

Many high quality RCTs in sports physical therapy are making false positive claims of treatment effect: a systematic survey

^{1,2} Bleakley Chris (0000-0001-9032-9691); ² Reijgers J; ² Smoliga James, M

1. School of Health Science, Faculty of Life and Health Science, Ulster University, Newtownabbey, Northern Ireland, BT370QB

2. Department of Physical Therapy, Congdon School of Health Science, High Point University, 1 University Parkway, High Point, NC 27260

Chris Bleakley associate professor Jonathan Reijgers spt James Smoliga professor

Correspondence address as above

Corresponding author: Chris Bleakley

Email c.bleakley@ulster.ac.uk

Word count: 2948

References: 32

Figures: 3

Key words: Null hypothesis significance testing, Randomised controlled trials, False Positive Risk, p-values

1 **Abstract**

2 Objective: To examine the risk of false positive reporting within high quality randomized
3 controlled trials (RCTs) in the sports physical therapy field.

4 Design: Cross-sectional

5 Methods: We searched the PEDro database for parallel design 2-arm RCTs reporting
6 positive treatment effects based on null hypothesis significance testing, and scoring >6/10
7 on the PEDro scale. No restrictions were made on pathology, intervention or outcome
8 variables. Sixty-two of 212 RCTs reported positive effects in at least one outcome
9 variable. We estimated False Positive Risk (FPR) using the FPR Web Calculator (version
10 1.5) based data on: *n* of participants, p-value, and effect size. For each study, FPR was
11 estimated using a range of prior probability assumptions: 0.2 (skeptical hypothesis), 0.5
12 and 0.8 (optimistic hypothesis).

13 Results: We calculated the FPR associated with 189 statistically significant findings
14 ($p < 0.05$) reported across 44 trials. The median FPR was 9% (25th-75th PCTL: 2-22%).
15 59% of statistically significant results (102/174) had FPR >5%, and 16% (28/174) had
16 FPR >50%. Changing the prior probability from skeptical to optimistic reduced the median
17 FPR from 30% (25th-75th PCTL: 9-54%) to 2% (25th-75th PCTL: 0.5-7%).

18 Conclusion: High quality RCTs using null hypothesis significance testing often
19 overestimated treatment effects. The median false positive risk (FPR) was 9% -- in one
20 in 10 trials, the researchers falsely concluded there was a treatment effect. Future RCTs
21 in sports physical therapy should be informed by pre study odds and a minimum FPR
22 estimation.

23

24 **Introduction**

25 High quality research can help clinicians and patients decide which treatments are likely
26 to be most effective.¹⁵ Successful replication of research findings is an integral part of the
27 scientific process, and represents a more robust evidence base for clinical decision
28 making. However, there is concern that the majority of published research claims are
29 false.¹⁷

30
31 In a survey of 1576 researchers, more than 70% had tried and failed to reproduce another
32 scientist's experiment, and more than half failed to reproduce their own experiments.¹ In
33 preclinical research, only 11 - 49% of research findings have been successfully replicated,
34 ¹⁰ with similar figures reported in psychological science.²⁷ Although evidence-based
35 practice should substantially improve the quality and cost of healthcare, serious concerns
36 regarding randomized controlled trial design and statistical analysis raise questions about
37 the validity of evidence-based interventions.

38
39 Experimental analysis in medicine is usually frequentist: conclusions informed by p
40 (probability) values generated from null hypothesis significance testing. However, many
41 researchers and clinicians are unable to define or accurately interpret p-values.⁵ Common
42 misconceptions are that a p-value represents 'the probability that the results occurred by
43 chance' or 'the probability that the null hypothesis (H0) is true'⁵ or 'the probability that the
44 hypothesis being tested is true.'²⁴ A p-value only represents the probability that the
45 obtained data, or more extreme values, could be obtained if H0 is true²⁴ – the probability

