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)	The effect of a concussion on subsequent baseline SCAT performance in professional rugby players: a retrospective cohort study in global elite Rugby Union
AUTHORS	Tucker, Ross; Falvey, Eanna; Fuller, Gordon; Brown, James; Raftery, Martin

VERSION 1 - REVIEW

REVIEWER	Todd Lyons Harvard Medical School/Boston Children's Hospital
REVIEW RETURNED	08-Mar-2020

GENERAL COMMENTS	Summary: In this analysis, the authors leverage a large database of SCAT testing performed on professional rugby players. They compare baseline SCAT performance between those players who were diagnosed clinically with a concussion between baselines (cases) to those who had no concussion (controls). The authors found those players diagnosed with a concussion endorsed fewer symptoms and had improved Digits Backward and Final concentration scores. The controls had improved performance in tandem gait. The authors note multiple plausible explanations for these findings including familiarity with the test. The results of these findings support that baseline neurologic testing on the SCAT is not clinically significantly affected by an intercurrent concussion and have implications for how we approach baseline testing for concussed athletes. Major Issues to Address: 1) As the manuscript is currently written cases appear to have to only have one baseline SCAT on either side of a concussion whereas controls had to have to 2 baselines. However, from the main analysis where change in scores were compared between cases and controls it seems to suggest that cases also had baselines before and after a concussion. This requires further clarification. Minor Issues to Address: 1) Consider just spelling out cases and controls in the manuscript. 2) Abstract (Participants): Consider writing out “Controls” the first time it appears before the abbreviation (CONT). 3) Abstract (Results): Once you have defined cases and controls consider using this terminology in the results. “Cases diagnosed with a concussion endorsed fewer symptoms…” 4) Abstract (Conclusion): I was waiting for a “so what” statement. It

	appears these results support continued use of baseline SCAT testing as a reliable marker of a players true baseline even after a concussion. 5) Introduction: The introduction is quite lengthy, but describes adequately the reason why this study is important. However, consider moving the third paragraph which contextualizes the magnitude of the issue of concussion in rugby to the first paragraph. 6) Methods (Aims): Define a “recent previous concussion in this section. 7) Methods (Study Design, Setting and Study Population): It is unclear to me that this is a cross-sectional study. It appears there is an exposure (concussion vs. no concussion) and an outcome (change in baseline SCAT vs. no change in baseline SCAT). Since you observed the difference after the exposure it appears to be a retrospective cohort study. 8) Methods: I comment the author for using a Bonferroni correction when looking for significant changes between groups. This strengthens their findings as not just being a result of chance alone. 9) Results: The results have often duplicated findings between the text and tables. Consider just summarizing findings briefly in text to support what is written in tables. 10) Results: The authors not referring to Table 1 that “No differences were observed for any sub-mode between CASES and CONT at PRE/T1.” Data to support no difference (p value, difference with 95% CI) are not shown to support this finding. 11) Discussion: The current first paragraph of the discussion just reiterates the already discussed Aims of the study. Consider using this to discuss the overall findings. 12) Discussion: It is interesting that balance was the only outcome in which cases performed more poorly than controls (albeit small), given that this is an objective measure (compared to symptom reporting) and is unlikely something individuals can practice or improve with familiarity. May be worth exploring this finding more.

REVIEWER	Stephen Carek, MD, CAQSM Prisma Health/University of South Carolina - Greenville Greenville, South Carolina
REVIEW RETURNED	11-Mar-2020

GENERAL COMMENTS	This was an interesting study which evaluated variations of SCAT5 baseline testing for rugby athletes who did and did not sustain a concussion during a set period of time. The authors conclusions suggested that those athletes who sustained a concussion during a season reported fewer symptoms and lower symptom severity score on their next baseline assessment, as well as improved several sub-mode tests. I do have a few questions regarding some components of this study: 1. Are the authors able to provide more detail about the population they are studying? What are some baseline demographics, history of concussions outside of the study period, and possibly previous SCAT testing (outside the study period noted). 2. For those athletes who did sustain a concussion, was there documented return to sport/improvement in scores before they underwent a repeat baseline test (particularly those who sustained a concussion 3 mo or less before the repeat baseline)? 3.Reference 5 is cited a good bit in the discussion, is there other literature to support the author's suggestions that familiarity and

	being accustomed with concussion symptoms affects reporting in the SCAT 5. Is this reproduced in other or athlete populations sports? 4. Can the authors elaborate more on the clinical implications of this? Do these changes validate/refute the need for repeat baseline testing? Or should a focus be on getting new baseline testing on those with a recent concussion?

REVIEWER	Professor Patria Hume Sport Performance Research Institute New Zealand, Auckland University of Technology, New Zealand I have previously conducted research on retired player health, and protective equipment, for World Rugby. I have co-authored technical reports with Martin Raftery for these projects for World Rugby.
REVIEW RETURNED	29-Mar-2020

GENERAL COMMENTS	Thank you for an interesting paper. However, I have suggestions to improve the paper. I have used comments within the pdf to make your revisions easier. For example: "Why is there a gap in the cases on the symptom figure? Explain in a footnote to the figure why there are missing values for the CASE DIST group (I assume) and why only for the symptom number and not the digits backwards and single leg errors." "The conclusion states only two of the variables instead of the larger number in the results. It is also not clear the direction of the changes. What the reader wants to know is the answer to the aim of "whether concussion affects subsequent baseline performance in professional rugby players" and HOW the baseline performance is affected - i.e. what variables from the SCAT got better. If familiarity is the reason - given the HIA1, 2, 3, then the T2 - make it more clear that the CONC have 3 extra SCAT assessments compared with the CONT. If there is no clinical significance of the results - then what is the variation in the SCAT that a clinician can consider not meaningful. I wanted to see what the variation is for a variable to be able to say the SCAT is affected by the prior concussion or not. I don't have a clear outcome that is meaningful to practice of using the SCAT - and the question that comes up is given the CONC group got better scores is the SCAT actually sensitive enough and should it be used. Clear recommendations to address these points are needed in the abstract, and more detailed discussion of the points with reference to test-retest (within weeks not a year) is needed in the main text discussion please." "This conclusion is just a restatement of the results. Please reword to provide the answer to the stated aim as indicated in the first paragraph of the discussion which is "whether this might have implications for the clinical utility and interpretation of return-to-play and diagnostic screens in the season after the concussion". "Annual baseline differences are 334 days for the controls. There would not be an expectation that this length of time would be subject to a learning effect. Please provide rationale for this explanation referring to the literature within the main text of the manuscript."

