The Bias for Statistical Significance in Sport and Exercise Medicine

Preprint (not peer reviewed)

David N Borg[1], Adrian G Barnett[1], Aaron R Caldwell[2], Nicole M White[1], Ian B Stewart[3]

[1]Australian Center for Health Services Innovation and Center for Healthcare Transformation, School of Public Health and Social Work, Queensland University of Technology, Brisbane, Queensland, Australia.
[2]Natick, MA, United States of America.
[3]School of Exercise and Nutrition Sciences, Queensland University of Technology, Brisbane, Queensland, Australia.

Please cite as: Borg DN, Barnett AG, Caldwell AR, White NM, Stewart IB. (2022). The Bias for Statistical Significance in Sport and Exercise Medicine. *OSF Preprints*. doi:osf.io/t7yfc.

Corresponding Author
Dr David N Borg
Queensland University of Technology, Faculty of Health, School of Public Health and Social Work, Australian Center for Health Services Innovation and Center for Healthcare Transformation, Brisbane, Queensland, Australia.
Email: dn.borg@qut.edu.au

Data and Code Availability
The datasets and R code used in this article are available at https://github.com/SciBorgo/sports-med-intervals. The code was adapted from https://github.com/agbarnett/intervals.

Acknowledgements
The authors would like to thank Dr Lucas BR Orssatto and Dr Robert Buhmann for their valued feedback.

Abstract
Aim: We aimed to examine the bias for statistical significance using published confidence intervals in sport and exercise medicine research. **Method:** The abstracts of 48,390 articles, published in 18 sports and exercise medicine journals between 2002 and 2022, were searched using a validated text-mining algorithm that identified and extracted ratio confidence intervals (i.e., odds ratios, hazard ratios and risk ratios). The text-mining algorithm identified 1,744 abstracts that included ratio confidence intervals, from which 4,484 intervals were extracted. After excluding ineligible intervals, the analysis used 3,819 intervals, reported as 95%

confidence intervals, from 1,599 articles. The cumulative distributions of lower and upper confidence limits were plotted to identify any abnormal patterns, particularly around a ratio of 1 (the null hypothesis for a ratio). The distributions were compared to those generated from unbiased reference data, which was not subjected to *p*-hacking or publication bias. Bias was also investigated using a histogram of *z*-values calculated from the intervals. **Results:** There was a marked change in the cumulative distribution of both lower and upper bound intervals just over (lower) and just under (upper) a ratio of 1. Twenty-five percent of lower bound intervals were between 1 and 1.2, which was higher than the 15% observed in the unbiased reference dataset. Sixteen percent of upper bound intervals were between a ratio of 0.9 and 1, which was over four times higher than the unbiased reference dataset. The excess of statistically significant results was also highlighted by the striking absence of *z*-values between –1.96 and +1.96, corresponding to *p*-values above 0.05. **Conclusion:** There was an excess of published research with statistically significant results just below the standard significance threshold of 0.05, which is indicative of publication bias. Transparent research practices, in particular the use of registered reports, are needed to reduce the bias in published sport and exercise medicine research. Researchers and peer reviewers need to direct their focus away from only statistically significant results when evaluating the suitability of manuscripts for publication.

Key words

Confidence interval, estimation, misconduct, *p*-hacking, publication bias, ratio, registered report, selective reporting

Introduction

Every sport and exercise medicine researcher should be aware that a statistically significant result is more likely to get published [1–3]. The selective publishing of statistically significant results has encouraged poor practices, including *p*-hacking, the generation of post hoc hypotheses, and data fabrication [4–7]. The bias towards the publication of significant results has also distorted evidence for scientific claims, with many null findings never making it to publication [8]. In exercise medicine, the focus on statistical significance has been shown to bias a researcher's perceptions and decision making during a study [1]. For example, the decision to collect more data when a result does not reach the specified significance threshold–usually a *p*-value of 0.05. In defense of researchers, significance-seeking behaviors may not always be overt, and can occur despite seemingly reasonable decisions being made [4,9].

Bias around statistical significance is often examined using *p*-curves [6]. A *p*-curve is a plot of the distribution of reported *p*-values that fall below a chosen threshold for defining statistical significance, most commonly 0.05. A left-skewed *p*-curve would indicate an excess of *p*-values that fall just below the chosen threshold, which is statistically implausible and indicates the presence of publication bias.[1]

There have been recent calls for researchers to replace *p*-values with confidence intervals in order to reduce the bias promoted by the overuse of *p*-values [10,11]. However, there is no empirical evidence that emphasizing confidence intervals over *p*-values reduces *p*-hacking and publication bias [12–14].

No previous study has used confidence intervals to examine bias regarding statistical significance in the sport and exercise medicine literature. We aimed to assess the presence of bias around the statistical significance threshold using ratio confidence intervals (i.e., odds ratios, hazard ratios, risk ratios) as these can be accurately extracted from published papers by automated tools. We hypothesized that there would be a marked change in the cumulative distribution of upper and lower bound intervals near a ratio of 1, which is the null hypothesis of no difference on a ratio scale.

