Original Research

Agreement among Six Methods of Predicting the Anaerobic Lactate Threshold in Elite Cross-Country Skiers

SEAN L. CARTER* and IAN NEWHOUSE‡

School of Kinesiology, Lakehead University, Thunder Bay, ONTARIO, CANADA

*Denotes undergraduate student author, ‡Denotes professional author

ABSTRACT

International Journal of Exercise Science 12(2): 155-172, 2019. The anaerobic lactate threshold (LTan) is used to prescribe training intensity and measure endurance capacity. The LTan identifies a critical point where small increases in workload result in large increases in blood lactate concentration. LTan is usually predicted through visual inspection of a blood lactate (bLa) vs workload plot. Numerous other methods for predicting LTan exist, and the literature lacks a consensus regarding validity of prediction methods. The purpose of this study was to assess the agreement among visual inspection (VI), maximum distance (Dmax) and modified maximum distance (Dmod) from the lactate curve, Baldari & Guidetti (BG), Dickhuth & Heck (DH) and Keul (K) methods for predicting the LTan. Blood lactate data was gathered from 8 male elite cross country skiers across two treadmill running incremental exercise tests. The above methods were used to predict LTan. Bland-Altman limits of agreement and Lin's Concordance Correlation Coefficient analyses were used to compare methods. Agreement was defined as 95% limits of agreement falling within a maximum allowed difference of ± 0.5 mM bLa between methods. No agreement was found among any of the prediction methods. Mean LTan calculated with the Dmax method was significantly different ($p < 0.05$) from mean LTan calculated using each other method. We conclude that the six methods for predicting LTan used in this study are not in agreement and should not be considered equivalent for exercise testing purposes. Future studies should compare agreement between LTan methods and the maximal lactate steady state to determine the most valid LTan prediction method.

KEY WORDS: Correlation, reliability, measurement, agreement, concordance, exercise, physiology, sport, sport science, statistical methods

INTRODUCTION

Elite cross-country skiers are required to train extensively at many different intensities in order to achieve the all-around fitness necessary for high performance. This high training load creates the need for an efficient prescription of training volume and intensity so as to maximize performance gains without producing fatigue. A landmark 1979 study by Kindermann, Simon, and Keul described a means of clinical assessment by which the heart rate could be related to energy metabolism (27). Based on the emerging concept of the lactate threshold, the researchers

recommended a clinical graded exercise test (GXT) where blood lactate (bLa) and heart rate (HR) are measured systematically with increases in workload. The authors argued that plotting the relationship between bLa, HR, and workload would allow coaches to relate real-time HR monitoring to the systemic accumulation of bLa, and thus prescribe training zones according to energy metabolism. Training load could then be optimized according to the desired adaptations of each energy system (27). This basic framework for prescription of training intensity is still common practice in today's sport science laboratory.

This framework was based on the theory of lactate accumulation. Lactate is a primary metabolite of pyruvate, and is therefore linked to energy production by anaerobic glycolysis (7,12). During exercise, increases in blood lactate concentration correspond with increases in exercise intensity (16). At low intensities, where energy production is thought to be accomplished in majority by oxidative phosphorylation (32), bLa concentration is low and unchanging. However, as workload increases, there exists a point on the plot of bLa vs. workload where bLa concentration begins to steadily increase (17). Historically, this point of marked increase was referred to as the "anaerobic threshold" due to the hypothesis that the accumulation of bLa was due to increasing contributions of anaerobic glycolysis to energy production (25, 27). The nomenclature surrounding this particular point has varied considerably over the past 30 years, but current consensus is that the point of increase as described here is found at workloads of around 65% of VO2 max, and at bLa concentrations between 1-2 mmol/L for most subjects (17, 36).

Kindermann et al. also identified a second point of interest on the plot of bLa against workload: the "aerobic-anaerobic" threshold, defined in their original text as a bLa concentration of around 4 mmol/L that corresponds to a workload around 80% of VO2 max in trained individuals (27). Training at this intensity was thought to provide the optimal stimulus for improvements in endurance capacity through both cardiorespiratory and metabolic mechanisms (27). It is easy to see how such a variable would contribute to the precision of training prescription: exact targeting of the threshold during a training bout would theoretically be the most efficient way to generate endurance gains (33, 34, 38).

In the nearly 40 years since the work of Kindermann et al, there has been considerable controversy surrounding the ideas and frameworks behind exercise prescription according to blood lactate accumulation (16, 29, 37). For further insight, contrast the 1985 papers of Brooks and Davis (7, 12). Concepts such as the lactate threshold were operationally defined in many (often contradictory) ways, resulting in a confusing literature that includes partially overlapping measures such as the anaerobic threshold, lactate minimum speed, onset of blood lactate accumulation, individual anaerobic threshold, and others (15, 39).

Despite this confusion, it is useful to distinguish between the two points on the bLa curve identified above. For the purposes of this paper, we adopt the terminology suggested by Faude, Kindermann, & Meyer (15) for referring to the two inflection points on the bLa curve generated by a GXT. We will refer to the first increase in bLa concentration above baseline levels as the aerobic lactate threshold (LTaer). We will refer to the second point on the bLa curve, where the

rate of lactate accumulation in the blood greatly exceeds the rate of elimination, as the anaerobic lactate threshold (LTan). This study will focus primarily on the methods by which the LTan can be determined from GXT data, though it is important to have distinct working definitions for each lactate threshold.

The maximal lactate steady state (MLSS) is conceptually similar to the LTan, but is operationally defined as the highest blood lactate concentration and work load that can be maintained over time without a continual blood lactate accumulation (4). It is distinct from the LTaer or the LTan in that it is determined from several fixed workload exercise tests of at least 30 minutes in duration, performed across several days (4, 15, 22). The MLSS correlates very strongly with competition performance in running and cycling events (18, 24). However, MLSS testing is costly and requires significant time (22).

Given the relative difficulty of MLSS testing, and the conceptual overlap between the MLSS and the LTan as defined above, the LTan as determined through a single-session short duration GXT is frequently used as an approximate measure of the MLSS. However, Bentley et al. and others have highlighted the impact of variations in test protocol, GXT stage length, and overall test duration on the absolute workload associated with ventilatory and lactate thresholds (3). It therefore seems likely that the LTan as measured through a GXT is an imprecise approximation of the MLSS.

The obvious move would be to assess the level of agreement between the MLSS and the LTan. If good agreement is found between the LTan determined through a GXT and the MLSS determined through the more laborious fixed workload exercise testing, then the LTan could be considered a valid approximation of the highest bLa concentration and workload which can be maintained without increases in bLa accumulation. Stated differently, measures of the LTan and MLSS could be considered interchangeable (6). However, comparing the agreement between the LTan and the MLSS requires a better understanding of LTan prediction methods than we currently have.

