International Journal of Sports Medicine
Thieme

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

Journal:	*International Journal of Sports Medicine*
Manuscript ID	IJSM-07-2024-10699-re.R1
Manuscript Type:	Review
Key word:	Averaging, Outliers, Interpolation, Cardiopulmonary exercise testing, Breath-by-breath, Reproducibility
Abstract:	Cardiopulmonary exercise testing involves collecting variable breath-by-breath data, 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 VO_2max. 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%, 95% CI: 4.1-5.0%) and 581 (7.0%, 6.4-7.5%) described outlier removal and interpolation, respectively. A random subset of 1,078 articles revealed 60.9% (57.9-63.7%) reported averaging methods. Commonly documented outlier cutoffs were ± 3 or 4 SD (39.1% and 51.6%, respectively). The dominating interpolation duration and procedure were one second (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.

The prevalence of gas exchange data processing methods: a semi-automated scoping review ~~Popularity And Prevalence Of Gas Exchange Data Processing Methods: A Scoping Review~~

Abstract

Cardiopulmonary exercise testing involves collecting variable breath-by-breath data, sometimes requiring data processing of outlier removal, interpolation, and averaging before later analysis. These data processing choices, such as averaging duration, ~~are known to~~ affect calculated values such as VO_2max. However, assessing the implications~~effects~~ 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~~351~~ articles, 376 (mean 4.5%, 95% CI: ~~± 0.~~4.1-5.0%) and 581 (7.0%, 6.4-7 ~~± 0~~.5%) described outlier removal and interpolation, respectively. A random subset of 1,078 articles revealed 60.9% (57.9-63.7%)~~An estimated 66.8 ± 2.8%~~ reported averaging methods. ~~(n = 1078).~~ Commonly documented outlier cutoffs were ± 3 or 4 SD (39.1% and 51.6%, respectively). The dominating interpolation duration and procedure were one second (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~~methods~~ will enhance reproducibility and facilitate future research comparing data processing choices.

1. Introduction

Clinicians and researchers ~~commonly~~ use cardiopulmonary exercise testing (CPET) to determine maximal aerobic capacity (VO_2max), ventilatory thresholds, and VO_2 kinetics. Such values help categorize fitness, predict disease risk, and guide exercise [1~~, p. 162~~]. Using CPET results to guide exercise, especially relative to thresholds, yields~~produces~~ better improvements given~~due to~~ 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 requires processing to manage its high variability [3]. BBB VO_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 at. [3] showed that the rate and depth of breathing account for most of the variation in VO_2 during both steady states and incremental exercise. Therefore,~~Calculating the above values often requires data processing when~~

~~CPET data is collected breath-by-breath (BBB) as it is highly variable [3].~~ 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 VO_2max and VO_2 plateau detection [3–17]. The importance of attaining a VO_2 plateau to accurately determine VO_2max has been highlighted because "secondary" VO_2max criteria often underestimate VO_2max [18], and thus misclassify cardiorespiratory fitness. Finally, we~~We~~ are unaware of research on the effects of data processing and locating ventilatory thresholds.

Outlier removal typically excludes~~Many studies remove outliers by finding~~ points beyond 2-~~±3 or ±~~4 standard deviations (SD) from~~beyond~~ the local mean [19].~~(i.e., a prediction interval).~~ 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 [2018]. More outliers appear than expected because of both conscious and unconscious alterations of breathing patterns, including swallowing and coughing [2018]. We are unaware of prior research that examines how different outlier removal strategies affect VO_2max, ventilatory thresholds, and VO_2 kinetics.

Interpolation, often to one-second intervals, is common in VO_2 kinetics research to "ensemble" average repeated transitions to minimize variability [20,2118,19]. Although this does not affect parameter estimates, one-second interpolation has been criticized for artificially narrowing confidence intervals as respiratory rates are usually below 60 breaths per minute, even near maximal exercise [22–25].~~[20–23].~~ As before, we are unaware of research specifically investigating how interpolation affects VO_2max and ventilatory threshold identification.

Data processing choices, such as averaging and interpolation, impact CPET variables or their confidence intervals. Earlier~~Existing~~ surveys [3] and studies [8] were~~are~~ small and focused on averaging methods only, finding time-based bin averages (e.g., 30-second averages) were popular. A more recent scoping review by Nolte et al. [19] found that nearly half of studies on VO2max 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].~~A larger sample can, therefore, better describe how often all data processing steps are described.~~

These low rates of reporting data processing steps~~Before conducting this scoping review, we anecdotally observed that many articles using CPET data did not report all data processing steps, especially outlier removal and interpolation details. This~~ 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 VO_2max 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~~24,25]~~ ~~Therefore, to assist with conducting future research on the effects of data processing on~~

~~CPET values and to evaluate the methodological reproducibility of research using BBB gas exchange data generally, we conducted a broad scoping review to identify the frequency of reporting, and popularity of outlier removal, interpolation, and data averaging methods~~.

