

Article

Applying Artificial Intelligence to Determine the Required Positions in Each Discipline for Overall Olympic Triathlon Success

Pablo García-González ^{1,*} , Luca A. Bianchini ², Andrea Fuk ³ , Simone Villanova ³,
José Antonio González-Jurado ¹  and Maria Francesca Piacentini ³ 

¹ Physical Performance & Sports Research Center (CIRFD), Pablo de Olavide University, 41013 Seville, Spain

² DBInformation S.P.A., Centro Direzionale Milanofiori Strada 4, Palazzo A, scala 2, 20057 Assago, Italy

³ Department of Movement, Human and Health Sciences, University of Rome “Foro Italico”, 00135 Rome, Italy

* Correspondence: pablogartri@gmail.com

Abstract

Background: The aim of this study was to determine the optimal discipline position in the overall result of Olympic-distance triathlon. **Methods:** Data were extracted for free from the API (Application Programming Interface) service on the World Triathlon website and collected using a custom Python code. Statistical and machine learning analyses were employed within a Jupyter Notebook file. Linear and polynomial regressions were calculated between the overall race position and final positions in each discipline. Descriptive statistics and machine learning analyses were computed to identify the average position and most likely average position required in each discipline, respectively. A heatmap correlation analysis was conducted between the best overall triathletes and the best discipline triathletes. Differences between the two sub-databases were assessed using the student's *t*-test. **Results:** Across all disciplines, the average position required in each segment remains consistently better than 13th place. The heat map shows a very small, negative correlation between the best time in each discipline and the overall best race time (*p*-values < 0.001). The student's *t*-test establishes significant differences for all disciplines and overall race time (*p*-values < 0.001). **Conclusions:** Consistently high-level performance across all disciplines is essential for ensuring a podium finish or race victory in an Olympic triathlon. Achieving the best time in each discipline is not required to contend for victory, although running appears to be a strong predictor of overall race outcome.

Keywords: athletic performance; machine learning; physical endurance; sports; statistics

1. Introduction

Triathlon is a multi-sport event composed of three sport disciplines: swimming, cycling, and running, in that order, in addition to two transition phases between the disciplines. The standard triathlon distance is the Olympic distance, which includes a 1.5 km swim, a 40 km cycling segment, and a 10 km run. Triathlon has been extensively studied from several different perspectives. From a physiological view, the physiological requirements of triathletes have been clearly mentioned and discussed [1,2]. Other studies have focused on the biomechanical [3,4], nutritional [5,6], and psychological [7] aspects of triathlon. Turning to the quantitative performance analysis of triathlon, it is worth highlighting some studies that analyzed triathlon and other multi-sports performance with a focus on different training periodization strategies [8], identifying optimal pacing



Received: 11 February 2026

Revised: 3 March 2026

Accepted: 6 March 2026

Published: 17 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

strategies [9–11], establishing cutoff values [12–14], comparing performance across competitions [15,16], and assessing the impact of the COVID-19 pandemic [17,18], among several other relevant topics.

The search for the most influential discipline in determining the overall result in triathlon has been widely studied. This topic was addressed in the different existing triathlon distances [19–22] and categories [12,23]. According to some authors, it seems evident that pacing strategies vary depending on the race distance [24]. In the Olympic distance, authors have not reached full agreement on the decisive discipline for overall success. Unlike longer distances, the Olympic distance allows for drafting during the cycling segment. Early studies in male triathletes highlighted the primary roles of swimming and running, suggesting that cycling has a reduced impact on overall race performance [25]. Further research on females reported different tactical strategies between sexes across disciplines, with successful elite females showing superior swimming and cycling abilities, thereby gaining a greater advantage at the start of the run compared to elite males [26]. Other researchers argued that the cycling drafting position could be the key determinant of performance in short-distance competitions, with pedal cadence strategy also influencing the final running phase [27]. These are clear examples of why the methodology and data considered in the analysis can show different outcomes. Most of these studies used conventional statistical approaches, providing inconsistent results. These methods typically assume linear and independent relationships between segment performance and overall ranking, which may oversimplify the multidimensional structure of the triathlon competition. In Olympic distance, drafting during cycling alters the race dynamics and performance in one segment directly conditions the tactical context of the subsequent one. The interaction between disciplines may also be more complex than previously modeled. Consequently, methodological constraints may partly explain the inconsistent findings reported in the literature.

This may suggest that an alternative is required to properly address the complexity of large-scale performance datasets. Machine learning techniques provide an alternative that does not require predefined assumptions about the structure of the data and can consider the interaction effects between disciplines. This flexibility may offer a more accurate estimation of the relative contribution of each segment to the final ranking in the Olympic distance.

Artificial intelligence has been applied to triathlon for several purposes [28–31]. It makes possible to model non-linear relationships, capture interaction effects, handle large datasets, and estimate relative feature importance. Two articles used artificial intelligence to determine the most important discipline in triathlon. Weiss et al. (2024) [32] focused on the Ironman 70.3 distance, concluding that non-drafted cycling is the most significant predictor of overall triathlon performance. Martínez-Sobrino et al. (2023) [33] focused on a different category: mixed-team relays. The authors did not identify papers using artificial intelligence in Olympic-distance races for segment position analysis.

To the authors' knowledge, no study has yet utilized artificial intelligence to identify the most influential segment position in the final result of an Olympic-distance triathlon. The aim of this study is to determine the optimal position in each discipline to achieve success in the Olympic-distance triathlon. Based on previous research, it was hypothesized that: (i) performing the best time in each discipline would not be necessary to achieve overall race success, and (ii) running would have the greatest impact on achieving a desired final position.

2. Materials and Methods

2.1. Data Collection

Data were extracted from the API service (Application Programming Interface) on the World Triathlon website (<https://www.triathlon.org/>; accessed on 26 May 2024) by signing up for the free service. Since the data were in the public domain and available on the Internet, no formal request was made to the Ethics Committee. To perform the data collection, a custom Python code (Python Software Foundation, Wilmington, DE, USA; ver. 3.12) was written for different operations. The Python codes used are available on the GitHub repository (https://github.com/AEONPHYTON/triathlon_analysis; accessed on 26 May 2024). Once the data had been extracted, cleanup operations were performed using Python as the programming language and Jupyter Notebook (Project Jupyter, San Francisco, CA, USA; ver. 7.2) as the working tool [34]. Two databases were created including Olympic distance races (one for males and one for females) of the elite category. Previous editions of the Olympics were also included in the database.

2.2. Data Extrapolation and Data Cleaning

Files identifying the races by the race number (prog_id and event_id) were downloaded and used to obtain the corresponding data. The dataset was restricted to elite and U23 athletes, with U23 retained due to their participation alongside elite competitors. Standard-distance events were considered, separated by sex. Different variables were generated to track the record of each athlete's position after each of the three disciplines. Athletes lacking data for any discipline or transition, as well as athletes who were disqualified (DQS), did not start (DNS), or did not finish (DNF), were removed. Races without complete podium results or with implausible overall times (lower than 1 h, 28 min and 20 s and higher than 2 h and 30 min) were also excluded. The lower threshold was defined based on historical performance records in the elite Olympic-distance triathlon, as finishing times below this value would go beyond currently observed physiological and competitive limits for this distance. The upper threshold was selected to exclude events characterized by atypical race conditions such as extreme weather or altered course profiles that are not representative of standard elite-level Olympic-distance competitions. These thresholds were conservative and represented a small proportion of the dataset. These procedures granted a heterogeneous but valid dataset including more than 30 years of races, from 6 November 1989 to 24 May 2024. For each entry, the dataset included the following variables: sex, race date, final overall position, swimming position, cycling position, running position, and transition-related indicators when available. Segment positions were defined as the athlete's rank at the completion of each discipline relative to the full race field. All preprocessing and analyses were performed using fixed parameters and deterministic procedures, allowing for the full replication of the results using the provided codebase.

