Response to opinion paper

Advocacy versus impartial scientific review: A problem for scientists and the courts

Larry D. Bowers *

United States Anti-Doping Agency, 1330 Quail Lake Loop, Colorado Springs, CO 80906, United States

A R T I C L E   I N F O

Article history:
Received 6 April 2009
Accepted 13 May 2009
Available online 19 May 2009

Keywords:
Gas chromatography–mass spectrometry
Gas chromatography–combustion–isotope ratio mass spectrometry
Steroids
Landis arbitration case

A B S T R A C T

Blackledge has reviewed a portion of the instrumental data in the Landis doping arbitration case. This rebuttal to his review presents an impartial review of the gas chromatography–combustion–isotope ratio mass spectrometric data in the case. The rebuttal also discusses the responsibilities of expert witnesses in courtroom testimony.

© 2009 Elsevier B.V. All rights reserved.

Mr. Blackledge has written an evaluation of the scientific data in the Landis case [1] based on no firsthand knowledge of the data available. In his acknowledgement, Mr. Blackledge clearly states his reliance on only the opinions of Landis' experts, which is particularly troubling in light of the Court of Arbitration for Sport (CAS) Panel's decision [2] in which they stated:

261. The Panel also finds much force in Respondent's contention that “Appellant's experts crossed the line, acting for the most part like advocates for Appellant's cause and not as scientists objectively assisting the Panel in the search for the truth.”

For those who are not experienced in providing expert testimony, this is as stern a rebuke from the bench as one is likely to encounter. The 3 jurists who heard Landis' appeal of his initial American Arbitration Association (AAA) conviction to CAS are extremely well respected international arbitrators who also hear cases outside of sport. 1 Mr. Blackledge, who apparently has not even read the decisions in Mr. Landis' cases, criticizes the process. I will not even respond to these uninformed allegations. But rather than simply take Mr. Blackledge to task for his partisan review of the instrumental data, I would like to provide an overview of the way that USADA and its experts reviewed the scientific data.

Before I can address the science issues, however, I need to briefly explain the broader context of the adjudication of anti-doping rule violations in sport (for an overview, see [3]). The World Anti-Doping Agency (WADA) Anti-Doping Program (Program) consists of the World Anti-Doping Code (Code), the List of Prohibited Substances, the International Standard for Testing, the International Standard for Laboratories, the International Standard for Therapeutic Use Exemptions, and the International Standard for Privacy and Personal Information. The Program was developed through an international consensus process involving athletes, sport leaders, and governments. More than 570 global sport organizations have incorporated the Code into their rules. The 110 governments (including the United States) who ratified the United Nations Education, Science and Cultural Organization's Anti-Doping Convention [4] have also agreed to support the Program in their laws. So the Program truly represents an international consensus on how anti-doping rule violations should be addressed.

The burden of proof in anti-doping cases is not the criminal standard of beyond reasonable doubt, but rather the “comfortable satisfaction” standard used for violation of professional standards by many medical, legal, and scientific organizations (including, for example, the American Association of Forensic Science). Athletes agree to abide by the anti-doping rules as part of the rules of their sport and as a pre-condition of their participation in the competition. Under the Code, arbitration is used to resolve anti-doping rule violations; this is consistent with the Ted Stevens Olympic and Amateur Sports Act that requires all disputes about eligibility in the United States to be heard by the AAA.