This research expands on that of Nolte et al. [19] but searches without date restriction and includes studies with CPET data beyond VO_2max 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 documents the data processing choices worth testing in future research investigating the effects of such choices on CPET values.

2. Methods

2.1 Protocol~~Design~~ and Registration

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

2.2 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.~~It is based on a dissertation chapter by the first author [26]. These methods [27] and results [28] are modeled on the PRISMA scoping review extension guidelines.~~ Eligible articles were original, peer-reviewed articles, with BBB gas exchange data, human participants, in English, ~~and~~ with a DOI. We imposed no date restriction to be as comprehensive as possible.

2.3~~2~~ Information Sources and Search

Data were collected~~We acquired data~~ from ~~the~~ Ovid-MEDLINE, Scopus, and Web of Science on 2022-06-27 ~~databases~~ with ~~the guidance of a university~~ librarian assistance. The electronic search strategies~~strategy~~ for all databases ~~the Ovid-MEDLINE database~~ can be found in the information sources and search section of the supplemental materials.

Our search output comprised article identifiers like DOIs. To find missing DOIs, we employed the PubMed Central ID Converter API [34~~29~~] using Python. Full texts were accessed via publisher text and data mining APIs using Python, unpaywall.org [35~~30~~] using the unpywall Python package, through custom-built web-scraping scripts, or manually. Our library subscription did not permit access to 1,549 articles.

2.43 Selection of Sources of Evidence

This study used a single screening process, requiring because it differs from most scoping reviews. It only requires an exercise test with BBB gas exchange data collection with exercise. The corresponding PRISMA flow diagram was created usingrather than a more complex assessment of the PRISMA2020 R package [36].overall methodology and intervention.

2.43.1 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.RegExs described below. These ML classifiers and RegExs helped identify ineligible articles. This computerized screening required converting full-text PDF and EPUB documents into plain text files. Plain text files were normalized by transforming text to lowercase, removing hyphenations and extra whitespace, and correcting some plain text conversion-induced errors.

Following the normalization, we identified and removed articles that failed to correctly convert into text format, spotted non-English articles using the fasttext Python module [3731] and employed a random forest classifier from the sklearn Python package [3832] to detect ineligible articles based on our criteria. SeeWe manually reviewed potentially ineligible articles flagged by the supplemental methods for text analysis and screening for additional ML model detailsclassifier.

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. Breath-by-breath 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 the data was collected BBB. In total, we identified 8,412417 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.

2.43.2 Data Charting Process

RegExs identified the presence of short phrases likely indicating that the authors described these methodological details. If present, we extracted a “snippet” of 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 hyphenation 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 and at the beginning or end of a word or sentence because the RegExs extracted a specific number of characters rather than words.

All~~We analyzed all~~ eligible BBB articles were analyzed for outlier and interpolation methods due to distinct descriptions and ~~because~~ fewer total articles ~~described these methods~~ (~5%).~~%) and the phrases were more distinct.~~ 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~~050~~ 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% confidence interval and a maximum margin of error (MOE) of ±3%, assuming a proportion of 0.5 for a conservative estimate. However, we raised 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 accuracy and the required corresponding samples: decreasing the MOE to ±2% with an assumed proportion of 0.5 would require another 1,333 samples.

2.~~4~~3.2.1 Outliers

Our outlier RegExs identified phrases like “swallowing”, “coughing”, “errant”, “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 ˙vo2 data from each step transition were initially edited to exclude errant breaths by removing values lying more than 4 sd” from [39~~33~~]. 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 ˙vo2 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.

2.43.2.2 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 [4034].

> the v̇ o2 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 v̇ o2 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 one second.

2.43.2.3 Averaging

We document averaging methods according to five criteria: type/units, subtype/calculation, amount, measure of center, and mean type. (Figure 1). 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 seconds but is 30 breaths for a breath average. Measure of center distinguishes between mean or median, and mean type delineates whole vs. trimmed mean. Trimmed (truncated) means exclude 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-second averages” and “averaged every 30 seconds” 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 VO_2max. We also recorded multiple averaging methods when the authors described the sampling interval and the transformation applied to it. For example, the snippet from [4135]

> ath method using the vmax respiratory gas analyzer (sensormedics, yorba linda, ca). vo2max 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 seconds and that VO_2max was calculated as the average of three 20-second intervals, or 60-seconds. For this article,

we documented one averaging method as a 20-second time bin whole mean and another as a 60-second time bin whole mean.