46 of the data, on the condition that the null hypothesis is true. For more help understanding
47 P values, see¹⁸

48

49 Misinterpreting the results of statistical tests makes it difficult to disentangle true from
50 false positive findings. Understanding and accurately applying appropriate statistics
51 defends against false discoveries.²⁴ Central, is quantifying the false positive risk (FPR) –
52 “the probability of observing a statistically significant p-value and declaring that an effect
53 is real, when it is not.”⁶ The FPR within different areas of biomedical science has been
54 conservatively estimated at 25%.²⁴ This means that in at least 1 in 4 studies, the
55 researchers falsely concluded a treatment effect. Others^{4, 5, 17} have used data simulations
56 to demonstrate experimental studies can carry a high FPR, even if their effect sizes are
57 large and/or p-values are less than commonly used thresholds such as $p < 0.01$.

58

59 The issue of irreproducible data has been discussed by scientists for decades.² However
60 it has received little attention in health care. No one has examined FPR using primary
61 data extracted from high-quality clinical experimental research. Given the criticism of a
62 weak evidence base for orthopedics and sports medicine,^{3, 14, 22, 26} our objective was to
63 estimate the false positive risk (FPR) of high-quality randomized controlled trials (RCTs)
64 in sports physical therapy. Our secondary objectives were to examine the relationship
65 between FPR and reported p-values by quantifying the number of studies with FPR >5%;
66 and to determine how FPR changed based on assumptions around the prior probability
67 of effect.

68

69 **Methods**

70 *Trial selection*

71 Trials were sourced from the Physiotherapy Evidence Base (PEDro), which is a freely
72 accessible database aiming to “guide users to trials that are more likely to be valid” and
73 “guide clinical practice.”¹⁹ In addition to serving as a database for clinical trials, PEDro
74 includes a 10-item scale quantifying study quality.^{14, 7}

75

76 We identified all RCTs scoring >6/10 and categorized in the subcategory of ‘sports’
77 (sports is defined by PEDro as “papers which specifically mention sports injuries as well
78 as conditions which commonly affect sports people (eg, ligament repairs).” Eligible RCTs
79 must have employed null hypothesis significance testing to determine evidence of effect
80 and a parallel group design. No restrictions were made on pathology, intervention type or
81 date of publication. We excluded RCTs with: healthy participants only; >2 intervention
82 groups; cross over, cluster or pilot study designs.

83

84 *Data extraction and management*

85 We extracted the following data from all eligible trials: population, number of participants,
86 primary diagnosis, intervention, comparison, outcome(s), allocation ratio, follow up time,
87 p-value, effect size, trial registration number, and a priori power calculation.

88

89 We subgrouped the trials as either 1). Positive: the attainment of a dichotomous threshold
90 of statistical significance ($p < 0.05$) in at least 1 outcome; or 2). Null: reporting no evidence
91 of effect ($p > 0.05$).

92

93 For all trials that reported evidence of effect (Positive studies), we extracted additional
94 data. First, we extracted details of between-group comparisons, making no restriction on
95 outcome construct or follow-up time. If there was a between-group comparison with a
96 positive statistically significant finding, we extracted the p-value, the number of
97 participants in each group, and when possible, we calculated the corresponding effect
98 size (Hedges g). If a trial reported a threshold of $p < 0.05$, rather than an exact p-value, we
99 assumed that the p-value was one decimal place below the threshold value (e.g. 0.049).

