PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Reporting and transparent research practices in sports medicine and orthopedic clinical trials: A meta-research study
AUTHORS	Schulz, Robert; Langen, Georg; Prill, Robert; Cassel, Michael; Weissgerber, Tracey

VERSION 1 – REVIEW

REVIEWER	Bullock, Garrett University of Oxford, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
REVIEW RETURNED	08-Feb-2022

GENERAL COMMENTS	General Comments The authors investigated the transparency of reporting in the top 25% of sports medicine journals concerning clinical trials. I commend the authors for undertaking such an important, overlooked, and clinically impactful scientific question. This is a well performed study. Please see specific comments below. Abstract Can you give further detail on what encompasses scientific rigor? Was this assesses through a risk of bias analysis or some of measure? Please clarify. Introduction The introduction is well detailed and the argument is sound. I commend the authors on a well written introduction. Methods Table 1 may want to be moved to further within the methods. Right now it is before the methods, without mention of it in the introduction. I commend the authors for registering this in the open science framework. I would consider including a subsection on inclusion/exclusion criterion. To improve readability. Why were top 25% of journals chosen for this sample? Please clarify. Page 5, line 27: Did you consider using a medical librarian to create the search strategy? Is this the entire search strategy? If not, this should be inputted into the appendix. I would consider potential missing articles from one search database. For example, did you check for fidelity in clinicaltrials.gov?

	Page 5, line 37: I would change this sentence to screening titles and abstracts. Page 6: Can you give a little more detail on open science access and data? I commend the authors for performing an a priori sample size calculation. Results General comment: The authors should consider adding references for each result, as this will improve a reader's ability to connect each study. Seems the actual figure (same with figure 2) is missing from the uploaded manuscript. I do commend the authors for using a flow diagram. Page 18: Line 28: I wholeheartedly agree that bar graphs should not be used as they bias data visualization and scientific communication. However, this statement should be within the discussion (and will be a great discussion point!). Discussion Page 22, Line 8: This is a really interesting point. I am not surprised by this, but is very salient. I would think pre-registration demonstrates increased scientific rigour through the entire process.

REVIEWER	McCrum, Christopher Maastricht Universitair Medisch Centrum+
REVIEW RETURNED	26-Apr-2022

GENERAL COMMENTS	Thank you for the opportunity to review this interesting and worthwhile study. The authors present results of a meta-research study of sports medicine and orthopaedic trials aiming to document current study reporting practices. The results mirror those of similar audits in other related and less-related fields in the general sense that some criteria are commonly reported but not to the extent that is typically desirable, while other criteria are rarely reported in sufficient detail. The discussion presents the authors' suggestions for potential interventions to improve the reporting standards in this field. Overall, the study methodology is appropriate to address the aims and the results are clearly presented and reported. I have some questions related to the sample and elements of the discussion and the recommendations for potential interventions as described below. I look forward to reading the authors' responses and hope these comments are useful for the authors. Major comments: Sample: Final paragraph of the introduction: "this meta-research study examined reporting among clinical trials published in the top 25% of sports medicine and orthopedics journals". From the introduction alone, it is not clear why only the top 25% was used, nor how "top" was determined. I understand that this is elaborated in the methods, but at least qualifying the type/method of ranking here would be helpful. Sample selection and screening: "This sampling strategy provides a broad overview of practices in the field while including high-impact journals, which have the potential to drive change." It is unclear what "drive change" refers to here – change in applied/clinical practice or change in reporting practice? Nevertheless, it is not apparent why either is relevant for the inclusion in this study. We know that impact factor and citations are

not an indication of study quality and many newer journals not yet incorporated in such rankings do place much more importance on open, transparent reporting. Can the authors comment on how truly representative they feel that their sample is?
Sample selection and screening: “The top 25% of journals (n=65) were entered into the PubMed search with article type (clinical trial) and publication date (2019/12:2020/08) filters”. There is a slight issue here in that the clinical trial article type may not always be sensitive or accurate. This issue might be compounded by the fact that, in this broader field, clinical trials are often not registered on clinical trial registries (https://doi.org/10.51224/cik.v1i3.43). As a result of this, could it be the case that you may be excluding some less-well reported manuscripts?

Figure 4:
While I understand the desire to give an example of what would be considered good reporting practice, I actually disagree that the example given is a good example of how to do/report an a priori power analysis. Taking a single effect size from an RCT, while common, is undesirable due to the risk that it overestimates the true effect size (single study effect, publication bias, inflated effect from small sample size, etc.) and that it gives no guarantee that the effect size is clinically relevant or indeed what the minimum effect size of interest is. It also doesn’t provide any justification of the type 1 and 2 error rate and while these are the normative values, many fields would consider 0.8 too low. I realise that my comments relate more to the appropriateness of the sample size calculation and not its reporting, but since this figure has been created in a shareable, “infographic” type of format, we can assume that this might be well circulated, in which case I would argue that the example given should not only tick all the desired reporting boxes but also be a good example of a well justified sample size calculation. I realise this is nit-picking, but hopefully the authors understand my concern. Perhaps they are already aware of it but Daniël Lakens’ article provides an excellent overview of these issues that could be drawn on for the purpose of this figure (https://doi.org/10.1525/collabra.33267).

Minor suggestions:
Introduction
The following is only my personal suggestion. Personally, I don’t generally like the use of puns or other similar features in titles that don’t help the reader identify the main content, aim or conclusion in the article. “The devil is in the details” doesn’t seem to help readers quickly understand the focus of the manuscript any more than the rest of the title, so I would suggest removing it for a more concise title. You may also consider going for a stronger statement based on the results of the study to have a more attention-grabbing title.
Somewhere in or between paragraphs 1 and 2 of the introduction, it might be information to include a sentence or two on the various types of bias that improved reporting practice would help address. Otherwise, the various criteria mentioned in the final paragraph of the introduction come a little out of the blue for less familiar readers.
Second last paragraph of the introduction: The authors may wish to update/add information from the following study, since it lends further support to the points made:
https://doi.org/10.51224/cik.v1i3.43 (full disclosure - I am a co-author on this paper and will not base my recommendation on the in- or exclusion of this citation).

