Competitive performance as a discriminator of doping status in elite athletes

James G. Hopker[1], Jim E. Griffin[2], Laurentiu C. Hinoveanu[1], Jonas Saugy[3], and Raphael Faiss[3]

[1]School of Sport & Exercise Sciences, University of Kent, Canterbury, Kent, UK

[2]Department of Statistical Science, University College London, London, UK

[3]Research & Expertise in Antidoping Sciences, University of Lausanne, Lausanne, Switzerland

June 21, 2023

PREPRINT not peer reviewed

Please cite as: Hopker JG, Griffin JE, Hinoveanu LC, Saugy J, Faiss R. (2023) Competitive performance as a discriminator of doping status in elite athletes. *SportRxiv*

Corresponding Author: James Hopker, School of Sport & Exercise Sciences, University of Kent, Canterbury, Kent, UK. email: J.G.Hopker@Kent.ac.uk

Abstract

As the aim of any doping regime is to improve sporting performance, it has been suggested that analysis of athlete competitive results might be informative in identifying those at greater risk of doping. This research study aimed to investigate the utility of a statistical performance model to discriminate between athletes who have a previous anti-doping rule violation (ADRV) and those who do not. We analysed performances of male and female 100m and 800m runners obtained from the World

Athletics database using a Bayesian spline model. Measures of unusual improvement in performance were quantified by comparing the yearly change in athlete's performance (delta excess performance) to quantiles of performance in their age matched peers from the database population. The discriminative ability of these measures was investigated using the area under the ROC curve (AUC) with the 50%, 75% and 90% quantiles of the population performance. The highest AUC values across age were identified for the model with a 75% quantile (AUC = 0.78-0.80). The results of this study demonstrate that delta excess performance was able to discriminate between athletes with and without ADRVs, and therefore could be used to assist in the risk stratification of athletes for anti-doping purposes.
Keywords: sports, modelling, biological passport, risk stratification, Bayesian, target testing, data analytics

1 Introduction

The current level of prevalence of doping in elite sport is unknown. Research studies involving anonymous athlete self-reports estimate the prevalence of doping within a 12-month period to be between 20% and 62% across a range of elite sports (de Hon et al., 2015; Ulrich et al., 2018). A recent study (Faiss et al., 2020) involving analysis of blood values taken from doping control tests at the Daegu (2011) and Moscow (2013) World Athletics Championships suggests the point prevalence may have been between 15 and 18% at the respective events. Nevertheless, despite the number of blood and urine samples taken from athletes across all sports remaining relatively consistent, with 241,430 taken in 2021 (267,645 in 2012, 278,047 in 2019) the percentage of those samples returning adverse analytical findings is falling (1.76% in 2012, 0.82% in 2020, 0.77% in 2021) (WADA, 2023). Therefore, given the aforementioned prevalence, questions can be raised about the efficiency of the current anti-doping policy and testing strategies of anti-doping organizations (ADOs) in identifying the right athletes, and testing them at the right time. As a consequence, there is a need to gather additional information on athletes to provide a forensic style intelligence led approach to anti-doping (Vernec, 2014). Such an approach would allow ADOs to make more informed decisions about assigning athletes to registered testing pools, better targeting of individual athlete tests, and ultimately more efficient distribution of their anti-doping resources. Indeed, anti-doping authorities such as the World Anti-Doping Agency and the Athletics Integrity Unit highlight the importance of an intelligence-led approach to anti-doping involving risk stratification of athletes based upon their athlete biological passport profile and performance (AIU, 2021; WADA,

2021).

Many factors can affect performance such as maturation (Allem and Hopkins, 2015), improved training (Haugen et al., 2018) and technological advances (Hébert-Losier, 2023). However, as the primary reasons for an athlete to dope is to artificially enhance their performance, it is intuitive to consider the analysis of their sporting performance as important information for ADOs to inform their anti-doping activities. To this effect, the most recent version of the International Standard for Testing and Investigations (WADA, 2021) highlights the use of sport performance history, including sudden major improvements and/or sustained periods of high performance as relevant factors indicating possible doping/increased risk of doping. Indeed, athletic performance has been shown to be sensitive to new anti-doping practices, such as the introduction of the ABP and out of competition doping tests in a range of sporting disciplines (Schumacher and Pottgiesser, 2009; Berthelot et al., 2010; Iljukov et al., 2020), suggesting that longitudinal monitoring of athlete performance is a viable method to inform anti-doping practice.

The main objective of what we have previously termed "the athlete performance passport" (APP) (Montagna and Hopker, 2018), is to distinguish between expected changes in sporting performance and disproportionate improvements which may be indicative of doping. We have previously developed a Bayesian hierarchical model to investigate both population and individual level longitudinal performance trajectories over time adjusted for age related changes (Griffin et al., 2022). Our work illustrated how individual performance progression could be modelled whilst allowing for confounders, such as atmospheric conditions, and could be fitted using Markov chain Monte Carlo. We calculate a term called *excess performance* by subtracting the population performance trajectory from the individual performance trajectory to show whether an athlete is performing better or worse than their age matched counterparts. Therefore as suggested above, sudden or unexpected changes in an athlete's level of excess performance might therefore be indicative of doping. Indeed, using this logic we have previously demonstrated the potential for distinguishing between the career performance trajectories of clean and doped athletes (Hopker et al., 2020). However, for use in targeted anti-doping efforts, it is necessary to identify athletes using a probability risk stratification approach. The objective of this study was therefore to validate the use of performance data to discriminate between athletes with and without previous anti-doping rule violations (ADRV). First, competitive performance results over 11-years were used to construct longitudinal profiles for individual athletes with and without ADRVs during this period, then the performance of our Bayesian model was tested using these profiles.

