The prevalence of gas exchange data processing methods: a semi-automated scoping review

Authors
Anton Hesse, Manix White, Christopher Lundstrom

Affiliations
Department of Kinesiology, University of Minnesota Twin Cities, Minneapolis, USA

Keywords
averaging, outliers, interpolation, cardiopulmonary exercise testing, breath-by-breath, reproducibility

received 25.07.2024
accepted 02.12.2024
published online 2024

Bibliography
Int J Sports Med
DOI 10.1055/a-2495-5364
ISSN 0172-4622

Georg Thieme Verlag KG, Oswald-Hesse-Straße 50, 70469 Stuttgart, Germany

Correspondence
Dr. Anton Hesse
Department of Kinesiology,
University of Minnesota Twin Cities,
Minneapolis,
USA
hesse151@umn.edu

Supplementary Material is available at https://doi.org/ 10.1055/a-2495-5364

ABSTRACT

Cardiopulmonary exercise testing involves collecting variable breath-by-breath data and sometimes requiring data processing of outlier removal, interpolation, and averaging before later analysis. These data processing choices, such as averaging duration, affect calculated values such as $\dot{V}O_2$max. However, assessing the implications of data processing without knowing popular methods worth comparing is difficult. In addition, such details aid study reproduction. We conducted a semi-automated scoping review of articles with exercise testing that collected data breath-by-breath from three databases. Of the 8,344 articles, 376 (mean: 4.5 % and 95 % confidence interval: 4.1–5.0 %) and 581 (mean: 7.0 % and 95 % confidence interval: 6.4–7.5 %) described outlier removal and interpolation, respectively. A random subset of 1,078 articles revealed (mean: 60.9 % and 95 % confidence interval: 57.9–63.7 %) the reported averaging methods. The commonly documented outlier cutoffs were ± 3 or 4 SD (39.1 and 51.6 %, respectively). The dominating interpolation duration and procedure were 1 s (93.9 %) and linear interpolation (92.5 %). Averaging methods commonly described were 30 (30.9 %), 60 (12.4 %), 15 (11.6 %), 10 (11.0 %), and 20 (8.1 %) second bin averages. This shows that studies collecting breath-by-breath data often lack detailed descriptions of data processing methods, particularly for outlier removal and interpolation. While averaging methods are more commonly reported, improved documentation across all processing steps will enhance reproducibility and facilitate future research comparing data processing choices.

Introduction

Clinicians and researchers use cardiopulmonary exercise testing (CPET) to determine maximal aerobic capacity ($\dot{V}O_2$max), ventilatory thresholds, and $\dot{V}O_2$ kinetics. Such values help categorize fitness, predict disease risks, and guide exercises [1]. Using CPET results to guide exercises, especially relative to thresholds, yields better improvements given more consistent and predictable metabolic responses [2]. Therefore, incorrectly calculating or identifying these values limits CPET benefits.

Breath-by-breath (BBB) data in CPET often require processing to manage its high variability [3]. BBB $\dot{V}O_2$ data can change by up to 86 % during a steady state, but changes to muscle blood flow or oxygen extraction cannot account for such rapid fluctuations [3]. Instead, Robergs et al. [3] showed that the rate and depth of breathing account for most of the variation in $\dot{V}O_2$ during both steady states and incremental exercise. Therefore, CPET data processing usually involves outlier removal, optional interpolation to regular intervals, and averaging to more accurately reflect whole-body metabolism [3]. Previous research has shown that data averaging influences CPET values. Averaging over longer durations reduces $\dot{V}O_2$max and $\dot{V}O_2$ plateau detection [3–17]. The importance of attaining a $\dot{V}O_2$ plateau to accurately determine $\dot{V}O_2$max has been

highlighted because "secondary" $\dot{V}O_2$max criteria often underestimate $\dot{V}O_2$max [18] and thus misclassify cardiorespiratory fitness. Finally, we are unaware of research on the effects of data processing and locating ventilatory thresholds.

Outlier removal typically excludes points beyond 2–4 standard deviations (SD) from the local mean [19]. These cutoffs are common because the relatively small sample size of BBB gas exchange data often contains more values beyond 3 or 4 SD than one would predict from an assumed Gaussian distribution [20]. More outliers appear than expected because of both conscious and unconscious alterations of breathing patterns, including swallowing and coughing [20]. We are unaware of prior research that examines how different outlier removal strategies affect $\dot{V}O_2$max, ventilatory thresholds, and $\dot{V}O_2$ kinetics.

Interpolation, often to 1-s intervals, is common in $\dot{V}O_2$ kinetics research to "ensemble" average repeated transitions to minimize variability [20, 21]. Although this does not affect parameter estimates, 1-s interpolation has been criticized for artificially narrowing confidence intervals (CIs) as respiratory rates are usually below 60 breaths/min, even near maximal exercise [22–25]. As before, we are unaware of research specifically investigating how interpolation affects $\dot{V}O_2$max and ventilatory threshold identification.

Data processing choices, such as averaging and interpolation, impact CPET variables or their CIs. Earlier surveys [3] and studies [8] were small and focused on averaging methods only, finding that time-based bin averages (e.g., 30-s averages) were popular. A more recent scoping review by Nolte et al. [19] found that nearly half of the studies on $\dot{V}O_2$max ramp protocols lacked data processing steps. They also found that only 4.3 and 4.5 % of papers reported interpolation and outlier removal strategies, respectively. Finally, they reported that scant studies employed the recommended moving-average or digital filter averaging options suggested in 2010 [3].

These low rates of reporting data processing steps may hamper reproduction or replication attempts, which have become a more prominent issue in science within the past decade [26, 27]. In addition, as summarized by Nolte et al. [19], using $\dot{V}O_2$max and similar values to classify fitness or evaluating patients for treatment requires practitioners to consider and state data processing choices as they may inadvertently misclassify or suboptimally select patients and treatments.

This research expands on that of Nolte et al. [19] but searches without date restriction and includes studies with CPET data beyond $\dot{V}O_2$max ramp tests. To accomplish greater breadth, we employed a semi-automated analysis to find more articles based on common text patterns before manually reading extracted subsections from each study. This review assesses the reporting frequency of outlier removal, interpolation, and averaging methods. The results emphasize the need for documentation to improve reproducibility and document the data processing choices worth testing in future research investigating the effects of such choices on CPET values.

Methods

Protocol and registration

This report followed the PRISMA scoping review and related guidelines [28–30]. This work was first registered with the Open Science Framework [31, 32] and was based on a dissertation chapter by the first author [33]. The code and most data for this project were available on GitHub.