Table 1: The text on NHST has undertones that imply NHST is inherently problematic, whereas I would argue that it is its inappropriate use that is problematic. I wonder if this could be slightly reworded.

Methods

Table 2: Data visualisation: This assessment criterion seems a little narrow since there are multiple issue that one could assess related to data visualisation (i.e. use of individual data points, method of indicating distribution, reproducibility of figures with data and code etc.). Why the specific focus on bar graphs?

Discussion

General comment that many of the results here agree with our recent study mentioned above in the broader sports science field. Comparison may be warranted to strengthen the authors' conclusions.

Page 18, second paragraph: "The benefits of data sharing for authors include more citations (64,65), and increased opportunities to collaborate with researchers who want to perform secondary analyses (66)." On reading this sentence, I realise that the reproducibility/trust aspects of data sharing are not mentioned in the manuscript. I think these reasons should take precedence over the ones stated here.

Page 18-19: I agree that this might be due to unrealistic/inflated effect sizes being used in effect size calculations. But it may also be due to the fact that researchers may be powering their studies on the average effect for a field (via meta-analysis or from a few previous studies) that are not necessarily inflated, but which are nevertheless not the minimum effect size of interest or the minimum clinically relevant effect size.

Page 19: "Focusing on the magnitude and precision of differences, instead making decisions based on p-value thresholds, reduces the likelihood of spurious findings (72,73)." I do not think the references used provide evidence to support this statement. The choice of reference here is also curious since one explicitly lends support to magnitude-based inference, which has been shown to increase the false positive risk (see work by Kristin Sainani), similar to what the authors write here for using the p-value.

Page 20, first two paragraphs. I don't think it is mentioned anywhere, but were the author guidelines or journal policies of the journals publishing the included articles taken into account in any way? For example, how many of the journals included here endorse CONSORT? How many include author guidelines related to the assessed criteria of the current study? This information could lend some preliminary evidence to how effective such guidelines and policies are, or at least inform on whether they are being enforced. For example, the journal Medicine and Science in Sports and Exercise explicitly requires a sample size calculation. A related question might be if there were authors overlapping in the included publications, since there may be some kind of clustering effect related to practice/reporting norms within research groups.

Page 20, second option: a limitation of this approach seems to be a lack of incentive for authors to do "extra work" for no return. In the first option, at least the journal can withhold publication until reporting criteria are met.

Page 20, third option. Perhaps I miss the point, but I find it difficult to see how dashboards as a journal or field will help directly incentivise individual researchers to improve their reporting practices, since there is no direct reward tied to it. Perhaps the

	authors can elaborate or give more examples on how implementation of options 2 and 3 might directly influence individual practice? Page 20, education paragraph: while as a STORK member, I agree that improving and disseminating educational material is vital in general, I am not sure that this will be one of the more successful approaches in isolation. Similar to my thoughts on options 2 and 3, there is still no direct reward and educational opportunities will only help improve the practices of those who are already interested in doing so and have already “bought in” to the value of such practices to some extent. In general, I miss a consideration of our academic incentive structure in this section on interventions, which I think is a major barrier and something that interventions need to account for in order to have a large effect on practice. Conclusions Just my personal preference, but I would suggest removing the first two sentences from the conclusion, since these are actually not a conclusion based on your results. Conclusion second paragraph – as mentioned above, I feel that the academic incentive structure has to be incorporated in some way into the discussion of these interventions. Figure S1: There appears to be an error in the first box.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1, Dr. Garrett Bullock, University of Oxford

Comments to the Author:
General Comments
The authors investigated the transparency of reporting in the top 25% of sports medicine journals concerning clinical trials. I commend the authors for undertaking such an important, overlooked, and clinically impactful scientific question. This is a well performed study. Please see specific comments below.
Author reply: We thank the reviewer for his careful review of our manuscript and his very helpful suggestions for improvement of the article.

Abstract
⌡ Can you give further detail on what encompasses scientific rigor? Was this assesses through a risk of bias analysis or some of measure? Please clarify.
Author reply:
We thank the reviewer for this suggestion. We added more information to clarify which criteria related to scientific rigor were assessed. As mentioned later in the manuscript, these criteria were selected based on CONSORT 2010. We added:
“Two independent reviewers assessed pre-registration, open data, and criteria related to scientific rigor, including randomization, blinding, and sample size calculations, as well as the study sample, and data analysis.”

Introduction
⌡ The introduction is well detailed and the argument is sound. I commend the authors on a well written

introduction.

Author reply:

We thank the reviewer for the encouraging words.

Methods

⌡ Table 1 may want to be moved to further within the methods. Right now it is before the methods, without mention of it in the introduction.

Author reply:

We have adjusted the text in the second to last paragraph of the introduction to more clearly emphasize the relevance of table 1 here, and to encourage readers to examine the table.

⌡ I commend the authors for registering this in the open science framework.

I would consider including a subsection on inclusion/exclusion criterion. To improve readability.

Why were top 25% of journals chosen for this sample? Please clarify.

Author reply:

We are thankful for this comment as this suggestion led to improved readability indeed. We implemented a new subsection on inclusion and exclusion criteria:

“Inclusion and exlcusion criteria

Two reviewers (RS, GL) screened titles and abstracts to exclude articles that were obviously not clinical trials, as defined by the ICMJE. […]”

Additionally, we rephrased the rationale for choosing the top 25% of journals to:

“This sampling strategy provides an overview of practices in the field , particularly among journals whose articles receive the most attention. The large number of journals included ensures that findings are not driven by practices or policies of individual journals.”