2 Methods

2.1 Data

We extracted 100 m and 800 m results for both male and female athletes from publicly available results databases of World Athletics including: athlete ID number, date of birth, sex, country of birth, country of representation, event details, performance result (time [s]), and finishing position. The 100 m data contained results from both male and female sprinters who had at least 5 competition results between 8th January 2011 and 28th August 2021. The database contained 2834 male athletes who have a personal best below 10.5 s and 1297 female athletes who have a personal best below 11.6 s. The male data set had 95,376 observed performances, with the female data set having 48,999 observations. The ages for males athletes ranged from 12 to 47 years, whereas females ranged from 12 and 42 years. The 800m data set contained results from both male and female middle distance runners who had at least 5 competition results between 1st January 2011 and 10th April 2022. The database contained 4382 male athletes (104,594 performance results) and 3760 female athletes (92,606 performance results). We also accessed publicly available sanction data to identify athletes with a previous anti-doping rule violation. This data comprised of the date and reason for the sanction. Only sanctions imposed for substance use that have been shown to have a performance enhancing effect in the concerned discipline (i.e. 100 m or 800 m) were included within the subsequent analysis.

2.2 Modelling Performance

Our methodology for modelling performance has been developed over several years (see (Montagna and Hopker, 2018; Griffin et al., 2022; Hopker et al., 2020, 2018)). We use the specification of a Bayesian spline model documented in Griffin et al. (2022) to construct performance trajectories for individual athletes. In brief, our model assumes individual performances can be represented as the sum of an individual performance trajectory, the effects of sport/discipline specific confounders and an observation error. The model is summarised by the equation below for M athletes, with $y_{i,j}$ indicating the j-th performance for athlete i at age $t_{i,j}$ (measured in years) and $x_{i,j}$ representing any observed confounders (e.g. atmospheric conditions) for that performance. We use n_i to denote the number of performances for individual i. The model is

$$y_{i,j} = h_i(t) + x_{i,j}\,\zeta + \epsilon_{i,j}, \qquad j = 1, \ldots, n_i,\ i = 1, \ldots, M$$

where h_i is the individual performance trajectory for the i-th individual, ζ are population level regression coefficients for the effects of confounders, and $\epsilon_{i,j}$ are observation errors which are assumed to follow a standard skew-t distribution (Azzalini and Capitanio, 2003). This error distribution, rather than the usual normal distribution, allows for the skewness and heavy tails observed in sporting performance data (i.e. poor performances lie much further from the median performance than exceptionally good performances). We express the individual performance trajectory $h_i(t)$ as the sum of two parts: the population performance trajectory $g(t)$ and the *excess performance* trajectory of the i-th athlete so that $h_i(t) = g(t) + f_i(t)$. The excess performance trajectory represents individual performances adjusted for the average performance of athletes within the population at the same age as well as any confounders and forms the basis of our risk stratification measure. The population performance trajectory $g(t)$ is modelled as a fourth-degree polynomial, which Griffin et al. (2022) find is sufficiently flexible for sporting performance, and $f_i(t)$ is flexibly modelled by separate Bayesian linear spline model for each athlete. The model is identified by assuming that the prior mean of $f_i(a)$ is 0, where a is the smallest integer age in the database.

2.3 Athlete Risk Stratification

We develop an athlete risk stratification measure using changes in excess performance, which adjusts individual performance for the expected effects of age and confounders, and therefore does not depend on absolute level of performance (which will be heavily influenced by physiological factors). To understand risk, we assume that an athlete who increases their level of competitive performance more rapidly than seen in the comparator population is likely to be at greater risk of doping and therefore warrant closer scrutiny by anti-doping organisations.

To make these ideas more precise, we define *delta excess performance* to be the change in excess performance for an athlete over a fixed period (we will use 1 year – although other values could be considered). For the delta excess performance between ages $j-1$ and j, we define $\Delta_{i,j} = f_i(j) - f_i(j-1)$ to be the one-year delta excess performance for athlete i observed at age j. If delta excess performance was observed, we could define a risk measure by considering how an athlete's level of delta excess performance at a given age compares to a fixed percentile of population distribution of delta excess performance across athletes at the same age. However, since the delta excess is not observed, we must use estimates to define our risk stratification measure. Our Bayesian analysis allows us to estimate delta excess performance for an individual at particular age by its posterior median and to quantify estimation error using its posterior distribution. Firstly, we estimate the population distribution of delta excess

performance at a particular age by the population distribution of the corresponding posterior median estimates. Secondly, rather than comparing an individual posterior median of excess performance to the population, we allow for estimation error by calculating the posterior probability that an athlete's delta excess performance exceeds a fixed percentile of the estimated population distribution. Specifically, we consider the 55th, 75th and 90th percentiles in our analysis and denote the corresponding risk scores as $M_i^1(j)$, $M_i^2(j)$ and $M_i^3(j)$ for athlete i at age j. Further details of this calculation using output from a Markov chain Monte Carlo algorithm are given in Appendix A.1. Under these risk scores, athletes with larger values will have a greater risk of doping.

ROC analysis was used to evaluate the ability of the risk scores to discriminate performance profiles as either leading to an ADRV or not ADRV in the next d years. We treat this as a binary classification problem for each integer age j and use the standard area under the ROC curve (AUC) as our metric of classification ability. This metric takes values between 0 and 1 with larger values associated with better discrimination. A value of 1 implies perfect discrimination and 0.5 is the same as guessing at random.

