**Reviewer responses**

**Reviewer 1**  
  
Comments to the Author  
The broad range of breath-to-breath data collection and treatment options combines with different exercise protocols often presents a confusing picture and interpretation to exercise testing.  Thus, this scoping review has an important imperative that, in certain respects, speaks to addressing reporting concerns regarding the rigor and reproducibility of data processing, analysis and interpretation.  The authors are to be congratulated for the breadth of their reach and employment of machine learning and other state-of-the art methods.  Against this enthusiasm are the following concerns which temper enthusiasm:

Where is the physiology? Why do these procedures matter?  Where is the convincing evidence that errors in reporting, interpretation and physiological understanding have resulted from the omissions etc. identified herein? Without convincing the reader of such any guidelines developed on the basis of the current findings have the danger of being perceived as meaningless over-reach.

*We have referred to Robergs 2010 (L27-33) which shows that the high variability in breath-by-breath (BBB) often requires processing to better reflect the slower changes in metabolism. We later reference Nolte et al.’s (2023) summary of misclassifying fitness and therefore suboptimally selecting patients for treatments (L80-83).*

2. One of the largest sources of error in exercise testing regards the VO2peak versus VO2max issue that, especially in patient populations, has raised substantial concerns especially of late.  The present article avoids the issue of validation tests which can be used to construct the classical VO2max plateau even in the absence of a levelling off in the CPET VO2 (see Poole et al. Eur J Appl Physiol. 2008 Mar;102(4):403-10 and Poole & Jones, J Appl Physiol (1985). 2017 Apr 1;122(4):997-1002). This is a crucial issue that some of the present findings – with appropriate and closer individual article scrutiny and expertise – might address.

*We highlight how identifying a VO2 plateau is preferable to using secondary criteria as the latter may underestimate VO2max. Given that some averaging durations are too long, this may interfere with locating a VO2 plateau and thus identifying one’s true VO2max (L35-54).*

3. Methods 2.1. Doesn’t the validation step remove many of the concerns, at least for VO2max?

*Section 2.1 addressed other content besides VO2 max and validation tests. We aren’t clear on what we should specifically address here. However, this study examined data processing beyond that for determining VO2max, so we feel that specifically addressing the interaction with validations tests is beyond the scope of this article.*

*We can see how data averaging choices may affect the need for a validation step if the data average duration were too long. As prior research shows this reduces the incidence of a VO2 plateau, averaging with 60 seconds may obscure a VO2 plateau that the subject would have shown with a shorter averaging duration. Please let us know if this point is more in line with your original comment.*

4. Page 2, line 55. Typo. “paywall”

*The name “unpywall” is a play on the unpaywall website name (*[*https://unpaywall.org/*](https://unpaywall.org/)*) and the Python programming language. The authors of this python package took out the first “a” in “unpaywall” to emphasize the package was written in python code.*

*See* [*https://pypi.org/project/unpywall/*](https://pypi.org/project/unpywall/) *for details.*

5. Page 4. Should probably complete individual words from the quotations.

*Our regular expressions (RegExs) extracted between 200-250 characters surrounding each phrase to form the “snippets.” We used this range of characters because it generally captured enough surrounding context to determine if the snippet provided evidence of data processing methodology. Since we built snippets based on characters (not words), the snippets often did not start or end exactly at the start or end of word. We elaborate on this in section 2.4.2.*

6. Page 5, lines 29-35. What is important about the specific RER.  Some physiological consideration necessary.

*The mention of RER is because the text surrounding the phrase “over 20-s” happened to contain the words “respiratory exchange ratio.” RER was included here because our RegExs first found the phrase “over 20-s”, then found a nearby mention of nearby text related to gas data, in this case, “vo2max”, and finally extracted the ~200 characters surrounding the phrase “over 20-s”. From this snippet of ~400 characters total we can determine that the authors used both 20 and 60-second bin averages. In addition to containing the details of averaging gas exchange data, this snippet contains some unrelated information, such as a mention of RER. The first paragraph of section 2.4.2 contains additional details on how we created these snippets. If this is still unclear, please let us know and we can further clarify.*

7. Page 5, line 51. And what exactly is the “anaerobic” threshold? Please inject some physiology here.

*The mention of the “anaerobic threshold” here is for the same reson as the reference to RER : some of the text extracted from each snippet contains information about other topics besides the data processing methods I hoped to document. Please see our response to your sixth comment for details.*

8. Page 8. Lines 51-3. Show numerical examples.  How does this impact interpretation and physiological conclusions based thereon?