2.3. Preliminary Statistical Analysis

Once the cleaning was completed, two databases (males and females) were obtained for conventional statistical and machine learning analysis. The selected competitions were: Continental Championships, World Championships, World Cup, Continental Cup, Recognized Event, Regional Championships, Recognized Games, Major Games (editions of the Olympics from 2000 to 2021), and World Championship Series. The reason for including different types of competitions was to obtain a large amount of data for analysis.

Two individual sub-databases were established. The first sub-database included only triathletes who achieved podium positions, along with their standings at the end of each discipline (podium3). The second sub-database focused on the top three performers in each discipline and their overall race times (discipline3).

Shapiro–Wilk, Kolmogorov–Smirnov, and Anderson–Darling tests were conducted to assess normality of distribution. The results indicated that the data sample did not follow a normal distribution in most cases. Nevertheless, all subsequent analyses were conducted in parallel using both parametric and non-parametric tests, producing similar results.

Linear and polynomial regressions were calculated between overall race position and final positions in each discipline for male and female triathletes. Descriptive statistics of the filtered data were computed to determine the average position required in each discipline to achieve the following outcomes in males and females: victory, podium finish, and 4th to 10th place. The Friedman test was used to determine whether significant differences existed among the positions in each discipline required to achieve a given final position.

Machine learning analyses to achieve probabilistic determination operated using the NGBoost software (NGBoost Development Team, open-source software; ver. 0.5.6). Unlike standard gradient boosting methods, NGBoost provides a full probabilistic prediction, rather than a single point estimate. The analysis provided the following values: Training Loss (TL) refers to the model's fit on the training dataset, where lower values indicate better performance. Validation Loss (VL) reflects the model's error on an unseen validation set; a constant value of 0.0000 indicates that no separate validation set was used in the run. The Scaling Factor (S) adjusts the magnitude of each model update per iteration. Norm (N) represents the magnitude of parameter updates applied during the optimization process. 40%, 60%, and 80% thresholds were selected to represent a low, medium, and high probability range. The analysis was restricted to a maximum of 60 positions, following ITU World Championship Series, where typical participation ranges between 55 and 65 athletes per sex. This threshold was chosen to reflect the realistic upper limit of competitors in elite-level events and to exclude positions unlikely to appear in such competitions.

A heatmap correlation analysis was conducted between the two sub-databases to identify and compare key differences. A student's *t*-test was performed to compare the two sub-databases. Correlations were established between the final race position and the position at the end of each discipline. Multiple polynomial regressions and one linear regression were employed to assess the stability of the R^2 values across all models.

3. Results

3.1. Number of Winners That Achieved First Position at the End of All Disciplines

The number of races included in the analysis was 1384 for males and 1387 for females, resulting in a total of 45,884 male triathletes and 37,106 female triathletes considered for the analysis. The number of events per year is shown in Figure 1. The number of times that females who finished 1st overall also placed 1st in the swim, bike, and run disciplines was 20 (5.77% of the total races analyzed), while it was 11 for male triathletes (1.45% of the total races considered) (Figure 2).

3.2. Optimal Position in the Different Disciplines to Achieve Winning Position, Podium, and 4th to 10th Place

A correlation exists between the final position and the position at the end of each discipline (Table 1). Different polynomial regressions are shown only for the purpose of comparing the R^2 values across them. R^2 remains relatively consistent across different regression models. R^2 is medium-low for swimming in both males and females, medium-low for men's cycling, slightly higher for women's cycling, and high for running in both sexes (Table 1). The positions required in each segment to achieve a specific overall race position are described in Table 2. The average position required in each segment is consistently lower than 13th place, across all disciplines, sexes, and all placing categories

(winners, podium, and 4th to 10th place) (Table 2). For both males and females, the best position observed in any discipline is always first position (Table 2).

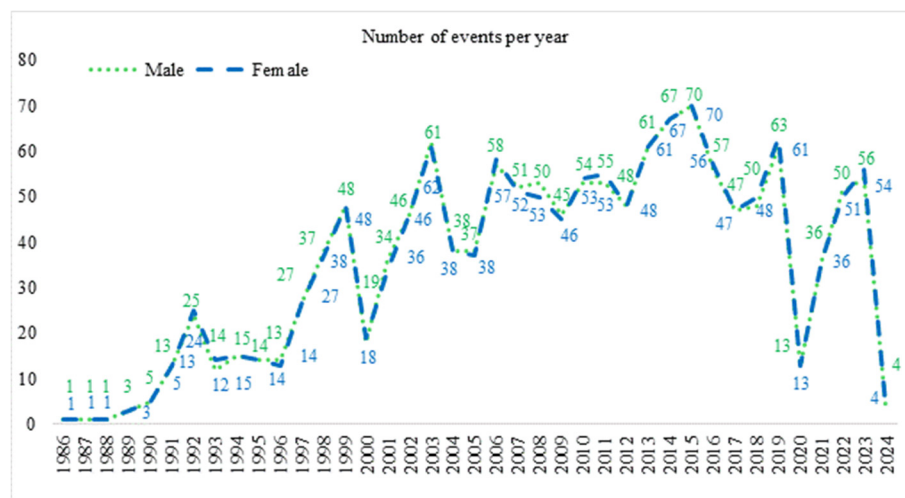


Figure 1. Number of events per year for males and females.

