**Expert Witness Report**
**Jonathan H. Foreman**
**February 18, 2010**

## Definitions and background

The Maine State Harness Racing Commission (MSHRC) regulation states:

> C. When test levels of Total Carbon Dioxide (TCO2) exceed the acceptable level, it is demonstrative that the horse has been administered an alkalinizing agent and constitutes a violation of the rules.

This regulation is, in itself, flawed, as the presence of a tCO2 concentration ≥37 (or ≥39 for a Lasix horse) does not prove that the horse received an alkalinizing agent. Why?

The origin of the "37 rule" is an Australian paper from 1993 in which Auer et al. showed a normal range of tCO2 in racing Standardbreds which they felt warranted a forensic prosecutorial tCO2 concentration of 37 (since then reduced to 36 in Australia). Since Lasix is not legal in Australian racing, Auer et al. (1993) did not argue for or against a Lasix allowance. The need for a Lasix allowance of 2 mmol/l over the regular tCO2 limit became accepted in many jurisdictions as a result of work in our laboratory in which we documented, in over 7000 pre-race samples, that Lasix caused a clear alkalinizing effect even in non-"doped" Standardbred horses (Frey et al. 1995). Recent work has shown that the Lasix alkalinizing effect is dose-dependent (Kline et al. 2006).

Considerable data exist now to show that normal untreated North American (Ontario) Standardbreds can have much wider variations in tCO2 concentrations than previously thought (Waller et al. 2010), or than previously documented using other methodologies on other continents (Auer et al. 1993). Waller et al. (2010) also showed that tCO2 concentrations vary significantly between different Standardbred racing stables. Many of Waller's normal untreated Standardbred horses would have tested "positive" in a jurisdiction with a tCO2 limit of 37 mmol/l (16 of her 211 horses tested between 37.0 and 42.9 mmol/l tCO2, a false positive rate of 7.6%).

In our laboratory Kauffman et al. (1999) documented that dietary changes cause transient changes in tCO2 which could result in horses testing ≥37. Waller et al. (2005) subsequently showed that plasma tCO2 decreased in response to eating a grain meal; conversely, it can be inferred that failure to eat a grain meal on race day may result in dietary-induced increases in tCO2. Clearly diet and dietary changes can have a significant effect on tCO2 concentrations.

## Establishing a population mean and range

The MSHRC states that sampling of Maine racing Standardbreds established a normal mean value of 30.5 ± 2.036 (mean ± standard deviation [SD]) mmol/l. Adding 3 SDs to that mean, investigators arrived at an upper limit of 36.6 mmol/l, according to the Commission's analysis. The argument appears to be that those data are sufficient to

establish and support the regulation of plasma tCO2 concentrations at 37. Unfortunately, the argument belies two huge errors in implementation.

First, the population "normals" were established using a different instrument (Beckmann ELISE analyzer) than that being used routinely in Maine in the latter half of 2010 (NOVA Biomedical CRT4 analyzer). Research in our laboratory has shown considerable variation in results when duplicates of plasma samples were tested for tCO2 using different laboratory analyzers (Greene et al. 1999).

The Commission's failure to re-establish normal values using a new analyzer violates a basic premise of all scientific investigation, that the normal values (or "normal range" or "reference range") of the subject population must be established using the same instrument which will be used subsequently for testing, whether that testing is for clinical or forensic purposes. Anytime an analyzer in a chemistry laboratory is replaced with a new model or a different manufacturer's equipment, it is incumbent upon that laboratory to establish a new normal range of population values for any analysis using that new analyzer. The MSHRC failed to perform that critical function in this situation, potentially to the detriment and harm of the harness trainers attempting to make a living in the state.