Methods

We used a validated text-mining algorithm [12,15] to extract confidence intervals (see “Ratio confidence intervals” box) from the abstracts of articles published in 18 sports and exercise medicine journals between 2002 and 2022, that are indexed in *MEDLINE* (Table 1). No ethical approval for the study was needed as we used publicly available data that is published to be read and scrutinized.

[1] In the absence of *p*-hacking and publication bias, all *p*-values below the commonly used significance threshold of 0.05 would be equally likely, rather than an excess of values just inside the threshold (e.g., 0.04). The shape of *p*-values will also depend on whether the null is true or not.

Ratio confidence intervals

Most confidence intervals are given as 95% intervals, which corresponds to a *p*-value threshold of 0.05. As a reminder, a 95% confidence interval is a range that should contain the true value on 95% of occasions if the data generating process could be repeated many times [12].

We extracted confidence intervals from three types of ratios: odds ratios (OR), hazard ratios and risk ratios. Irrespective of the type of ratio, a value of 1 indicates the null hypothesis [19].

Considering odds ratios, these can be used to compare the relative odds of the occurrence of an event of interest (e.g., sustaining an injury), given exposure to a treatment of interest (e.g., injury prevention exercises) [19], with values interpreted as:

OR = 1, the null hypothesis, that is, performing the injury prevention exercises is not associated with being injured;
OR < 1, performing the injury prevention exercises is associated with lower odds of being injured; and
OR > 1, performing the injury prevention exercises is associated with higher odds of being injured.

In practice, the 95% confidence interval is often used as a proxy for statistical significance if the interval does not include the null hypothesis value of 1 [19].

Below are two examples from our dataset of how ORs are used in practice.

Example 1: The authors were interested in the association of body mass index with the risk of developing hypertension. Risk of hypertension was a categorical variable with two levels, no risk and risk. The authors found that "*...the association of BMI was greatly attenuated (OR = 1.04 [95% CI, 0.99–1.09]) when fitness also was included in the model*" (PubMed ID 17909393). The 95% confidence interval spanned OR values from 0.99 to 1.09, therefore, including the null hypothesis of 1. The *p*-value reported for this interval was 0.1.

Example 2: A study described long-term outcomes of neurogenic bowel dysfunction in adults with pediatric-onset spinal cord injury. The use of colostomy was an outcome of interest, with two levels (not used and used). The authors found that "*...over time, the likelihood of using colostomy (OR = 1.071; 95% CI, 1.001–1.147) increased*" (PubMed ID 27473299). The 95% confidence interval spanned OR values from 1.001 to 1.107, therefore, excluding the null hypothesis of 1. The *p*-value reported for this interval was 0.047.

Eighteen journals were selected from a list of the top 100 journals in the subject area of Physical Therapy, Sports Therapy and Rehabilitation on Scimago [16]. We chose any journal that included the word 'medicine' in the name and appeared in *MEDLINE*. The extraction was restricted to original articles and reviews. Our focus was on journals that appeared in *MEDLINE* over the past two decades, but to increase the sample size we also included three journals that appeared after 2002 and continued to 2022. These three journals were: Research in Sports Medicine (appears from 2005 onwards), Sports Medicine and Arthroscopy Review (2006 onwards) and European Journal of Physical and Rehabilitation Medicine (2008 onwards).

The text-mining algorithm was designed to recognise regular expressions that authors use to report statistical ratios. For example, "OR = 0.42, 95% CI = 0.16–1.13", where 'OR' is the odds ratio and 'CI' the confidence interval. The text-mining algorithm has previously been used to extract ratio confidence intervals to identify reporting errors [15] and to investigate bias in ratio confidence intervals in the medical literature [12]. In the current study, the text-mining algorithm was highly accurate with a true positive percentage of 99%, in 100 abstracts, sampled at random. In the one missed observation, it was unclear whether the reported interval was an interquartile range or a confidence interval.

Confidence intervals were excluded from the analysis when: there was a boundary violation, that is, when the ratio point estimator was outside the confidence interval; the lower bound was below zero, which is not possible for ratios; and when the level of confidence interval was not reported.

Data Analysis

Graphical summaries were used to examine the presence of bias in the distribution of intervals, particularly around the significance threshold, that is, a ratio of 1 (see "Ratio confidence intervals" box). The cumulative distributions of lower and upper bounds for all confidence intervals were plotted to highlight changes without the need for smoothing. We also calculated the percentage of lower bound intervals that were just above a ratio of 1 (i.e., within +0.1 and +0.2 of 1) and the percentage of upper bound intervals that were just below a ratio of 1 (i.e., within –0.1 and –0.2 of 1).

For comparison, we plotted the cumulative distributions alongside those generated from an unbiased reference dataset [17]. The unbiased dataset contains thousands of analyses not subjected to *p*-hacking or publication bias, and therefore, provides a reference for the shape of the distributions if all study results were published and no bias was present [12]. To compare our results to the field of medicine, we also plotted the cumulative distributions of the extracted intervals against the results (abstracts only) published by Barnett & Wren [12].

We plotted the distributions in 5-year blocks to investigate whether there was any change in the cumulative distributions of lower and upper intervals over time. We used 5-year blocks because the sample size was insufficient to generate cumulative distributions for each year.

