How does one teach anti-doping officials about evidence-based decision making?

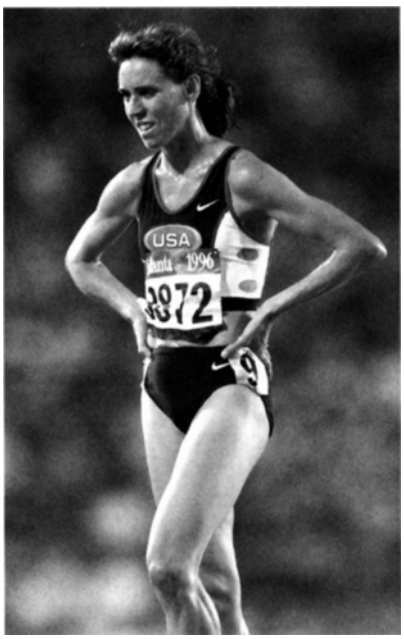
Inferences about Testosterone Abuse among Athletes

Donald A. Berry and LeeAnn Chastain

Drug testing of athletes has two purposes: to prevent artificial performance enhancement (known as doping) and to discourage the use of potentially harmful substances. Significant advancements in laboratory tests for doping have developed during the 30-year history of such organized efforts, and competitors at the Olympic Games of 2002 benefited from such scientific advancements. However, developing a good test is only part of an effective and fair process.

As is well known in medicine, the outcome of a test never plays a solo role in diagnosis. It represents only one factor to be considered. Physicians arrive at a diagnosis by weighing the likelihood that a test result correctly identifies a disorder or disease in the context of other relevant information. Such information includes the patient's symptoms, if any, the patient's characteristics, such as age and gender, prior history of screening and disease, etc. The characteristics of the test procedure must also be taken into account, including its sensitivity, specificity, and inherent variability. In the best of worlds, physicians start from a set of differential diagnoses based on what is known at that point in time. They may plan further tests, depending on the results of the first test, and they update the likelihood of each of the possible diagnoses as new information becomes available. Simply put, a medical diagnosis is a statistical inference.

Similarities between the processes of diagnosing a complex medical disorder and determining whether or not an athlete's test result is indicative of doping are especially strong when the alleged infraction is the use of a substance that is a normal physiologic component, such as testosterone. In interpreting diagnostic tests, however, physicians have an inferential advantage: if a patient's sample shows an elevated marker that might indicate a serious disease, the truth becomes known when the patient turns out to have or not to have the disease. This contributes to databases of information from both diseased and healthy individuals. Whether or not an athlete has abused a substance will rarely become known by other objective evidence. There is no "gold standard" when testing for substance abuse.



Mary Decker Slaney at the 1996 Summer Olympics. (AP Photo/Doug Mills)

The appropriate tool for decision making under uncertainty is Bayes' rule. Its contribution to evidence-based medicine has been described by many authors, and is thoroughly supported in two articles by Steven N. Goodman, published in the *Annals of Internal Medicine* in 1999. The application of Bayes' rule is well understood in medicine, and is appropriate for anti-doping science, as well. Failure to recognize relevant statistical issues when making inferences from laboratory tests leads to flawed conclusions in the case of an accusation of

testosterone use. These issues will be discussed in the context of the case of the USA Track and Field (USATF) and runner Mary Slaney.

Mary Decker Slaney was a world-class runner for more than 20 years. She was the 1983 world champion at 1,500 and 3,000 meters, and held the fastest times for American women in distances from 800 to 10,000 meters. At the time of her induction into the USATF Hall of Fame in 2003—nearly 20 years after she captured these records—Slaney still held the North and Central American women's record for outdoor track at 1 mile, 1,500 meters, and at 3,000 meters. She also continued to hold the 1-mile record for this world region on an indoor track. In the mid-1980s, Slaney underwent treatment for exercise-induced asthma, in addition to multiple surgeries to repair bones in her left foot and remove degenerated tissue from her Achilles tendon. Rehabilitation and a return to elite competition were slow, but in 1996, at the age of 37, Slaney was making a comeback in distance running. Her efforts were halted by a suspension due to alleged doping consistent with testosterone use. In this article we raise questions about the appropriateness of such suspensions.

Performance Enhancement and the Anti-Doping Movement

Combating doping among athletes is a worldwide endeavor, involving the establishment in 1999 of a World Anti-Doping Agency (WADA) to work with the Olympic committees and sports federations. More than 25 accredited laboratories are involved in the analysis of over 90,000 athletes' urine samples annually, at a cost of millions of dollars.

Athletes may be tested for all forms of doping, and face sanctions ranging from several months of suspension from competition to permanent ineligibility. The finding of a foreign, banned substance or its metabolites in a bodily sample is legal evidence of doping, and is a relatively straightforward inference. A rather different circumstance exists, however, when the substance in question is not foreign, but is normally present in the athlete's body. Such is the case with testosterone.