The use of ADRVs rather than the (unobserved) true doping statuses of athletes has some important implications. We can only consider whether an athlete receives an ADRV over a period of d years and so we will also define the doping status of an athlete over the same period. We define the "doping" group to contain athletes who are, at some time during the period, involved with a doping regime that is designed to increase their performance over time, rather than those involved "one-off" instances of doping. We will refer to all athletes not in this doping group as "clean". The period doping prevalence levels discussed in the introduction imply that many doping athletes will never receive an ADRV and so the group without an ADRV will contain both doping and clean athletes. As a consequence, if our risk stratification measure was successful at discriminating between doping and clean athletes, we could still achieve a low AUC measure since many doping athletes do not receive an ADRV in the corresponding period. For example, if the risk measure could perfectly discriminate between doping and clean athletes, then athletes who are doping but have not received an ADRV will be recorded as misclassified. This will lead to an AUC value below 1 (potentially far below 1). We quantify how the level of the mislabelling of doping athletes as without an ADRV affects the AUC metric in section 2.4 with further details provided in Appendix A.2.

The difference between the group of athletes with ADRVs and the group of doping athletes (without ADRVs) also leads to the following trade-off in the choice of d.

Firstly, the doping group contains athletes who are not doping at age j but subsequently start doping. For these athletes, our risk stratification measures will be small since the performance before age j will not be affected by doping. As d increase, the number of such athletes will tend to increase and so increasingly affect our estimate of the AUC implying a smaller value of d is preferable. Secondly, since the number of athletes with ADRVs will be small relative to the total number of athletes, the accuracy of the ROC (and the AUC measure) deteriorates as d become smaller implying that a larger value of d will be preferable to avoid a very small doped group. We consider $d = 3$, $d = 5$ and $d = 8$ to investigate this trade-off. In order to maximise the number of ADRVs recorded for a given value of d, we combined data across combinations of discipline and sex (i.e. 100m males and females & 800m males and females). Table 1 shows the number of "doped" athletes under this definition for different values of d at a range of ages. To summarize, we compared the ability of the risk measures $M_i^1(j)$, $M_i^2(j)$ or $M_i^3(j)$ to discriminate between athletes with a anti-doping rule violation (ADRV) over the following d years and the wider population of athletes without an ADRV over the following d years under the AUC metric.

Age	ADRV cases (3 years)	ADRV cases (5 years)	ADRV cases (8 years)
18	3	4	7
19	4	5	11
20	5	10	12
21	3	8	12
22	13	18	22
23	17	20	26
24	14	18	23
25	13	19	22
26	12	16	16
27	9	13	13
28	11	11	
29	6	9	
30	7	7	

Table 1: The number of identified ADRV cases across age intervals.

2.4 The effects on the AUC of the ROC curve of doping athletes without an ADRV

As we discussed in the section 2.3, some doping athletes will not receive an ADRV which will effect estimation of the AUC of the ROC curve. To investigate this effect further, we will distinguish between the true status of an athlete (which we will call either truly clean or truly doping) and the observed status of an athlete determined by ADRVs (which we will call either observed clean or observed doping). The true doping status could correspond to the one described in the previous section, but the analysis can be used with any definition of doping over a period. The approach makes several assumptions

- There are no false positives and so a truly clean athlete will never have an ADRV.
- The probability that a truly doped athlete has an ADRV (the prevalence of ADRV's in the truly doping group) is q and is the same for all doped athletes. We refer to q as the doping detection rate.
- The prevalence of doping is w.

Under these assumptions, the prevalence of ADRVs is $w\,q$ and so depends on the doping detection rate and the prevalence of doping and has the value. To understand these two values consider the following example. Suppose that the prevalence over a period of one year is 21.2% Petróczi et al. (2022). If all doping athletes only take part in a doping regime for four weeks randomly distributed throughout the year, every athlete was tested once at random throughout the year, and the test was perfectly accurate (i.e. the test result was positive if the athlete was doping), then the doping detection rate would be $4/52 = 1/13$ and the probability of an athlete receiving an ADRV would be $1/13 \times 21.2\% = 1.6\%$. This is just an example and, in practice, there are several potential confounders, such as the presence of false negatives at the testing stage, variation in doping regimes, time between doping and anti-doping test, variations in testing times etc. As a consequence, it is difficult to identify the size of the athlete population sub-group who are doping but don't have an ADRV. Therefore, within our model, we assume both the prevalence and proportion of doping athletes within the sub-group to be relatively stable over time, and therefore the probability of detection to increase over the observation time period (i.e. 3, 5 or 8 years) as more athletes will test positive. This approach therefore allows us to accommodate for the aforementioned uncertainties in identification of truly doping athletes. Therefore, the probability of an athlete receiving an ADRV (if doping) would simply be calculated by dividing the number of athletes with ADRVs by the number of athletes who are

defined as doping but without ADRVs, i.e., having established the size of the ADRV group, the choice of prevalence can be used to establish the rate of doping detection (i.e. the probability of an athlete receiving an ADRV if doping).

We distinguish between the AUC calculated using the true labels (either truly clean and truly doped) which we will call AUC_{true} and the AUC calculated using the observed labels (either observed clean and observed doped) which we will call $\text{AUC}_{\text{observed}}$. Under the assumptions above, we can relate these two metrics by:

$$\text{AUC}_{\text{observed}} = (1 - r)\,\frac{1}{2} + r\,\text{AUC}_{\text{true}}$$

where $r = \frac{1-w}{1-wq}$ is the prevalence of truly clean athletes in the observed clean group. This shows that $\text{AUC}_{\text{observed}}$ is always smaller than AUC_{true} and that the difference is controlled by the value of r. This implies that we need to be careful about how we interpret AUC metrics for the risk classification if the proportion of observed clean athletes who are truly doping is large.