Eligibility criteria

This scoping review surveyed gas exchange data processing choices in original, peer-reviewed studies, summarizing the reporting frequency and methods for outlier removal, interpolation, and averaging. Full-text files could not be shared due to licensing restrictions. Eligible articles were original, peer-reviewed articles, with BBB gas exchange data, human participants, in English, with a DOI. We imposed no date restriction to be as comprehensive as possible.

Information sources and search

Data were collected from Ovid-MEDLINE, Scopus, and Web of Science on 6/27/2022with librarian assistance. The electronic search strategies for all databases can be found in the information sources and search section of supplemental materials.

Our search output comprised article identifiers like DOIs. To find missing DOIs, we employed the PubMed Central ID Converter API [34] using Python code. Full texts were accessed via publisher texts and data mining APIs using Python, unpaywall.org[35], using the unpywall Python package, through custom-built web-scraping scripts or manually. Our library subscription did not permit access to 1,549 articles.

Selection of sources of evidence

This study used a single screening process, requiring only BBB gas exchange data collection with exercise. The corresponding PRISMA flow diagram was created using the PRISMA2020 R package [36].

Text analysis and screening

Despite database search filters, we screened additional non-English, non-human, and non-original articles such as reviews, meta-analyses, and protocol registrations, in addition to case studies. We manually analyzed a subset of articles to help build machine learning (ML) classifiers and construct regular expressions (RegExs) described below. Regular expressions identify specific text sequences. A familiar RegEx example is searching a document using cmd/ctrl + F. These ML classifiers and RegExs helped in identifying ineligible articles. This computerized screening required converting full-text PDF and EPUB documents into plain text files. Plain text files were normalized by converting the text to lowercase, removing hyphenations and extra whitespace, and correcting some plain text conversion-induced errors.

Following normalization, we identified and removed articles that failed to correctly convert into a text format, spotted non-English articles using the fasttext Python module [37] and employed a random forest ML classifier from the sklearn Python package [38] to identify ineligible articles based on our criteria. See the supplemental methods for text analysis and screening for additional ML model details.

Next, we identified BBB articles using RegExs. Articles were considered BBB articles if their text contained variations of the phrase "breath-by-breath", or if their text included the make or model of a known BBB analyzer. BBB brands and analyzers we included were Oxycon and Carefusion brands, Medgraphics Ultima, CPX, CCM,

and CardiO$_2$ models, Sensormedics Encore and 2900 models, Cosmed quark, k4, and k5 models, and the Minato RM-200, AE-280S, AE-300S, and AE-310S models. Some metabolic carts have both BBB and mixing chamber modes. If not described, we assumed that the data were collected BBB. In total, we identified 8,412 articles.

Within this subset, we performed a similar RegEx search for studies that documented using Douglas Bags or mixing chambers and excluded those articles. The full details are described in the "data charting process" section.

Data charting process

RegExs identified short phrases likely indicating that the authors described these methodological details. If present, we extracted a "snippet" of the text surrounding those phrases for later manual analysis by obtaining approximately 200 surrounding characters. We then recorded the methods from these snippets. In all cases, methods were only considered documented if the snippets provided at least some specific information. For example, articles stating outlying breaths were removed but without describing the outlier criteria were considered "not described." Finally, we read the full-text article to accurately document the data when snippets were ambiguous. Full-text articles without snippets were not read and there methods were documented as "not described."

The data charting subsections below provide text extraction examples. Extracted texts were normalized to lowercase, with end-of-line hyphenations and unnecessary white space removed before capitalizing certain keywords for readability. Therefore, formatting varies and may include unconventional spacing and Unicode characters. Finally, the snippets may not start or end at the beginning or end of a word or a sentence because the RegExs extracted a specific number of characters rather than words.

All eligible BBB articles were analyzed for outlier and interpolation methods due to distinct descriptions and fewer total articles (~5 %). In contrast, we analyzed a random subset of articles using a random number generator in Python to document data averaging methods because far more articles described their averaging methods. Early estimates as we developed our RegExs were that ~60 % or 5,047 articles had some averaging details. Furthermore, the phrases associated with averaging methods are more generic and often refer to other study aspects, such as heart rate averaging periods. Given the large number of articles, we needed a minimum sample size of 1,068 based on a 95 % CI and a maximum margin of error (MOE) of ± 3 %, assuming a proportion of 0.5 for a conservative estimate. However, we increased this to 1,100 in anticipation of finding ineligible articles that eluded our previous text screening. The chosen MOE was selected as a balance between the accuracy and the required corresponding samples: decreasing the MOE to ± 2 % with an assumed proportion of 0.5 would require another 1,333 samples.

Outliers

Our outlier RegExs identified phrases like "swallowing", "coughing", "errant", and "aberrant", and references to the "local mean," "prediction interval," or a specific standard deviation limit such as ± 3 or ± 4. For example, our RegExs found "errant"; "local mean"; and "breath-by-breath $\dot{V}O_2$ data from each step transition were initially edited to exclude errant breaths by removing values lying more than 4 sd" from ref. [39]. We gathered snippets surrounding those phrases and combined them when overlapping, thus producing

> *y[hb + mb] data (Quaresima & Ferrari, 2009). expressed as 2.5 data analysis and kinetic modelling the breath-by-breath $\dot{V}O_2$ data from each step transition were initially edited to exclude errant breaths by removing values lying more than 4 sd from the local mean determined using a five-breath rolling\n\x0c1932 Breese et al. and deoxy[hb + mb] responses were subavera*

We recorded the outlier limit as ± 4 SD and the outlier function as a rolling 5-breath whole mean average.

Interpolation

Nearly all articles describing interpolation methods used variations of "interpolate." The remaining phrases were infrequent and inconsistent enough that interpolation methods were only described for those articles when discovered by chance. To illustrate interpolation documentation, our RegExs extracted the snippet from ref. [40].

> *the $\dot{V}O_2$ data from gd and gl exercise bouts were modeled to characterize the oxygen uptake kinetics following the methods described by Bell et al (2001). Breath-by-breath $\dot{V}O_2$ data were linearly INTERPOLATed to provide second-by-second values. Phase 1 data (i.e. the cardiodynamic component), from the first ~20 s of exercise, were omitted from the kinetics analysis because phase 1 is not directly repres*

We documented the interpolation type as "linear" and the interpolation time as 1 s.