Scientists have attempted to detect artificial increases in testosterone concentrations through the establishment of a "normal urinary range" for the ratio of testosterone glucuronide to epitestosterone glucuronide (known as the T/E ratio). Human physiologic processes do not convert testosterone to epitestosterone, its 17 alpha-epimer, and doping scientists have found that the urinary concentrations of the two analytes are usually about the same. Urinary testosterone concentrations are known to fluctuate significantly in males and females. And little is known about epitestosterone because it has not been studied clinically as an indicator of health or disease. Despite these limitations, following the 1980 Olympic Games, doping scientists set the threshold for positive testosterone doping at a T/E ratio greater than 6:1. (Descriptions of the complicated methods used to measure testosterone and epitestosterone concentrations in urine have been published elsewhere; some references are listed at the end of this article.)

Biological Issues

Many issues associated with urinary excretion of testosterone and epitestosterone are poorly understood, especially in the female. Research is limited, and some authors have urged caution when assessing the female T/E profile. Variations in steroid profiling (resulting in altered T/E ratio) are associated with ethnicity; age; sex; circadian rhythm; training and competition; diet; nutritional supplementation; environmental factors; alcohol ingestion; enzyme deficiencies; decreased epitestosterone excretion; menstruation; pregnancy; birth control pills; other hormonal therapy; consumption of meat from animals supplemented with anabolic steroids; polycystic ovary syndrome (a common endocrine disorder); and other pathologic medical conditions.

Christiane Ayotte, Professor and Director of the Montreal Doping Control Laboratory, member of WADA's Health, Medical and Research Committee, and expert witness for the USATF at Mary Slaney's hearing, reported in her research studies that some women exhibit physiologic T/E ratios greater than 6. Professor Ayotte's studies were published in the International Amateur Athletic Federation's (IAAF) *New Studies in Athletics*, 1997. Another expert witness for the USATF at Mary Slaney's hearing was Don H. Catlin, Director of the Olympic Analytical Laboratory at UCLA. In an article published in *Clinical Chemistry* in 1997, Catlin reported that the T/E ratio above which could not be attributed to normal physiologic variations was a mean ratio of 15. Mary Slaney's T/E ratios, which resulted in her suspension, were 9.5 and 11.6.

The Case of Mary Decker Slaney

Mary Slaney submitted a routine urine sample at the 1996 U.S. Olympic Trials in Atlanta. Urinalysis at the accredited laboratory in Montreal, Canada, indicated that her T/E ratio exceeded the upper limit of 6.

Sports officials regarded this to be proof of testosterone doping. However, in September 1997 a Doping Hearing Board of the USATF exonerated Slaney. (The Hearing Board was presented with the arguments given in this article by one of the authors [DAB].) In April 1999 the IAAF (known officially as the International Association of Athletics Federations as of 2001), exerting its stated right to overrule a member federation, ruled that Slaney "was guilty of a doping offense . . . and that the USATF's Doping Hearing Board decision . . . was erroneous." Slaney filed a suit against the IAAF and the U.S. Olympic Committee (USOC) in U.S. District Court, seeking to reverse this ruling and to obtain an injunction against the use of the T/E ratio on women athletes until a better understanding of the causes of altered ratios could be reached. The decision from the court, issued on March 27, 2001, was that it did not have subject matter jurisdiction over the state-law claims against the USOC. The court ruled that it was obligated to recognize the valid arbitration in which Slaney had participated with the IAAF, and that although arbitrators were not bound by rules of evidence, the issue decided in the arbitration could not be relitigated.

Anti-doping officials behave as though a T/E ratio greater

than 6 documents illegal steroid use to the exclusion of other possible explanations. This implication was made explicit by the IAAF in the April 1999 ruling declaring Mary Slaney's wins and achievements from June 17, 1996, through June 16, 1998, null and void. Such a conclusion based solely on this indirect laboratory measurement is inconsistent with common clinical or forensic practice.

False Positives

Any laboratory test can give false positive results, whether because of inherent variability in the testing procedure, inherent biologic variability, or inappropriate handling of source material. If 1% of nonusers are found to have a urinary T/E ratio above the established threshold of 6:1 (a figure consistent with at least some of the testimony presented in the Slaney case), then an average of one of 100 athletes' tests will result in a positive reading even if none have used an illegal substance. If 90,000 urine samples are tested (the annual rate referenced previously) then 900 will test positive, even if none of the athletes used illegal substances. Using a more stringent cut-point to identify only 0.1% of users will still incorrectly identify 90 nonusers.

