

## COMMENTARY

## The science of doping

The processes used to charge athletes with cheating are often based on flawed statistics and flawed logic, says **Donald A. Berry**.

Recently, the international Court of Arbitration for Sport upheld doping charges against cyclist Floyd Landis, stripping him of his title as winner of the 2006 Tour de France and suspending him from competition for two years. The court agreed with the majority opinion of a divided three-member American Arbitration Association (AAA) panel and essentially placed a stamp of approval on a laboratory test indicating that Landis had taken synthetic testosterone. Although Landis asserts his innocence, his options for recourse have all but dried up.

Already, in the run-up to this year's Olympic Games, vast amounts of time, money and media coverage have been spent on sports doping. Several doping experts have contended that tests aren't sensitive enough and let dozens of cheaters slip through the cracks. And some athletes are facing sanctions. Upon testing positive for clenbuterol, US swimmer Jessica Hardy was held back from the Olympic team and faces a two-year ban from the sport. She is attesting her innocence. China has already banned several athletes, some of them for life, on doping charges. Indeed, many world-class athletes will find their life's accomplishments and ambitions, their integrity and their reputations hinging on urine or blood tests. But when an athlete tests positive, is he or she guilty of doping? Because of what I believe to be inherent flaws in the testing practices of doping laboratories, the answer, quite possibly, is no.

In my opinion, close scrutiny of quantitative evidence used in Landis's case show it to be non-informative. This says nothing about Landis's guilt or innocence. It rather reveals that the evidence and inferential procedures used to judge guilt in such cases don't address the question correctly. The situation in drug-testing labs worldwide must be remedied. Cheaters evade detection, innocents are falsely accused and sport is ultimately suffering.

### Prosecutor's fallacy

One factor at play in many cases that involve statistical reasoning, is what's known as the prosecutor's fallacy<sup>1</sup>. At its simplest level, it concludes guilt on the basis of an observation that would be extremely rare if the person were innocent. Consider a blood test that perfectly matches a suspect to the perpetrator of a crime. Say, for example, the matching profile occurs in just 1 out of every 1,000 people.



Floyd Landis (centre) was disqualified after winning the 2006 Tour de France.

A naive prosecutor might try to convince a jury that the odds of guilt are 999:1, that is, the probability of guilt is 0.999. The correct way to determine odds comes from Bayes rule<sup>2-4</sup> and is equal to 999 times  $P/(1-P)$  where  $P$  is the 'prior probability' of guilt. Prior probability can be difficult to assess, but could range from very small to very large based on corroborating evidence implicating the suspect. The prosecutor's claim that the odds are 999:1 implies a prior probability of guilt equal to 0.5 (in which case  $P$  and  $1-P$  cancel). Such a high value of  $P$  is possible, but it would require substantial evidence. Suppose there is no evidence against the suspect other than the blood test: he was implicated only because he was from the city where the crime occurred. If the city's population is one million then  $P$  is 1/1,000,000 and the odds of his guilt are 1001:1 against, which corresponds to a probability of guilt of less than 0.001.

The prosecutor's fallacy is at play in doping cases. For example, Landis's positive test result seemed to be a rare event, but just how rare? In doping cases the odds are dictated by the relative likelihood of a positive test assuming the subject was doping ('sensitivity') against a positive result assuming no doping (which is one minus 'specificity'). Sensitivity and specificity are crucial measures that must be estimated

with reasonable accuracy before any conclusion of doping can be made, in my opinion.

The studies necessary to obtain good estimates are not easy to do. They require known samples, both positive and negative for doping, tested by blinded technicians who use the same procedures under the same conditions present in actual sporting events. In my view, such studies have not been adequately done, leaving the criterion for calling a test positive unvalidated.

### Laboratory practices

Urine samples from cyclists competing in the 2006 Tour de France were analysed at the French national anti-doping laboratory (LNDD) in Châtenay-Malabry. This is one of 34 laboratories accredited by the World Anti-Doping Agency to receive and analyse test samples from athletes. The LNDD flagged Landis's urine sample following race stage 17, which he won, because it showed a high ratio of testosterone to epitestosterone.

Based on the initial screening test, the LNDD conducted gas chromatography with mass spectrometry, and isotope ratio mass spectrometry on androgen metabolites in Landis's sample. Such laboratory tests involve a series of highly sophisticated processes that are used to identify the likelihood of abnormal levels of plant-based androgen metabolites (from dietary or

pharmaceutical sources) in a urine sample. The goal is to differentiate from endogenous androgen metabolites normally found in urine.

Mass spectrometry requires careful sample handling, advanced technician training and precise instrument calibration. The process is unlikely to be error-free. Each of the various steps in handling, labelling and storing an athlete's sample represents opportunity for error.

In arbitration hearings, the AAA threw out the result of the LNDD's initial screening test because of improper procedures. In my opinion, this should have invalidated the more involved follow-up testing regardless of whether or not sensitivity and specificity had been determined. Nevertheless, the AAA ruled the spectrometry results sufficient to uphold charges of doping.

During arbitration and in response to appeals from Landis, the LNDD provided the results of its androgen metabolite tests for 139 'negative' cases, 27 'positive' cases, and Landis's stage 17 results (see Fig. 1). These data were given to me by a member of Landis's defence team. The criteria used to discriminate a positive from a negative result are set by the World Anti-Doping Agency and are applied to these results in Fig. 1b and d. But we have no way of knowing which cases are truly positive and which are negative. It is proper to establish threshold values such as these, but only to define a hypothesis; a positive test criterion requires

further investigation on known samples.

The method used to establish the criterion for discriminating one group from another has not been published, and tests have not been performed to establish sensitivity and specificity. Without further validation in independent experiments, testing is subject to extreme biases. The LNDD lab disagrees with my interpretation. But if conventional doping testing were to be submitted to a regulatory agency such as the US Food and Drug Administration<sup>5</sup> to qualify as a diagnostic test for a disease, it would be rejected.

