planned undertook the review. CB and JS extracted data. CB undertook much of the analysis and JS verified the analytical methods.

All authors discussed the results and contributed to the final manuscript.

Competing interests

Authors have no competing interests to declare

References

1. Gribble PA, Bleakley CM, Caulfield BM, et al. 2016 consensus statement of the International Ankle Consortium: prevalence, impact and long-term consequences of lateral ankle sprains. *Br J Sports Med* 2016;50(24):1493-95. doi: 10.1136/bjsports-2016-096188
2. Hootman JM, Dick R, Agel J. Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. *J Athl Train* 2007;42(2):311-9.
3. Hupperets MD, Verhagen EA, Heymans MW, et al. Potential savings of a program to prevent ankle sprain recurrence: economic evaluation of a randomized controlled trial. *Am J Sports Med* 2010;38(11):2194-200. doi: 10.1177/0363546510373470
4. Hiller CE, Nightingale EJ, Raymond J, et al. Prevalence and impact of chronic musculoskeletal ankle disorders in the community. *Arch Phys Med Rehabil* 2012;93(10):1801-7. doi: 10.1016/j.apmr.2012.04.023
5. Gribble PA, Bleakley CM, Caulfield BM, et al. Evidence review for the 2016 International Ankle Consortium consensus statement on the prevalence, impact and long-term consequences of lateral ankle sprains. *Br J Sports Med* 2016;50(24):1496-505. doi: 10.1136/bjsports-2016-096189
6. Anandacoomarasamy A, Barnsley L. Long term outcomes of inversion ankle injuries. *Br J Sports Med* 2005;39(3):e14; discussion e14. doi: 10.1136/bjsm.2004.011676
7. Waterman BR, Owens BD, Davey S, et al. The epidemiology of ankle sprains in the United States. *J Bone Joint Surg Am* 2010;92(13):2279-84. doi: 10.2106/JBJS.I.01537

8. Knowles SB, Marshall SW, Miller T, et al. Cost of injuries from a prospective cohort study of North Carolina high school athletes. *Inj Prev* 2007;13(6):416-21. doi: 10.1136/ip.2006.014720
9. Arnold BL, Wright CJ, Ross SE. Functional ankle instability and health-related quality of life. *J Athl Train* 2011;46(6):634-41.
10. Valderrabano V, Hintermann B, Horisberger M, et al. Ligamentous posttraumatic ankle osteoarthritis. *Am J Sports Med* 2006;34(4):612-20. doi: 10.1177/0363546505281813
11. Hintermann B, Boss A, Schafer D. Arthroscopic findings in patients with chronic ankle instability. *Am J Sports Med* 2002;30(3):402-9. doi: 10.1177/03635465020300031601
12. Hashimoto T, Inokuchi S. A kinematic study of ankle joint instability due to rupture of the lateral ligaments. *Foot Ankle Int* 1997;18(11):729-34. doi: Doi 10.1177/107110079701801109
13. Wikstrom EA, Hubbard-Turner T, McKeon PO. Understanding and treating lateral ankle sprains and their consequences: a constraints-based approach. *Sports Med* 2013;43(6):385-93. doi: 10.1007/s40279-013-0043-z
14. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *BMJ* 1998;316(7126):201. doi: 10.1136/bmj.316.7126.201
15. Wilson DH. Treatment of soft-tissue injuries by pulsed electrical energy. *Br Med J* 1972;2(5808):269-70.
16. Doherty C, Bleakley C, Delahunt E, et al. Treatment and prevention of acute and recurrent ankle sprain: an overview of systematic reviews with meta-analysis. *Br J Sports Med* 2017;51(2):113-25. doi: 10.1136/bjsports-2016-096178
17. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
18. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* 2014;1(3):140216. doi: 10.1098/rsos.140216
19. Heneghan C, Goldacre B, Mahtani KR. Why clinical trial outcomes fail to translate into benefits for patients. *Trials* 2017;18(1):122. doi: 10.1186/s13063-017-1870-2
20. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* 2017;4(12):171085. doi: 10.1098/rsos.171085
21. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. *Biol Rev Camb Philos Soc* 2017;92(4):1941-68. doi: 10.1111/brv.12315
22. Bleakley C, Smoliga JM. Validating new discoveries in sports medicine: we need FAIR play beyond p values. *Br J Sports Med* 2020 doi: 10.1136/bjsports-2019-101797 [published Online First: 2020/06/26]
23. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About *p*-Values. *The American Statistician* 2019;73:192-201.
24. Copay AG, Eyberg B, Chung AS, et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. *JBJS Rev* 2018;6(9):e2. doi: 10.2106/JBJS.RVW.17.00160
25. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535. doi: 10.1136/bmj.b2535 [published Online First: 2009/07/21]
26. Kamper SJ, Moseley AM, Herbert RD, et al. 15 years of tracking physiotherapy evidence on PEDro, where are we now? *Br J Sports Med* 2015;49(14):907-9. doi: 10.1136/bjsports-2014-094468
27. Plisky PJ, Gorman PP, Butler RJ, et al. The reliability of an instrumented device for measuring components of the star excursion balance test. *N Am J Sports Phys Ther* 2009;4(2):92-9.
28. Ramagopalan SV, Skingsley AP, Handunnetthi L, et al. Funding source and primary outcome changes in clinical trials registered on ClinicalTrials.gov are associated with the reporting of a statistically

