Noname manuscript No.
(will be inserted by the editor)

Result-based talent identification in road cycling: Discovering the next Eddy Merckx

David Van Bulck · Arthur Vande Weghe · Dries Goossens

Received: date / Accepted: date

Abstract In various sports large amounts of data are nowadays collected and analyzed to help scouts with identifying talented young athletes. In contrast, the literature on result-based talent identification in road cycling is remarkably scarce. The purpose of this paper is to provide insight into the possibilities of the use of publicly available data to discover new talented Under-23 (U23) riders via statistical learning methods (linear regression and random forest techniques). At the same time, we try to find out the main determinants of success for U23 riders in their first years of professional cycling. We collect results for more than 25,000 road cycling races from 2007-2018 and consider more than 2,500 riders from over 80 countries. We use the data from 2007 to 2017 to train and validate our models, and use the data from 2018 to predict how well U23 riders will perform in their first three elite years. Our results reveal that past U23 race results appear to be important predictors of future cycling performance.

Keywords Scouting · Talent identification · Professional Road Cycling · Performance

1 Introduction

One of the main goals of scouting is to identify young prospects, who could make a fine addition to professional teams (Schumaker et al., 2010). In road cycling, while there is little disagreement on Eddy Merckx being the greatest of all time (Cherchye and Vermeulen, 2006), there is much more debate on who will become

D. Van Bulck
Ghent University, Faculty of Economics and Business Administration
E-mail: david.vanbulck@ugent.be

A. Vande Weghe
E-mail: arthur.vandeweghe@gmail.com

D. Goossens ✉
Ghent University, Faculty of Economics and Business Administration
E-mail: dries.goossens@ugent.be

the “next Eddy Merckx”. For instance, at the age of 19, the Belgian rider Remco Evenepoel was already given that label in the media (Farrand, 2018).

Talent discovery in professional road cycling is currently done by national cycling federations, professional teams, and rider agents. Some of the national cycling federations, such as Cycling Australia and British Cycling, have started a talent identification program, which focuses on young cyclists (typically starting at the age of 7), subjecting them to basic physical tests, and providing support with respect to training (Hopker, 2016). The best professional cycling teams have their own scout and development team. While for various sports the ability of scouts to identify talented individuals should not be underestimated, the process is often subjective (e.g. based on intuition or gut feeling) and much of the expert knowledge remains tacit (Cobley et al., 2020, Vaeyens et al., 2009, Williams and Reilly, 2000). In other worldwide sports such as football or basketball, agents or intermediates can help match young athletes with teams and advertise them to team managers, however, Brocard and Larson (2016) state that this practice is anecdotal in cycling.

Effectively and efficiently identifying young talents from a massive pool of potentials is quite a challenge. Note that the International Cycling Union reports a million licensed bike riders, with more than 1,500 of them professional riders[1]. While the academic literature on talent identification has boomed in recent years (for an overview, see e.g. Johnston et al. (2018), Vaeyens et al. (2008)), and the use of advanced data gathering tools and statistical methods to find talented athletes has become the standard in many sports such as soccer (Boon and Sierksma, 2003, Pappalardo et al., 2019), football (Lehman, 2020), basketball (Manisera et al., 2020), or archery (Muazu Musa et al., 2019), the contributions with respect to cycling are rare. In fact, as far as we are aware, there is no quantitative large-scale talent identification system for (road) cycling described in the literature.

The contribution of this paper is to develop a computer-aided system to assist scouts in professional road cycling to make a first selection of talented new riders that - given further training and experience - are likely to become top professional riders. Given that professional cyclists reach peak performance at a relatively late age (29.5 years, Longo et al. (2016)) and that several studies have shown that predictions about future success tend to be more accurate when made closer to the time of peak performance (Mostaert et al., 2021, Vaeyens et al., 2008), we focus on the Under-23 (U23) age category. For more than 2,500 U23 riders from over 80 countries, we analyze their results in more than 25,000 races from 2007 to 2018. We acknowledge that more detailed data on races (e.g. level of competition) and race events (e.g. position data for each rider on short time frames), as well as physiological data on the riders (e.g. weight, maximal oxygen uptake) could be of great value, however, currently no such data is (publicly) available, and certainly not on the scale required for talent identification. Hence, this paper works with variables based on race results and basic personal data (age and nationality) , and quantifies future success of riders in terms of the UCI points (see Section 2) collected in their first three years of elite racing. As riders do not participate in all races, not all races may be equally difficult to win. While alternative rankings exist that explicitly cope with this issue, e.g. Bradley-Terry paired-comparison models

[1] https://www.uci.org/docs/default-source/publications/2019-uci-rapport-annuel-inside-english-web.pdf

(Bradley and Terry, 1952), we believe that the UCI rankings sufficiently capture this aspect as more prestigious races, that typically attract the best riders, receive more points.

The talent identification problem can be approached from two angles: predicting the rider's race results in terms of UCI points, and assessing the probability that the rider will belong to the best professional riders of his age group. The former is tackled by means of a linear regression and a random forest regression model, while we use a random forest probabilistic classification model to achieve the latter. The linear regression also provides insight in the determinants of success for U23 riders in their first years as professionals.

The remainder of this paper is organized as follows. Section 2 provides a gentle introduction into the history and organizational side of professional road cycling. Section 3 then discusses the related literature, and Section 4 presents the data that we use to train our talent identification system. Next, Section 5 proposes the regression and classification based talent identification models and derives the main determinants of talented U23 riders. All models are benchmarked against a naive baseline model that makes use of podium places only. Section 6 assesses the quality of the proposed models and constructs a ranking of riders that are predicted to break through as professional riders. Finally, conclusions follow in Section 7.

2 Background in cycling

In this section, we briefly sketch the context of cycling, as well as the UCI points ranking. For a more in-depth introduction and background in professional road cycling, we refer to Mignot (2016) and Rebeggiani (2016). Details on the UCI points ranking can be found in the UCI Cycling Regulations[2].

Professional road cycling has a rich history with the first races dating back to the late nineteenth century. Throughout the twentieth century, racing became more popular in European countries such as France, Belgium, Italy, and Spain. As a reaction to the internationalization of the sport, a number of national cycling federations founded the 'Union Cycliste International (UCI)'. To date, the UCI is the world governing body of cycling and is responsible for, inter alia, organizing the cycling calendar, issuing racing licenses, and enforcing race rules and anti-doping regulations.