In many cases, authors did not explicitly use the terms “average” or “mean” to describe their averaging methods, but we documented their methods when implied. For example, the snippet from [42~~36~~] reading

> 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 VO_2peak using the last 30 seconds of exercise data. We documented such phrasing as a 30-second time-bin whole mean average.

2.~~4~~3.3 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, those 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.

2.~~4~~3.3.1 Outliers

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

2.~~4~~3.3.2 Interpolation

We recorded the interpolation type (linear, cubic, Lagrange, specifically *un*interpolated, and other) and time frame (e.g., every one second).

2.~~4~~3.3.3 Averaging

We noted the following averaging types: Time, breath, breath-time, time-breath, time-time, digital filter, ensemble, (explicitly) *un*averaged, and other. Averaging subtypes included bin, rolling, bin-roll, 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 mean or median. Finally, we noted if the mean was a whole or trimmed.

2.~~4~~3.4 Synthesis of Results

Counts, proportions~~percentages~~, and Agresti-Coull ~~margin of error (~~95% confidence intervals~~)~~ were calculated for the reporting frequency of each data processing method using R R version 4.1.2 [43] in the~~[37] and~~ RStudio IDE version 2023.6.1.524 [43].

When articles reported multiple methods, we only counted this article once for calculating overall reporting proportions.[38].

3. Results

3.1 Selection of Sources of Evidence

Figure 12 shows the selection of sources of evidence flowchart. The initial searchDuring our analysis, we identified 50,7301,352 ineligible articles; 21,715 remained for retrieval after removing duplicates and . We cross-referenced those without DOIs. A total of 8,344against the breath-by-breath 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.and removed another 354, leading to 8,351 articles.

3.2 Characteristics and Results of Individual Sources of Evidence

The PRISMA Extension for Scoping Reviews checklist normally requires a section to report the characteristics and results of individual sources of evidence, usually in a table format, including citations [27]. Given the vast nature of this scoping review, readers can instead view web links to our outlier, interpolation, and averaging data charting spreadsheets.

3.3 Synthesis of Results

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

3.3.1 Outliers

Of the 8,344351 articles, 376 (4.5%, 95% CI: ± 0.4.1-5.0%) reported outlier removal methods. TheOf the articles reporting their outlier methods, the most prevalent reported methods were ±3 (39.1%) and ±4 (51.6%) SDstandard deviations, respectively (Figure 23).

Only 102 (1.2%, 95% CI: 1.± 0-1.5.2%) articles reported details of the function they used to calculate their outlier limit. Of those, breath-based averages (n = 76, 74.5%) 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.

3.3.2 Interpolation

Of 8,344 articles We found that 581 (7.0%, 95% CI: 6.4-7.± 0.5%) out of 8,351 specified their interpolation, with one-second intervals as methodology. When reported, the most common (interpolation time was one second (n = 527, 93.9%). Around half of reportedAlthough the majority of articles reporting interpolation procedures included the did not explicitly specify their interpolation method (n = 314, 54.0%), with linear

interpolation aswas the most popular stated method (n = 247, 92.5%) (see Table 1 and Figure 34).

Table 1: Most prevalent specified interpolation methods by type (a) and by time (b).

3.3.3 Averaging

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. We recorded 656 (60.9%, 95% CI: 57.9-63.7%) articles withthat 852 (66.8 ± 2.8%) reported some details of their data averaging methods. Of these, 14 articles reported more than one averaging method. Time averages dominated in popularity (91.5%) (Table 12). Bin averages proved the most widespread averaging subtype (89.9%) (Table 12). Together, time-bin (86.8%) was the most frequent type-subtype averaging method combination.

Table 12: Averaging methods by type (a) and subtype (b).

When incorporating averaging amounts, 30-, 60-, 15-, and 10-second bin averages (Figure 45) 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.

4. Discussion

4.1 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 removalRemoving outliers is key forimportant to VO_2 kinetics and similar research requiringwith rapid intensity changes because they rely on high temporal resolution. Outlier removal is also relevant for maximal exercise testing as outliers near the end of a test may influence VO_2max or VO_2peak. Previous research indicates that a VO_2max below the 20th percentile for age and sex increases the risk of all-cause mortality [4439], so accurate determinations of VO_2max are important for individuals with low cardiorespiratory fitness: an erroneous breath yielding an overestimated VO_2max may subdue the urgency to improve cardiovascular health for low-fitness individuals. Although we are unaware of studies examining VO_2max 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 VO_2 plateau determinations. Such methods test if neighboring VO_2 values or a VO_2 vs. 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. [10,45–4840–43]. Though data averaging dampens their influence, outliers present near the conclusion of a maximal test could plausibly interfere with mathematical VO_2 plateau determination.

