SPORT'S INFLUENCE ON ATHLETES' PHYSICAL MARKERS

RESEARCH PAPER

Student Number: 2938740

Statistics for Data Science

25th November 2020

ABSTRACT

Biomarkers and biostatistics play an important role in detecting a range of physical and non-physical issues. Female athletes are incredibly susceptible to a condition known as sports anaemia and early detection is key to preventing this condition affecting the athlete. It was found that 30% female athletes in this study falling below the recommended early warning detection level (Ferr < 40ng/ml) with male athletes markedly less affected (11%). A range of physical and non-physical biomarkers were analysed and statistically significant differences were found for the means of all bar one bi-omarker when factored by the sex and sport of the athlete ($p < 0.05$). Such information can be used to optimize training, rest and workloads to specific sex and sport combinations as well as monitoring and taking preventative action before athletes suffer performance and/or health degradations due to overtraining or inappropriate gender training methods.

1. INTRODUCTION

Sport is becoming increasingly data driven at the elite levels as national programs, teams and individuals seek to exploit every ounce of potential and athletes often talk about performance in terms of the one percent that distinguishes the world class athlete from the chasing pack

Gerodimos et al. [1] studied the body composition of the Greek national basketball team and saw significant differences in body fat percentages between player positions as well as competition levels (age-grade through to elite). Data gained from such studies helps coaches and nutritional experts to develop a player to their full potential.

Biomarkers also provide a wealth of data and potential predictors. A Swiss study by Clénin et al. [2] found that iron deficiency occurred in 52% of the studied adolescent, female athletes, compared to 2.2% of the general population from national data. White blood cell count is suspected to be lower among endurance athletes, as found by Horn et al. [3] following a 10-year study. Horn et al. also noted that most of the studies relating to an athlete's haematology are focused on red blood cells as they play a vital role in oxygen delivery. This leaves white blood cell count relatively less understood.

The purpose of this study is two-fold. Firstly, to investigate whether the biomarkers of an athlete are influenced by any of the following factors: Sex, Sport, or a combination of both. Secondly, the data will be analysed to see if it is possible to fit a predictive model to the iron ferritin response variable.

2. DATA

A dataset was provided containing the biomarkers of 202 athletes. The origin of the dataset is unknown nor is the standard of the athletes involved. The 11 biomarkers are continuous variables and fall into two distinct categories. The first deals with the physical makeup of the athlete, namely height, weight, and measurements of body composition (lean mass, bodyfat, BMI, and skin folds). The second category consists of non-physical measurements associated with the area of haematology (white and red cell count, hemoglobin, hematocrit, and iron ferritin levels). It is unknown whether

the blood tests that provided the non-physical data were taken “at rest”, “before”, “during” or “after” exercise. The unit of measurement is also missing from the blood test but can be inferred with sufficient background research (e.g. WCC appears to be “per cubic liter”). The data has two factors defined by the sex of the athlete and their chosen sport. There are no missing values however it is noted that not all sports are represented by both sexes.

2.1.EXPLORATORY DATA ANALYSIS

As seen in figure 1, the dataset contains 102 male and 100 female athletes, the number of athletes is unevenly distributed across the sports and in the cases of gymnastics, netball and water polo it is noticed that the data only contains data for a single sex.