Mr. Landis was charged with an anti-doping rule violation based on the results of a gas chromatography–combustion–isotope ratio mass spectrometric (GC-C-IRMS) analysis conducted at the WADA-accredited Laboratoire National de Dépistage du Dopage (LNDD). That analysis showed that a metabolite of testosterone in his urine sample had significantly less $^{13}$C than his other endogenous steroids. The
application of GC-C-IRMS to anti-doping has been reviewed [5]. To briefly describe the method, after a very selective clean-up procedure using, e.g., HPLC, the hydroxyl groups of steroids isolated in a number of fractions are derivatized with acetic anhydride and separated on a GC column. On exit from the column, the steroids are combusted to CO₂ and H₂O in a CuO furnace, the H₂O removed, and the CO₂ introduced into a special mass spectrometer that continuously monitors m/z 44 (12C16O₂), 45 (13C16O₂; 12C17O16O) and 46 (12C18O16O) when isotopes of carbon are analyzed. Given that the amount of 13C in nature averages 1.1% of the 12C, one would expect that the m/z 45 signal would be about 100-fold smaller than the m/z 44 signal. Because the differences in 13C content are even smaller than the 1.1%, the ratio of the 13C/12C is reported as δ values, the units of which are “per mil” (‰). Once a compound (or mixture of compounds) is combusted in the furnace, there is no way to make any statement about what contributed to the isotope ratio. The laboratory must, therefore, have another way to ensure that the compounds of interest are pure and have been identified. This can be done by splitting the flow post column and directing a portion to a molecular mass spectrometer or by analyzing the derivatized fractions in a separate GC-MS instrument. LNDD did the latter. It should be pointed out that no matter which way the process is done, the retention times will not be identical in the GC-MS and GC-C-IRMS chromatograms primarily because of the post-column hardware required to combust the compounds.

It should be pointed out that GC-C-IRMS was within the scope of LNDD’s ISO/IEC 17025 accreditation and that an external GC/C-IRMS expert had assessed their procedure only months before the Landis sample was analyzed. According to the International Laboratory Accreditation Cooperation, accreditation is “a formal recognition that an organization is competent to perform certain specified tasks.” [6] The LNDD staff has also published GC-C-IRMS articles in the peer-reviewed literature [7–10].

The steroids involved in the Landis case were 5α-androstane-3x-ol-17-one (Andro), 3β-androstane-3x-ol-17-one (Eto), 5α-androstan-3β,17β-diol (5αAdiol), 5β-androstan-3β,17β-diol (5βAdiol), 5β-androstan-3x-ol-11,17-dione (11-ketoEtio; an endogenous reference compound (ERC)), and 5β-pregn-3x,20x-diol (5Pdiol; ERC). All 4 of the 19 carbon steroids are metabolites of testosterone and its precursors. Pharmaceutical testosterone is synthesized from plant precursors that are depleted in 13C. Thus when 13C-depleted testosterone or its precursors are taken, the δ values of the steroids metabolically downstream become more negative. The ERC compounds 11-ketoEtio and 5Pdiol are formed in the metabolism of glucocorticosteroids and thus their δ would not be affected by the ingestion of testosterone or its precursors. Thus the use of differences between the δ values (Δδ) of the suspect steroid and an ERC is a sensitive indicator of changes in 13C depletion not related to diet or other physiological causes. Testosterone metabolism makes up about 10% of the Andro and Eto pool, while the 5αAdiol and 5βAdiol pool are predominantly formed through reduction of testosterone. Piper et al. [11] showed that the 5α-stereoids are more affected by transdermal administration of testosterone than the 5β-stereoids; for injections there was less of a difference [12]. Thus the greater change in δ values for 5αAdiol and Andro observed in Mr. Landis’ urine (Table 1) are explainable.

Mr. Blackledge demonstrates his lack of understanding of the fundamental concepts of quality control by asking why “it’s OK to screw up 1 of 4 measured values for a blank sample ...” when 1 positive result in an athlete’s sample is a doping violation. If ± 0.8‰ is 2 standard deviations above or below the mean, then 5 out of 100 measurements would be expected to fall outside the ± 0.8% range. If one then measures 4 analytes in a quality control sample, each of which can fall outside the ± 0.8% range 5% of the time, it is a straightforward calculation to compute that the probability of one of the values being outside of the ± 0.8% range would be 17.1%. This concept is well understood in clinical chemistry where multiple controls are run in immunoassay. In this case, the analytical batch could be rejected nearly 20% of the time for no good reason. A similar calculation shows that the probability that ≥ 2 of the values will be outside of the 2 SD range is 1.4%. By allowing 1 value to fall outside of the 0.8% limit in any run, the negative quality control will be accepted 98.6% of the time when there is nothing wrong—reasonably close to the 95% used in most quality control charts. Unfortunately, there is no ability to allow between 1 and 2 of the measurements to be outside of the ± 0.8% limits. Analysis of population reference ranges and the Δδ values seen in testosterone administration studies, on the other hand, determined the WADA threshold. And measurement uncertainty is added on top of the threshold. So there is no possibility that an analytical error, particularly in the light of the quality control measures undertaken by LNDD, resulted in Mr. Landis’ adverse finding.