Secondly, the Commission's use of the population mean ± 3 standard deviations (SDs) allows for an unacceptable inherent level of error resulting potentially in the prosecution of innocent trainers. It allows for false positives at least once in every 370 tests. Table 1 illustrates the inherent statistical false positive rate when the mean has increasing numbers of standard deviations added to it to establish a high end of acceptable error. The simple use of the mean ± 4 SDs would make the chance of a random error to be on the order of 1 per 15,788 as opposed to 3 SDs resulting in a random error of 1 per 370 horses.

| **TABLE 1** | **% normal horses outside that range defined by the mean + X# of SDs** | **% normal horses outside that range defined by the mean + X# of SDs** | **Frequency of false positive test, or the fraction of normal horses outside the range defined by the mean + X# of SDs** |
|---|---|---|---|
| Mean ± 0.674 SD | 50.00% | 50.00% | 1 / 2 |
| Mean ± 1 SD | 68.26% | 31.73% | 1 / 3.15 |
| Mean ± 2 SDs | 95.45% | 4.55% | 1 / 21.97 |
| Mean ± 2.58 SDs | 99.00% | 1.00% | 1 / 100 |
| Mean ± 3 SDs | 99.73% | 0.27% | 1 / 370.39 |
| Mean ± 4 SDs | 99.9937% | 0.000063% | 1 / 15,788 |
| Mean ± 5 SDs | 99.999943% | 0.000057% | 1 / 1,744,278 |

**Day-to-day variation in measurements**

All of the above is based on the assumption that the test being used is accurate. In Maine the test being used in not accurate for forensic purposes in the manner in which it is being

used. Specifically, the test has considerable day-to-day variation which is not being taken into account properly when screening tests are determined to be "positive."

Table 2 summarizes a statistical analysis of all patient or horse samples (not "calibration" or reference samples) in the State's exhibit entitled "Maine2 QC Test Results NOVA CRT 4 010311." All horse samples were divided, by day of analysis, into two groups: days where there were no screening samples ≥37, and days where there were screening sample results ≥37. No "calibration" sample results were included; only actual horse sample results were used. [Lasix information was not provided and was not necessary for this analysis, as inclusion of Lasix samples ≥37 for this analysis was in the state's favor, not the trainers' favor, in determining the day-to-day variation in the instrument.]

| **TABLE 2**<br><br>Group | **"Low" test days: No tCO2 ≥37 mmol/l on that day of testing** | **"High" test days: At least one tCO2 ≥37 mmol/l on that day of testing** | **Modified "High" test days: All values ≥37 mmol/l eliminated from analysis** |
|---|---|---|---|
| Mean (± SD) | **31.9** (4.3) | **35.2** (3.4) | **33.6** (2.6) |
| P value when compared to "Low" test day values | Not applicable | **P<0.001** | **P<0.001** |

The mean values calculated are summarized in Table 2. All of these new population means are considerably higher than the Commission's mean (30.5 ± 2.036) determined using a different instrument (Beckman ELISE) in the previous MSHRC population study. Clearly the NOVA tests higher than the Beckman values on which the regulation is based.

The mean (average) values for the two groups are presented in Table 2, showing that on "Low" test days the mean tCO2 concentration was 31.9 (Table 2, column 2) and on "High" test days the mean tCO2 concentration was 35.2 (Table 2, column 3). A simple statistical test called a "signed rank test" showed that the two groups were mathematically ("statistically") different from one another (P<0.001), a finding which may not be a surprise since some values in the second group were ≥37. [The "signed rank test" was used because a test of "normality" showed that the data were not normally-distributed, meaning that a test using continuous data, such as a simple "Student's t test," was inappropriate.]

However, even when all the values ≥37 in the second group, the "High" test day group, were eliminated from the analysis (Table 2, column 4), there was still a clear mathematical difference in mean tCO2 concentration between "Low" (31.9) and "High" (33.6) test days (P<0.001).