	"Re word this section so it is clear that there were two baseline annual tests taken as T1 and T2 for both the CONT and the CONC. The CONC sustained a concussion between T1 and T2 whereas the CONTROL did not sustain a concussion. There is no need to use the confusing PRE and POST versus the T1 and T2. On first reading given the HIA 1, 2, 3 information I thought analyses were going to be of the changes in the SCAT over the T1, HIA1, HIA2, HIA3 and T2 times, but this was not the case. Only annual T1 and T2 differences are presented." "It is not clear from this sentence whether you excluded a control/case that had different SCAT3 to SCAT5 baseline scores (as individuals), or whether you allowed a difference, but then compared them with a similar control/case. That is, T1 was SCAT3 for both case/control, and T2 was SCAT5 for both case/control. Please provide further details in the introduction on the differences in means and SD for the SCAT measures when going from SCAT3 to SCAT5." "Please be consistent in using CONT for the control group once you have specified the abbreviation the first time. See highlighting in these two paragraphs that shows the three ways the same group has been named." "Delete the median values as they do not add any more useful information than the mean and SD. Removal of the medians will also allow the merging of tables 1 and 2, which would enable easier reading of the raw values and then the change scores." "In the methods and tables please state if the two 5–word groups for a total of 10 words per trial was used or not as per SCAT options. If unknown state unknown. The variation in method will increase the mean and SD scores." "For ease of reading please delete the % after each value in the column as you only need to provide it in the header of the column." "Why has figure 1 only provided a few of the variables - with no rationale as to why those ones were chosen?" "Remove the non significant p values from the text. Keep in the tables. You need the Bonferroni adjusted p value threshold stated in the table footer, so it is clear for people that the usual $p<0.05$ is not the significance threshold. Otherwise when the reader sees not sig at a P=0.014 they will think it is an error)." "Remove the capital C for Concussed players as there is no need for capitalisation, and it is not used for the "POST concussion group" for example." "Throughout the text simplify to T2 and T1 for CONT and CONC. There is not need to have a PRE and POST and it makes the reading of the text confusing. Change "were unchanged POST vs PRE concussion, and T2 vs T1" to "were unchanged T2 vs T1 for both CONT and CONC groups". " "Use CONC throughout not CASE. " "In Table 3 please remove case from this column and replace cohort with CONC."

	"Given the sample sizes are small at 18 or 20 in several of the analysis, is the power too low for this type of analysis for those variables and they should be deleted from the table?" "This highlighted sentence should be deleted given the first sentence re no significant differences. You have a threshold for significance, so only report as either significant or not." "Again delete the tendency paragraphs... focus on the threshold of significance you have chosen." "Given these aims please clearly report in the results and the discussion: 1. CONC versus CONT at T1 = no changes in any SCAT scores (I would hope). 2. CONC versus CONT at T2 - a clear difference - resulting in the need to understand why - which it seems the CONC are BETTER than the CONT at T2 - so the question is therefore is it familiarity due to the extra SCAT tests during the HIA assessments. Therefore show the CONC SCAT scores for T1, HIA1, HIA2, HIA3 and T2. Only if this figure shows a clear clustering of those with recent or distant concussions - using cluster analysis to therefore set the threshold for the time for recent - then provide the recent versus distant concussion analyses." "This paragraph is the key one that outlines why the study has been conducted. It needs to go into the introduction of the paper." "Here you have acknowledged the small sample size - so then why present the data given such limitations?" "Id like to see better recommendations at the end of each section. Should clinicians use the number of symptoms as a useful part of their assessment or not. What number of symptoms should be the threshold? Does a change of 1, 2, 3 etc matter or not?" "I do not like the "post-concussion" terminology used in this paper as it implies and immediate post concussion assessment - ie HIA1, 2, or 3. It needs to be clear that the T2 time is being talked about which is when players have been cleared to return to play from their prior concussion and this is on average *#* days." "There is too much repetition of results being reporting in the discussion. Focus on the key points and relate to other literature to build a case as to whether the SCAT and the sub-components should be used or not based on the variation in baseline testing at T1 and T2, and the effects of repeated HIA testing for the CONC participants." "This is a limitation as the rationale is that repeat exposure to SCAT is affecting the T2 results compared with the T1 results - yet there is no indication of the number of prior tests. The number of SCAT tests must be in the WR database? Can't these analyses be conducted looking at SCAT changes over time for players?" "There seems to be too many limitations to make any meaningful interpretation of the differences between the two baseline annual tests useful." "Unfortunately it is just not clear what the magnitude of the changes for each of the variables should be considered useful clinically given potential learning effects etc."

	"WR clearly have this data for the HIA1, 2, and 3. These data and analyses should be provided for the players in the CONC group, so the short term variation in the SCAT due to concussion can be shown. This would then hopefully show that the SCAT is sensitive to concussion, and that the T1 and T2 SCAT scores are better than the scores during the short term concussion time. Provide a graph for the CONC group of the T1, HIA1, 2 3, then T2 scores, so we can see how SCAT changes. This would provide more useful information to enable interpretation of the T2 changes for CONC and CONT." "Please provide a final conclusion that addresses the key aim of "whether any SCAT5 sub-modes were affected by a previous concussion, and whether this might have implications for the clinical utility and interpretation of return-to-play and diagnostic screens in the season after the concussion." The reviewer provided a marked copy with additional comments. Please contact the publisher fo full details.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name

Todd Lyons

Institution and Country

Harvard Medical School/Boston Children's Hospital

Please state any competing interests or state ‘None declared’:
None

Please leave your comments for the authors below
Summary: In this analysis, the authors leverage a large database of SCAT testing performed on professional rugby players. They compare baseline SCAT performance between those players who were diagnosed clinically with a concussion between baselines (cases) to those who had no concussion (controls). The authors found those players diagnosed with a concussion endorsed fewer symptoms and had improved Digits Backward and Final concentration scores. The controls had improved performance in tandem gait. The authors note multiple plausible explanations for these findings including familiarity with the test. The results of these findings support that baseline neurologic testing on the SCAT is not clinically significantly affected by an intercurrent concussion and have implications for how we approach baseline testing for concussed athletes.

Thanks for your review and the issues you’ve raised. We’ve made pretty much of all of them, and below are some responses to those, more detailed explanations. I think the paper is improved as a result, so thanks for the input

Major Issues to Address:
1) As the manuscript is currently written cases appear to have to only have one baseline SCAT on either side of a concussion whereas controls had to have to 2 baselines. However, from the main

analysis where change in scores were compared between cases and controls it seems to suggest that cases also had baselines before and after a concussion. This requires further clarification.

Thanks for your review and the suggestions and questions asked, which we respond to here, and make changes to the manuscript that we hope improve its quality.

With respects to this first question, I think it's just a misunderstanding. When we wrote "either side of a concussion", what we mean is that there is a SCAT baseline test BOTH before and after the concussion. As in for instance if a football team scores a touchdown either side of half-time, it would have scored both before and after. However, we appreciate that "either" might also be taken to mean one side or the other, and so we've gone into the manuscript and changed this, so that it now reads:

For the purposes of the present study, cases were identified as all professional players who had a baseline screen both before and after a documented concussion between 2017 and 2019 (CONC T1 and CONC T2).

Hopefully this makes it clear that every person included in the study had TWO screens, a T1 and T2 screen, with the difference being that in CONC, a concussion occurred between them.

We have also, under the advice of other reviewers, tried to change the abbreviations used, so that PRE and POST are now replaced by T1 and T2, and that cases are now referred to as CONC and Controls as CONT. So we effectively have CONC T1 vs CONC T2, and CONT T1 vs CONT T2, rather than PRE and POST and a mix of CONC, CASES, CONT. We hope this is more logical and easier to understand now.

Minor Issues to Address:
1) Consider just spelling out cases and controls in the manuscript.

Hopefully this is now dealt with adequately under the change described above. Other reviewers advised conclusion with the use of PRE, POST, CASE, CONC and CONT, so we have reduced it right down to T1 and T2, always meaning pre and post, and CONT and CONC, meaning controls and concussions.

2) Abstract (Participants): Consider writing out "Controls" the first time it appears before the abbreviation (CONT).

Done, thanks

3) Abstract (Results): Once you have defined cases and controls consider using this terminology in the results. "Cases diagnosed with a concussion endorsed fewer symptoms…"

We've done this, though we've using CONT and CONC when using the abbreviations, and where appropriate, we have written it as suggested – "Cases (or players, in some instances) diagnosed with concussion", thanks.