We are currently unaware of research that has tested this, but outliers may interfere with submaximal thresholds found using algorithms, especially if they exist near likely breakpoints. Threshold algorithms often fit piecewise linear regressions and solve for the lowest sums of squares [49–5144–46]. 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 calculated outlier limit, it also affects where values are considered outliers. The most popular strategy we observed were rolling breath averages and standard deviation with 3-5 breaths and ±3-4 SD. Though rarely noted, we suspect these are popular as they are based on 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~~We are unaware of a recommended outlier removal function but encourage stating such details~~.

Lower~~We find the low~~ interpolation reporting ~~more reasonable because this procedure~~ is expected, given its relevance primarily~~most relevant~~ to less frequent VO_2 kinetics studies. However, the V-slope method, one of the most common methods for determining the first ventilatory threshold, interpolates data in their original method [5045]. Importantly, the V-slope algorithm is only 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 confidence interval shrinkage, we recommend authors specify interpolation details. Although the uncertainty of the VO_2 at thresholds are not commonly reported, doing so with appropriate confidence intervals may be prudent as traditional thresholds may be better described as transitions [52,53].~~it may be prudent for future papers to specify interpolation or lack thereof~~.

Most studies use one-second 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 [5447]. Though the choice of linear vs. cubic interpolation may be~~likely~~ small, we recommend authors specify the interpolation type for improved reproducibility.

While~~Despite a much higher percentage of papers describing at least some of their~~ averaging methods are more frequently reported, about 40%~~, a third~~ of studies lacked documentation~~examined in this review neglected to document their process~~. Data averaging likely contributes more to the final calculated values of VO_2max and other variables than do outlier removal and interpolation as averaging suppresses potential outliers itself by combining those points with additional observations. Research~~. Indeed, the research~~ on the effect of interpolation on VO_2 kinetics parameters shows that interpolation does not significantly affect the values of parameter estimates [22–2520–23]. 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 VO_2max 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~~[12~~] offers a strategy to compare two VO_2max values obtained with different averaging methods. Without such corrections, one could misclassify cardiorespiratory fitness based on VO_2max if VO_2max 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 VO_2max values.

Importantly, the normative data offered by the American College of Sports Medicine [1~~, table 4.9, pp. 88-93~~] is based on a regression of VO_2 vs. time-to-exhaustion using a modified Balke protocol and equations developed from [55~~48~~] and [56~~49~~] (Cooper Institute, personal communication, 9/2021), rather than directly measured. The system used to create the regression for males [56~~49~~] and females [55~~48~~] averaged the data every minute and every 30 seconds, respectively. Given that, stating the averaging methods used may allow for better comparisons to normative data.

The most frequent, fully specified data averaging method, the 30-second time average, fits the maximum recommended duration~~guideline~~ by Robergs [3]. However, Robergs et al. recommended a 30-s rolling rather than a bin average. The same study~~[3]~~ also recommended ~~a~~the 15-breath rolling average or, preferably, the low-pass digital filter, but we only documented these methods two (0.2 ~~±0.3%)~~ and one (0.1 ~~±0.2~~%) times, respectively. Despite the most popular method not exceeding the maximum recommended duration, at least for calculating VO_2max, 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 rates of 4.5% was similar, while our interpolation reporting rate of 7.0% was slightly higher than 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-second 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 VO_2max.

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.

4.2 Limitations

While~~This study presents the most~~ extensive, this study's ~~review of gas exchange data processing methods to date. However, due to its~~ scope limited, ~~not every article received a~~ detailed examination of each article, meaning, ~~which means~~ 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 priorthat referred to previous works for their data processing techniques were also marked as "not described" for simplicity. Citing other work may help meetWe realize authors must balance adequate methodological documentation with journal word counts, butor character limits. Yet, 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 are available to facilitate transparent analyses [58–60].

can mean missing details that prevent readers from fully reproducing the methods used [50]. 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,Finally, it is possible that a few ineligible articles may have eluded our screening. We also predict some studies may have used mixing chambers without specifying [19]. Taken together, our results are not entirely comprehensive and 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-second time-bin average was used to calculate VO_2max or a steady-state exercise period. Therefore, this review cannot estimate the prevalence of different processing methods for specific analyses, such as VO_2max. Previous research shows that 30-second averages for VO_2max are the most common [3,8,19], presumably as they remove short-term fluctuations. We anticipate that VO_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 changepoints between ventilatory variables. Nevertheless, this is the first study we know of to document data processing methods *besides* those used to calculate VO_2max.

4.3 Conclusions

This scoping review found that data processing methods were seldom reported for outlier removal and interpolation, and that averaging reporting, though much higher, could further improve. The results reflect prevalent methods. While prevalence should not be conflated with quality, knowing the prevalent methods allows testing their can allow others to test the influence on data processing. We in this field by comparing relevant options. Finally, we hope these results motivate better others to improve their methodological documentation and, thus, reproducibility in this field.