A standard statistical practice is to make adjustments in cut-points to account for the fact that many tests were carried out. To ensure that the probability of one or more false positives in a program of 100 tests is less than 1% a particular individual's test result would have to be in the upper 99.99th percentile of nonusers tested. That is, no more than 1 in 10,000 nonusers tested should have a T/E ratio above the cut-point. If 90,000 urine samples are tested, then the appropriate cut-point is the 99.99999th percentile. The implications are profound. And there are possibly insurmountable problems in applying such a standard. One is statistical: such an extreme percentile is not even approximately known. Mounting a study to identify a value of T/E for which only one in 10 million nonusers tests higher would be an enormous undertaking.

The trade-off between finding an innocent person guilty and protecting the integrity of sport is far from obvious. It is a societal issue beyond the scope of this article. But it seems unreasonable to impugn 1% or even a tenth of a percent of nonusers who happen to be tested.

Sensitivity and Specificity of Laboratory Tests

The diagnostic accuracy of any laboratory test is defined as the ability to discriminate between two types of individuals—in this case users and nonusers. *Specificity* and *sensitivity* characterize diagnostic tests. In the anti-doping analogy regarding the T/E ratio, specificity is the proportion of nonusers with $T/E < 6$ (as discussed above) and sensitivity is the proportion of users with $T/E \geq 6$. Estimating these proportions requires collecting and tabulating data from the two reference samples, users and nonusers, as well as data from the appropriate subpopulation for that athlete. ("Users" will, of course, vary in dose, duration, and substance used. Bayes' rule allows for learning about the degree of use on the basis of test results, but we do not consider this more general setting here.)

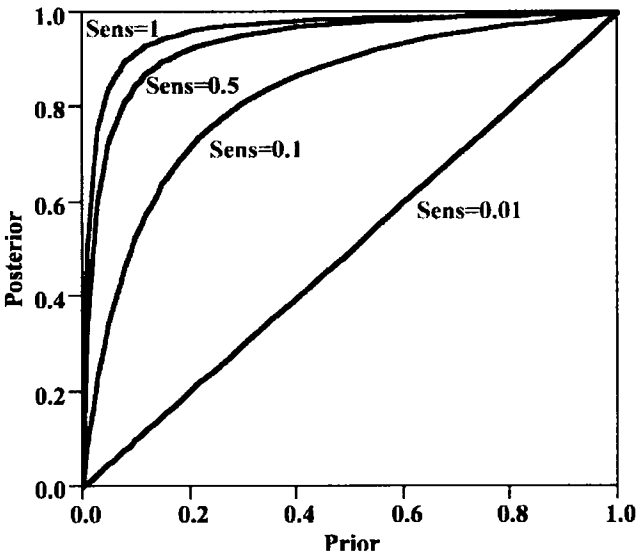


Figure 1. Relationship between posterior probability and prior probability for four different values of sensitivity. All curves assume specificity is 99%.

Sensitivity must be addressed in the context of the effect of a particular banned substance upon the T/E ratio. In an article published in the *Journal of Mass Spectrometry* in 2000, David J. Borts and Larry D. Bowers, studying the use of advanced laboratory methods to measure the T/E ratio, stated that the "measurement of a prohibited substance [in the urine matrix] requires identification of the compound either by full scan or by consistency of ion abundance ratios between a reference material and the unknown" (p. 56). Anti-doping laboratories, however, fail to address sensitivity or to recognize its relevance. Specificity, generally addressed by keeping records of tests performed on athletes, introduces a clear bias of unknown magnitude because some athletes in their databases may be users but have "normal" T/E. For example, Asian males have baseline urine concentrations of testosterone and T/E that can be so low that testosterone doping may not elevate their T/E to that of some nonusers among Caucasian athletes.

Bayes' Rule and the Probability of Guilt

Bayes' rule is a necessary inferential tool for relating experimental evidence to conclusions, such as whether someone has a disease or has used a particular substance. Applying Bayes' rule requires determining the test's sensitivity and specificity. It also requires a pre-test (or prior) probability that the athlete has used a banned substance.

Consider a population of 1,000 athletes with the following characteristics: 100 are users (prior probability = 10%), and of these, 50 would test positive (sensitivity = 50%). Of the 900 nonusers, nine would test positive (specificity = 99%). If an athlete tests positive, then the probability she is a user is 50/59 = 84.7% (positive predictive value, or PPV).

Where "prior" indicates the prior probability of being a user, Bayes' rule says

$$\mathrm{PPV} = \frac{\mathrm{sens} \times \mathrm{prior}}{\mathrm{sens} \times \mathrm{prior} + (1 - \mathrm{spec}) \times (1 - \mathrm{prior})}$$

In the above example with 1,000 athletes:

$$\frac{50}{59} = \frac{(50/100) \times (100/1000)}{(50/100) \times (100/1000) + (9/900) \times (900/1000)}$$

If the sensitivity is 1 (in which case all users have positive tests) and the other factors are unchanged, then the PPV is

$$\frac{1 \times 0.1}{1 \times 0.1 + 0.01 \times 0.9} = \frac{100}{109} = 91.7\%$$

Bayes' rule relates "inverse probabilities": the probability of substance abuse given the test result, with the probability of the test result given substance abuse. But these two probabilities are not equal. Nor is the probability of substance abuse given the test result (posterior probability) the same as the probability of testing negative given that the athlete is a nonuser (specificity). In the example, the latter is 99% and the former is 84.7%.