A bias for statistical significance was further investigated using a histogram plot of *z*-values calculated from the extracted 95% confidence intervals. In theory, *z*-values should follow a standard Normal distribution, with a mean of 0 and a standard deviation of 1. *Z*-values outside the range of –1.96 and +1.96 correspond to two-tailed *p*-values less than 0.05. In the absence of bias, we would expect the extracted *z*-values to approximately follow a Normal distribution. For each confidence interval a *z*-value was calculated using the equation: $z = \log(\text{mu})/\text{se}$, where 'mu' is the mean estimate and 'se' is the standard error [18].

All analyses were undertaken in R [18]. The datasets and R code used to produce our results are available at https://github.com/SciBorgo/sports-med-intervals. The code was adapted from https://github.com/agbarnett/intervals.

Results

Abstracts from 48,390 unique articles, published in 18 sports and exercise medicine journals between 2002 and 2022, were searched for ratio confidence interval pairs. The text-mining algorithm identified 1,744 unique abstracts from 16 of the 18 journals that included ratio confidence intervals, from which 4,484 intervals were extracted. Table 1 provides a list of the journals searched and the number of intervals extracted from these journals.

We removed interval pairs due to a boundary violation (n=104; 2.3%), a negative lower bound (n=14; 0.3%), or a missing level of confidence (n=508; 11.3%), leaving 3,858 interval pairs. In terms of missing data, the percentage of intervals missing the level of confidence decreased over time (Supplement 1 Panel A) and was as high as 26.3% for one journal (Supplement 1 Panel B). Five journals had over 20% intervals missing the level of confidence interval. When the level of confidence was provided, most intervals were given as 95% confidence intervals (n=3819/3858; 99%), with 90% (n=30/3858; 0.8%) and 99% (n=9/3858; 0.2%) intervals also reported.

Focusing on 95% confidence intervals, 3,819 interval pairs were extracted from 1,599 articles. The cumulative distribution of these 3,819 intervals showed that there was an excess of statistically significant results, with a clear inflection point in the distribution of lower bounds just over a ratio of 1, and to a lesser extent, upper bounds just below 1 (Figure 1; Table 2). This distinct distributional pattern was very similar to that observed in medical research (Figure 1). The excess of statistically significant results has changed little over time (Figure 2).

The excess of statistically significant results was clearly highlighted by the marked under-representation of *z*-values between –1.96 and +1.96, corresponding to *p*-values greater than 0.05, which is the commonly used significance threshold (Figure 3). Figure 3 clearly shows the enormous absence of published null results, and the distribution would be smoother and more like a standard Normal distribution if there was no bias.

Table 1. List of journals searched, and the number of articles and intervals extracted from these journals between 2002 and 2022.

Journal	Articles with ratio estimates (n=1,744)	Intervals extracted (n=4,484)
American Journal of Sports Medicine	375 (21.5%)	1048 (23.4%)
Archives of Physical Medicine and Rehabilitation	263 (15.1%)	690 (15.4%)
British Journal of Sports Medicine	269 (15.4%)	660 (14.7%)
Medicine and Science in Sports and Exercise	178 (10.2%)	464 (10.3%)
Journal of Science and Medicine in Sport	138 (7.9%)	345 (7.7%)
Scandinavian Journal of Medicine and Science in Sports	94 (5.4%)	235 (5.2%)
Clinical Journal of Sport Medicine	89 (5.1%)	228 (5.1%)
Journal of Rehabilitation Medicine	74 (4.2%)	184 (4.1%)
American Journal of Physical Medicine and Rehabilitation	63 (3.6%)	153 (3.4%)
Sports Medicine	38 (2.2%)	123 (2.7%)
Journal of Sports Medicine and Physical Fitness	37 (2.1%)	90 (2.0%)
International Journal of Sports Medicine	40 (2.3%)	84 (1.9%)
Physician and Sportsmedicine	34 (1.9%)	74 (1.7%)
European Journal of Physical and Rehabilitation Medicine	24 (1.4%)	40 (0.9%)
Journal of Sports Science and Medicine	16 (0.9%)	39 (0.9%)
Research in Sports Medicine	12 (0.7%)	27 (0.6%)

Note. Five journals ranked in the top 100 in the subject area of Physical Therapy, Sports Therapy and Rehabilitation were not searched because they published editorial-style or narrative review articles (i.e., Physical Medicine and Rehabilitation Clinics of North America) or featured in *MEDLINE* toward the end of the studied period (i.e., Science and Medicine in Football appears from 2020 onwards; BMJ Open Sport and Exercise Medicine appears from 2015 onwards; Sports Medicine and Health Science appears from 2019 onwards; and Sports Medicine Open appears from 2015 onwards). We searched Sports Medicine and Arthroscopy Review and Clinics in Sports Medicine. However, these journals contained no papers with ratio intervals reported in the abstract over the study period.

Table 2. The percentage of 95% confidence interval lower bounds just above and below a ratio of 1, in the sports and exercise medicine (i.e., the current study), an unbiased reference dataset and in medicine.