4) Abstract (Conclusion): I was waiting for a "so what" statement. It appears these results support continued use of baseline SCAT testing as a reliable marker of a players true baseline even after a concussion.

Yes, thanks, this was also suggested by another reviewer, and it's clear we didn't 'pull the trigger' on a conclusions, so to speak. We've gone back and tried to be more decisive. Your interpretation is

correct, and we've written a new conclusion to the abstract (and this is reflected in the paper) that hopefully achieves this.

5) Introduction: The introduction is quite lengthy, but describes adequately the reason why this study is important. However, consider moving the third paragraph which contextualizes the magnitude of the issue of concussion in rugby to the first paragraph.

Done, thank you

6) Methods (Aims): Define a "recent previous concussion in this section.

Done. It has been defined as a concussion occurring within a year or season between two annually conducted baseline screens. We excluded concussions that occurred more than year prior to T2, since the two SCAT5s that would be compared in this way would be more than two years apart. This is explained in the methods.

7) Methods (Study Design, Setting and Study Population): It is unclear to me that this is a cross-sectional study. It appears there is an exposure (concussion vs. no concussion) and an outcome (change in baseline SCAT vs. no change in baseline SCAT). Since you observed the difference after the exposure it appears to be a retrospective cohort study.

You're quite right, it is. We had previously analysed some other SCAT data that was cross sectional and I absent-mindedly just retained the description for this study. I've corrected it as suggested, thank you.

8) Methods: I comment the author for using a Bonferroni correction when looking for significant changes between groups. This strengthens their findings as not just being a result of chance alone.

Thank you. While 'frustrating' to have so many findings that fall under the 0.05 threshold, we felt it necessary to apply the correction.

9) Results: The results have often duplicated findings between the text and tables. Consider just summarizing findings briefly in text to support what is written in tables.

Thanks, we've tried to scale back on the text in the revised manuscript. The same is true in the discussion, which had too much repetition of the results. So the revised paper now has fewer words in the results. We believe this makes it easier to follow, and cleaner.

10) Results: The authors note referring to Table 1 that "No differences were observed for any sub-mode between CASES and CONT at PRE/T1." Data to support no difference (p value, difference with 95% CI) are not shown to support this finding.

You're right, we didn't shown the 95% CI in the original. Our approach on that first analysis was to analyse the medians, where we found no statistical differences. Of interest, we now analysed the means, as prompted by your question, and report the 95% CI, and that the Immediate memory scores were greater in CONC than CONT both at T1 and T2. This is now also mentioned in the discussion, though it does not change the overall finding that IM scores are unaffected by a concussion relative to a control (as per Table 2). The addition of the 95% CI evaluation also highlights the differences we subsequently explain in the rest of the results, namely the lower symptoms at T2, and the higher cognitive scores at T2.

11) Discussion: The current first paragraph of the discussion just reiterates the already discussed Aims of the study. Consider using this to discuss the overall findings.

Yes, this is a stylistic thing, in a way – some authors like to restate the aim to frame the discussion. We've changed it as recommended, though we still feel it adds some value to re-orient the reader as to the purpose of the study, so we leave the first section of the paragraph in, but have tried to bridge faster to the discussion points and key implications, and hope that we've done so acceptably.

12) Discussion: It is interesting that balance was the only outcome in which cases performed more poorly than controls (albeit small), given that this is an objective measure (compared to symptom reporting) and is unlikely something individuals can practice or improve with familiarity. May be worth exploring this finding more.

Indeed, that is an important element of balance sub-modes that others do not have. We are mindful of the length, so we didn't go into it in much more detail, but we have shortened the manuscript in the results section, and so we have, in the revised manuscript, devoted a paragraph to at least mentioning this factor and highlighting the element you raise.

Reviewer: 2

Reviewer Name

Stephen Carek, MD, CAQSM

Institution and Country

Prisma Health/University of South Carolina - Greenville
Greenville, South Carolina

Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below
This was an interesting study which evaluated variations of SCAT5 baseline testing for rugby athletes who did and did not sustain a concussion during a set period of time. The authors conclusions suggested that those athletes who sustained a concussion during a season reported fewer symptoms and lower symptom severity score on their next baseline assessment, as well as improved several sub-mode tests. I do have a few questions regarding some components of this study:

Thank you for your review, and the questions you've asked below. We have attempted to address these in as much detail as possible, and incorporate your comments in your manuscript revision. Unfortunately, some of what you have asked is 'unknowable", and we recognize that these are limitations of the study. Nevertheless, between our responses here, and the resultant revisions, we hope that you will consider the revised manuscript suitable for revision, even with some limitations!

1. Are the authors able to provide more detail about the population they are studying? What are some baseline demographics, history of concussions outside of the study period, and possibly previous SCAT testing (outside the study period noted).

Unfortunately not. It has long been a limitation, with data protection and logistic issues standing in the way, that we are not able to gather information on the language, or age, or playing experience of the cohort. The HIA process used in the sport does ask the clinicians to indicate the player's date of birth, but this is omitted in well over half of cases. It might be possible, if the data can be de-identified sufficiently, to manually insert the ages of players, and possibly even to go into archives and study their playing careers. However, we believe that this would provide a massively disproportionate investment for the return.

Put simply, given that the HIA process is followed globally, the governing body for the sport needs to have a policy that can be implemented globally. Part of the reason we have never "aggressively" explored getting all these demographic data is because while it would have substantial academic interest, it would not likely change practice too much, given that players from all nations, and at all ages, ethnicities, concussion histories etc would still have to be treated by the same HIA process. So unless there was good reason to stratify players by age, concussion history, etc, and treat them differently, World Rugby would not pursue this given its complexity – it is in effect a pay-off between getting incredible data (which I would naturally love) and getting a process that is actually followed and serves its purpose for player welfare. Finding this balance is indeed tricky!

We have acknowledged this limitation in the limitations section of the paper, and we recognize that future research may look to plug some of these gaps in knowledge and context to findings like those in this study.

In terms of including players who have performed a range of baseline screens between 1 and 4, we chose to include them all because again, we want to have a finding that we can apply to our jurisdiction, of global rugby, and we recognize that moving forward, two players, playing for the same team in say, France or England, may be vastly different in age, experience, concussion history and familiarity with the SCAT, but they need to (for now, at least) be treated the same way by the HIA process.

In that regard, that this study is part of a batch of studies seeking to answer these questions – one of our companion papers is in fact an exploration of whether the SCAT results are different in a player who has done two of them, compared to a player who has done four or five. This "learning effect" study is thus one being considered, but we wanted to make this a paper that basically answers "Does a player who has experienced a concussion present with a different baseline result the following season?", and all those demographic factors are certainly important factors contributing to any finding, possibly explaining them, but unfortunately, we don't have them at this stage.

This is a long answer, apologies, to a question whose simple answer is "no", but I wanted to explain why it was and what the future may hold in terms of research.

2. For those athletes who did sustain a concussion, was there documented return to sport/improvement in scores before they underwent a repeat baseline test (particularly those who sustained a concussion 3 mo or less before the repeat baseline)?

By definition within the HIA process and the General Return to Play protocols that teams are obliged to follow in the sport, yes. The problem for us is that we don't have access to these scores, as they belong to teams. Every player who is concussed enters a graduated return to play process, which precludes them from playing again until their symptoms are normal, and they report with no abnormalities. The follow up baseline screening (T2 in our manuscript) is conducted at the start of the following season, so in every single case, a concussion occurring in Season 1 would necessarily be followed by a full recovery prior to a return to play, and then by an off-season period of at least two

months, prior to the T2 baseline assessment. We do not have access to these medical screening processes that are administered within each club, unfortunately, so we can't for instance track how long it takes to return to normal after the injury. All we are able to assess is what is done within the World Rugby system for baseline and in-match diagnostic screening.