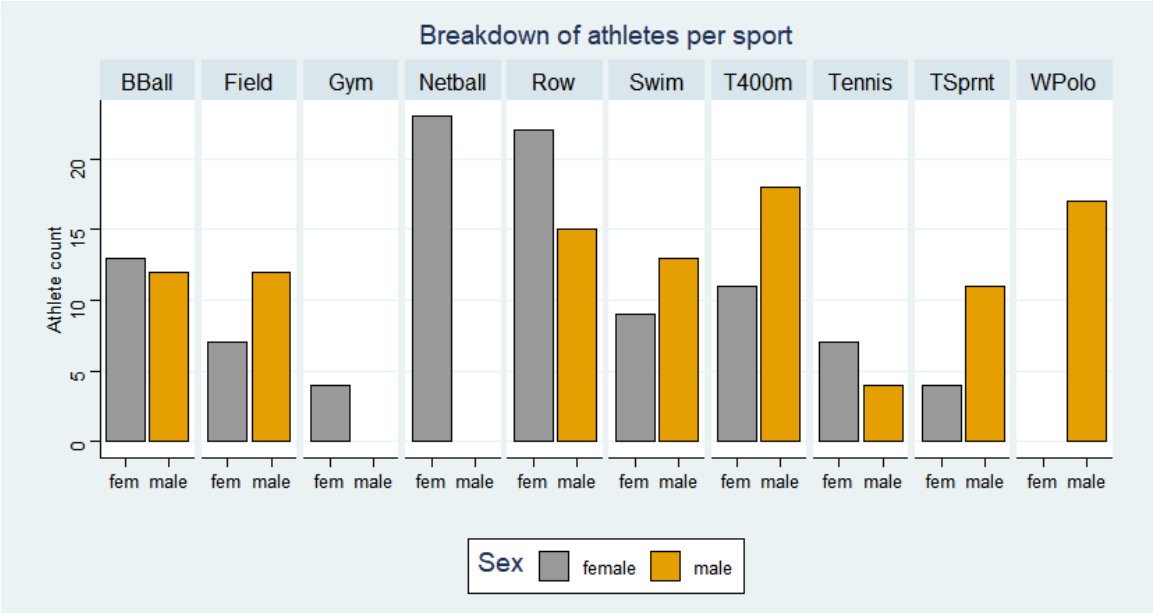

Figure 1 – breakdown of athletes by sex and sport

At all levels of stratification boxplots appear to suggest that the WCC response variable is the only one showing any sign of equality, with clear differences noted across all other response variables. This would match expectation for physical markers given the known differences between men and women. For non-physical markers, there is no observational expectation. There is significant variation between the boxplots for the sex:sport factor combination as seen in figure 2 of the supplementary information document.

A subsequent analysis of outliers (figure 5 supplementary document and appendix A) revealed 37 outlying values when looking at the total dataset, or when stratifying by the sex factor, 23 female and 40 male total outlying values. Some athletes produced multiple outlying values but after adjusting for this effect we see that 17% of our dataset produce outlying values. Athletes, by nature, tend to lie at the extreme end of the genetic spectrum therefore it would be more unusual not to find outliers in the data but care should be taken not to dismiss potential warning signals in the athlete haematology by removing outlying values from the analysis.

3. METHODOLOGY

3.1.HYPOTHESIS TESTING

The first question that this study seeks to answer is whether there are any statistically significant differences between the response variables when stratified by sex and/or sport. The null and alternative hypotheses are given as:

H_0: *There are no differences in the means of the response variables* ***at the given factor***

H_1: *There is (at least) one difference in the means of the response variables* ***at the given factor***

The hypotheses of the response variables will be tested at the 5% significance level using the following methods:

- Difference of means between the sex factor via t-test. The test will be two sample, two-tailed.
- Difference of means between the sports factor via one-way ANOVA.
- Both of the above plus interaction of the sex:sport factors via two-way ANOVA.

  (The data will also be tested for an interaction effect between athlete sex and sport).

  Note: *Where the null hypothesis is rejected during ANOVA testing, the individual results will be subjected to the Bonferroni multiple comparisons test, given as: alphaPC = alphaFW/K, where K is the number of comparisons*

The variance of the response variables is assumed to be constant between the factor levels with hypotheses:

H_0: There are no differences in the variance of the response variables **at the given factor level**

H_1: There is (at least) one difference in the variance of the response variables **at the given factor level**

The distribution of the response variables between the factor levels is assumed to be normal with hypotheses:

H_0: *The response variable is normally distributed* ***at the given factor level***

H_1: *The response variable is not normally distributed* ***at the given factor level***

3.2.PREDICTING IRON FERRITIN LEVELS

Obtaining a predictive model for the athlete's iron ferritin levels from continuous variables will require fitting a multiple regression model to the data. A baseline model featuring all response variables will be created (except iron ferritin) and then refined, if possible, using the drop1 method. The model variables will the lowest F-statistic will be iteratively dropped. The overall model will need to be significant through the p-value at the 5% level. The dataset will be randomly partitioned into an 80% training set and a 20% test set. The formula for the regression model will take the form:

$$Y = \alpha + \boldsymbol{\beta}_1 \boldsymbol{x}_1 + \boldsymbol{\beta}_2 \boldsymbol{x}_2 + \ldots + \boldsymbol{\beta}_i \boldsymbol{x}_i$$