Only two conclusions can be argued, and the two conclusions stand in stark juxtaposition. First, if the instrument is completely accurate and does not vary from day-to-day, then the population mean clearly varies from day-to-day. In that case, the whole premise that it is possible to establish a cutoff for testing "positive" based on a population mean is obviously faulty if the population mean varies from day-to-day. Second, the conclusion must be that the test varies in accuracy from day-to-day, as do all chemical tests, and the chances of a horse testing "positive" are much better on days when the instrument is testing "high" than on days when the instrument is testing "low." In a single phrase, in Maine as the test is currently being used, horses test high when the machine tests high, and horses test low when the machine tests low.

Obviously, the latter conclusion is much more likely than the first one, but in either case one can only conclude that the test, as implemented and interpreted, is flawed.

**Accuracy is more than just linearity**

A paper attached in a separate document is entitled "Validation and comparison of two methods of measuring lactate in equine plasma" by Dr. Prawit Butudom from our laboratory (2010). While lactate analysis may not seem pertinent to the current argument, the paper illustrates quite well that there is much more to establishing accuracy over a critical range of analysis than merely proving linearity. In Butudum's paper, all of the following laboratory measures were used to validate and to compare two methods of measuring plasma lactate concentration in standardized solutions and in plasma samples:

- Coefficient of variation
- % Recovery
- Linearity
- Parallelism
- Direct comparison of the two methods using paired analyses

To opine that linearity alone is sufficient to document accuracy is belied by common sense. The instrument may test incrementally along the line it has drawn for itself in its software (it may be "linear"), but that does not mean that the values produced are accurate or true to the real value of tCO2 in the sample. The only way to determine that the instrument is accurate is to perform measurements using external standardized solutions which are run contemporaneously with the patient or suspect sample (see recommendations below).

The current MSHRC methodology has no available written Standard Operating Procedure (SOP). However, the data printouts in "Maine2 QC Test Results NOVA CRT 4 010311" clearly indicate that when a horse tests "positive" on a screening test, the laboratory personnel run a calibrator solution which tests roughly 13 and 23 mmol/l. Without an SOP, it is impossible to tell how the laboratory personnel use or interpret those calibration data. Do they adjust the patient or suspect value for the standard curve which could be constructed using the ~13 and ~23 sample results? Apparently not, as there is no evidence that those standard curves exist, nor is there any evidence that those standard curves are used to adjust instrument output values to obtain true or accurate

values. How far off 13 and 23 would those values have to be for laboratory personnel to determine that the instrument was not performing accurately that day? Again, there is no documentation (no SOP) of how far of 13 and 23 mmol/l is "too far off" to consider the instrument accurate that day.

**What is the solution to the problem?**
The following are minimal steps which must be achieved in order to improve (and defend) the manner in which the laboratory and the Commission are conducting and interpreting tCO2 test results.

- The laboratory must develop a written SOP documenting for all personnel how to perform this test, including how to interpret the results from the instrument.
- Without external calibrators, all we know is that the instrument makes incremental measurements along a line (it is "linear" according to the manufacturer), but we still do not know if the instrument is measuring the true (or accurate) value for that sample at that point in time. The laboratory must use external (non-NOVA Biomedical) standard solutions with U.S. National Institute of Standards and Technology (NIST) certification. The NIST-certified standards must have values bracketing the range of interest for the test being performed. In the case of tCO2 testing, such standard solutions exist with known NIST-certified tCO2 values of 10, 20, 30, and 40 mmol/l tCO2 concentrations (Casco-NERL Diagnostics, East Providence, Rhode Island 02914).
- Once the laboratory adopts the inclusion of appropriate external NIST standards, then the NIST standards must be run (in several replicates) when a confirmatory test is performed in order to create a "standard curve" of known tCO2 values between 10 and 40, thus bracketing the points of interest at 37 and 39 mmol/l.
- Once the replicates of the patient or horse sample have produced an apparent tCO2 value, then that value must be plotted on the standard curve to correct it for the way the instrument is running that day.
    - If the instrument is running high that day (the NIST standards are testing higher than their certified values), then the suspect value obtained must be adjusted downward, according to the standard curve, to arrive at the true or accurate value for that horse on that day.
    - If the instrument is running low that day (the NIST standards are testing lower than their certified values), then the suspect value obtained must be adjusted upward, according to the standard curve, to arrive at the true or accurate value for that horse on that day.
    - To illustrate the correction of measured values to true or accurate values, please see Figures 1 and 2 below. Figure 1 illustrates the following:
        - The open circles represent the ideal or perfect "line of identity" where the external NIST standards tested exactly as published, with 30 mmol/l testing 30 and 40 testing 40 on that series of measurements.
        - The reality is that the NIST standards will almost never test along the line of identity. The solid circles and line represent the actual results from NIST standards during a series of measurements.