$\text{AUC}_{\text{observed}}$	AUC_{true} $(w = 0.212, q = 0.5)$	AUC_{true} $(w = 0.212, q = 0.3)$	AUC_{true} $(w = 0.212, q = 0.1)$
0.78	0.82	0.83	0.85
0.75	0.78	0.80	0.81
0.70	0.73	0.74	0.75
0.68	0.70	0.71	0.72
0.65	0.67	0.68	0.69
0.60	0.61	0.62	0.62
0.55	0.56	0.56	0.56
0.50	0.50	0.50	0.50

Table 2: The effect of mislabelling of doping status on AUC values for a doping prevalence (w) of 21.2% Petróczi et al. (2022) with high to low doping detection rates (q).

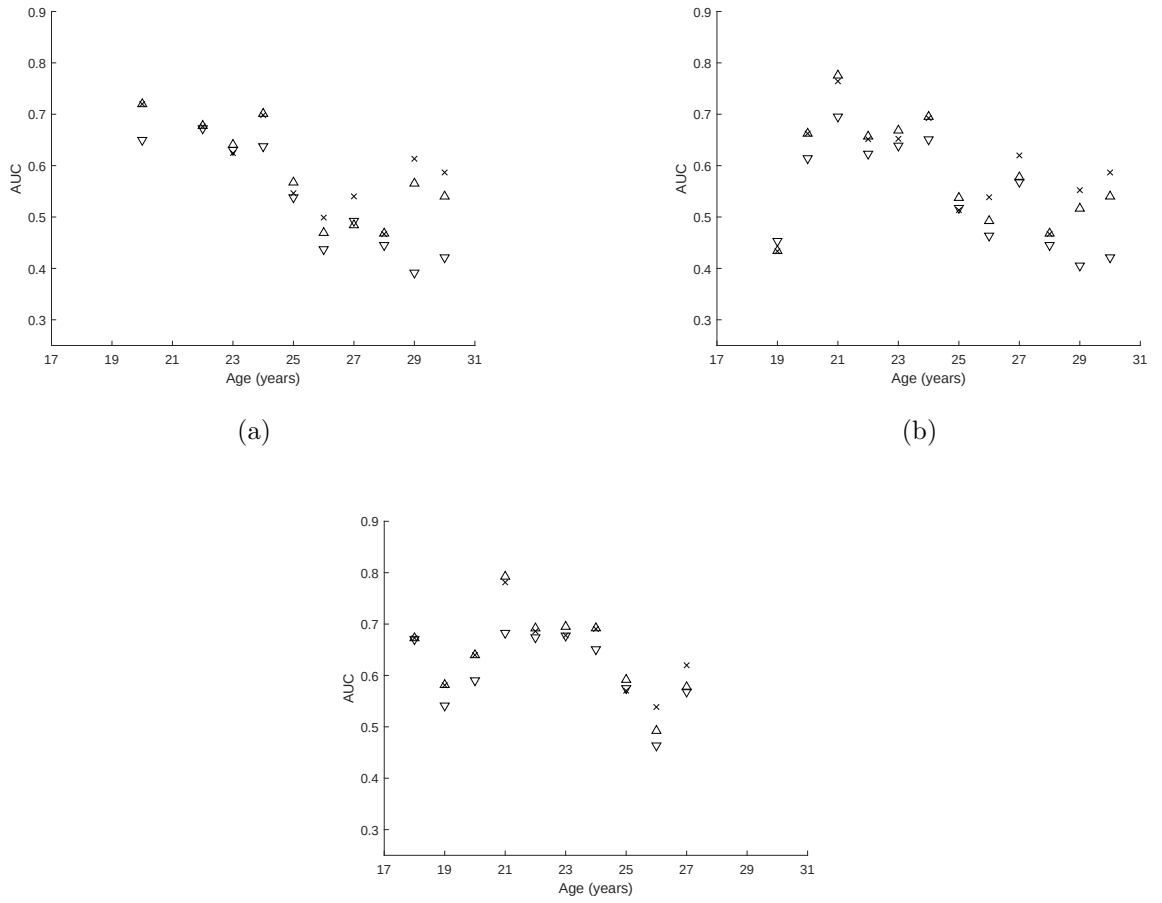
To illustrate the effect of doping athletes without an ADRV on the value of the AUC metric we used estimates of doping prevalence from the work of Petróczi et al. (2022). These researchers used a randomised response technique to estimate a doping probability in the previous 12-month period of 21.2% from athletes participating at the World Athletics Championship in Daegu, South Korea. In Table 2 we demonstrate the impact of changes in doping detection on the ability of our performance model to discriminate between doped and non-doped athletes considering Petróczi et al's

probability of doping. Given the WADA 2021 Testing Figures Report (WADA, 2023) the total percentage of adverse findings (0.65%) suggests detection is low assuming the prevalence is as high as documented in research analysing both point prevalence from abnormal blood profiles (15-18%: Faiss et al., 2020) and period prevalence from anonymous athlete self-reports (21.2%: Petróczi et al., 2022). As such, assuming a period prevalence of 21.2% and a low detection rate (q = 0.1) an $\text{AUC}_{\text{observed}}$ of 0.75 equates to an AUC without mislabelling, AUC_{true}, of 0.81. Although this difference seems quite small, the AUC metric will usually only take values between 0.5 and 1 and so if interpreted in this context, the observed change from 0.75 to 0.81 is relatively large, with values close to 0.80 suggesting very good performance.