Averaging

We document averaging methods according to five criteria: type/units, subtype/calculation, amount, measure of center, and mean type. Type/units refer to the averaging units of time, breath, and digital filters. Subtype/calculation involves specific computations like bin and rolling averages or digital filter forms. The amount is the unit quantity. For example, 30 for a time average is 30 s but it is 30 breaths for a breath average. The measure of center distinguishes between a mean or a median and mean type delineates whole vs. a trimmed mean. Trimmed (truncated) means excluding a number of the highest and lowest values in the quantity before averaging the remaining data.

Descriptions of averaging methods are also considerably more diverse and generic than outlier and interpolation descriptions. For example, "30-s averages" and "averaged every 30 s" invite complexity, leading to more snippets referring to averaging something besides BBB gas exchange data. Given that, we required that the text snippets include a reference to gas data such as the text "O_2," "breath," "gas," "ventilation," etc.

In contrast to previous studies, we also documented every averaging method we found per paper instead of only describing the averaging method for $\dot{V}O_2$max. We also recorded multiple averaging methods when the authors described the sampling interval and

the transformation applied to it. For example, the snippet from ref. [41]

> *ath method using the vmax respiratory gas analyzer (sensormedics, yorba linda, ca).$\dot{V}O_2$max was defined as the mean of the three highest values of the averaged oxygen consumption measured consecutively OVER 20-S intervals. A total of 98 % of the subjects achieved the respiratory exchange ratio of ■⌂■1.1. Electrocardiography was recorded throughout the exercise test using cardiosoft software (ge medical systems,*

states that oxygen consumption was measured every 20 s and that $\dot{V}O_2$max was calculated as the average of three 20-s intervals or 60 s. For this article, we documented one averaging method as a 20-s time bin whole mean and another as a 60-s time bin whole mean.

In many cases, authors did not explicitly use the terms "average" and "mean" to describe their averaging methods, but we documented their methods when implied. For example, the snippet from ref. [42]

> *red using a continuously monitored electrocardiograph. blood pressure was measured at the end of each workload increment using an automatic sphygmomanometer. Peak v9o2 was defined as the v9o2 measured DURING THE LAST 30 S of peak exercise. Oxygen pulse was calculated by dividing v9o2 by cardiac frequency. The anaerobic threshold was detected using the v-slope method [16]. The ventilatory equivalent for carbon dioxide w*

states they calculated $\dot{V}O_2$peak using the last 30 s of exercise data. We documented such phrasing as a 30-s time-bin whole mean average.

Data items

In all cases, articles that did not return any phrases were documented as "not described" for their respective data processing category. If snippets did not refer to the data processing category or if the snippet lacked sufficient information, these data processing variables were documented as "not described." For example, interpolation variables were denoted as "not described" if interpolation was acknowledged but without details for the interpolation type or time.

Outliers

We documented the outlier limit, for example, ± 3 standard deviations, and any outlier function used to compute the outlier limit, if described.

Interpolation

We recorded the interpolation types (linear, cubic, Lagrange, specifically uninterpolated, and other) and time frame (e.g., every 1 s).

Averaging

We noted the following averaging types: Time, breath, breath-time, time-breath, time–time, digital filter, ensemble, (explicitly) unaveraged, and others. Averaging subtypes included bin, rolling, binroll, rolling-bin, Butterworth low-pass, Fast Fourier Transform (FFT), and Savitsky–Golay. Next, we recorded the time in seconds or the number of breaths. We recorded the measure of center as the mean or the median. Finally, we noted if the mean was a whole or trimmed.

Synthesis of results

Counts, proportions, and Agresti-Coull 95 % CIs were calculated for the reporting frequency of each data processing method using R version 4.1.2 [43] in the RStudio IDE version 2023.6.1.524 [43]. When articles reported multiple methods, we only counted this article once for calculating overall reporting proportions.

Results

Selection of sources of evidence

▶ **Fig. 1** shows the selection of sources of evidence flowchart. The initial search identified 50,730 articles; 21,715 remained for retrieval after removing duplicates and those without DOIs. A total of 8,344 articles analyzed were included in the interpolation and outlier analysis. After removing 22 ineligible articles that we discovered during data documentation from the original 1,100 random articles, we analyzed 1,078 articles for our averaging analysis.

Characteristics and results of individual sources of evidence

Given the vast nature of this scoping review, readers can view web links to our outlier, interpolation, and averaging data charting spreadsheets.

Synthesis of results

We present our results according to the reporting prevalence followed by the specific characteristics when reported.

Outliers

Of the 8,344 articles, 376 (4.5 %, 95 % CI: 4.1–5.0 %) reported outlier removal methods. The most prevalent reported methods were ± 3 (39.1 %) and ± 4 (51.6 %) SD (▶ **Fig. 2**).