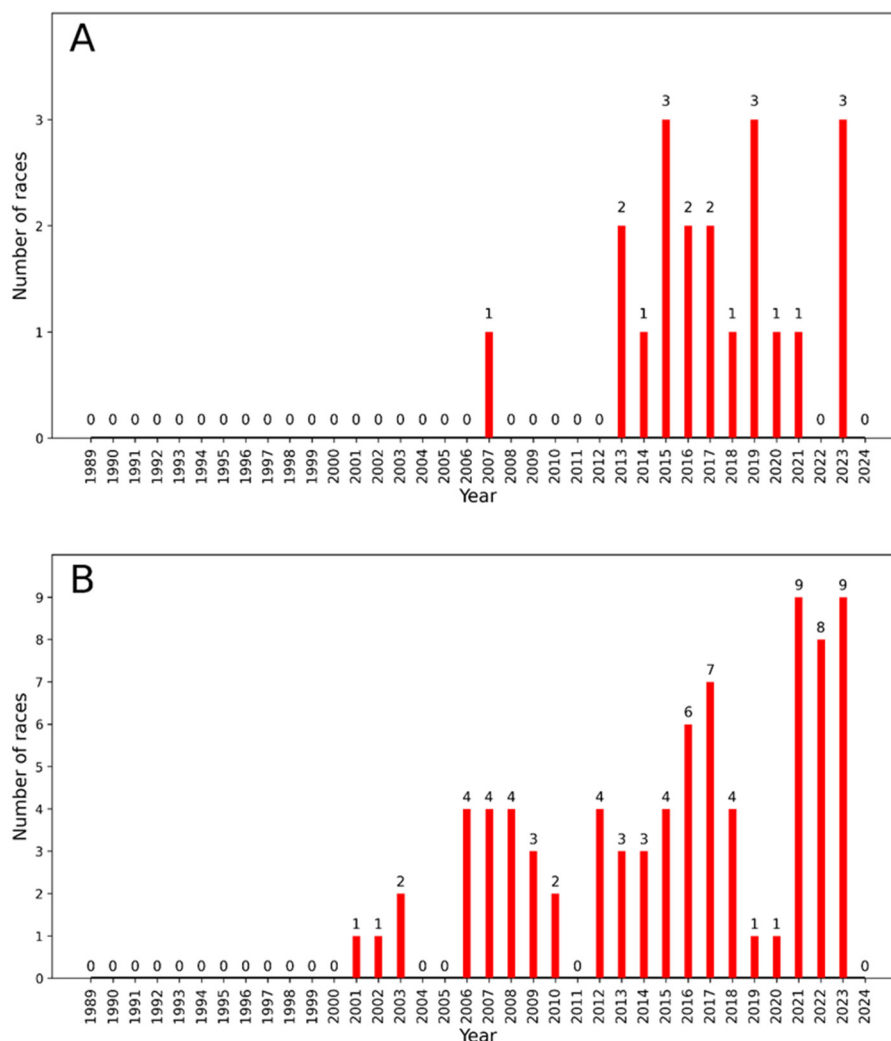


Figure 2. Number of race winners that achieved first position at the end of each discipline. Males (A) and females (B).

Table 1. R² and *p*-values for different regressions between overall race and disciplines final position for male and female triathletes.

		Swim (<i>p</i> -Value; R ²)	Bike (<i>p</i> -Value; R ²)	Run (<i>p</i> -Value; R ²)
Linear regression	Male	<0.001; 0.29	<0.001; 0.31	<0.001; 0.71
	Female	<0.001; 0.41	<0.001; 0.58	<0.001; 0.71
Polynomial (degree 2)	Male	<0.001; 0.29	<0.001; 0.31	<0.001; 0.71
	Female	<0.001; 0.41	<0.001; 0.58	<0.001; 0.71
Polynomial (degree 3)	Male	<0.001; 0.29	<0.001; 0.31	<0.001; 0.71
	Female	<0.001; 0.41	<0.001; 0.58	<0.001; 0.71
Polynomial (degree 4)	Male	<0.001; 0.29	<0.001; 0.31	<0.001; 0.71
	Female	<0.001; 0.41	<0.001; 0.58	<0.001; 0.71
Polynomial (degree 5)	Male	<0.001; 0.29	<0.001; 0.31	<0.001; 0.71
	Female	<0.001; 0.41	<0.001; 0.58	<0.001; 0.71

Polynomial (degree 2, 3, 4 and 5) = Polynomial regressions of two, three, four, and five degrees, respectively. R² = coefficient of determination; *p*-value = *p*-value (linear or polynomial regression); Significance level = 0.05.

Table 2. Male and female individual discipline final positions necessary to achieve overall race victory, podium, and 4th to 10th place.

	Swim	Bike	Run
Male Winner			
Mean ± SD (Min; Max)	7.72 ± 8.62 (1, 93)	8.81 ± 10.69 (1, 111)	2.78 ± 4.15 (1, 85)
<i>p</i> -value	<0.001		
Male Podium			
Mean ± SD (Min; Max)	8.56 ± 9.08 (1, 105)	9.36 ± 10.43 (1, 111)	3.77 ± 4.28 (1, 89)
<i>p</i> -value	<0.001		
Male 4th to 10th Place			
Mean ± SD (Min; Max)	12.23 ± 10.22 (1, 134)	12.05 ± 9.89 (1, 112)	8.54 ± 5.51 (1, 99)
<i>p</i> -value	<0.001		
Female Winner			
Mean ± SD (Min; Max)	5.16 ± 5.97 (1, 66)	4.42 ± 5.54 (1, 47)	2.26 ± 2.77 (1, 48)
<i>p</i> -value	<0.001		
Female Podium			
Mean ± SD (Min; Max)	5.97 ± 6.25 (1, 66)	5.14 ± 5.64 (1, 53)	3.37 ± 3.32 (1, 63)
<i>p</i> -value	<0.001		
Female 4th to 10th Place			
Mean ± SD (Min; Max)	9.63 ± 7.48 (1, 75)	8.61 ± 6.12 (1, 64)	8.02 ± 4.62 (1, 69)
<i>p</i> -value	<0.001		

Swim-Bike-Run = Required average position, standard deviation, minimum position, maximum position. *p*-value corresponding to Friedman test; significance level = 0.05.

Regarding the machine learning technique, TL decreases from 0.85 to approximately 0.47 over five iterations. N decreases from 6.82 to 0.77 in males and to 1.48 in females over the same number of iterations. VL remains consistently at 0.0000 for both males and females. Figure 3 illustrates the probabilistic determination of victory, podium finish, and 4th to 10th place. Each axis represents a discipline: swim (top right), bike (top left), and run (bottom). Each colored line corresponds to an outcome category at a specific probability

threshold. Points closer to the center represent better (lower) average positions. To achieve 4th to 10th place, the lines are visibly farther from the center compared to podium or race victory outcomes. The most probable position to achieve a 4th to 10th place finish is around position 30 across all disciplines and probability thresholds in both males and females.