3 Results and Discussion

We considered the ability of the risk measures described in subsection 2.3 to correctly classify performance profiles as receiving or not receiving an ADRV over a d years. Figure 1 shows how the discriminatory performance of the risk measures (as measured by the AUC metric) changes depending upon whether we consider athletes receiving an ADRV in the following 3, 5 or 8 years.

For example, when considering the model performance over a 3-year period, the AUC value for age 19 quantifies the ability of the risk measures to classify an athlete who is 19 years of age as either having or not having an ADRV in the subsequent 3 years (i.e. between ages 20 and 22). If we consider a 5-year period, we consider between the ages of 20 and 24, and an 8-year period, between the ages of 20 and 27 years. As can be seen from Figure 1, the AUC values are fairly stable for the different measures and whether the 3-, 5- or 8-year observation period is used. The risk measure M^2 (which uses a threshold of 75%) and the 5- and 8-year periods give slightly higher AUC values on average than other choices. Therefore, we recommend the use of this risk measure. All risk measures perform better for the ages 19 to 23 than 24 to 29. For ages 19 to 23, the AUC metric is between 0.65 and 0.70, which suggests that the risk measures can discriminate between athletes with and without an ADRV. Particularly since, as discussed in subsection 2.4, this is an underestimate of the AUC if we had access to the true doping status of athletes. For ages 24 to 29, the AUC metric is stable between 0.55 and 0.65 which suggests that the risk metrics are not able to consistently discriminate between ADRV and non-ADRV athletes for these ages. However, as shown in Table 1, it is important to acknowledge that there are a much greater numbers of ADRVs for ages 19 to 23 compared to ages 24 to 29. This may also reflect that the detection probability is lower between ages 24 to 29



(a)

(b)

(c)

Figure 1: AUC values for model performance at different ages over periods of 3 years (a), 5 years (b) and 8 years (c) at thresholds of 55% ($M^1 = \text{x}$), 75% ($M^2 = \triangle$) and 90% ($M^3 = \triangledown$) using delta excess performance in 100m and 800m athletes. Only age points that have more than 5 ADRV athletes are considered within the AUC analysis.

and so the underestimation of the AUC is larger for these ages.

ROC analysis allows us to consider the overall ability of a risk measure to discriminate between doped and clean athletes. It is also interesting to consider how we can choose a threshold for a given risk measure above which an athlete is considered particularly high risk of doping based upon their delta excess performance. To provide an example, we will concentrate on risk measure M^2 for the 100 m. We want to choose a threshold for the posterior probability that the delta excess performance falls outside (greater than) the 75% quantile risk measure across all athletes, and are

therefore at greater risk of doping. We used False Positive and True Positive rates to identify the posterior probability level which minimised false positives and maximised identification of athletes with ADRVs. In order to assess the specificity of the model using the 75% quantile at different age points, we assessed the False Positive rate across different probability levels for delta excess performance. The true positive rate ranges between 0.20 and 0.67 across the ages due to the changes in the number of observed true positives (i.e. ADRVs) recorded at each age, and athletes within the database. As an example, at the age of 21 years using a period of 3 years and a false positive rate of 0.1, a posterior probability threshold of 0.8 results in a true positive rate of 0.57. Incorporating all athlete's performance profiles in our sample (across years 2011 to 2022) would result in approximately 10% of 100m sprinters being flagged per year for delta excess performance. This level of prevalence is based upon our observation of athletes who receive an ADRV over a fixed number of years, which will be an under-estimate of the true doping prevalence, and is lower than has been reported by previous self-report and randomised response studies (de Hon et al., 2015; Ulrich et al., 2018)), due to the assumed high rate of false negatives.

3.1 Application to the individual athlete

The output from our model is in the form of individual performance trajectories (adjusted for covariates such as seasonality and wind effects), and is presented across four different sets of analysis. Figure 2 illustrates the performance trajectories for two 100m athletes, one with and one without an ADRV. The data points in the first column represent the raw performance times of each individual adjusted for covariates. The second column represents the data adjusted by the posterior mean population performance trajectory, month and wind effects. The third column shows the delta excess performance and the fourth column is the probability that the delta excess performance exceeds the 75% quantile of the population distribution.

As can be seen from Figure 2 the athlete with ADRV (top row) demonstrates a negative excess performance (panel b), suggesting that their performance is better than anticipated given the performance level of age-matched peers. Similarly, the delta excess performance in this example is greater than the 75% quantile (panel d), suggesting that their performance is evolving at a faster rate than anticipated at the time of their ADRV, and appears to be unabated, even after returning to competition following their doping ban. The athlete's level of excess performance continues to increase as they age, reaching 0.6s by the age of 34 years i.e. their performance decline with age is much slower compared to their age matched peers. Linked with this, there is a high probability that the athletes have exceeded the

75% quantile for delta excess performance at the time of their ADRV, acknowledging the uncertainty within the model estimates. Specifically, setting the probability of delta excess at 0.9 would flag Athlete A's performances at the ages of 21-23 and 26-27 years. By comparison, the athlete without the ADRV (Figure 2 bottom row) who has a similar absolute performance level, still demonstrates excess performance suggesting that their performance is consistently about 0.3s better than their age matched counterparts, but their delta excess performance is 0s, which indicates that their career evolves at the anticipate rate for their age. As a consequence, there is a very low probability that the athlete would exceed the 75% quantile delta excess performance. Therefore, we would conclude that the athlete is a high level sprinter that is performing better than their age matched counterparts, but at a low risk of doping.