The heart of Landis’ defense, parroted by Mr. Blackledge was their claim that LNDD could not reliably identify the peaks in the GC-C-IRMS chromatograms by reference to the GC-MS chromatograms of the sample fractions because 2 different column types and temperature programs were used on the 2 instruments. Landis’ defense team did not raise the column type issue at the initial AAA hearing in 2007 (probably because it was discovered by an advocate blogging on the Trust but Verify and/or “Wiki defense” website after the AAA hearing). USAADA had noted the column types reported for the GC-MS method (Agilent 19091s-433 (HP-5ms): 0.25 mm × 30 m; 0.25 μm dₜ) and the GC-C-IRMS method (J&W DB17-MS: 0.25 mm × 30 m; 0.25 μm dₜ) were not the same during its review of the documentation packages. The laboratory director of the LNDD laboratory was specifically asked whether a different column was used and he personally assured USAADA that the same column type (DB17-MS) was used in both instruments. Given my personal experience with steroid chromatography on the columns in question, I knew that the elution order would be the same although the peak resolution on the Agilent HP-5ms column could be somewhat decreased (e.g., [12]). To assure himself that the same column was used, Dr. Brenna, an expert for USAADA, purchased both columns after the AAA hearing and ran them on both GC-MS and GC-C-IRMS instruments in his own laboratory. He included the results of these tests in his written witness statement for the CAS hearing. The retention behavior of the two columns using the LNDD temperature program was clearly different (as expected), and the GC-MS data presented in the documentation package could only have come from a DB17-MS column. This is the kind of scientific work

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dd values obtained in the Landis case.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Blank urine “A” negative quality control</td>
</tr>
<tr>
<td>Blank urine “B” negative quality control</td>
</tr>
<tr>
<td>Landis urine “A” bottle</td>
</tr>
<tr>
<td>Landis urine “B” bottle</td>
</tr>
<tr>
<td>Landis Tour de France other seven samples</td>
</tr>
<tr>
<td>Highest value*</td>
</tr>
<tr>
<td>Lowest value*</td>
</tr>
</tbody>
</table>

*The shaded boxes represent the measurements that were adverse analytical findings. The value of 3.51% fell below the “decision limit” which was the WADA threshold (3%) plus the expanded measurement uncertainty (0.8%). In workplace testing, for example, this would have been considered a positive since it is not be statistically different than the first value and the re-test value is not required to be above the threshold value. 4/7 5αAdiol Δδ measurements were above the threshold; 2/7 5β(Adiol Δδ measurements were above the threshold. 1/7 Andro or Eto measurements were above the threshold limit. The samples were not declared an adverse analytical finding initially because the testosterone/epitestosterone ratio was less than 4:1 and thus the GC-C-IRMS test was not applied.

*Two samples from the early stages of the Tour had Δδ values for all analytes within the population “reference range”. The intra-individual variations between the lowest and highest differences are larger than expected.
that assists the arbitrators in determining the truth in cases, not diatribes about what might have or should have been. During the CAS hearing, a service engineer testified that he had installed his Agilent 19091s-433 column during an annual verification of instrument performance, had changed the column name in the Agilent 6890 Instrument Control Parameters for documentation, and had forgotten to change the column name back after he removed his column. As anyone familiar with the Agilent 6890 is aware, the column listing in the Instrument Control Parameters section of the GC method is a text field and has nothing to do with control of the analysis. Presented with two pieces of evidence, the arbitrators ruled the Landis experts incorrect in concluding that two different column were used. The panel’s opinion has been available since June 30, 2008. Even more troubling is that I personally informed Mr. Blackledge that a number of his statements were factually incorrect at a breakfast roundtable he presented at the International Association of Forensic Science meeting in New Orleans in July, 2008. Yet Mr. Blackledge continues to make these statements in presentations and in print.