There is presently no accepted gold standard for determining the LTan from GXT data. Many methods exist for quickly and easily predicting the LTan from laboratory test data, but the relative precision of each method has been inadequately researched to date, and there is no consensus on the best LTan prediction method for use in the sport laboratory (15). The discussion is further hampered by a lack of consensus concerning regression analysis of the bLa curve: various authors have argued for regression analysis using linear splines (35), a 3rd degree polynomial regression (5, 11), a continuous exponential function (23), and a log-log transformation of the data (31).

The precision of the prediction method becomes relevant if we accept the hypothesis that targeting the HR and workload that correspond to the LTan during a training bout is the best way to generate gains in endurance performance (33, 34). Take the example of exercise prescription for endurance cycling. If a particular method generates predicted values for bLa

and workload at LTan that overestimate the true LTan by 0.2 mM and 20 Watts respectively, then a training program centered around targeting the LTan will require the cyclist to train at a workload that is 20 Watts over his true threshold throughout the duration of his training bout. Since workloads above the LTan can only be sustained for a matter of minutes (27), a cyclist with this prescription will fatigue faster during the training bout, spend less time at the target workload, and potentially be exposed to undesirably high training loads when compounded over a period of training. The reverse is true if the LTan prediction method tends to underestimate the true LTan.

Any exercise prescription, training program, or method of monitoring training load that is based on a measure of the LTan will be confounded by an LTan prediction method that demonstrates anything resembling the above bias. We would consider such an LTan prediction method to be invalid regardless of how well it correlates with sport competition performance or the MLSS, as this is a practically relevant degree of imprecision. Thus the precision of the LTan prediction method is of importance to the clinical physiologist, sport scientist, coach, and endurance athlete.

The purpose of this study was to suggest a methodology for assessing the precision and validity of LTan concepts through analysis of agreement, and to address the question of whether LTan prediction methods can be considered interchangeable for the purposes of exercise testing. If we found agreement between two LTan prediction methods, then it would not matter which method of the two was chosen when analyzing GXT data. Conversely, if two methods were not in agreement, then further validation studies that compared each method to the MLSS would be necessary. We assumed that the true LTan could be found within the range of LTan values generated by six commonly used prediction methods. We also assumed that agreement between prediction methods (defined here as 95% limits of agreement of no more than ± 0.5 mM bLa between methods), would mean convergence around the true LTan value. We hoped that we would be able to identify LTan prediction methods that were more likely to agree with the MLSS in subsequent studies.

The present study examines six LTan prediction methods (detailed in the Method). The maximum distance from lactate curve (Dmax) method appears frequently in exercise testing literature, but is influenced by starting GXT workload such that the LTan value given by this method depends significantly on the GXT protocol used (10). The modified Dmax method (Dmod) fixes this problem and correlates well with the MLSS, but predicts an LTan that is on average 23 Watts lower than MLSS (41). The subjectivity of the visual inspection (VI) method is evident, and there are considerable problems with inter-rater reliability (43), though it is the most frequently used LTan prediction method (35). The Keul Individual Anaerobic Threshold (K) correlates well with the MLSS in runners, cyclists, and rowers (2, 20, 21). However, all three studies reported significant differences in power output or speed (0.2 m/s, 21 Watts, and 31 Watts higher respectively) between the K LTan and the MLSS. The original study detailing the Baldari & Guidetti Individual Anaerobic Threshold (BG) method returned bLa values at the LTan that matched subsequent MLSS testing (1), though no further research has been done on

the method. The Dickhuth & Heck Individual Anaerobic Threshold method (14) has not been compared to the MLSS or correlated with exercise performance.

We expected there to be agreement between the Dmod, VI and K LTan prediction methods based on their shared curve regression technique, and a lack of agreement between these methods and the other three methods used. We hypothesized that the bias inherent in the Dmax method (10) and the absence of validity or reliability studies for the DH method would result in practically relevant differences in agreement (95% limits of agreement of greater than ± 0.5 mM bLa) between these two methods and the other four methods used.

Bland-Altman limits of agreement analysis quantifies the scatter about a line of perfect agreement between two measures, and allows an estimate of mean bias and expected variation between two measures for 95% of cases (6). We primarily used analysis of agreement among prediction methods because traditional measures of validation (e.g. tests of significance or the interclass correlation coefficient) do not address questions of variability or precision between two measures of the same quantity (6, 30). It is likely that there will be a strong linear correlation between data generated by two LTan prediction methods, given that both methods assess the same quantity. Nevertheless, the agreement between those two methods could be low despite strong correlation. That is to say, two methods for assessing the LTan could return significant variability in predicted LTan values despite high correlation and low mean difference between those methods (6). We also used a repeated measures ANOVA to compare mean LTan values for each method, in order to highlight the inadequacy of tests of significance in terms of answering the question of agreement (6, 30).

To our knowledge, agreement analysis among LTan prediction methods has only been done in three studies to date. Davis, Roznek, DeCicco, Carizzi, and Pham assessed agreement among three methods of predicting the LTaer in 8 male and 6 female cyclists using Bland-Altman analysis and the intraclass correlation coefficient (13). Correlational analysis revealed strong correlations between methods (r values between r=0.97 and r=0.99). Based on this, the authors concluded that there was agreement between the methods. However, Bland-Altman analysis revealed limits of agreement of ± 12 Watts between log-log and VI methods, and ± 20 Watts between log-log and a definition-based method. Thus predicted LTaer workload for 95% of cases differed as much as 20 Watts between methods, which we interpret as a practically relevant lack of agreement.

Cerda-Kohler et al. assessed agreement between VI, Dmax, Dmod, and log-log methods of lactate threshold prediction in professional soccer players (9). The authors compared treadmill speed and HR at predicted LTan, and defined agreement as no more than ± 1 km/h and ± 10 bpm between methods respectively. No agreement was found among methods for speed at LTan, but there was limited agreement between VI, Dmax, and Dmod methods for HR at LTan.

Hauser, Adam, and Schulz assessed agreement between a fixed 4 mmol/L bLa concentration prediction method, the DH and VI prediction methods, and the MLSS in 57 male cyclists (19).

The authors compared power output at predicted LTan with power output at MLSS, and determined a general lack of agreement between LTan prediction methods and the MLSS.

METHODS

Participants
This study used secondary data collected by the Lakehead University Exercise Physiology Lab in conjunction with the Thunder Bay National Development Centre during their spring and summer testing sessions. Participants were a convenience cohort of 8 male elite cross country skiers sampled over two testing sessions separated by four months. Training during the four-month period between testing sessions consisted of running 2-3 times per week, strength training 2-3 times per week, and rollerskiing 2-4 times per week. Both running and rollerskiing sessions consisted of workloads designed to improve aerobic capacity and anaerobic capacity.

Because testing occurred before and after a four-month period of intensive training, it was assumed that maximal aerobic speed, running speed at LTan, time to exhaustion and peak blood lactate during a treadmill GXT would be sufficiently altered for each participant (8, 42) such that data from both testing sessions could be pooled. Thus data from both testing sessions (16 bLa data sets) were subject to analysis in order to increase statistical power.

