The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 3

Statistical Classification of Abnormal Blood Profiles in Athletes

Pierre-Edouard Sottas, *Swiss Laboratory for Doping
Analyses, Université de Lausanne, 1005 Lausanne,
Switzerland*

Neil Robinson, *Swiss Laboratory for Doping Analyses,
Université de Lausanne, 1005 Lausanne, Switzerland*

Sylvain Giraud, *Swiss Laboratory for Doping Analyses,
Université de Lausanne, 1005 Lausanne, Switzerland*

Franco Taroni, *School of Criminal Justice, Université de
Lausanne, 1015 Lausanne, Switzerland*

Matthias Kamber, *Swiss Federal Office for Sports, 2532
Macolin, Switzerland*

Patrice Mangin, *Institute of Legal Medicine, Université de
Lausanne, 1005 Lausanne, Switzerland*

Martial Saugy, *Swiss Laboratory for Doping Analyses,
Université de Lausanne, 1005 Lausanne, Switzerland*

Recommended Citation:

Sottas, Pierre-Edouard; Robinson, Neil; Giraud, Sylvain; Taroni, Franco; Kamber, Matthias; Mangin, Patrice; and Saugy, Martial (2006) "Statistical Classification of Abnormal Blood Profiles in Athletes," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 3.

DOI: 10.2202/1557-4679.1011

Statistical Classification of Abnormal Blood Profiles in Athletes

Pierre-Edouard Sottas, Neil Robinson, Sylvain Giraud, Franco Taroni, Matthias Kamber, Patrice Mangin, and Martial Saugy

Abstract

Blood doping has been challenging the scientific community since the early 1970's, where it was demonstrated that blood transfusion significantly improves physical performance. Here, we present through 3 applications how statistical classification techniques can assist the implementation of effective tests to deter blood doping in elite sports. In particular, we developed a new indirect and universal test of blood doping, called Abnormal Blood Profile Score (ABPS), based on the statistical classification of indirect biomarkers of altered erythropoiesis. Up to 601 hematological profiles have been compiled in a reference database. Twenty-one of them were obtained from blood samples withdrawn from professional athletes convicted of blood doping by other direct tests. Discriminative training algorithms were used jointly with cross-validation techniques to map these labeled reference profiles to target outputs. The strict cross-validation procedure facilitates the adherence to medico-legal standards mandated by the World Anti Doping Agency (WADA). The test has a sensitivity to recombinant erythropoietin (rhEPO) abuse up to 3 times better than current generative models, independently whether the athlete is currently taking rhEPO or has stopped the treatment. The test is also sensitive to any form of blood transfusion, autologous transfusion included. We finally conclude why a probabilistic approach should be encouraged for the evaluation of evidence in anti-doping area of investigation.

KEYWORDS: statistical classification, blood doping, biomarkers, cross-validation, forensic standards, scientific evidence

1 INTRODUCTION

In anti-doping research, statistical quantities are often used to describe the prevalence of abnormal biological profiles (Casoni *et al.* , 1993; Vergouwen *et al.* , 1999; Videman *et al.* , 2000; Stray-Gundersen *et al.* , 2003), but they are rarely used to propose a decision strategy. In this domain, statistical pattern classification (Duda *et al.* , 2001) looks particularly promising and, surprisingly, has been limited to the use of simple generative models (Parisotto *et al.* , 2001; Gore *et al.* , 2003). As opposed to generative models, discriminative training algorithms are designed to approximate the behavior of a function that maps a vector of features into one of several classes by looking at several input-output examples of the function. The optimization process can be supervised, in the sense that the output - representing the goal to achieve - is fixed a priori. In anti-doping research, the vector of input features represents the physiological information derived from analyses performed on an athlete, such as hematological markers of altered erythropoiesis in case of blood doping. As output, two class labels are defined: class $C = 0$ for a sample withdrawn from a clean athlete, and class $C = 1$ for a sample coming from an athlete who has abused of an illegal substance following WADA prohibited list. This can be interpreted as a two-category classification problem. The development of supervised algorithms is then only made possible through the creation of a database of reference patterns labeled from an independent method. For example, such patterns can be obtained from longitudinal studies implying doped as well as control subjects on a voluntary basis.

In this article, we present 3 applications of statistical classification techniques with the aim to foster research against blood doping. Firstly, we introduce discriminative training algorithms as well as their respective validation methodology in the context of homologous blood transfusion. Patterns obtained from flow cytometry quantitation techniques are classified to allow an unbiased interpretation of the homologous blood transfusion test. As a second application, we use the same methodology to classify indirect markers of altered erythropoiesis for the detection of rhEPO abuse. Reference patterns were collected in a longitudinal study implying 32 amateur athletes, some of them received a discontinuous rhEPO treatment. The key point of the study was to adjust the rhEPO doses so that the volunteers present hematocrit levels very close to 50%. Indeed, professional sportsmen have most probably adapted their behavior, taking intermittent rhEPO treatments and/or intravenous injections of small doses. Finally, with patterns obtained from samples withdrawn from professional athletes who were simultaneously convinced of doping by 2 other direct tests, we propose a new indirect test, called Abnormal Blood Profile Score, that is sensitive to multiple blood doping practices. In particular, the test is sensitive to any form of blood transfusion, autologous transfusion included.

The paper is divided in 3 sections, one per application. Each of these sections is further separated in an introduction, a subsection presenting the methods and finally the results. We conclude with a general discussion why a probabilistic approach should be encouraged for the evaluation of evidence in anti-doping area of investigation.

2 INTERPRETATION OF FLOW CYTOMETRY RESULTS

2.1 Introduction

Blood transfusion is used by athletes to increase the number of red blood cells (RBC) and subsequently for better muscle oxygenation. Autologous transfusion is the collection and reinfusion of the subject's own blood or blood components. Although an improved physical performance by blood transfusion was already demonstrated in the early 70's (Eklom *et al.* , 1972) with known consequences on several indirect hematological parameters (Berglund *et al.* , 1987), autologous transfusion has remained undetectable so far. Recent rumors in newspapers suggest that autologous transfusion, while officially prohibited by the WADA, remains present in endurance sports.