A related issue in Mr. Blackledge’s opinion piece was that Dr. Meier-Augenstein purportedly could not figure out which peak corresponded to which steroid based on the data provided in the documentation package. Although the columns used in the GC-MS and GC-C-IRMS procedures are in fact the same, the LNDD did use different temperature programs for the two instruments. Interestingly, the person(s) who prepared the section in the Landis brief on temperature program differences did not understand the concept of “final time” in Agilent 6980 programming and ignored it. These statements from that brief appeared verbatim in at least 1 of Landis’ expert witness statements. I have plotted the 2 temperature programs, taken from pages USADA 124 and USADA 152 in the LNDD “A” documentation package, in Fig. 1. While not identical, the programs are similar particularly in the temperature region where the steroids of interest elute. The flow rates are somewhat different, as LNDD adjusts the head pressure in the GC-C-IRMS system to set the retention time of the 5α-androstanol-acetate retention time marker to about 860 s And, of course, the GC-C-IRMS column is terminated at atmospheric pressure whereas the GC-MS column terminates in a vacuum. The LNDD ran a standard called “Mix Acetate” that contained 5α-androstanol-acetate, Etio-acetate, Andro-acetate, 5αAdiol-diacetate, 5βAdiol-diacetate, 11-ketoEtio-acetate, and 5Pdiol-diacetate in their GC-MS method. So we have established, based on the retention time and full scan mass spectra of each peak, the identity of the seven steroids on the GC-MS chromatogram. The LNDD produced documentation showing a full scan of all six steroids in Mr. Landis’ sample. Notable in the scans was not only the near identity of the spectra between the reference material and Mr. Landis’ sample fraction, but also the absence of any other significant fragment ions in the steroid peak spectra. This clearly documented that no other compounds were eluting with the steroid in the liquid in the vial about to be introduced into the GC-C-IRMS instrument.

In their procedure, LNDD transfers the capped autosampler vials containing the derivatized fractions obtained in the sample clean-up first to the GC-MS then to the GC-C-IRMS. So exactly the same solutions were analyzed by both techniques. LNDD ran a different standard, “Mix Cal Acetate”, in the GC-C-IRMS method. The 5α-androstanol-acetate, Etio-acetate, 5βAdiol-diacetate, and 11-ketoEtio-diacetate in this standard had 13C δ values measured in a reference laboratory, and the δ values covered the range from −16‰ to −33‰. This is the range of δ values that might be measured in the method and documents that the IRMS instrument is measuring δ values accurately. The retention time of these 4 steroids is established from the standard injection as well. How can we be sure that the peaks in the GC-C-IRMS correspond to the ones in the GC-MS? In Fig. 2, the retention times of the 4 known steroid standards from the GC-C-IRMS are plotted against the retention times of the same 4 known steroid standards from the GC-MS (closed diamonds). Given the experimental conditions on the two instruments, one would expect a slightly curved line with an offset due to the additional GC-C-IRMS plumbing. One can then interpolate what the retention times of the Andro-acetate, 5αAdiol-diacetate, and 5Pdiol-diacetate should be from their retention times on the GC-MS (open diamonds). The interpolation agrees with the observed GC-C-IRMS retention times within ±8 s. Since a blank urine sample that must contain naturally produced Andro, Etio, 5αAdiol, 5βAdiol, 11-ketoEtio, and 5Pdiol was run through the same preparation steps, their retention times must be reflected in the blank urine chromatogram. For anyone with steroid GC expertise, it should be known the 5β steroids always precede 5α steroids on a DB17-MS column and they will be beside each other in the chromatogram. Since we have the retention time of 5βAdiol-diacetate from the “Mix Cal Acetate” reference material, the peak next to it must be 5αAdiol-diacetate. The agreement between the 5αAdiol-diacetate GC-C-IRMS retention time and the “unknown peak” in Mr. Landis’ sample was 1336.6 vs 1337.2 s (a difference of 0.6 s and well within the ±0.1 min generally acknowledged as sufficient for consistency in GC). Similar comparisons can be made for the other 2 steroids not contained in the “Mix Cal Acetate.” Further additional confidence can be obtained from the relative magnitude of the peak heights observed on both instruments. Dr. Meier-Augenstein argued that because the ionization process in GC-MS is different than measuring the moles of CO2 in GC-C-IRMS, the peak heights could be different. While his argument is theoretically correct, it is moot in this case because we are comparing steroid isomers having nearly identical ionization cross-sections and
the same number of carbon atoms. (In raising this point, was Dr. Meier-Augenstein attempting to assist the panel in understanding complex science or was he acting as an advocate?) Looking at the totality of the information, all of the scientific complaints from Landis' experts were addressed in a similar logical manner.