We did try, incidentally, to find cases where the concussion was very recent, occurring within a week or two of T2, but unfortunately no such cases exist, because as explained above, the baseline tests are invariably done after a long period without play, during pre-seasons. There are occasional situations where a player gets a baseline screen during a season, and these are the cases where the gap is smaller (our REC group), but none of these were concussed players within 2 weeks of T2.

3.Reference 5 is cited a good bit in the discussion, is there other literature to support the author's suggestions that familiarity and being accustomed with concussion symptoms affects reporting in the SCAT 5. Is this reproduced in other or athlete populations sports?

Yes, indeed, perhaps over-reliant on reference 5. I have included other references that support possible learning effects, in the revised manuscript.

4. Can the authors elaborate more on the clinical implications of this? Do these changes validate/refute the need for repeat baseline testing? Or should a focus be on getting new baseline testing on those with a recent concussion?

Upon reading all three reviews, it is clear that we did not "pull the trigger" enough on the implications. We were to some degree being circumspect, because while we have some provocative findings, which we believe to be important, we also know that things like the demographic factors, the influence of time between injury and T2, and the history combined with complexity of recovery, make it difficult to change the policy. So in a way perhaps typical of the speed of policy change, we hesitate to make strong conclusions around implications.

However, prompted by the reviews, we have tried to do this. In the Discussion, at the end of each sub-section, there is now a new concluding paragraph (highlighted red in the revision), that tries to more strongly conclude what we would advise. Briefly, we do not think that the scale of the changes in symptom report or cognitive performance is large enough to change the process of annual baseline screening, and so we recommend that it continue, but that doctors be mindful that improved scores (fewer symptoms and higher cognitive scores) be explored in detail, given how they may betray a player who is deliberately underreporting symptoms or is more motivated to perform better in cognitive tasks. This may take the form of follow up, in-depth screens, and possible repetition of cognitive tests using a different number of word sequence.

We feel this is as far as one can go on the practical advice, because we believe that the threshold to change the process needs to be quite high if that change has any risk of undermining player welfare.

In terms of the tool, we believe our data confirms the need for more difficult cognitive tasks, or more options for number sequences, and include this in the paper too.

We hope that these three concluding paragraphs in each section are stronger and commit us to a position around baseline testing, with practical advice for what the findings may mean when it is conducted.

Reviewer: 3

Reviewer Name

Professor Patria Hume

Institution and Country

Sport Performance Research Institute New Zealand, Auckland University of Technology, New Zealand

Please state any competing interests or state ‘None declared’:
I have previously conducted research on retired player health, and protective equipment, for World Rugby. I have co-authored technical reports with Martin Raftery for these projects for World Rugby.

Please leave your comments for the authors below
Thank you for an interesting paper. However, I have suggestions to improve the paper. I have used comments within the pdf to make your revisions easier.

Thanks for your time in reading and reviewing the paper so extensively. We have gone through each one in turn, and below outline our responses, and also explain where changes you suggest have been made to the manuscript.

Many of the changes or comments are addressed as part of the other reviewer comments, and many others have been dealt with by changes in response to yours (for instance, rewriting sections of the discussion has removed sections of the text that were previously highlighted for a comment).

Below follows the responses, and some clarification where possible. Thank you.

For example:

"Why is there a gap in the cases on the symptom figure? Explain in a footnote to the figure why there are missing values for the CASE DIST group (I assume) and why only for the symptom number and not the digits backwards and single leg errors."

That was an error, an artefact of the way the data was set out in the graphing programme used to draw the graphs, sorry. Because we had varying numbers of cases for symptoms (which had to be assessed in a SCAT5, so n = 143) and the digits backwards and single leg errors (which can be compared between SCAT3 and SCAT5, so n = 501), the end result was a table that was “staggered” with a gap of 368 rows between the sub-groups for symptoms. When graphed, that gap appears as the space you see in the figure. All it required was to take exactly the same data and simply change the arrangement in the graph software, and the appearance of the graph was made similar to the other two. It was not, therefore, a function of missing values – just rearrangement of existing values. The new Figure 1 is included in the revision.

"The conclusion states only two of the variables instead of the larger number in the results. It is also not clear the direction of the changes. What the reader wants to know is the answer to the aim of "whether concussion affects subsequent baseline performance in professional rugby players" and HOW the baseline performance is affected - i.e. what variables from the SCAT got better. If familiarity is the reason - given the HIA1, 2, 3, then the T2 - make it more clear that the CONC have 3 extra SCAT assessments compared with the CONT. If there is no clinical significance of the results - then what is the variation in the SCAT that a clinician can consider not meaningful. I wanted to see what the variation is for a variable to be able to say the SCAT is affected by the prior concussion or not. I don't have a clear outcome that is meaningful to practice of using the SCAT - and the question that

comes up is given the CONC group got better scores is the SCAT actually sensitive enough and should it be used. Clear recommendations to address these points are needed in the abstract, and more detailed discussion of the points with reference to test-retest (within weeks not a year) is needed in the main text discussion please."

We appreciate that all these variables are of interest, and perhaps we were too mindful of word limits in the abstract and produced an excessively limited and then only high-level summary of the data. We have gone back and revised the abstract in an attempt to bring out these important findings.

With reference to the suggestion about detailed discussion of the "points with reference to a test-retest (within weeks not a year", this data unfortunately does not exist. We specifically looked for what the smallest gap was between a concussion and a follow up baseline screen, and it's not less than two weeks. Indeed, this is almost by definition, because the baseline screening is pre-season, and the concussions happen after them, in-season. We had hoped to find some cases of international players who might receive baseline screens when in international camps, possibly shortly after a concussion, but these cases did not exist. Further, there is a conceptual problem with using concussions that are very close to baseline screens, because then you have the prospect that a still-symptomatic player is being assessed, and so whatever differences are found in T2 compared to T1 might be the result of both learning effects and the actual brain injury. Our analysis of REC vs DIST cases was an attempt to explore this, but as you've noted, it's limited by size. So even though we have a relatively large cohort for analysis, we would lose statistical power when we begin to constrain time frames too much.

The end result is that we do not have a study design that allows us to tease out the size of the learning effect, or the size of the change that would be clinically meaningful. One of the outcomes of this (and other) analyses is the recognition that a specific study may be required to assess this.

The discussion section of the revised manuscript has been added to, and there are now conclusion paragraphs in each sub-section that we hope tackles most of the issues you raised above. In particular, we have concluded each sub-section with clinical implications and recommendations that we hope are stronger, more "commital".

We do have some hesitation in committing to a specific value for a sub-mode change that might be clinically significant. This is an important question, we agree, but it's not answerable within this study design (I don't know of a study that has answered it, in the context of prior concussion). In the discussion, we do have a section that explains that we're not in a position to assess the magnitude of the changes that may be attributed to learning/experience, and the same concept is true of the magnitude of what a clinically significant change is. Until such time as we know what normal variation of repeated SCAT5 results is (when tests are a year apart), and how a concussion compares, it will be impossible to commit to a value that answers this.