	Sport and exercise medicine 2002–2022 (n=3,819)	Unbiased dataset (n=279,876) †	Medicine 1976–2019 (n=968,289) ‡
Lower bound intervals			
Ratio >1.0 and <1.1	16.2%	8.7%	16.8%
Ratio >1.0 and <1.2	25.3%	14.9%	26.4%
Upper bound intervals			
Ratio >0.9 and <1.0	10.0%	2.5%	7.1%
Ratio >0.8 and <1.0	16.0%	3.5%	11.8%

Note. A ratio of 1 is the null hypothesis.

† The unbiased dataset included thousands of analyses not subjected to *p*-hacking or publication bias, was taken from [17].

‡ Data from the field of medicine were taken from [12].

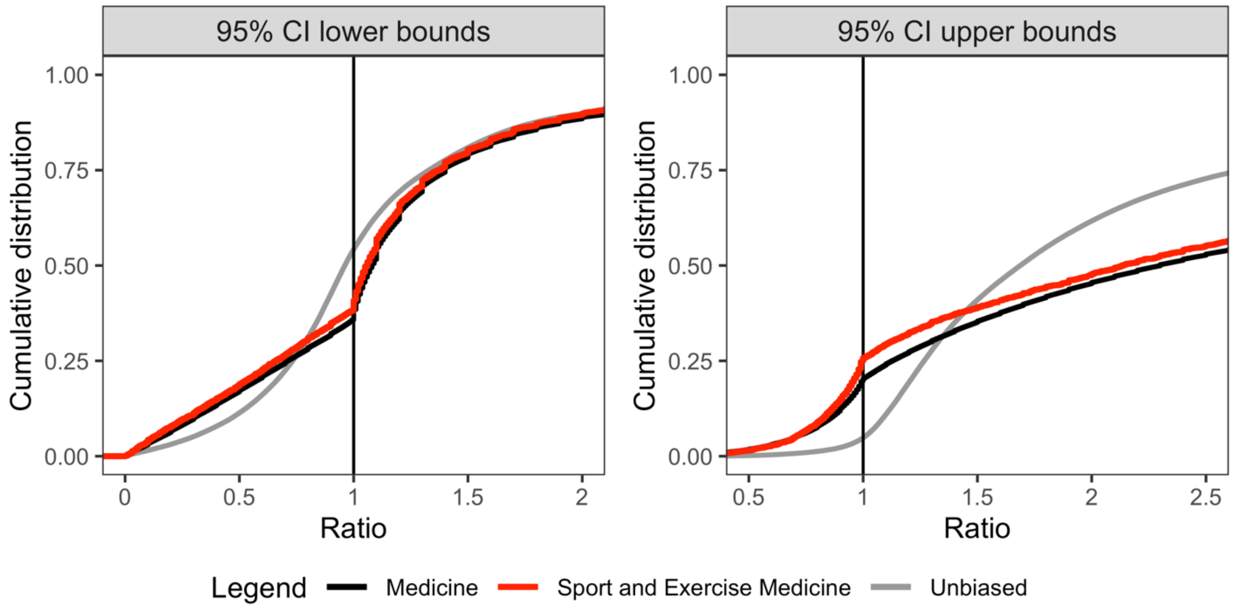


Figure 1. Empirical cumulative distributions for ratio confidence intervals from the abstracts of articles published in sports and exercise medicine journals between 2002 and 2022 (red), the abstracts of articles published in medical journals between 1976 and 2019 (black), and from an unbiased reference dataset (grey). Lower bounds are shown on the left panel and upper bounds on the right panel. To be statistically significant, lower intervals need to be above 1, and upper intervals need to be below 1. The x-axes are restricted to focus on changes around the significance threshold of 1 (vertical line). Note the marked change in the distribution of intervals from sports and exercise medicine around a ratio of 1, which is not present in the distribution from the unbiased dataset. The marked change around a ratio of 1 was also evident for intervals from medicine.