References

[1] Pescatello LS. ACSM's guidelines for exercise testing and prescription. 9th ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health; 2014 p. 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 VO2 Max. J Exerc Physiol Online 2003; 6

[5] Sousa A, Figueiredo P, Oliveira N, et al. Comparison Between Swimming VO2peak and VO2max 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 V̇o 2peak and V̇o 2max 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 V̇O2 time-averaging interval on the reproducibility of V̇O2max 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 VO2 max and the VO2 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 VO2max 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 V̇O2max 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 VO2max 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 VO2peak Is Independent of Exercise Protocol. Med Sci Sports Exerc 2017; 49: 1911–1916. doi:10.1249/MSS.0000000000001301

[14] Jesus K de, Guidetti L, Jesus K de, et al. Which Are The Best VO2 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 VO2peak 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 VO2 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 V̇O $_{2max}$: V̇O $_{2peak}$ 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~~[18]~~] 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~~19~~] Keir DA, Murias JM, Paterson DH, et al. Breath-by-breath pulmonary O2 uptake kinetics: effect of data processing on confidence in estimating model parameters. Exp Physiol 2014; 99: 15111522. doi:10.1113/expphysiol.2014.080812

[22~~20~~] 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~~21~~] Francescato MP, Cettolo V, Bellio R. Confidence intervals for the parameters estimated from simulated O2 uptake kinetics: effects of different data treatments. Exp Physiol 2014; 99: 187195. doi:10.1113/expphysiol.2013.076208

[24~~22~~] Francescato MP, Cettolo V. The 1-s interpolation of breath-by-breath O2 uptake data to determine kinetic parameters: the misleading procedure. Sport Sci Health 2019; 16: 193. doi:10.1007/s11332-019-00602-9

[25~~23~~] Francescato MP, Cettolo V, Bellio R. Interpreting the confidence intervals of model parameters of breath-by-breath pulmonary O2 uptake. Exp Physiol 2015; 100: 475475. doi:10.1113/EP085043

[26~~24~~] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med 2016; 8. doi:10.1126/scitranslmed.aaf5027

[2725] Open Science Collaboration. Estimating the reproducibility of psychological science. Science (1979) 2015; 349: aac4716. doi:10.1126/science.aac4716

[28[26] *blinded for peer review*

[27] 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

[2928] 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; Im Internet: https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/; Stand: 06.04.2022

[35] Unpaywall: an open database of 20 million free scholarly articles. Im Internet: https://unpaywall.org/; Stand: 24.07.2024

[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[29] NCBI. ID Converter API. PubMed Central (PMC) 2022; Im Internet: https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/; Stand: 06.04.2022

[30] Unpaywall: An open database of 20 million free scholarly articles. Im Internet: https://unpaywall.org/; Stand: 24.07.2024

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

[3832] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12: 28252830

[3933] 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

[4034] 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; jeb.186585. doi:10.1242/jeb.186585

[4135] 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

[4236] 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

[4337] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021

[44[38] Posit team. RStudio: Integrated Development Environment for R. Boston, MA: Posit Software, PBC; 2022

[39] 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

[4540] Robergs RA. An exercise physiologist's "contemporary" interpretations of the" ugly and creaking edifices" of the VO2max concept. J Exerc Physiol Online 2001; 4: 144

[4641] 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

[4742] 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

[4843] Yoon B-K, Kravitz L, Robergs R. VO2maxV O2max, protocol duration, and the VO2V O2 plateau. Med Sci Sports Exerc 2007; 39: 11861192

[4944] 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

[5045] 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

[5146] 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[47] 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

[5548] 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

[5649] 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

[5750] 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. 2023; 8: 5089. doi:10.21105/joss.05089

Table and Figure Captions

Table Captions

Table 1a: Averaging ~~1: Most prevalent specified interpolation methods by~~ type by counts (N~~(a~~) and proportions (%).~~by time (b).~~

Table 1b~~2~~: Averaging ~~methods by type (a) and~~ subtype by counts (N) and proportions (%).~~(b).~~

Figure Captions

Figure 1: ~~Flowchart depicting the four major components of averaging method documentation.~~

~~Figure 2:~~ Selection of sources of evidence flow diagram per the PRISMA 2020 statement [30].~~flowchart. Dashed lines point to articles that were removed. Solid lines indicate the path of articles that remained in the analysis. Details are provided in the main text. "Unobtained" articles were those unavailable due to University library licensing limitations or faulty download links. "Resolvable" files could be analyzed by RegExs and ML.~~

Figure 2~~3~~: Counts and percentages of outlier limits when specified.

Figure 3~~4~~: Most prevalent specified interpolation methods by both type and time.

Figure 4~~5~~: 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.

Table 1a: Averaging type 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 by counts (N) and proportions (%).