4. RESULTS

4.1.T-TEST BY FACTOR: SEX

The White Cell Count (WCC) was the only response variable that did not produce sufficient evidence to reject the null hypothesis in favour of the alternate hypothesis at the 5% level with P(t(0.897) > (0.05, 198.27df)) = 0.3709. The 95% confidence interval confirms this with the true mean equal to zero lying within the range (-0.272, 0.726). For all other response variables there is sufficient evidence to reject the null hypothesis at the 5% level in favour of the alternate hypothesis. All confidence intervals confirm this. Refer to section 2 of the supplementary document for the full results.

A QQ-Plot showed that whilst some evidence of non-conformity to the normal distribution was found at the extremes of the distribution plots we remember that the Central Limit Theorem for large numbers allows us to accept H_0 with n=202. Welch's t-test does not assume equal variance and was preferred to the student's t-test in this case.

The outlying data values where then removed from the dataset and the tests were re-run to understand if the outliers were having any effect on the data. The results were the same with WCC being the only response variable to hold the null hypothesis and at stronger level, with P(t(-0.129 > (0.05, 170.14df)) = 0.8972. The 95% confidence interval confirms this with the true mean equal to zero lying near the centre of the range (-0.504, 0.442).

The results show that athletic biomarkers not only vary according to the sex of the athlete, but some also show higher levels of variance between the sexes. Furthermore, we note that female athletes have significantly lower iron ferritin levels and lean muscle mass whilst possessing higher levels of body fat (outliers included):

Biomarker	Female average	Male average
Iron Ferritin	56.96	96.40
Weight	67.34	82.52
Lean Body Mass	54.89	74.66
Percentage Bodyfat	17.85	9.25
Sum of Skin Folds	51.42	86.97

Figure 2 – Significantly different means of response variables by sex

4.2.ONE-WAY ANOVA BY FACTOR: SPORT

Welch's one-way ANOVA produced sufficient evidence to reject the null hypothesis in favour of the alternate hypothesis for all response variables at the 5% level. The p-values for every response factor were significant at the 1% level. Refer to section 3 of the supplementary document for the full results. Welch's ANOVA was preferred as Levene's test of homogeneity of variance was significant at the 5% level for all but one response variable finding sufficient evidence to reject H_0. Welch's ANOVA is more robust under such circumstances as it does not assume equal variance.

The assumption of normality was tested with QQ-Plots and H_0 was not rejected. Two response variables were transformed via square root to achieve normality and this transformation was carried forward into the Welch ANOVA. The results show significant differences between sports with up to 20 significant differences in sports combinations noted across all response variables. The Iron Ferritin and White Blood Count biomarkers shows the fewest differences with two and three sport combinations respectively differing from the null hypothesis.

4.3.TWO-WAY ANOVA BY FACTORS: SEX & SPORT

Two-way ANOVA was performed on the dataset after removing outliers as it yielded better results given the reduced sample size per sex:sport factor pairing. The ANOVA test is considered robust and the results verified those results observed in the Welch t-test and one-way ANOVA however caution is advised given some significant results on Levene and Shapiro-Wilk's test of the assumptions. In this case a non-parametric test such as a Kruskal-Wallis would be more appropriate. Only the variable for White Cell Count (WCC) was not significant against the sex factor with F(1, 158) = 0.18, p-value = 0.8931; and in post-hoc testing for main effect F(1, 158) = 0.022, p-value= 0.882; we conclude that there is insufficient evidence to reject the null hypothesis that the means of the WCC response variable vary by sex at the 5% significance level. Figure 17 from the supplementary document finds that the interaction effect for sex:sport was significant ($p < 0.05$) for four of the response variables and affected female athletes more than male athletes.

4.4.MULTIPLE LINEAR REGRESSION MODEL

High degrees of multicollinearity were found in the data particularly between 3 of the 5 non-physical variables (RCC:Hc:Hg; $r > 0.89$), with physical variables SSF:Bfat ($r = 0.97$) and Wt:LBM ($r = 0.92$) also extremely well correlated. No response variables showed any significant degree of correlation with the iron ferritin variable ($r^{max} = 0.3532$). The partitions were checked for athlete sex factor bias (train = 48/52; test 40/60).