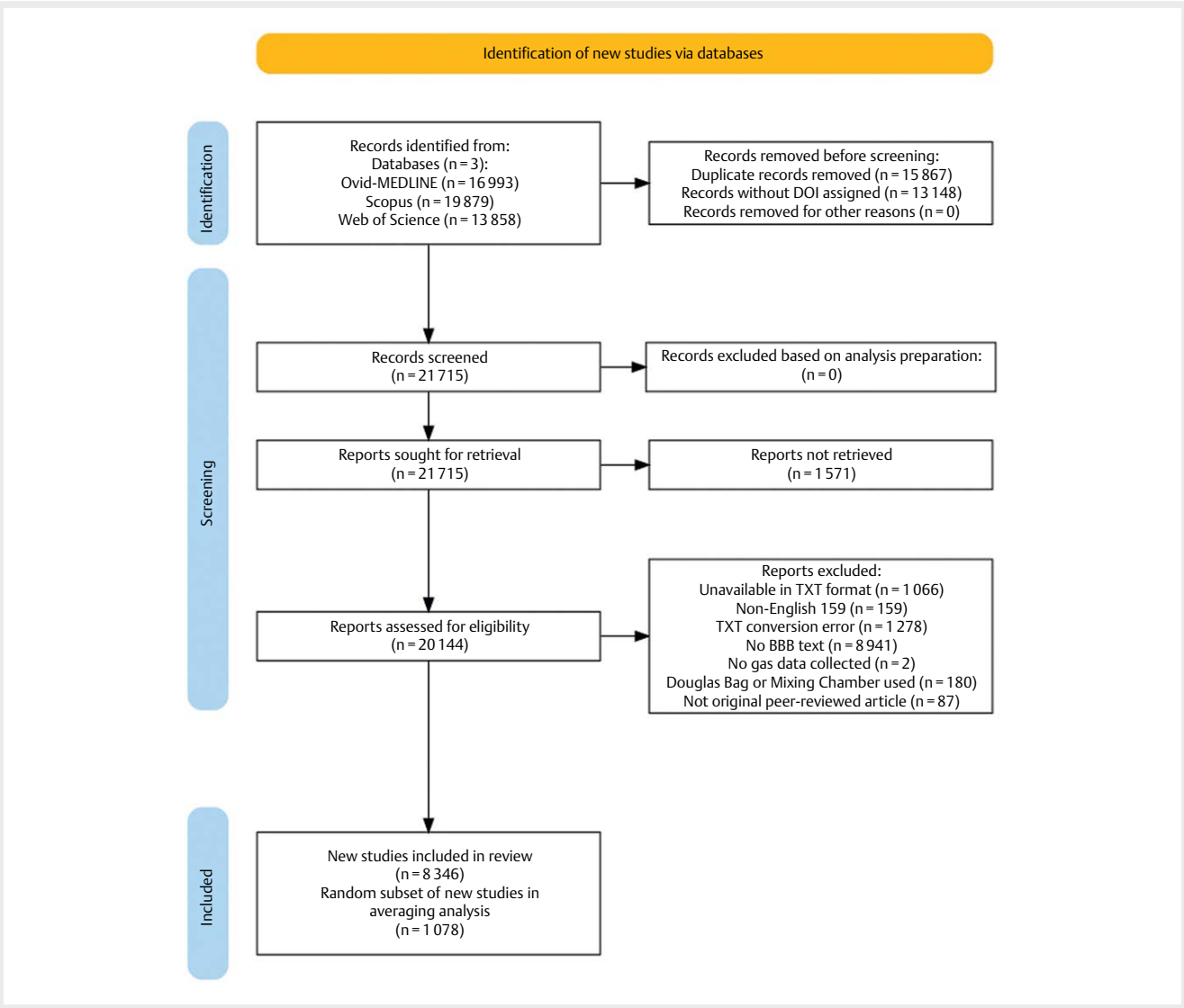
Only 102 (1.2 %, 95 % CI: 1.0–1.5 %) articles reported details of the function that they used to calculate their outlier limits. Of those, breath-based averages (n = 76, 74.5 %) and then time-based averages (n = 15, 14.7 %) were the most common for calculating outlier boundaries. Specifically, 5-breath averages (n = 54, 52.9 %) were the most prevalent functions to calculate outlier limits.

Interpolation

Of 8,344 articles, 581 (7.0 %, 95 % CI: 6.4–7.5) specified interpolation, with 1-s intervals as the most common (93.9 %). Around half of the reported interpolation procedures included the method (n = 314, 54.0 %), with linear interpolation as the most popular (n = 247, 92.5 %; ▶ **Fig. 3**).

Averaging

We recorded 656 (60.9 %, 95 % CI: 57.9–63.7 %) articles with averaging methods. Of these, 14 articles reported more than one averaging method. Time averages dominated in popularity (91.5 %; ▶ **Table 1**). Bin averages proved the most widespread averaging



▶ **Fig. 1** Selection of sources of evidence flow diagram per the PRISMA 2020 statement [30].

subtype (89.9 %; ▶ **Table 1**). Together, time–bin (86.8 %) was the most frequent type-subtype averaging method combination.

When incorporating averaging amounts, 30-, 60-, 15-, and 10-s bin averages (▶ **Fig. 4**) were the most popular. The "other" methods category accounted for the second highest share of the total, but this represents many rarely used averaging methods.

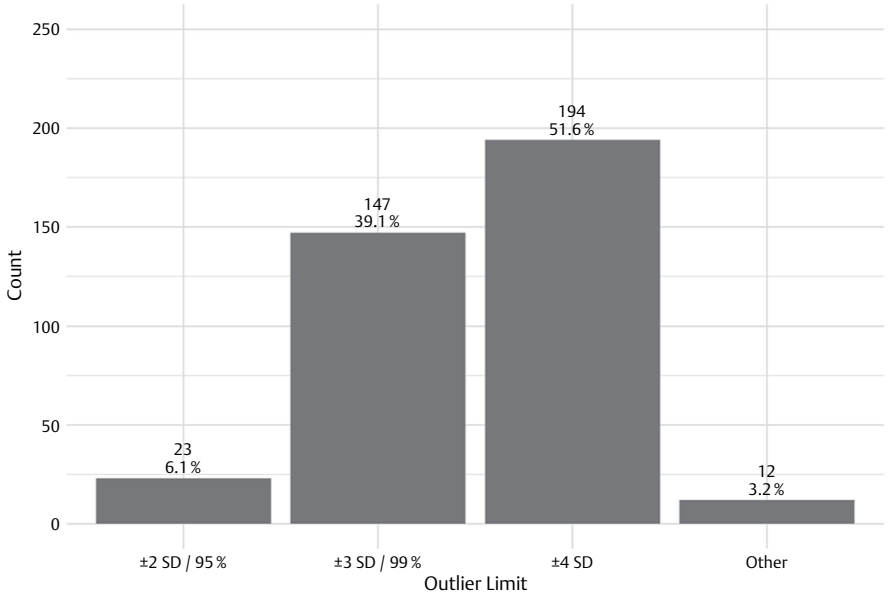
Discussion

Summary of evidence

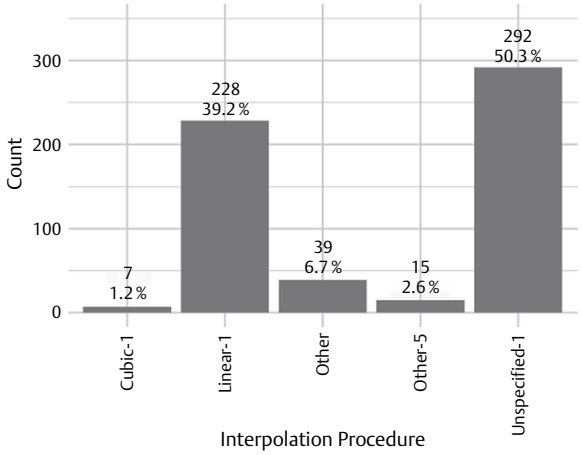
This review shows that gas exchange data processing methods are infrequently reported for outlier removal and interpolation. We consider outlier removal documentation important as it applies to many exercise test analyses. Outlier removal is the key for $\dot{V}O_2$ kinetics research requiring a high temporal resolution. Outlier removal is also relevant for maximal exercise testing as outliers near the end of a test may influence $\dot{V}O_2$max or $\dot{V}O_2$peak. Previous research indicates that a $\dot{V}O_2$max below the 20th percentile for age and sex increases the risk of all-cause mortality [44], so the accurate determinations of $\dot{V}O_2$max are important for individuals with low cardiorespiratory fitness: an erroneous breath yielding an overestimated $\dot{V}O_2$max may subdue the urgency to improve the cardiovascular health for low-fitness individuals. Although we are unaware of studies examining $\dot{V}O_2$max misclassification due to different outlier removal strategies, averaging duration can influence which patients are deemed eligible for heart transplantation [6].