Another form of transfusion, homologous transfusion, involves withdrawing blood from another person and infusing it into a receiver compatible for ABO and Rhesus groups. However, in most cases, donor and receiver are not compatible for minor red blood cell antigens and analysis of antigenic expression patterns by flow cytometry is able to reveal the presence of mixed RBC populations in transfused recipients (Nelson *et al.* , 2002). The test of homologous blood transfusion is now applied routinely in several WADA accredited laboratories for doping control analyses, including our laboratory.

Detection of homologous blood by flow cytometry results in a collection of 8 to 10 histograms, depending on the number of antibodies used for screening. Two histograms from the 8 to 10 should present a bimodal population to declare a blood transfused. Figure 1a shows 3 typical histograms of the fluorescence intensity channel; Figure 1b represents 3 histograms obtained after analysis of bloods withdrawn from professional athletes. Two peaks characterize a bimodal antigenic expression, hence a mixed RBC population. Several analytical aspects, such as noise in electronic devices, time delay between transfusion and blood withdrawal, sample conservation procedure, user-dependent methodology and variable antigenic expression levels, prevent perfect inter-observer reproducibility. An objective interpretation is difficult in the case of two overlapping peaks (see Figure 1b) and when the minor RBC population is small as compared to the major RBC population. These aspects may lead to a decrease of the final sensitivity of the test, since any doubtful

result is returned negative to avoid any false positive.

2.2 Methods

A database of 804 flow cytometry histograms was created following a method described in details elsewhere (Nelson *et al.*, 2002). These histograms were obtained after analysis of control bloods coming from the industrial package ID-DiaPannel (DiaMed AG, Morat, Switzerland). We have also included histograms obtained from bloods of the local blood bank (Centre de Transfusion de Sang, Lausanne, Switzerland). All bloods were phenotyped using a blood agglutination test (Coombs, 1970). Control bloods have a minor RBC population of about 10%, whereas the proportion of the minor RBC population of the bloods coming from the local bank is not known. All bloods were anonymous for the sake of privacy protection. Known RBC phenotype allowed us to attach a class label to each of the 804 histograms, class $C = 0$ for 424 histograms for which we were certain they had an unimodal expression pattern, and $C = 1$ for the remaining 380. Flow cytometry analysis was performed on a FC500 Beckman-Coulter instrument with software CXP 2.0 (Beckman-Coulter Int. SA, Nyon, Switzerland).

The statistical analysis was carried out on Matlab version 6.1.0, with Statistics Toolbox version 3.0, Signal Processing Toolbox version 5.1 and Optimization Toolbox version 2.1.1. For classification with support vector machines (SVMs), we also used the independent package SVM-KMT Toolbox for Matlab (Canu *et al.*, 2003).

2.3 Results

We built a pattern classification algorithm that outputs a value enabling the classification of each histogram as unimodal or bimodal. If the algorithm finds 2 or more bimodal histograms among the 8 to 10 histograms, the blood sample will be classified as transfused. We tried 3 different feature extraction methods on the histograms made of 1024 bins:

1. low-pass signal filtering
2. fitting with a sum of 2 asymmetric bell-shaped functions
3. a combination of the first two

Low-pass filtering aims to remove high-frequency noise from the histogram (8-order Chebyshev Type 1 low-pass filter at $1/2R$ times the original sampling rate, hyper-parameter R chosen following the performance on a validation set). This pre-processing results in $N = 1024/2^R$ features.

Figure 1: Analysis of antigenic expression of red blood cells for the detection of homologous blood transfusion. A) 3 histograms obtained by flow cytometry quantitation techniques on bloods with known phenotype. Two peaks characterize a bimodal antigenic expression, hence a mixed RBC population. B) 3 histograms obtained from professional athletes blood samples collected during major sport events, 2nd and 3rd histograms are returned as single population or suspicious if interpreted on visual criteria only. C) Histogram fitting as feature extraction for classification. Blue dots are raw data, red and green lines results of the fit. D) Results of the classifier on 797 histograms after leave-one-out cross-validation. Each bar counts the number of histograms at the score given by the algorithm. The blue area corresponds to 422 single RBC population samples, the red area to 375 samples with bimodal expression pattern respectively. Red outputs have a larger variance, because correlated to the amount of blood transfused. Final test sensitivity is still enhanced because involving between 8 to 10 antigens for screening, while 2 positive histograms are needed to return a blood sample as transfused.

The second feature extraction method aims to parameterize the peaks present in the histogram. An asymmetrical bell-curved function is used to fit the 1024-points histograms that have been preliminary divided by their maximum value. We found good results with a skew-normal function with 4 parameters (Azzalini & Capitanio, 1999). Initial conditions of the non-linear least squares fitting problem are chosen so that the distribution is symmetrical and centered around the maximal value of the 1024 points, with the underlying idea to favor fitting of the major peak. A second fit with the same functional form is accomplished on the residual of the first fit. Finally, original data is fitted with the sum of 2 skew-normal functions with initial parameters taken as the parameters got after the first 2 fits (see Figure 1c). This 3-steps procedure is motivated by the aim of avoiding local minima of the mean squared error (MSE). Nonetheless, if the MSE of the residual of the last fit is superior to a given threshold value ($MSE > 0.001$ in practice), classification is stopped and the histogram is classified in class $C = 0$. This happened 7 times on the 804 histograms, 2, respectively 5, times for a histogram of class $C = 0$, respectively $C = 1$. This feature extraction algorithm decreased the 1024 original points to $N = 8$ parameters, 4 characterizing the main peak, 4 the potential second peak.

The 3rd extraction method is the concatenation of the first two, with the features vector made of all parameters obtained by the two methods. As validation techniques require high computer time demand, we studied only the case $R = 6$ in this last method.

We then built several supervised classifiers that take one of these 3 feature representations of the inputs and map it to a classification label $C = 0$ or $C = 1$. We chose classification algorithms known to be robust to overfitting (S.Haykin, 1998; Duda *et al.*, 2001): a linear discriminant analysis (LDA) and Support Vector Machines (SVMs) (S.Haykin, 1998; Duda *et al.*, 2001; Schoelkopf & Smola, 2001). An excellent generalization performance is essential to reach the quality standards needed in medico-legal settings. In contrast to what its name suggests (Duda *et al.*, 2001), LDA is a generative model that assumes Gaussian conditional density functions of the parameters. SVMs are on the other hand a discriminative model that attempts to maximize the quality of the output with a training cost function. As the use of discriminative training is new in anti-doping research, the aim is also to contrast results of both approaches. Performance is evaluated following:

1. the normalized area between the receiving operating characteristic (ROC) curve and the no-discrimination line;
2. the sensitivity level reached before to make any false positive on the database.

