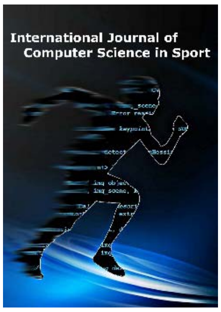
International Journal of Computer Science in Sport

DE GRUYTER OPEN

Volume 16, Issue 2 (Special Edition), 2017

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.1515/ijcss-2017-0009



Predicting Elite Triathlon Performance: A Comparison of Multiple Regressions and Artificial Neural Networks

Hoffmann, M.[1], *Moeller, T.*[2], *Seidel, I.*[1,3], *Stein, T.*[1]

[1] *BioMotion Center, Institute of Sports and Sports Science, Karlsruhe Institute of Technology*

[2] *Institute for Applied Training Science*

[3] *Olympic Training Centre Lower Saxony*

Abstract

Two different computational approaches were used to predict Olympic distance triathlon race time of German male elite triathletes. Anthropometric measurements and two treadmill running tests to collect physiological variables were repeatedly conducted on eleven male elite triathletes between 2008 and 2012. After race time normalization, exploratory factor analysis (EFA), as a mathematical preselection method, followed by multiple linear regression (MLR) and dominance paired comparison (DPC), as a preselection method considering professional expertise, followed by nonlinear artificial neural network (ANN) were conducted to predict overall race time. Both computational approaches yielded two prediction models. MLR provided $R^2 = 0.41$ in case of anthropometric variables (predictive: pelvis width and shoulder width) and $R^2 = 0.67$ in case of physiological variables (predictive: maximum respiratory rate, running pace at 3-mmol·L^{-1} blood lactate and maximum blood lactate). ANNs using the five most important variables after DPC yielded $R^2 = 0.43$ in case of anthropometric variables and $R^2 = 0.86$ in case of physiological variables. The advantage of ANNs over MLRs was the possibility to take non-linear relationships into account. Overall, race time of male elite triathletes could be well predicted without interfering with individual training programs and season calendars.

KEYWORDS: PROFESSIONAL TRIATHLETES, PERFORMANCE PREDICTION, RACETIME NORMALIZATION, FACTOR ANALYSIS, DOMINANCE PAIRED COMPARISON

Introduction

Performance prediction in training-intensive sports like triathlon (a combination of swimming, cycling, and running) could be beneficial for optimizing training protocols and identifying talent. It is therefore important to identify performance parameters predicting triathlon race performance (Landers, Blanksby, Ackland, & Smith, 2000), such as anthropometric and physiological parameters based on laboratory tests (Schabort, Killian, St Clair Gibson, Hawley, & Noakes, 2000).

Several studies have shown the importance of maximum oxygen uptake (VO_2max) and anaerobic thresholds (Millet, Vleck, & Bentley, 2009; Millet, Vleck, & Bentley, 2011) in endurance running or running in triathlon. These parameters showed significant correlations to race performance (Bassett, 2000; McLaughlin, Howley, Bassett, Thompson, & Fitzhugh, 2010). Similar results were found for swimming and cycling (Millet et al., 2009; Sleivert & Rowlands, 1996). However, these variables only have a prerequisite function, and are not performance predictors in homogenous samples, because of the small variation between athletes (Bassett, 2000; Sleivert & Rowlands, 1996; Stratton et al., 2009). Nonetheless, blood lactate concentrations from treadmill or cycle ergometer tests were useful parameters in predicting triathlon performance independent of athletes' performance level (Schabort et al., 2000; Van Schuylenbergh, Eynde, & Hespel, 2004). Besides such physiological factors, anthropometric variables such as percent body fat, body mass index (BMI), and the circumferences of several parts of the body could also be important for performance in triathlon races (Knechtle, Wirth, Rüst, & Rosemann, 2011) and therefore for performance prediction.

Unlike in the individual sports of swimming, cycling, and running, which constitute triathlon, the relationship between one or a combination of anthropometric and physiological parameters and overall race time with regard to performance prediction have rarely been investigated in triathlon. Schabort et al. (2000) used multiple linear regressions to predict overall Olympic distance triathlon race time of the South African national team by using physiological parameters such as peak treadmill running speed [$km \cdot h^{-1}$] and blood lactate value at 4 $W \cdot kg^{-1}$ body mass on a cycle ergometer. The correlation between predicted and actual race time was highly significant ($r = 0.90$, $p < 0.001$). Multiple regression analysis ($R^2 = 0.98$; SEE = 0.95 [min]) was also used by Van Schuylenbergh et al. (2004) to predict sprint distance triathlon performance of male physical education students. In each of these two studies, subjects competed in the same triathlon competition, which likely caused the high explanation of variance (R^2) because of comparable conditions. Nonetheless, this kind of experimental design is rarely possible with elite triathletes due to their individual season calendar. Artificial neural networks (ANNs) are an alternative computational approach for performance predicition. Edelmann-Nusser, Hohmann, and Henneberg (2002) as well as Silva et al. (2007) showed that artificial neural networks could be a valuable method for performance modelling, without the restrictions of distribution and independence of variables. Edelmann-Nusser et al. (2002) predicted the 200 m backstroke time of an elite female swimmer in the finals of the Olympic games by using artificial neural networks (multi-layer perceptrons) based on collected training data. The accuracy of the results of this approach were attributed to the fact that "the adaptive behavior of the system athlete is quite a complex, non-linear problem" (Edelmann-Nusser et al., 2002). However, multiple linear regression analyses have been more widely used to develop prediction models. Linear regressions require linear relationships between independent variables and a dependent variable, whereas artificial neural networks could also handle non-linear relationships based on a different model architecture. Nevertheless, ANNs have rarely been used to predict race-performance, possibly because the network design of an ANN requires a lot of specifications concerning the number of neurons, layers, training algorithm etc. (Zhang, Eddy Patuwo, & Y. Hu, 1998).