Outliers could also affect mathematical $\dot{V}O_2$ plateau determinations. Such methods test if neighboring $\dot{V}O_2$ values or a $\dot{V}O_2$ vs. a time slope does not change or increase by more than a set rate (e.g., 50 mL/min) at the end of a maximal test. [10, 45–48]. Although data averaging dampens their influence, outliers present near the conclusion of a maximal test could plausibly interfere with mathematical $\dot{V}O_2$ plateau determinations.

We are currently unaware of research that has tested this, but outliers may interfere with submaximal thresholds found using al-



▶ **Fig. 2** Counts and percentages of outlier limits when specified.



▶ **Fig. 3** Most prevalent specified interpolation methods by both type and time.

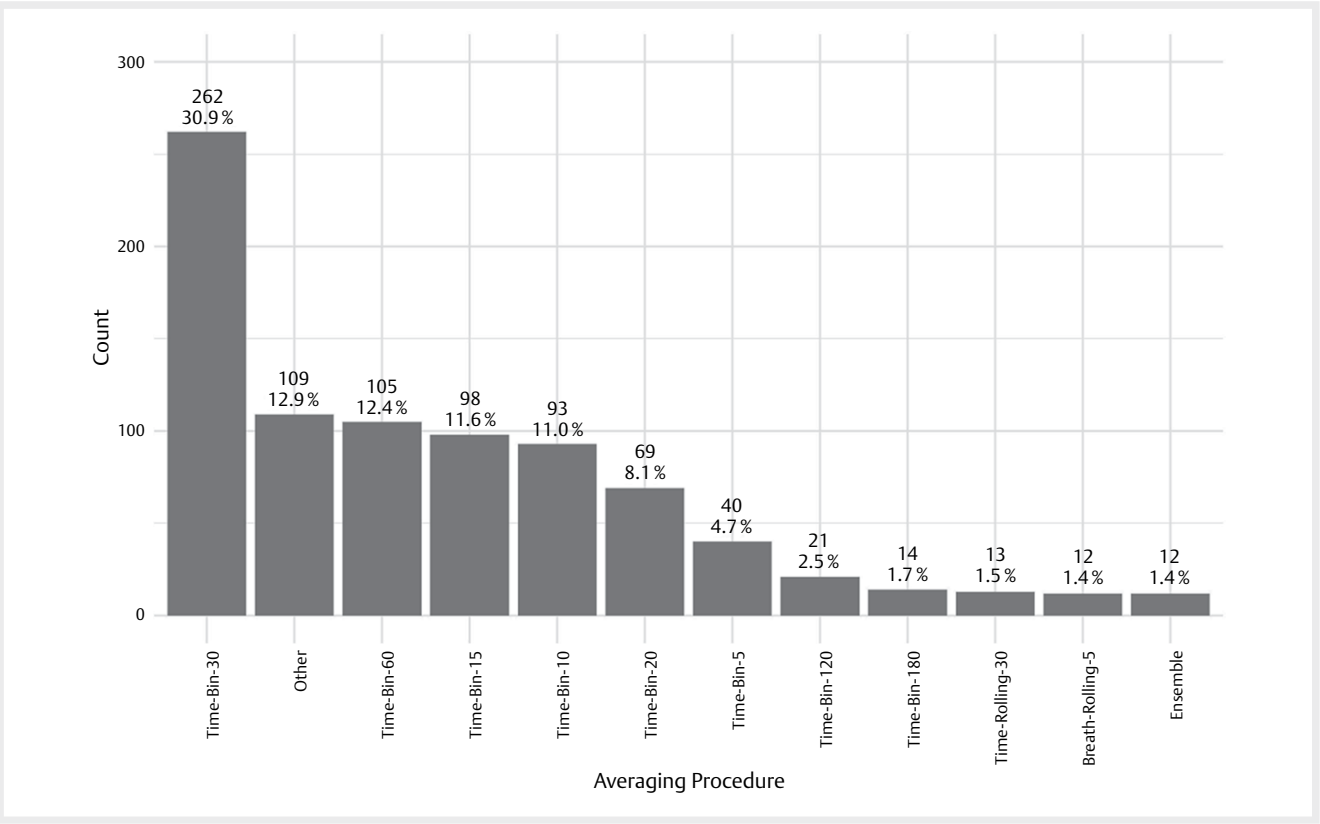
▶ **Table 1** (A) Averaging type by counts (*N*) and proportions (%). (B) Averaging subtype by counts (*N*) and proportions (%).

Averaging type	***N***	**%**
Time	776	91.5
Breath	38	4.5
Ensemble	13	1.5
Breath–time	8	0.9
Other	4	0.5
Digital filter	3	0.4
Time–time	3	0.4
Unaveraged	2	0.2
Time–breath	1	0.1
Table 1B		
Averaging subtype	*N*	%
Bin	744	89.9
Rolling	58	7.0
Rolling–bin	16	1.9
Bin–roll	7	0.8
Fast Fourier transform	2	0.2
Butterworth low-pass	1	0.1

gorithms, especially if they exist near likely breakpoints. Threshold algorithms often fit piecewise linear regressions and solve for the lowest sums of squares [49–51]. Points near the edges of the regression lines have more leverage when solving for the best-fit line and, therefore, are more likely to influence the slope or intercept. Such changes could alter the intersection point of the piecewise regression and thus the threshold values.

Finally, even fewer articles reported the outlier limit calculation function. As the function chosen impacts the calculated outlier limit, it also affects where values are considered outliers. The most popular strategy we observed was rolling breath averages and a standard deviation with 3–5 breaths and ± 3–4 SD. Although rarely noted, we suspect that these are popular as they are based on the published literature [20], are relatively easy to program, and may come pre-installed on some metabolic cart software. However, this should not preclude investigating other outlier functions within this domain or using digital filtering methods that specifically dampen high-frequency oscillations.