significant primary outcome: a cross-sectional study. *F1000Res* 2015;4:80. doi: 10.12688/f1000research.6312.2
29. Ramagopalan S, Skingsley AP, Handunnetthi L, et al. Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res* 2014;3:77. doi: 10.12688/f1000research.3784.1
30. Bleakley C, Reijgers J, Smoliga JM. Many High-Quality Randomized Controlled Trials in Sports Physical Therapy Are Making False-Positive Claims of Treatment Effect: A Systematic Survey. *J Orthop Sports Phys Ther* 2020;50(2):104-09. doi: 10.2519/jospt.2020.9264
31. Longstaff C, Colquhoun D. http://fpr-calc.ucl.ac.uk/. [accessed 01-02-2019.
32. Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J Rheumatol* 2001;28(2):400-5.
33. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11(2):171-84. doi: 10.1586/erp.11.9
34. Martin RL, Irrgang JJ. A survey of self-reported outcome instruments for the foot and ankle. *J Orthop Sports Phys Ther* 2007;37(2):72-84. doi: 10.2519/jospt.2007.2403
35. Eechaute C, Vaes P, Van Aerschot L, et al. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: a systematic review. *BMC Musculoskelet Disord* 2007;8:6. doi: 10.1186/1471-2474-8-6
36. Wright CJ, Linens SW, Cain MS. Establishing the Minimal Clinical Important Difference and Minimal Detectable Change for the Cumberland Ankle Instability Tool. *Arch Phys Med Rehabil* 2017;98(9):1806-11. doi: 10.1016/j.apmr.2017.01.003
37. Alcock GK SP. Validation of the Lower Extremity Functional Scale on Athletic Subjects with Ankle Sprains *Physiother Canada* 2002;Fall 233-40.
38. Fraser JJ, Koldenhoven RM, Saliba SA, et al. Reliability of Ankle-Foot Morphology, Mobility, Strength, and Motor Performance Measures. *Int J Sports Phys Ther* 2017;12(7):1134-49.
39. Sekir U, Yildiz Y, Hazneci B, et al. Reliability of a functional test battery evaluating functionality, proprioception, and strength in recreational athletes with functional ankle instability. *Eur J Phys Rehabil Med* 2008;44(4):407-15.
40. Devoogdt N, Cavaggion C, Van der Gucht E, et al. Reliability, Validity, and Feasibility of Water Displacement Method, Figure-of-Eight Method, and Circumference Measurements in Determination of Ankle and Foot Edema. *Lymphat Res Biol* 2019 doi: 10.1089/lrb.2018.0045
41. Rohner-Spengler M, Mannion AF, Babst R. Reliability and minimal detectable change for the figure-of-eight-20 method of, measurement of ankle edema. *J Orthop Sports Phys Ther* 2007;37(4):199-205. doi: 10.2519/jospt.2007.2371
42. Searle A, Spink MJ, Chuter VH. Weight bearing versus non-weight bearing ankle dorsiflexion measurement in people with diabetes: a cross sectional study. *BMC Musculoskelet Disord* 2018;19(1):183. doi: 10.1186/s12891-018-2113-8
43. Alghadir AH, Anwer S, Iqbal A, et al. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *J Pain Res* 2018;11:851-56. doi: 10.2147/JPR.S158847
44. Büttner F, Toomey E, McClean S, et al. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. *Br J Sports Med* 2020 doi: 10.1136/bjsports-2019-101863 [published Online First: 2020/07/22]
45. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;310(20):2191-4. doi: 10.1001/jama.2013.281053