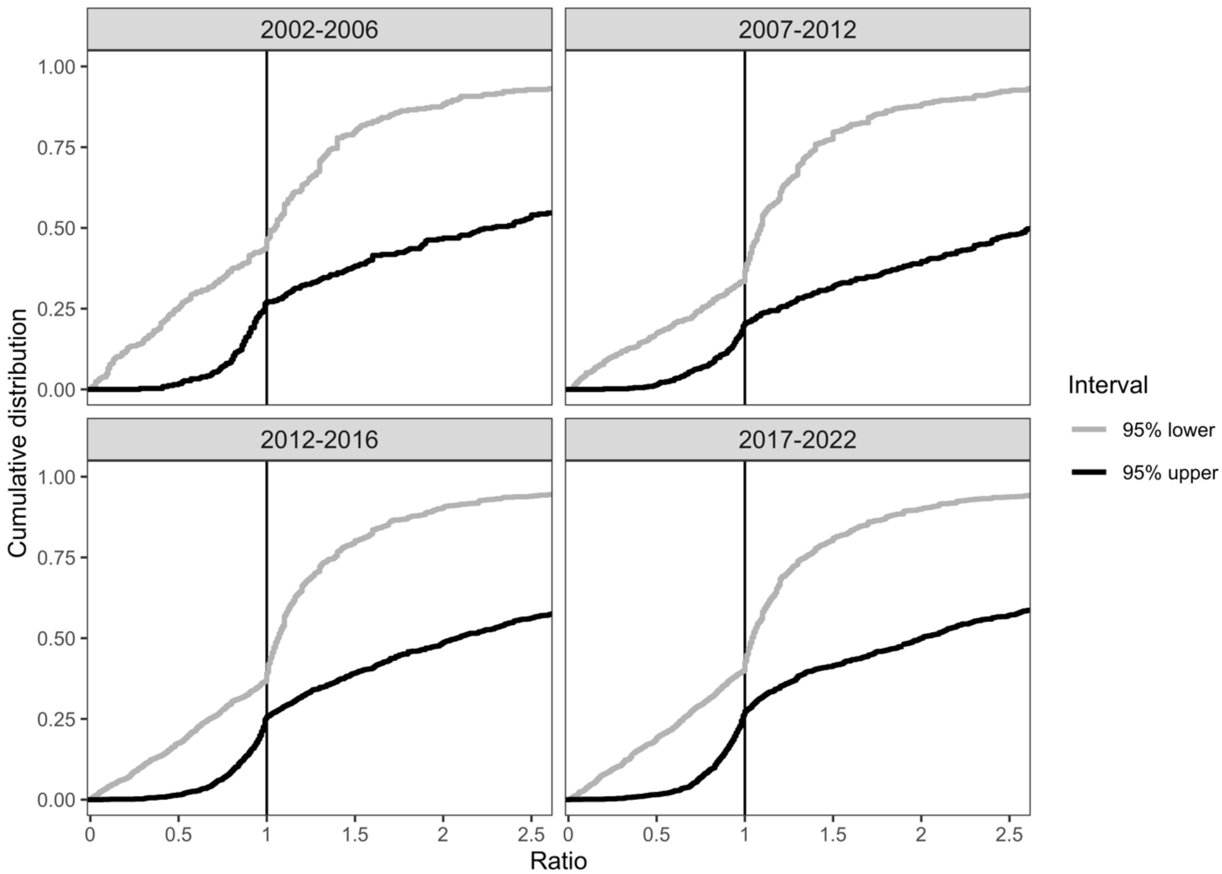


Figure 2. Empirical cumulative distributions in 5-year blocks for ratio confidence intervals from the abstracts of articles published in Sports and Exercise Medicine journals between 2002 and 2022. Lower bounds are shown in gray and upper bounds in black. To be statistically significant, lower intervals need to be above 1, and upper intervals need to be below 1. The x-axes are restricted to focus on changes around the significance threshold of 1 (vertical line). Note that the distributions become smoother across the panels due to the number of intervals published in those years and decimal place reporting.

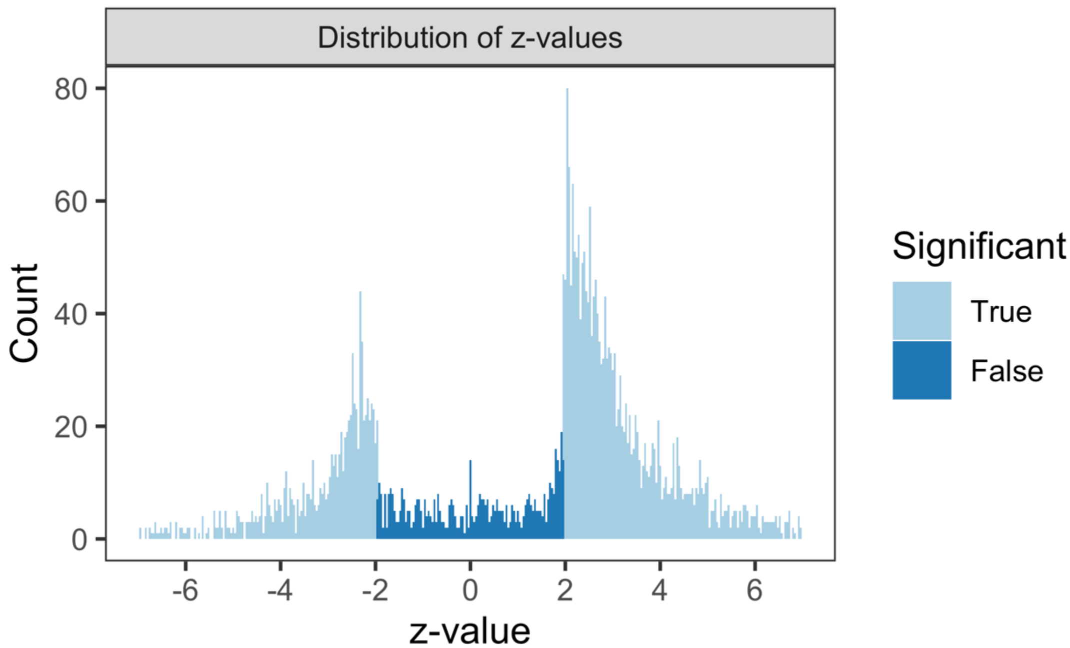


Figure 3. The distribution of z-values from 3,819 intervals. There was an under-representation of *z*-values between –1.96 and 1.96, corresponding to a *p*-value of 0.05, which is the commonly used significance threshold. The absence of published null results is striking. In the absence of bias, the distribution would be expected to be smoother and more like a standard Normal distribution. Note, histograms group data into "bins" of equal width to create a distribution impression of continuous data. A user is required to specify a bin width, which depending on the choice, can create different impressions of the same data. We generated a high-resolution histogram using the bin width of 0.04, which provides a fair impression for our context.