We have mentioned this in the discussion of the paper, along with reference to a study we have recently published where doctors frequently apply clinical judgement to overrule abnormal sub-mode tests during HIA1 screens. This would suggest, and this has always been acknowledged, that the screen is simply a guide for the clinician and is not really intended to be so sensitive or specific as to dictate a diagnosis.

In any event, the long and short of it is that an assessment of how large a change needs to be in order to be deemed significant is not in the scope of this paper. We can and have speculated, particularly in the revised manuscript, that the typical and average changes we see (one symptom, 0.23 digits backwards) are likely NOT meaningful, and thus do NOT necessitate a change in process or clinical practice, other than that a clinical now needs to be mindful that improved scores might

reveal learning, motivation and 'deception' effects, and consider that follow up or repeat testing should be conducted, and we hope that is sufficient.

"This conclusion is just a restatement of the results. Please reword to provide the answer to the stated aim as indicated in the first paragraph of the discussion which is "whether this might have implications for the clinical utility and interpretation of return-to-play and diagnostic screens in the season after the concussion".

The discussion has been rewritten now, trimmed significantly, and with more emphasis on practical conclusions and implications, which we think speaks more to the aims of the study.

"Annual baseline differences are 334 days for the controls. There would not be an expectation that this length of time would be subject to a learning effect. Please provide rationale for this explanation referring to the literature within the main text of the manuscript."

Players regularly report that they learn the sub-modes in preparation for the screens, because they know that they'll be assessed and any diagnostic results interpreted against those screens, so there is a possibility of a learning effect. What needed to be explained better is that it's not necessarily the once-off screen at each season's baseline testing day that creates this and we have amended this section to make this clearer.

"Re word this section so it is clear that there were two baseline annual tests taken as T1 and T2 for both the CONT and the CONC. The CONC sustained a concussion between T1 and T2 whereas the CONTROL did not sustain a concussion. There is no need to use the confusing PRE and POST versus the T1 and T2. On first reading given the HIA 1, 2, 3 information I thought analyses were going to be of the changes in the SCAT over the T1, HIA1, HIA2, HIA3 and T2 times, but this was not the case. Only annual T1 and T2 differences are presented."

Done. The section in the methods that describes the cohorts has been rewritten, in response to previous reviews also. We think it is now clearer that T1 refers to the initial baseline test, T2 to a second test approximately one year later, which either follows a concussion in that intervening period, or not, in CONC and CONT, respectively.

We have also, throughout the document, referred only to CONC and CONT, and to T1 and T2.

"It is not clear from this sentence whether you excluded a control/case that had different SCAT3 to SCAT5 baseline scores (as individuals), or whether you allowed a difference, but then compared them with a similar control/case. That is, T1 was SCAT3 for both case/control, and T2 was SCAT5 for both case/control. Please provide further details in the introduction on the differences in means and SD for the SCAT measures when going from SCAT3 to SCAT5."

This section has also been rewritten in the revised manuscript, and hopefully now clarifies what you are asking. I was not 100% sure of your query here, but I think I understand it, and basically, what sometimes happened is that a player's T1 was done as a SCAT3, and their T2 was a SCAT5. This meant that some of the sub-modes for that player could not be included in a paired analysis, because SCAT3 uses 5 word lists and SCAT 5 a 10 word list, and because SCAT 3 and SCAT 5 ask for symptoms in a different way.

As a result, we had to exclude players where the sub-mode was not the same at T2 as T1 and our sample size for those sub-modes is reduced. Our analysis of each sub-mode is only done where we have paired data – same length words, same SCAT form for symptoms. This is why out of 501

concussed players, we have 143 symptom results – many were assessed at T1 using SCAT3, and at T2 using SCAT5, which makes the T1 vs T2 comparison illegitimate.

Note that it's not always an issue of SCAT3 vs SCAT5 – there are some SCAT5s that also used a 5-word list, and so it was more on the basis of the same sub-mode being paired to allow comparisons. The new paragraph explaining that is:

In some cases, the baseline screens at T1 and T2 were different versions of the SCAT, which changed from SCAT3 to SCAT5 during the sampling period, or used a different word length list. For our analysis of word-list sub-modes (Immediate memory and Delayed recall) and symptom endorsement, we analysed only paired and directly comparable sub-modes, and thus exclude all cases and controls where a sub-mode was assessed differently in T2 compared to T1. We report the number of available paired comparisons in the results.

"Please be consistent in using CONT for the control group once you have specified the abbreviation the first time. See highlighting in these two paragraphs that shows the three ways the same group has been named."

Yes, thanks for this suggestion, we have made the change throughout. It's now CONC and CONT, T1 and T2. We did get into something of a mess trying to explain how PRE and POST were T1 and T2 depending on cases and controls, and thanks for the suggestion to clarify.

The only small concern is that CONT and CONC are very similar, three quarters the same, but it is still simpler than it was.

"Delete the median values as they do not add any more useful information than the mean and SD. Removal of the medians will also allow the merging of tables 1 and 2, which would enable easier reading of the raw values and then the change scores."

We believe that the median must be reported in the descriptive summary, since the subsequent analyses of change from T1 to T2 were used non parametric statistics by virtue of most sub-modes being non-normally distributed. So we believe that the median needs to be reported in order to show that the values were not different between CONC and CONT at the outset (T1 comparison). We agree it would make for an simpler single table, but statistically, we believe that we should report the median, otherwise the only basis for assessing the sub-mode scores is the change scores. Presenting the medians is also consistent with our previous studies (ours and other) report scores for these sub-modes.

"In the methods and tables please state if the two 5–word groups for a total of 10 words per trial was used or not as per SCAT options. If unknown state unknown. The variation in method will increase the mean and SD scores."

I am not 100% sure what you are enquiring here, sorry? The list of words used is provided by the SCAT, and is either 5 or 10 unique words, generally divided along the lines of a SCAT 3 having 5 and a SCAT5 having 10 (though not always, since some SCAT5s used 5 words in their early iterations). For our CONT group, we used only SCAT5s, which takes care of this problem, but for the CONC, the clinician conducting the baseline would have applied the instructions as per the SCAT they were using. As soon as 10 word lists were introduced, the clinician would have read 10 words as per the list provided in the SCAT, rather than the same list of five words, twice.

"For ease of reading please delete the % after each value in the column as you only need to provide it in the header of the column."

Done, thank you. This change was also made in Table 3.

"Why has figure 1 only provided a few of the variables - with no rationale as to why those ones were chosen?"

The figure is meant to be illustrative, because the method of analysis is effectively comparing the distributions of scores within each sub-mode. So really, we just wanted to provide an illustration of one symptom outcome, one cognitive sub-mode, and one balance sub-mode, because we feel the reader will be in a better position to conceptualize what the analysis is doing as a result. As it happens, we have presented the two sub-modes that vary significantly between CONT and CONC, and then with single leg sub-modes, the outcome most different between REC and DIST, but the main purpose was illustrative, and we hope that it will be appropriate to retain the figure. We did not, for instance, see any merit in including the other ten sub-modes that were NOT different in a paper already heavy on data, because Table 2 and Table 3 adequately present the results from those sub-modes. This is now explained in the revised manuscript, for clarity, thanks.

"Remove the non-significant p values from the text. Keep in the tables. You need the Bonferroni adjusted p value threshold stated in the table footer, so it is clear for people that the usual p<0.05 is not the significance threshold. Otherwise when the reader sees not sig at a P=0.014 they will think it is an error)."

Done, thanks.

"Remove the capital C for Concussed players as there is no need for capitalisation, and it is not used for the "POST concussion group" for example."