        - The conversion line (arrow) shows that an expected value of 37 would represent an actual or accurate test value of 39.4 mmol/l if properly converted using the standard curve.
    - Figure 2 illustrates the following:
        - Again, the open circles represent the ideal or perfect "line of identity" where the external NIST standards tested exactly as published.
        - The solid circles and line represent the actual results from NIST standards in a series of measurements.
        - One conversion line (arrows) shows that an uncorrected measured value of 39.2 mmol/l converts to an accurate value ("true tCO2") <37 and would not trigger a "positive" test report.
        - The other conversion line (arrows) shows that an uncorrected measured value of 39.8 mmol/l converts to an accurate value ("true tCO2") >37 and would trigger a "positive" test report.
- The NIST-certified standard-corrected value must then be reported as the fair and accurate value for that horse that day.
- The Commission must determine if the appropriate values for prosecution are 37 and 39 mmol/l. Considerable recent research (Kauffman et al. 1999; Waller et al. 2005, 2010) should make the Commission re-consider its current cutoffs of 37 and 39 mmol/l.
    - The Commission must repeat its randomized controlled survey of Maine racing Standardbreds to determine a more appropriate, instrument-specific population mean ± 3 (or hopefully ± 4) SDs (see Table 1) to determine fairly and more accurately the correct enforcement levels in their industry using the same instrument and the improved methodologies that it adopts.

**Selected References**

Auer DE, Skelton KV, Tay S, Baldock FC. Detection of bicarbonate administration (milkshake) in Standardbred horses. *Australian Vet J* 1993;70:336-340.

Butudom P, Foreman JH, Kline KH, Whittem EL. Validation and comparison of two methods of measuring lactate in equine plasma. *Equine Vet J* 2010;Suppl 42:155-160.

Frey LP, Kline KH, Foreman JH. Effects on prerace exercise, frusemide, sex and ambient temperature on blood sodium, bicarbonate and pH values in Standardbred horses. *Equine Vet J* 1995;27:170-173.

Greene AS, KH Kline, Foreman JH. Comparison of three blood gas machines for determination of plasma $TCO_2$ in horses administered sodium bicarbonate. *Proceedings of the Equine Nutrition and Physiology Symposium* 1999;16:380.

Kauffman KE. Kline KH, Foreman JH, Lyman JT. Effects of diet on plasma tCO2 in horses. *Proceedings of the Equine Nutrition and Physiology Symposium* 1999;16:363-364.

Kline K, Fitzpatrick D, Taddei L, Sukata A. Effect of dose of furosemide on plasma TCO2 changes in Standardbred horses. *J Equine Vet Sci* 2006;26:317-321.

Waller A, Smith KJ, Ecker GL, Geor R, Lindinger MI. Cyclical plasma electrolyte and acid-base responses to meal feeding in horses over a 24-h period. *Equine and Comparative Exercise Physiol* 2005;2:159-169.

Waller A, Pearson W, Lindinger MI. Factors contributing to plasma TCO2 and Aacid-base state in Ontario Standardbred racehorses. *Equine Vet J* 2010;Suppl 42:592-600.