Discussion

We used a validated text-mining algorithm to extract over 4,000 ratio confidence intervals from nearly 1,700 sport and exercise medicine articles between 2002 and 2022. We plotted the cumulative distribution of lower and upper 95% confidence interval bounds to identify whether there were any abnormal changes in the distributions around the null hypothesis ratio of 1, which could be indicative of bias. As expected, there was a large excess of published research with statistically significant results, just below the standard significance threshold of 0.05. This excess of results just below the significance threshold would not occur if published results were completely unbiased. Transparent research practices are needed to reduce the bias in published sport and exercise medicine research. This includes the use of registered reports [20], ending the practice of continuing data collection until reaching significance, and the sharing of data and code. There is a pressing need for peer reviewers, editors, and journals to direct the rewards of publication away from statistical significance and onto scientific rigor.

Despite a smaller sample size, our findings in sports and exercise research are consistent with observations in medical research, where a large excess of lower and upper bound intervals around a ratio of 1 has been reported [12]. We observed an abnormal change in the direction of the cumulative distribution around a ratio of 1, which is unlikely to occur in the absence of bias (Figure 1). We found that a quarter (25.3%) of lower bound intervals were between a ratio of 1 and 1.2, which was similar to medicine [12] but was much higher than the unbiased reference dataset (14.9%). Alarmingly, the percentage of upper bound intervals just below a ratio of 1 was higher than in medical research, and four times higher than the unbiased reference dataset (Table 2). Similarities of the bias in confidence intervals between medicine and sport and exercise medicine is further supported by the highly unusual distribution of *z*-values, characterized by a stark absence of non-significant *z*-values (Figure 3), which was also observed in medicine (see Figure 1 in [21]).

Only focusing on statistically significant results is harmful for new discoveries because it distorts the literature by emphasizing an arbitrary threshold rather than rigor. Significant results with a small *p*-value are often mistakenly viewed as valid, reliable and meaningful [22], yet the exclusive focus on significance can lead to an overestimation in the magnitude of an effect [21,23]. The bias in the magnitude of an effect decreases as a function of a study's sample size, which is worrying in the field of sport and exercise medicine, as sample sizes are often small, and therefore, bias is likely to be large [24]. The overestimation of effects and subsequent distortion of evidence for scientific claims can lead to wasted resources, as researchers direct their attention toward unworthy areas, for which there is little evidence [4,8]. Worse, unproven, or ineffective treatments may be promoted which can directly harm the public and lower trust in scientific institutions.

If researchers focused on estimation, rather than significance, the overestimation of effects could be reduced [25]. This would require researchers to think more carefully about their analysis and interpretation [26]. Recently, there has been advocacy for adopting an unconditional interpretation of statistical results [25,27]. This approach would involve focusing on the estimation of effects rather than statistical significance and focusing on the uncertainty around the estimated effect (e.g., the confidence interval width). It is believed that this unconditional estimation approach would avoid the problem of oversimplifying results into significance and non-significance [25]. However, there is no empirical evidence that shows requiring researchers to adopt such an approach reduces bias and improves the interpretation of statistical results.

Improvements in research transparency are urgently needed. This includes pre-registration, the use of registered reports [20,28] and the public sharing of data and code [29]. As a reminder, a registered report is a type of journal article where authors outline their study plan, including methods and analyses, which undergoes peer review and if passed the journal commits to publishing the results. In psychology, registered report studies produced far fewer positive results (44%) than non-registered report studies (96%), with “positive” being statistically significant [8]. The success of registered reports in practice requires investment from several parties [30]. Reviewers and editors need to hold researchers to their pre-specified plan, allowing for reasonable exceptions due to unforeseen changes [31]. To be effective, researchers must use registered reports. However, their use in the field remains sparse. For example, in the three years since Science and Medicine in Football introduced registered reports, the journal has received none [32]. This is not an isolated example. In the related field of sports science, only several registered reports have been published in the Journal of Sports Sciences since their introduction [33]. Registered reports are a long-term solution to improving research transparency, requiring systemic uptake by the field.

Journals have a critical role to play in research transparency, particularly through policy and mandates [34]. For example, including the option for registered report submissions and mandating data and code sharing, with only minimal rare exceptions. When the original data cannot be shared for privacy or other reasons, a simulated data set based on the original may be sufficient to reproduce the study results [29]. We hope that the establishment of the Society for Transparency, Openness, and Replication in Kinesiology (STORK) will improve research transparency in the field, including the widespread use of registered reports.