Both computational approaches have a major limitation while working with measurement data from elite athletes: large numbers of independent variables require many sets of data, which are rarely available while working with elite athletes. Therefore, a preselection of parameters is necessary to reduce the number of independent variables. If there are only a few variables with non-linear relationships, a purely statistical approach like an exploratory factor analysis can be used to preselect variables before computing a prediction model. In case of ANNs, which could also handle non-linear relationships, a dominance paired comparison based on the expertise of professional triathlon coaches could be beneficial, since this method utilizes a more subjective point of view and practical experiences.

In summary, the prediction of individual overall race time in elite Olympic distance triathlon competition, by using several anthropometric and physiological parameters as well as different computational approaches, has not been investigated to date. Previous studies mostly tested recreational triathletes (Kohrt, Morgan, Bates, & Skinner, 1987; Millet et al., 2011; Miura, Kitagawa, & Ishiko, 1997; Sleivert & Wenger, 1993) because of the availability of a larger number of potential athletes. National squads normally consist of 4–5 athletes, which makes it very difficult to get a sufficient sized sample. Moreover, elite athletes are often reluctant to participate in experiments. In addition, individual training programs and different season calendars complicate experimental laboratory studies with elite athletes. Therefore, the first aim of this study was to assess whether overall Olympic distance triathlon race time of elite athletes could be predicted using regular performance diagnostics. The second aim was to compare two computational approaches and determine whether one is better than the other. A purely statistical approach consisting of an exploratory factor analysis to preselect variables in combination with a multiple linear regression to predict overall race time was compared to an expertise-based non-linear approach consisting of a dominance paired comparison as a preselection method in combination with an artificial neural network to predict overall race time. In both cases several anthropometric and physiological variables measured during laboratory tests over a period of four years in German male elite Olympic distance triathletes were used.

Methods

Subjects

Eleven male German elite triathletes (age: 23.38 ± 2.79 years) competing in national or international championships were included in this study. Written informed consent in the form of an athlete agreement between each national squad triathlete and the German national triathlon association (DTU), as well as a cooperation agreement with the Institute for Applied Training Science (Leipzig, Germany), which is responsible for classic performance diagnostics of elite athletes in Germany, were mandatory. Participation in the performance diagnostics was voluntary and the triathletes could opt out at any time. After data acquisition, all statistical analyses were conducted anonymously. Table 1 shows descriptive characteristics (mean value and standard deviation (SD) as well as the coefficient of variation (CV = (SD/Mean)*100)) of the triathletes.

Table 1. Descriptive variables of German elite triathletes (N = 11).

	mean ± SD	**CV (%)**
Anthropometric		
age [yrs]	23.38 ± 2.79	11.93
body height [cm]	187.0 ± 2.90	1.55
body weight [kg]	74.46 ± 4.28	5.75
seat height [cm]	96.38 ± 1.59	1.65
shoulder width [cm]	40.16 ± 2.24	5.58
pelvis width [cm]	28.65 ± 1.61	5.62
thorax width [cm]	28.27 ± 1.29	4.56
thorax depth [cm]	21.06 ± 1.41	6.70
Quetelet Index [g·cm^{-1}]	398.15 ± 21.32	5.35
BMI [kg·m^{-2}]	21.29 ± 1.17	5.50
body fat [%]	10.70 ± 1.36	12.71
body fat [kg]	8.00 ± 1.35	16.88
lean body mass [kg]	66.46 ± 3.27	4.92
Physiological		
VO$_2$max [mL·min^{-1}]	5457.67 ± 292.56	5.36
VO$_2$max [mL·min^{-1}·kg^{-1}]	72.02 ± 4.29	5.96
PL3 [m·s^{-1}]	5.08 ± 0.23	4.53
max running pace [m·s^{-1}]	5.22 ± 0.27	5.17
max running pace mobi [m·s^{-1}]	6.92 ± 0.17	2.46
LA max mobi [mmol·L^{-1}]	9.18 ± 1.30	14.16
VCO$_2$ max mobi [mL]	6472.75 ± 431.74	6.67
max distance mobi [m]	1762.69 ± 136.70	7.76
RMV max mobi [mL·min^{-1}]	187.73 ± 12.40	6.61
RR max mobi [breaths·min^{-1}]	63.18 ± 10.10	15.99
BLC 3 min [mmol·L^{-1}]	8.08 ± 1.31	16.21
BLC 6 min [mmol·L^{-1}]	9.13 ± 1.29	14.13
BLC 10 min [mmol·L^{-1}]	8.62 ± 1.38	16.01
normalized overall race time Olympic distance [min]	113.79 ± 3.21	2.82

Notes: PL3 = running pace at 3-mmol·L^{-1} blood lactate; mobi = mobilization test; LA max mobi = maximum blood lactate in mobilization test; RMV max Mo = maximum respiratory minute volume in mobilization test; RR max Mo = maximum respiratory rate in mobilization test; BLC 3, 6, 10 min = blood lactate concentration 3, 6, 10 min after load in mobilization test, respectively

Experimental Procedure

The data in this study were derived from laboratory tests performed between 2008 and 2012 at the Institute for Applied Training Science (Leipzig, Germany) within the frame of national squad investigations. Because elite triathletes were tested at various time slots based on their competition calendar, the distribution of tests was not consistent. Overall, 23 men completed 58 laboratory tests between 2008 and 2012. The iterative approach to select valid sets of variables was based on the following requirements: (1) complete sets of variables of the laboratory tests and (2) finished Olympic distance triathlon races within 8 weeks after each single performance diagnostic. Twenty-five sets of variables from eleven triathletes fulfilled these criteria and were eventually used.