Participants were required to have competition experience at the National level or above, have previous experience with treadmill testing and bLa sampling, and incorporate running into their training at least 2 times weekly. Participants were excluded if injuries were present that did not allow them to complete the testing protocol, or if they exhibited a subjectively high HR due to nervousness during testing. Ethical approval for this study was received from the Lakehead University School of Kinesiology Ethics Review Committee.

Sample size was determined based on calculations for confidence intervals at 95% limits of agreement as outlined in Bland & Altman (6). A sample of 16 data sets gives a confidence interval for 95% limits of agreement of approximately 0.85 times the standard deviation of the difference between two methods (6). This is a large confidence interval for limits of agreement analysis, but the lower range still falls outside our definition of agreement between methods (below), thus allowing the question of agreement to be assessed.

Protocol
Blood lactate data was collected during a modified Bruce protocol treadmill graded exercise test. Athletes performed a 10-minute warmup at a heart rate less than 120 bpm, followed by 5 minutes of rest. A bLa sample was taken at the end of the rest period to ensure starting values below 2mM. Athletes began running on a treadmill at 1% grade. Previous testing data for each athlete was used to estimate a starting speed that would allow the athlete to complete roughly 8 stages (e.g. 10 km/h start if athlete can complete full stage at 18 km/h). Each stage lasted 3 minutes, allowing for a roughly 24-minute test protocol. BLa was measured at the end of every stage and HR was recorded at the end of every minute. During bLa measurement, subjects were

stationary for approximately 30 seconds before the next stage began. Speed was increased by 1 km/h at the end of every stage. The test was stopped when bLa measures reached 6.0 mM or above. Blood lactate measurement was performed with a Lactate Pro portable blood analyzer, which was found to be an accurate and reliable field test for blood lactate concentration (40).

Prediction methods were performed as follows:
The VI method subjectively determines the location of LTaer and LTan on a plot of bLa versus workload by visually identifying the two inflection points on the lactate curve (35).

The Dmax method applies a 3rd degree polynomial regression to the lactate curve. The LTan is the point on the regression curve which is maximally distant from the line which connects the two endpoints of the curve (11).

The Dmod method determines the LTan in a similar manner to Dmax, but the line is drawn between the LTaer and the endpoint of the curve. The LTaer is defined for the purposes of this method as the point that corresponds to the workload before the first rise in bLa (5, 15).

The BG method defines the LTan as the data point that corresponds to the second increase in bLa of greater than 0.5 mmol/L (1).

The DH method defines the LTan as the data point that corresponds to a 1.5 mmol/L increase in bLa from the LTaer (14, 15). The LTaer was determined for the purposes of DH as described previously for Dmod.

The LTan under the K method is given by the tangent to the blood lactate curve at 51° of inclination. For the purposes of this study, the K LTan was calculated by determining the requisite slope of the 3rd degree polynomial regression curve, and subsequently determining the bLa value (in increments of 0.1 mmol/L) at the point of corresponding slope (15, 26).

Statistical Analysis
Bland-Altman limits of agreement (LoA) analysis was used to quantify mean bias and 95% LoA (expressed as mean ± 1.96 *SD*) between two prediction methods. The maximum allowed difference between prediction methods was set at a mean difference of no more than ± 0.5 mM bLa, highlighting what we judged to be a clinically relevant difference in bLa concentrations when guiding exercise prescription. Agreement between methods was defined as upper and lower 95% LoA falling within the maximum allowed difference.

This analysis technique relies on the assumption that the distribution of differences between two prediction methods is approximately normal (6). The Shapiro-Wilk test revealed non-normal distributions of differences between prediction methods for all comparisons. We therefore normalized our data using a log transformation (6). On a logarithmic scale, the maximum allowed difference between methods is ± 0.30 units from the line of zero mean difference. LoA

between methods must therefore fall within this maximum allowed difference in order for methods to be considered in agreement.

Lin's Concordance Correlation Coefficient (CCC) was used to verify the agreement between LTan prediction methods, with 95% confidence intervals. This method calculates the degree to which two sets of continuous data fall along a line of y=x when plotted against one another (28), which is analogous to Bland & Altman's concept of the "line of agreement" (6). Agreement was defined as CCC > 0.90.

Additionally, a repeated measures analysis of variance (ANOVA) was used to compare means. The Shapiro-Wilk test and Mauchly's test were used to confirm assumptions of normality and sphericity respectively. A two-tailed Bonferroni post-hoc analysis was performed (α = 0.05). Microsoft Excel software was used to process the bLa data. Bland-Altman, CCC, and ANOVA analyses were all performed with IBM SPSS version 23 software.

RESULTS

No agreement (i.e., 95% LoA within ± 0.30 units from zero mean difference on the logarithmic scale) was found between LTan prediction methods. Method 1 versus method 2 (M_1 - M_2) comparisons showed that M_2 could be expected to differ greatly from M_1 across all prediction methods (Table 1). LoA were found outside our defined maximum allowable difference between methods (Figures 1-3).

Table 1. Mean difference and 95% Limits of Agreement among VI, Dmod, Dmax, BG, DH, and K anaerobic lactate threshold prediction methods.

M_1 - M_2	Mean Difference (log)	Upper LoA	% above M_1	Lower LoA	% below M_1
Dmod - VI	-0.38	0.32	108	-1.07	92
Dmax - VI	0.04	0.34	41	-0.25	23
BG - VI	-0.36	0.33	39	-1.05	65
DH - VI	-0.29	0.26	30	-0.84	57
K - VI	-0.28	0.27	31	-0.82	56
Dmax - Dmod	-0.10	0.30	35	-0.51	40
BG - Dmod	-0.41	0.33	38	-1.14	32
DH - Dmod	-0.48	0.17	18	-1.13	68
K - Dmod	-0.46	0.19	21	-1.10	67
Dmax - BG	-0.04	0.62	87	-0.71	51
Dh - BG	-0.32	0.57	78	-1.22	71
K - BG	-0.30	0.46	59	-1.06	65
Dmax - DH	-0.21	0.34	41	-0.77	53
K - DH	-0.48	0.32	37	-1.28	72
Dmax - K	-0.21	0.58	79	-1.00	63