*We feel it is beyond the scope of this analysis to show how manipulating averaging method, interpolation, and outlier removal strategies affect VO2max. Nevertheless, we have added “as averaging suppresses potential outliers itself by combining those points with additional observations.” Since averaging combines multiple values together, while outlier removal typically affects singular values, we feel this reasoning supports our assertion that the choice of data average usually has a greater impact than the choice of outlier removal on the final calculated VO2max (L517-518).*

9. Page 9.  “30 s average” So, is this good?  How does it fit with physiology and clinical assessment?

*I have clarified that although most people aren’t choosing averaging durations that are “too long.” that the vast majority of people are still not following the best-practices set out by Robergs (2010). (L541-568).*

**Reviewer 2**

# General Comments

The article uses a semi-automated screening approach to describe the reporting of data processing strategies for gas exchange measurements in the published literature. The relevance of this meta- research article is clear, but a recent key reference with a comparable aim is missing (Nolte 2023). The text-mining approach allows for a large sample size of studies, but may require validation. The

presentation of the methods and results could be improved, and some important aspects should be discussed.

*When the study by Nolte et al. (2023) came out, I had previously finished writing the scoping review portion of my dissertation around May/June of 2023. Since defending my dissertation in September 2023 I had not searched for studies previous to our scoping review. I am glad to know of Nolte et al. (2023) and have included comparisons to their work throughout this work.*

# Specific Comments

(Continuous line numbering would be helpful in the source document)

*The new version has line numbers.*

Abstract

I believe it is important to mention in the abstract that you used automatic screening procedures to assist your review. This is both a strength and a weakness of your research, but it is certainly a quite unique feature that warrants inclusion in the abstract (and possibly even in the title).

*Title and abstract now include “semi-automated.”*

Please indicate at some point that you are providing CIs/margins of error here.

*We have clarified the use of Cis in the abstract*

P1 19 Please state that the 1078 articles are a random subset of the larger sample.

*This random subset number has been added to the abstract*

Introduction

I generally like the reasoning and style of the introduction. Some minor comments:

P1 49 Please give a reference for this claim. Is this not what you are essentially trying to investigate?

*It appears to us that Nolte et al. (2023) was the first to systematically report the prevalence of reported outlier removal strategies. This line was informed from personal experience. We have now referenced Nolte et al. (2023.) (L57)*

P2 9 This is true as long as respiratory rates are lower than 60 per min, which is of course common in most test settings. However, during exhaustion, trained athletes may also reach rates of >60, in which case one-second interpolation would widen the confidence interval.

*I have specified that the most common effect is narrowing confidence intervals as most subjects do not surpass 60-breaths per minute (L66-67).*

P2 13 You are missing a key reference here (Nolte 2023). They also investigated the reporting of data filtering and interpolation. However, their sample size was also small compared to your work.

*Reference to Nolte et al. (2023) added.*

P2 20 The low reporting rate of data processing methods is not only something that can be observed anecdotally but has also been demonstrated in published meta-research (Midgley 2007, Nolte 2023).

*At the time I first wrote this, I only had my personal anecdotal experience with respect to outlier removal and interpolation reporting. I have changed the text as this is no longer anecdotal but documented.*

Methods:

You are missing the section ‘Protocol and registration’ from the PRISMA-ScR here (see the sources you cited).

*We have formally added this section at the beginning of our methods.*

Please provide the full RegExs in the supplemental material.

*These RegExs will be provided on the GitHub repository which is now public. Key files with RegExs are bbb\_regex.py and scrips within the regex\_analysis folder. The link to this repository is on line 121. It may be obscured by track changes.*

As one of the main reasons for your research here is to improve the replicability and reproducibility of other studies, I highly encourage you to make your own research reproducible by making your Python and R scripts publicly available.

*We have made the corresponding github repository public and have provided a link at the start of the methods section.*

You are using a semi-automated screening and data extracting procedure. Have you performed any validation for this on a subset of articles? If so, what were the performance metrics?

*We didn’t have a formal subset validation test, but we have included cross-validation accuracy values and other details in a supplemental methods document (L179-180).*

P2 42 Please provide a reason for not using a date restriction. Others may argue that processing strategies may have changes over time with changes in technology and education. This is also something you should discuss in the discussion section.

*The reason to forgo a date restriction was to cast as wide a net as possible (L127-128).*

*We note how our lack of date restriction likely impacts some of the different results in our study compared to previous works. For example, we showed that 60-s averaging was the next most popular method, while others did not.*

P2 47 Please provide the search dates for your data base searches.

*The search date on 2022-06-27 has been added to the “Information Sources and Search” section*

P2 49 From what I can see, you provide the search strategies for all three sources in the supplemental materials?