The anthropometric characteristics of each triathlete were selected and determined based on the information provided by Tittel and Wutscherk (1972) and Knussmann and Barlett (1988). Body height and segment lengths and widths were measured using precise measuring instruments and valid measurement regulations, and provided the basis to calculate the various indices (Tittel & Wutscherk, 1972). Body fat was determined by measuring skin fold thickness of ten skin folds with a caliper (Tittel & Wutscherk, 1972); lean body mass could then be calculated. The anthropometric variables mentioned in Table 1 (except age and body weight, which are only listed for a better characterization of the sample) were used for computation.

For the physiological parameters, the triathletes had to perform two different motorized treadmill running tests under laboratory conditions (gradient of 0°). First, a classic step test with an individual initial speed between 4 and 4.5 $m \cdot s^{-1}$ depending on general performance was conducted. The step length was 4 km, with an increasing rate of 0.25 $m \cdot s^{-1}$ between two consecutive steps. The test was stopped after a maximum of four steps. One day later, a maximum mobilization test with an initial speed of 5 $m \cdot s^{-1}$, an increasing rate of 0.25 $m \cdot s^{-1}$ per step, and a step length of 30 s until voluntary exhaustion was performed. In both tests, blood lactate measurements and spirometry were conducted. Pulmonary and respiratory gas-exchange parameters were measured using a calibrated breath-by-breath gas-analyzer (Cortex METAMAX 3B and Cortex METALYZER 3B). The physiological variables considered for computation are shown in Table 1 (except relative VO_2max, which is only listed for a better characterization of the sample).

Data analysis

Normalization of race time

Normalization was necessary to obtain comparable individual race times independent of the various triathlon races in which the subjects participated. These normalized race times were fundamental to all following analyses, since they accounted for the slightly different competition calendars of each elite triathlete. Races with a maximum time lag of 8 weeks to each single performance diagnostic were selected from official results (www.triathlondata.org). To guarantee a similar race progress, the minimum requirement was participation in the German, European, or World championships as well as in races within the World Triathlon Series (WTS).

The reference factor was calculated as the mean value of overall race times of the Top 10 athletes in WTS between 2009 and 2012. All finished races within each year were considered. The resulting mean value for Olympic distance triathlon race time was used to normalize each individual race time.

reference factor = mean (overall race times of Top 10 WTS athletes of all races within the WTS 2009, 2010, 2011, 2012)

Up to two races of the WTS are sprint distance triathlons. To use these race times, the same approach was applied to determine a factor transforming sprint distance triathlon race time into an Olympic distance equivalent.

Statistical methods

The statistical analyses applied after race time normalization could be divided into two computational approaches to identify performance-relevant parameters and predict overall race times of German male elite triathletes:

- A purely statistical approach consisting of an exploratory factor analysis, to preselect important anthropometric and physiological parameters, and multiple linear regressions to identify performance-relevant parameters and predict overall race times of German male elite triathletes.
- An expertise-based non-linear approach consisting of a dominance paired comparison with four professional German triathlon coaches, to preselect important anthropometric and physiological parameters, and the application of artificial neural networks to predict overall race times of German male elite triathletes.

The converse implementation of the preselection and prediction methods (e.g. dominance paired comparison and multiple linear regressions) were deemed unsuitable because of their different fields of application, based on the linear and non-linear relationships between the independent variables and the dependent variable, normalized overall race time.

The purely statistical approach could be divided into two consecutive steps: An exploratory factor analysis (EFA) was first applied to preselect relevant independent variables, followed by a multiple linear regression (MLR) to determine potential prediction models and the priority of the used parameters. An EFA helps to uncover structures in large sets of variables. This allows a preselection of parameters with high correlations among themselves and similar explanation of variance to the same underlying factor. For small sample sizes, which are inevitable while working with elite triathletes, a reduction of variables can improve the results in MLR, and prevent multicollinearity. Therefore, an EFA was conducted using the ‘principal component’ method. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy of 0.726 (based on anthropometric variables) and 0.697 (based on physiological variables) show a “middling” suitability (Kaiser & Rice, 1974) for both EFAs. A varimax rotation led to the final solution, with variables sorted by the size of factor loadings related to a general factor. With this step, variables such as relative VO_2max (less descriptive than absolute VO_2max) and the maximum blood lactate concentration in classic step test (less descriptive than maximum blood lactate concentration in mobilization test or blood lactate concentrations after load) could be excluded with a minimal loss of information. Based on these results, a stepwise multiple linear regression analysis (backward method, default exclusion criteria: probability of F to remove ≥ 0.1) was used to detect the relationships between independent variables and overall race time in Olympic distance triathlon. Each parameter had to be significant ($p < 0.05$). To avoid multicollinearity, Variance Inflation Factor (VIF) was checked with a cut-off of 10 (Hair, 1995). Additionally, the normality of residuals was examined via normal distribution plots, and residual independence and homoscedasticity were determined by plotting the residuals against the estimated data. Furthermore, Cook’s Distance was used with a cut-off ≥ 1 to identify and remove influential cases in case of homoscedasticity (Heiberger & Holland, 2004). The coefficient of determination (percentage of variance explained; R^2) and the standard error of the estimate (SEE) were used to evaluate the models. The adjusted R^2, in particular, allows a comparison between several MLR models, considering the number of variables used in each case.