447 46. Mathieu S, Boutron I, Moher D, et al. Comparison of registered and published primary outcomes in
448 randomized controlled trials. *JAMA* 2009;302(9):977-84. doi: 10.1001/jama.2009.1242
449 47. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with
450 incentives for truth telling. *Psychol Sci* 2012;23(5):524-32. doi: 10.1177/0956797611430953
451 48. Nwachukwu BU, Runyon RS, Kahlenberg CA, et al. How are we measuring clinically important
452 outcome for operative treatments in sports medicine? *Phys Sportsmed* 2017;45(2):159-64. doi:
453 10.1080/00913847.2017.1292108
454 49. Cocks K, King MT, Velikova G, et al. Quality, interpretation and presentation of European
455 Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in
456 randomised controlled trials. *Eur J Cancer* 2008;44(13):1793-8. doi: 10.1016/j.ejca.2008.05.008
457 50. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A
458 Reassessment. *Front Hum Neurosci* 2017;11:390. doi: 10.3389/fnhum.2017.00390
459 51. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the
460 what, why, when and how. *Int J Epidemiol* 2017;46(2):746-55. doi: 10.1093/ije/dyw320
461 52. Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on
462 ClinicalTrials.gov: cross sectional study. *BMJ* 2012;344:d7373. doi: 10.1136/bmj.d7373
463 53. Goldacre B, Drysdale H, Dale A, et al. COMPare: a prospective cohort study correcting and
464 monitoring 58 misreported trials in real time. *Trials* 2019;20(1):118. doi: 10.1186/s13063-019-
465 3173-2
466 54. Matthews R. Why should clinicians care about Bayesian methods? *J Stat Plan Inference* 2001;94:43-
467 58. doi: doi:10.1016/S0378-3758(00)00232-9

468

469

470

471

472

473

474 **Figure 1**

475 Flow diagram summarizing trial selection

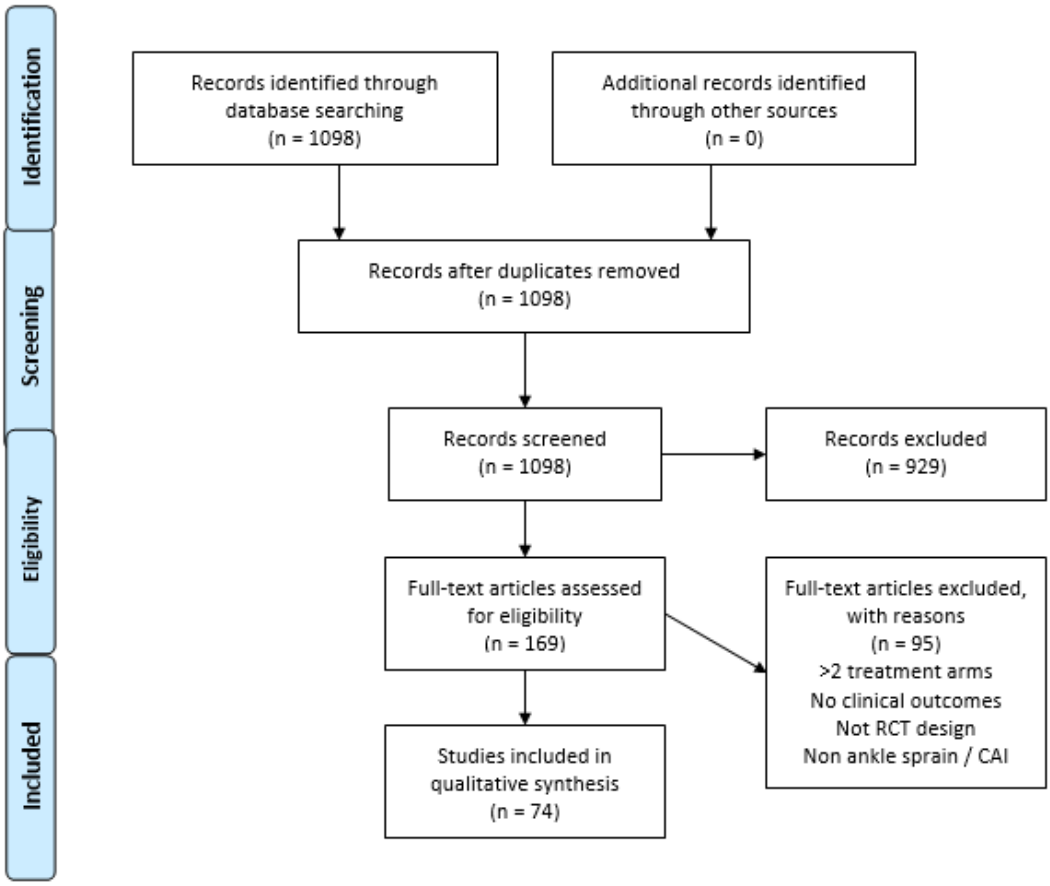
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1098)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 1098)
Records screened (n = 1098)
Records excluded (n = 929)
Full-text articles assessed for eligibility (n = 169)
Full-text articles excluded, with reasons (n = 95)
>2 treatment arms
No clinical outcomes
Not RCT design
Non ankle sprain / CAI
Studies included in qualitative synthesis (n = 74)