*Good catch. The text in the information sources and search section now matches our supplemental materials document with all electronic search strategies (L131-132).*

P2 56 In my opinion, the number of excluded articles belongs in the results section. Perhaps you could write something as ‘Articles that could not be accessed through our library or via unpaywall were excluded from the analysis’ here.

*Please let us know if the updates Results-Selection of Sources of Evidence section explains this.*

P3 24 Please add more context on the random forest classifier(s?). How large was your training set and did you have a test set?

*See the supplemental methods file.*

P3 15 Not all readers may be familiar with the term RegExs.

*I have compared regular expressions to cmd/ctrl+F as most readers are familiar with this action (L148-149).*

P3 26 Did you manually review every article flagged ineligible?

*Yes, we manually reviewed every article flagged as ineligible. See the supplemental methods.*

P3 32 Some metabolic carts offer both a BBB and mixing chamber mode. How did you deal with these?

*We made the assumption that the data was collected breath by breath (L186-188). We stated this and added a line in our limitations (L594-595).*

P3 51 Did you also read the full-text if no snipped could be found?

*Aside from random spot checking, we did not read the full text if no snippet could be found. We considered such methods as “not described.” (L200-201).*

P4 9 How did you generate the random sample?

*We used python to generate a random sample. We set a seed when generating this sample. (L210).*

P4 13 Thank you for providing a precision-based sample size calculation, something I have not seen often so far. I could fully reproduce your results. Please include an explanation how you determined your input parameters (margin of error and assumed proportion).

*We chose 0.5 for a proportion as it provides for a more conservative estimate because it maximizes the standard error. That is, it maximizes the p \* (1-p) part of the formula.*

*The margin of error was selected based on a balance between accuracy and the number of samples required to analyze to attain such a margin of error.*

*The details are explained in text (L231-233).*

P5 26 How often was this the case and how did you consider this in your analysis when given proportions?

*We included that we observed this about 14% of the time in our results.*

*In Methods-Synthesis of Results section, we’ve specified that when articles reported multiple methods, we only counted that article once for the purpose of the proportion of articles reporting averaging methods (L331-332).*

P6 34 You are calculating binomial confidence intervals, presumably using the classic Wald method. Please specify the method used and explain the reason for choosing it. Especially since the Wald interval has rather poor properties and some serious limitations that also apply to your results, e.g. overshoot (Brown 2001).

*We were previously using the Wald method. We were ignorant to the erratic behavior of binomial confidence intervals, especially at low proportions. Our updated confidence intervals are now calculated using the Agresti-Coull method, per the recommendation by Brown et al. (2001)*

Results

P6 43 I am a bit surprised by the relatively low number of initial search results (21981). Given that you had no date restrictions, I would have expected a larger number of studies. Is there anything in your search strategy that may limit the number of studies retrieved?

*After conforming to the PRISMA 20202 flowdiagram guidelines, our results more clearly show that our initial search results were far more expansive. This is the new figure 1.*

*Nevertheless, our search strategy included some mention of exercise modality, as our early searches showed it helped identify studies with an exercise test. See the supplemental materials.*

P6 52 Thanks for providing the raw data of your research. I recommend uploading your data to a data repository with a persistent DOI instead of using Google Sheet links, which may become inaccessible in the future.

*We are still working on preparing this data to add to our institution’s data repository.*

Discussion

As you are given recommendations on reporting of the strategies, please discuss them in light of the recently published reporting recommendations (Nolte 2023).

*See just before the limitations section. Our results agree with Nolte’s in that we see that most authors have* not *historically followed those best practices.*

*Our proportions for the popularity of different methods do not always match Nolte and similar sources, but they are often close.*

*L569-581*

P8 11 This is even more important for individuals that are less fit: If evaluations for operations are based on cardiopulmonary test parameters (Johnson 1998).

*We have added a sentence and reference to Johnson 1998.*

P8 21 The influence of data processing on VT determination is indeed an interesting topic. I look forward to seeing this research in the future.

*I’m looking to publish on this topic next! My preview is that it doesn’t have a large effect on average, but there can be great variability. Certain algorithms, at least in combination with certain x-y relationships, are more inconsistent than others.*

P8 30 In the absence of a recommendation for outlier removal, please discuss the strategies you have found in the literature. On what basis were these strategies likely chosen? And how ‘good’ might the +/- 4 SD strategy be? There are some important aspects to discuss (e.g., do you need outlier removal if you use filtering instead of averaging?, is there a potential value of using simultaneous confidence bands [Liebl 2023] or Bayesian credible intervals?)

*We’ve noted how these methods are based on previous literature and because it’s not that hard to program. (L480-485).*

P8 35 Discuss your reporting prevalence for outlier removal and interpolation with the prevalence found by Nolte (2023).