A leave-one-out cross-validation procedure was applied to validate the method (Hjorth, 1994; S.Haykin, 1998; Duda *et al.*, 2001): the performance is estimated

Table 1: Performance of flow cytometry histograms classification after leave-one-out cross-validation as a function of the pre-processing and classifier algorithms

	<i>LF</i> ^a		<i>CF</i>		<i>LF + CF</i>		<i>EA</i>	
	1 ^b	2	1	2	1	2	1	2
LDA	0.32	0.04	0.84	0.01	0.80	0.04	0.76	0.04
SVM	0.97	0.71	0.94	0.87	0.94	0.41	0.96	0.93

^a LF=low-pass filtering, CF= curve fitting, EA= ensemble averaging.

^b 1: area between ROC and the no-discrimination line, a value of 1 indicates a perfect classification, a null value is obtained by random guessing. 2: sensitivity level found before to make any false positive on the 804 histograms.

as if the sample put aside was a new sample to classify. We opted for a leave-one-out cross-validation instead of a K -fold cross-validation, since the former method mimics real conditions of the application of an anti-doping test. In details, among the 803 histograms of the training set (all histograms except one), 402 histograms were randomly selected to set up the estimation set, while the 401 remaining histograms formed the validation set. The estimation set is used for adjusting model parameters in the classifier. Hyper-parameters of the feature extraction (R for low-pass filtering) and classification algorithms (kernel types and parameters for SVMs) were chosen following generalization performance on the validation set. After a last estimation of the model parameters on the training set with hyper-parameters fixed, the algorithm was finally applied to the last histogram (the 804th) as if this was new data. This whole process was repeated 804 times, with a new histogram being tested at each iteration. For each histogram, the algorithm outputs a value enabling the classification.

As shown in Table 1, SVMs gave systematically the best results for both performance criteria. If the outputs with the first 2 pre-processing techniques are linearly combined to produce an overall output which is then used as a classifier - we simply took here the sum of the two outputs without cross-validating the result - the sensitivity level achieved 93% before to make any false positive (column 4). This idea is reminiscent of a committee machine known as ensemble-averaging (S.Haykin, 1998). Values obtained for the 797 histograms - all histograms except those which pre-processing failed - by this classifier are shown in Figure 1d. If a threshold value is defined as the largest value obtained by a histogram of class $C = 0$, the algorithm is able to classify more than 9 histograms over 10 in the histograms of class

$C = 1$. In practice, it is possible to include a margin to further decrease the risk to make a false positive. False negatives are characterized by a small area of the second peak, a situation that occurs when the amount of blood transfused remains small as compared to original blood. The sensitivity decreased to 66% (6 over 9 samples) when the ratio of the area of the second peak (minor RBC population) and the area of the main peak (major RBC population) is smaller than 2%. This means that 66% of these histograms are returned positive with a certainty close to 100%. As opposed to indirect testing whose sensitivity depends on the amount of blood transfused (see Section 4), these numbers confirm that the direct test of homologous blood transfusion remains sensitive when the quantity of transfused blood is small. Keeping in mind that only 2 from the 8 to 10 histograms should be positive to declare the blood sample positive, we can expect a sensitivity high enough to detect homologous blood transfusion to a point where performance enhancement is negligible.

3 Indirect test of *rhEPO*

3.1 Introduction

Current blood doping practices are versatile (Cazzola, 2000; Ashenden, 2002; Gaudard *et al.*, 2003). Besides blood transfusions, an unscrupulous athlete may choose between the intake of erythropoietic stimulants or hormones such as *rhEPO*, and the transfusion of hemoglobin-based substitutes. All these doping practices are known to enhance oxygen transfer to the muscles through a modification of athlete's oxygen transport capacity.

Modern indirect approaches of blood doping detection are mostly descriptive, or use simple generative models of classification (Audran *et al.*, 1999; Parisotto *et al.*, 2001). This is surprising because the strict validation techniques as well as a generally higher accuracy proper to discriminative models may permit to reach the forensic standards required by the WADA. Current most elaborated indirect tests of *rhEPO* (Gore *et al.*, 2003) use a LDA to derive several models sensitive to the disturbance of 5 hematological parameters caused by *rhEPO* injections. Several hematological parameters, such as reticulocytes and erythropoietin, are known to present larger values than the basal level if measured during *rhEPO* treatment, and lower values if measured after the cessation of *rhEPO* injections. From a pattern classification point of view, this is a typical XOR problem that a LDA cannot solve (S.Haykin, 1998; Duda *et al.*, 2001). Developers bypassed this problem with the development of algorithms sensitive during one of the two phases only, with an ON-score sensitive during the treatment phase and an OFF-score sensitive during the "wash-out" phase. Unfortunately, we do not know a priori if the athlete is cur-

rently abusing of rhEPO or has just stopped the treatment. A clear limit between both regions is also dependent on treatment timing and methodology adopted by the athlete. Actually, a discontinuous rhEPO treatment demonstrated that both models suffer serious limitations to detect sportsmen cheating with rhEPO (Robinson *et al.*, 2001; Robinson, 2003). In this second application, we tested if discriminative algorithms that do not need this assumption can better support the identification of athletes who have received a discontinuous rhEPO treatment. To this aim, a database of reference patterns was created from a longitudinal study implying some amateur athletes treated to rhEPO as well as control subjects. The key point of the study was to have volunteers with hematocrit levels very close to 50%. Indeed, with current cut-off limits fixed by several international sport federations, unscrupulous professional athletes seem to adapt their rhEPO treatment according to their hematocrit or hemoglobin levels easily measurable with a micro centrifuge.