A baseline linear regression model explaining 22% of the variance ($R^2 = 0.2233$). Attempts to refine the model failed to yield any improvements with the best resultant model having ($R^2 = 0.1879$), possibly affected by non-normality at the upper tail. The model gives F = 12.19, p-value = 3.238e-7, there we reject the null hypothesis that $\beta = 0$.

The final model is given as: Ferr = 307.5102 – 2.1874*Bfat – 2.0229*Ht + 2.1946*Wt + ε

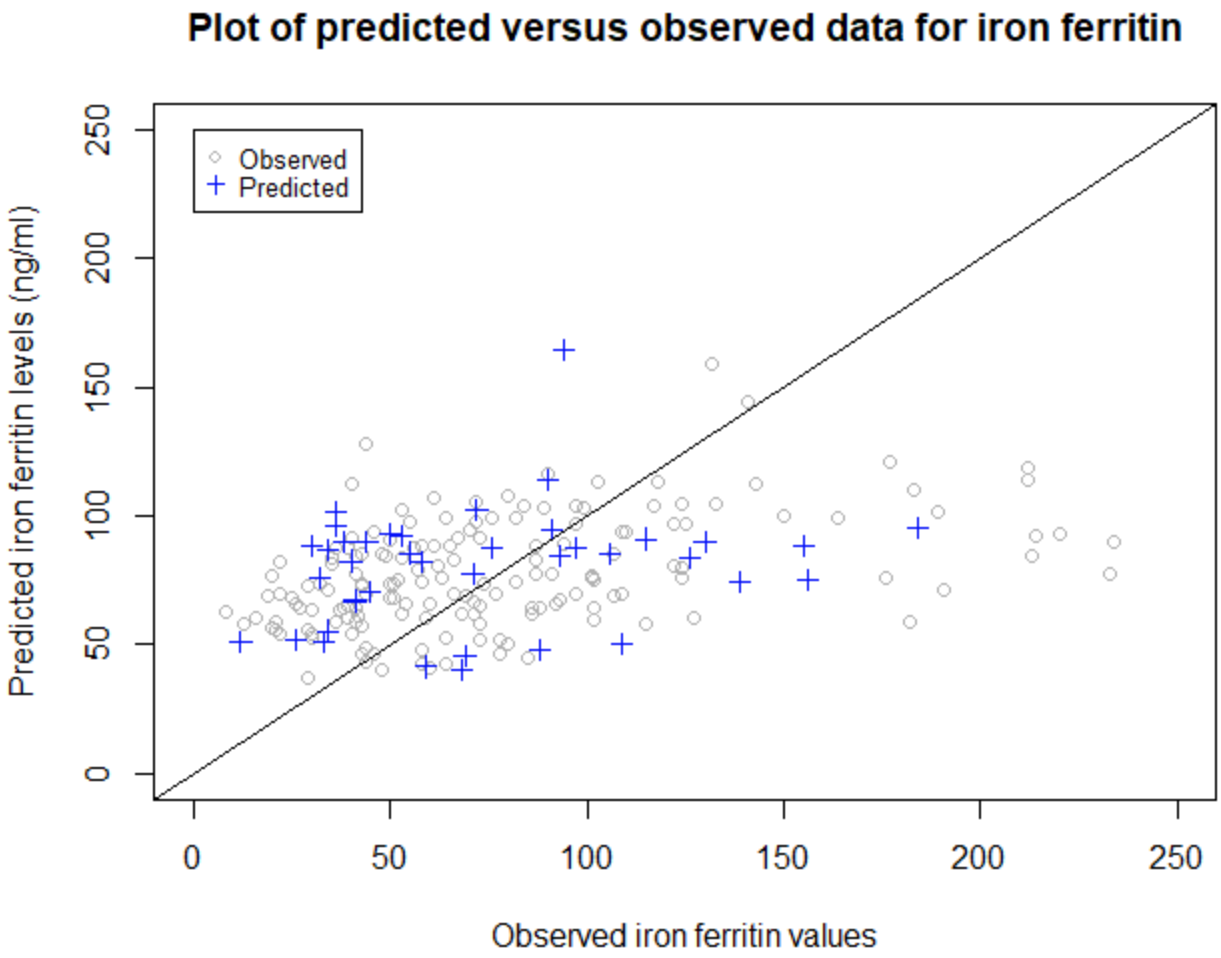

Figure 3 – Overlay of predicted iron ferritin levels

When the predictive values were overlaid with the observed values it provided R^2 = -0.0639 yet the model did appear to produce a good fit given the limited data. More data is required before any meaningful conclusions can be made. Refer to section 5 of the supplementary document for full results and plots. The model over-predicts at < 60ng/ml and under-predict at > 60ng/ml.