The sampling frame was selected to capture journals that most scientists in the field would regard as reputable, while minimizing the likelihood of including predatory journals which may be listed lower on the SciMago Journals Ranking List. This approach also included enough journals that observed estimates were unlikely to be largely attributable to unique practices from a single journal. The top 25% journals ranked by SJR (comparable to impact factor metric) were not chosen because we expected those articles to be of higher or lower quality than journals with a lower rank. We used the SJR as a proxy measure for the attention that articles in those journals are receiving. Due to word limit restrictions, we cannot provide more information in the manuscript text.

⌡ Page 5, line 27: Did you consider using a medical librarian to create the search strategy? Is this the entire search strategy? If not, this should be inputted into the appendix.

I would consider potential missing articles from one search database. For example, did you check for fidelity in clinicaltrials.gov?

Author reply:

We fully agree that the search strategy in systematic reviews and meta-analysis is of extraordinary importance. However, this meta-research study differs from more traditional systematic reviews with meta-analyses in some important aspects. Instead of attempting to find all relevant articles on a specific topic, we used a journal-based sample. All journals in our sample (n=65) were indexed in PubMed, which allowed us to use one database only. We agree with the reviewer that multiple databases are essential for topic based searches, as different databases index different journals. However, the use of a single database is appropriate for journal based searches in which all journals are indexed in the database that is searched. While medical librarians can be invaluable for selecting keywords for topic based searches, identifying all articles published in a predefined list of journals and specified time period does not require extensive expertise. While we checked clinicaltrials.gov to assess criteria related to pre-registration, our sample focussed on clinical trials that were published in the top 25% of journals in the field. While investigating the nature of (unpublished) clinical trials in clinical trial registries would certainly be a very interesting and valuable area of research, this was not the aim of the present study. We also included a statement in our limitations section that our results

may not be generalizable to journals that are not indexed by PubMed:
“We examined the top 25% sports medicine and orthopedics journals; hence our findings may not be generalizable to journals that are not indexed by PubMed, lower tier journals, non-English journals, or unpublished trials. The use of the clinical trial filter may have led to the exclusion of a small number of trials that were incorrectly classified upon indexing.”

⌡ Page 5, line 37: I would change this sentence to screening titles and abstracts.
Author reply:
We thank the reviewer for careful reading. We implemented this suggestion as proposed:
“The search was run on September 16, 2020. All articles (n=175 from 27 journals) were uploaded into Rayyan (RRID:SCR_017584; 36) to screen titles and abstracts.”

⌡ Page 6: Can you give a little more detail on open science access and data?
I commend the authors for performing an a priori sample size calculation.
Author reply:
We thank the reviewer for the helpful comment. We added more details on information on open access and open data that was extracted from the trial publications in the text:
“We also abstracted additional open science criteria, focusing on the open access status of the trial publication, whether a data availability statement was included and whether data were deposited in a public repository (37,38).”

Results
⌡ General comment: The authors should consider adding references for each result, as this will improve a reader’s ability to connect each study.
Author reply:
We thank the reviewer for his careful review.
We added a reference to the related figure in each subsection of the results.
Moreover, the single ratings for each criterion in each included individual trial are available in the referenced OSF-project. We added a statement on the option to see all individual ratings to the beginning of the results section:
“Additional details regarding sample selection and screening, data abstraction, and sample size calculation as well as all ratings for each individual study can be found in the supplemental materials.”

⌡ Seems the actual figure (same with figure 2) is missing from the uploaded manuscript. I do commend the authors for using a flow diagram.
Author reply:
In our version of the manuscript, all figures are included. Unfortunately, the submission system does not allow the figures to be embedded into the text, but requires figures to be attached at the end of the manuscript. We have double checked to ensure that all figures are accessible to the reviewer in the uploaded revision.

Page 18: Line 28: I wholeheartedly agree that bar graphs should not be used as they bias data visualization and scientific communication. However, this statement should be within the discussion (and will be a great discussion point!).
Author reply:
We thank the reviewer for underlining the importance of transparent data visualization. Due to word count restrictions, we did not discuss the related results in the original manuscript. We have added a brief discussion of this data in the revised manuscript, as suggested:
“Our study shows that more than one-fifth of the included trials used bar graphs to visualize continuous data. While this practice is common in many fields (75), these figures are problematic because many different data distributions can lead to the same summary statistics shown in bar graphs. Researchers should use data visualisations that show the data distribution, such as dot plots,

box plots, or violin plots (50,51).”

Discussion

- Page 22, Line 8: This is a really interesting point. I am not surprised by this, but is very salient. I would think pre-registration demonstrates increased scientific rigour through the entire process.

Author reply:
We thank the reviewer for his expression of interest. We also hope that this work can stimulate follow-up research in this area.

Reviewer: 2, Dr. Christopher McCrum, Maastricht Universitair Medisch Centrum+

Comments to the Author:
Thank you for the opportunity to review this interesting and worthwhile study. The authors present results of a meta-research study of sports medicine and orthopaedic trials aiming to document current study reporting practices. The results mirror those of similar audits in other related and less-related fields in the general sense that some criteria are commonly reported but not to the extent that is typically desirable, while other criteria are rarely reported in sufficient detail. The discussion presents the authors’ suggestions for potential interventions to improve the reporting standards in this field. Overall, the study methodology is appropriate to address the aims and the results are clearly presented and reported. I have some questions related to the sample and elements of the discussion and the recommendations for potential interventions as described below. I look forward to reading the authors’ responses and hope these comments are useful for the authors.

Major comments:
Sample:

- Final paragraph of the introduction: “this meta-research study examined reporting among clinical trials published in the top 25% of sports medicine and orthopedics journals”. From the introduction alone, it is not clear why only the top 25% was used, nor how “top” was determined. I understand that this is elaborated in the methods, but at least qualifying the type/method of ranking here would be helpful.

Author reply:
We thank the reviewer for his careful review of our manuscript. Reasons for selecting the top 25% of journals are explained in our responses to the first reviewer. We added the following part to the first sentence of the last paragraph of the introduction:
“Therefore, this meta-research study examined reporting among clinical trials published in the top 25% of sports medicine and orthopedics journals, as determined by Scientific Journal Rank.”