3.2 Methods

The study was approved by the ethics committee of our hospital. The nature of the experiment was explained to the volunteers and their written consent was obtained. A full medical examination was undertaken before the study. 32 healthy Caucasian males took part in this randomized double blind study. Their mean age was 26 years (range 19 - 36), height 1.80 m (range 1.71 - 1.91) and weight 76.1 kg (range 59.8 - 98.6) respectively. All volunteers were requested not to abuse of alcohol, caffeine and were not allowed to smoke while they could take drugs (aspirin, paracetamol...) only after medical advice.

The 32 amateur athletes were randomly assigned to 4 groups before the administration phase: 8 subjects received a low treatment of rhEPO (group R40), 8 subjects a treatment with moderate rhEPO doses (group R80), 8 subjects received placebo (group P) and the last 8 subjects (group NT) knew they were receiving no treatment. The R40 and R80 groups were given subcutaneous injections of EPREX, three times a week (40 IU/kg for R40, 80 IU/kg for R80). The treatment in both groups was nevertheless individualized as follows: full dosage until the haematocrit reached 45 %; then, only half the dose and substitution of rhEPO by isotonic saline when the haematocrit exceeded 50 %. The P group received physiological saline instead of rhEPO. All groups also received intravenous infusions of 100 mg of iron (Venofer) once a week and took daily tablets of 15 mg of B12 (Vitarubin) and 1 mg of folic acid (Folvite). The study lasted 60 days from the first day of treatment, with an administration phase of 30 days.

Blood samples were collected twice a week. All 508 bloods samples were analyzed in our laboratory. Hematological parameters including red and reticulocyte cell populations were measured with an automatic analyzer (Cell-Dyn 4000, Abbott

Diagnostics Division, Baar, Switzerland). Precision of the analyzer was checked on a daily basis via blood controls. All serum erythropoietin (DPC-Buhlmann Laboratories, Schoenenbuch, Switzerland) and soluble transferrin receptor (Nichols Diagnostics Division, Allschwil, Switzerland) assays were performed at the end of the study on the Immulite and Advantage analysers respectively.

All statistical simulations were carried out on Matlab version 6.1.0. with Statistics Toolbox version 3.0. For this application, we also used two independent packages for Matlab: ICA-DTU Toolbox version 1.5 (Hansen *et al.* , 2001) and SVM-KMT Toolbox (Canu *et al.* , 2003).

3.3 Results

From the total of 508 bloods, 330 were control bloods, with 253 samples withdrawn from the P and NT groups, and 77 from R40 and R80 before treatment. These 330 hematological profiles were labeled with class $C = 0$. The 178 remaining blood profiles from R40 and R80 were influenced by rhEPO injections and therefore labeled with class $C = 1$.

From the 65 initial hematological and biochemical parameters, 13 were selected as potential markers of altered erythropoiesis according to literature (Parisotto *et al.* , 2001; Robinson *et al.* , 2001). In the future, it could be interesting to see if overall performance can be increased with dedicated feature selection algorithms (Guyon & Elisseeff, 2003). The 13 selected markers are: red blood cell count (RBC), optically measured red blood cell count (RBCo), hemoglobin (HGB), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin content (MCHC), hematocrit (HCT), red cell distribution width (RDW), reticulocyte count (RETC), reticulocyte percentage (RETI%), immature reticulocyte fraction (IRF), erythropoietin (EPO) and soluble transferrin receptor (sTFR). Reticulocyte values were standardized to adjust for instrument bias (Ashenden *et al.* , 2004), which was statistically non-significant here.

A different distribution was found for all 13 parameters if separated following their class C (2-sample Kolmogorov-Smirnov test, $p < 0.0001$), at the exception of MCH ($p = 0.17$). This suggests that MCH is not a good marker of rhEPO intake. On the remaining 12 parameters, we estimated the number of independent components with a Bayesian Information Criterion (BIC) approximation (Hand & Yu, 2001). On the 508 blood profiles, BIC suggests between 7 to 11 ($p = 97\%$) independent components, with a maximum at 8 components ($p = 49\%$). Despite that 99% of the variance is explained in the first 8 dimensions of a PCA analysis, we nevertheless let data unchanged 1) for the ease of interpretation, and 2) to keep the opportunity to work with a smaller number of variables. Note that the dependence between the markers is implicitly taken into account in the results presented below.

Parametric models have been proposed to describe the variability of some hematological parameters (Parisotto *et al.* , 2001; Sharpe *et al.* , 2002). It has been proposed to take the square root of reticulocyte values (RETI%, RETC) and the logarithm of EPO and sTFR to be closer to a normal distribution. Here, we found evidence against normality for several transformed variables before rhEPO treatment, both on the groups formed with the total of 330 profiles of class $C = 0$ and on the groups formed with the 32 initial profiles before rhEPO treatment (2-sided Lilliefors test, $p < 0.03$ for $\sqrt{\text{RETI\%}}$, $p < 0.02$ for $\sqrt{\text{RETC}}$, $p < 0.02$ for $\log(\text{sTFR})$, $n=32$). Therefore, we decided to use instead non-parametric techniques of density estimation (Silverman, 1986; Duda *et al.* , 2001). There is also no theoretical justification that hematological parameters should have a specific parametric distribution after discontinuous rhEPO treatment. We nevertheless used the proposed distributions as kernel window functions. The densities of the 12 hematological parameters were firstly estimated on the 330 control samples of class $C=0$. The mean and 95% reference range obtained on the computed densities for the control samples are shown in Table 2. They are in good agreement with ranges calculated for elite athletes in a large-scale study (Sharpe *et al.* , 2002), except for sTFR, most probably due to a lack of standardization between different methods. No significant departure was observed on the estimates with a jackknife resampling method on the 32 athletes. These preliminary tests suggest that 1) the kernel density approach is effective; 2) sample size (32 athletes) is large enough; 3) inter-instrument bias remains limited for all markers, except for sTFR.

After kernel density estimation of the parameters on the samples of class $C = 1$, we used a naive Bayesian approach to classify the data (Friedman *et al.* , 1997). Naive Bayesian classification assumes independent binomial conditional density models. This model gave excellent results in many complex real-world situations even if the assumption is clearly violated (Hand & Yu, 2001). Note that the evident violation of this assumption does not preclude the validity of the results presented here. As output, a log-likelihood ratio indicates how likely the blood profile is characteristic of rhEPO treatment. Here, decision making is based on the log-likelihood ratio alone, without including the prevalence of both classes as an additional criterion in the decision. This digression from a traditional Bayesian decision scheme is considered in the final discussion.