There are two kinds of road races: one-day races, which are settled in a single event, and stage races, where a general classification (GC) is made based on each riders' accumulated race time over all stages (although individual stage victories are also considered valuable wins). The most prestigious one-day races are known as classics or monuments, such as Tour of Flanders and Paris-Roubaix. Although a typical stage race takes one week, the so-called grand tours last for three weeks and are organized in France (Tour de France), Italy (Giro d'Italia), and Spain (Vuelta a España). The UCI categorizes each road race depending on its relevance and difficulty, with World Tour races being the highest level races. Besides, there are two race categories that are restricted to young non-professional riders only:

[2] www.uci.org/docs/default-source/rules-and-regulations/part-ii-road/2-roa-20210101-e.pdf

Race / Rank	1	2	3	4	5	6	7	8	9	10
Tour de France GC	1,000	800	675	575	475	400	325	275	225	175
Tour de France Stage	120	50	25	15	5	0	0	0	0	0
Monument	500	400	325	275	225	175	150	125	100	85
U23 World Championship	200	150	125	100	85	70	60	50	40	35

Table 1: Illustration of UCI points awarded in 2020 for top ten results for various race types. Note that in many races, UCI points are also awarded for results outside the top ten.

U23 (for riders under the age of 23), and juniors (for riders under the age of 19). Note, however, that since 2016, any rider under the age of 23 can race the U23 World Championship, including professionals.

The UCI distinguishes three tiers of teams, and this classification largely determines access to the races. The World Teams are the highest category in professional road cycling: they are subject to strict regulations and can participate in the most important races. The second and third tier of professional cycling are formed by the Pro Teams and Continental Teams. As an exception, the World Championships are run amongst national teams.

While modern day road cycling is a team sport in the sense that riders (so-called domestiques) usually work together with their teammates, sacrificing their own chances to accomplish success for their team leader, race results rank riders individually (team time trials excepted). The UCI manages the UCI World Ranking, where in each UCI race riders can collect UCI points according to their race rank. A higher rated race will result in the successful riders receiving more points (and more riders being awarded UCI points), as illustrated in Table 1. The UCI World Ranking accumulates the points collected over the last year on a rolling basis (i.e. each race result replaces last year's result), updated every week. Note that this ranking includes all male riders from the World Tour down to U23 riders, and is not to be confused with the UCI World Tour ranking (discontinued in 2018), which included only riders from World Teams, and their results in World Tour races.

Riders and teams highly value UCI points. This is not so much due to the trophy awarded to the end-of-the year UCI World Ranking leader, as to the fact that UCI points determine the order of the team cars, the number of riders that each nation can delegate to the World Championships, and which Pro Teams can participate in World Tour races. Consequently, on an individual level, riders can use their UCI points to negotiate a higher wage or a team transfer.

3 Related Literature

As discussed in Section 1, the literature on talent scouting in cycling is limited. However, several important contributions have been made in four related research fields.

A first and prominent part of the literature focuses on the various physiological (e.g. maximal oxygen uptake), psychological (e.g. behavioural dynamics), and physical factors (e.g. air and rolling resistance) that influence cycling power output, velocity, and winning probability. For a general overview of these factors we refer to Atkinson et al. (2003), Faria et al. (2005a,b) and Phillips and Hopkins (2020), and we refer to Lucia et al. (2003) for an overview of the physiological aspects in the Tour de France. For scouts, these studies could be of interest since they could help to identify the relevant physiological measures that are required to excel at a later age (Menaspà et al., 2010, Svendsen et al., 2018), or to orient cyclists towards their best discipline (see Mostaert et al. (2020)).

Another strand of literature focuses on predicting race outcomes. Olds et al. (1995) are one of the first to use simulation tools to predict cycling times in a road time trial based on physiological, biophysical, and environmental variables. Olds (1998) later extends this model to predict the winning chances of a breakaway group. Rodríguez-Gutiérrez (2014) shows that the leaders of elite cycling teams achieve better performance not only because they have greater abilities but also because they get support and help from their domestiques. The author defines rider performance as the total number of points earned adjusted by the total number of kilometers ridden in a season and performs among others a two-stage least squares analysis, instrumenting for physiological features and rider and team quality. Kholkine et al. (2020) focus on predicting the winners of the 2018 and 2019 editions of the Tour of Flanders. They consider features based on past rider performance in similar races (time difference relative to the winner), long-term rider's profile (e.g. total career points), and the quality of a rider's team (total points collected by the team in the previous year).

A third and related strand of literature is to rank riders or teams at the end of a multi-stage race or season. Instead of simply looking at the total number of victories or UCI points collected, Cherchye and Vermeulen (2006) and Rogge et al. (2013) take into account the fact that professional cyclists pursue multiple objectives that cannot be traded off easily (e.g. stage wins versus second places in GC's). In particular, Cherchye and Vermeulen (2006) propose a robust ranking method using only ordinal information regarding the importance of the different objectives and come up with an all-time ranking of Tour de France participants listing Eddy Merckx, Bernard Hinault, and Lance Armstrong on top. Furthermore, a rider's individual performance is not determined by his individual characteristics only but also by team characteristics, since team leaders benefit from their domestiques' help. Analyzing results from the Tour de France (2002-2005), Prinz and Wicker (2012) come to the conclusion that team managers should pay attention to the composition of the team: having only one strong team captain and several good domestiques turned out more effective than having several star riders (i.e. potential captains) in a team. Rogge et al. (2013) evaluate the performance of cycling teams in the Tour de France (2007-2011) using data envelopment analysis. They conclude that teams that focus on the general classification are often more efficient than teams that focus on sprint stage wins, or on the hilly (transition) stages. Also interesting is the study by Hsia (2017) who uses pairwise comparisons of past race results in the UCI Mountain Bike World Cup to rank riders and predict future race outcomes. Compared to the traditional UCI point-based ranking, their method may better reflect ability of riders as competitors may enter a different number of races and the level of competition may vary between races.

