Can we trust doping tests?

In the hunt for doped athletes, little account is taken of the risk of condemning an innocent athlete. The doping hunters appear to be more concerned with catching all the guilty than with considering the risk that some test results may be false positives.

In 2010, a Norwegian athlete (Erik Tysse) tested positive for the banned substance CERA (Continuous Erythropoietin Receptor Activator), according to Laboratorio Anti-doping FMSI in Rome. The laboratory maintained that the substance had been found in the athlete's urine. The method used was isoelectric focusing (IEF). Both the A and the B sample were reported to test positive, both only after repeated analyses. The case was treated by the Norwegian Confederation of Sports' Prosecution Committee, which decided to report it to the Adjudication Committee. Tysse was found guilty, but he appealed. In the following, we will present some reflections on doping tests in general and the testing in this case in particular.

Biological tests

Biological tests can be characterised in terms of two different properties: reproducibility and validity. Reproducibility is the ability of the test method to produce the same result in repeated measurements. Measurements are subject to two different sources of error: random and systematic. It is important that these errors, which are often unavoidable, are minimised.

Systematic errors may be unimportant. It is of little importance whether there is systematic error associated with cuff measurement of intra-arterial blood pressure performed according to standard medical practice since all clinical tests with risk analyses and treatment efficacy are performed using cuff measurement.

Efforts are made when performing doping tests to take account of random error by analysing both an A sample and, if this tests positive, a B sample. The A and B samples are identical; the original sample is simply divided between two bottles. The two samples are analysed using the same test method, but sometimes an additional analysis is made of the B sample. This naturally gives an inadequate basis for estimating reproducibility. Two samples are not sufficient for this purpose. The B sample is also tested if the A sample tests positive. This creates expectations with respect to what should be found, which can create a bias. For example, the head of the laboratory in Rome is reported to have said that it would be a scandal if the A sample was not confirmed with a positive B sample. If the B sample did not confirm the positive A sample, 25 people at the laboratory in Rome would lose their jobs (1).

Another factor is that repetition with a B sample is only a control for certain types of error, i.e. those associated with the laboratory's ability to carry out the test correctly. If the reason for the test result is that the athlete has unusual values due to aspects of his or her physiology or intake of nutrients, neither an A nor a B sample will result in higher precision. Nor will any contamination or mix-up of samples necessarily be detected.

The doping hunters assume that if both A and B samples are positive, it is a foregone conclusion that the athlete is guilty. This is unsatisfactory as a matter of principle. Perfect reproducibility is not enough if the test lacks validity.

Test validity

Validity concerns whether the test results say something about what we want to know. Does a mammogram say anything about the occurrence of breast cancer? Does a tuberculosis test say anything about whether a person is infected with *Mycobacterium tuberculosis*?

Validity is expressed by sensitivity and specificity. Sensitivity is an expression of the ability of the test to yield a positive result if the disease in question is present or if doping has taken place. A sensitivity of 0.9 will mean that one of ten persons with the disease is not observed or identified. This is called a «false negative». Specificity is an expression of the ability of the test to yield a negative result for persons who do not have the disease or who are not doped. A specificity of 0.9 will mean that one of ten tests will conclude that the disease/substance is present in individuals who are not sick or doped. These are called «false positives».

It is important that both sensitivity and specificity are as close to 100 per cent as possible. But it is important to recognise that attempts to increase the sensitivity of a test will reduce specificity, and vice versa. There is no *a priori* optimal balance between the two. The optimal balance depends on the nature of the disease or of the condition and the relative costs associated with them. A very high sensitivity is important if there is an effective treatment for a serious disease. In doping work, it is specificity that is important. For some doped athletes to slip through the net is a less serious matter than the trauma occasioned by the conviction of an athlete who is not doped. A very high specificity and the possibility of documenting it with statistically satisfactory empiri are therefore crucial legal safeguards. Without knowing the specificity, it is fundamentally impossible to conclude whether disease (or in this case a forbidden substance) is present.

If, for example, 90 per cent of the sick persons test positive, can we then conclude that 90 per cent of those testing positive are sick? It is not uncommon to hear a conclusion of this kind. But it is not correct, and we demonstrate this in Table 1. Let us assume we have the following values: sensitivity 0.9, specificity 0.9, pre-test probability (prevalence) 0.01.

The pre-test probability of a disease is the assumed prevalence of the disease in the population being tested, a probability based on previous experience and knowledge. Given these figures, 100 persons out of a population of 10 000 would be sick. Of those who are sick, 90 test positive. Of those who are healthy, 990 test positive. Of the 1 080 who test positive, there are 90 who are sick – i.e., only 8.3 per cent of those who test positive are sick.

What happens if we assume a pre-test probability that is lower, for example 0.005? Table 2 shows that the positive predictive value of the test is then reduced from 8.3 per cent to 4.3 per cent.

The probability of a positive test indicating disease is thus dependent on the prevalence of the disease in the population being tested. This means, for example, that the probability of a given mammogram shadow indicating breast cancer for a randomly selected group of women in their 50s is different from that for a group of women who have a family history of breast cancer.

The doping case in question

In the case in question, the biggest problem is that we do not know the validity of the test. The validity of the CERA test has been estimated recently (2). It shows a clear interplay between sensitivity and specificity and demonstrates a clear possibility of a false positive result, depending on the threshold that is chosen for defining positivity. However, this analysis applies to use of an ELISA method, a different type of analysis from that used in the case in question (IEF).