The Iron Ferritin variable contains high levels of variance and outlying values therefore fitting an accurate and reliable linear model based on this dataset will be difficult. Attempts to transform the data and remove the skew and centre the model around zero made the model worse, as did introducing a sex factor as per results of the previous tests.

4.5.CONCLUSION / DISCUSSION

The results consistently rejected the null hypothesis across all tests except for the WCC response variable. The two-way ANOVA, whilst less reliable due to discrepancies with assumption testing, arrived at similar conclusions and produced some significant interactions at the sex:sport level. The conclusion is that the sex of the athlete and their chosen sport are significant factors in determining the physical and non-physical attributes of the presented biomarkers.

The results show that men and women are different and that they possess different physical and non-physical attributes for their chosen sports. The QQ Plots for swimming and field also alluded to multiple distributions within the sports which could be explained by the sub-divisions of those sports. Long distance requires endurance (e.g. 1500m swim, heptathletes) whilst short distance and power events require strength and explosive power (100m swim, shot putt). The training regimes that maximise these different characteristics will inevitably lead to different physiologies, a difference no better highlighted than comparing a petite, lean and powerful female gymnast to the archetypical tall, muscular, heavy set shot putter to the muscular yet wiry middle distance endurance runner.

Section 6.1 of the supplementary document appears to provide support for the findings by Horn et al. [3] that athletes training for sports involving aerobic endurance possess lower than average white cell counts. In this study it looks to be true for some of the female athletes (swimming, tennis, and basketball) with 400m showing up for male athletes (sprint endurance training as a possible factor). Iron ferritin is an important marker for detecting sports anaemia and conditions associated with elevated iron ferritin levels, such as Haemochromatosis and cancer amongst others. Two female athletes possess elevated iron ferritin levels greater than 150ng/ml (suggested cut-off for normal females is 120ng/ml according to University of Rochester Medical Centre [4]). No males exceed the upper limit of 250ng/ml but seven record higher than 200ng/ml. However, it is not unusual for elite endurance athletes to have iron ferritin levels higher than 300ng/ml.

At the other end of the scale, sports anemia is a very real problem for athletes with the condition most prevalent in amongst female athletes, characterized by low iron ferritin levels. Whilst 20ng/ml is the cut-off for clinical diagnosis, Team USA (the US Olympic and Paralympic Committee) [5] quote clinical research on their website stating that iron deficiency is better identified in athletes if a cut-off value of 40ng/ml and is more appropriate for endurance athletes. Under these conditions 11 male and 30 female athletes from our study would fall under this cut-off with 1 male and 4 females possessing clinically low levels of iron ferritin leading to sports anaemia.

From the initial boxplots it was clear that one should expect to see differences between male and female athletes, as well as differences between some sports. This expectation was proved in the subsequent analysis. The two-way ANOVA with multiple comparisons highlighted that there were significant results at the sex:sport interaction level as expected. Basic human physiology dictates that certain body shapes are better equipped for the rigours of high-level sport and this extends to the effect that training and competition has on the human body, with some sports being more physically demanding than others. The boxplots also show that the white cell count was most likely to satisfy the null hypothesis

and this proved true at the sex factor level. Whilst physical differences could be expected from common knowledge, the differences in the non-physical, haematology variables were surprising.

The data explored in this study, albeit a limited amount compared to the wealth of opportunity on offer, seems to tally well with the quoted research and provided a fascinating insight into the realm of biostatistics.