### The problem with multiples

Landis seemed to have an unusual test result. Because he was among the leaders he provided 8 pairs of urine samples (of the total of approximately 126 sample-pairs in the 2006 Tour de France). So there were 8 opportunities for a true positive — and 8 opportunities for a false positive. If he never doped and assuming a specificity of 95%, the probability of all 8 samples being labelled 'negative' is the eighth power of 0.95, or 0.66. Therefore, Landis's false-positive rate for the race as a whole would be about 34%. Even a very high specificity of 99% would mean a false-positive rate of about 8%. The single-test specificity would have to be increased to much greater than 99% to have an acceptable false-positive rate. But we don't know the single-test specificity because the

appropriate studies have not been performed or published.

More important than the number of samples from one individual is the total number of samples tested. With 126 samples, assuming 99% specificity, the false-positive rate is 72%. So, an apparently unusual test result may not be unusual at all when viewed from the perspective of multiple tests. This is well understood by statisticians, who routinely adjust for multiple testing. I believe that test results much more unusual than the 99th percentile among non-dopers should be required before they can be labelled 'positive'.

Other doping tests are subject to the same weak science as testosterone, including tests for naturally occurring substances, and some that claim to detect the presence of a foreign substance. Detecting a banned foreign substance in an athlete's blood or urine would seem to be clear evidence of guilt. But as with testing for synthetic testosterone, such tests may actually be measuring metabolites of the drug that are naturally occurring at variable levels.

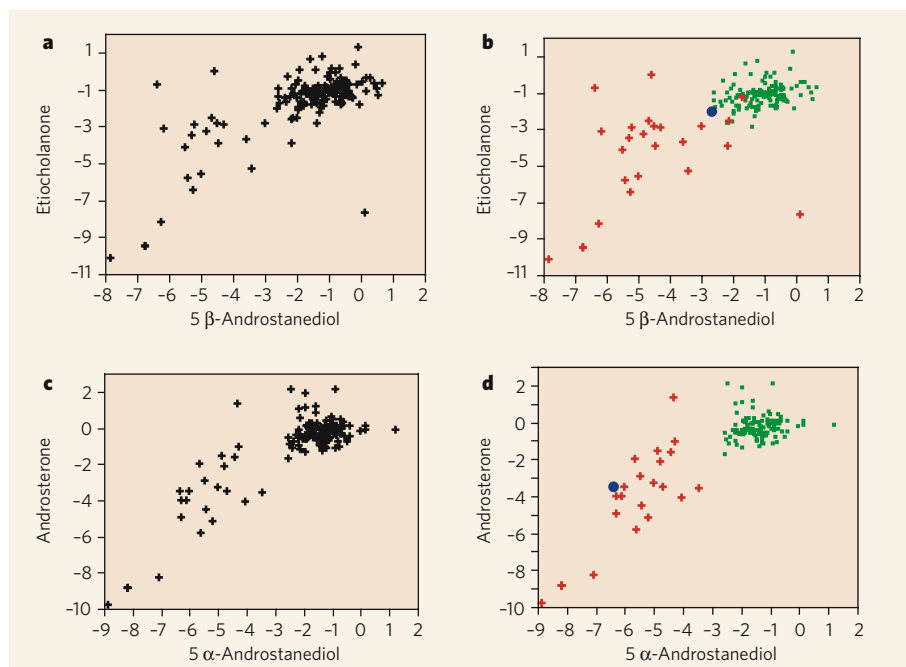
Whether a substance can be measured directly or not, sports doping laboratories must prospectively define and publicize a standard testing procedure, including unambiguous criteria for concluding positivity, and they must validate that procedure in blinded experiments. Moreover, these experiments should address factors such as substance used (banned and not), dose of the substance, methods of delivery, timing of use relative to testing, and heterogeneity of metabolism among individuals.

To various degrees, these same deficiencies exist elsewhere — including in some forensic laboratories. All scientists share responsibility for this. We should get serious about interdisciplinary collaborations, and we should find out how other scientists approach similar problems. Meanwhile, we are duty-bound to tell other scientists when they are on the wrong path. ■

**Donald A. Berry** is head of the Division of Quantitative Sciences, chair of the Department of Biostatistics and Frank T. McGraw Memorial Chair of Cancer Research, MD Anderson Cancer Center, University of Texas, 1400 Pressler Street, Houston, Texas 77030-1402, USA.  
e-mail: [dberry@mdanderson.org](mailto:dberry@mdanderson.org)

1. Buchanan, M. The prosecutor's fallacy. *The New York Times* (16 May 2007).
2. Berry, D. A. *Stat. Sci.* **6**, 175–205 (1991).
3. Berry, D. A. *Statistics: A Bayesian Perspective* (Duxbury Press, California, 1996).
4. Berry, D. A. & Chastain, L. A. *Chance* **17**, 5–8 (2004).
5. <http://www.fda.gov/cdrh/osb/guidance/1620.pdf>

**Donald Berry testified for the defence team of 1996 Olympian Mary Decker Slaney before a doping hearing board in 1997. He was paid for his time. See Editorial, page 667.**



**Figure 1 | Metabolite data.** Plots show the distribution of 167 samples of the metabolites etiocholanone and 5  $\beta$ -androstanediol (a, b), and androsterone and 5  $\alpha$ -androstanediol (c, d). Panels b and d show samples the French national anti-doping laboratory (LNDD) designate to be 'positive' (red crosses) or 'negative' (green dots); the values from Landis's second sample from stage 17 is shown as a blue dot. Axes display delta notation, expressing isotopic composition of a sample relative to a reference compound.