A fourth and final part of the literature uses junior race results to infer the main determinants of elite cycling success. Schumacher et al. (2006) investigate for over 8,000 riders and 100 nationalities whether riders that achieved success in the junior world championships are on average more likely to achieve top 10 places in any of the elite word championships or grand tours. For several track cycling disciplines, their results confirm this hypothesis, however, for road cycling they did not find a significant trend between junior and elite success. Svendsen et al. (2018) investigate for 80 Norwegian cyclists whether there are any statistical differences on the level of training, performance, or physiological data between juniors that became World Tour riders in their first year of elite rider and juniors that did not. Their main findings suggest that junior riders who reached the world tour level scored significantly better in the junior national championship and have a higher maximal aerobic power. Similarly, Mostaert et al. (2021) investigate for over 300 Belgian cyclists whether there is a link between U15, U17, and junior race results and the chance to later become a member of a professional elite racing team. Their findings suggest that every top 10 result in one of 6 considered U17 or junior races increases this chance with respectively 3-5% and 6%. Interestingly, no significant relation was found for the U15 category.

Although the studies in the previous paragraph look at between-group differences and can therefore be used to determine which characteristics may be relevant for talent identification, none of them can be used directly to identify talent on an individual basis. In fact, apart from a master's thesis by Maton (2020), we are not aware of any quantitative large-scale talent identification system for (road) cycling described in the literature. Maton (2020) uses junior and U23 race results to predict the total number of PCS points scored in the first two elite years. As opposed to our approach, the author chooses not to aggregate race results and instead introduces for each race separately a variable that gives the best ever result of a rider in that race. As a consequence, dedicated and computationally rather expensive value imputation techniques are needed to deal with missing values since junior and U23 riders usually participate in only a selection of all races.

4 Data and variables

Our study makes use of a dataset of race results between January 2007 and December 2018, which was obtained through the courtesy of CQ Ranking[3], and appended with data publicly available from ProCyclingStats. We chose 2007 as a starting point because we could find data only for the top-tier races prior to 2007, while young riders typically have less competitive races on their schedule in their first years after they leave the U23 category. Furthermore, cycling was plagued with doping cases prior to 2007 (Wagner, 2010). Recall that the Tour de France officially has no winner in the years 1999-2005, which were dominated by Lance Armstrong, and also its 2006 edition saw several riders denied the right to participate due to Operation Puerto and its apparent winner, Floyd Landis, disqualified. We believe this may have had an effect on the performance of young riders entering the professional racing circuit prior to 2007 (see also Lentillon-Kaestner and Carstairs (2010)).

From the dataset, we were able to derive the variables given in Table 2 for riders who took part in at least one U23 race in that period. Note that when a rider appears in a U23 race in a certain year, he is considered a U23 for the entirety

[3] `www.cqranking.com`

of that year. The dataset is split into two parts. The first parts consists of 2,308 riders that participated in U23 races between January 2007 and December 2017. This data will be used to train and test our models (see Section 5). The second part consists of 270 riders that participated in U23 races between January 2018 and December 2018, but not later (the 2019 U23 Worlds Championship excepted, since that race is open to professional riders as well). In other words, for these riders, 2018 was their last season as U23 and hence it remains to be seen how these riders will perform in the future. Some promising riders together with their estimated performance are listed in Section 6.

We measure the performance of riders in their first three years after their U23 career (possibly as a professional rider) by the average yearly number of UCI points collected in this period (NeoProfUCI). We acknowledge that alternative measures (e.g. CQ points or PCS points) could be considered as well, but given the importance and official status of UCI points (see Section 2) and the fact that they can be collected in all UCI races by a fairly large proportion of the riders that finish the race, we think this is a reasonable choice. Note that 20% (19%) of the riders only had one (two) year(s) in which they scored UCI points, due to the fact that they left the U23 category after 2016; in these cases, we take the average over the years in which they scored UCI points.

Table 2 also lists the independent variables that we consider in this study. All result-based variables are averaged over the number of years the rider was active in the U23 category. In case of the collected UCI points (U23UCI), we use a weighted average, where more recent years receive a higher importance. We also consider the number of victories (U23top1) and other podium places (U23top3), as winning (and to a lesser extent obtaining a podium spot) is paramount in cycling. In order to take into account which riders came close to winning, or in cycling terms “rode the finale”, we count the rider’s number of top 20 results. Although the cut-off point is somewhat arbitrary, we believe that using top 20 spots is more meaningful than e.g. looking at time gaps or speed differences, because these depend heavily on the race circuit, the type of race (e.g. in a stage race, limiting the time gap is of the utmost importance for some riders, whereas in a one-day race it is not a goal on itself), and the weather conditions. Hence, they don’t necessarily reflect the strength difference of the riders. Based on labels given by CQ ranking, we track top 20 results separately for various race types: sprints (17%), mountain stages (9%), time trials (15%), and general classifications (10%); the other races we label as hills (49%). We also consider the age at which the rider rode his first U23 race (Age_Started), and the number of years the rider has been active in U23 races (U23Years). As Mostaert et al. (2021) have shown that there is no relative age effect in cycling categories above 15 years old, we did however not control for the month in which an athlete was born. Finally, we include the rider’s continent as a control variable, in which we make a distinction between the best five European countries according to the number of UCI points collected in the period 2007-2017 (i.e. Spain, Italy, Belgium, France, Netherlands) and the other European countries, by considering them as separate continents. Indeed, we believe that the rich tradition of cycling and the highly developed training infrastructure in these five countries may give their youth riders an advantage.

We opted not to include variables for the U23 team(s) for which the rider has been riding. This would create too many variables to get meaningful results, and issues with riders who don’t have a team (or race in mixed teams for a considerable

Variable name	Description
NeoProfUCI	Average yearly number of UCI Points during first 3 years as a pro (dependent variable)
U23UCI	Weighted average of UCI Points during U23 career, more recent years get higher weights
U23top1	Average number of victories during U23 career
U23top3	Average number of top 3 finishes (besides victories) during U23 career
U23top20_Sprints	Average number of top 20 finishes in a sprint stage during U23 carreer
U23top20_Mountains	Average number of top 20 finishes in a mountain stage during U23 career
U23top20_TTs	Average number of top 20 finishes in a time trial during U23 career
U23top20_Hills	Average number of top 20 finishes in a hilly race during U23 career
U23top20_GCs	Average number of top 20 finishes in a general classification during U23 career
Age_Started	Age at which the rider first appears in the database
U23Years	The number of years the rider raced in U23 races
Continent	Continent of the rider (categorical: Top5Europe, Europe, Africa, Asia, Oceania, America)

Table 2: Variable description.

number of races). Furthermore, we don't think that the impact of the team in U23 cycling is as pronounced as in professional road cycling (Cabaud et al., 2016), since most U23 riders have a "free role" in their team rather than a task as domestique, which makes sense as they all want to shine in order to be contracted by a top-tier team.