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

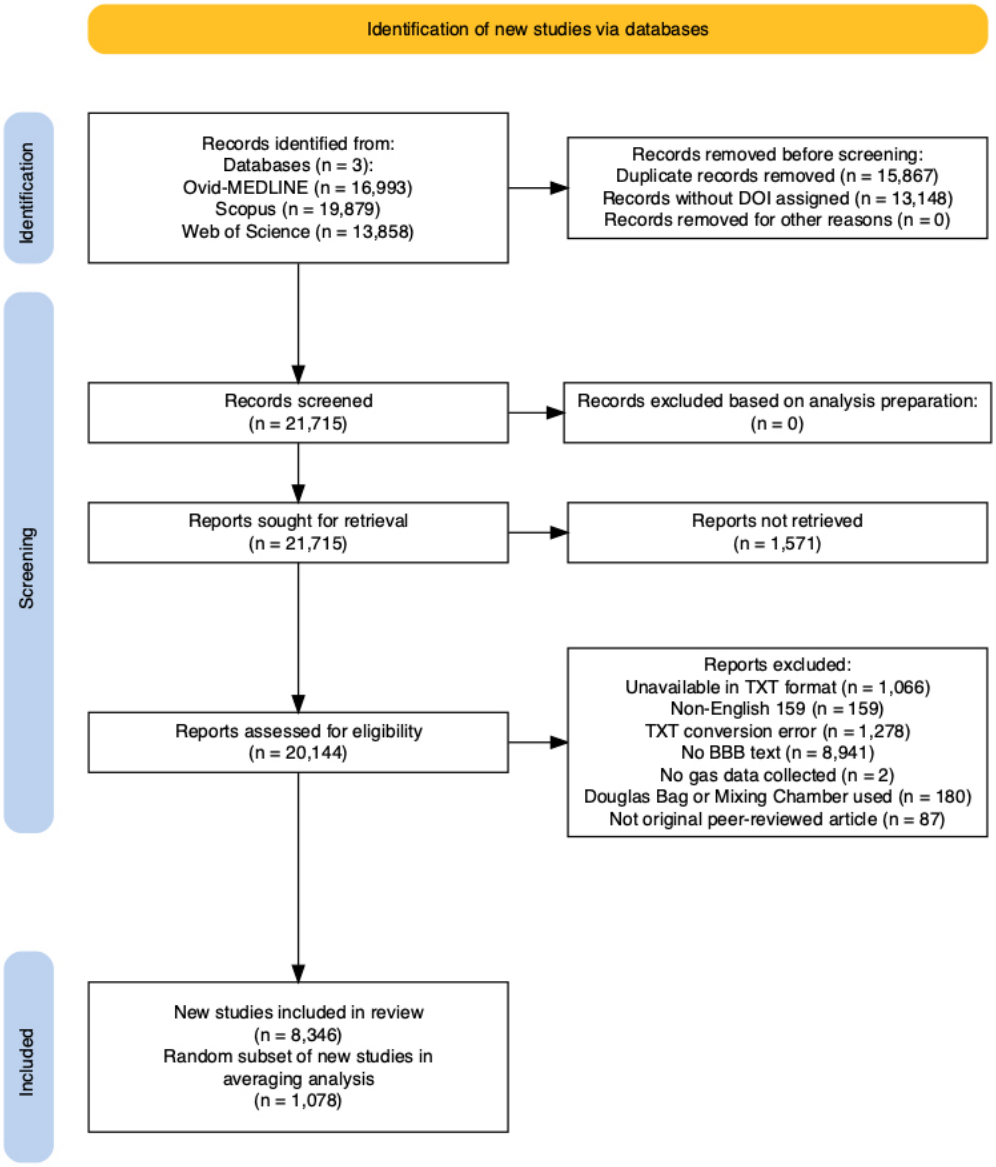

Figure 1: Selection of sources of evidence flow diagram per the PRISMA 2020 statement [30].

264x310mm (72 x 72 DPI)