The expertise-based non-linear approach also consisted of two conscecutive steps: A dominance paired comparison was first conducted to identify performance-relevant parameters, based on the expertise of four professional German triathlon coaches, followed by the computation of artificial neural networks (ANNs) to determine potential prediction models. A dominance paired comparison helped raters to prioritize influencing variables in a systematic and objective way. Thus, personal preferences and subjective influences could be avoided with regard to prioritization of the independent variables. Each national triathlon coach had to rate the significance of each variable against all others. The overall sum score was used for the final prioritization. To ensure solvability of the numerous connections in the artificial neural networks with regard to the sample size, the five most relevant variables were finally selected. Two dominance paired comparisons were conducted (for anthropometric and physiological variables separately). The selected relevant parameters were used to compute two-layer feedforward artificial neural networks as a non-linear approach to predict overall race time in Olympic distance triathlon of elite triathletes. In general, ANNs have the ability to learn relationships between variables in complex, non-linear contexts. A multi-layer perceptron with one input layer (one input neuron for each independent variable), one hidden layer (two neurons), and one output layer (one neuron for the dependent variable, normalized overall race time), as shown in Figure 1, was selected as a universal approach (Hornik, Stinchcombe, & White, 1989). To minimize mean squared error, Levenberg-Marquardt algorithm was used as a training algorithm due to its attribute of robustness (Marquardt, 1963). The dataset was randomly divided into datasets for training (80% of the sample), validation (10% of the sample), and testing (10% of the sample). In the training process, a set of input-output patterns was used to adjust the weights of all interconnections between the neurons in an ANN. The validation set is mainly used to avoid over fitting in the learning process. The test data is finally used to predict an output, which should be within an acceptable margin compared to the actually given output. The presented results below involving the entire sample. The coefficient of determination (R^2) and the standard error of the estimate (SEE) were used to evaluate the models. The SEE was calculated to ensure comparability between both computational approaches, even though it is not common in ANNs.

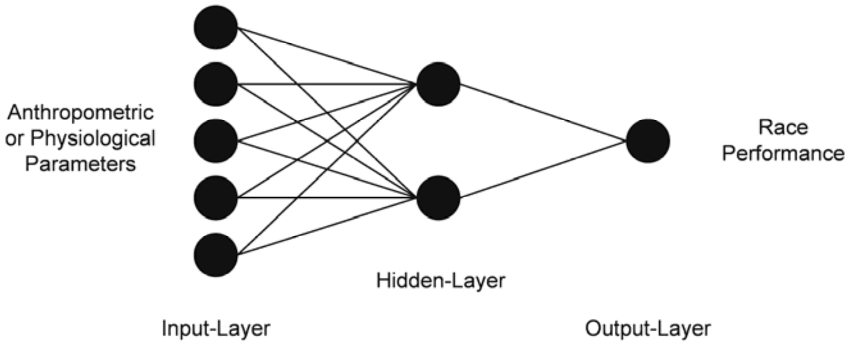


Figure 1. Internal characteristics of an Artificial Neural Network consisting of five Input-Neurons, two Hidden-Neurons, and one Output-Neuron.

SPSS Statistics (Version 22, IBM) and MATLAB (Version R2015b, MathWorks) with Neural Network Toolbox were used for statistical analyses. The level of significance was set to $p < 0.05$.

Results

Normalization of race time

The normalization of race times yielded a mean ± standard deviation of overall race time in Olympic distance triathlon of 6827.57 ± 192.56 [s] (approximately 1:54 h) for male elite triathletes. The conversion factor for sprint distance race times into an Olympic distance equivalent is 2.08 ± 0.03.

Preselection of variables

Exploratory factor analysis

EFA yielded four factors in case of anthropometric variables and three factors in case of physiological variables. Tables 2 and 3 show the variables sorted by the size of factor loadings related to the general factor, and after varimax rotation. A suppression level of 0.5 was used to point out decisive variables (Hair, 1995) and to exclude variables with poorer explanation to one general factor (e.g. relative VO_2max).

Most of the variables showed a strong relationship to one single factor. Lean body mass was related to body composition and height. Running pace at 3-mmol·L^{-1} blood lactate and maximum running pace in classic step test were both related to respiration and velocity as well as respiration and velocity in the mobilization test. The variables in Table 2 and Table 3 were therefore used to compute the following multiple linear regression analyses.

Table 2. Varimax-rotated factor loadings of exploratory factor analysis for anthropometric variables.

	Factor 1	Factor 2	Factor 3	Factor 4
BMI	.908			
Quetelet Index	.863			
lean body mass	.777		.551	
thorax depth	.613			
body fat %		.968		
body fat kg		.839		
pelvis width		.583		
body height			.930	
seat height			.802	
shoulder width				.888
thorax width				.852
possible factor interpretation	„body composition“	„body fat“	„height“	„segment width“

Table 3. Varimax-rotated factor loadings of exploratory factor analysis for physiological variables.

	Factor 1	Factor 2	Factor 3
LA max mobi	.966		
BLC 6 min	.959		
BLC 10 min	.921		
BLC 3 min	.824		
VCO_2 max mobi		.870	
VO_2 max		.834	
PL3		.708	.591
max running pace		.682	.523
RR max mobi			.788
max distance mobi			.689
max running pace mobi			.684
RMV max mobi			.655
possible factor interpretation	„lactate“	„respiration and velocity“	„respiration and velocity mobi“

Dominance paired comparison

The dominance paired comparisons as a second preselection approach yielded five parameters concerning anthropometric and physiological parameters, shown in Table 4. Anthropometric parameters mostly described the body composition of the athletes. The selection of physiological parameters consisted of respiratory, lactate, and velocity-related variables.