Our retrospective analysis of competitive performance data in athletes with and without ADRVs provides an indication that longitudinal monitoring of competition results has a valuable role to play in the fight against doping in sports. Specifically, by combining this type of performance monitoring with other sources of data (e.g. biological, whereabouts, social networks etc...), there is the potential to improve the effectiveness and efficiency of anti-doping programs and bring greater certainty to the process of athlete risk stratification. In turn, athletes with a higher probability of doping risk would therefore be subject to closer scrutiny by anti-doping organisations. Moreover, given the longitudinal nature of our modelling approach and comparison to the age-matched population performance trajectory, even though an athlete may have been "clean" for many years, it is possible to "detect" an abnormal change in excess performance when doping occurs at the latter part of a career to sustain a given performance level. However, it is important to acknowledge that our model currently only considers athletic competition results in isolated disciplines. As a consequence, there is potential to miss important performance related information where an athlete competes over multiple events (e.g. 100m & 200m or 800 & 1500m). Future research is needed to consider how performance related information can be shared across different events to construct a complete performance profile for individual athletes.

4 Conclusions

This study demonstrates the utility of performance monitoring to discriminate between athletes with historical ADRVs and those without. Specifically, we demonstrate how our model could be utilised to identify athletes who are at greater risk of doping. However, it is important to recognise that high levels of delta excess performance are

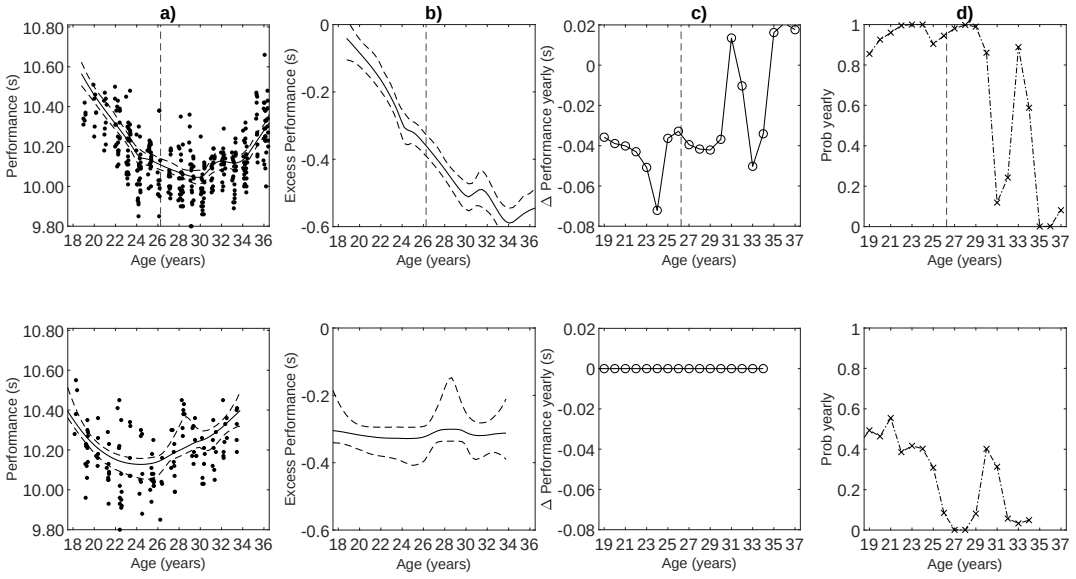


Figure 2: Illustrative performance model plots from a male sprinter with an ADRV (top row) and a male sprinter without an ADRV (bottom row). Plot a) Athlete raw performance with median (solid line) and confidence intervals (dashed lines), b) Athlete excess performance with median (solid line) and 95% credible interval (dashed lines), c) Yearly delta excess performance, d) Probability of yearly delta excess performance to exceed 75th percentile of the population. Dashed vertical line illustrates the timing of athlete A's ADRV.

not sufficient to *prove* an athlete is doping, and that information obtained from this type of analysis should be integrated with other data as part of a wider intelligence gathering approach to anti-doping.

A Appendix

A.1 Calculating the risk measure from Markov chain Monte Carlo output

Run the Markov chain Monte Carlo algorithm in Griffin et al. (2022), we will use $\theta^{(s)}$ to represent the s-th posterior sample of a parameter θ and assume that there are S samples. We define a and b to be smallest and largest integer ages in database respectively. We can calculate the risk measures in the following way:

1. For $i = 1, \ldots, M$ and $j = a, \ldots, b$, calculate a posterior sample for $\Delta_{i,j}$ by $\Delta_{i,j}^{(s)} = f_i^{(s)}(j) - f_i^{(s)}(j-1)$ for $s = 1, \ldots, S$.
2. For $i = 1, \ldots, M$ and $j = a, \ldots, b$, calculate the posterior median of $\Delta_{i,j}$, denoted $\text{med}\left(\Delta_{i,j}\right)$, by taking the sample median of $\Delta_{i,j}^{(1)}, \ldots, \Delta_{i,j}^{(S)}$.
3. We calculate the percentile of the $\text{med}\left(\Delta_{i,j}\right)$ restricted to athletes with performances in the period from j to $j+1$. Let $i_1, \ldots, i_j$ be the indices of the athletes with a performance in that period and calculate the given percentile (50% for $M_{i,j}^1$, 75% for $M_{i,j}^2$ and 90% for $M_{i,j}^3$) of $\text{med}\left(\Delta_{i_1,j}\right), \ldots, \text{med}\left(\Delta_{i_j,j}\right)$ which is written q_j
4. For $i = 1, \ldots, M$ and $j = a, \ldots, b$, calculate, the risk measure for i-th athlete in period j as the posterior probability that $\Delta_{i,j}^1$ is greater than q_j which can be calculated by
$$\frac{1}{S}\sum_{s=1}^{S} \text{I}\left(\Delta_{i,j}^{(s)} > q_j\right)$$
where $\text{I}(x) = 1$ if x is true and 0 otherwise.