Lower interpolation reporting is expected, given its relevance primarily to less frequent $\dot{V}O_2$ kinetics studies. However, the V-slope method, one of the most common methods for determining the



▶ **Fig. 4** Prevalence of complete averaging procedures. The numbers in each column label are in seconds for time averages and the number of breaths for breath averages. The "other" column represents methods that accounted for less than 1 % of the total stated methods.

first ventilatory threshold, interpolates data in their original method [50]. Importantly, the V-slope algorithm is only a part of the overall V-slope method, so it can be unclear if authors interpolated data when citing the V-slope method. Given this and the artificial CI shrinkage, we recommend authors specify interpolation details. Although the uncertainty of the $\dot{V}O_2$ at thresholds is not commonly reported, doing so with appropriate CIs may be prudent as traditional thresholds may be better described as transitions [52, 53].

Most studies use 1-s linear interpolation, but different time frames and styles, such as cubic interpolation, may yield different results. Cubic spline interpolation produces a smooth curve but may slightly "overshoot" measured values [54]. Although the choice of linear vs. cubic interpolation may be small, we recommend authors specify the interpolation type for improved reproducibility.

While averaging methods are more frequently reported, about 40 % of studies lacked documentation. Data averaging likely contributes more to the final calculated values of $\dot{V}O_2$max and other variables than do outlier removal and interpolation as averaging suppresses potential outliers itself by combining those points with additional observations. Research on the effect of interpolation on $\dot{V}O_2$ kinetics parameters shows that interpolation does not significantly affect the values of parameter estimates [22–25]. Although we are unaware of studies comparing the effect of outlier removal or leaving data as-is before proceeding with other calculations, the known impact of data averaging on $\dot{V}O_2$max and the inherent dampening effect of averaging on outliers itself suggests that data averaging is the most important of the three steps when the goal is to reflect the underlying whole-body metabolic rate. Therefore, researchers should state their gas exchange data averaging methods to improve research reproducibility and study comparisons.

Stating averaging methods can also help correctly classify cardiorespiratory fitness against normative data. Research by Martin-Rincon et al. [11] offers a strategy to compare two $\dot{V}O_2$max values obtained with different averaging methods. Without such corrections, one could misclassify cardiorespiratory fitness based on $\dot{V}O_2$max if $\dot{V}O_2$max values were calculated with a sufficiently different sampling interval than that used to generate the normative data. However, this correction applies to group data [11] and different averaging strategies also affect individual $\dot{V}O_2$max values.

Importantly, the normative data offered by the American College of Sports Medicine [1] is based on a regression of $\dot{V}O_2$ vs. time-to-exhaustion using a modified Balke protocol and equations developed from ref [55] and [56] (Cooper Institute, personal communication, 9/2021), rather than directly measured. The system used to create the regression for males [56] and females [55] averaged the data every minute and every 30 s, respectively. Given this, stating the averaging methods used may allow for better comparisons to normative data.

The most frequent, fully specified data averaging method, the 30-s time average, fits the maximum recommended duration by Robergs [3]. However, Robergs et al. recommended a 30-s rolling rather than a bin average. The same study also recommended a 15-breath rolling average or, preferably, the low-pass digital filter, but we only documented these methods two (0.2) and one (0.1 %) times, respectively. Despite the most popular method not exceeding the maximum recommended duration, at least for calculating $\dot{V}O_2$max, the vast majority of papers neglected to follow current guidelines.

This review did not document all aspects of the recommended reporting guidelines proposed by Nolte et al. [19]. We did not record the exact metabolic cart used, the measurement mode, the software used, nor the data processing rationale. Our outlier reporting rate of 4.5 % was similar, while our interpolation reporting rate of 7.0 % was slightly higher than that of Nolte's 4.3 %. Our averaging reporting rate was also ~5 % higher at 60.9. Our results are similar to previous works showing that time-bin averages are the most popular, with 30-s averages as the most common overall [3, 8, 19]. The order in popularity of other methods differs from these other studies. This may be attributed to our wider date range and including data processing beyond that to describe the $\dot{V}O_2$max.

Although more complex data processing strategies have existed for several years, we suspect time-bin averages are still favored in part due to subjective reasons like tradition [3]. We speculate that practitioners also chose time-bin averages if more complex options are not integrated within metabolic cart software.

Limitations

While extensive, this study's scope limited detailed examination of each article, meaning that some data processing descriptions might have been missed due to the limitations of our RegExs, leading us to categorize these as "not described." Articles citing prior works for data processing were marked as "not described" for simplicity. Citing other works may help meet word counts, but methodological shortcut citations may hinder reproducibility [57]. The best practice yet would be for authors to publish all data and code when possible. Several open-source gas data software packages are available to facilitate transparent analyses [58–60].

Next, by chance, we found rare examples of articles using the median as the measure of center as we built our RegExs. However, we did not document any such cases in our random sample. A larger random sample would likely find these and other rare data averaging methods. Also, a few ineligible articles may have eluded our screening. We also predict that some studies may have used mixing chambers without specifying [19]. Taken together, our results may slightly underestimate data processing methods' true reporting frequency.

Another limitation of this scoping review is that our results do not indicate how different data processing methods were used. For example, we did not distinguish if a 60-s time-bin average was used to calculate $\dot{V}O_2$max or a steady-state exercise period. Therefore, this review cannot estimate the prevalence of different processing methods for specific analyses, such as $\dot{V}O_2$max. Previous research shows that 30-s averages for $\dot{V}O_2$max are the most common [3, 8, 19], presumably as they remove short-term fluctuations. We anticipate that $\dot{V}O_2$ kinetics and similar analyses which rely on high temporal resolution data likely employ averaging with less smoothing. Moderate smoothing may assist with detecting ventilatory thresholds as an optimal signal-to-noise ratio may highlight the systematic change points between ventilatory variables. Nevertheless, this is the first study we know of to document data processing methods besides those used to calculate the $\dot{V}O_2$max.

Conclusions

This scoping review found that data processing methods were seldom reported for outlier removal and interpolation and that averaging reporting, although much higher, could further improve. The results reflect prevalent methods. While prevalence should not be conflated with quality, knowing prevalent methods allows testing their influence on data processing. We hope that these results motivate better methodological documentation and reproducibility in this field.