5 Models

This section proposes various statistical models to predict and understand the performance of riders in the first three years after their U23 career. To this end, we make use of the first part of the dataset (period 2007-2017) which we further split into a training set (80%) used to train the models and a test set (20%) used to validate the models. All models were coded within the statistical software package R and were tuned for best performance using grid search in combination with a 10-times repeated 10-fold cross validation resulting in a total of 100 folds.

5.1 Linear Regression

Our first model is a linear regression model where we estimate the regression coefficients via ordinary least squares and a backward variable selection strategy. This means that we initially include all independent variables in the model and then repeatedly remove the least significant variable (but not the intercept) and refit the model until all the remaining variables have a p-value lower than a predefined

Variable name	Estimate	Std. Error	p-value	VIF
(Intercept)	-22.60	5.87	< 0.001	
U23UCI	0.98	0.12	< 0.001	3.22
U23top3	21.40	3.53	< 0.001	2.47
U23top1	30.57	4.82	< 0.001	2.03
U23top20_Hills	4.72	1.54	0.002	2.27
U23top20_Mountains	8.92	2.78	0.001	1.34
U23Years	6.26	2.02	0.002	1.16
Africa	-36.21	13.11	0.006	1.05
America	36.08	8.17	< 0.001	1.12
Top5Europe	21.09	4.64	< 0.001	1.20

Table 3: Regression coefficients estimated via ordinary least squares.

significance level (0.05 in our case). Table 3 presents the estimated values of the regression coefficients. Each regression coefficient represents the marginal and *ceteris paribus* effect of an independent variable, i.e. the change in the dependent variable when the corresponding independent variable increases by one unit while all other independent variables are held constant. Because of the ceteris paribus condition, the absence of multicollinearity in a regression model is important and can be measured by the Variance Inflation factor (VIF). In order for a model to have a meaningful interpretation the VIF of each variable should be at most 5 (see e.g. James et al. (2013)), which is clearly the case as shown in the last column of Table 3.

When analyzing the regression coefficients in Table 3, it is apparent that the number of UCI points collected by U23 riders serves as an important predictor for their future career: the average number of UCI points collected in the first three elite years is expected to increase with 0.98 points for every point the rider scores on average during his U23 career. Regarding the U23top20 variables, we observe that an additional top 20 result in a mountain stage is almost worth double compared to an additional top 20 result in a hill stage. One likely explanation is that there are considerably more hill races than mountain races (see Section 4). Hence, an additional top 20 result in one of the many hill races is worth less than a good ranking in one of the few mountain races. Because time differences are typically largest in mountain races, another explanation could be that climbing skills are good predictors for future top results in grand tours; note that the U23top20_GC variable was deleted during backward variable selection.

The positive regression coefficient of Top5Europe supports the hypothesis that the rich tradition of cycling and highly developed training infrastructure in the top five European countries may give their youth riders an advantage over Asian, Oceanian, and other European countries (these continent variables were removed from the model in the backward variable selection step and hence serve as the base category). The negative regression coefficient for Africa might suggest that it is more difficult for African U23 riders to enter the mainly European elite road racing circuit. For instance, no African-registered team participated in a Grand Tour until 2015, and at the time of writing, there is still only one African-registered team (Team Qhubeka - Assos) among the World and Pro Teams. The opposite is true for riders from America, which has several US-registered World and Pro teams, and an established cycling tradition in Colombia, with successful riders like Rigoberto Uran and Nairo Quintana as inspiring stars. Although in line with findings by Van Reeth (2016) on the internationalization of the peloton, the regression coefficients for Africa and America should be interpreted with caution since there are only 57 African and 155 American U23 riders in our database.

Finally, the positive coefficient of the U23Years variable suggests that riders are more likely to be successful in the first three years of their elite career if they have more experience in the U23 category (e.g. because they started road cycling at the U23 level at a young age).

5.2 Random Forest Regression

One of the main limitations of linear regression is that it assumes a linear relationship between the dependent and independent variables, and that it is relatively sensitive towards outliers. One approach that copes well with the limitations and pitfalls from linear regression is the random forest approach. As its name suggests, a random forest is a combination of a number of decorrelated decision trees (typically 500 or 1000), where every tree is created by randomly selecting a subset of the independent variables at each split and using only a subset of the observations. The main advantages of random forest algorithms are that they do not assume any data distribution (unlike linear regression the method is non-parametric), the method is fairly intuitive and flexible, and typically only few parameters need to be tuned (e.g. number of trees, number of candidate independent variables at each split, and number of observations used in each tree).

Although it is not trivial to understand how predictions are made by a random forest regression model, some insights can be derived by conducting a variable importance analysis. For each independent variable, the second column of Table 4 shows the estimated percentage increase in Mean Squared Error (MSE) when this variable would be omitted from the model. It is clear that the U23UCI variable is the most useful variable in predicting elite road cycling success. Nevertheless, the high values for several other variables show that it is wise for scouts to take into account several of the other independent variables as well. The high value for the time trial variable may be explained by the fact that time trial skills may help to end up high in the final rankings of a grand tour. Also interesting to see is that the combined number of second and third positions provides more information to the model than the total number of victories. Finally, in contrast to the Top5Europe variable, Table 4 hints that the categorical variables Oceania, Europe, and America are not so reliable to base predictions on.

5.3 Random Forest Classification

Instead of predicting UCI points on a continuous scale, we can also address our talent identification problem as a classification problem by labeling a rider as talented if his performance in terms of the dependent variable NeoProfUCI is higher than a predefined value. When determining this value, there is a trade-off between on the one hand labeling too many riders as talented and on the other hand not having enough samples to learn from. In our model, we label a rider

	% Increase MSE	Mean Decrease Accuracy
U23UCI	16.56	50.00
U23top1	7.77	24.92
U23top3	13.87	27.12
U23top20_Sprints	3.98	15.07
U23top20_Mountains	6.02	25.17
U23top20_TTs	11.64	20.04
U23top20_Hills	11.89	33.03
U23top20_GCs	8.79	22.68
Age_Started	4.19	10.41
U23Years	7.24	16.15
Top5Europe	13.66	11.69
Europe	0.31	8.68
Africa	3.46	-0.60
Asia	5.13	6.28
Oceania	-2.20	-0.87
America	-0.71	1.90

Table 4: Variable importance in the random forest regression (% increase in Mean Squared Error) and classification models (Mean Decrease Accuracy).

as talented if the NeoProfUCI variable is within the top-20%; this corresponds to labeling a rider as talented if he scored more than 75 UCI points in the first three years after leaving the U23 category. Once the classification labels are created, we construct a random forest model similar to the one discussed in Section 5.2, but with the main difference that we predict the probability that a rider is talented instead of the dependent variable NeoprofUCI. Moreover, since there are four times more regular riders than talented riders we correct for class imbalance by stratified sampling and balanced class weights.