Mean difference, upper LoA and lower LoA are expressed on a logarithmic scale. Upper and lower LoA are also expressed for M_2 values as percentages above and below M_1 values. For example, the VI method might return values as much as 108% above and 92% below the Dmod method for 95% of cases.
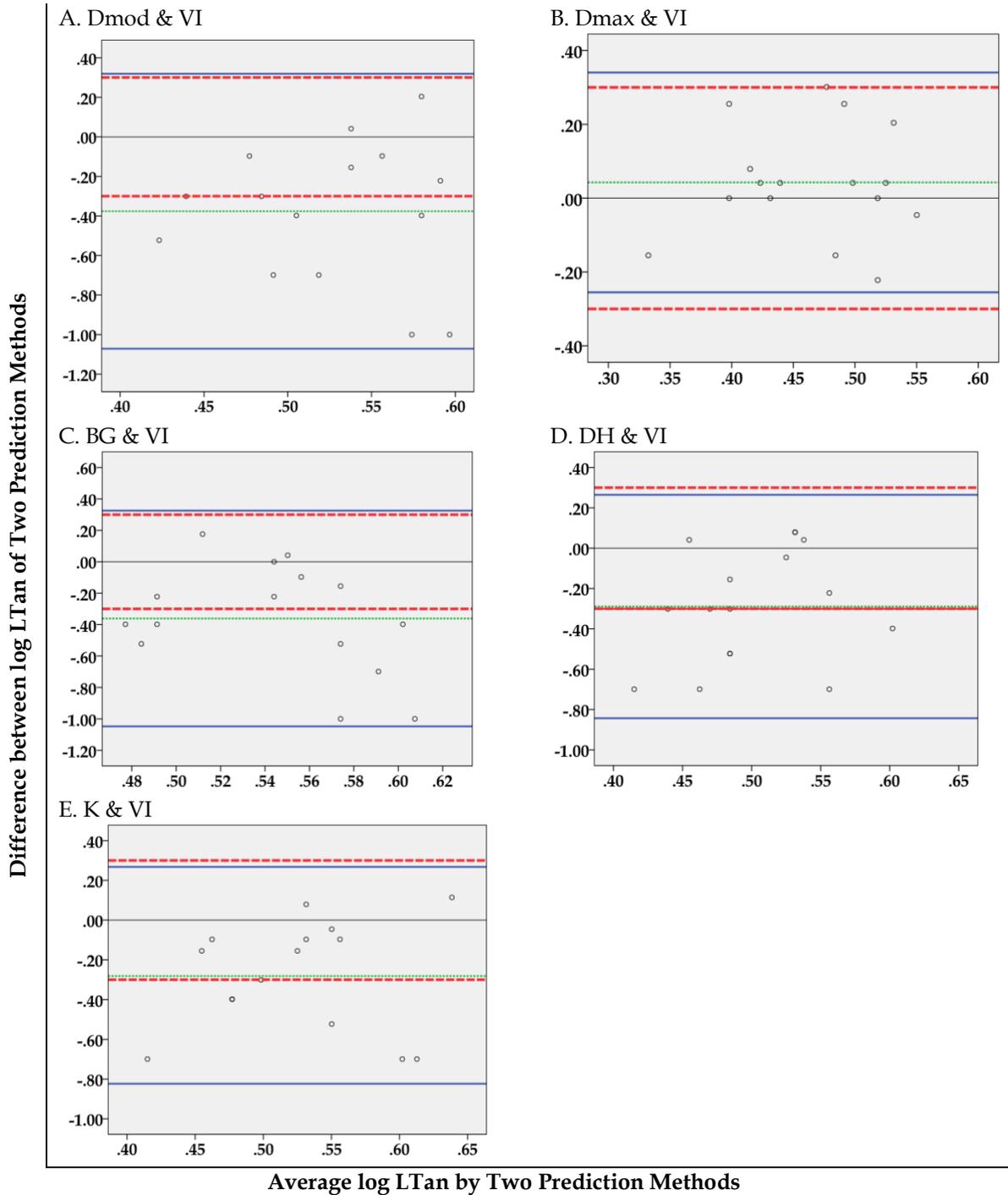