Acknowledgement

We thank library staff Scott Marsalis and Cody Hennesy for their advice and support.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Pescatello LS. ACSM's guidelines for exercise testing and prescription. 9th ednPhiladelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health; 2014: 162

[2] Jamnick NA, Pettitt RW, Granata C et al. An Examination and Critique of Current Methods to Determine Exercise Intensity. Sports Med 2020; 50: 1729–1756. DOI: 10.1007/s40279-020-01322-8

[3] Robergs RA, Dwyer D, Astorino T. Recommendations for Improved Data Processing from Expired Gas Analysis Indirect Calorimetry. Sports Med 2010; 40: 95–111. DOI: 10.2165/11319670-000000000-00000

[4] Robergs RA, Burnett AF. Methods Used to Process Data from Indirect Calorimetry and Their Application to $\dot{V}O_2$ Max. J Exerc Physiol Online 2003; 6: 44–57

[5] Sousa A, Figueiredo P, Oliveira N et al. Comparison Between Swimming $\dot{V}O_2$peak and $\dot{V}O_2$max at Different Time Intervals. Open Sports Sci J 2010; 3: 22–24. DOI: 10.2174/1875399X01003010022

[6] Johnson JS, Carlson JJ, VanderLaan RL et al. Effects of Sampling Interval on Peak Oxygen Consumption in Patients Evaluated for Heart Transplantation. Chest 1998; 113: 816–819. DOI: 10.1378/chest.113.3.816

[7] Sell KM, Ghigiarelli JJ, Prendergast JM et al. Comparison of $\dot{V}O_2$ peak and $\dot{V}O_2$ max at Different Sampling Intervals in Collegiate Wrestlers. J Strength Cond Res 2021; 35: 2915–2917. DOI: 10.1519/JSC.0000000000003887

[8] Midgley AW, McNaughton LR, Carroll S. Effect of the $\dot{V}O_2$ time-averaging interval on the reproducibility of $\dot{V}O_2$ max in healthy athletic subjects. Clin Physiol Funct Imaging 2007; 27: 122–125. DOI: 10.1111/j.1475-097X.2007.00725.x

[9] Astorino TA. Alterations in $\dot{V}O_2$ max and the $\dot{V}O_2$ plateau with manipulation of sampling interval. Clin Physiol Funct Imaging 2009; 29: 6067. DOI: 10.1111/j.1475-097X.2008.00835.x

[10] Astorino TA, Robergs RA, Ghiasvand F et al. Incidence of the oxygen plateau at $\dot{V}O_2$max during exercise testing to volitional fatigue. J Exerc Physiol Online 2000; 3: 112

[11] Martin-Rincon M, González-Henríquez JJ, Losa-Reyna J et al. Impact of data averaging strategies on $\dot{V}O_2$ max assessment: mathematical modeling and reliability. Scand J Med Sci Sports 2019; 29: 1473–1488. DOI: 10.1111/sms.13495

[12] Martin-Rincon M, Calbet JAL. Progress Update and Challenges on $\dot{V}O_2$max Testing and Interpretation. Front Physiol 2020; 11: 1070. DOI: 10.3389/fphys.2020.01070

[13] Scheadler CM, Garver MJ, Hanson NJ. The Gas Sampling Interval Effect on $\dot{V}O_2$peak Is Independent of Exercise Protocol. Med Sci Sports Exerc 2017; 49: 1911–1916. DOI: 10.1249/MSS.0000000000001301

[14] de Jesus K, Guidetti L, de Jesus K et al. Which Are the Best $\dot{V}O_2$ Sampling Intervals to Characterize Low to Severe Swimming Intensities? Int J Sports Med 2014; 35: 1030–1036. DOI: 10.1055/s-0034-1368784

[15] Hill DW, Stephens LP, Blumoff-Ross SA et al. Effect of sampling strategy on measures of $\dot{V}O_2$peak obtained using commercial breath-by-breath systems. Eur J Appl Physiol 2003; 89: 564–569. DOI: 10.1007/s00421-003-0843-1

[16] Smart NA, Jeffriess L, Giallauria F et al. Effect of duration of data averaging interval on reported peak $\dot{V}O_2$ in patients with heart failure. Int J Cardiol 2015; 182: 530–533. DOI: 10.1016/j.ijcard.2014.12.174

[17] Matthews JI, Bush BA, Morales FM. Microprocessor Exercise Physiology Systems vs a Nonautomated System. Chest 1987; 92: 696–703. DOI: 10.1378/chest.92.4.696

[18] Poole DC, Jones AM. Measurement of the maximum oxygen uptake $\dot{V}O_{2max}$: $\dot{V}O_{2peak}$ is no longer acceptable. J Appl Physiol 2017; 122: 997–1002. DOI: 10.1152/japplphysiol.01063.2016

[19] Nolte S, Rein R, Quittmann OJ. Data Processing Strategies to Determine Maximum Oxygen Uptake: a Systematic Scoping Review and Experimental Comparison with Guidelines for Reporting. Sports Med 2023; 53: 2463–2475. DOI: 10.1007/s40279-023-01903-3

[20] Lamarra N, Whipp BJ, Ward SA et al. Effect of interbreath fluctuations on characterizing exercise gas exchange kinetics. J Appl Physiol 1987; 62: 20032012. DOI: 10.1152/jappl.1987.62.5.2003

[21] Keir DA, Murias JM, Paterson DH et al. Breath-by-breath pulmonary O_2 uptake kinetics: effect of data processing on confidence in estimating model parameters. Exp Physiol 2014; 99: 15111522. DOI: 10.1113/expphysiol.2014.080812

[22] Benson AP, Bowen TS, Ferguson C et al. Data collection, handling, and fitting strategies to optimize accuracy and precision of oxygen uptake kinetics estimation from breath-by-breath measurements. J Appl Physiol 2017; 123: 227242. DOI: 10.1152/japplphysiol.00988.2016

[23] Francescato MP, Cettolo V, Bellio R. Confidence intervals for the parameters estimated from simulated O_2 uptake kinetics: effects of different data treatments. Exp Physiol 2014; 99: 187195. DOI: 10.1113/expphysiol.2013.076208

[24] Francescato MP, Cettolo V. The 1-s interpolation of breath-by-breath O_2 uptake data to determine kinetic parameters: the misleading procedure. Sport Sci Health 2019; 16: 193. DOI: 10.1007/s11332-019-00602-9