Done, thank you

"Throughout the text simplify to T2 and T1 for CONT and CONC. There is not need to have a PRE and POST and it makes the reading of the text confusing. Change "were unchanged POST vs PRE concussion, and T2 vs T1" to "were unchanged T2 vs T1 for both CONT and CONC groups". "

"Use CONC throughout not CASE. "

"In Table 3 please remove case from this column and replace cohort with CONC."

We have made these changes throughout, thank you

"Given the sample sizes are small at 18 or 20 in several of the analysis, is the power too low for this type of analysis for those variables and they should be deleted from the table?"

"This highlighted sentence should be deleted given the first sentence re no significant differences. You have a threshold for significance, so only report as either significant or not."

"Again delete the tendency paragraphs... focus on the threshold of significance you have chosen."

Statistically, yes, we suspect we are underpowered for this analysis. And so we have deleted the “tendency paragraphs” from the Results section as advised. This has significantly slimmed down the results section, thanks.

The analysis in question, though, is quite important and we'd like to retain it. When we conducted these analyses, we immediately recognized the potential for the time between the injury and T2 to affect the results. Had we not done this exploration of REC vs DIST, I strongly suspect it would have been the first thing most reviewers would have asked for and insisted be done. So that table (Table 3) is important, we feel, to show that we have explored how this factor may affect the results. That it is not significant for the submodes is an important outcome, in terms of developing the explanations offered for our other significant findings.

With respects to its implications, we have mentioned this tendency finding once in the discussion, because we are suggesting the hypothesis that the symptoms endorsement and cognitive performance differences we find are the result of subjective reporting that is affected by time between injury and report, and by learning that is also influenced by the time between injury and screen. The same is true for single leg balance errors' in more recent cases.

These are not strong findings, you're right, and as such, are removed entirely from the Results, and do not feature in the abstract. However, in explaining the significant findings in the paper, they lend support to the "why", the explanation for what was different. Therefore, we believe Table 3 serves an important purpose, in two respects – it anticipates the request to explore this effect, which we believe is legitimate and likely strong, and second, it offers support to the theory for why we have found certain other differences in Table 2. On balance, then we would like to retain Table 3 – the study would be missing an important analysis without it. And because it gives us suggestions for future research, which we think is important.

"Given these aims please clearly report in the results and the discussion: 1. CONC versus CONT at T1 = no changes in any SCAT scores (I would hope).

We have reported this in the results, linked to Table 1. It is now also included in the Discussion, as advised

2. CONC versus CONT at T2 - a clear difference - resulting in the need to understand why - which it seems the CONC are BETTER than the CONT at T2 - so the question is therefore is it familiarity due to the extra SCAT tests during the HIA assessments. Therefore show the CONC SCAT scores for T1, HIA1, HIA2, HIA3 and T2. Only if this figure shows a clear clustering of those with recent or distant concussions - using cluster analysis to therefore set the threshold for the time for recent - then provide the recent versus distant concussion analyses."

Thank you, we have made changes in this regard and believe that our revised discussion addresses these issues, and makes a much clearer case, including in the Introduction, for why familiarity may be the factor driving the improvement.

That said, we do not believe that HIA1, HIA2 and HIA3 scores add great merit to this discussion, and would only complicate the analysis and paper if they were to be included. That is because in a player suspected of concussion, the results of HIA1, HIA2 and HIA3 cannot be interpreted as having any value for learning effects when they may change directly because of the head injury. Ideally, a study would be required where players who are known not to have experienced a concussion undergo the three tests, and then the true impact of learning on subsequent baseline screens may be known.

This comes back to our earlier response about quantifying the magnitude of learning or clinically meaningful changes – it's a very important question, one that our own analysis has revealed to be important, but it's one that would require a bespoke study to answer. This is now emphasized in the discussion, and we hope it's adequate.

"This paragraph is the key one that outlines why the study has been conducted. It needs to go into the introduction of the paper."

Yes, you're right, this paragraph makes an important conceptual point, and we have now moved that important concept into the introduction, where it sets up the key study question better, thank you.

"Here you have acknowledged the small sample size - so then why present the data given such limitations?"

For the reasons described above, regarding its importance to respond to an inevitable question about the finding.

We really feel that omitting this data leaves the paper "incomplete", because any reader would assess it and ask whether the time between injury and T2 has an effect? Showing this anticipates that question and answers it. Also, we feel that we need to show the results because some of them, admittedly limited by small sample size, do lend support to the hypotheses offered for why symptoms are reduced and cognitive scores improved. In effect, the analysis in Table 3 links the results of this study to what previous research has found, and provides us with a stronger rationale to explain the changes we have measured as significant.

We hope that this is sufficient to keep the analysis in the manuscript. We do recognize that in our original manuscript, we were overplaying the tendency results in the Results section, and those are now deleted as you have suggested. We thus retain Table 3, a very brief summary of that Table in the Results (with no mention of non-significant findings), and then one mention in the Discussion, which we believe is appropriate as a "hypothesis generating exercise" and a prompt for what future research should look to assess.

"Id like to see better recommendations at the end of each section. Should clinicians use the number of symptoms as a useful part of their assessment or not. What number of symptoms should be the threshold? Does a change of 1, 2, 3 etc matter or not?"

Each sub-section of the discussion now has a new concluding paragraph added. That conclusion makes some recommendations, and while we appreciate what you are asking regarding committing to a specific change that matters, we respectfully maintain that to do so would be speculative and wrong based on this study.

What we can say, and have tried to do so definitively, is that improved symptom and cognitive scores may be a function of deliberate under-reporting, familiarity and learning, and so clinicians must proceed with caution when a baseline result produces these changes. This includes possible repetition of testing using different word lists or number sequences for cognitive modes, and more in-depth evaluation of symptoms, rather than accepting these improvements at face value.

We believe that this advice, now made prominently in the discussion, will greatly improve the clinical utility of the baseline screens.

"I do not like the "post-concussion" terminology used in this paper as it implies and immediate post concussion assessment - ie HIA1, 2, or 3. It needs to be clear that the T2 time is being talked about which is when players have been cleared to return to play from their prior concussion and this is on average *#* days."

Fair point. The revised manuscript does not use the term, so this problem should have been alleviated by the edits we have made. Also, note that we don't know when the player is cleared to

return to play – the GRTP protocol does not use documents that exist within a WR database so we have no access to that date. We can thus only assess from time of injury to T2.

"There is too much repetition of results being reporting in the discussion. Focus on the key points and relate to other literature to build a case as to whether the SCAT and the sub-components should be used or not based on the variation in baseline testing at T1 and T2, and the effects of repeated HIA testing for the CONC participants."

Agreed, and the revised manuscript has less such repetition, and we hope that this achieves the intended outcome

"This is a limitation as the rationale is that repeat exposure to SCAT is affecting the T2 results compared with the T1 results - yet there is no indication of the number of prior tests. The number of SCAT tests must be in the WR database? Can't these analyses be conducted looking at SCAT changes over time for players?"

It is, and the range is between 0 and 3 previous tests for the players in this cohort (T1 is the first ever SCAT for some players, while it is the 4th test for some others). A key point is that this number was also not different between CONC and CONT, in that there were similar proportions of players in the two groups with one, two and three previous SCAT assessments before T1 in the present study. We have included a mention of this in that paragraph of the limitations.

Broadly speaking, the selection of what data to include was a real dilemma in identifying the cohort that we should use for analysis.