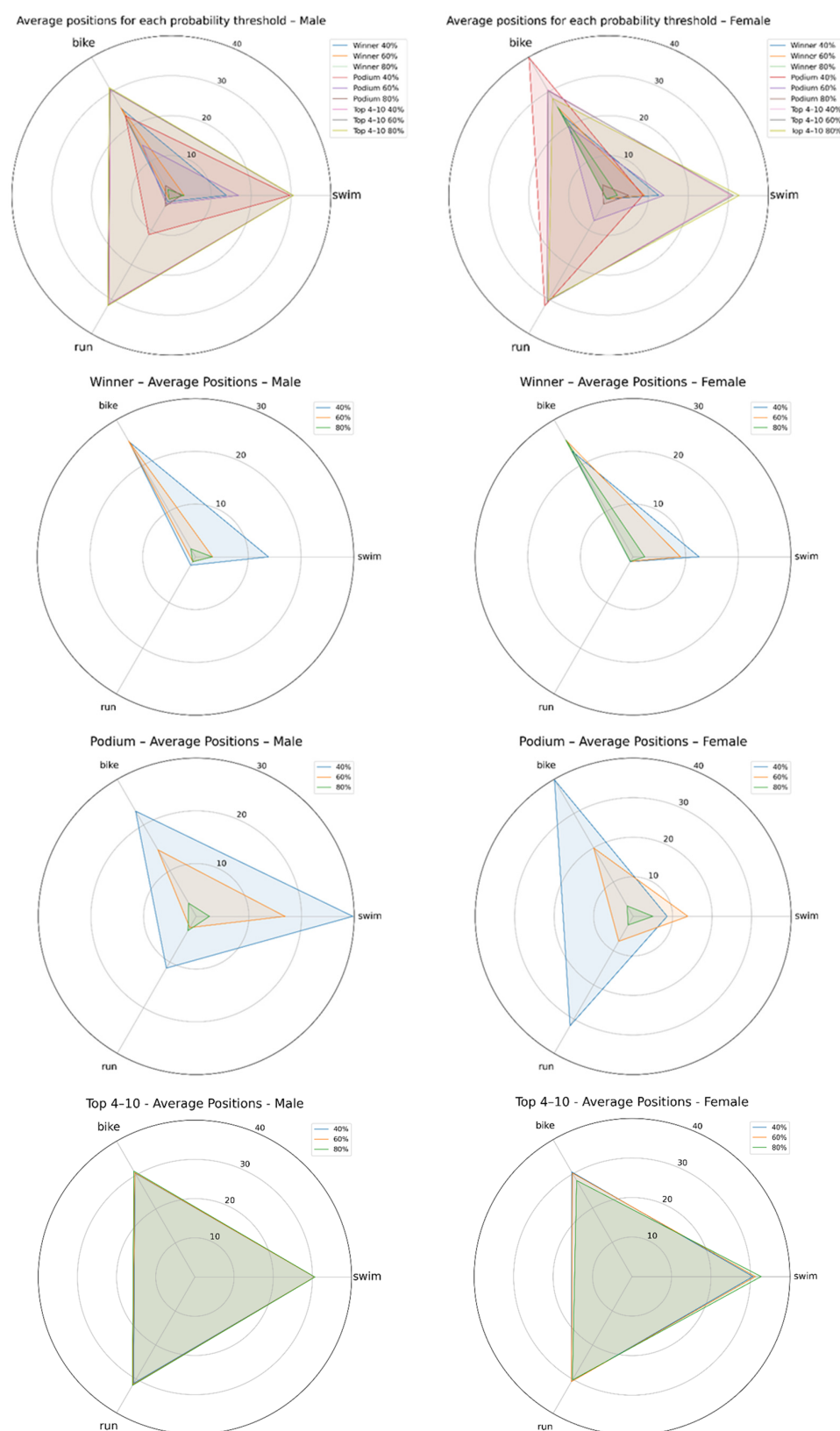


Figure 3. Radar chart (machine learning) with average position by probability threshold.

In the male category, the average positions required to achieve a podium were (29.74 ± 18.13) , (23.02 ± 18.88) , and (11.33 ± 19.40) for swimming, cycling, and running, respectively, at the 40% probability threshold (Figure 3). At the 60% threshold, these positions were (16.85 ± 18.93) , (14.56 ± 16.46) , and (2.46 ± 0.60) , respectively (Figure 3). At the 80% threshold, the average positions were (2.50 ± 1.41) , (2.88 ± 0.96) , and (3.12 ± 0.34) , respectively (Figure 3). In the case of race victory, the average positions were (13.82 ± 14.64) , (25.18 ± 20.25) , and (1.82 ± 6.40) , respectively, at the 40% threshold; (3.24 ± 3.56) , (25.40 ± 19.20) , and (1.02 ± 0.16) , respectively, at the 60% threshold; and (3.15 ± 2.58) , (1.77 ± 1.36) , and (1.08 ± 0.28) , respectively, at the 80% threshold (Figure 3).

For the females, the average positions required to achieve a podium finish were (8.61 ± 12.24) , (39.80 ± 21.30) , and (31.93 ± 22.71) for swimming, cycling, and running, respectively, at the 40% probability threshold (Figure 3). At the 60% threshold, these positions were (13.85 ± 17.71) , (19.95 ± 20.00) , and (7.29 ± 15.78) , respectively (Figure 3). At the 80% threshold, the average positions were (5.00 ± 6.73) , (3.05 ± 0.78) , and (2.53 ± 0.61) , respectively (Figure 3). In the case of race victory, the average positions were (12.55 ± 16.15) , (23.10 ± 20.75) , and (1.05 ± 0.25) , respectively, at the 40% threshold; (9.00 ± 14.79) , (25.55 ± 20.63) , and (1.00 ± 0.06) , respectively, at the 60% threshold; and (2.21 ± 1.33) , (25.57 ± 20.53) , and (1.00 ± 0.00) , respectively, at the 80% threshold (Figure 3).

Radar charts were used to visualize the average race position across the three disciplines. Each axis represents one discipline (swim, cycle, run), and the radial distance indicates the mean finishing position within each segment for athletes above a given predicted probability threshold (40%, 60%, 80%). Differences between polygons therefore reflect changes in the average segment positioning as the probability threshold increases. This approach enables a multidimensional comparison of performance profiles across disciplines rather than examining each split independently.

Radar charts illustrating the average position of athletes in swim, bike, and run segments at different probability thresholds (40%, 60%, and 80%) from the machine learning model. Blue area (40%): represents athletes with a predicted probability $\geq 40\%$. Orange area (60%): represents athletes with a predicted probability $\geq 60\%$. Green area (80%): represents athletes with a predicted probability $\geq 80\%$. The top panels show the overall positions for males and females, while the bottom panels show 4th to 10th places for both sexes.

3.3. Comparison of Podium Database and Top Three Discipline Performers Database

The heat map shows a very small correlation between the total time of the top performers in each discipline and the total time of the race winners ($r = 0.004$ and $r = -0.012$ for male and female triathletes respectively) (Figure 4). In addition, the best time in each discipline is marginally and negatively correlated with the overall best race time ($r = -0.019$, $r = 0.043$ and $r = -0.006$ for swim, bike, and run times, respectively, in males/ $r = -0.011$, $r = 0.173$, and $r = -0.003$ for swim, bike, and run times, respectively, in females) (Figure 4). The student's *t*-test establishes significant differences for all disciplines and the overall race time in both sexes (*p*-values < 0.001) (Table 3).

Heat map showing the Pearson correlation coefficients (*r*) between performance times in the Podium3 (top three finishers) and Discipline3 (top three performers in each discipline) databases for male and female athletes. Each cell displays the correlation coefficient (*r*) and *p*-value (*P*) between the respective swim, bike, run, and total times. The significance level was set at 0.05. Darker red cells indicate stronger positive correlations, while lighter or grayish cells indicate weaker or negative correlations. NaN values indicate that the comparison does not make statistical sense.