For each independent variable, the third column of Table 4 shows the estimated decrease in the mean accuracy of a tree, i.e. the percentage of predictions correctly made for the stratified sample considered in the tree, when the variable is left out the model. Conclusions similar to the ones of the previous section can be derived.

6 Results

This section experimentally evaluates the performance of the models. First Section 6.1 evaluates the quality of the models for the period 2007-2017. Section 6.2 then uses the models to predict the first three years of professional performance of riders that participated in their last U23 race in 2018. We compare these predictions against the available race results for the period 2019-2021.

6.1 Model validation

In order to evaluate how well the two regression models fit the data, we consider three different goodness-of-fit measures. The R^2 indicates to what extent the created model explains the variance of the dependent variable, while the Mean

	R^2	MAE	RMSE	Log Loss	AUC
Naive Reg.					
CV Training	0.336 ± 0.105	49.91 ± 5.38	93.33 ± 19.20		
Test	0.318	55.67	116.90		
Linear Reg.					
CV Training	0.396 ± 0.076	46.85 ± 4.91	89.35 ± 20.15		
Test	0.387	54.13	110.21		
RF Reg.					
CV Training	0.397 ± 0.077	44.51 ± 5.09	89.06 ± 20.89		
Test	0.378	52.19	111.63		
Naive Class.					
CV Training				0.524 ± 0.045	0.819 ± 0.037
Test				0.514	0.827
RF Class.					
CV Training				0.324 ± 0.043	0.883 ± 0.032
Test				0.354	0.855

Table 5: Prediction and classification measures. For the cross-validated training set, the mean ± standard error is calculated over all folds.

Absolute Error (MAE) and Root Mean Square Error (RMSE) directly measure the error of the predictions made. Since in RMSE the prediction errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. Table 5 provides the results for the cross-validated training set (CV) and test set. We observe no notable performance differences in terms of R^2 between the linear regression model and the random forest regression model. Compared to a naive regression model ('Naive Reg.') which uses as independent variables only the podium place variables U23top1 and U23top3, the R^2 values are however substantially better. Table 5 shows that the linear regression and random forest models make on average a prediction error of around 53 UCI points in the test set, and that these errors are in general less extreme than these of the naive model (lower RMSE). While the differences in MAE and RMSE may look small, they can make a difference in practice as these values are averaged over all riders. Indeed, the majority of the U23 riders in our dataset score only very few UCI points in their first years as a professional, and hence their typically small absolute prediction errors average out larger differences in predictions that are present for the most talented riders that score a decent amount of UCI points.

In order to assess the performance of the random forest classification model, we consider a naive logistic regression model ('Naive Class.') that again uses as independent variables only the podium place variables. Figure 1 shows the resulting calibration plot: a model is considered well calibrated if for any level the predicted probability of a rider to be talented corresponds more or less with the actual probability that the rider is talented. The calibration plot hints that the models are a bit too optimistic with regard to the future success of riders, especially for the higher end predictions. Then again, we do not directly interpret the predicted probabilities. Moreover, for a talent identification system it seems better to be (overly) optimistic rather than pessimistic as scouts may recognize the false

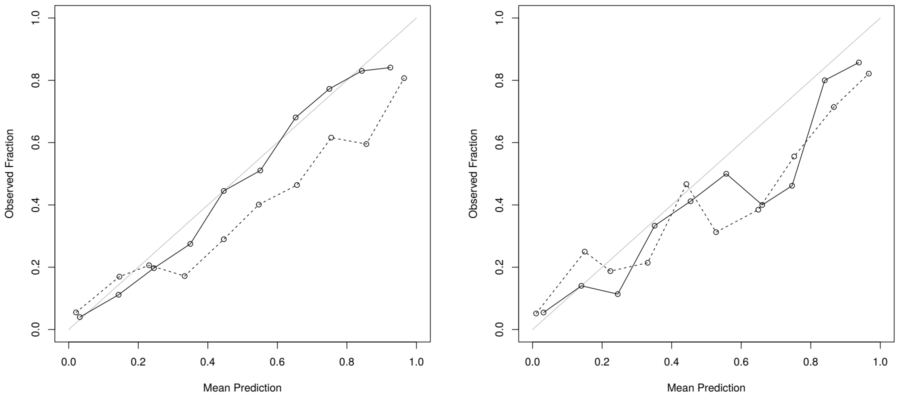

Fig. 1: Calibration plots for the random forest classification model (solid line) and the naive logistic regression model (dashed line) on the cross-validated training set (left) and the test set (right). Riders are first grouped into 10 bins based on the probability forecast. The horizontal axis then shows the mean probability forecast in each bin, while the vertical axis shows the actual relative frequency of talented riders in each bin. The diagonal line presents a perfectly calibrated model.

positives relatively simply whereas false negatives may correspond to talented riders that remain undetected. When inspecting the log loss measure that penalizes a prediction more heavily when the predicted class probability diverges further from the actual label (see Table 5), we see that the random forest classification model achieves a substantially better log loss than the naive logistic regression model.

Instead of looking at the probability that a rider is talented, we may also label a rider as ‘talented’ if the predicted probability is above a predefined threshold value, and as ‘regular’ otherwise. There is, however, an important trade-off to be made when choosing this value: a higher threshold results in better specificity (i.e. the percentage of regular riders identified as such) but this comes at a cost in terms of sensitivity (i.e. the percentage of talented riders identified as such). It is therefore interesting to inspect the Receiver Operating Curve (ROC) which plots the sensitivity against the specificity for different threshold values of the model (see Figure 2). A model that always predicts the correct label is situated in the top left corner, whereas a purely random model or a model that always predicts the same class has a ROC corresponding to the diagonal line. For reasons outlined in the previous paragraph, scouts may particularly be interested in models with a high sensitivity and a reasonable specificity. The area under the ROC (AUC) therefore is a measure for the performance of a model: the naive model has an AUC of 0.827 on the test set whereas the random forest classification model has an AUC of 0.855 on the test set (see Table 5). The random forest classification model is thus clearly the better of the two models.