For the ease of interpretation, the log-likelihood ratio was scaled and divided by the value at which no false positive was found on the 330 blood profiles of class $C = 0$. A negative ratio indicates a high chance that the hematological profile was obtained from an unmanipulated blood, an output between 0.0 and 1.0 suggests a suspicious blood profile, whereas no value exceeding 1.0 was obtained on the 330 control samples. Kernel density estimation as well as naive Bayesian classification were validated by cross-validation, with the training set composed of all samples

Table 2: Reference ranges and mean of 12 hematological parameters in amateur male athletes after kernel density estimation

blood parameter	95% reference range	mean
RETI% [%]	0.58-2.22	1.32
HGB [g/L]	134-170	151
EPO [mU/mL]	5.0-26.8	13.2
sTFR [nmol/L]	4.64-25.4	14.4
HCT [%]	39.1-49.7	43.9
RBC [$10^6/\mu\text{L}$]	4.12-5.93	4.94
RBCo [$10^6/\mu\text{L}$]	4.09-5.85	4.89
MCV [fL]	81.5-98.2	89.1
MCHC [g/dL]	32.6-36.0	34.3
RDW [%]	10.9-13.1	11.9
RETC [$10^3/\mu\text{L}$]	31.2-106.5	64.9
IRF [%]	0.13-0.36	0.24

except those coming from one athlete. This procedure was repeated 32 times, once per athlete, as if the data from the test set was new data. The same computer intensive "leave-one-athlete-out" validation method was also used for classification with SVM on raw data, except that the training set was further separated in two distinct sets to estimate SVMs hyper-parameters (kernel's type and parameters).

As summarized in Table 3, we tried 5 different algorithms: a first version with the 12 initial markers, the second with 4 markers that are believed to be biologically relevant (HGB, RETI%, EPO, sTFR), a third version that takes practical and financial aspects into account with 10 parameters that can be obtained "online" on a portable analyzer (all markers except EPO and sTFR), and finally 2 versions with 7 and 3 parameters respectively that we will use in the 3rd application below. In order to compare the performance with existing models, the rate of detection is plotted in Figure 2 at a high level of specificity [0.99 – 0.997], for the 5 SVM models as well as the OFF-hr, OFF-hre and ON-he scores (Gore *et al.*, 2003) applied on the 508 hematological profiles. A gain in sensitivity up to 3-4 times is obtained with the new models, without any explicit a priori assumption taken on athlete's state and treatment timing. Elevated rates of detection at a high specificity are obtained when the number of erythropoiesis markers is large. Two reasons may explain this: 1) as put forward by the BIC analysis, erythropoiesis seems to be a complex biological

Table 3: Classification performance of blood profiles after discontinuous treatment to recombinant erythropoietin ^a

model	# parameters	naive Bayesian		SVM	
1	12	0.88	0.47	0.88	0.5
2	4	0.83	0.22	0.76	0.25
3	10	0.86	0.36	0.83	0.48
4	7	0.83	0.45	0.77	0.41
5	3	0.77	0.31	0.73	0.26

^a Performance as a function of the number of hematological parameters and classifier algorithms, after "leave-one-athlete-out" cross-validation procedure. First value is the area between the ROC curve and the no-discrimination line, the second the sensitivity level found at 100% specificity on 508 blood profiles measured on 32 athletes, with 330 control samples and 178 positive at rhEPO.

process whose description needs a large number of hematological parameters; 2) analytical variability may have a lower impact on sensitivity if correlated parameters are measured simultaneously but somewhat independently.

4 Abnormal Blood Profile Score

4.1 Introduction

Because of the recent progress of biotechnology and the development of products indistinguishable from endogenous proteins and hormones, it appears inevitable to develop indirect tests that base their decision on the consequences of the substance intake, and not on the substance itself (Cazzola, 2000; Ashenden, 2002). The recent development of Dynepo by the industry, a EPO based product that possesses human glycosylation, is a good illustration. Although recently approved in several countries, this product is in theory undetectable by any rhEPO direct test (Wide *et al.* , 1995; Lasne & de Ceaurriz, 2000). This should be contrasted with the indirect rhEPO test that we developed in the previous section that remains valid for the detection of Dynepo.

Another advantage of indirect testing is the possibility to make the test sensitive to multiple doping practices. An example of an indirect and universal test is the so-called Hematocrit rule, a simple rule that prevents athletes with elevated HCT

Figure 2: Rate of detection of indirect tests to recombinant erythropoietin at a low level of false positives. Previous generative models: a) green dashed line: ON-he, b) red dashed line: OFF-hre, c) red solid line OFF-hr. New discriminative training algorithms after "leave-one-athlete-out" cross-validation procedure: models with d) 3, e) 4, f) 7, g) 10, and h) 12 hematological parameters in yellow solid line, black dashed line, blue solid line, black solid line and blue dashed line respectively. Models with solid lines use haematological markers that can be obtained online on a portable analyzer. Rates calculated from 508 blood samples collected on 32 amateur athletes, 178 samples are positive because of discontinuous rhEPO treatment.

from competing. Although its elevated proportion of false-positives (up to 2.5% for a limit fixed at 50%) has attracted justified criticism, this naive rule has the interesting property to be independent of the doping practice.

In this 3rd application, we show that the same property of universality can be obtained thanks to the techniques peculiar to statistical classification, but this time without the poor sensivity/specificity relation of the Hematocrit rule. To this aim, we developed a test that uses some hematological profiles obtained from bloods withdrawn from professional athletes that have been deemed by a direct and independent test to be positive to either rhEPO or blood transfusion. Both rhEPO urine test (Lasne & de Ceaurriz, 2000) and homologous blood transfusion test (Nelson *et al.*, 2002) are carried out in our laboratory on a regular basis for doping control analysis. Sometimes, we also collect hematological profiles simultaneously on the same athletes. If the direct test returns a positive result, it is then possible to attribute a class label $C = 1$ to the corresponding hematological profile. Interestingly enough, the knowledge of the class label from an independent method makes the use of supervised pattern classification algorithms possible.