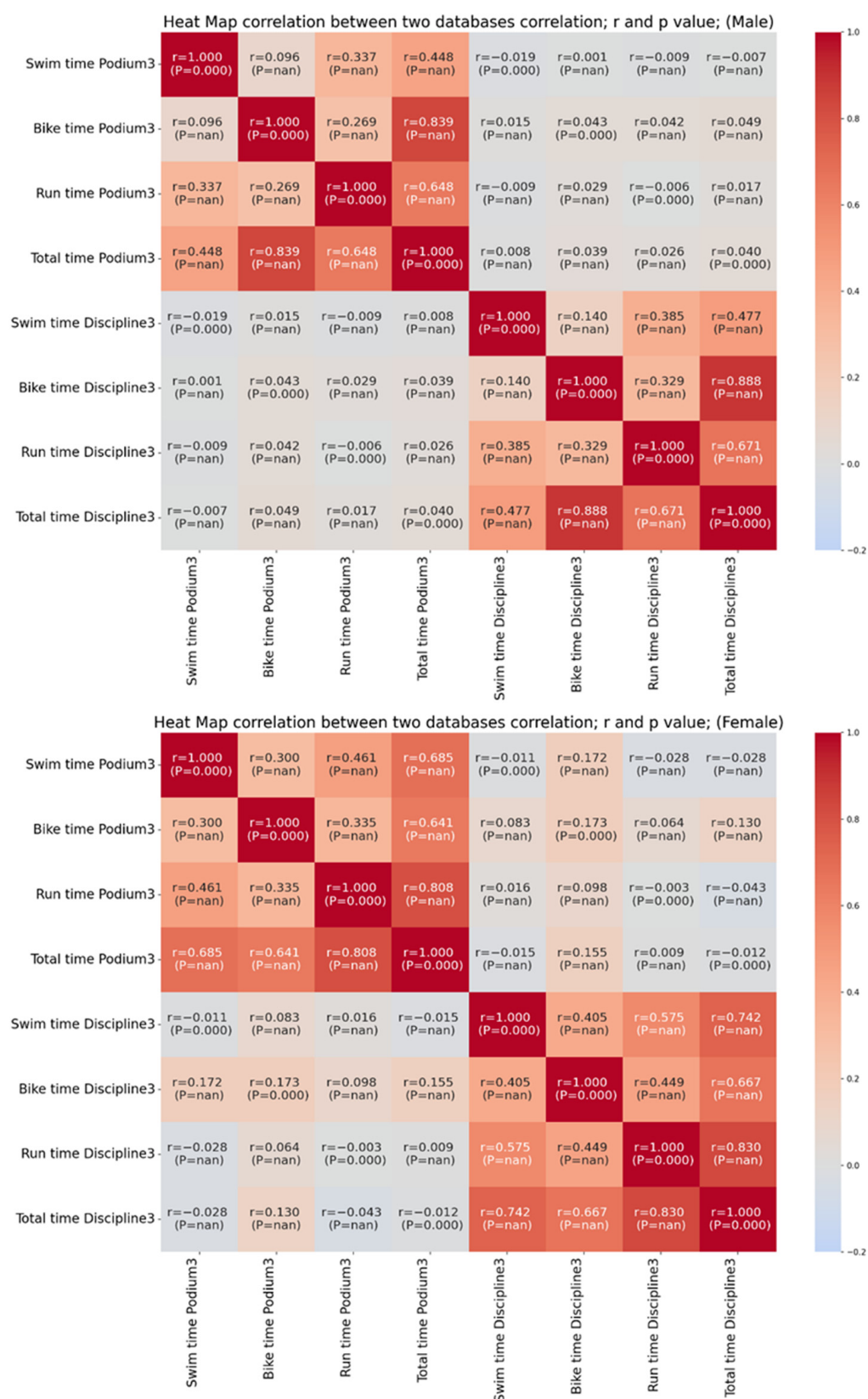


Figure 4. Heat map correlations between podium database (Podium3) and top three discipline performers database (Discipline3).

Table 3. Comparison between podium database and top three discipline performers database.

		Between Two Datasets					Combined Dataset	
	Database	Performance (s) $\pm$ SD (s)	Difference (s)	T	R ²	p	R ²	p
Male	Swim time	Discipline 3	1109.34 $\pm$ 90.15	28.86	11.95	0.54	<0.001	0.009
		Podium 3	1138.20 $\pm$ 95.55					
	Bike time	Discipline 3	3534.13 $\pm$ 274.15	59.37	8.18	0.70	<0.001	0.013
		Podium 3	3593.50 $\pm$ 284.12					
	Run time	Discipline 3	1941.12 $\pm$ 122.40	30.41	9.07	0.57	<0.001	0.007
		Podium 3	1971.53 $\pm$ 135.27					
	Total time	Discipline 3	6584.59 $\pm$ 386.50	175.67	18.18	0.43	<0.001	0.035
		Podium 3	6760.26 $\pm$ 374.71					
	Swim time	Discipline 3	1190.34 $\pm$ 186.50	49.57	10.57	0.463	<0.001	0.009
		Podium 3	1239.91 $\pm$ 186.92					
Female	Bike time	Discipline 3	3864.57 $\pm$ 651.35	106.37	6.84	0.546	<0.001	0.021
		Podium 3	3970.94 $\pm$ 584.63					
	Run time	Discipline 3	2272.61 $\pm$ 320.28	65.33	7.95	0.485	<0.001	0.009
		Podium 3	2272.61 $\pm$ 320.28					
	Total time	Discipline 3	7461.15 $\pm$ 897.42	123.31	5.76	0.475	<0.001	0.006
		Podium 3	7584.46 $\pm$ 805.51					

Discipline 3 = database corresponding to the best three performers of the discipline; Podium 3 = database corresponding to the podium positions; Performance = mean performance time. SD = standard deviation. Difference = absolute mean differences; T = *t*-test value; R² = R² value from *t*-test; *p* = *p*-value (*t*-test); Significance level = 0.05.

4. Discussion

The aim of this study was to determine the necessary position in each discipline to achieve success in elite Olympic-distance triathlon races for both sexes. To the best of our knowledge, this is the first article to date that has employed artificial intelligence to address this question. The most significant findings of our research are that performing the best split time in individual disciplines is not crucial for winning professional Olympic triathlon competitions, although performing well in running seems more important. Nevertheless, performing consistently well across all disciplines is essential for ensuring a podium finish or race success.

A correlation exists between the final position and the position at the end of each discipline (Table 1). R² remains relatively consistent across different regression models. The R² is medium-low for swimming in both males and females, medium-low for men's cycling, slightly higher for women's cycling, and high for running in both sexes (Table 1). Based on this initial analysis, it can be assumed that strong performances in both swimming and cycling are important, but a race victory is largely contingent on running performance. In other words, shorter running times significantly increase the likelihood of victory, provided the triathlete has attained competent placements in swimming and cycling. These findings align with what has been described by many authors and with popular belief. Some researchers attribute this fact to the influence of drafting during the cycling segment [35]. Moreover, our data confirm what has already been reported by [36], who suggest that cycling is somewhat more important for female triathletes although they exhibit a slightly higher R² (Table 1). Similarly, Piacentini et al. (2019) [37] reported that being in front upon arrival at the second transition and executing an optimal run are key factors in securing victory. Piacentini et al. (2019) [37] also reported that females tend to have more difficulty

reaching the lead group during the cycling segment, which lends additional importance to this discipline compared to men. Nonetheless, it is also important to consider that studies considering Olympic triathlon races where drafting was not permitted produced similar results [38]. García-González and González-Jurado (2025) [15,16] compared the times achieved in the three main triathlon competitions of the 21st century (drafting allowed) and found that medalists and other participants had similar performances in swimming and cycling, but the medalists performed better in running, confirming previous research [36,37]. For running, some authors have accentuated the importance of a correct pacing strategy, emphasizing that it is crucial to stick to a pre-determined pace strategy rather than following the lead group [9].