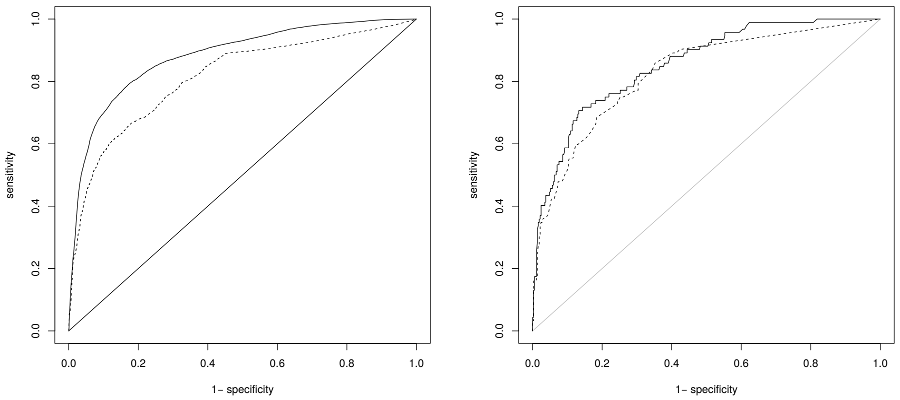

Fig. 2: Receiver Operating Curve (ROC) for the random forest classification model (solid line) and the naive logistic regrssion model (dashed line) on the cross-validated training set (left) and the test set (right).

6.2 Predicting talented riders

Table 6 shows the most promising talents among the 270 riders that had 2018 as the last year in which they raced in the U23 category. The first two columns give the actual rank of the riders based on the average UCI points collected in 2019 and 2020 together with their name and their age in 2019. The third to fifth column provide the UCI points collected and indicate the type of team that contracted the rider in the given year (one star refers to a World Team, two stars to a Pro Team, and three stars to a Continental Team). At the time of writing, the 2021 season has not yet ended and the fifth column therefore includes only the UCI points collected between the first of January and the first of July. For each of the models, the table also gives the top-20 predictions of the yearly UCI points collected in the first three years after the U23 career (for the linear and random forest regression models), and the probability that the rider will belong to the best 20% neo-professionals (for the logistic regression and random forest classification models).

It is fair to say that, based on the available results from 2019 to 2021, our models did not overlook the top talents. Indeed, all ten best performing young professionals in our dataset were predicted as a top-30 rider (i.e., top 11%) by at least one of our models. Tadej Pogačar, who was ranked second and third by the non-naive regression models, largely surpassed the predicted UCI points, among others becoming the youngest rider to win the Tour de France since World War II. Also Marc Hirshi (Flèche Wallone win, Tour de France stage win, bronze at the Worlds), Jay Hindley (stage win and second place overall in Giro d'Italia), Neilson Powless (San Sebastián Classic win), Michael Storer (stage win in Vuelta a España, stage win and first place overall in Tour de l'Ain), and Jasper Philipsen (multiple stage wins in Vuelta a España, and six stage podium places in the Tour de France) managed to taste victory in top-tier races in their first years as a professional rider.

		UCI Points			Linear Reg.		RF. Reg.		RF. Class.		Naive Reg.		Naive Class.	
Rank	Name (Age)	2019	2020	2021	Rank	UCI	Rank	UCI	Rank	UCI_20	Rank	UCI	Rank	UCI_20
1	Pogačar, T. (21)	2065*	3055*	2313*	3	560	2	472	14	0.788	4	367	4	0.999
2	Hirschi, M. (21)	796*	1430*	271*	8	361	11	264	20	0.769	13	241	13	0.983
3	Vlasov, A. (23)	671**	1399*	1180*	29	164	31	168	36	0.663	49	97	49	0.682
4	Hindley, J. (23)	438*	1155*	30*	9	334	13	255	19	0.769	14	204	14	0.963
5	Philipsen, J. (21)	906*	602*	672**	1	593	1	498	4	0.816	1	524	1	1.000
6	Madouas, V. (23)	642*	426*	155*	11	314	23	200	29	0.698	102	53	102	0.432
7	Dunbar, E. (23)	516*	241*	121*	16	261	18	230	18	0.77	36	119	35	0.783
8	Kämna, L. (23)	80*	595*	58*	31	161	22	200	25	0.725	57	83	57	0.611
9	Lambrecht, B. (22)	649*			2	571	4	418	16	0.774	3	382	3	0.999
10	Mcnulty, B. (21)	325**	294*	112*	7	376	8	324	11	0.798	12	245	12	0.986
11	Dainese, A. (21)	445***	104*	10*	37	141	39	148	48	0.607	34	120	36	0.775
12	Geniets, K. (22)	279*	205*	280*	137	36	138	32	138	0.126	113	43	113	0.385
13	Paret-Peintre, A. (23)	214*	256*	532*	49	114	52	122	49	0.581	110	47	106	0.406
14	Taminiaux, L. (23)	423**	28**	5**	74	84	78	73	90	0.282	103	50	103	0.419
15	Powless, N. (23)	301*	146*	180*	4	465	5	405	9	0.798	9	301	9	0.996
16	Moschetti, M. (23)	169*	255*	157*	12	308	7	335	13	0.789	5	353	5	0.999
17	Vingegaard, R. J. (23)	306*	116*	606*	38	141	45	137	33	0.673	33	120	33	0.791
18	Affini, E. (23)	316*	82*	150*	63	88	72	82	85	0.308	54	86	54	0.623
19	Malecki, K. (23)	229***	158*	158*	26	180	29	178	3	0.824	26	154	26	0.89
20	Gidich, Y. (23)	363*	8*	90*	5	440	6	342	7	0.807	11	254	11	0.988
25	Mäder, G. (22)	101*	354*	356*	17	252	10	274	8	0.803	16	195	16	0.953
26	Stannard, R. (21)	72*	254*	160*	6	419	3	435	2	0.831	2	422	2	1.000
30	Pronskiy, V. (21)	250***	40*	98*	24	192	21	212	53	0.554	20	164	20	0.907
32	Eenkhoorn, P. (22)	126*	157*	23*	19	243	14	250	1	0.845	21	162	21	0.904
35	Padun, M. (23)	233*	40*	130*	18	249	19	227	15	0.776	17	191	18	0.948
37	Dewulf, S. (22)	110*	156*	29*	15	263	15	244	10	0.798	10	278	10	0.993
45	De Decker, A. (23)	118**	86**	6***	39	140	27	189	17	0.773	28	143	28	0.861
48	Storer, M. (22)	91*	103*	32*	14	296	17	231	27	0.706	22	158	22	0.897
54	Areruya, J. (23)	78**	93**	80**	10	318	9	287	12	0.794	8	301	8	0.996
56	Johansen, J. (20)	59***	108**	0**	23	200	30	176	34	0.669	19	170	19	0.916
59	Peak, B. (21)	62***	90*	45*	32	154	33	158	22	0.75	15	201	15	0.961
79	Vanhoucke, H. (22)	5*	110*	126*	20	211	20	214	30	0.697	18	191	17	0.949
100	Kanter, M. (22)	2*	65*	0*	13	297	16	241	5	0.815	6	318	6	0.997
112	Cullaigh, G. (23)	54***	0***	115*	42	129	41	143	6	0.81	30	136	30	0.839
178	Galdoune, A. A. (23)	4	0	4	22	202	12	255	21	0.753	7	316	7	0.997