We had the option of excluding all players with multiple SCATs prior to the T1 and T2 CONT and CONC, such that we analysed only players whose T1 was their first ever SCAT and whose T2 would be their second ever SCAT, but the impact this would have a massive reduction in sample size, from approximately 500 cases to fewer than 100, and a reduction in CONT of around 80% too.

Ultimately, we decided that for the purpose of this study, namely “Does a concussion affect the next baseline SCAT in a professional player?”, we should include these cases of multiple tests because that’s what is likely to happen, more and more in time, and because it’s not different between the groups. That is, every single player who has been part of a professional team for more than one year will have undergone many SCATs by the time they are concussed in a given season. It is that player’s response to the concussion that was of interest, regardless of whether it was their fourth or first or even eighth SCAT (one day).

So from a policy-perspective, we don’t wish to distinguish between players who have had only one SCAT in their lives and those who’ve had many, mostly because we don’t think it’s feasible to stratify the policy or HIA process between them, unless there is good reason to do so. And that “good reason” is going to emerge from a series of studies on baseline SCATs, of which this paper is but one component.

On the other hand, where you are exactly right is that there is an ‘academic’ or intellectual basis for wanting to explore how “test exposure” might influence test performance. Ultimately, we decided that we would pursue a number of papers, intended to ask a number of questions. This paper is one of them – how does a concussion affect subsequent baseline screen results?

Others are:

- How does baseline performance change over repeated testing in players who are not concussed? This is where we would analyse baseline performance in that cohort of players with 2 or more SCATs, conducted annually, to determine if there is a learning effect;
- Does baseline sub-mode performance change with exercise compared to rest?
- Is baseline performance different in men vs women?

In other words, you've identified an important question, one that we are trying to answer as part of a collection of studies, but we felt that we would be better served identifying one question (concussion effects) and focusing on that one in this paper.

We hope that this is understandable. The incorporation of two of the above questions (concussion and multiple tests) would, we felt, make the paper excessively data heavy, and so we acknowledge that a companion paper should assess this issue moving forwards to understand exactly if and how learning effects impact on findings (note: Without going into detail of what we have found in that paper, I can say that we've not found significant differences in sub-mode performances in the third or fourth SCAT compared to the first in players who have done that many. But again, this was beyond the scope of this paper, which we decided should focus on a concussed and control group without trying to adjust for exposures.

We have elaborated on this in the discussion, in this paragraph, which now also includes that mention you advised of the number of prior tests.

We also did not exclude players who would have performed multiple SCATs prior to the analysis period in this study, whether in the form of the off-field screen, diagnostic screens or baseline assessments. Our cohort included players with between zero and three SCAT assessments prior to their T1 assessment in the present study, though the proportion of players with these prior tests was similar between CONC and CONT. For this study, we chose to include players irrespective of previous SCAT history, because the formalized use of the SCAT5 in rugby means that all players in their second season of rugby who undergo the diagnostic and baseline are going to be exposed to repeated testing. Thus, for the sake of external validity of our findings, we felt it necessary to include all players, irrespective of their previous SCAT baseline history. Future research may explore how multiple tests, rather than two consecutive tests, influence performance as a result of potential learning effects.

"There seems to be too many limitations to make any meaningful interpretation of the differences between the two baseline annual tests useful."

We don't believe this to be the case, and would suggest it is an unfavourable reading of our open acknowledgment of the limitations. Yes, we cannot quantify the size of a learning effect, and yes, we acknowledge that there are limitations, but we don't think that this removes meaning from the result. The reality, in terms of external validity, is that professional rugby players undergo a baseline test approximately a year apart, every year. This study has established that some of those sub-mode scores are different in players who suffered a concussion in that year, compared to those who have not. That is a meaningful finding, and despite the limitations, which really affect our ability to quantify the mechanism, it retains its meaning.

"Unfortunately it is just not clear what the magnitude of the changes for each of the variables should be considered useful clinically given potential learning effects etc."

No, and this study or any like it would never be in a position to quantify the size of these effects. We believe it is a significant finding that the symptom and cognitive sub-mode scores change, and we do

believe we have adequately (in the revised manuscript) explained why these may occur, but we must also be realistic and accept that an ecological study such as this will not allow us to determine how large a change needs to be before it is clinically useful. This is something future studies do need to explore, but given the initial 'bluntness' of the tool, the fact that clinicians apply judgment to it regularly, and the normal variation in these outcomes, we believe we have stopped short, in this manuscript, of speculating beyond what the study and data allow.

"WR clearly have this data for the HIA1, 2, and 3. These data and analyses should be provided for the players in the CONC group, so the short term variation in the SCAT due to concussion can be shown. This would then hopefully show that the SCAT is sensitive to concussion, and that the T1 and T2 SCAT scores are better than the scores during the short term concussion time. Provide a graph for the CONC group of the T1, HIA1, 2 3, then T2 scores, so we can see how SCAT changes. This would provide more useful information to enable interpretation of the T2 changes for CONC and CONT."

We would argue that this would not show short term variation in the SCAT. It would show a mixture of the short-term variation in the SCAT combined with the impact of a brain injury on the performance of a SCAT, which would set up a circular argument making it impossible to discern how SCAT sub-mode performance may change as a result of learning, as opposed to brain injury.

Ideally, the study that should be conducted would assess a group of players who are NOT concussed over time frames similar to what is done on players who are suspected of concussion and thus enter the HIA. That group would provide a better reference for how sub-mode scores change over time, with learning, or randomly.

But we were deliberately reluctant to include the diagnostic screen performances in this study, because they cannot be backed, with any degree of confidence, to be a reliable comparator for what occurs in baseline performances.

"Please provide a final conclusion that addresses the key aim of "whether any SCAT5 sub-modes were affected by a previous concussion, and whether this might have implications for the clinical utility and interpretation of return-to-play and diagnostic screens in the season after the concussion."

We hope that the revised manuscript achieves this. The concluding paragraph has been revised, and we have also, as mentioned, included a concluding "implication paragraph" in each sub-section of the discussion. We acknowledge, as mentioned earlier, that we stopped short of committing to this position but thanks to your and other reviews, we have tried to be more definitive in our revision.

I appreciate that your position may be to want more definitive, including a commitment to a specific value, but we sincerely do not believe that we can credibly do that in this study, given its design and some of the limitations. That said, we don't believe that the study fails to provide "any clarity", or is so limited as to detract from its meaning. We believe the finding is important, and we believe we have a good basis, from a large cohort, to offer possible explanations that expand upon previous research in this area. We do have to stop short of quantifying the relative contribution of those reasons, and the clinical cut-off where a change is significant, though we are committed to exploring these in other studies (some of which are in press as this one is submitted).

We therefore would suggest that we have produced an important finding, and our revised manuscript offers clinical recommendations for those conducting baseline and diagnostic screens, and guidance to them to assess any reductions in symptoms and improvements in cognitive function as potential indicators of deliberate player behaviour, and we think this overall message improves the utility of the baseline screen.

We hope that this is communicated to your satisfaction, limitations notwithstanding, and that our revisions are satisfactory.