- Sample selection and screening: “This sampling strategy provides a broad overview of practices in the field while including high-impact journals, which have the potential to drive change.” It is unclear what “drive change” refers to here – change in applied/clinical practice or change in reporting practice? Nevertheless, it is not apparent why either is relevant for the inclusion in this study. We know that impact factor and citations are not an indication of study quality and many newer journals

not yet incorporated in such rankings do place much more importance on open, transparent reporting. Can the authors comment on how truly representative they feel that their sample is?
Author reply:
With this sampling frame, we wanted to ensure to include reputable journals in the field while minimizing the likelihood of including any predatory journals which are more likely to be listed lower on ranking lists. The top 25% journals ranked by SJR (comparable to impact factor metric) were not chosen because we expected those articles to be of higher or lower quality than journals with a lower rank. We used the SJR because it gives an idea of the average attention that articles are receiving in the respective journals. With random sampling, there would have been the risk of excluding journals whose articles receive lots of attention, possibly leaving room for criticism that we cannot speak for a field while neglecting the most reputable journals with the most widely circulated articles. While most meta-research sampling strategy yields specific weaknesses, we deliberately chose this sampling strategy during the study design- and registration phase because we concluded that this strategy would provide the best compromise. And of course, we addressed the limits of the representativity in the limitations section of our manuscript ("We examined the top 25% sports medicine and orthopedics journals; hence our findings may not be generalizable to journals that are not indexed by PubMed, lower tier journals, or non-English journals.")
Due to word limit restrictions, we cannot provide more information in the manuscript text but we rephrased the quoted sentence to avoid misinterpretation of the rationale for our sampling frame as follows:
"This sampling strategy provides a broad overview of practices in the field while ensuring to include those journals whose articles receive the most attention."

⎠ Sample selection and screening: "The top 25% of journals (n=65) were entered into the PubMed search with article type (clinical trial) and publication date (2019/12:2020/08) filters". There is a slight issue here in that the clinical trial article type may not always be sensitive or accurate. This issue might be compounded by the fact that, in this broader field, clinical trials are often not registered on clinical trial registries (https://doi.org/10.51224/cik.v1i3.43). As a result of this, could it be the case that you may be excluding some less-well reported manuscripts?
Author reply:
We share your reservations regarding the use of publication type filters in principle. However, in the present study, the objective was not necessarily to find every single studies that was ever published on a certain subject. Basic filters for clinical trial publication types can be highly sensitive (10.5195/jmla.2020.912, 10.1136/bmj.38446.498542.8F), and the false-positives were excluded during manual review by two independent researchers. The circumstance that the 48% of unregistered trials that we found with our search strategy is surprisingly high already and at least does not strongly suggest that using our search strategy led to exantensive, systematic exclusion of unregistered trials.
Yet, as we generally share the thoughts of the reviewer on that mater, we added a statement regarding this matter to the limitations section.
"We examined the top 25% sports medicine and orthopedics journals; hence our findings may not be generalizable to journals that are not indexed by PubMed, lower tier journals, non-English journals, or unpublished trials. The use of the clinical trial filter may have led to the exclusion of a small number of trials that were incorrectly classified upon indexing."

⎠ Figure 4: While I understand the desire to give an example of what would be considered good reporting practice, I actually disagree that the example given is a good example of how to do/report an a priori power analysis. Taking a single effect size from an RCT, while common, is undesirable due to the risk that it overestimates the true effect size (single study effect, publication bias, inflated effect from small sample size, etc.) and that it gives no guarantee that the effect size is clinically relevant or indeed what the minimum effect size of interest is. It also doesn't provide any justification of the type 1 and 2 error rate and while these are the normative values, many fields would consider 0.8 too low. I

realise that my comments relate more to the appropriateness of the sample size calculation and not its reporting, but since this figure has been created in a shareable, “infographic” type of format, we can assume that this might be well circulated, in which case I would argue that the example given should not only tick all the desired reporting boxes but also be a good example of a well justified sample size calculation. I realise this is nit-picking, but hopefully the authors understand my concern. Perhaps they are already aware of it but Daniël Lakens’ article provides an excellent overview of these issues that could be drawn on for the purpose of this figure (https://doi.org/10.1525/collabra.33267).
Author reply:
We thank the reviewer for his very knowledgeable comment. In general, we absolutely agree with the standpoint that all elements of an a priori sample size calculation should be justified individually, and we truly appreciate Daniel Lakens’ excellent article on sample size justification.
However, our findings show that many studies lack a very basic level of reporting the single elements of sample size calculations. We created this infographic to provide an easily accessible entry point to start learning more about a priori sample size calculations. Similar to many other assessments that have been conducted in this study, like the blinding or randomization, we assessed the reporting of sample size calculations but not its quality necessarily.
We want to avoid overloading the infographic to prevent that it appeals only to those who already have a certain degree of interest, understanding, and knowledge about a priori sample size calculations.

Moreover, while we agree that a statistical power of 0.8 can rightfully be considered too low in some cases or fields, the actual statistical power of rehabilitation trials (10.1016/j.apmr.2020.02.017) or clinical trials in general (10.1016/j.jclinepi.2018.06.014) likely rather centers around 0.15.