Limitations

We summarized confidence intervals graphically and descriptively, rather than with any statistical model. About 11% of the extracted interval pairs were missing the level of confidence (Supplement 1). Although we excluded these from the analysis, we found that the cumulative distribution of these 508 interval pairs was very similar to the main analysis, see Supplement 2 [35]. Our sample size was smaller than previous similar work [12]. The relatively small sample size precluded some analyses, such as examining the cumulative distribution across journals, with some journals contributing less than 100 intervals to the data (Table 1). Nonetheless, our results clearly highlight the extent of bias in ratio confidence intervals in sport and exercise medicine, and can be considered robust given the similarity to observations in medicine where nearly 1 million ratio interval pairs were examined [12].

Conclusion

There was an excess of published research with results that were just below the standard significance threshold of 0.05, which clearly shows publication bias. Transparent research practices are needed to reduce the bias in published sport and exercise medicine research, such as the use of registered reports and the sharing of study materials, including data and code. The successful implementation of registered reports in practice requires investment from authors and journal editors, and will need journal policy changes.

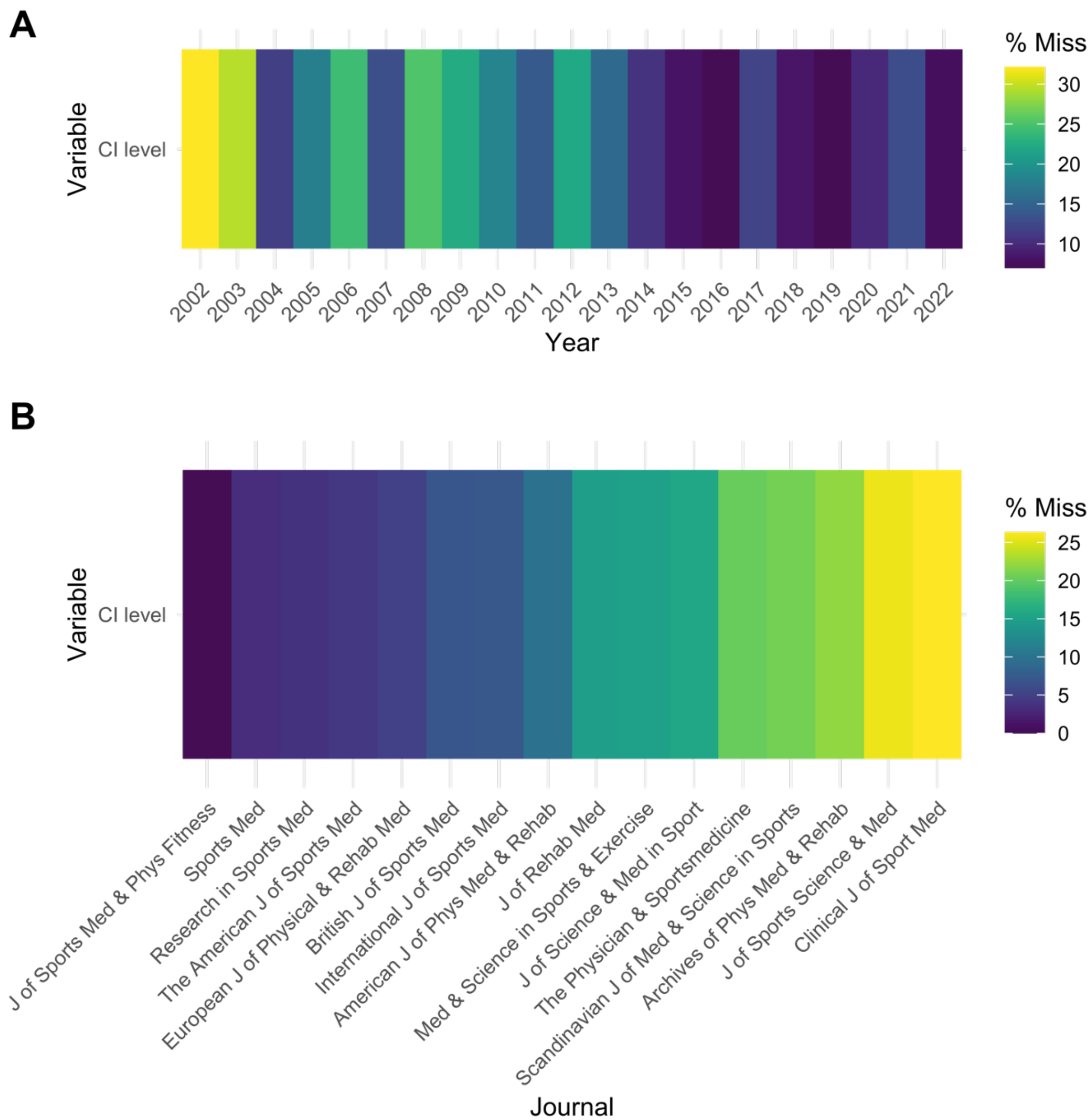
References

1. Buchanan TL, Lohse KR. Researchers' perceptions of statistical significance contribute to bias in health and exercise science. Meas Phys Educ Exerc Sci. 2016;20:131–9.
2. Emerson GB, Warme WJ, Wolf FM, Heckman JD, Brand RA, Leopold SS. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. Arch Intern Med. 2010;170:1934–9.
3. Twomey R, Yingling V, Warne J, Schneider C, McCrum C, Atkins W, et al. The Nature of Our Literature: A Registered Report on the Positive Result Rate and Reporting Practices in Kinesiology. Commun Kinesiol. 2021;1(3):1–17.
4. Büttner F, Toomey E, McClean S, Roe M, Delahunt E. 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;54:1365–71.
5. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009;4:e5738.
6. Simonsohn U, Nelson LD, Simmons JP. p-curve and effect size: Correcting for publication bias using only significant results. Perspect Psychol Sci. 2014;9:666–81.
7. Sainani KL, Borg DN, Caldwell AR, Butson ML, Tenan MS, Vickers AJ, et al. Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. Br J Sports Med. 2021;55:118–22.
8. Scheel AM, Schijen MR, Lakens D. An excess of positive results: Comparing the standard Psychology literature with Registered Reports. Adv Methods Pract Psychol Sci. 2021;4:25152459211007468.
9. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, 2013. Available: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
10. Elkins MR, Pinto RZ, Verhagen A, Grygorowicz M, Söderlund A, Guemann M, et al. Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. Phys Ther. 2022;68:1–4.

11. Ranstam J. Why the P-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage. 2012;20:805–8.
12. Barnett AG, Wren JD. Examination of CIs in health and medical journals from 1976 to 2019: an observational study. BMJ Open. 2019;9:e032506.
13. Hoekstra R, Morey RD, Rouder JN, Wagenmakers E-J. Robust misinterpretation of confidence intervals. Psychon Bull Review. 2014;21:1157–64.
14. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. Psychon Bull Review. 2016;23:103–23.