100

101 *Estimating the false positive risk*

102 We calculated FPR using the False Positive Risk Web Calculator (version 1.5)²³ For
103 further details of the analysis script and simulated examples of FPR calculations see ^{5, 6}.
104 Calculating FPR requires imputation of the prior probability that there is a real effect
105 [P(H1)] for a given treatment. In all trials, we initially assumed that P(H1), was 0.5 – that
106 there was a 50% probability a treatment intervention had a positive underlying effect
107 before the trial was conducted.^{4, 5}

108

109 We ran additional simulations based on extreme prior probabilities of $P(H1) = 0.2$, where
110 the chances of a positive effect are very small (a skeptical hypothesis), and $P(H1) = 0.8$
111 where chances of effect are almost certain (an optimistic hypothesis). We also applied a
112 reverse Bayesian approach:^{5, 25} using observed p-values to determine the prior probability
113 that would be required to achieve a FPR of 5%. In all cases FPR estimations were
114 calculated using the p-equals method,²³ which is the probability of observing a statistically

115 significant finding that is due to chance for a single result, rather than trying to estimate
116 the long term error rate (lifetime FPR).

117
118 We calculated FPR for primary and secondary outcomes where applicable. When trials
119 included multiple outcome measures but did not clearly specify a primary outcome, we
120 assigned a primary outcome based on the nature of the research question and the
121 following definition:²⁸ 'a specific key measurement(s) or observation(s) used to measure
122 the effect of experimental variables in a study. We examined the relationship between all
123 reported p-values and the corresponding FPR using descriptive statistics, scatter and
124 violin plots.

125

126 **Results**

127 There were 212 RCTs scoring >6/10 within the 'sport' subcategory on PEDro. Ninety trials
128 were excluded for the following reasons: not parallel design (2 group) randomized
129 controlled trial (n=56); healthy participants/no clinical outcomes (n=23); non-English
130 language (n=9); abstract/full text not available (n=2).

131

132 We included 122 RCTs; 49% (n=60/122) reported a null finding, and 51% (n=62/122)
133 reported positive effects from at least one outcome (Figure 1). Full trial details can be
134 found in the Supplemental data file. There were few differences between the subgroups
135 (positive vs null) in primary diagnoses and treatment interventions (Figure 1). The majority
136 of RCTs included participants with tendinopathy (n=47 studies), musculoskeletal pain

137 (n=19 studies) or ligament/joint problems (n=21 studies). Electro-physical agents (n=48),
138 rehabilitation (N=37) and manual therapy (n=17) were the most common interventions.

139

140 **Insert Figure 1**

141 **Diagnosis and Primary Treatment***

142

143 **False Positive Risk**

144 In trials reporting positive effects (n=62), 67% compared two different physiotherapeutic
145 approaches, and 33% used either sham or placebo controls. The mean sample size was
146 n=57.3 (SD=35.2; range 16-172). Twenty-nine percent of trials (18/62) were prospectively
147 registered; 64% (40/62) reported using *a priori* sample size calculation. The majority of
148 sample size estimations included alpha (Type 1 error) and beta (Type 2 error) levels of
149 5% and 20% respectively; and the anticipated *a priori* effect size used was 0.9 on average
150 (SD 0.4, range 0.2- 2.2).

151

152 We could not calculate FPR in 18 trials due to missing data. In the remaining 44 trials, we
153 calculated FPR associated with 189 between-group comparisons reported as statistically
154 significant. Lower p-values were associated with lower FPR (Figure 2). The mean FPR
155 (based on prior probability of 0.5) was 25.2% (SD 34.3). As the data were not normally
156 distributed, the median FPR of 9% is more representative of the data's central tendency
157 (25th-75th percentile: 2-24%). Sixty-three percent of reported p-values (119/189) were
158 associated with FPRs greater than 5%; 18% (35/189) had a FPR greater than 50%.

159

160 Using a reverse Bayesian approach, 57% (68/119) of statistically significant findings
161 (primary or secondary outcomes) would require prior probabilities greater than 0.8, if
162 FPRs of 5% were to be achieved. FPR patterns were similar when examining only primary
163 outcomes, with mean and median FPRs of 22.9% (SD 36.1) and 5% (25th-75th percentile:
164 1-22%) respectively.

165

166 **Insert Figure 2**

167 **P-value vs False Positive Risk**

168 [Data relate to 189 positive effects reported from high quality RCTs (n=44); FPR based
169 on a prior probability of 0.5; Dashed line = reference if p-value was equal to FPR.]