**Figure 1**
The open circles represent the ideal or perfect "line of identity" where the external NIST standards tested exactly as published, with 30 mmol/l testing 30 and 40 testing 40 on that series of measurements. The solid circles and line represent the actual results from NIST standards during a series of measurements. The conversion line (arrow) shows that an expected value of 37 would represent an actual or accurate test value of 39.4 mmol/l if properly converted using the standard curve.

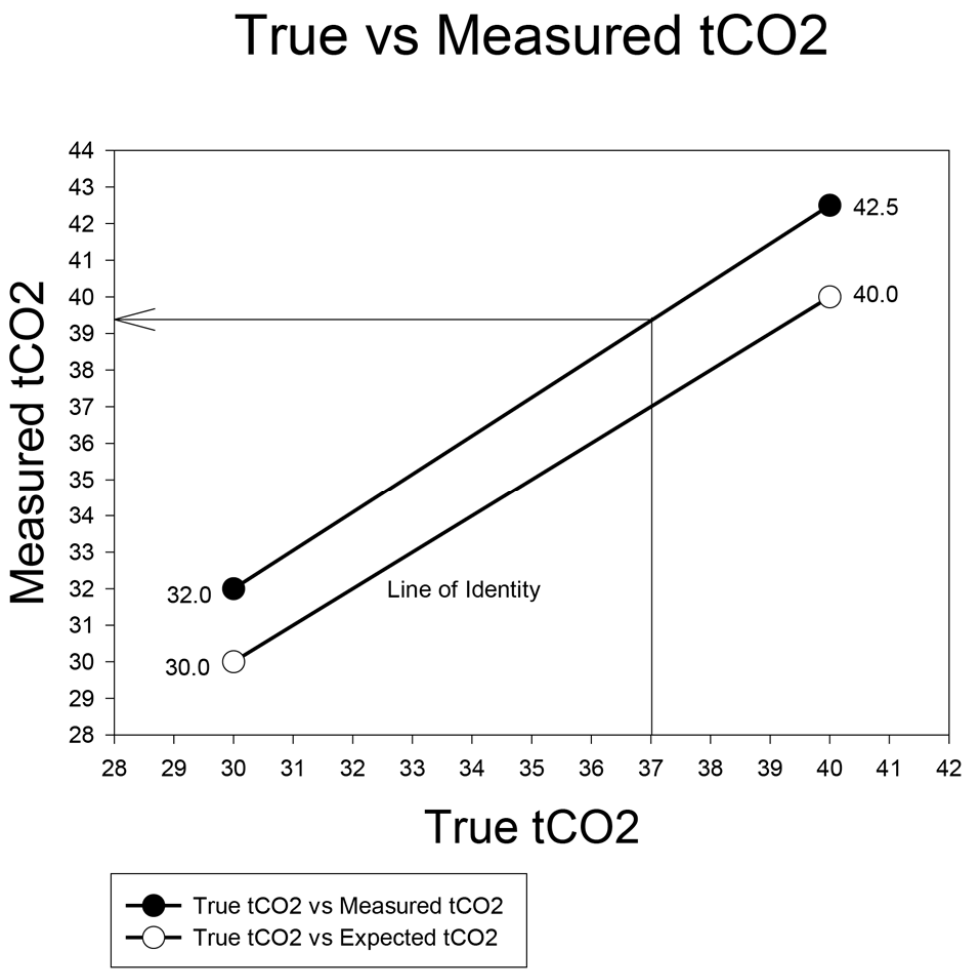

**Figure 2**
The open circles represent the ideal or perfect "line of identity" where the external NIST standards tested exactly as published, with 30 mmol/l testing 30 and 40 testing 40 on that series of measurements. The solid circles and line represent the actual results from NIST standards during a series of measurements. One conversion line (arrows) shows that an uncorrected measured value of 39.2 mmol/l converts to an accurate value ("true tCO2") <37 and would not trigger a "positive" test report. The other conversion line (arrows) shows that an uncorrected measured value of 39.8 mmol/l converts to an accurate value ("true tCO2") >37 and would trigger a "positive" test report.