Table 4. Results of dominance paired comparisons with national triathlon coaches for anthropometric and physiological variables.

Five most important parameters	
anthropometric	**physiological**
body weight [kg]	absolute VO_2max [$mL \cdot min^{-1}$]
BMI [$kg \cdot m^{-2}$]	relative VO_2max [$mL \cdot min^{-1} \cdot kg^{-1}$]
body fat [%]	running pace at 3-$mmol \cdot L^{-1}$ blood lactate [$m \cdot s^{-1}$]
body fat [kg]	maximum running pace [$m \cdot s^{-1}$]
lean body mass [kg]	maximum running pace in mobilization test [$m \cdot s^{-1}$]

Performance prediction models

Multiple linear regression

Statistical assumptions of multiple linear regression (normal distribution of regression residuals, homoscedasticity) were validated by assessment and testing of residuals. Multiple linear regression analysis after EFA revealed that, among anthropometric parameters, pelvis width and shoulder width were the best predictors of overall race time in Olympic distance triathlon. The R² showed an explanation of variance of 40.5% of overall race time by the anthropometric based model. The multiple linear regression model after EFA based on physiological parameters included running pace at 3-mmol·L^{-1} blood lactate, maximum lactate, and maximum respiratory rate in the mobilization test. The physiological model showed a higher R² of 66.5, with a lower SEE in comparison (Table 5). The results led to two equations predicting overall Olympic distance triathlon race time for male elite triathletes:

Predicted race time [s] based on anthropometric variables = 7643.56 - 80.889 × (pelvis width [cm]) + 37.388 × (shoulder width [cm])

Predicted race time [s] based on physiological variables = 8521.03 + 8.556 × (maximum respiratory rate [breaths·min^{-1}]) - 332.80 × (running pace at 3-mmol·L^{-1} blood lactate [m·s^{-1}]) - 61.658 × (maximum blood lactate [mmol·L^{-1}])

Table 5. Parameter and model estimates of multiple linear regression analyses for male elite triathletes.

	value	β-coefficient	R²	Adjusted R²	SEE [s]	p-value	VIF
EFA + MLR (anthropometric)			0.405	0.351	155.14	0.003	
Constant	7643.56						
SW	37.39	0.434				0.025	1.199
PW	-80.89	-0.674				0.001	1.199
EFA + MLR (physiological)			0.665	0.582	117.27	0.003	
Constant	8521.03						
PL3	-332.80	-0.474				0.018	1.065
LA max Mo	-61.66	-0.450				0.028	1.161
RR max Mo	8.56	0.505				0.014	1.103

Notes: SEE = standard error of the estimate, VIF = Variance Inflation Factor, SW = shoulder width; PW = pelvis width; PL3 = running pace at 3-mmol·L^{-1} blood lactate; LA max Mo = maximum blood lactate in mobilization test; RR max Mo = maximum respiratory rate in mobilization test; general format for multiple regression equation: y = constant + value1 × variable1 + value2 × variable2 + …

Artificial neural networks

The artificial neural network computed after dominance paired comparison using the anthropometric variables body weight, BMI, lean body mass, and absolute and relative body fat explained 43.4% of the variance in overall race time (R² = 0.43; SEE = 144.56 [s]). The artificial neural network after dominance paired comparison using the physiological variables maximum running pace, running pace at 3-mmol·L^{-1} blood lactate, absolute and relative VO$_2$max, and maximum running pace in mobilization test explained 86.2% of the variance in overall race time ($R^2 = 0.86$; SEE = 91.82 [s]). Both artificial neural networks, with their specific characteristics, could be used to predict overall Olympic distance triathlon race time based on a single input pattern.

Discussion

The aim of the current study was to assess whether overall Olympic distance triathlon race time of male elite athletes could be predicted using regular performance diagnostics, and to compare two different computational approaches. Anthropometric and physiological variables measured during routine laboratory tests provided a database for the prediction, without interfering with individual training programs and season calendars of the elite triathletes. Both the combinations assessed (an exploratory factor analysis and multiple linear regression, and a dominance paired comparison and artificial neural network), yielded prediction models of overall triathlon race time.

Assessment of parameters

Table 1 shows the homogeneous appearance of elite triathletes within the sample. Anthropometric characteristics had only small variations, except for body fat [% and kg], which became obvious because of a larger CV. The physiological variables showed a partially similar distribution: VO$_2$max [mL·min^{-1} and mL·min^{-1}·kg^{-1}] showed a small variation because of its premising function in samples consisting of elite triathletes. Maximum lactate value and maximum respiratory rate in mobilization tests as well as the lactate values after load showed higher CVs. Different individual strengths in the three disciplines likely affected the results of running-specific step tests.

The selection of parameters has an important effect on the prediction results. Body height, body weight, and resulting BMI, as well as age were in accordance to the reports of Hue (2003), Schabort et al. (2000) and Hue, Le Gallais, Boussana, Chollet, and Prefaut (2000) (slightly lower body height and weight) as well as Ackland, Blanksby, Landers, and Smith (1998) (slightly older, smaller, and lighter). VO$_2$max [mL·min^{-1}·kg^{-1}] as gross criterion of endurance performance was slightly lower than that reported by Hue, Le Gallais, Chollet, and Préfaut (2000) and Schabort et al. (2000), and similar to that reported by Hue (2003). Lactate values could not be compared because of various specifications such as defined bounds, running paces, or power outputs while cycling. McLaughlin et al. (2010) showed a considerably slower running pace at 3-mmol·L^{-1} blood lactate (4.41 m·s^{-1}), which is likely because their sample consisted of well-trained but non-elite triathletes. In summary, our set of variables seemed to be accurately selected and showed values similar to those reported in other studies using male elite triathletes.