We fully support the concept of the smallest effect sizes of interest as one way to evaluate which effect sizes are of interest for a given study (six ways presented in Lakens 2022 alone, https://doi.org/10.1525/collabra.33267), which is why we included the smallest effect sizes of interest/ smallest meaningful effect as one of the examples in the section “What should be reported” of Figure 4.
We chose to use a single study as reference for the expected effect in our example for several reasons:
1. Figure 4 is about the essential aspects that should be reported in a priori sample size calculations. We wanted to keep the information as accessible as possible. A single study seems an easier reference for an expected effect compared to (the potentially new concept) SESOI.
2. We tried to keep our assessment and recommendations as close to CONSORT as possible, as done for most assessments in our study.
see https://doi.org/10.1136/bmj.c869:
“Item 7a. How sample size was determined
Examples—“To detect a reduction in PHS (postoperative hospital stay) of 3 days (SD 5 days), which is in agreement with the study of Lobo et al17 with a two-sided 5% significance level and a power of 80%, a sample size of 50 patients per group was necessary, given an anticipated dropout rate of 10%. To recruit this number of patients a 12-month inclusion period was anticipated.” […]
Elements of the sample size calculation are (1) the estimated outcomes in each group (which implies the clinically important target difference between the intervention groups); (2) the α (type I) error level; (3) the statistical power (or the β (type II) error level); and (4), for continuous outcomes, the standard deviation of the measurements. The interplay of these elements and their reporting will differ for cluster trials and non-inferiority and equivalence trials.)”
3. Similar to using effect estimates from single studies or meta-analyses that need to be controlled for possible publication-bias, it must be ensured that the justification of the SESOI is relevant to the planned study when using SESOI (https://journals.sagepub.com/doi/10.1177/2515245918770963). In both cases, further engagement with the topic is necessary to adequately justify the expected effect

size for a priori sample size calculations. As our study focusses predominantly on aspects of reporting, we chose the more accessible option.

However, as we highly value the comment that was made by the reviewer, we added a section on the need for indivdually justifying all elements of a sample size calculation and included a referral to Daniel Lakens article on sample size justification.

Therefore, we added the following statement to the figure caption:

“This infographic focuses on key elements a priori sample size calculations that should be reported in clinical trial publication. However, it is important to note that each element should be justified individually, including the thresholds for type 1 and type 2 errors, and the expected effect size. Daniel Lakens free article on sample size justification provides an excellent overview of aspects to consider when planning empirical research studies (62).”

Furthermore, we changed the “What should be reported”-section of Figure 4 to:

“The justification of the expected effect size with literature reference (e.g. smallest effect size of interest, meta-analysis, single study)”

And we changed the example-section to:

“We performed an a priori sample size calculation for our primary outcome O, comparing intervention and placebo. We set the Type I error at α = 0.05 and the Type II error at β = 0.1.

Cohen's d = 0.5 is considered the smallest effect size of interest for studies comparing O between intervention and placebo (X et al. 2001), and was used for the sample size calculation of the present study. Therefore, the required number of participants per group is 84.”

Minor suggestions:

Introduction

The following is only my personal suggestion. Personally, I don’t generally like the use of puns or other similar features in titles that don’t help the reader identify the main content, aim or conclusion in the article. “The devil is in the details” doesn’t seem to help readers quickly understand the focus of the manuscript any more than the rest of the title, so I would suggest removing it for a more concise title. You may also consider going for a stronger statement based on the results of the study to have a more attention-grabbing title.

Author reply:

We admire the honest feedback and thank reviewer for sharing his personal feelings on this matter. We used the saying “The devil is in the details” in the title for this manuscript because we wanted to create an engaging title that still represents a rather accurate summary of our results. However, your feelings on the title show that this is not working for everyone and we are happy with replacing our current title. However, in our author group, we are still undecided on the best alternative title. We have listed three different options underneath and would be glad to receive feedback from you on which title of those you would like best.

- “Reporting and transparent research practices in orthopaedics and sports medicine clinical trials: A meta-research study”

⌡ Somewhere in or between paragraphs 1 and 2 of the introduction, it might be information to include a sentence or two on the various types of bias that improved reporting practice would help address. Otherwise, the various criteria mentioned in the final paragraph of the introduction come a little out of the blue for less familiar readers.

Author reply:

We thank the reviewer for this insightful advice. We added a selection of biases that can be prevented or better detected with comprehensive reporting:

“Comprehensive reporting may prevent biases like selective reporting, selection bias, attrition bias, or wrong sample size bias, or make them easier to detect.”

⌡ Second last paragraph of the introduction: The authors may wish to update/add information from the

following study, since it lends further support to the points made: https://doi.org/10.51224/cik.v1i3.43 (full disclosure - I am a co-author on this paper and will not base my recommendation on the in- or exclusion of this citation).
Author reply:
We thank the reviewer for this relevant note. In the introduction, we highlight selected literature that informed the design of our study. As the article mentioned was published after posting our preprint on MedRxiv and after our submission to BMJ Open, we were not able to include this before.
We added the reference to the discussion:
“A recent study in kinesiology shows even lower rates of pre-registration, data-availability statements, and data sharing in open repositories (64).”

⌡ Table 1: The text on NHST has undertones that imply NHST is inherently problematic, whereas I would argue that it is its inappropriate use that is problematic. I wonder if this could be slightly reworded.
Author reply:
We are thankful for the important advice from the reviewer. We have rephrased the sentence to:
“While NHST is standard practice in many fields, the International Committee of Medical Journal Editors warns against the inappropriate use and sole reliance on NHST due to several shortcomings of using this approach inappropriately (32).”

Methods
⌡ Table 2: Data visualisation: This assessment criterion seems a little narrow since there are multiple issue that one could assess related to data visualisation (i.e. use of individual data points, method of indicating distribution, reproducibility of figures with data and code etc.). Why the specific focus on bar graphs?
Author reply:
We thank the author for this relevant comment. We certainly agree that there is a large number of possible assessments that could be conducted on data visualization. The options proposed by the reviewer could form the basis for a larger study on visualization practices.
Since data visualization was not the primary focus of the study, we chose to examine one aspect of data visualization to avoid lengthening an already intensive data abstraction form. We chose the topic of bar graphs for continuous data because prior studies from our group and others have shown that these graphs are commonly used. This issue has generated considerable attention in recent years, leading to policy changes in many journals, and simple solutions are readily available (10.1371/journal.pbio.1002128, 10.1161/CIRCULATIONAHA.118.037777, 10.1101/2022.03.14.484206).
However, we agree that the data visualization part was slightly detached from the remaining manuscript. Therefore, we added the following sentences to the discussion:
“Our study shows that more than one-fifth of the included trials used bar graphs to visualize continuous data. While this practice is common in many fields (75), these figures are problematic because many different data distributions can lead to the same summary statistics and bar graphs. Researchers should use data visualisations that show the data distribution such as dot plots, box plots, or violin plots (50,51).”