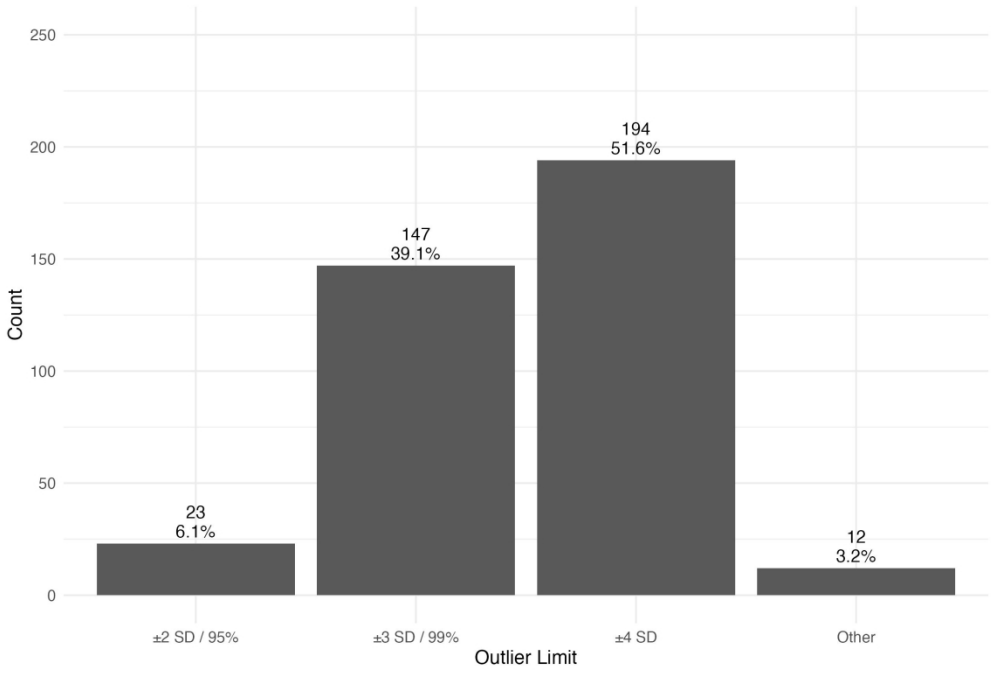

Figure 2: Counts and percentages of outlier limits when specified.

917x623mm (72 x 72 DPI)

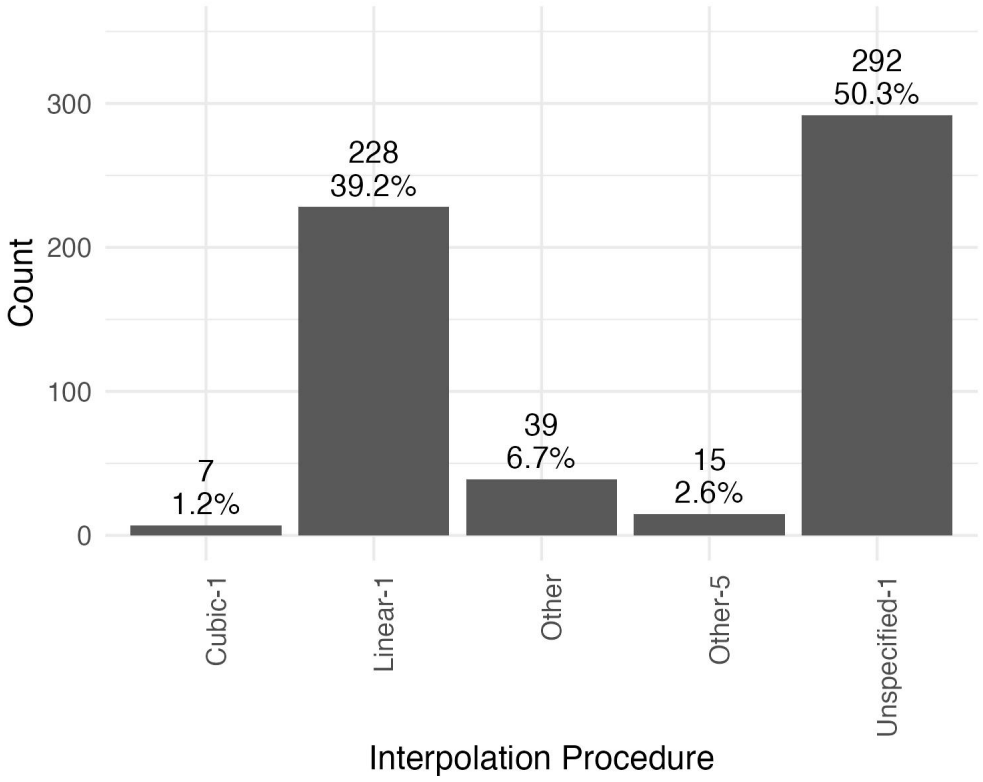

Figure 3: Most prevalent specified interpolation methods by both type and time.