A.2 The effects on the AUC of the ROC curve of doping athletes without an ADRV

In this appendix, we provide more details on understanding the effect of doping athletes without ADRVs on the ROC curve and the AUC metric including mathematical details.

For a randomly chosen athlete, we define the random variables O to represent the observed status of that athlete (clean/doped) and Y to represent the true status of that athlete (clean/doped). We define $O = 1$ if the athlete is observed doped and $O = 0$ if the athlete is observed clean (and similarly for Y). The assumption in subsection 2.4 can be expressed as

- There are no false positives and so a truly clean athlete will never have an ADRY implying that $\Pr(O = 0|Y = 0) = 1$.
- The probability that a doped athlete has an ADRVs is q and is the same for all doped athletes. This implies that $\Pr(O = 1|Y = 1) = q$ and so $\Pr(O = 0|Y = 1) = 1 - q$
- The prevalence of doping is w which implies that $\Pr(Y = 1) = w$ or $\Pr(Y = 0) = 1 - w$.

Gneiting and Vogel (2022) show how the theoretical ROC curve can be written in terms of the probability distributions of the risk measure for the clean and doped groups. If we consider the truly clean and doped groups, the distribution of the risk measure for the truly clean and truly doped groups are denoted F_{true} and G_{true}. The ROC curve for these true groupings can be written as

$$R_{\text{true}}(p) = 1 - G_{\text{true}}\left(F^{-1}_{\text{true}}(1-p)\right), \qquad 0 < p < 1.$$

Similarly, we can define a theoretical ROC curve under the observed groupings. This involves the distribution of the risk measure for the observed clean and observed doped groups which are denoted F_{observed} and G_{observed}. We can link these distributions to F_{true} and G_{true}. Firstly,

$$\begin{aligned}
\Pr(O = 0) &= \Pr(O = 0|Y = 1)\Pr(Y = 1) + \underbrace{\Pr(O = 0|Y = 0)\Pr(Y = 0)}_{1-w} \\
&= (1-q)w + 1 - w = 1 - q\,w \\
\Pr(O = 1) &= \Pr(O = 1|Y = 1)\Pr(Y = 1) + \underbrace{\Pr(O = 1|Y = 0)\Pr(Y = 0)}_{0} = q\,w
\end{aligned}$$

$$\begin{aligned}
&\Pr(X \le x, O = 0) \\
=&\Pr(X \le x|O = 0, Y = 0)\ \Pr(O = 0|Y = 0)\Pr(Y = 0) \\
&+ \Pr(X \le x|O = 0, Y = 1)\ \Pr(O = 0|Y = 1)\Pr(Y = 1) \\
=&F_{\text{true}}(x)\,(1-w) + G_{\text{true}}(x)\,(1-q)\,w
\end{aligned}$$

and

$$\begin{aligned}
&\Pr(X \le x, O = 1) \\
=&\underbrace{\Pr(X \le x|O = 1, Y = 0)\ \Pr(O = 1|Y = 0)\Pr(Y = 0)}_{0} \\
&+ \Pr(X \le x|O = 1, Y = 1)\ \Pr(O = 1|Y = 1)\Pr(Y = 1) \\
=&G_{\text{true}}(x)\,q\,w.
\end{aligned}$$

This allows to calculate F_{observed} and G_{observed} as

$$F_{\text{observed}}(x) = \Pr(X \le x | O = 0) = \frac{\Pr(X \le x, O = 0)}{\Pr(O = 0)} \tag{1}$$

$$= \frac{F_{\text{true}}(x)\,(1-w) + G_{\text{true}}(x)\,(1-q)\,w}{1 - q\,w}$$

$$= r\,F_{\text{true}}(x) + (1-r)\,G_{\text{true}}(x) \tag{2}$$

$$G_{\text{observed}}(x) = \Pr(X \le x | \emptyset = 1) = \frac{\Pr(X \le x, O = 1)}{\Pr(O = 1)} = G_{\text{true}}(x) \tag{3}$$

where $r = \dfrac{1-w}{1-w\,q}$. This could be used to express the theoretical ROC curve for the observed groups, which is

$$R_{\text{observed}}(p) = 1 - G_{\text{observed}}\left(F^{-1}_{\text{observed}}(1-p)\right), \qquad 0 < p < 1,$$

in terms of F_{observed} and G_{observed} (although, this does not lead to a simple expression).

We now consider how $\text{AUC}_{observed}$ is related to AUC_{true}. Firstly, we can show that, $\text{AUC}_{\text{observed}}$ can be expressed as

$$\text{AUC}_{\text{observed}} = \int_0^1 R_{\text{observed}}(p)\; dp = 1 - \int_0^1 G_{\text{observed}}\left(F^{-1}_{\text{observed}}(1-p)\right)\; dp$$

$$= \int_0^1 F_{\text{observed}}\left(G^{-1}_{\text{observed}}(p)\right)\; dp\,.$$

A proof of this result is given in Appendix B. Using (2) and (3), we get

$$\text{AUC}_{\text{observed}} = \int_0^1 F_{\text{observed}}\left(G^{-1}_{\text{observed}}(p)\right)\; dp$$

$$= \int_0^1 (1-r)\,G_{\text{true}}\left(G^{-1}_{\text{true}}(p)\right) + r\,F_{\text{true}}\left(G^{-1}_{\text{true}}(p)\right) dp$$

$$= \int_0^1 (1-r)\,p\,dp + r\int_0^1 F_{\text{true}}(G^{-1}_{\text{true}}(p))\,dp$$

$$= (1-r)\,\frac{1}{2} + r\,\text{AUC}_{\text{true}}.$$

B Proof of expression for $\text{AUC}_{\text{observed}}$

Consider

$$\text{AUC}_{\text{observed}} = 1 - \int_0^1 G_{\text{observed}}\left(F^{-1}_{\text{observed}}(1-p)\right).$$