Discussion
⌡ General comment that many of the results here agree with our recent study mentioned above in the broader sports science field. Comparison may be warranted to strengthen the authors’ conclusions.
Page 18, second paragraph: “The benefits of data sharing for authors include more citations (64,65), and increased opportunities to collaborate with researchers who want to perform secondary analyses (66).” On reading this sentence, I realise that the reproducibility/trust aspects of data sharing are not mentioned in the manuscript. I think these reasons should take precedence over the ones stated here.

Author reply:
We thank the author for sharing his thoughts on this topic. As proposed for other parts of the manuscript, we wanted to focus here on incentives for the individual researcher, as the individual researcher has to carry out the additional work of sharing data openly. However, we strongly agree with your opinion that the idealistic rationale of more trustworthy research outputs should not be neglected, although we want to avoid by all means implying that research outputs without open data are automatically less trustworthy. Therefore, we added this aspect to the quoted sentence:
“The benefits of data sharing for authors include more citations (66,67), likely increased trustworthiness (68), and increased opportunities to collaborate with researchers who want to perform secondary analyses (69).”

⌡ Page 18-19: I agree that this might be due to unrealistic/inflated effect sizes being used in effect size calculations. But it may also be due to the fact that researchers may be powering their studies on the average effect for a field (via meta-analysis or from a few previous studies) that are not necessarily inflated, but which are nevertheless not the minimum effect size of interest or the minimum clinically relevant effect size.
Author reply:
We believe that the explanation proposed by the reviewer may be a separate problem, rather than an explanation for the observed phenomenon. If the effect size estimates used in power calculations are accurate (not inflated), then the studies should be adequately powered and should detect effects of the expected size used in the power calculation. The fact that the true effect size of interest may be smaller or larger is a separate issue.

⌡ Page 19: “Focusing on the magnitude and precision of differences, instead making decisions based on p-value thresholds, reduces the likelihood of spurious findings (72,73).” I do not think the references used provide evidence to support this statement. The choice of reference here is also curious since one explicitly lends support to magnitude-based inference, which has been shown to increase the false positive risk (see work by Kristin Sainani), similar to what the authors write here for using the p-value.
Author reply:
We thank the reviewer for this important point. We chose these references because they highlighted many different statistical approaches that can be used with NHST (reporting standardized effects sizes and providing confidence intervals of point estimates) or instead of NHST (proper Bayesian analysis, not MBI). We did not intend to imply any support for magnitude-based inferences with this statement but wanted to emphasize that the p-value should only be one element amongst many when evaluating the effectiveness of different interventions. We eliminated the word magnitude from the sentence to avoid any confusion over whether this might refer to MBI:
“Instead of making decisions based on p-values alone, reporting the size and precision of effects in combination with the p-value provides a more complete representation of the results and reduces the likelihood of spurious findings.”

⌡ Page 20, first two paragraphs. I don’t think it is mentioned anywhere, but were the author guidelines or journal policies of the journals publishing the included articles taken into account in any way? For example, how many of the journals included here endorse CONSORT? How many include author guidelines related to the assessed criteria of the current study? This information could lend some preliminary evidence to how effective such guidelines and policies are, or at least inform on whether they are being enforced. For example, the journal Medicine and Science in Sports and Exercise explicitly requires a sample size calculation. A related question might be if there were authors overlapping in the included publications, since there may be some kind of clustering effect related to practice/reporting norms within research groups.
Author reply:

The reviewer raises two very interesting questions here. Studies on the effectiveness of journal policies on certain reporting characteristics would need a different sampling frame, in which journals were selected based on policies to ensure adequate representation of journals with different policy types. A longitudinal sample would also be helpful here to evaluate the impact of policy changes. A much larger sample would also likely be required to answer this question. We chose to conduct a descriptive and exploratory study that aims to provide an overview of reporting practices within the field of orthopaedics and sports medicine and provide a basis for subsequent work on the topics that are most important for the field. At the study design- and registration stage, we decided against a number of possible comparisons or associations that could be investigated because there had been no reliable field-specific estimates of reporting prevalences, which would have been crucial for planning a study that investigates the effects of certain characteristics on reporting quality.

⌡ Page 20, second option: a limitation of this approach seems to be a lack of incentive for authors to do “extra work” for no return. In the first option, at least the journal can withhold publication until reporting criteria are met.
Author reply:
We share the authors concern that a broader overhall of incentives is needed to encourage and support authors in adopting more rigorous practices. In the absence of systemic, widespread changes, we have attempted to emphasize “selfish” reasons for improving practice throughout the article (e.g. direct benefits to the author). We view automated screening of preprints on preprint servers as a scalable way of increasing awareness of good reporting and transparent research practices. One might speculate that public automated screening reports may form an incentive for good reporting in themselves, as clinical trial authors can achieve more favorable public reports by updating their preprints with more transparent reporting. However, meta-research studies would be needed to assess this claim.