VERSION 2 – REVIEW

REVIEWER	Todd Lyons Boston Children's Hospital and Havard Medical School
REVIEW RETURNED	03-May-2020

GENERAL COMMENTS	Summary: Since their initial submission, the authors have attempted to address the issues raised during the first review of the manuscript. My major concern on the first review as about differences between cases and controls. The authors have done an adequate job in explaining the cases and controls. Minor Points: • This manuscript continues to be incredibly long with some redundancy. • Abstract (Page 4, Line 46): the abbreviation T2 appears in the abstract, but has not been defined anywhere prior. • Introduction: Consider moving some of the discussion around the testing of players to the methods, as it is more a description of where the data comes from, rather than why this study is important, or why it was performed. • Methods (Page 10 Lines 12): A new abbreviation PRE appears and is capitalized, but doesn't seem to appear again in the manuscript or be defined. • Results (Page 14): Conssider just putting the description of the directionality of the Odds Ratio and explanation of what the p-value evaluates in the table, and use this to describe the actual results… "While there was a trend towards… in some sub-modes, none of these reached significance using the Bonferonni corrected p-value." • Results (Figure 1): Consider renaming x-axis "player" or "participant" instead of "case" as it is confusing as you have "cases" and "controls" color indicated on the figure. • Discussion: In the interest of shortening the manuscript some of the sub-heading conclusions are duplicative with the "Clinical implications" section.

REVIEWER	Stephen Carek University of South Carolina School of Medicine Greenville Greenville, SC, USA
REVIEW RETURNED	25-May-2020

GENERAL COMMENTS	Thank you for the additions and providing changes. Very interesting study and appreciate the discussion. No additional changes/questions were identified. After review, feel that this paper is ready for publication.

REVIEWER	Professor Patria Hume Sports Performance Research Institute New Zealand, Auckland University of Technology, New Zealand I have previously conducted research on retired player health, and protective equipment, for World Rugby. I have co-authored technical reports with Martin Raftery for these projects for World Rugby.
REVIEW RETURNED	02-Jun-2020

GENERAL COMMENTS	The authors have by their own reporting, completed 90% of the reviewers' requests. The authors have clearly considered and taken on board the reviewers' comments when comparing the manuscript versions. Thank you for correcting Figure 1, revising the abstract, for clarifying and making consistent the terminology for the cohort groups, for clarifying the methods (in particular the available paired comparisons) and improving the analyses with 95% CI, for concluding each sub-section with clinical implications and recommendations, and for providing clarification around the learning effects with references. The paper is now correctly described as a retrospective cohort study. The compromise solution regarding the sub-powered analysis is suitable given the limitations have now been better outlined and rationale provided. Minor editorial changes: The title does not need a capital B for baseline SCAT. page 4 line 9: Delete word "present" in the abstract before study (also on page 9 line 20). page 5 line 27: Change report to reporting. Correct the tense to past tense throughout in the manuscript. For example, is to was (page 9 line 20), are to were (page 9 line 27).

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name

Todd Lyons

Institution and Country

Boston Children's Hospital and Havard Medical School

Please state any competing interests or state 'None declared':
None

Please leave your comments for the authors below
Summary: Since their initial submission, the authors have attempted to address the issues raised during the first review of the manuscript. My major concern on the first review as about differences between cases and controls. The authors have done an adequate job in explaining the cases and controls.

Thank you, I'm glad we have sorted out those issues. The paper is a lot clearer now, so thanks for helping in that regard.

Minor Points:

- This manuscript continues to be incredibly long with some redundancy.

- Abstract (Page 4, Line 46): the abbreviation T2 appears in the abstract, but has not been defined anywhere prior.

We've moved it into a bracket after "second assessment", hopefully this is sufficient to identify it as the coding used for that second test. I tried to find another place within the abstract to put it (or mention T1 and T2), but couldn't without exceeding word limits, so I'm hopeful this is adequate now.

- Introduction: Consider moving some of the discussion around the testing of players to the methods, as it is more a description of where the data comes from, rather than why this study is important, or why it was performed.

Done. We moved that paragraph that describes the HIA to the opening paragraph of the Methods section. This required a small rewording, highlighted in the revision.

- Methods (Page 10 Lines 12): A new abbreviation PRE appears and is capitalized, but doesn't seem to appear again in the manuscript or be defined.

Missed that one, sorry. It's been changed to "the initial" assessment.

- Results (Page 14): Conssider just putting the description of the directionality of the Odds Ratio and explanation of what the p-value evaluates in the table, and use this to describe the actual results… "While there was a trend towards… in some sub-modes, none of these reached significance using the Bonferonni corrected p-value."

Thanks we've done that for the OR in the preceding paragraph, saying that if OR is greater than 1, it indicates that score is more likely to have improved in T2 of CONC than in T2 of CONT.

With respects to the trends and not quite reaching significance, we tried to downplay those based upon Reviewer 3s advice in the first round of reviews, and so we took a lot of those mentions out. I've retained the one for symptoms, as suggested.

- Results (Figure 1): Consider renaming x-axis "player" or "participant" instead of "case" as it is confusing as you have "cases" and "controls" color indicated on the figure.

Yes, good point. We've made this change, thanks.

- Discussion: In the interest of shortening the manuscript some of the sub-heading conclusions are duplicative with the "Clinical implications" section.

Thanks, we've done what we can do shorten it without taking away what we feel are the key sections. We've moved a paragraph out of the main discussion section under symptoms, truncated it by about half and then included that shorter section in the limitations section. This reduces repetition. We were previously asked by a reviewer to conclude each section in the discussion with a clinical implication section, so that has probably led to some duplication, and we're trying to juggle those two

"stylistic" approaches of consolidating all implications in one place, as opposed to discussing implications per section. Hopefully the latest version does that a bit better.

Reviewer: 2
Reviewer Name

Stephen Carek

Institution and Country

University of South Carolina School of Medicine Greenville
Greenville, SC, USA

Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below
Thank you for the additions and providing changes. Very interesting study and appreciate the discussion. No additional changes/questions were identified. After review, feel that this paper is ready for publication.

Thank you very much, and thanks for your constructive input on the first submission which we do think has helped its quality!

Reviewer: 3
Reviewer Name

Professor Patria Hume

Institution and Country

Sports Performance Research Institute New Zealand, Auckland University of Technology, New Zealand

Please state any competing interests or state 'None declared':
I have previously conducted research on retired player health, and protective equipment, for World Rugby. I have co-authored technical reports with Martin Raftery for these projects for World Rugby.

Please leave your comments for the authors below
The authors have by their own reporting, completed 90% of the reviewers' requests. The authors have clearly considered and taken on board the reviewers' comments when comparing the manuscript versions. Thank you for correcting Figure 1, revising the abstract, for clarifying and making consistent the terminology for the cohort groups, for clarifying the methods (in particular the available paired comparisons) and improving the analyses with 95% CI, for concluding each sub-section with clinical implications and recommendations, and for providing clarification around the learning effects with references. The paper is now correctly described as a retrospective cohort study.

Thank you for the re-review, and for your first review. Your suggestions were certainly beneficial, so we are pleased that you've approved of our responses to them, and we agree that the paper is better as a result.
The compromise solution regarding the sub-powered analysis is suitable given the limitations have now been better outlined and rationale provided.

Minor editorial changes:
The title does not need a capital B for baseline SCAT.

Thanks, this has been changed to lower case

page 4 line 9: Delete word “present” in the abstract before study (also on page 9 line 20).

Done, thank you

page 5 line 27: Change report to reporting.

Done thanks

Correct the tense to past tense throughout in the manuscript. For example, is to was (page 9 line 20), are to were (page 9 line 27).

We’ve gone through and made this change throughout, thank you

VERSION 3 – REVIEW

REVIEWER	Todd Lyons Harvard Medical School/Boston Children's Hospital
REVIEW RETURNED	18-Jun-2020

GENERAL COMMENTS	The authors have addressed all of the concerns I have raised on my previous reviews. The paper describes an important topic that will be of use to many clinicians who care for athletes longitudinally at risk of concussion.