15. Georgescu C, Wren JD. Algorithmic identification of discrepancies between published ratios and their reported confidence intervals and P-values. Bioinform. 2018;34:1758–66.
16. Scimago Journal & Country Rank. Available from: https://www.scimagojr.com/
17. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci. 2018;376:20170356.
18. R Core Team. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria. Version 4.1.3; 2022. https://www.r-project.org.
19. Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatr. 2010;19:227.
20. Chambers C. The registered reports revolution Lessons in cultural reform. Signif. 2019;16:23–7.
21. van Zwet EW, Cator EA. The significance filter, the winner's curse and the need to shrink. Stat Neerl. 2021;75:437–52.
22. Berner D, Amrhein V. Why and how we should join the shift from significance testing to estimation. J Evol Biol. 2022;35:777–87.
23. van Zwet E, Schwab S, Greenland S. Addressing exaggeration of effects from single RCTs. Signif. 2021;18:16–21.
24. Hutchins KP, Borg DN, Bach AJ, Bon JJ, Minett GM, Stewart IB. Female (Under) Representation in Exercise Thermoregulation Research. Sports Med Open. 2021;7:1–9.
25. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020;20:1–13.
26. White NM, Balasubramaniam T, Nayak R, Barnett AG. An observational analysis of the trope "A p-value of< 0.05 was considered statistically significant" and other cut-and-paste statistical methods. PLoS One. 2022;17:e0264360.
27. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat. 2019;73:235–45.
28. Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, et al. Moving sport and exercise science forward: A call for the adoption of more transparent research practices. Sports Med. 2020;50:449–59.
29. Borg DN, Bon JJ, Sainani KL, Baguley BJ, Tierney NJ, Drovandi C. Comment on: 'Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices.' Sports Med. 2020;50:1551–3.
30. Scheel AM. Registered Reports: a process to safeguard high-quality evidence. Qual Life Res. 2020;29:3181–2.
31. Singh B, Fairman CM, Christensen JF, Bolam KA, Twomey R, Nunan D, et al. Outcome reporting bias in exercise oncology trials (OREO): A cross-sectional study. medRxiv. 2021:1–47.

32. Impellizzeri FM, McCall A, Meyer T. Registered reports coming soon: our contribution to better science in football research. Sci Med Footb. 2019:87–8.
33. Abt G, Boreham C, Davison G, Jackson R, Wallace E, Williams AM. Registered reports in the journal of sports sciences. J Sports Sci. 2021:1789–90.
34. Hansford HJ, Cashin AG, Wewege MA, Ferraro MC, McAuley JH, Jones MD. Open and transparent sports science research: the role of journals to move the field forward. Knee Surg Sports Traumatol Arthrosc. 2022;1–3.
35. Borg DN, Nguyen R, Tierney NJ. Missing data: current practice in football research and recommendations for improvement. Sci Med Footb. 2021;1–6.

Supplement 1. Missing data overview plot. Panel A shows the percentage of missing data, for the variable confidence interval (CI) level, each year between 2002 and 2022. Panel B shows the percentage of missing data for each journal. J = Journal, Med = Medicine, Phys = Physical.



Supplement 2. Empirical cumulative distributions for ratio confidence intervals that were missing the level of confidence. To be statistically significant, lower intervals need to be above 1, and upper intervals need to be below 1. The x-axes are restricted to focus on changes around the significance threshold of 1 (vertical line).