Normalization of race time

The mean and standard deviation of normalized overall race time in Olympic distance triathlon for male triathletes (6827.57 ± 192.56 [s]; approximately 1:54 h) calculated in this study were comparable to those reported by Landers et al. (2000). A closer look at the Top 10 ranked athletes in the WTS from 2009 until 2012 showed that the mean ± SD of overall race time were consistent with the values used, considering that German elite triathletes commonly have a Top 20 position in WTS races.

Preselection of variables

The purely statistical approach using exploratory factor analysis is devoid of subjective influences, and prioritizes variables based on their influence to a general factor. Therefore, variables with a small variance will typically be sorted out. This could be why only two anthropometric and three physiological variables provided a significant contribution in the computed linear regression models, which is unfavorable because it could result in a lack of explanation of variance. Additionally, a sufficient sized sample must be available. In contrast, the dominance paired comparison does not have high demands regarding the number of coaches consulted. An objective prioritization based on professional expertise seems to be a plausible preprocessing step, if combined with ANNs to model complex and non-linear patterns. A reduction of variables similar to an exploratory factor analysis could therefore not be achieved and the maximum number of variables used in the computational model must be specified manually. As a prime example, VO_2max is a common parameter characterizing the endurance of heterogeneous groups and predicting performance (Butts, Henry, & Mclean, 1991; Miura et al., 1997). This could be why national triathlon coaches select absolute and relative VO_2max as predictive parameters. In homogenous groups, VO_2max normally has only premising instead of predicting character, because of the small variation (Sleivert & Rowlands, 1996). This could possibly be a drawback of subjective assessments compared to the exploratory factor analysis as a purely statistical approach, which sorted out VO_2max.

Performance prediction

Landers et al. (2000) underlined the importance of identifying parameters predicting race performance. Besides potentially supporting the creation of new training programs, the information provided by performance prediction models could also be used in the field of talent diagnostics. Considering that the small and homogenous sample limits generalizability, the reported performance prediction models showed that specific influencing parameters generally exist. These parameters could allow more objective talent selection by defining minimum physical requirements (e.g. for specific age groups). Talent identification programs could also use information on advantageous anthropometric requirements to direct young athletes to the sport of triathlon. The design of training programs could be influenced by focusing on optimal training levels (e.g. to improve identified lactate levels).

The combination of a professional triathlon coach survey and ANNs provided two performance prediction models with medium and large explained percentages of variance, respectively (anthropometric: $R^2 = 0.43$; physiological: $R^2 = 0.86$). In comparison, the MLR showed clearly poorer results (anthropometric: $R^2 = 0.41$; physiological: $R^2 = 0.67$). Therefore, the predictions using ANNs outperformed those from the purely statistical approach comprising factor analysis and multiple regressions. Furthermore, a closer look at the SEE (based on MLR: anthropometric: 155.14 [s]; physiological: 117.27 [s]; based on ANN: anthropometric: 144.56 [s]; physiological: 92.82 [s]) revealed that these are smaller than the performance variation of individual athletes (e.g. SD of race time of Javier Gomez during WTS 2014: 200.93 [s]) and of the Top 10 athletes in WTS 2014 (SD of race time: 217.83 [s]), which confirms the results of the performance prediction models.

The first MLR model yielded the anthropometric parameters pelvis width and shoulder width as significant predictors of overall race time in elite Olympic distance triathlon. These two variables could theoretically have an impact on running economy (Barnes & Kilding, 2015). Shoulder width seems to be a predictor for swimming performance, which is necessary to be in the first group getting out of the water. In contrast, pelvis width should be smaller, which was already shown for distance runners (Anderson, 1996; Williams, Cavanagh, & Ziff, 1987), and is therefore plausible in connection with the importance of the run part in elite Olympic distance triathlon. The ANN model used the five anthropometric parameters, body weight, BMI, lean body mass, and absolute as well as relative body fat, which were identified through dominance paired comparison as most important for overall race time in elite Olympic distance triathlon. Parameters such as body height or BMI normally show too small variations to get significant results in small and homogenous samples (Table 1). In the present study, the triathlon coaches were partially responsible for young athletes in national squads, where the mentioned variables have a higher influence and a greater variance than in elite triathletes.

The second MLR model yielded the physiological parameters running pace at 3-mmol·L^{-1} blood lactate, maximum lactate, and maximum respiratory rate in mobilization test as significant predictors of overall race time in elite Olympic distance triathlon. The ANN model used the five physiological parameters, maximum running pace, running pace at 3-mmol·L^{-1} blood lactate, maximum running pace in mobilization test, and absolute as well as relative VO$_2$max identified by a dominance paired comparison as most important for overall race time in elite Olympic distance triathlon. Both approaches identified running pace at 3 mmol·L^{-1} blood lactate as important for overall race time. This variable describes the possibility of an athlete to realize a higher pace with the same utilization of metabolic processes. The mentioned lactate interval is mainly used while competing in Olympic distance triathlon, and therefore leads directly to a faster race time. Some studies identified VO$_2$max or ventilatory thresholds as important for performance prediction in heterogeneous groups (Butts et al., 1991; Miura et al., 1997). This could be why national triathlon coaches select absolute and relative VO$_2$max as predictive parameters, particularly for young athletes. In homogenous groups, VO$_2$max normally has only premising instead of predicting character because of only small variation in VO$_2$max (Sleivert & Rowlands, 1996). In contrast, maximum values such as maximum lactate and maximum respiratory rate in mobilization as well as maximum running pace allow a valid assessment of anaerobic capacities. EFA and MLR as well as DPC and ANN used these kind of variables, which seems to be plausible: nearly all races of the WTS were actually won during the running discipline, especially in the final spurt. High lactate values and high running paces could therefore be important factors for overall race time. The maximum respiratory rate could also influence this kind of race situation, because a selectively high oxygen uptake is required to prevent the formation of lactate.