The necessary positions in each segment to achieve a specific overall race position following descriptive statistics are described in Table 2. The average position required in each segment is consistently lower than 13th place. Males generally need to rank within the top eight in swimming, the top ten in cycling, and the top three in running to win or reach the podium. Females are required to achieve top-five positions in swimming and cycling. Instead, they need a top-three position in running, similar to male athletes (Table 2). This highlights that athletes who excel in running, compared to those who do so in swimming and cycling, have higher odds of winning the race or achieving the podium. These results suggest that finishing near the front in swimming is not as important as achieving the fastest times in cycling or running. These results contrast with earlier work emphasizing the predictive value of the front position at the end of swimming for race success in males [36]. Using machine learning, they estimated segment time gaps for placing tiers (podium, 4th–10th, others). Their suggested ~10 s gap to the swim leader for males may imply a lower position than those we observed for podium positions (8.56 ± 9.08). In females, their proposed 20–30s (podium) and 10–20s (4th–10th) gaps align more closely with our findings (positions 5.97 ± 6.25 and 9.63 ± 7.48 , respectively) (Table 2). Our results also show that high performance in swimming, as well as in the other disciplines, is mandatory for triathlon success. It should be noted that the results described so far follow a descriptive statistical analysis approach, and we will now proceed to explain the results achieved by the machine learning analysis.

Based on the analysis conducted with NGBoost, several noteworthy conclusions can be drawn. As the probability threshold increases (from 40% to 80%) for both sexes, the average swim and bike positions show a moderate improvement while the run position consistently remains at 1st place, confirming that running performance has special importance for the final victory. Athletes who aim at the podium must still perform well across all three disciplines, although the acceptable range of positions is wider, especially in swimming and cycling. Finally, average swimming and cycling positions are significantly worse for athletes aiming to arrive between the 4th to 10th positions. Running positions are also less competitive for this group. Running seems to be the most important discipline independently of sex, aligning with the descriptive analysis.

Our analysis also reveals discipline-specific patterns by sex for race victories. A better swimming position is generally required than a cycling one to win or achieve a female podium finish. The same pattern appears across the 40%, 60%, and 80% probability thresholds, except at the 80% threshold for podium finishes, where the cycling position is more important than the swimming one (3.05 ± 0.78 in cycling versus 5.00 ± 6.73 in swimming). Female winners still maintain an average bike position around the 25th place at the 80% probability threshold, suggesting that recovery during the race is possible. Therefore, it can be stated that achieving a better swimming position is more advantageous for securing a podium position in comparison to cycling. A similar pattern is observed for the male winning position, again with the exception being at the 80% threshold (1.77 ± 1.36

in cycling versus 3.15 ± 2.58 in swimming). However, positions in cycling and swimming positions are very similar when aiming for the podium. It appears that both disciplines require a comparable level of performance in the male category to achieve a podium finish, suggesting that a better cycling position is needed compared to for females.

Both males and females need to achieve top-three and top-five swimming positions at the 80% probability threshold to secure victory. On the other hand, the average swimming position requirement at the 40% threshold drops to around the top fifteen. These findings suggest that the probability of winning is higher when athletes are already in leading positions after swimming and cycling, and that scenarios involving poor swimming, moderate cycling, and exceptional running are rare among winners. This does not lead us to the same conclusion achieved in our descriptive statistics presented in Table 2. The implementation of artificial intelligence in the analysis underlines the importance of achieving an advantageous position exiting the water and the relevance of cycling for male triathletes, but not for females. Therefore, our study agrees with others [36] in highlighting the importance of being at the front exiting the water, as they proposed that a maximum difference of ten seconds from the leader for success. The similar approach used in their paper compared to ours, both using machine learning techniques, is likely the reason why our results match now and did not do so with the traditional statistics approach also employed in this study.

Female winners can occasionally offset a weaker cycling performance with exceptional swimming and running results. In contrast, male winners more consistently occupy top positions across all three disciplines. Cycling positions also show greater variability compared to swimming and running. In many races, especially those allowing drafting, large athlete packs form during the bike leg, meaning that finishing 10th or 30th in cycling may correspond to the same time if the athlete remains within the lead pack. This dynamic helps explain why the cycling position is less predictive of final outcomes compared to swimming and running positions.

Swimming depends on anaerobic power and start intensity, but once athletes enter the lead pack, pacing stabilizes and variability decreases. In drafting-legal cycling, aerodynamic benefits reduce the direct impact of individual power output, allowing athletes to conserve energy within groups and weakening its predictive value. Running, however, occurs under accumulated glycogen depletion and neuromuscular fatigue, making it highly dependent on fatigue resistance and metabolic efficiency. Small physiological differences are therefore amplified in the final segment, which may explain its stronger association with overall race success. Some authors proved that triathlon running is harder than control running at the same speed in terms of oxygen consumption, respiratory frequency, ventilation rate, and heart rate [39]. Prior exercise may elevate markers of muscle damage, such as creatine kinase and lactate dehydrogenase, potentially enabling higher-level triathletes to experience smaller performance decrements compared to their competitors.

This machine learning analysis quantifies the likelihood of success for specific combinations of positions, making the findings actionable for real scenario decision-making in training and race planning. It also demonstrates that classical statistics, such as those described above and in other research [15,16], and machine learning lead to consistent conclusions, for example by stating the running segment as a key discipline for Olympic triathlon success. However, this study adds some conclusions, such as the importance of the cycling segment or the relevance of the position at the end of the swimming segment, which is a contribution of this paper.

Given the importance of running, we tested whether the best discipline performers (including the best running performers) were the same as the race winners. Two individual sub-databases were established for the analysis. The first sub-databases included only triathletes who achieved podium positions, along with their standings at the end of each

discipline (podium3). The second sub-databases focused on the top three performers in each discipline and their overall race times (discipline3).

The heat map shows an almost nonexistent correlation between the total time of the top performers in each discipline and the total time of the race winners ($r = 0.04$ and $r = 0.012$ for male and female triathletes, respectively) (Figure 4). In addition, the best time in each discipline is marginally correlated with the overall best race time (Figure 4). This proves that podium achievers are not necessarily the best performers in the individual segments, although they were close. This occurs for both males and females.

Likewise, the student's *t*-test establishes significant differences between the podium (podium3) and the top three discipline performers databases (discipline3) for all disciplines and overall race time in both sexes (p -values < 0.001) (Table 3). This provides conclusive evidence that both databases are significantly different and supports the observations made in Figure 4. The data show that outstanding individual performance in each discipline is not crucial for success in an Olympic triathlon, and that consistent high-level performance across all disciplines is essential for ensuring a podium finish or race victory. Once again, these statements reinforce the results previously reported by similar investigations [37].

This study has some limitations. Given the nature of the study, three limitations can be highlighted. First, there are a number of missing values not published by the ITU. This means that, despite careful data cleaning, there is a possibility that incorrect records were included in the analysis. Moreover, this study adopts a quantitative approach, which implies a lack of consideration for physical or other nature-related variables, such as training, previous experience, circuit layout, or climatic conditions. Finally, analyzing competitions over a long period, including competitions where drafting was allowed and others where it was forbidden, may have a small influence on the results, although preliminary analysis shows that the results would be similar to those reported in this study. Regarding the use of machine learning techniques, missing split times and potential recording inconsistencies may introduce bias into probability estimations. Furthermore, machine learning models identify statistical associations rather than causal mechanisms; therefore, the findings should be interpreted as probabilistic tendencies rather than deterministic rules of performance.