⌡ Page 20, third option. Perhaps I miss the point, but I find it difficult to see how dashboards as a journal or field will help directly incentivise individual researchers to improve their reporting practices, since there is no direct reward tied to it. Perhaps the authors can elaborate or give more examples on how implementation of options 2 and 3 might directly influence individual practice?
Author reply:
We agree that dashboards are a monitoring tool, and additional steps are required to integrate this tool into a broader rewards and incentives structure. The EU-Trials Tracker is one example of a dashboard that made a significant impact on research and policy making (10.1136/bmj.k3218). Publicity surrounding these dashboards has raised awareness about unreported trials, leading to polical pressure and new policies to require timely reporting. We have clarified these points in the text:
“Dashboards may offer a third option for monitoring changes in practice over time, and raising awareness about the importance of specific reporting practices among researchers, policymakers and the public. When used to inform increntives systems, dashboards may potentially contribute to improved reporting. Dashboards may work best in combination with other measures, like policy changes, incorporating practices described in dashboards into researcher assessments, or rewarding researchers for improving reporting.. Policymakers and the scientific community can use dashboards to evaluate the effectiveness of interventions to improve scientific practice. […]”

⌡ Page 20, education paragraph: while as a STORK member, I agree that improving and disseminating educational material is vital in general, I am not sure that this will be one of the more successful approaches in isolation. Similar to my thoughts on options 2 and 3, there is still no direct reward and educational opportunities will only help improve the practices of those who are already interested in doing so and have already “bought in” to the value of such practices to some extent. In general, I miss a consideration of our academic incentive structure in this section on interventions, which I think is a major barrier and something that interventions need to account for in order to have a

large effect on practice.
Author reply:
We thank the reviewer for sharing his thoughts on the very important but difficult subject of achieving behavioral change in clinical trial reporting practice. We agree that incentives are a very important aspect of achieving behavioral change. However, incentivization most often requires extensive systemic efforts on regulatory, institutional, and research funding levels that are discussed in several dedicated articles in detail (10.1080/00461520.2021.1902329, 10.1186/s13104-022-05942-3, 10.1186/s13104-022-05932-5, 10.1371/journal.pbio.3000576). While these aspects are highly important, we want to focus our recommendations in this article on topics that are more directly related to clinical trial authors and clinical trial reporting.
However, we added the following sentence to the first paragraph of Options for systemic interventions to improve reporting:
“Transparent research practices and reporting need to be incentivized on different levels and by different stakeholders in the academic research lifecycle (79,80). Persistent reporting deficiencies (12,21) indicate that endorsement without enforcement is insufficient (81,82), and engaging individuals, journals, funders, and institutions is necessary to improve reporting (79,83).”
More in-depth discussions of general incentivization strategies to improve scientific rigor and reproducibility are better covered in specialized articles on that topic that have a different target audience.

Conclusions
⌡ Just my personal preference, but I would suggest removing the first two sentences from the conclusion, since these are actually not a conclusion based on your results.
Author reply:
We thank the reviewer for his helpful comment. We removed the first two sentences from the conclusion.

⌡ Conclusion second paragraph – as mentioned above, I feel that the academic incentive structure has to be incorporated in some way into the discussion of these interventions.
Author reply:
We thank the reviewer for expressing his thoughts on this topic here again. While we still think that the very complex topic of academic incentive structure should better be discussed in dedicated outlets, we added the topic of incentives to the last paragraph:
“[…] the persistent lack of detailed reporting suggests that education and existing guidelines alone are not working. Better incentives, further interventions, and other innovative approaches are needed to improve clinical trial reporting further.”

VERSION 2 – REVIEW

REVIEWER	Bullock, Garrett University of Oxford, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
REVIEW RETURNED	02-Jul-2022

GENERAL COMMENTS	General Comments This study investigated the transparency and reporting of research practices in sports and orthopaedics. This is an important endeavor, as this is essential for moving the profession forward and creating improved methods and repeatability of these methods. Please see specific comments below.

Introduction
The second half of the first paragraph is a little confusing. It is not clear if this information on allocation concealment, randomization, etc, was from your study or a previous study. Please clarify.

Overall, the introduction has very good information and a thorough literature review has been performed. However, I think the argument needs to be made clearer why this specific study needs to be performed. The authors give good information on the past studies that have investigated poor methods; but, detailing where there are gaps for this meta-research study will strengthen the argument for this study. Are the studies not comprehensive? Or missing specific sports? Too old with the large amount of publications that have been disseminated recently?
Methods
I commend the authors for registering on OSF.
Sorry if I missed this. While this is a meta-research and ethics are not needed, can you add a statement explicitly saying this?
Why only 9 months for exploration (instead of a year, or other time measure), can this be justified by previous references?
Concerning search strategy: it is not quite clear what the exact search strategy used for this study. The date and the use of the top 25% of journals is apparent? But the search strategy is not clear.

Page 9, space needed between ‘amongjournals’
Page 9; line 28: misspelling of ‘aee’ ..are?

In inclusion exclusion criteria section, the screening process and inclusion/exclusion criteria are mixed together, which makes it hard to follow. Suggest making two separate sections. Further, more detail is needed for the inclusion/exclusion section.
I commend for examining open science practices in addition to CONSORT. The BIH reference (36) which was referenced for open science, the link was an error.

I commend the authors for reporting protocol deviations.

Page 10, line 50; space needed between ‘the first’

Can you explain further why if the outcome was changed after the study began that this was deemed a retrospective study? Would this be more deviation from the protocol, or a posteriori hypotheses creating (if they did the change to make the study ‘better’), or something else?

Results
It may be recommended to include referencing for all included studies within the results.
Data visualization: The authors give recommendations for data visualization within the results. While the reviewer agrees with your recommendations, this should be a talking point in the discussion, not in the results. You can state that no other visualization methods were used in RCT’s, to give better highlight to this within the results.

Discussion
I commend the authors for giving specific recommendations to improve open and transparent reporting.
The limitations are comprehensive.

REVIEWER	McCrum, Christopher Maastricht Universitair Medisch Centrum+
REVIEW RETURNED	20-Jun-2022

GENERAL COMMENTS	I thank the authors for their responses to my comments and their revisions to their manuscript. My comments have been addressed and I have nothing further to add.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2, Dr. Christopher McCrum, Maastricht Universitair Medisch Centrum+

Comments to the Author:
I thank the authors for their responses to my comments and their revisions to their manuscript. My comments have been addressed and I have nothing further to add
Author reply:
⌡ We are glad that we were able to address the reviewer's comments to his satisfaction and thank the reviewer for his valuable contribution to improve our manuscript.