Limitations

The sample in this study was elite, small, and homogenous, which limits generalizability to other triathlete cohorts. However, generalizability of results to other triathlete cohorts was not the aim of this study; we focused on elite athletes. National squads for triathlon are generally small; compared to other sports, only 4 -5 athletes are included in the elite Olympic distance triathlon squad each year, and elite athletes are often reluctant to participate in experiments. Additionally, individual training schedules and differences in season calendars complicate experimental laboratory studies with this special population. Therefore, one of our aims was to assess whether overall Olympic distance triathlon race time of elite athletes could be predicted using regular performance diagnostics. To overcome the drawback of having a small number of available athletes, we developed an algorithm that helped us to increase the number of datasets used in the statistical analyses, by collecting performance diagnostics over a period of four years. We only included data sets if two requirements were fulfilled: (1) availability of a complete set of variables from the laboratory tests and (2) a finished Olympic distance triathlon race within 8 weeks after each single performance diagnostic. However, despite this improvement, no prediction models could be determined by combining anthropometric and physiological variables due to the sample size.

Further, we did not take the results of laboratory tests in swimming and cycling into account. This was because the protocols slightly changed over the period of interest (2008-2012), which led to inconsistent data sets. Therefore, we decided to exclude these sources of information to avoid a further reduction of the sample size. Nevertheless, the general tactical behavior in elite Olympic distance triathlon races allows the use of running diagnostics alone to generate meaningful results. The swimming and cycling disciplines in elite Olympic distance triathlon only have premising function whereas the running discipline is normally the critical factor for success (Fröhlich, Klein, Pieter, Emrich, & Gießling, 2008; Vleck, Burgi, & Bentley, 2006). Therefore, the results of the present prediction models, with only running diagnostics as physiological parameter, could be considered appropriate. However, we are planning to incorporate more comprehensive data sets (swimming, cycling, and running diagnostics) in future studies, since the test protocols for swimming and cycling have now been standardized.

Conclusion

Two different approaches to determine performance prediction models of overall race time in elite Olympic distance triathlon were developed without interfering with individual training programs, through triathlete participation in a standardized experimental study and the identification of important parameters collected through laboratory tests. According to these models, the combination of an exploratory factor analysis and multiple linear regression provided appropriate explanations of variance in case of anthropometric (R^2 = 0.41) and physiological (R^2 = 0.67) variables. These were selected with a strong analytical procedure, using variables with a greater variance. The corresponding SEEs of 155.14 [s] (anthropometric variables) and 117.27 [s] (physiological variables) showed acceptable results when compared to performance variations of individual athletes (e.g. SD of race time of Javier Gomez during WTS 2014: 200.93 [s]) and of the Top 10 athletes in WTS 2014 (SD of race time: 217.83 [s]), and therefore confirmed the results of the performance prediction models.

The advantage of ANNs compared to MLRs is the possibility to take non-linear relationships into account and to model more complex patterns. Therefore, the trained ANNs considering expertise of professional triathlon coaches through dominance paired comparison as preselection method could preferably be used to predict individual race time based on the values of an actual performance diagnostic. The explanations of variance and the standard errors of the estimate in case of anthropometric (R^2 = 0.43; SEE = 144.56 [s]) and physiological (R^2 = 0.86; SEE = 91.82 [s]) variables were an improvement over those of the the purely statistical approach.

Finally, the results of the present study show that future research should focus on collecting larger samples, and on the developmental process of young triathletes, with a focus on the influence on performance prediction models. Information from previous races, such as overall or split times and training indicators, could also enhance prediction (Gilinsky, Hawkins, Tokar, & Cooper, 2014).

Acknowledgments

We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

References

Ackland, T. R., Blanksby, B. A., Landers, G., & Smith, D. (1998). Anthropometric profiles of elite triathletes. *Journal of Science and Medicine in Sport*, *1*(1), 52–56. https://doi.org/10.1016/S1440-2440(98)80008-X

Anderson, T. (1996). Biomechanics and running economy. *Sports Medicine*, *22*(2), 76–89.

Barnes, K. R., & Kilding, A. E. (2015). Running economy: measurement, norms, and determining factors. *Sports Medicine - Open*, *1*(1), 357. https://doi.org/10.1186/s40798-015-0007-y

Bassett, D. R. (2000). Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine & Science in Sports & Exercise*, *32*(1), 70–84. https://doi.org/10.1097/00005768-200001000-00012

Butts, N. K., Henry, B. A., & Mclean, D. (1991). Correlations between VO2max and performance times of recreational triathletes. *Journal of Sports Medicine and Physical Fitness*, *31*(3), 339–344.

Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, *2*(2), 1–10. https://doi.org/10.1080/17461390200072201

Fröhlich, M., Klein, M., Pieter, A., Emrich, E., & Gießling, J. (2008). Consequences of the Three Disciplines on the Overall Result in Olympic-distance Triathlon. *International Journal of Sports Science and Engineering*, *2*(4), 204–210.