Figure 1. Agreement between A) Dmod & VI, B) Dmax & VI, C) BG & VI, D) DH & VI, and E) K & VI methods of predicting the anaerobic lactate threshold. Limits of agreement (LoA) for 95% of cases fell outside the defined maximum allowed difference across all prediction methods. All graphs are presented on a logarithmic scale to correct for an increase in the scatter of difference between methods with an increase in average between methods.
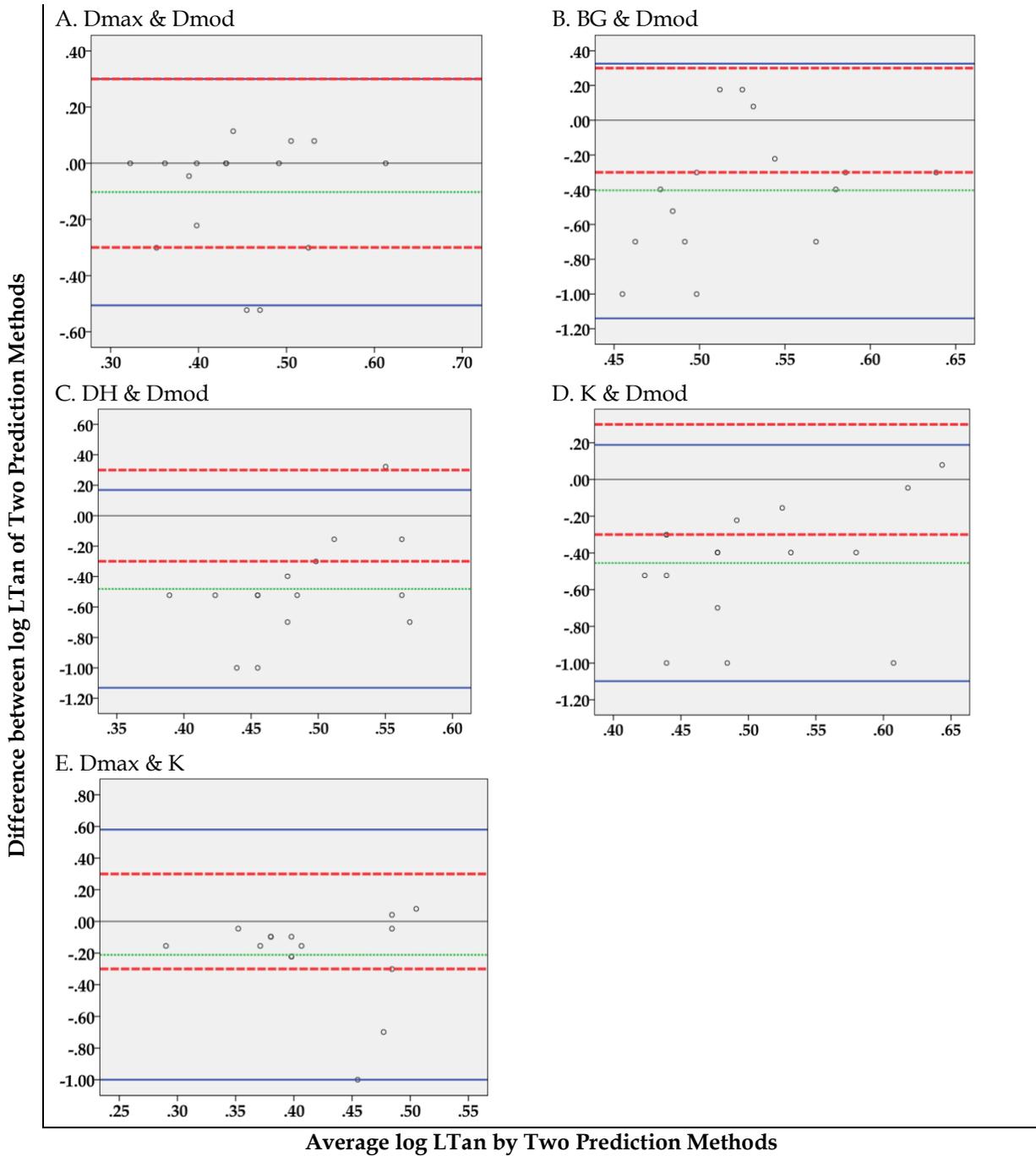
Figure 2. Agreement between A) Dmax & Dmod, B) BG & Dmod, C) DH & Dmod, D) K & Dmod, and E) Dmax & K methods of predicting the anaerobic lactate threshold. Limits of agreement (LoA) for 95% of cases fell outside the defined maximum allowed difference across all prediction methods. All graphs are presented on a logarithmic scale to correct for an increase in the scatter of difference between methods with an increase in average between methods.
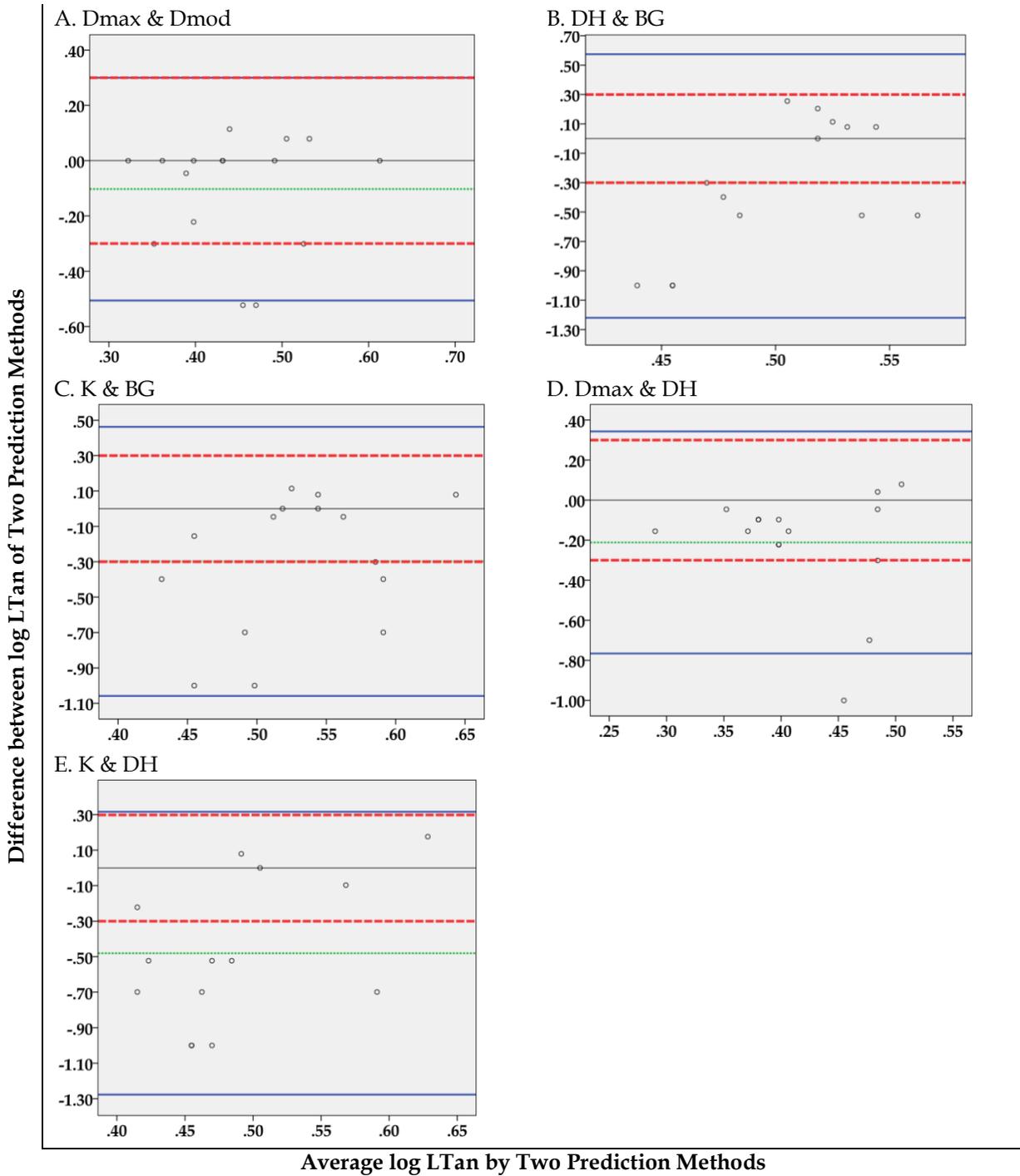
Figure 3. Agreement between A) Dmax & BG, B) DH & BG, C) K & BG, D) Dmax & DH, and E) K & DH methods of predicting the anaerobic lactate threshold. Limits of agreement (LoA) for 95% of cases fell outside the defined maximum allowed difference across all prediction methods. All graphs are presented on a logarithmic scale to correct for an increase in the scatter of difference between methods with an increase in average between methods.

Analysis using Lin's Concordance Correlation Coefficient confirmed a lack of agreement between any of the LTan prediction methods (CCC << 0.9).

Table 2. Lin's Concordance Correlation Coefficient reveals no agreement among VI, Dmod, Dmax, BG, DH, and K methods for predicting the anaerobic lactate threshold.

	CCC	95% Confidence Intervals	
		lower	upper
Dmod & VI	0.22	-0.24	0.60
Dmax & VI	0.08	-0.07	0.23
BG & VI	0.02	-0.46	0.50
DH & VI	0.26	0.01	0.48
K & VI	0.33	-0.10	0.66
Dmax & Dmod	0.35	0.14	0.53
BG & Dmod	0.20	-0.25	0.58
DH & Dmod	0.26	-0.10	0.56
K & Dmod	0.62	0.21	0.84
Dmax & BG	0.04	-0.11	0.19
DH & BG	0.00	-0.24	0.24
K & BG	0.23	-0.21	0.59
Dmax & DH	0.14	-0.16	0.42
K & DH	0.44	0.13	0.68
Dmax & K	0.14	-0.16	0.57