[25] Francescato MP, Cettolo V, Bellio R. Interpreting the confidence intervals of model parameters of breath-by-breath pulmonary O_2 uptake. Exp Physiol 2015; 100: 475475. DOI: 10.1113/EP085043

[26] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean. Sci Transl Med 2016; 8: 341ps12. DOI: 10.1126/scitranslmed.aaf5027

[27] Open Science Collaboration. Estimating the reproducibility of psychological science. Science (1979) 2015; 349: aac4716. DOI: 10.1126/science.aac4716

[28] Tricco AC, Lillie E, Zarin W et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018; 169: 467–473. DOI: 10.7326/M18-0850

[29] Peters MDJ, Marnie C, Tricco AC et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020; 18: 2119–2126. DOI: 10.11124/JBIES-20-00167

[30] Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021; 18: e1003583. DOI: 10.1371/journal.pmed.1003583

[31] Foster ED, Deardorff A. Open Science Framework (OSF). J Med Libr Assoc 2017; 105. DOI: 10.5195/jmla.2017.88

[32] Blinded for peer review. Scoping review of gas exchange data processing procedures in published literature. 2022. DOI: 10.17605/OSF.IO/A4VMZ

[33] Blinded for peer review. Data Processing Methods and their Effects on the Limits of Agreement and Reliability of Automated Submaximal Threshold Calculations. 2023.

[34] NCBI. ID Converter API. 2022; https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/ [stand: 6.4.2022].

[35] Unpaywall: an open database of 20 million free scholarly articles. https://unpaywall.org/ [stand: 24.7.2024]

[36] Haddaway NR, Page MJ, Pritchard CC et al. PRISMA2020 : An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews 2022; 18: e1230. DOI: 10.1002/cl2.1230

[37] Bojanowski P, Grave E, Joulin A et al. Enriching word vectors with subword information. arXiv Preprint 2016: arXiv:160704606

[38] Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12: 28252830

[39] Breese BC, Saynor ZL, Barker AR et al. Relationship between (non) linear phase II pulmonary oxygen uptake kinetics with skeletal muscle oxygenation and age in 11–15 year olds. Exp Physiol 2019; 104: 1929–1941. DOI: 10.1113/EP087979

[40] Hartman ME, Ekkekakis P, Dicks ND et al. Dynamics of pleasure-displeasure at the limit of exercise tolerance: conceptualizing the sense of exertional physical fatigue as an affective response. J Exp Biol 2018; 222: jeb.186585. DOI: 10.1242/jeb.186585

[41] Hassinen M, Lakka TA, Savonen K et al. Cardiorespiratory Fitness as a Feature of Metabolic Syndrome in Older Men and Women. Diabetes Care 2008; 31: 1242–1247. DOI: 10.2337/dc07-2298

[42] Deboeck G, Niset G, Lamotte M et al. Exercise testing in pulmonary arterial hypertension and in chronic heart failure. Eur Respir J 2004; 23: 747–751. DOI: 10.1183/09031936.04.00111904

[43] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

[44] Blair SN, Kohl HW, Barlow CE et al. Changes in physical fitness and all-cause mortality: a prospective study of healthy and unhealthy men. JAMA 1995; 273: 10931098. DOI: 10.1001/jama.1995.03520380029031

[45] Robergs RA. An exercise physiologist's "contemporary" interpretations of the" ugly and creaking edifices" of the $\dot{V}O_2$max concept. J Exerc Physiol Online 2001; 4: 144

[46] Myers J, Walsh D, Buchanan N et al. Can maximal cardiopulmonary capacity be recognized by a plateau in oxygen uptake? Chest 1989; 96: 1312–1316. DOI: 10.1378/chest.96.6.1312

[47] Myers J, Walsh D, Sullivan M et al. Effect of sampling on variability and plateau in oxygen uptake. J Appl Physiol 1990; 68: 404–410. DOI: 10.1152/jappl.1990.68.1.404

[48] Yoon B-K, Kravitz L, Robergs R. $\dot{V}O_2$max, protocol duration, and the $\dot{V}O_2$ plateau. Med Sci Sports Exerc 2007; 39: 11861192

[49] Jones RH, Molitoris BA. A statistical method for determining the breakpoint of two lines. Anal Biochem 1984; 141: 287–290. DOI: 10.1016/0003-2697(84)90458-5

[50] Beaver WL, Wasserman K, Whipp BJ. A new method for detecting anaerobic threshold by gas exchange. J Appl Physiol 1986; 60: 2020–2027. DOI: 10.1152/jappl.1986.60.6.2020

[51] Orr GW, Green HJ, Hughson RL et al. A computer linear regression model to determine ventilatory anaerobic threshold. J Appl Physiol 1982; 52: 1349–1352. DOI: 10.1152/jappl.1982.52.5.1349

[52] Pethick J, Winter SL, Burnley M. Physiological evidence that the critical torque is a phase transition, not a threshold. Med Sci Sports Exerc 2020; 52: 2390–2401. DOI: 10.1249/MSS.0000000000002389

[53] Ozkaya O, Balci GA, As H et al. Grey zone: a Gap Between Heavy and Severe Exercise Domain. J Strength Cond Res 2022; 36: 113–120. DOI: 10.1519/JSC.0000000000003427

[54] Zhang Z, Martin CF. Convergence and Gibbs' phenomenon in cubic spline interpolation of discontinuous functions. J Comput Appl Math 1997; 87: 359–371. DOI: 10.1016/S0377-0427(97)00199-4

[55] Pollock ML, Foster C, Schmidt D et al. Comparative analysis of physiologic responses to three different maximal graded exercise test protocols in healthy women. Am Heart J 1982; 103: 363373

[56] Pollock ML, Bohannon RL, Cooper KH et al. A comparative analysis of four protocols for maximal treadmill stress testing. Am Heart J 1976; 92: 39–46. DOI: 10.1016/S0002-8703(76)80401-2

[57] Standvoss K, Kazezian V, Lewke BR et al. Shortcut citations in the methods section: frequency, problems, and strategies for responsible reuse. PLoS Biol 2024; 22: e3002562. DOI: 10.1371/journal.pbio.3002562

[58] Mattioni Maturana F. whippr: tools for manipulating gas exchange data. 2024.

[59] Hesse A. gasExchangeR: analyze gas exchange data from cardiopulmonary exercise tests. 2023.

[60] Nolte S. spiro: An R package for analyzing data from cardiopulmonary exercise testing. J Open Source Softw 2023; 8: 5089. DOI: 10.21105/joss.05089