Gilinsky, N., Hawkins, K. R., Tokar, T. N., & Cooper, J. A. (2014). Predictive variables for half-Ironman triathlon performance. *Journal of Science and Medicine in Sport*, *17*(3), 300–305. https://doi.org/10.1016/j.jsams.2013.04.014

Hair, J. F. (1995). *Multivariate data analysis with readings* (4th ed). Englewood Cliffs, N.J.: Prentice Hall.

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-plus, R, and SAS*. New York: Springer.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hue, O., Le Gallais, D., Boussana, A., Chollet, D., & Prefaut, C. (2000). Performance level and cardiopulmonary responses during a cycle-run trial. *International Journal of Sports Medicine*, *21*(4), 250–255. https://doi.org/10.1055/s-2000-8883

Hue, O., Le Gallais, D., Chollet, D., & Préfaut, C. (2000). Ventilatory threshold and maximal oxygen uptake in present triathletes. *Canadian Journal of Applied Physiology*, *25*(2), 102–113.

Hue, O. (2003). Prediction of Drafted-Triathlon Race Time From Submaximal Laboratory Testing in Elite Triathletes. *Canadian Journal of Applied Physiology*, *28*(4), 547–560. https://doi.org/10.1139/h03-042

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Knechtle, B., Wirth, A., Rüst, C. A., & Rosemann, T. (2011). The Relationship between Anthropometry and Split Performance in Recreational Male Ironman Triathletes. *Asian Journal of Sports Medicine*, *2*(1), 23–30.

Knussmann, R., & Barlett, H. L. (1988). *Anthropologie : Handbuch der vergleichenden Biologie des Menschen [Anthropology : Handbook of comparative biology of humans]* (2nd ed). Stuttgart: Fischer.

Kohrt, W. M., Morgan, D. W., Bates, B., & Skinner, J. S. (1987). Physiological responses of triathletes to maximal swimming, cycling, and running. *Medicine & Science in Sports & Exercise*, *19*(1), 51–55.

Landers, G. J., Blanksby, B. A., Ackland, T. R., & Smith, D. (2000). Morphology and performance of world championship triathletes. *Annals of Human Biology*, *27*(4), 387–400.

Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441. https://doi.org/10.1137/0111030

McLaughlin, J. E., Howley, E. T., Bassett, D. R., Thompson, D. L., & Fitzhugh, E. C. (2010). Test of the classic model for predicting endurance running performance. *Medicine & Science in Sports & Exercise*, *42*(5), 991–997. https://doi.org/10.1249/MSS.0b013e3181c0669d

Millet, G. P., Vleck, V. E., & Bentley, D. J. (2009). Physiological differences between cycling and running: lessons from triathletes. *Sports Medicine*, *39*(3), 179–206. https://doi.org/10.2165/00007256-200939030-00002

Millet, G. P., Vleck, V. E., & Bentley, D. J. (2011). Physiological requirements in triathlon. *Journal of Human Sport and Exercise*, *6*(2 Suppl.), 184–204. https://doi.org/10.4100/jhse.2011.62.01

Miura, H., Kitagawa, K., & Ishiko, T. (1997). Economy during a simulated laboratory test triathlon is highly related to Olympic distance triathlon. *Journal of Human Sport and Exercise*, *18*(4), 276–280. https://doi.org/10.1055/s-2007-972633

Schabort, E. J., Killian, S. C., St Clair Gibson, Hawley, J. A., & Noakes, T. D. (2000). Prediction of triathlon race time from laboratory testing in national triathletes. *Medicine & Science in Sports & Exercise*, *32*(4), 844–849.

Silva, A. J., Costa, A. M., Oliveira, P. M., Reis, V. M., Saavedra, J., Perl, J., & Marinho, D. A. (2007). The Use of Neural Network Technology to Model Swimming Performance. *Journal of Sports Science & Medicine*, *6(1)*, 117–125.

Sleivert, G. G., & Rowlands, D. S. (1996). Physical and physiological factors associated with success in the triathlon. *Sports Medicine*, *22*(1), 8–18.

Sleivert, G. G., & Wenger, H. A. (1993). Physiological predictors of short-course triathlon performance. *Medicine & Science in Sports & Exercise*, *25*(7), 871–876.

Stratton, E., O'Brien, B. J., Harvey, J., Blitvich, J., McNicol, A. J., Janissen, D.,. . . Knez, W. (2009). Treadmill Velocity Best Predicts 5000-m Run Performance. *International Journal of Sports Medicine*, *30*(1), 40–45.

Tittel, K., & Wutscherk, H. (1972). *Sportanthropometrie : Aufgaben, Bedeutung, Methodik und Ergebnisse biotypologischer Erhebungen [Sports Anthropology : tasks, meanings, methodology and results of biotypological surveys]*. Leipzig: Barth.

Van Schuylenbergh, R., Eynde, B. V., & Hespel, P. (2004). Prediction of sprint triathlon performance from laboratory tests. *European Journal of Applied Physiology*, *91*(1), 94–99. https://doi.org/10.1007/s00421-003-0911-6

Vleck, V. E., Burgi, A., & Bentley, D. J. (2006). The consequences of swim, cycle, and run performance on overall result in elite olympic distance triathlon. *International Journal of Sports Medicine*, *27*(1), 43–48. https://doi.org/10.1055/s-2005-837502

Williams, K. R., Cavanagh, P. R., & Ziff, J. L. (1987). Biomechanical studies of elite female distance runners. *International Journal of Sports Medicine*, *8 Suppl 2*, 107–118.

Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks. *International Journal of Forecasting*, *14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7