For our first endeavor into the realm of accuracy-based testing for testosterone and estradiol, we are very pleased with the wealth of information we were able to collect. In this participant summary report, we hope to give an initial overview of what we found, as well as a description of the grading criteria we used. We assure you, however, that a great deal more information will be forthcoming as we analyze the data in more detail.

A bit of background on these samples is in order. Each sample came from a single individual. As you know, in order to maximize the range of testosterone and estradiol concentrations, we collected samples from a young man, an older man, a young woman, and an older woman. In addition, to achieve a range of cortisol concentrations, two samples were drawn in the morning, and the other two were drawn in the late afternoon. We hoped that we might also see a range in TSH concentrations and SHBG concentrations. We were reasonably sure that the calcium concentration range would be relatively small.

The primary focus of this exercise was testosterone and estradiol, because we had arranged to have the true values for these analytes determined at the CDC by a reference measurement procedure. That is, with minimally processed human blood samples and concentrations determined by the reference method, we could establish TRUTH, and then we could compare each laboratory’s measurement to this value.

For the other analytes, we did not establish the true values by reference methods (either because we did not arrange to do so (calcium and cortisol) or because they do not yet exist (SHBG and TSH)). However, we believe that the data we collected is still extremely interesting and valuable, because we have values on minimally processed human serum samples, so we can make some comments about the “harmonization” of the methods represented; that is, the degree to which they get the same values on the same samples.

Because this represents the first iteration of this survey, the working group, after considerable discussion, decided to make all the grading EDUCATIONAL. In the detailed descriptions for each analyte that follow, we describe the specific criteria we would have applied, along with the range of values that would have been deemed GOOD PERFORMANCE. We felt we needed more time to make a final decision on grading, and we didn’t want to delay any further getting these results back to you.

What follows is a more detailed description of the results for each analyte.
**Calcium**

The calcium results were remarkably good. As expected, the range of concentrations among the four samples was small (~8.8 to 9.4 mg/dL), but the agreement among peer group means was excellent. For all samples, the range of method means was very small; in fact, for all but ABS-04, the differences in mean values was less than 0.1 mg/dL. As a result of this excellent harmonization, we can look at the all method/all instrument lines for each sample to see that the SD and CV for calcium is very good (roughly 0.2 mg/dL and 2.5%).

Although we were tempted to grade this exercise more stringently in light of this excellent performance, in the end we chose to apply the same criterion we use for our regular proficiency survey – within 1.0 mg/dL of the peer group mean. *With this criterion, almost 100% of participating laboratories submitted ACCEPTABLE results on all challenges.*

**Cortisol**

We were gratified that our morning and afternoon draws succeeded in providing such a good range of cortisols. In contrast to the calcium results, there are some peer group differences reflected in this data, but they are minimal and, we would assert, have little clinical significance. That is, we think it is reasonable to conclude that, at least among the methods represented in this Survey, the agreement among different methods on real human serum samples is excellent.

In the absence of a reference measurement procedure, we used the same criterion we use on our regular proficiency survey – within 25% of the peer group mean. *With this criterion, almost 100% of participating laboratories submitted ACCEPTABLE results. If the all method mean was used instead of the peer group means and the 25% limit is maintained, the pass rates were 91.2%, 94.7%, 98.2%, and 98.2% for ABS-01, ABS-02, ABS-03, and ABS-04, respectively.*

**Estradiol**

The range of sample concentrations came as a surprise to us. Samples ABS-03 and ABS-04 were from men, so the relatively low concentrations made sense. But Samples ABS-01 and ABS-02 were from women, one of whom was relatively young. After confirming the ages and genders of the subjects as well as re-checking all the reference analyses and transcriptions, we finally came up with what believe is the true explanation. The younger woman was confirmed to be taking oral contraceptives, which, of course, could explain the very low estradiol value by the reference method as well the somewhat higher values by immunoassay (possible cross-reactivity with the estrogen derivatives used in the oral contraceptives and other endogenous substances). After insuring that the older woman had not taken any steroid hormone replacement for many years, we believe that the very low value in her case is real and that two of the peer groups were measuring something other than estradiol.
From these samples, we think we can infer that estradiol measurements in men (or in the concentrations typically seen in men) may be adequate, but that at least two of the current immunoassay techniques provide falsely high estradiol results in women with very low estradiol concentrations. Whether this is clinically significant or not is less clear.