Table 6: Talent predictions as made by the different models for the top-20 riders ranked according to the average UCI points in 2019 and 2020, with the age of the rider in 2019 between parentheses. Note that the 2021 UCI points only include race results before July 1st. The number of stars refers to the type of team a rider is under contract with in a specific year: one star refers to a World Team, two stars to a Pro Team, and three stars to a Continental Team.

Bjorg Lambrecht would probably also have turned out a rider of that caliber, if he had not died after crashing into a concrete culvert in the 2019 Tour de Pologne. Lennard Kämna and in particular Alexandr Vlasov and Jonas Vingegaard (second place overall and three podium places in the Tour de France) performed very well in their neoprof years, but were not particularly ranked high by any model. The naive models, on the other hand, missed 4 out of 10 top talents when looking at their top 30. Furthermore, all but three of the top-20 riders predicted by any of our (non-naive) models were hired by a World Team. One of these three, Jasper Philipsen, was contracted by the prestigious Pro Team Alpecin-Fenix, by many considered to be a better team than several World Tour Teams.

Finally, Table 6 also includes a number of riders who did not (yet) live up to the expectations of the models, although in most cases, these are riders that were ranked highly by only one of the three models. For instance, riders such as Pascal Eenkhoorn, Yevgeniy Gidich, and Robert Stannard have collected a fair share of UCI points and demonstrated their potential at times, but did not collect a (notable) victory yet. We would like to point out that, due to the Covid-19 pandemic, the cycling season 2020 (and to a lesser extend also 2021) has been crippled, as many (smaller) races were not organized. Since neo-pros typically are scheduled for those races, we believe that they may not have received the same opportunities to collect UCI points as in normal seasons, which may to some degree explain why some riders have not performed as predicted.

7 Conclusion

We provided an overview on the related literature on talent identification in cycling and developed three statistical methods that allow to identify talented U23 riders, based on publicly available data (race results, age, and nationality). Despite the fact that this data is not very detailed, our results are quite encouraging, as we are able to predict all top-10 successful professional riders from a set of riders that had 2018 as their final year in the U23 category. This result is interesting in light of the study by Schumacher et al. (2006), who found that results in junior races were not a significant predictor for success in elite races.

At the same time, our models provide some deeper insight in what makes a success in their first years as a professional more likely for a U23 rider. We found that this is not just determined by the number of podium finishes, but also by the UCI points collected, top 20 results in mountain and hill stages, experience in the the U23 category, as well as the rider's continent.

We see our models as a valuable tool for cycling scouts or agents, as they can relatively easily be applied to a massive set of riders and data, to highlight the most promising talents. These talents could then be invited for a lab test, to collect the physiological data necessary to get a better impression of the potential of the rider (see e.g. Svendsen et al. (2018)). As the final position in a race does not fully reveal a rider's impact on the race, it is clear that our models would benefit from more detailed position data such as the total number of kilometers a rider was present at the front of the peloton. Although such data is currently not yet (publicly) available, the rise of sensor data and data-driven summarization techniques may change this in the near future (see e.g. Verstockt et al. (2020)).

Finally, the attentive reader will have noticed that our models do not answer the question in the introduction whether or not Remco Evenepoel is the new Eddy Merckx. The reason is that Evenepoel never rode an U23 race, moving from the juniors category directly to the professional level. However, with that in mind, and taking into account that he already won a World Tour stage race, a classic, and a gold and silver medal at the European and Worlds time trial championship respectively in his first two seasons as a professional rider, his talent is undoubtedly exceptional.

Acknowledgements David Van Bulck is a post-doctoral research fellow funded by the Research Foundation – Flanders (FWO) [1258021N].

References

Atkinson G, Davison R, Jeukendrup A, Passfield L (2003) Science and cycling: current knowledge and future directions for research. Journal of Sports Sciences 21:767–787

Boon BH, Sierksma G (2003) Team formation: Matching quality supply and quality demand. European Journal of Operational Research 148:277 – 292

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39:324–345

Brocard J, Larson D (2016) Agents in professional road cycling. In: Van Reeth D, Larson DJ (eds) The economics of professional road cycling, Springer, pp 147–163

Cabaud B, Scelles N, Morrow S, François A (2016) Modeling performances and competitive balance in professional road cycling. In: Van Reeth D, Larson DJ (eds) The economics of professional road cycling, Springer, pp 257–283

Cherchye L, Vermeulen F (2006) Robust rankings of multidimensional performances: An application to Tour de France racing cyclists. Journal of Sports Economics 7:359–373

Cobley S, Baker J, Schorer J (2020) Talent identification and development in sport: an introduction to a field of expanding research and practice. In: Cobley S, Baker J, Schorer J (eds) Talent Identification and Development in Sport: international perspectives, Routledge, pp 1–16

Faria EW, Parker DL, Faria IE (2005a) The science of cycling: Factors affecting performance – part 2. Sports medicine 35:313–338