Finally, we also exploited hematological profiles obtained from bloods withdrawn from 11 professional athletes who participated to a general Swiss study to fight against doping. We can add with confidence these profiles in class $C = 0$, since these 11 athletes signed statements of informed consent to participate to the study and were extensively controlled during a period of 2 years. From a statistical point of view, this procedure has also the advantage to sample directly in the targeted population. In that case, the possible bias that may exist between amateur athletes and professional athletes is non-existent.

4.2 Methods

Both rhEPO urine test (Lasne & de Ceaurriz, 2000) and homologous blood transfusion test (Nelson *et al.*, 2002) are performed routinely in our laboratory (ISO 17025 accreditation, WADA accreditation).

We used 2 instruments for the measurements of the hematological parameters on bloods withdrawn from athletes simultaneously deemed positive by the rhEPO urine test: a Sysmex R-500 (Sysmex, Norderstedt, Germany) for the RETI% parameter, and a Beckman Coulter Act-Series (Beckman Coulter International, Nyon, Switzerland) for HGB and HCT parameters. These measurements were carried out between 2001 and 2003.

A portable Sysmex XT-2000i (Sysmex, Norderstedt, Germany) was used for the measurement of the hematological parameters on bloods positive to the homologous blood transfusion test. 7 hematological parameters were collected on this analyzer (RBC, HGB, HCT, MCV, MCH, MCHC, RETI%). It was shown that this machine

and the machine used for the rhEPO study show virtually no inter-instrument bias for the RETI% parameter (Ashenden *et al.*, 2004), a result confirmed by our own data on another control group of 29 amateur sportsmen (data not shown).

22 professional athletes, 11 males and 11 females, participated to a general study to fight against doping, called "Doping Freier Spitzen Sport (DPSS)". This study was mandated and organized by the Swiss Federal Office for Sports (BASPO). The professional athletes were extensively followed and tested during a period of 2 years, with control samples withdrawn in real testing conditions. Each athlete was tested 17 times in average during the study (controls in both urine and blood matrices). About half of the samples was collected out-of-competition, while the second half was withdrawn prior or during competition events. In particular, 72 blood samples were collected on the 11 male athletes during the study. The hematological parameters were obtained with a Beckman Coulter GENS (Beckman Coulter International, Nyon, Switzerland). 7 parameters (RBC, HGB, HCT, MCV, MCH, MCHC, RETI%) were set aside, the same parameters as available by the Sysmex XT-2000i. Standardization of these parameters was applied to quantify instrument bias following the 95% ranges given by the manufacturer, so as to reproduce reference ranges obtained in a large-scale study (Sharpe *et al.*, 2002).

All statistical analysis have been carried out on Matlab version 6.1.0. with Statistics Toolbox version 3.0. We also used the independent SVM-KMT Toolbox (Canu *et al.*, 2003) for Matlab.

4.3 Results

We were able to collect hematological profiles from bloods withdrawn from 10 professional athletes whose urine was simultaneously deemed to be positive by the direct rhEPO urine test. 9 of these urine samples tested positively to rhEPO and 1 urine sample to Aranesp. 9 from the 10 blood profiles were characterized by high values of the 3 hematological parameters (HGB, HCT, RETI%), confirming that the urine direct test is mostly sensitive when the athlete has recently taken an EPO based product. Unfortunately, since EPO and sTFR parameters cannot be obtained online, we were not able to compute ON-scores (Gore *et al.*, 2003).

All 10 values obtained by the naive Bayesian classifier (as presented in the last section with 3 parameters) were positive, i.e. all were suspicious. 5 from the 10 were larger than the cut-off limit. For the SVM model with 3 parameters, 9 from the 10 outputs were suspicious, with 4 exceeding the cut-off. Only 1 from the 10 samples would have been targeted as suspicious by the OFF-score with a value of 127.

11 blood samples positive to the homologous blood transfusion direct test were also analyzed. These 11 blood samples were withdrawn from 9 different male pro-

fessional athletes. These 11 hematological profiles have a class label $C = 1$. The same 7 parameters were also measured on the 11 professional athletes participating to the DFSS study, adding 72 hematological profiles of class $C = 0$. We applied the test developed for rhEPO detection with 7 markers to these 83 hematological profiles. This is challenging since 1) all samples with known class label have been withdrawn from professional athletes; 2) hematological profiles have been measured with different analyzers; 3) doping product is different (the test has been developed to detect rhEPO whereas the profiles are characteristic of blood transfusion). On the 72 profiles of class $C = 0$, the naive Bayesian classifier gave 62 negative values, 10 suspicious values between 0 and 1, and 0 values exceeding the threshold limit 1. On the 11 profiles of class $C = 1$, the same classifier outputted 2 negative values, 4 values between 0 and 1, and 5 values exceeding 1. This means the classifier reaches a sensitivity of about 45% without any false positive. For the SVM classifier, all outputs were negative for the 72 profiles of class $C = 0$, whereas 3 negative values, 7 suspicious and 1 exceeding the cut-off limit were obtained on the 11 transfused bloods. Note that no standardization of the parameters would have also lead to no false positive with both models, despite that the models were developed with data coming from another instrument (Abbott Cell-Dyn 4000 vs Beckman Coulter Gens). However, only data obtained from a rigorous analytical control with calibrated instruments and proper standardization should be used to assure good generalization properties of the algorithms on new bloods to test.

Even though we cannot exclude that the athletes with transfused blood had also taken rhEPO injections, these results suggest that hematological profiles obtained from transfused bloods share strong similarities with profiles obtained from athletes who have abused from rhEPO. This is not surprising from a physiological point of view since both doping methods share the same goal of increasing oxygen carrier capacity, with in turn a similar increase of maximal aerobic power (Ekblom & Berglund, 1991). This statistical confirmation on a model involving 7 indirect markers gives hope on the development of a universal test of blood doping. To this aim, we grouped all data to build a reference database that includes the hematological profiles of the athletes - all with the 7 markers that can be measured on a portable analyzer - and their class status $C = 0$ or $C = 1$ independently of the blood doping practice (rhEPO or blood transfusion):

1. 508 blood profiles from the discontinuous rhEPO study, 330 of class $C = 0$, 178 of class $C = 1$.
2. 11 blood profiles from professional athletes positive at the homologous blood transfusion test, all of class $C = 1$
3. 72 blood profiles from professional athletes participating to the DFSS study,

all of class $C = 0$

This reference database is composed of 591 blood profiles, with 402 control samples of class $C = 0$ and 189 samples of class $C = 1$. Then, we applied the same pattern classification algorithms, i.e. naive Bayesian after kernel density estimation and SVMs. Again, the same computer intensive "leave-one-athlete-out" cross-validation procedure is applied to validate the method. This subsampling technique gives surety on the generalization properties of the algorithm.