529x422mm (72 x 72 DPI)

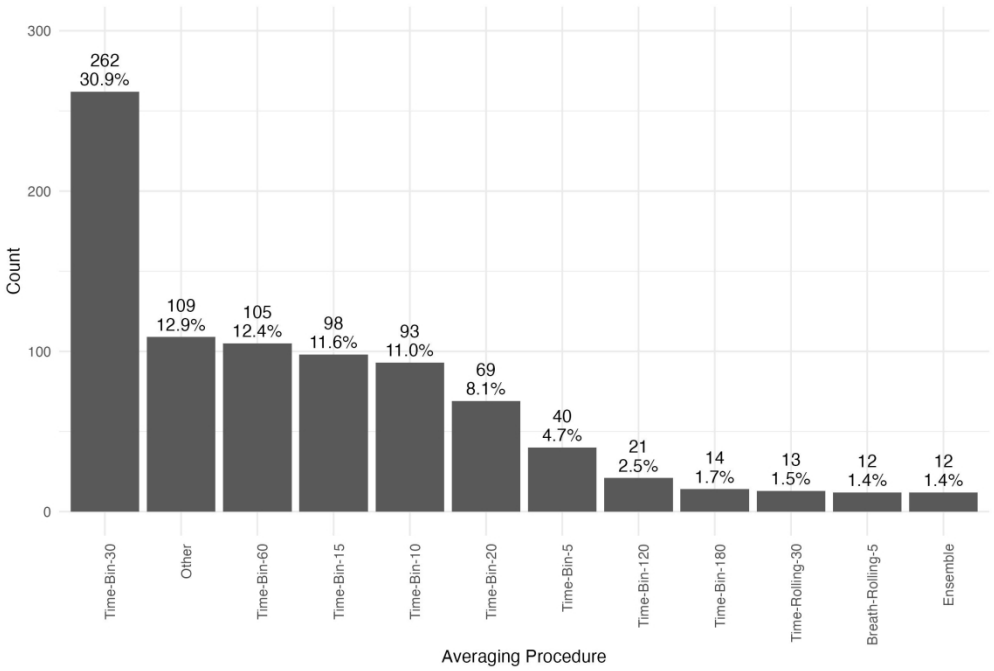

Figure 4: Prevalence of complete averaging procedures. The numbers in each column label are in seconds for time averages and the number of breaths for breath averages. The "other" column represents methods that accounted for less than 1% of the total stated methods.

917x621mm (72 x 72 DPI)

Supplementary material: Popularity and prevalence of gas exchange data processing methods: a semi-automated scoping review

Supplemental Methods

Information Sources and Search

Scopus search:

(TITLE-ABS-KEY (*"oxygen consumption"* OR *"oxygen uptake"* OR *vo2*) AND TITLE-ABS-KEY (*cycli OR *bicycl** OR *run* OR *runn** OR *treadmill** OR *swim** OR *ski* OR *skie** OR *skiing* OR *ergometer** OR *row**) AND TITLE-ABS-KEY (*exer**)) AND (LIMIT-TO (DOCTYPE , *"ar"*)) AND (LIMIT-TO (LANGUAGE , *"English"*)) AND (LIMIT-TO (SRCTYPE , *"j"*)) AND (LIMIT-TO (EXACTKEYWORD , *"Human"*))**

Web of Science Search

((ALL=("oxygen consumption" or "oxygen uptake" or VO2)) AND ALL=(cycli* or bicycl* or run or runn* or treadmill* or swim* or ski or skie* or skii or ergometer* or row*)) AND ALL=(exer*) and **English** (Languages) and **Articles** (Document Types)

Ovid-MEDLINE

1. ("oxygen consumption" or "oxygen uptake" or VO2).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
2. (cycli* or bicycl* or run or runn* or treadmill* or swim* or ski or skie* or skii or ergometer* or row*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
3. 1 and 2
4. 3 and exer*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
5. 4 and "Humans".sa_suba.
6. 5 and "Journal Article".sa_pubt.
7. 6 not ("Review" or "Systematic Review" or "Meta-Analysis").sa_pubt.

Text Analysis & Screening

We principally employed two machine learning (ML) models using a random forest classifier through the Python sklearn package [1]. One model predicted if the article was original peer-reviewed research vs. a review, meta-analysis, case study, protocol registration, etc. The other predicted if the subjects were human or non-human.

We initially built the training data for these models while manually reading articles to construct our regular expressions (RegExs). We then used the models to predict the eligibility of the remaining unread articles. Given that our electronic search primarily obtained articles that matched our search criteria, a relatively small fraction of the unread articles were predicted to be ineligible. We manually verified the eligibility for this small number of articles. Our classifiers also calculated an eligibility probability. We also manually reviewed articles predicted eligible within ~15% of the decision boundary.

We then iteratively rebuilt our classifiers with more data from the manual eligibility verification and from other ineligible articles found incidentally while reading their full text for context when charting our data items. Through rebuilding larger models, we discovered additional ineligible articles missed by previous classifier iterations.

Both models were assessed using repeated stratified 5-fold cross-validation as our training data favored eligible articles. The original peer-reviewed research model contained 1,505 samples while the human model 1,919. The mean accuracy for the original research and human models were 83.9% and 93.4%, respectively.

References

[1] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011; 12: 2825–2830