Faria EW, Parker DL, Faria IE (2005b) The science of cycling: Physiology and training – part 1. Sports medicine 35:285–312

Farrand S (2018) Remco Evenepoel: Don't call me the next Eddy Merckx. Cyclingnews URL `https://www.cyclingnews.com/news/remco-evenepoel-dont-call-me-the-next-eddy-merckx`

Hopker J (2016) Identifying and developing talent in cycle sports. Aspetar Sports Medicine Journal 5:416–422

Hsia R (2017) Ranking and prediction for Cycling Canada. Master's thesis, Simon Fraser University

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer

Johnston K, Wattie N, Schorer J, Baker J (2018) Talent identification in sport: a systematic review. Sports Medicine 48:97–109

Kholkine L, De Schepper T, Verdonck T, Latré S (2020) A machine learning approach for road cycling race performance prediction. In: Brefeld U, Davis J, Van Haaren J, Zimmermann A (eds) International Workshop on Machine Learning and Data Mining for Sports Analytics, Springer, pp 103–112

Lehman B (2020) Projecting NFL potential from college career performance curve. MIT Sloan Sports Analytics Conference URL `https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5f6a66d98d495c62ab669ad0_SSAC2020-Paper-Projecting-NFL-Potential-final.pdf`

Lentillon-Kaestner V, Carstairs C (2010) Doping use among young elite cyclists: a qualitative psychosociological approach. Scandinavian Journal of Medicine and Science in Sports 20:336–345

Longo AF, Siffredi CR, Cardey ML, Aquilino GD, Lentini NA (2016) Age of peak performance in Olympic sports: A comparative research among disciplines. Journal of Human Sport and Exercise 11:31–41

Lucia A, Earnest C, Arribas C (2003) The Tour de France: a physiological review. Scandinavian Journal of Medicine & Science in Sports 13:275–283

Manisera M, Sandri M, Zuccolotto P (2020) Advances in basketball statistics. In: Ley C, Dominicy Y (eds) Science meets sports: When statistics are more than numbers, Cambridge Scholars Publishing, pp 19–52

Maton M (2020) The Next Egan Bernal: Predicting Elite New Professionals in Road Cycling Using Data Analysis of Youth Series Races. Master's thesis, Ghent University

Menaspà P, Sassi A, Impellizzeri FM (2010) Aerobic fitness variables do not predict the professional career of young cyclists. Medicine and science in sports and exercise 42:805–812

Mignot J (2016) The history of professional road cycling. In: Van Reeth D, Larson DJ (eds) The economics of professional road cycling, Springer, pp 7–31

Mostaert M, Laureys F, Vansteenkiste P, Pion J, Deconinck FJ, Lenoir M (2020) Discriminating performance profiles of cycling disciplines. International Journal of Sports Science & Coaching 16:110–122

Mostaert M, Vansteenkiste P, Pion J, Deconinck FJ, Lenoir M (2021) The importance of performance in youth competitions as an indicator of future success in cycling. European Journal of Sport Science In Press, URL `doi.org/10.1080/17461391.2021.1877359`

Muazu Musa R, Abdul Majeed A, Taha Z, Abdullah M, Husin Musawi Maliki A, Azura Kosni N (2019) The application of artificial neural network and k-nearest neighbour classification models in the scouting of high-performance archers from a selected fitness and motor skill performance parameters. Science & Sports 34:e241 – e249

Olds T (1998) The mathematics of breaking away and chasing in cycling. European journal of applied physiology and occupational physiology 77:492–497

Olds TS, Norton KI, Lowe EL, Olive S, Reay F, Ly S (1995) Modeling road-cycling performance. Journal of Applied Physiology 78:1596–1611

Pappalardo L, Cintia P, Ferragina P, Massucco E, Pedreschi D, Giannotti F (2019) Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transaction on Intelligent Systems and Technology 10:59

Phillips KE, Hopkins WG (2020) Determinants of cycling performance: a review of the dimensions and features regulating performance in elite cycling competitions. Sports Medicine - Open 6:6–23

Prinz J, Wicker P (2012) Team and individual performance in the Tour de France. Team Performance Management 18:418–432

Rebeggiani L (2016) The organizational structure of professional road cycling. In: The economics of professional road cycling, Springer, pp 33–54

Rodríguez-Gutiérrez C (2014) Leadership and efficiency in professional cycling. EDP, Economic Discussion papers URL econpapers.repec.org/article/jsfintjsf/v_3a9_3ay_3a2014_3ai_3a4_3ap_3a315-330.htm

Rogge N, Van Reeth D, Van Puyenbroeck T (2013) Performance evaluation of Tour de France cycling teams using data envelopment analysis. International Journal of Sport Finance 8:236–257

Schumacher YO, Mroz R, Mueller P, Schmid A, Ruecker G (2006) Success in elite cycling: A prospective and retrospective analysis of race results. Journal of Sports Sciences 24(11):1149–1156

Schumaker RP, Solieman OK, Chen H (2010) Sports knowledge management and data mining. Annual Review of Information Science and Technology 44:115–157

Svendsen IS, Tønnesen E, Tjelta LI, Ørn S (2018) Training, Performance, and Physiological Predictors of a Successful Elite Senior Career in Junior Competitive Road Cyclists. International Journal of Sports Physiology and Performance 13:1287 – 1292

Vaeyens R, Lenoir M, Williams AM, Philippaerts RM (2008) Talent identification and development programmes in sport. Sports medicine 38:703–714

Vaeyens R, Güllich A, Warr CR, Philippaerts R (2009) Talent identification and promotion programmes of olympic athletes. Journal of Sports Sciences 27:1367–1380

Van Reeth D (2016) Globalization in professional road cycling. In: Van Reeth D, Larson DJ (eds) The economics of professional road cycling, Springer, pp 165–205

Verstockt S, Van den broeck A, Van Vooren B, De Smul S, De Bock J (2020) Data-driven summarization of broadcasted cycling races by automatic team and rider recognition. In: icSPORTS 2020, 8th International Conference on Sport Sciences Research and Technology Support, Proceedings, SCITEPRESS, pp 13–21

Wagner U (2010) The International Cycling Union under siege – anti-doping and the biological passport as a mission impossible? European Sport Management Quarterly 10:321–342

Williams A, Reilly T (2000) Talent identification and development in soccer. Journal of Sports Sciences 18:657–667