On the validation set, we used an ensemble averaging method (S.Haykin, 1998) that mixes outputs of both algorithms to yield a universal score that we call Abnormal Blood Profile Score (ABPS). ABPS performance is given by an area between the ROC curve and the no-discrimination line of 0.84. The algorithm achieved 45% sensitivity before to make any false positive on the full database. Furthermore, it correctly tested positively 9 of the 11 transfused samples (82%) before to make a false positive on the 402 control bloods.

Under the assumption that the consequences of blood transfusion on the hematological parameters are independent of the type of transfusion, we can expect ABPS to be sensitive to any form of blood transfusion, autologous transfusion included. This is a reasonable hypothesis since the 7 secondary markers do not depend on what may differentiate the donor's blood from the recipient's blood, such as a different antigenic expression pattern.

2 rationales give us confidence on the generalization properties of the algorithm on future bloods to test:

1. a significant part of the blood samples have been withdrawn directly from the targeted population with rigorous analytical control;
2. the strict cross-validation procedure that we employed here mimics real conditions of testing.

The sensitivity of the test depends of course on the amount of blood reinfused by the unscrupulous athlete. In the 11 bloods of our database, we do not know how much blood the 9 different professional athletes reinfused, but we know that these quantities are representative of current practices of some immoral athletes. Therefore, we can expect a sensitivity/specificity relation very close to the relation we obtained here (82% sensitivity without any false positive) on future bloods to test, for a doping practice that has challenged the scientific community since the early 1970's.

Finally, ABPS can be easily updated if new blood profiles measured with a standardized phlebotomy procedure from any WADA accredited laboratory is added to the database. For example, if a new blood doping method is discovered in the near

future, blood profiles induced by this method can be integrated as reference profiles to make ABPS sensitive to it. A fast reaction to deter blood doping in sport is therefore possible in this discriminative training procedure.

5 DISCUSSION: a probabilistic framework in anti-doping realm

We showed through 3 applications how validation techniques peculiar to discriminative pattern classification algorithms can be used to foster anti-doping research. In particular, these 3 applications suggest that statistical classification is a good alternative to provide the robustness and reliability needed for interpreting evidence, and, subsequently, facilitating the adherence to medico-legal standards mandated by WADA.

However, the probabilistic framework that we proposed here prevents decision making with 100% specificity. In other areas of forensic sciences, such as DNA profiling, transfer evidence (i.e. paint, fibers and glass fragments) and paternity cases, a probabilistic framework has become widely accepted for the evaluation of evidence (Aitken & Taroni, 2004). In these areas, the use of statistics has proven to facilitate the adherence to international medico-legal standards (Fienberg, 1989). Forensic scientists and jurists have abandoned the idea of "absolute certainty" and "exact science" in legal decision-making with scientific evidence. Those terms are essentially meaningless (Berger & Berry, 1988; Evett, 1996). Therefore, forensic scientists accompany their findings with appropriate explanations and thus contribute, in particular among non-experts, to avoid an unjustified impression of precision and conclusive force.

When experts are called upon to express an opinion on the value of evidence, only one form of reasoning could be applied which derives from the fundamental laws of probability. An effect has already been observed (e.g. blood parameters measured) and one wants to know the (posterior) probability that it is the result of a specified cause (e.g. blood manipulation). This is a line of reasoning against the causal direction and Bayes' theorem handles this point.

To obtain the posterior probability for each possible cause of an effect, the decision maker must multiply prior probabilities for each cause (e.g. the athlete has manipulated his blood, or the athlete has not manipulated his blood) by the probability of the effect given each cause (the so-called likelihood ratio). This fundamental approach calls for an estimation of the prior probabilities based on an evaluation of the background information of the case. In anti-doping area of investigation, the prevalence looks at first sight a good alternative for estimating prior probabilities. However, the forensic scientist is not in the same position as the decision maker. In fact, he lacks relevant information to assess prior probabilities. Therefore, one

of the views held in a significant part of forensic literature is that the value of the evidence is appropriately described by a likelihood ratio, which is thought to provide a transparent and logical measure of how well the evidence permits one to discriminate among competing hypotheses (Edwards, 1970).

In anti-doping area of investigation, the main advantages of a probabilistic approach are: 1) evidentiary values in indirect testing could be expressed through probability figures, 2) the probabilistic approach in evidence evaluation is uniform for all types of scientific evidence, 3) the scientific evidence should always be considered as one element to be incorporated with or corroborated by other elements in the case, such as testimony, medical records history, a particular physiological or pathological condition of the athlete, etc., and 4) the duties of decision makers (a judge) and experts (a scientist) are clear and distinct; scientists have to be asked to give their opinion on the evidence alone, whereas jurists should judge the alleged facts.

Moreover, a probabilistic framework is fully compatible with WADA code, as illustrated by this citation from the prohibitive list of substances and methods: "*Unless the athlete can demonstrate that the concentration was due to a physiological or pathological condition, a sample will be deemed to contain a prohibitive substance (as listed above) where the concentration of the prohibited substance or its metabolites and/or relevant ratios or markers in the athlete's sample so exceeds the range of values normally found in humans so that it is unlikely to be consistent with normal endogenous production*".

References

- Aitken, C., & Taroni, F. 2004. *Statistics and the evaluation of evidence for forensic scientists*. 2nd edn. Chichester: John Wiley & Sons.
- Ashenden, M. J. 2002. A strategy to deter blood doping in sport. *Haematologica*, **87**, 225–234.
- Ashenden, M. J., Sharpe, K., Damsgaard, R., & Jarvis, L. 2004. Standardisation of reticulocyte values in an antidoping context. *Am J Clin Pathol*, **121**(6), 816–825.
- Audran, M., Gareau, R., Matecky, S., Durand, F., Chenard, C., Sicard, M.T., Marion, B., & Bressolle, F. 1999. Effects of erythropoietin administration in training athletes and possible indirect detection in doping control. *Med Sci Sports Exercise*, **31**(5), 639–645.
- Azzalini, A., & Capitanio, A. 1999. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society B*, **61**, 987–989.
- Berger, J.O., & Berry, D.A. 1988. Statistical analysis and the illusion of objectivity. *Amer Scient*, **76**, 159–165.