Making the change of variable $(1-p) \to p$ leads to

$$\text{AUC}_{\text{observed}} = 1 - \int_0^1 G_{\text{observed}}\left(F^{-1}_{\text{observed}}(p)\right)\; dp.$$

Assuming that F_{observed} and G_{observed} are continuous implies that the composition of the functions G_{observed} and F^{-1}_{observed} is continuous and invertible. Furthermore, $G_{\text{observed}}\left(F^{-1}_{\text{observed}}(0)\right) = 0$ and $G_{\text{observed}}\left(F^{-1}_{\text{observed}}(1)\right) = 1$ since F_{observed} and G_{observed} are distribution functions. We can apply Laisant's integral formula for inverse functions (Laisant, 1905) to derive the result

$$\int_0^1 \left(G_{\text{observed}}\left(F^{-1}_{\text{observed}}\right)\right)(p)\,dp + \int_0^1 \left(G_{\text{observed}} \circ F^{-1}_{\text{observed}}\right)^{-1}(p)\,dp = 1{\cdot}1 - 0{\cdot}0 = 1$$

or, due to the properties of the inverse of a function composition,

$$\int_0^1 G_{\text{observed}}\left(F^{-1}_{\text{observed}}(p)\right)dp + \int_0^1 F_{\text{observed}}\left(G^{-1}_{\text{observed}}(p)\right)dp = 1.$$

This final equation implies that

$$\text{AUC}_{\text{observed}} = 1 - \int_0^1 G_{\text{observed}}\left(F^{-1}_{\text{observed}}(p)\right)\,dp = \int_0^1 F_{\text{observed}}\left(G^{-1}_{\text{observed}}(p)\right)\,dp.$$

References

AIU (2021). Press release: Competition manipulation is a threat to sport integrity: Aiu identifies multiple illegitimate qualifying performances for the tokyo 2020 olympic games.

Allem, S. and W. Hopkins (2015). Age of peak competitive performance of elite athletes: a systematic review. *Sports Medicine 45*, 1431–1441.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B 65*, 367–389.

Berthelot, G., M. Tafflet, N. El. Helou, and et al. (2010). Athlete atypicity on the edge of human achievement: Performances stagnate after the last peak, in 1988. *PLoS One 5*, 1–8.

de Hon, O., H. Kuipers, and van Bottenburg M. (2015). Prevalence of doping use in elite sports: a review of numbers and methods. *Sports Medicine 45*, 57–69.

Faiss, R., J. Saugy, N. Zollinger, and et al. (2020). Prevalence estimate of blood doping in elite track and field athletes during two major international events. *Frontiers in Physiology 11*, 1–11.

Gneiting, T. and P. Vogel (2022). Receiver operating characteristic (roc) curves: equivalences, beta model, and minimum distance estimation. *Machine learning 111*, 2147–2159.

Griffin, J. E., L. C. Hinoveanu, and J. G. Hopker (2022). Bayesian modelling of elite sporting performance with large databases. *Journal of Quantitative Analysis in Sports 18*(4), 253–268.

Haugen, T., G. Paulsen, S. Seiler, and O. Sandbakk (2018). New records in human power. *International Journal of Sports Physiology and Performance 13*, 678–686.

Hopker, J., J. Griffin, J. Brookhouse, J. Peters, Y. O. Schumacher, and S. Iljukov (2020). Performance profiling as an intelligence-led approach to antidoping in sports. *Drug Testing and Analysis 12*(3), 402–409.

Hopker, J., Y. Schumacher, M. Fedoruk, and et al. (2018). Athlete performance monitoring in anti-doping. *Frontiers in Physiology 9*, 1–4.

Hébert-Losier, K. Pamment, M. (2023). Advancements in running shoe technology and their effects on running economy and performance - a current concepts overview. *Sports Biomechanics 3*, 335–350.

Iljukov, S., J. Kauppi, A. Uusitalo, and et al. (2020). Association between implementation of the athlete biological passport and female elite runners' performance. *International Journal of Sports Physiology and Performance 15*, 1231–1236.

Laisant, C.-A. (1905). Intégration des fonctions inverses. *Nouvelles annales de mathématiques : journal des candidats aux écoles polytechnique et normale 4e série, 5*, 253–257.

Montagna, S. and J. G. Hopker (2018, July). A bayesian approach to the use of athlete performance data within anti-doping. *Frontiers in Physiology 9*(884).

Petróczi, A., M. Cruyff, O. de Hon, D. Sagoe, and M. Saugy (2022). Hidden figures: Revisiting doping prevalence estimates previously reported for two major international sport events in the context of further empirical evidence and the extant literature. *Frontiers in Sports and Active Living 4*.

Schumacher, Y. and T. Pottgiesser (2009). Performance profiling: a role for sport science in the fight against doping? *International Journal of Sports Physiology and Performance 4*, 129–133.

Ulrich, R., H. G. Pope, L. Cléret, A. Petróczi, T. Nepusz, J. Schaffer, G. Kanayama, R. Comstock, and P. Simon (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine 48*, 211–219.

Vernec, A. (2014). The athlete biological passport: an integral element of innovative strategies in antidoping. *British Journal of Sports Medicine 48*, 817–819.

WADA (2021). Testing and investigations international standard.

WADA (2023). 2021 anti-doping testing figures.