Based on our findings, athletes targeting victory should aim to exit the swim within the leading group and maintain a cycling position that preserves tactical contact with the front pack, as scenarios involving recovery from distant positions appear statistically rare. For podium ambitions, maintaining consistent top-ten positioning across the swim and cycle segments substantially increases success probability, while running capacity remains the decisive differentiator in the final outcome.

Sex-specific patterns suggest that female athletes may benefit from prioritizing front-pack positioning during swimming to a greater extent, whereas male athletes require a more homogeneous performance across the swim and cycle segments. Coaches may therefore structure race simulations and tactical training sessions around position-based objectives rather than solely around time-based metrics, particularly in drafting-legal competitions where pack dynamics influence cycling variability.

For federations and talent development programs, these findings emphasize the importance of developing high-level running capacity without neglecting early race positioning skills. Athlete transitions to longer distances should also consider that the positional dynamics observed in Olympic-distance racing may not translate directly to non-drafting formats.

Future research should extend artificial intelligence applications from retrospective analysis toward predictive and monitoring frameworks in Olympic-distance triathlons. Integrating contextual variables (e.g., course profile, environmental conditions) could enable a real-time estimation of podium probabilities and support tactical decision-making during races.

5. Conclusions

Olympic triathlon success seems to be influenced by an athlete's position throughout the race. To win an Olympic triathlon, performing the best split times in one or more individual disciplines is not determinant. Performing well in only two disciplines is insufficient to secure a top-three position, as success is characterized by multidimensional consistency. Triathletes must achieve a higher level of performance in running than in swimming or cycling to secure any placing category. This relative performance difference is smaller in females.

Author Contributions: Conceptualization, L.A.B. and P.G.-G.; methodology, L.A.B.; software, L.A.B.; validation, P.G.-G., J.A.G.-J. and M.F.P.; formal analysis, L.A.B.; investigation, L.A.B. and P.G.-G.; resources, J.A.G.-J.; data curation, L.A.B.; writing—original draft preparation, P.G.-G.; writing—review and editing, L.A.B., A.F., S.V., J.A.G.-J. and M.F.P.; visualization, P.G.-G.; supervision, J.A.G.-J. and M.F.P.; project administration, M.F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research exclusively used publicly available data retrieved from the official World Triathlon website through its open API service. The dataset contained competition results and performance metrics. No interaction with human participants occurred, and no intervention or data collection involving human subjects was conducted by the researchers. Consequently, the study analyzed secondary, openly accessible performance data that are routinely published for informational and sporting purposes.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were extracted for free from the API (Application Programming Interface) service on the World Triathlon website (<https://www.triathlon.org/>; accessed on 26 May 2024) and collected using a custom Python code available on the GitHub repository (https://github.com/AEONPHYTON/triathlon_analysis; accessed on 26 May 2024).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI ver. 5.3) to assist with language refinement of some paragraphs. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the content of the published article.

Conflicts of Interest: Author Luca A. Bianchini was employed by DBinformation S.P.A. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Suriano, R.; Bishop, D. Physiological attributes of triathletes. *J. Sci. Med. Sport* **2010**, *13*, 340–347. [[CrossRef](#)]
2. Millet, G.P.; Vleck, V.E.; Bentley, D.J. Physiological requirements in triathlon. *J. Hum. Sport Exerc.* **2011**, *6*, 184–204. [[CrossRef](#)]
3. Cuba-Dorado, A.; Álvarez-Yates, T.; Carballo-López, J.; Iglesias-Caamaño, M.; Fernández-Redondo, D.; García-García, O. Neuromuscular changes after a Long Distance Triathlon World Championship. *Eur. J. Sport Sci.* **2023**, *22*, 1838–1848. [[CrossRef](#)]
4. Ambrosini, L.; Presta, V.; Galli, D.; Mirandola, P.; Vitale, M.; Gobbi, G.; Condello, G. Interlink Between Physiological and Biomechanical Changes in the Swim-to-Cycle Transition in Triathlon Events: A Narrative Review. *Sports Med.-Open* **2022**, *8*, 129. [[CrossRef](#)]
5. Jeukendrup, A.E.; Jentjens, R.L.P.G.; Moseley, L. Nutritional Considerations in Triathlon. *Sports Med.* **2005**, *35*, 163–181. [[CrossRef](#)]
6. Miguel-Ortega, Á.; Rodríguez-Rodrigo, M.A.; Mielgo-Ayuso, J.; Calleja-González, J. Triathlon: Ergo Nutrition for Training, Competing, and Recovering. *Nutrients* **2025**, *17*, 1846. [[CrossRef](#)]
7. Bales, J.; Bales, K. Triathlon: How to Mentally Prepare for the Big Race. *Sports Med. Arthrosc. Rev.* **2012**, *20*, 217–219. [[CrossRef](#)]
8. Clemente-Suárez, V.J.; Ramos-Campo, D.J. Effectiveness of reverse vs. Traditional linear training periodization in triathlon. *Int. J. Environ. Res. Public Health* **2019**, *16*, 2807. [[CrossRef](#)]