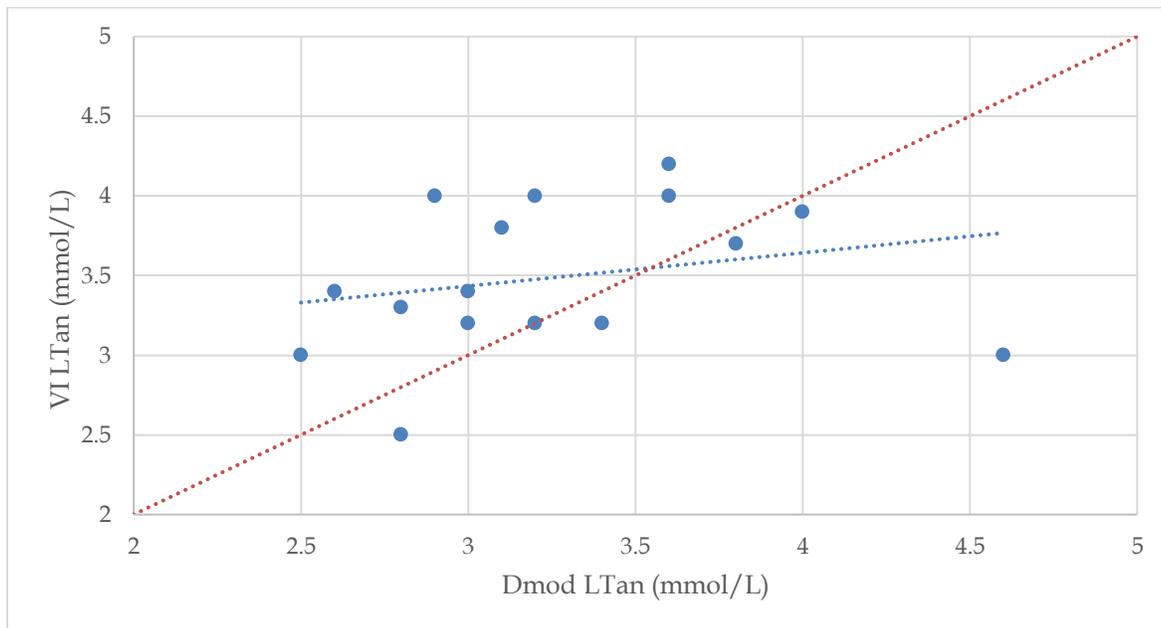A concordance coefficient of <0.90 is considered poor concordance.

Figure 4. Visualization of Bland & Altman's "line of agreement" (orange line) using mock data. This is analogous to a line of y=x when two LTan prediction methods are plotted against one another. Both Bland-Altman analysis and Lin's Concordance Correlation Coefficient are means of quantifying the degree of scatter (blue points) about the line of agreement between two measures of the same quantity. The two methods plotted here are linearly correlated (blue line) despite demonstrating poor agreement.

A repeated measures ANOVA (Figure 4) showed that there was a significant difference in mean predicted LTan across methods, $F(1, 5) = 16.43$, $p < 0.01$. Post-hoc comparison using the Bonferroni correction indicated a significant difference between Dmax ($M = 2.39$, $SD = 0.52$) and VI ($M = 3.49$, $SD = 0.47$, $p < 0.05$), Dmod ($M = 3.26$, $SD = 0.55$, $p < 0.05$), BG ($M = 3.52$, $SD = 0.50$, $p < 0.05$), DH ($M = 2.89$, $SD = 0.37$, $p < 0.05$), and K ($M = 3.24$, $SD = 0.69$, $p < 0.05$) methods respectively.
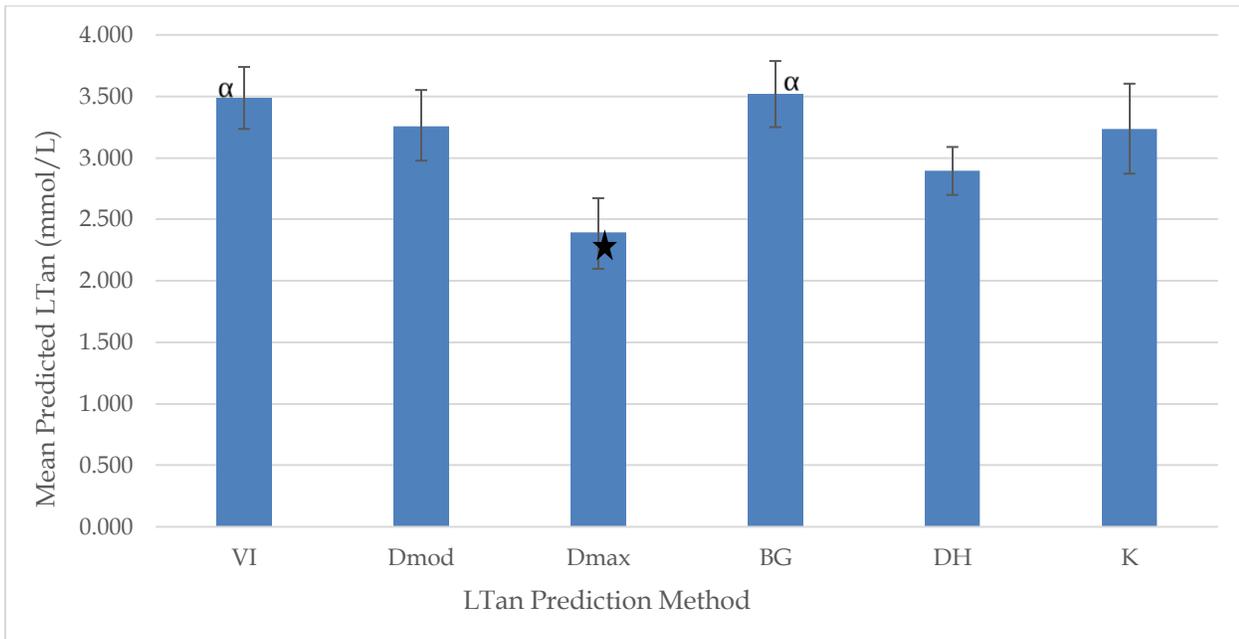

Figure 5. Mean predicted LTan across prediction methods. *****Denotes significant difference ($p < 0.05$) from each other prediction method. ᵃDenotes significant difference ($p < 0.01$) from DH.

DISCUSSION

The purpose of this study was to suggest Bland-Altman limits of agreement as methodology for examining the precision and validity of LTan prediction methods. We demonstrated this methodology by examining the agreement among six methods of predicting the LTan.

The main finding was a lack of agreement between all methods, defined as 95% limits of agreement that exceeded the maximum allowed difference of ± 0.5 mM bLa, or ± 0.30 units on our logarithmic scale (Figures 1-3, Tables 1 & 2). We expected to see agreement between the VI, Dmod, and K methods because all three apply a 3rd degree polynomial regression to the bLa curve generated from GXT data. This was not the case (Tables 1 & 2). For example, when compared to the Dmod method, the VI method showed both a mean difference (-0.38 units) that exceeded the maximum allowed difference of ± 0.30 units, and limits of agreement for 95% of cases that indicated the VI method could overestimate the Dmod method by as much as 108% or underestimate the same by as much as 92% (Figure 1A, Table 1). Stated differently, the mean difference between VI and Dmod methods exceeded the maximum allowed difference between methods for an entire normally distributed range of variation in predicted LTan values. This

represents significant lack of agreement between the two methods, and similar results were found among the other four methods as well (Tables 1 & 2).