170

171 The lowest FPR occurred when the prior probability of effect was assumed as 0.8, with
172 median risk of 2% (25th-75th percentile: 0.6-7%) (Figure 3). False positive risk increased
173 when prior probabilities of 0.2 were assumed: median risk of 29% (25th-75th percentile: 9-
174 56%).

175

176 **Insert Figure 3**

177 **FPR based on 3 different prior probability levels [P(H1)=0.2, P(H1)=0.5; P(H1)=0.8]**

178 [In all calculations, data relate to 189 positive effects reported from high quality RCTs
179 (n=44)]

180

181 **Discussion**

182 We found that 63% of statistically significant findings ($p < 0.05$) in the sports physical
183 therapy literature generated FPRs greater than 5%. Repeated simulations of t-tests
184 suggest that if one uses $p = 0.05$ to conclude a discovery, one will be wrong at least 30%
185 of the time.⁴ False discoveries (claiming a treatment effect is real when it isn't) may be
186 minimized through better understanding of the FPR. This is the first time that the
187 healthcare literature has been audited to determine the FPR using primary data extracted
188 from higher quality clinical experimental research. The median FPR was 9% (25th-75th
189 percentile: 2-24%), suggesting that approximately one in every 10 trials in the sports
190 physical therapy field have falsely concluded a treatment effect.

191
192 There have been a range of proposals to help minimize unsubstantiated claims of
193 effectiveness in research. One option has been to lower p-values thresholds to $p \leq 0.001$,
194 to keep false discovery rates below 5%.⁴ Recently the American Society of Statisticians
195 released a number of recommendations aimed at improving use of null hypothesis
196 significance testing.³² The core objective of the American Society of Statisticians is to
197 progress research beyond 'all or nothing' hypothesis tests, which may be particularly
198 important if the theoretical predictions within a study are weak.³⁰

199
200 Clinical decisions *should not* be made solely on a p-value.³² Many of the positive
201 statistically significant conclusions from high-quality RCTs in sports physical therapy are
202 probably no more than suggestive. Researchers must also understand that null
203 hypothesis significance testing is only designed to work efficiently in the context of long-
204 run repeated testing (exact replication).³⁰ A single significant result should not be

205 concluded as a “scientific fact.” The result should be interpreted as something worthy of
206 further investigation,^{12, 31} particularly if it was derived from a secondary outcome.

207
208 There is no consensus on how best to communicate results of testing scientific
209 hypotheses. RCTs in orthopedics and sports medicine have traditionally used a
210 frequentist approach based on deductive inference. Our calculation of FPR involved
211 application of Bayes’ Theorem, where the central tenet is to consider how current data
212 alter our “prior probability”, to generate a new, “posterior probability.” We initially used a
213 “non-informative” prior probability of 50%, meaning that we assumed an even odds of
214 treatment effect. As we audited clinical studies from a diverse field, there may be
215 situations when hypotheses are more skeptical or optimistic. Therefore, we calculated
216 FPRs based on both low [$P(H1) = 0.2$] and high [$P(H1) = 0.8$] prior probabilities. As
217 expected, when prior probabilities were shifted closer to zero, the FPR was inflated; when
218 we assumed a high prior probability of effect, 75% of findings had FPRs <8%.

219
220 There continues to be debate around the relative merits of a frequentist and Bayesian
221 approach to statistical analysis. Our findings highlight how Bayesian thinking and
222 conditional probabilities can affect the interpretation of null hypothesis significance
223 testing.⁴ For example, a statistically significant finding generated from a RCT examining
224 the effects of jugular vein compression devices²⁹ on concussion incidence in contact
225 sports (skeptical prior) should be interpreted with more caution than a statistically
226 significant finding from a RCT testing the analgesic effects of topical cooling after a
227 musculoskeletal injury (optimistic prior). In effect, Bayesian logic ensures that the

228 skeptical prior example requires more 'extreme' data before treatment effectiveness can
229 be concluded. In contrast, the traditional frequentist approach, does not differentiate
230 between these two research questions.