9. Etxebarria, N.; Wright, J.; Jeacocke, H.; Mesquida, C.; Pyne, D.B. Running your best triathlon race. *Int. J. Sports Physiol. Perform.* **2021**, *16*, 744–747. [[CrossRef](#)]
10. Hausswirth, C.; Le Meur, Y.; Bieuzen, F.; Brisswalter, J.; Bernard, T. Pacing strategy during the initial phase of the run in triathlon: Influence on overall performance. *J. Appl. Physiol.* **2010**, *108*, 1115–1123. [[CrossRef](#)]
11. Stjepanovic, M.; Knechtle, B.; Weiss, K.; Nikolaidis, P.T.; Cuk, I.; Thuany, M.; Sousa, C.V. Changes in pacing variation with increasing race duration in ultra-triathlon races. *Sci. Rep.* **2023**, *13*, 3692. [[CrossRef](#)]
12. Ledanois, T.; Hamri, I.; De Laroche Lambert, Q.; Libicz, S.; Toussaint, J.F.; Sedeaud, A. Cutoff value for predicting success in triathlon mixed team relay. *Front. Sports Act. Living* **2023**, *5*, 1096272. [[CrossRef](#)]
13. Gadelha, A.B.; Sousa, C.V.; Sales, M.M.; Rosa, T.d.S.; Flothmann, M.; Barbosa, L.P.; Aguiar, S.d.S.; Olher, R.R.; Villiger, E.; Nikolaidis, P.T.; et al. Cut-off values in the prediction of success in olympic distance triathlon. *Int. J. Environ. Res. Public Health* **2020**, *17*, 9491. [[CrossRef](#)]
14. Barbosa, L.P.; Sousa, C.V.; Aguiar, S.d.S.; Gadelha, A.B.; Nikolaidis, P.T.; Villiger, E.; Knechtle, B. The beginning of success: Performance trends and cut-off values for junior and the U23 triathlon categories. *J. Exerc. Sci. Fit.* **2022**, *20*, 16–22. [[CrossRef](#)] [[PubMed](#)]
15. García-González, P.; González-Jurado, J.A. Analysis and comparison of male triathlon performance at the Olympic Games, International Triathlon Union World Championship, and European Triathlon Union European Championship over a period of twenty-three years, considering both medalists and overall triathletes participants. *Gazz. Medica Ital.-Arch. Sci. Medicine* **2025**, *184*, 835–843.
16. García-González, P.; González-Jurado, J.A. Analysis and Comparison of Female Triathlon Top-Class Performance at Three Main Competitions over 23 Years Considering Race Position. *Ann. Appl. Sport Sci.* **2025**, *13*, e1490. [[CrossRef](#)]
17. García-González, P.; González-Jurado, J.A. Has Covid-19 pandemic influenced the performance of top-class triathletes in the Sprint and Olympic distance of the ITU World Triathlon Championship Series? *Retos* **2024**, *57*, 137–146. [[CrossRef](#)]
18. García-González, P.; González-Jurado, J.A. Has COVID-19 influenced the performance of top-class athletes in the ITU World Duathlon and World Aquathlon Championship? *Rev. Andal. Med. Deport.* **2024**, *17*, 151–158. [[CrossRef](#)]
19. Nikolaidis, P.T.; Valero, D.; Weiss, K.; Villiger, E.; Thuany, M.; Sousa, C.V.; Andrade, M.; Knechtle, B. Predicting overall performance in Ironman 70.3 age group triathletes through split disciplines. *Sci. Rep.* **2023**, *13*, 11492. [[CrossRef](#)]
20. Figueiredo, P.; Marques, E.A.; Lepers, R. Changes in contributions of swimming, cycling, and running performances on overall triathlon performance over a 26-year period. *J. Strength Cond. Res.* **2016**, *30*, 2406–2415. [[CrossRef](#)]
21. Lepers, R.; Knechtle, B.; Knechtle, P.; Rosemann, T. Analysis of ultra-triathlon performances. *Open Access J. Sports Med.* **2011**, *2*, 131–136. [[CrossRef](#)] [[PubMed](#)]
22. Olaya, J.; Fernández-Sáez, J.; Østerlie, O.; Ferriz-Valero, A. Contribution of segments to overall result in elite triathletes: Sprint distance. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8422. [[CrossRef](#)]
23. Quagliarotti, C.; Gaiola, D.; Bianchini, L.; Vleck, V.; Piacentini, M.F. How to Form a Successful Team for the Novel Olympic Triathlon Discipline: The Mixed-Team-Relay. *J. Funct. Morphol. Kinesiol.* **2022**, *7*, 46. [[CrossRef](#)]
24. Wu, S.S.X.; Peiffer, J.J.; Brisswalter, J.; Nosaka, K.; Lau, W.Y.; Abbiss, C.R. Pacing strategies during the swim, cycle and run disciplines of sprint, Olympic and half-Ironman triathlons. *Eur. J. Appl. Physiol.* **2015**, *115*, 1147–1154. [[CrossRef](#)]
25. Vleck, V.E.; Bürgi, A.; Bentley, D.J. The consequences of swim, cycle, and run performance on overall result in elite olympic distance triathlon. *Int. J. Sports Med.* **2006**, *27*, 43–48. [[CrossRef](#)]
26. Vleck, V.E.; Bentley, D.J.; Millet, G.P.; Bürgi, A. Pacing during an elite Olympic distance triathlon: Comparison between male and female competitors. *J. Sci. Med. Sport* **2008**, *11*, 424–432. [[CrossRef](#)]
27. Hausswirth, C.; Brisswalter, J. Strategies for Improving Performance in Long Duration Events Olympic Distance Triathlon. *Sports Med.* **2008**, *38*, 881–891. [[CrossRef](#)]
28. Martínez-Gramage, J.; Albiach, J.P.; Moltó, I.N.; Amer-Cuenca, J.J.; Moreno, V.H.; Segura-Ortí, E. A random forest machine learning framework to reduce running injuries in young triathletes. *Sensors* **2020**, *20*, 6388. [[CrossRef](#)]
29. Rothschild, J.A.; Stewart, T.; Kilding, A.E.; Plews, D.J. Predicting daily recovery during long-term endurance training using machine learning analysis. *Eur. J. Appl. Physiol.* **2024**, *124*, 3279–3290. [[CrossRef](#)] [[PubMed](#)]
30. Thuany, M.; Valero, D.; Villiger, E.; Forte, P.; Weiss, K.; Nikolaidis, P.T.; Andrade, M.S.; Cuk, I.; Sousa, C.V.; Knechtle, B. A Machine Learning Approach to Finding the Fastest Race Course for Professional Athletes Competing in Ironman® 70.3 Races between 2004 and 2020. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3619. [[CrossRef](#)] [[PubMed](#)]
31. Thuany, M.; Valero, D.; Villiger, E.; Fernandes, M.S.S.; Forte, P.; Weiss, K.; Nikolaidis, P.T.; Cuk, I.; Knechtle, B. A study of the fastest courses for professional triathletes competing in IRONMAN® triathlons. *Hum. Mov.* **2024**, *25*, 148–160. [[CrossRef](#)]
32. Weiss, K.; Valero, D.; Andrade, M.S.; Villiger, E.; Thuany, M.; Knechtle, B. Cycling is the most important predictive split discipline in professional Ironman® 70.3 triathletes. *Front. Sports Act. Living* **2024**, *6*, 1214929. [[CrossRef](#)]
33. Martínez-Sobrino, J.; Veiga, S.; Del Cerro, J.S.; González-Ravé, J.M. What is the Most Important Leg and Discipline in Triathlon Mixed Team Relays? *J. Hum. Kinet.* **2023**, *89*, 269–278. [[CrossRef](#)]

34. Rule, A.; Birmingham, A.; Zuniga, C.; Altintas, I.; Huang, S.-C.; Knight, R.; Moshiri, N.; Nguyen, M.H.; Rosenthal, S.B.; Pérez, F.; et al. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Comput. Biol.* **2019**, *15*, e1007007. [[CrossRef](#)] [[PubMed](#)]
35. Landers, G.J.; Blanksby, B.A.; Ackland, T.R.; Monson, R. Swim Positioning and its Influence on Triathlon Outcome. *Int. J. Exerc. Sci.* **2008**, *1*, 96–105. [[CrossRef](#)] [[PubMed](#)]
36. Ofoghi, B.; Zeleznikow, J.; Macmahon, C.; Rehula, J.; Dwyer, D.B. Performance analysis and prediction in triathlon. *J. Sports Sci.* **2016**, *34*, 607–612. [[CrossRef](#)]
37. Piacentini, M.F.; Bianchini, L.A.; Minganti, C.; Sias, M.; Di Castro, A.; Vleck, V. Is the bike segment of modern olympic triathlon more a transition towards running in males than it is in females? *Sports* **2019**, *7*, 76. [[CrossRef](#)] [[PubMed](#)]
38. Horne, M.J. The relationship of race discipline with overall performance in sprint and standard distance triathlon age-group world championships. *Int. J. Sports Sci. Coach.* **2017**, *12*, 814–822. [[CrossRef](#)]
39. Millet, G.P.; Vleck, V.E. Physiological and Biochemical Adaptations to the Cycle to Run Transition in Olympic Triathlon: Review and Practical Recommendations for Training. *Br. J. Sports Med.* **2000**, *34*, 384–390. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.