The dataset contained a relatively high number of outlying values (16% male, 18% female) which skewed the results somewhat and proved rather difficult to remove in any meaningful way. The QQ Plots and Shapiro-Wilk tests disagreed on five variables in the two-way ANOVA which urges caution however the ANOVA is considered robust. The results were not marginal and proved very significant ($p < 0.01$) on the sex (except WCC) and sport factors. The small sample sizes at the sex:sport level prevented a more interesting interrogation of the data and one area for further investigation would be the sub-divisions within swimming and field sports as mentioned. This would require a larger data sample and in general it would be better to have a larger sample size ($n > 100$) for each sex:sport level in order to perform this type of in-depth analysis as well as some reference data to the general population in order to mitigate against the fact that athletes tend to be extreme outliers in their own right.

5. REFERENCES

[1] Gerodimas et al. Body composition characteristics of elite male basketball players. *Journal of human movement studies* 2005. Link: https://www.researchgate.net/publication/47649268_Body_composition_characteristics_of_elite_male_basketball_players

[2] Clénin et al. Iron deficiency in sports – definition, influence on performance and therapy. *Consensus of statement of the society of Swiss Sports Medicine* October 2015. Link: https://doi.org/10.4414/smw.2015.14196

[3] Horn et al. Lower white blood cell counts in elite athletes training for highly aerobic sports. *Australian Institute of Sport* Jul 2010. Link: https://pubmed.ncbi.nlm.nih.gov/20640439/

[4] University of Rochester Medical Centre. *Health Encyclopedia*. Online Resource
Link: https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=167&ContentID=ferritin_blood

[5] Team USA (US Olympic & Paralympic Committee. Research article. Link: https://www.teamusa.org/USA-Triathlon/News/Blogs/Multisport-Lab/2019/August/27/What-Endurance-Athletes-Should-Know-About-Iron-Deficiency-Anemia-and-Ferritin-Screening

APPENDIX A

The dataset was analysed against a custom R function based around calculating mild and extreme outliers from the inter-quartile range (see section 1.4 supplementary document for more information). The outliers are described as:

Lower Extreme: *Lower Quartile - 3 * Inter-Quartile Range*

Lower Mild: *Lower Quartile - 1.5 * Inter-Quartile Range*

Upper Mild: *Upper Quartile + 1.5 * Inter-Quartile Range*

Upper Extreme: *Upper Quartile + 3 * Inter-Quartile Range*

The data was stratified by Sex before analysis as per known biological differences.

The exact composition of the data is not known; however, it can be inferred from the values that it likely contains data relating to elite or sub-elite athletes.

LE – Lower Extreme | LM – Lower Mild | UM – Upper Mild | UE – Upper Extreme

Response Variable	Number Outliers	Sport & Classification
Ht	6 Female 1 Male	3 gymnasts (LM), 1 rower (LM), 2 basketballers (UM) 1 basketballer (UM)
Wt	2 Female 1 Male	1 gymnast (LM), 1 field athlete (UM) 1 field athlete (UM)
LBM	4 Female 3 Male	3 gymnast (LM), 1 field athlete (UM) 1 rower (LM), 1 basketballer (UM), 1 field athlete (UM)
SSF	1 Female 6 Male	1 netballer (UM) 3 field athletes, 3 water polo players (all UM)
BMI	2 Female 6 Male	2 field athletes (UM) 5 field athletes (4UM, 1UE), 1 water polo player (barely UM)
%Bfat	1 Female 7 Male	1 netballer (UM) 3 field athletes (2UM, 1UE), 4 water polo players (UM)
RCC	4 Female 6 Male	1 track 400m, 1 field athlete, 1 track sprinter, 1 tennis player (all UM) 1 track 400m, 1 basketballer, 1 swimmer (all LM), 2 track printers (UM, UE), 1 field athlete (UM)
WCC	1 Female 3 Male	1 rower (UM) 3 water polo players (UM)
Hc	1 Male	1 sprinter (UE)
Hg	4 Male	1 field athlete, 1 sprinter, 1 tennis player (all UM), 1 track sprinter (UE)
Ferr	2 Female 2 Male	1 field athlete (UM), 1 tennis player (UE) 1 water polo player, 1 tennis player (both UM)

The physical outliers could reasonably be explained by the physical requirements of the given sport through observation of the athlete alone (e.g. shot putters are big and bulky).

The non-physical outliers are known to be affected by athletic training, extreme fitness levels and the timing of the blood test in relation to exercise (an unknown quantity). The identified sports do not appear to be random and should be included in the study to understand these biomarkers in greater detail (i.e. potential warnings).