231
232 A key limitation of Bayes' Theorem is the uncertainty when determining what a suitable
233 prior probability should be. One solution is a reverse Bayesian approach,²⁵ where the
234 observed p-value is used to calculate the prior probability required to achieve a specific
235 or minimal false positive risk (eg. 5%). This approach allows the researcher to determine
236 whether the calculated prior probability is plausible or not. It has been suggested that 0.5
237 (or a 50:50 chance of success) might be the largest prior probability that can be
238 legitimately assumed.⁵ In our analysis, approximately 60% of positive (statistically
239 significant, $p < 0.05$) outcomes would require prior probabilities greater than 0.8 to achieve
240 FPRs of 5%. Such extreme prior probabilities are likely unacceptable as they represent
241 situations where a researcher is almost certain of treatment success (a non-zero effect),
242 before the experiment is even initiated.

243
244 Trials with positive outcomes are published more often, and more quickly, than trials with
245 negative findings.¹⁶ The proportion of positive results in published scientific literature may
246 be as high as 86%.⁹ In our analysis of high-quality RCTs within sports physical therapy,
247 we found an equal ratio of trials reporting positive and null effects. Although this might
248 suggest that publication bias is not an issue within the sports physical therapy field, there
249 were no trials reporting negative or harmful effects of an intervention. There may also be
250 publication bias in lower quality studies, which we excluded. Trial registration is

251 considered an effective way to control publication bias,²⁰ and can help to prevent cherry-
252 picking statistically significant results later. We found that only 29% of sports physical
253 therapy trials were prospectively registered. It is important that this figure eventually
254 increases to 100%. A broader and more complex challenge is that often, many trials have
255 discord between the original registry data and the published data, despite registration.¹¹
256 Additional solutions have been proposed including: improved CONSORT compliance,
257 from both researchers and editorial boards, and improvement to the post-publication peer
258 review process. ¹¹

259
260 The evidence base for orthopaedics and sports medicine has been criticized for
261 inappropriate participant selection³ and high risk of bias.²² Issues related to undefined
262 primary endpoints and multiple comparisons have plagued the literature,²² but their
263 relevance has been difficult to quantify. Our results suggest that methodological
264 shortcomings may be leading researchers in orthopaedics, sports medicine and sports
265 physical therapy astray in their conclusions, and negatively influencing evidence-based
266 practice.

267

268 **Limitations**

269 A recent audit of the PEDro database (The Physiotherapy Evidence Database (PEDro;
270 <http://www.pedro.org.au>)) listed over 23 049 RCTs, of which 1098 have been undertaken
271 in sports-related disciplines.¹⁹ We limited inclusion to RCTs archived within the PEDro
272 database and used a cut off of >6/10 (on the PEDro scale) to define high quality. Our
273 audit was limited to results from single experiments and we did not fully consider false

274 discoveries relating to other important sources such as the use of multiple treatment arms,
275 analysis of multiple outcomes, and multiple analyses of the same outcome at different
276 times.²¹ FPR is likely to increase if lower quality methodological designs are employed,⁵
277 therefore our FPR estimations are likely conservative in the broader context of all clinical
278 trials. We did not focus on false negative findings or outcomes deemed to be surrogate
279 in nature (e.g. biomarkers). We acknowledge the importance of directing future work in
280 this area; our primary focus was on the risk of false positive findings regarding outcomes
281 that reflect real-clinical settings.