*L569-581*

P8 41 Well argued. However, I am wondering how important the confidence interval shrinkage is to current published research. I am not aware of many articles providing confidence intervals for individual level gas exchange parameters (often, rather the point estimate is treated as the true value and confidence intervals are only constructed around the group mean, i.e., reflecting sampling uncertainty on the population level).

*In our experience, we also haven’t often seen confidence intervals reported with the threshold value. Previous research shows these thresholds are better described as “transitions,” so we argue a confidence interval may better capture that uncertainty L505-508).*

Responded P8 46 ‘Though the effect of interpolation is likely small, …’

*L511-513*

P9 12 Write the first author name out here

*First author name added (L528).*

P9 14 These conversion formulas only work on the group level. When it comes to the individual patient/athlete, misclassifications can still occur because the effect of data processing strategies can vary on the individual level.

*We reread Martin-rincon and discussed the group-level limitation (L532-533).*

P9 25 Please discuss your finding of averaging strategies reporting with the results found in other research (Midgley 2007, Robergs 2010, Nolte 2023).

*We note the similarity of the most popular methods (L573-577).*

P9 27 Robergs (2010) recommends a digital filter, and the other strategies only as a substitute if researchers are unable to apply a filter. Also, he recommends a 30s-rolling average instead of the 30s- bin average you found to be most prevalent in the literature.

*We’ve noted how the duration matches but the calculation method (bin vs. rolling) does not (L545-568).*

P9 29 Please discuss the prevalence of data processing strategies you have found in the research. Why were some common despite not recommended? Why were some uncommon despite recommendations?

*We think it’s in part because it’s what’s been done in the past (citing Robergs 2010) and because that might be what the metabolic cart software offers. L578-581*

P9 40 A way to deal with word limits may be to put detailed methodology in the supplemental

material of a research article. However, this is something that could not be detected by your method.

*We mention how our method does not capture data that is referenced by other works (L858-586)*

P9 41 An even better approach for ensuring reproducibility than just reporting is to make data and code available. There is some existing open source software for analysis of gas exchange data you could point to here.

*Both are good points. We have mentioned publishing all code and data. We have also linked to 3 R packages capable of analyze gas exchange data. (L588-590)*

P9 45 I wonder how you would construct a larger sample, as you had no date and test parameter limitation in your search.

*We could have analyzed a larger random subset. L593.*

P9 55 You could discuss this in more detail. E.g., would you expect that data processing does differ based on the parameter of interest in the literature. And more importantly, should it differ?

*We have noted that vo2 kinetics and similar analyses, for example, ventilatory threshold detection, likely employ somewhat less smoothing than averaging to find VO2max. L604-623).*

P10 5 You can add a limitation regarding ambiguity in processing strategy reporting (and therefore your data extraction). I suppose there were many cases in which the exact data processing strategy was not clearly defined in the articles and you had to imply the original authors’ meaning.

*At the start of the 2nd to last paragraph in the limitations section, we mention that this study did not record the purpose of each data processing strategy. In some cases, as you mentioned, this was unclear. Please let us know if this sufficiently addresses this point. L569-577)*

Figures

Fig 1: This figure rather confuses me instead of clarifying your research. I do not think it is necessary for your article.

*We originally included this figure to document the several steps in processing breath-by-breath data. We discussed your suggestion to remove this figure and agree it can be removed as we describe this process in the main text.*

Fig 2: Please use a PRISMA flowchart for this purpose. Also provide how many results you obtained from which source and how many duplicates you removed. You could also include the additional step of the random sample in the flowchart for clarity.

*Figure 2 (now figure 1) now follows the PRISMA 2020 flowdiagram guidelines.*

*This figure now states how many articles were obtained from each source and how many duplicates were removed.*

Fig 3/4: Maybe combine these two into one figure with two subfigures.

*We see the reasoning here as outlier removal and interpolation were both reported far less often. However, we’d like to stick with a separate figure for outliers, interpolation, and averaging because the paper overall has repeated subsections dedicated to each of those steps. Therefore, it feels consistent to keep the outlier and interpolation figures separate.*

Tab 1: I find this table rather redundant given the result section and Fig 4.

*Although table 1 separates the interpolation time from it type, we agree this can be omitted.*

References

Brown et al. (2001). Stat Sci [10.1214/ss/1009213286](https://doi.org/10.1214/ss/1009213286)

Liebl & Reimherr (2023). J R Stat Soc Series B Stat Methodol [10.1093/jrsssb/qkad026](https://doi.org/10.1093/jrsssb/qkad026)

Nolte et al. (2023). Sports Med [10.1007/s40279-023-01903-3](https://doi.org/10.1007/s40279-023-01903-3)