Interpreting the results of Bland-Altman analysis is not intuitive because of the logarithmic scale. However, Lin's Concordance Coefficient (CCC) (28) provides a quick and easy way of quantifying agreement (Table 2), and confirms the results of Bland-Altman analysis. The CCC provides an estimate of the degree to which values generated by two measures of the same quantity fall along a line of y=x when plotted against each other (Figure 4). A complete lack of agreement (CCC << 0.90) among all methods highlights significant variation in predicted LTan values. Stated differently, none of the LTan prediction methods converged on a line of y=x when plotted against one another (Table 2), and thus cannot be relied upon to give equivalent results when used to analyze the same data set.

Earlier, we assumed that the true LTan could be found within the range of LTan values approximated by the six prediction methods, and that agreement between prediction methods would mean convergence around a "true" LTan value. A complete lack of agreement between all six methods suggests that this assumption may not be valid, and the true LTan value may not be approximated by any of the prediction methods used in this study. Future work which compares LTan prediction methods with the MLSS is necessary to answer this question.

We hypothesized correctly that the Dmax method would not be in agreement with any of the other methods (Table 1). Other authors have highlighted the influence of the GXT protocol on the value of the predicted LTan by the Dmax method (3, 10). We demonstrated that the Dmax method does not agree with other LTan prediction methods used to analyze the same bLa data (Table 1), and that the mean predicted LTan by Dmax is significantly lower (p<0.05) than the mean LTan predicted by each other method (Figure 5). According to our analysis, the Dmax method (M = 2.39, SD = 0.52) returns a mean predicted LTan that is more than 0.5 mM lower than all other methods: the next closest is the DH method (M = 2.89, SD = 0.37), and mean LTan by the BG method (M = 3.52, SD = 0.50) is more than 1 mM higher than mean LTan by Dmax (Figure 5), on a scale that is only practically relevant from 1 to 6 mM bLa.

We suggest that taken together, this provides sufficient evidence to discount the Dmax method from use in the sport laboratory, as it produces predicted LTan values that are not even in the ballpark relative to other methods, regardless of whether it is total agreement or mean predicted LTan that is considered. We hoped that we would be able to identify LTan prediction methods that should be subject to further comparisons with the MLSS. Based on the above, we do not consider it likely that the Dmax method will demonstrate good agreement with the MLSS in future studies.

This study also proposed to examine whether any LTan prediction methods could be used interchangeably in the sport laboratory. It can be conclusively stated that the six LTan prediction methods examined here are not equivalent and should not be considered interchangeable for the purposes of clinical exercise testing or training prescription. Practically speaking, the results

of this study suggest that the value of an athlete's predicted LTan could depend as much on the prediction method used to analyze the GXT data as on the physiological characteristics of the athlete in question. This finding (here described in elite cross-country skiers) was consistent with the results of Davis et al. (13) in cyclists and Cerda-Kohler et al. (9) in professional soccer players.

Mean comparison using a repeated measures ANOVA highlights the discrepancy between analysis of agreement and tests of significance when assessing the precision and validity of LTan prediction methods. The mean LTan by the Dmax method was significantly lower ($p < 0.05$) than mean LTan by all other methods. However, comparison using *only* tests of significance would have lead one to conclude that there was no relevant difference in predicted LTan among VI, Dmod, BG, and K methods. Bland-Altman analysis and use of the CCC showed significant, practically relevant differences in agreement (greater than ± 0.5 mM bLa) between all of the LTan prediction methods (Figures 1-3, Tables 1& 2). This is why we suggest using Bland-Altman analysis for any future work comparing LTan prediction methods to each other or to the MLSS.

A limitation of this study was the assumption made regarding the source data. Data was sourced from 8 male participants who completed two GXT tests separated by four months of training. We assumed that participant data from the second GXT could be compared alongside data from the first GXT, as four months of training would result in relevant differences in several variables that affect the bLa curve (8, 42). We made this assumption so as to increase statistical power. However, this assumption led to non-normal distributions of mean difference between prediction methods. Nevertheless, we believe that our analysis remains valid as we corrected for this assumption using a log transformation, which allowed for Bland-Altman analysis of agreement. In addition, Lin's Concordance Correlation Coefficient does not rely on assumptions of normality, and confirms a lack of agreement. However, we were unable to demonstrate the pervasive problem of inappropriate use of the intra-class correlation coefficient (ICC) (Figure 4) that is present in the literature on validity of LTan prediction methods, given that the ICC relies on the assumption of normality.

A second limitation was the use of treadmill running tests to assess the LTan in elite cross-country skiers, given that only roughly 30-40% of the participants' training was running. The physiological principle of specificity indicates that it would be best to test elite cross country skiers using a rollerski treadmill. However, we do not believe that this would have resulted in relevant differences in the outcome of our study, as we were interested in differences produced through data analysis rather than between study participants. Nevertheless, future studies should consider testing cross-country skiers with sport-specific protocols.

A useful direction of further inquiry would be to challenge the underlying assumptions of the VI, Dmod, BG, DH, and K methods in the light of the changing consensus of the paradigm and underlying physiology of lactate accumulation (16). In order to assess the validity of the VI, Dmod, BG, DH, and K methods, we recommend that future research assess the LTan using these methods, and examine the agreement between these values and the bLa value identified by MLSS testing. We recommend that agreement between predicted LTan and MLSS be assessed

using Bland-Altman analysis or Lin's Concordance Correlation Coefficient, and suggest that only LTan prediction methods that demonstrate good agreement with the MLSS be considered valid for the purposes of clinical exercise testing or prescription of training workload.

REFERENCES

1. Baldari C, Guidetti LA. simple method for individual anaerobic threshold as predictor of max lactate steady state. Med Sci Sports Ex 32(10): 1798-802, 2000.

2. Beneke R. Anaerobic threshold, individual anaerobic threshold, and maximal lactate steady state in rowing. Med Sci Sports Ex 27(6): 863-867, 1995.

3. Bentley DJ, Newell J, Bishop D. Incremental exercise test design and analysis: implications for performance diagnostics in endurance athletes. Sports Med 30(7): 575-586, 2007.

4. Billat VL, Sirvent P, Py G, Koralsztein JP, Mercier J. The concept of maximal lactate steady state: a bridge between biochemistry, physiology, and sport science. Sports Med 33(6): 407-426, 2003.

5. Bishop D, Jenkins DG, Mackinnon LT. The relationship between plasma lactate parameters, Wpeak and 1-h cycling performance in women. Med Sci Sports Ex 30(8): 1270-5, 1998.