282

283 **Recommendations for future research**

284 Future reports should include exact figures for p-values rather than thresholds ($p < 0.05$)
285 and avoid using the term significant.⁴ We were often unable to calculate FPR due to
286 missing data. It is essential that researchers accompany reported p-values with effect
287 sizes, corresponding confidence intervals, and ideally a minimum false positive risk
288 estimation. It is important that there is a continued focus on the mandatory preregistration
289 of study protocols, publication of pre-study power calculations and effect sizes, including
290 any negative findings.

291

292 While the proper use of statistics will help to minimize false discoveries in research, there
293 are other factors currently influencing the risk of erroneous findings in the sports
294 physiotherapy field. It is possible that the existing academic system in sports physical
295 therapy (like many other areas of healthcare) might increase the risk of erroneous or
296 selective publishing, because career milestones such as promotion or tenure are often

297 determined by the volume of researchers' publication record.¹³ Journal editors, reviewers
298 and grant-review committees may also favor scientific findings that are confirmatory, clear
299 and complete² — limiting the chances of disseminating negative or contradictory research
300 findings. We encourage researchers to examine FPR in other disciplines of health care.

301
302 To calculate FPR, we used an online calculator that uses post-hoc statistical power to
303 inform FPR values. It is possible that some studies recorded very large effect sizes due
304 to sampling variation, which consequently overestimates statistical power (a posteriori)
305 and potentially inflates the FPR estimate. Future research could include additional FPR
306 estimations using a range of statistical power parameters (partially post hoc power).⁸

307
308 **Conclusion**
309 Research conclusions should not be based solely on Null Hypothesis Significance Testing
310 (NHST) and p-values. Over 60% of statistically significant findings ($p < 0.05$) reported in
311 the physiotherapy literature, carried FPRs greater than 5% and the median FPR was 9%
312 (assuming a prior probability of 0.5).

313
314 **Ethics approval and consent to participate:** Not applicable
315 **Consent for publication:** Not applicable
316 **Availability of data and material:** Data sharing is not applicable to this article as no
317 datasets were generated or analyzed during the current study
318 **Competing interests:** The authors declare that they have no competing interests"
319 **Funding:** Nil

320 **Authors' contributions:** All authors certify that they have participated sufficiently in the
321 work to take public responsibility for the content, including participation in the concept,
322 design, analysis, writing, or revision of the manuscript. CB and JS were involved in the
323 concept, design and writing. CB and JR were involved in the analysis. All authors were
324 involved in final submission and revision of the manuscript.

325 **Acknowledgements:** Not applicable

326

327

328

329 **Key points**

330 **Findings**

331 Many of the positive statistically significant conclusions from high-quality RCTs in sports
332 physiotherapy are probably no more than suggestive. We estimate the median false
333 positive risk (FPR) in this field to be 9% (25th-75th percentile: 2-24%).

334 **Implications**

335 Research conclusions should not be based solely on Null Hypothesis Significance Testing
336 (NHST) and p-values. The risk of making a false claim of treatment effectiveness can be
337 reduced through, more rigorous consideration of pre study odds (ie. the chances that a
338 treatment will work a priori) and reporting of FPR (a posteriori).

339 **Cautions**

340 This audit was limited to high quality, 2-arm RCTs. We also did not consider other sources
341 of false discoveries in research such as: the use of multiple treatment arms, analysis of
342 multiple outcomes, and multiple analyses of the same outcome at different time points.