An essential factor in Bayes' rule is the prior probability of disease, or guilt of substance abuse, depending on the application. In the above example, changing the prior probability from 10% to 90% means changing the posterior probability from 84.7% to 450/451 = 99.8 %.

Figure 1 shows the relationship between posterior probability, prior probability, and sensitivity, for a fixed specificity of 99%. Such a figure allows for using Bayes' rule in reverse. For example, a hearing board might not want to rule against an athlete unless her probability of guilt is at least 95%. The prior probability must be at least 20%, even for perfect sensitivity. And if the standard is 99%, then the prior probability must be at least 50%.

Prevalence of disease is relatively easy to estimate, depending on the patient population. Prevalence of substance abuse is not. There is an inevitable subjective aspect of assigning a prior probability. A hearing board made up of an athlete's peers is especially appropriate for making such assignments. For example, assuming they know nothing about the athlete beyond the information presented at the hearing, they might regard testosterone substance abuse to be rare and so the positive predictive value would be at most moderately large. On the other hand, if they know testosterone substance abuse to be widespread, then the probability of abuse would be larger, based on a positive test.

Conclusion

Conclusions about the likelihood of testosterone doping require consideration of three components: specificity and sensitivity of the testing procedure, and the prior probability of use. As regards the T/E ratio, anti-doping officials consider only specificity. The result is a flawed process of inference. Bayes' rule shows that it is impossible to draw conclusions about guilt on the basis of specificity alone. Policy-makers in the athletic federations should follow the lead of medical scientists who use sensitivity, specificity, and Bayes' rule in interpreting diagnostic evidence.

Judging whether Mary Slaney was guilty as alleged is not the point of this article. Our focus has been the flawed inferential process used in such cases. Wonderfully modern technological advances in the laboratory should not be marred by erroneous statistical interpretations. In the short term, the bar for urinalysis of testosterone doping must be set much higher (T/E much higher than 6). The threshold should be determined in light of the number of athletes to be tested, pegged to the number of false positives that are acceptable to society, and based on the limitations of current laboratory methods. If the sizes of current databases are too small to allow for good estimates of proportions above large thresholds, then the laboratory test should be suspended until such information becomes available.

Additional Reading

American Medical Association. 1987. "Scientific Issues in Drug Testing." *Report J of the Council on Scientific Affairs, JAMA*; 257(22):3110-3114.

Berry, D.A. 1991. "Inferences Using DNA Profiling in Forensic Identification and Paternity Cases." (with discussion) *Stat Science*; 6:175-205.

Berry, D.A. 1994. "DNA, Statistics and the Simpson Case." *Chance*; 7(4):9-12.

Berry, D.A., Hochberg, Y. 1999. "Bayesian Perspectives on Multiple Comparisons." *J Stat Planning Inference*; 82:215-227.

Black, D.L. 2001. "Doping Control Testing Policies and Procedures: a Critique." In Wilson, W., Derse, E., eds. *Doping in Elite Sport: The Politics of Drugs in the Olympic Movement*. Champaign, Ill: Human Kinetics Pub; 29-42.

Kammerer, R.C. 2001. "What is Doping and How is it Detected?" In Wilson, W., Derse E., eds. *Doping in Elite Sport: The Politics of Drugs in the Olympic Movement*. Champaign, Ill.: Human Kinetics Pub; 3-28.

Mary Decker Slaney v. The International Amateur Athletic Federation and the United States Olympic Committee. 244 F.3d 580 (U.S. 7th Cir. 2001).

Sackett, D.L. 1997. *Evidence-Based Medicine: How to Practice and Teach EBM*. New York: Churchill Livingstone.

Strauss, R.H., Liggett, M.T., Lanese, R.R. 1985. "Anabolic Steroid Use and Perceived Effects in Ten Weight-trained Women Athletes." *JAMA*; 253(19):2871-2873.

van de Kerkhof, D.H., de Boer, D., Thijssen, J.H.H., Maes, R.A.A. 2000. "Evaluation of Testosterone/ epitestosterone Ratio Influential Factors as Determined in Doping Analysis." *J Anal Toxicol*; 24:102-115.

Zweig, M.H., Campbell, G. 1993. "Receiver-operating Characteristic (ROC) Plots: a Fundamental Evaluation Tool in Clinical Medicine." *Clin Chem*; 39(4):561-577.