Figure 2

Area plots subgrouping *p*-values (n=966) by: level of significance (A), primary outcomes (B) and pre-registration (C)

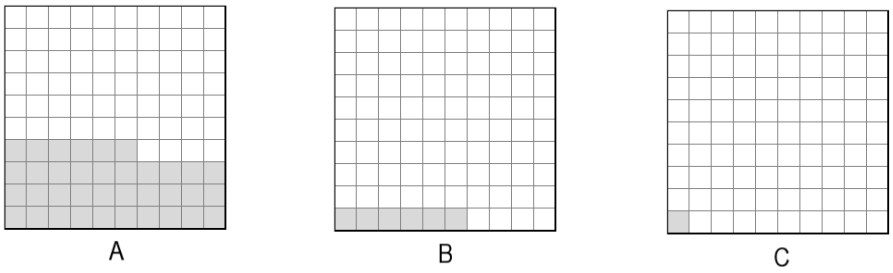

Figure 2 footnote

Each square represents ~10 *p*-values generated from between-group comparisons.

White squares = No statistical significance (p>0.05)

Shaded squares represent:

A). Statistically significant – primary or secondary outcomes

B). Statistically significant - primary outcomes only, any trial

C). Statistically significant - primary outcomes, pre-registered trials only

Figure 3

Violin plot summarizing False Positive Risk in trials reporting significant ($p<0.05$) effects in their primary outcome

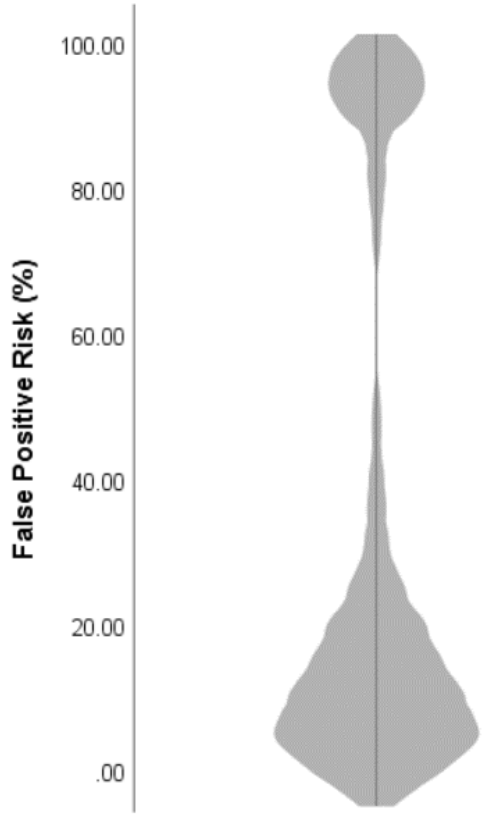

Figure 4

Venn diagram illustrating N trials meeting one than one FAIR criteria