343 **References**

- 344 1. Baker M. Reproducibility crisis: Blame it on the antibodies. *Nature*. 2015;521(7552):274-276.
- 345 2. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*.
346 2012;483(7391):531-533.
- 347 3. Bleakley C, MacAuley D, McDonough S. Are sports medicine journals relevant and applicable to
348 practitioners and athletes? *Br J Sports Med*. 2004;38(5):E23.
- 349 4. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values.
350 *R Soc Open Sci*. 2014;1(3):140216.
- 351 5. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open*
352 *Sci*. 2017;4(12):171085.
- 353 6. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About *p*-Values. *The*
354 *American Statistician*. 2019;73:192-201.
- 355 7. de Morton NA. The PEDro scale is a valid measure of the methodological quality of clinical trials:
356 a demographic study. *Aust J Physiother*. 2009;55(2):129-133.
- 357 8. Dziak J, Dierker L, Abar B. The interpretation of statistical power after the data have been
358 gathered. *Current Psychology*. 2019.
- 359 9. Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*.
360 2012;90(3):891-904.
- 361 10. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research.
362 *PLoS Biol*. 2015;13(6):e1002165.
- 363 11. Goldacre B, Drysdale H, Dale A, et al. COMPare: a prospective cohort study correcting and
364 monitoring 58 misreported trials in real time. *Trials*. 2019;20(1):118.
- 365 12. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45(3):135-140.
- 366 13. Grimes DR, Bauch CT, Ioannidis JPA. Modelling science trustworthiness under publish or perish
367 pressure. *R Soc Open Sci*. 2018;5(1):171511.
- 368 14. Harris JD, Cvetanovich G, Erickson BJ, et al. Current status of evidence-based sports medicine.
369 *Arthroscopy*. 2014;30(3):362-371.
- 370 15. Heneghan C, Goldacre B, Mahtani KR. Why clinical trial outcomes fail to translate into benefits
371 for patients. *Trials*. 2017;18(1):122.
- 372 16. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to
373 statistical significance or direction of trial results. *Cochrane Database Syst Rev*.
374 2009(1):MR000006.
- 375 17. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
- 376 18. Kamper SJ. Interpreting Outcomes 2-Statistical Significance and Clinical Meaningfulness: Linking
377 Evidence to Practice. *J Orthop Sports Phys Ther*. 2019;49(7):559-560.
- 378 19. Kamper SJ, Moseley AM, Herbert RD, Maher CG, Elkins MR, Sherrington C. 15 years of tracking
379 physiotherapy evidence on PEDro, where are we now? *Br J Sports Med*. 2015;49(14):907-909.
- 380 20. Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration: looking back and moving ahead.
381 *Lancet*. 2007;369(9577):1909-1911.
- 382 21. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials:
383 the what, why, when and how. *Int J Epidemiol*. 2017;46(2):746-755.
- 384 22. Lohmander LS, Roos EM. The evidence base for orthopaedics and sports medicine. *BMJ*.
385 2015;350:g7835.
- 386 23. Longstaff C, Colquhoun D. <http://fpr-calc.ucl.ac.uk/>. Accessed 01-02-2019.
- 387 24. Marino MJ. How often should we expect to be wrong? Statistical power, P values, and the
388 expected prevalence of false discoveries. *Biochem Pharmacol*. 2018;151:226-233.

- 389 25. Matthews R. Why should clinicians care about Bayesian methods? *J Stat Plan Inference*.
390 2001;94:43-58.
- 391 26. Moseley AM, Elkins MR, Janer-Duncan L, Hush JM. The Quality of Reports of Randomized
392 Controlled Trials Varies between Subdisciplines of Physiotherapy. *Physiother Can*.
393 2014;66(1):36-43.
- 394 27. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological
395 science. *Science*. 2015;349(6251):aac4716.
- 396 28. Ramagopalan S, Skingsley AP, Handunnetthi L, et al. Prevalence of primary outcome changes in
397 clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res*. 2014;3:77.
- 398 29. Smoliga JM, Wang L. Woodpeckers don't play football: implications for novel brain protection
399 devices using mild jugular compression. *Br J Sports Med*. 2018.
- 400 30. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A
401 Reassessment. *Front Hum Neurosci*. 2017;11:390.
- 402 31. Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of
403 near significant P value becoming more significant with extra data. *BMJ*. 2014;348:g2215.
- 404 32. Yaddanapudi LN. The American Statistical Association statement on P-values explained. *J*
405 *Anaesthesiol Clin Pharmacol*. 2016;32(4):421-423.

406

407

408

409

410