6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327(8476): 307-310, 1986.

7. Brooks GA. Anaerobic threshold: review of the concept and directions for future research. Med Sci Sports Ex 17(1): 22-34, 1985.

8. Carter H, Jones AM, Doust JH. Effect of 6 weeks of endurance training on the lactate minimum speed. J Sports Sci 17(12): 957-967, 1999.

9. Cerda-Kohler H, Burgos-Jara C, Ramirez-Campillo R, Valdes-Cerda M, Baez E, Izquierdo M. Analysis of agreement between 4 lactate threshold measurement methods in professional soccer players. J Strength Cond Res 30(10): 2864-70, 2016.

10. Chalmers S, Esterman A, Eston R, Norton K. Standardization of the Dmax method for calculating the second lactate threshold. Int J Sports Physiol 10(7): 921-926, 2015.

11. Cheng B, Kuipers H, Snyder AC, Keizer HA, Jeukendrup A, Hesselink M. A new approach for the determination of ventilatory and lactate thresholds. Int J Sports Med 13(7): 518-22, 1992.

12. Davis JA. Anaerobic threshold: review of the concept and directions for future research. Med Sci Sports Ex 17(1): 6-21, 1985.

13. Davis JA, Rozenek R, DeCicco DM, Carizzi MT, Pham PH. Comparison of three methods for detection of the lactate threshold. Clin Physiol Func Imaging 27(6): 381-384, 2007.

14. Dickhuth H, Huonker M, Munzel T, Drexler H, Berg A, Keul J. Individual anaerobic threshold for evaluation of competitive athletes and patients with left ventricular dysfunction. Adv Ergometry 173-179, 1991.

15. Faude O, Kindermann W, Meyer T. Lactate threshold concepts: how valid are they? Sports Med 39(6): 469-490, 2009.

16. Gladden LB. Lactate metabolism: a new paradigm for the third millennium. J Physiol (Lond) 558(1): 5-30, 2004.

17. Golnick PD, Bayley WM, Hodgson DR. Exercise intensity, training, diet, and lactate concentration in muscle and blood. Med Sci Sports Ex 18(3): 334-40, 1986.

18. Harnish CR, Swensen TC, Pate RR. Methods for estimating the maximal lactate steady state in trained cyclists. Med Sci Sports Ex 33(6): 1052-1055, 2001.

19. Hauser T, Adam J, Schulz H. Comparison of calculated and experimental power in maximal lactate-steady state during cycling. Theor Bio Med Model 27: 11-25, 2014.

20. Heck H. Laktat in der leistungsdiagnostik. Schorndorf: Hofmann; 1991.

21. Heck H, Hess G, Mader A. Comparative study of different lactate threshold concepts. Dtsch Z Sportmed 36(1-2): 19-25, 40-52, 1985.

22. Hering GO, Hennig EM, Riehle HJ, Stepan J. A lactate kinetics method for assessing the maximal lactate steady state workload. Front Physiol 9: 310, 2018.

23. Hughson RL, Weisiger KH, Swanson GD. Blood lactate concentration increases as a continuous function in progressive exercise. J Appl Physiol 62(5): 1975-81, 1985.

24. Jones AM, Doust JH. The validity of the lactate minimum test for determination of the maximal lactate steady state. Med Sci Sports Ex 30(8): 1304-1313, 1998.

25. Jones NL, Ehrsam RE. The anaerobic threshold. Exerc Sports Sci Rev 10: 49-83, 1982.

26. Keul J, Simon G, Berg A, Dickhuth HH, Goerttler I, Kübel R. Bestimmung der individuellen anaeroben schwelle zur leistungsbewertung und trainingsgestaltung. Dtsch Z Sportmed 30: 212-218, 1979.

27. Kindermann W, Simon G, Keul J. The significance of the aerobic-anaerobic transition for the determination of work load intensities during endurance training. Eur J Appl Physiol 42(1): 25-34, 1979.

28. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1): 255, 1989.

29. Lindinger MI, Heigenhauser GJF, McKelvie RS, Jones NL. Blood ion regulation during repeated maximal exercise and recovery in humans. Am J Physiol 262(1): 126–136, 1992.

30. Liu J, Tang W, Chen G, Lu Y, Feng C, Tu XM. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch Psychiatry 28(2): 115-120, 2016.

31. Lundberg MA, Hughson RL, Weisiger KH, Jones RH, Swanson GD. Computerized estimation of lactate threshold. Comput Biomed Res 19(5): 481-486, 1986.

32. Mader A, Heck H. A theory of the metabolic origin of "anaerobic threshold." Int J Sports Med 7(Suppl 1): 45-65, 1986.

33. McLellan TM, Skinner JS. The use of the aerobic threshold as a basis for training. Can J Appl Sport Sci 6(4): 197-201, 1981.

34. Meyer T, Lucia A, Earnest CP, Kindermann W. A conceptual framework for performance diagnosis and training prescription from submaximal gas exchange parameters - theory and application. Int J Sports Med 26(Suppl 1): 38-48, 2005.

35. Newell J, Higgins D, Madden N, Cruickshank J, Einbeck J, McMillan K, McDonald R. Software for calculating blood lactate endurance markers. J Sports Sci 25(12): 1403-1409, 2007.

36. Powers SK, Howley ET. Exercise Physiology: Theory and Application to Fitness and Performance. New York, NY: McGraw-Hill Education; 2015.

37. Richardson RS, Noyszewski EA, Leigh JS, Wagner PD. Lactate efflux from exercising human skeletal muscle: role of intracellular PO2. J Appl Physiol 85: 627–634, 1998.

38. Seiler KS, Kjerland GO. Quantifying training intensity distribution in elite endurance athletes: is there evidence for an "optimal" distribution? Scand J Med Sci Sports 16(1): 49-56, 2006.

39. Svedahl K, MacIntosh BR. Anaerobic threshold: the concept and methods of measurement. Can J Appl Physiol 28(2): 299-323, 2003.

40. Tanner RK, Fuller KL, Ross ML. Evaluation of three portable blood lactate analyzers: Lactate Pro, Lactate Scout and Lactate Plus. Eur J Appl Physiol 109(3): 551-9, 2010.

41. Van Schuylenbergh R, Vanden Eyende B, Hespel P. Correlations between lactate and ventilatory thresholds and the maximal lactate steady state in elite cyclists. Int J Sports Med 25(6): 403-408, 2004.

42. Vorup J, Tybirk J, Gunnarsson TP, Ravnhold T, Dalsgaard S, Bangsbo J. Effect of speed endurance and strength training on performance, running economy and muscular adaptations in endurance-trained runners. Eur J Appl Physiol 116(7): 1331-1341, 2016.

43. Yeh MP, Gardner RM, Adams TD, Yanowitz FG, Crapo RO. Anaerobic threshold: problems of determination and validation. J Appl Physiol 55(4): 1178-86, 1983.