- Berglund, B., P., Hemmingsson, & G., Birgegard. 1987. Detection of autologous blood transfusions in cross-country skiers. *Int. J. Sports Med.*, **8**(2), 66–70.
- Canu, S., Grandvalet, Y., & Rakotomamonjy, Y. 2003. *SVM and Kernel Methods Matlab Toolbox*.
- Casoni, I., Ricci, G., Ballarini, E., Borsetto, C., Grazi, G., Guglielmini, C., Manfredini, F., Mazzoni, G., Patracchini, M., & Paoli, E. V. De. 1993. Hematological indices of erythropoietin administration in athletes. *Int. J. Sports Med.*, **14**, 307–311.
- Cazzola, M. 2000. A global strategy for prevention and detection of blood doping with erythropoietin and related drugs. *Haematologica*, **85**, 561–563.
- Coombs, R.R.A. 1970. History and evolution of the anti-globin reaction and its application in clinical and experimental medicine. *Am. J. Clin. Path.*, **53**, 131–135.
- Duda, R.O., Hart, P.E., & Stork, D.G. 2001. *Pattern classification*. 2nd edn. New York: Wiley-Interscience.
- Edwards, A.W.F. 1970. Likelihood. *Nature*, **227**, 92.
- Eklom, B., & Berglund, B. 1991. Effects of erythropoietin administration on maximal aerobic power. *Scand J Med Sci Sports*, **1**, 88–93.
- Eklom, B., Goldberg, A., & Gullbring, B. 1972. Response to exercise after blood loss and reinfusion. *J. Appl. Physiol.*, **33**, 175–180.
- Evet, I.W. 1996. Expert evidence and forensic misconceptions of the nature of exact science. *Sci. and Just.*, **36**, 118–122.
- Fienberg, S.E. 1989. *The Evolving Role of Statistical Assessments as Evidence in the Courts*. Heidelberg: Springer.
- Friedman, N., Geiger, D., & Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning*, **29**, 131–163.
- Gaudard, A., Varlet-Marie, E., Bressolle, F., & Audran, M. 2003. Drugs for increasing oxygen and their potential use in doping: a review. *Sports Med*, **33**, 187–212.
- Gore, C.J., Parisotto, R., Ashenden, M.J., Stray-Gundersen, J., Sharpe, K., Hopkins, W., Emslie, K.R., Howe, C., Trout, G.J., Kazlauskas, R., & Hahn, A.G. 2003. Second-generation blood tests to detect erythropoietin abuse by athletes. *Haematologica*, **88**, 333–344.
- Guyon, I., & Elisseeff, A. 2003. An introduction to variable and feature selection. *J Mach Learn Res*, **3**, 1157–1182.
- Hand, D.J., & Yu, K. 2001. Idiot's Bayes - not so stupid after all? *Int Stat Review*, **69**, 385–399.

- Hansen, L.K., Larsen, J., & Kolenda, T. 2001. Blind Detection of Independent Dynamic Components. vol. 5.
- Hjorth, J.S.U. 1994. *Computer Intensive Statistical Methods Validation, Model Selection, and Bootstrap*. London: Chapman & Hall.
- Lasne, F., & de Ceaurriz, J. 2000. Recombinant erythropoietin in urine. *Nature*, **405**, 635.
- Nelson, M., Ashenden, M., Langshaw, M., & Popp, H. 2002. Detection of homologous blood transfusion by flow cytometry: a deterrent against blood doping. *Haematologica*, **87**, 881–882.
- Parisotto, R., Wu, M., Ashenden, M.J., Emslie, K.R., Gore, C.J., Howe, C., Kazlauskas, R., Sharpe, K., Trout, G.J., Xie, M., & Hahn, A.G. 2001. Detection of recombinant human erythropoietin abuse in athletes utilizing markers of altered erythropoiesis. *Haematologica*, **86**, 128–137.
- Robinson, N. 2003. *Blood manipulation in endurance sports*. Ph.D. thesis, University of Lausanne.
- Robinson, N., Saugy, M., Buclin, T., Gremion, G., & Mangin, P. 2001. The interpretation of secondary blood markers can get hazardous in case of a discontinuous rhEPO treatment. *Haematologica*, **6**, 87.
- Schoelkopf, B., & Smola, A.J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Boston: MIT press.
- Sharpe, K., Hopkins, W., Emslie, K.R., Howe, C., Trout, G.J., Kazlauskas, R., Ashenden, M.J., Gore, C.J., Parisotto, R., & Hahn, A.G. 2002. Development of reference ranges in elite athletes for markers of altered erythropoiesis. *Haematologica*, **87**, 1248–1257.
- S.Haykin. 1998. *Neural Networks: A Comprehensive Foundation*. 2nd edn. Upper Saddle River: Prentice Hall.
- Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Stray-Gundersen, J., Videman, T., Penttila, I., & Lereim, I. 2003. Abnormal Hematologic Profiles in Elite Cross-Country Skiers: Blood Doping or? *Clin J Sport Med*, **13**, 132–137.
- Vergouwen, P.C., Collee, T., & Marx, J.J. 1999. Hematocrit in elite athletes. *Int. J. Sports Med.*, **20**, 538–541.
- Videman, T., Lereim, I., Hemmingsson, P., Turner, M.S., Rousseau-Bianchi, M.P., Jenoure, P., Raas, E., Schonhuber, H., Rusko, H., & Stray-Gundersen, J. 2000. Changes in hemoglobin values in elite cross-country skiers from 1987 to 1999. *J. Scand J Med Sci Sports*, **10**, 98–102.
- Wide, L., Bengtsson, C., Berglund, B., & Ekblom, B. 1995. Detection in blood and urine of recombinant erythropoietin administrated to healthy men. *Med. & Sci. in Sports & Exerc.*, **27**, 1569–1576.