The accused's counsel has attempted repeatedly to obtain validity values for the CERA test method that was used, but in vain (Mr Kjenner, Advocate, personal communication). A letter of 29 October 2010 to

Mr Kjenner from Professor Hemmersbach, Director of the Norwegian Doping Control Laboratory states: «WADA is not aware of any false positive case reported for CERA using the IEF method.» It is an interesting reply, in that it is an acknowledgement that not even WADA knows the validity of the CERA test.

None of the references supplied by Hemmersbach deal with the question of false positivity. We therefore have to limit ourselves to simulating values. The sensitivity and specificity values are given in Table 3. The pre-test probability is 0.001, i.e. one of 1 000 athletes is expected to be doped. The estimate is based on figures from the records of the Adjudication Committee (case 25/10) which state: «The Rome laboratory reports that it has performed about 5000 analyses, six of which have tested positive for CERA.» The values in Table 3 show that there is a less than 50 per cent probability of a positive test indicating actual doping.

The case documents also reveal that the accused could document a so-called normal blood profile (3). It is a known fact that if an athlete's blood profile is suspicious, that alone is a strong indicator of actual doping. This means that if the test population consists of athletes with normal blood profiles, the pre-test probability of positivity will be substantially lower than in an unselected population. The positive predictive value will then be even lower. In Table 3, example 2, we have simulated the supposed prevalence for such a group of athletes as 0.0005. The positive predictive value will then be 32.2 per cent, which is far from constituting a basis for a conviction.

On the other hand, let us assume a pre-test probability of 25 per cent (for example cyclists with suspicious blood profiles). Table 3, example 3: Here the positive predictive value of 99.68 will be sufficient for a conviction.

Table 1 Relationship between sensitivity (0.9), specificity (0.9), prevalence (0.01) and positive predictive value in a population of 10 000

	Test +	Test –	Total
Condition +	90	10	100
Condition –	990	8 910	9 900
Total	1 080	8 920	10 000
Positive predictive value	0.083		

Table 2 Relationship between sensitivity (0.9), specificity (0.9), prevalence (0.005) and positive predictive value in a population of 10 000

	Test +	Test –	Total
Condition +	45	5	50
Condition –	995	8 955	9 950
Total	1 040	8 960	10 000
Positive predictive value	0.043		

Table 3 Relationship between sensitivity, specificity, prevalence and positive predictive value. Current doping case

Example	Sensitivity	Specificity	Prevalence	Positive predictive value
1 General test population	0.95	0.999	0.001	0.487
2 Normal blood profile	0.95	0.999	0.0005	0.322
3 Suspicious blood profile	0.95	0.999	0.25	0.9969

Conclusion

When it comes to the case in question, we must express concern: From the Adjudication Committee's ruling it is clear that they are not certain of the facts of the case and that their decision is based on a perception that the laboratory has followed the WADA's procedures. The chairman of the Adjudication Committee, Lars Erik Frisvold, said quite frankly during the verbal proceedings that much of the discussion that took place was over his head. Is it acceptable adherence to due process of law that the Adjudication Committee, lacking its own expertise, blindly accepts WADA's procedures?

For doping tests, as for all other biological tests, knowledge of the validity of the test is fundamental. Validity is expressed through sensitivity and specificity. Whether sensitivity or specificity is the more important depends on the condition the test is intended to reveal. Both are important for doping testing. It is important to catch all cheats, but it is also important to avoid convicting innocent persons. Perhaps it is better to let ten doped athletes go free than to convict one who is innocent.

In this article we have concentrated on doping with CERA, and have had to conclude that the control system has drawn its conclusions without being able to document the validity of the test. It appears that it is assumed a priori that specificity is 100 per cent for a combination of A and B samples, which would exclude the possibility of false positive results. We have shown that lack of documentation of the validity of the test makes a conviction unreasonable and may result in a miscarriage of justice. The necessity of documenting the validity of tests applies to all doping tests – to ensure legal protection for athletes.

Hans Th. Waaler
hanswaaler@gmail.com
Harald Siem
Odd O. Aalen

Hans Th. Waaler (born 1926) has the titles of PhD and Professor Emeritus.
Conflicts of interest: None declared.

Harald Siem (born 1941) is a medical doctor, has an MSc in public health, and is employed in the Department of Global Health, Norwegian Directorate of Health. He has been district doctor for the municipality of Aukra and has worked at the Institute for General Medicine at Oslo University, on the Oslo Board of Health, for the Employers' Association, and for ten years with global health in Geneva.
Conflicts of interest: None declared.

Odd O. Aalen (born 1947) is professor at the Department of Biostatistics, University of Oslo.
Conflicts of interest: None declared.

Bibliography
1. Protokoll fra domsutvalget i Norges idrettsforbund og Norges olympiske og paralympiske komité. Sak 25/10. Oslo: Norges idrettsforbund, 2010.
2. Lamon S, Giraud S, Egli L et al. A high-throughput test to detect C.E.R.A. doping in blood. J Pharm Biomed Anal 2009; 50: 954–8.
3. ADN (Anti-Doping-Norge). Sak 25/10: ADN mot Erik Tysse. Oslo: Anti-Doping-Norge, 2010.

Received 18 May 2011, first revision submitted 16 June 2011, approved 18 August 2011. Medical editor Anne Kveim Lie.