To provide an additional example of Blackledge's lack of competence in GC-C-IRMS, he asks in the caption for Fig. 4 of his paper “What caused the sudden rise in baseline at about 760 s?” A competent scientific review would recognize that this occurs because the first 750 s of column effluent in the GC-C-IRMS analysis are not directed to the combustion furnace and IRMS since the carbon in the injection solvent could deplete the CuO in the furnace (LNDD “A” documentation package, page USADA 152). (Note: the chromatograms in Fig. 4 of Blackledge’s paper are not from the same preparatory HPLC fraction and thus cannot be compared.) As yet another example of Mr. Blackledge not getting his facts right, he states than Dr. Simon Davis was present for the analysis of Landis’ B sample. Had he actually studied the documentation available to him, Mr. Blackledge would have seen that Dr. Douwe de Boer (along with two of Mr. Landis' lawyers and two additional scientific experts) attended the B analysis (LNDD “B” documentation package, pages USADA 250-1). Dr. de Boer noted in his report (page USADA 368) that “The impression of the expert regarding the analytical performance of the B-sample analysis was that the LNDD worked in a transparent and professional way and according to transparent and professional procedures.” This is not to say that Dr. de Boer did not have some questions regarding some aspects of the analysis (which were addressed during the hearings), but at least he maintained scientific perspective in his assessment. Interestingly, Dr. de Boer never testified for Mr. Landis’ team in either hearing.

While I could continue to critique Mr. Blackledge's understanding of the science behind the Landis case point by point, I will not take further space to do so. I would like to end here with a few comments. First, lawyers are paid to argue their client's position in court or arbitration. Expert witnesses, no matter who pays for their time, have a duty to assist the arbitrators or judges in understanding complex scientific issues. This is not necessarily what the client's lawyer-advocate wants the expert to do, and it is easy for the scientist to be seduced into becoming a part of the defense team. In my opinion, an expert is required to impartially review the totality of the documentation in order to reach an informed opinion. The increasing partisanship of expert witnesses is a serious concern to both the legal and scientific profession. Mr. Blackledge and I agree on one thing. A large amount of time and money were expended in the Landis case. In my opinion, Landis’ advisors and experts significantly increased costs by arguing positions (column identity, bacterial contamination, lifting rings on the magnet, etc., etc.) despite the fact that their arguments were refuted by other data in the more than 4000 pages of documentation provided to them. The CAS arbitrators recognized this advocacy and chastised Landis’ experts for it. The rigorous, impartial scientific review undertaken by USADA’s experts focused on the totality of the data and considered how all the data fit together.

As Paul Harvey used to say “Now you know the rest of the story.”

Acknowledgements

I would like to acknowledge discussions during the preparation for the Landis hearings with three true experts in GC-C-IRMS: Prof Dwight Matthews, who invented the technique during his Ph.D. dissertation and has been a continual contributor to the literature since; Prof Tom Brenna, who has contributed to both instrumental and software development and an understanding of lipid metabolism using stable isotope techniques; and Dr. Janine Jumeau, who headed the development of the original VG GC-C-IRMS instrument. All three acted as expert witnesses for USADA during the course of the two Landis hearings.

References