Reviewer: 1, Dr. Garrett Bullock, University of Oxford

Comments to the Author:
General Comments
This study investigated the transparency and reporting of research practices in sports and orthopaedics. This is an important endeavor, as this is essential for moving the profession forward and creating improved methods and repeatability of these methods.
Please see specific comments below.
Author reply:
⌡ We thank the reviewer for his positive feedback and the helpful comments to further improve our manuscript. We hope that we could address these comments to his satisfaction.

Introduction
The second half of the first paragraph is a little confusing. It is not clear if this information on allocation concealment, randomization, etc, was from your study or a previous study. Please clarify.
Author reply:
⌡ We thank the reviewer for this suggestion. We rephrased this sentence slightly to make clear that we are talking about previous studies at this point:
“While reporting has improved over time, major deficiencies that can impair translation are still common (11,12). These previous studies show that details needed to assess the risk of bias were missing from many published trials. More than half of all trials failed to address allocation concealment, and almost one third of studies did not address blinding of participants and personnel (12).

Overall, the introduction has very good information and a thorough literature review has been performed. However, I think the argument needs to be made clearer why this specific study needs to be performed. The authors give good information on the past studies that have investigated poor

methods; but, detailing where there are gaps for this meta-research study will strengthen the argument for this study. Are the studies not comprehensive? Or missing specific sports? Too old with the large amount of publications that have been disseminated recently?
Author reply:
⌡ We appreciate the advice by the reviewer in this regard. All mentioned aspects hold true, earlier studies are either from different fields, focus on a few specific aspects of reporting, or are not recent enough to remain relevant in the fast evolving field of reproducible research methods and reporting. Due to word count restrictions, we kept the section on the gap in the literature rather short and concentrated on providing more basic background knowledge. Following the reviewers comment, we added the following sentences to the manuscript:
“However, these results are only available for a relative narrow set of criteria, and it is unclear whether whether these results are still applicable in recently published literature and for a broader range of journals.
[…]
Yet, available studies either lack currency, assessed a small number of criteria or are not specific to orthopedics and sports medicine. Comprehensive data on current reporting practices of orthopedics and sports medicine clinical trials are lacking.”

Methods
I commend the authors for registering on OSF.
Sorry if I missed this. While this is a meta-research and ethics are not needed, can you add a statement explicitly saying this?
Author reply:
⌡ The statement on ethics approval is inluded in the manuscript on page 25:
“Research Ethics Approval
This is a meta-research study that does not involve human or animal participation and did not require ethical approval.”

Why only 9 months for exploration (instead of a year, or other time measure), can this be justified by previous references?
Author reply:
⌡ We started in August 2020 and went monthwise backwards until the desired sample size was approximately reached. This procedure is described in more detail in the supplemental file under “Sample Size Calculation”.

Concerning search strategy: it is not quite clear what the exact search strategy used for this study. The date and the use of the top 25% of journals is apparent? But the search strategy is not clear.
Author reply:
⌡ The search strategy, including the search string, search engine and used filters is described in more detail in the supplemental file under “Sample selection and screening process”.

Page 9, space needed between ‘amongjournals’
Page 9; line 28: misspelling of ‘aee’ ..are?
Author reply:
⌡ We thank the author for his careful reading. We corrected the spelling errors as suggested.

In inclusion exclusion criteria section, the screening process and inclusion/exclusion criteria are mixed together, which makes it hard to follow. Suggest making two separate sections. Further, more detail is needed for the inclusion/exclusion section.
Author reply:
⌡ We understand the reviewers’ perspective in this regard. However, given the overall structure of the manuscript and that there relatively few exclusion and inclusion criteria, we would prefer to not further

subdivide this section. Inclusion and exclusion criteria are often closely linked and it is not uncommon to address these in a single section.

I commend for examining open science practices in addition to CONSORT. The BIH reference (36) which was referenced for open science, the link was an error. I commend the authors for reporting protocol deviations. Page 10, line 50; space needed between 'the first'
Author reply:
⌡ We thank the author for his careful review. We corrected these errors.

Can you explain further why if the outcome was changed after the study began that this was deemed a retrospective study? Would this be more deviation from the protocol, or a posteriori hypotheses creating (if they did the change to make the study 'better'), or something else?
Author reply:
⌡ This is an important question. We would like to clarify that only the registration was classified as retrospective; not the study itself. This classification was made any time the study was registered after the first patient was recruited, regardless of whether the authors openly declared protocol deviations. There are many good reasons to change parameters of a registered study after its registration, but there is also the possibility that these changes are motivated by a desire to make results look more interesting (cherry-picking, selective reporting). Objectively differentiating between these scenarios is usually not easily possible by only inspecting the published trials and their registrations. Therefore we simply assessed whether the registration was completed before enrollment of the first participant (https://embassy.science/wiki/Theme:5e34933a-293e-447a-9ab4-9299a152e8a5), which is in accordance with the standard definition of prospective registrations and the classifications in trial registries like the ANZTR or DRKS.

Results
It may be recommended to include referencing for all included studies within the results.
Data visualization: The authors give recommendations for data visualization within the results. While the reviewer agrees with your recommendations, this should be a talking point in the discussion, not in the results. You can state that no other visualization methods were used in RCT's, to give better highlight to this within the results.
Author reply:
⌡ This meta-research study included 163 trials and it would not be feasible, or informative, to reference all of them within the paper. There is no discussion of the details of individual trials included in the study; we only present summary statistics on the quality of reporting for the entire sample. However, interested readers can find information on all included studies in the open data file, deposited on the data repository.

We thank the reviewer for this important advice regarding the data visualization section. As suggested, we removed the recommendation sentence from the results section. This information is already captured in the discussion section.
"Bar graphs were used to display continuous data in 21% (CI 15-21%; n=34) of trials."

Discussion
I commend the authors for giving specific recommendations to improve open and transparent reporting.
The limitations are comprehensive.