In this case, since we do have true values established by the reference method, we decided to grade against this value using a clinically established criterion of 21.6%. Acceptable values for these samples, then, are listed in the following table. With this criterion, the overall pass rates for samples 01 through 04 were: 13.5%, 2.5%, 54.1%, and 34.5%.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Reference Value</th>
<th>Acceptable range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS-01</td>
<td>3.5</td>
<td>2.7 – 4.3</td>
</tr>
<tr>
<td>ABS-02</td>
<td>1.7</td>
<td>1.3 – 2.1</td>
</tr>
<tr>
<td>ABS-03</td>
<td>26.9</td>
<td>21.1-32.7</td>
</tr>
<tr>
<td>ABS-04</td>
<td>19.4</td>
<td>15.2 – 23.6</td>
</tr>
</tbody>
</table>

**Testosterone**

As was the case with estradiol, we were a little surprised by the range of concentrations in our samples. By including both a young man and an old man, we hoped to get a low normal sample (~300 ng/dL) and a high normal sample (~700 ng/dL). It turned out that both samples had values in the 300-400 range. As expected, the samples drawn from the women had low values.

On sample ABS-01, with a reference value of 26.5, there were distinct differences in the mean values from the 3 peer groups. On sample ABS-02, with a reference value of 7.4, one of these peer groups “disappeared”, because many of those participants reported values less than their detection limit (which, in this case, was a correct answer); the mean values for the other two peer groups were 20.2 and 33.4, which are clearly UNACCEPTABLE results. Presumably, these methods are detecting something other than testosterone. Whether this represents a serious clinical problem is a different question, but laboratories using these methods on clinical samples from females need to be aware of this issue, which has been reported previously.

The data from samples ABS-03 and ABS-04 again show potentially significant differences in the peer group means. For ABS-03, the peer group means were 299.9, 362.6, and 337.0; for ABS-04, 312.5, 356.9, and 310.6. Notably, the peer group with the highest mean value on ABS-03 was also the highest value on ABS-04, but the peer group with the lowest value on ABS-03 was not the peer group with the lowest value on ABS-04. Although these differences in means are not insignificant, they are much lower than reflected in the conventional Survey, where differences of 2.5-fold are seen and have been the subject of considerable criticism. Presumably, the excess differences among peer group means seen on the conventional Survey can be ascribed to matrix effects.
In this case, since we do have true values established by the reference method, we decided to grade against this value using a clinically established criterion of 14.0%. Acceptable values for these samples, then, are listed in the following table. **With this criterion, the overall pass rates for samples 01 through 04 were: 34.3%, 9.8%, 77.9%, and 80.6%. When an arbitrary acceptable limit of 20% was used, the overall pass rates were 41.8%, 9.8%, 95.6%, 97.0%.

We think it is worth emphasizing that the vast majority of testosterone assays are done in conjunction with the work-up of potential male hypogonadism, where a value of ~300 becomes a threshold for further evaluation. The fact that 77.9% and 80.6% of laboratories participating in this exercise had values within 14% of the true value suggests that the state of the art in testosterone testing for this application, though less than ideal, is far better than alleged in a recent article.³

<table>
<thead>
<tr>
<th>Sample</th>
<th>Reference Value</th>
<th>Acceptable range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS-01</td>
<td>26.5</td>
<td>22.8 - 30.2</td>
</tr>
<tr>
<td>ABS-02</td>
<td>7.4</td>
<td>6.4 – 8.4</td>
</tr>
<tr>
<td>ABS-03</td>
<td>351.0</td>
<td>301.9 – 400.1</td>
</tr>
<tr>
<td>ABS-04</td>
<td>322.0</td>
<td>276.9 – 367.1</td>
</tr>
</tbody>
</table>

**SHBG**

Only 22 laboratories submitted data on SHBG, and no method was represented by 10 or more labs, so we have no peer group data. Furthermore, there is currently no reference method for SHBG, so we cannot do accuracy grading. What we can say is that we achieved a good range of concentrations (~10 to 130 nmol/L) and that the interlaboratory variability for each sample seemed reasonably good (CVs of ~10-15%).

We wanted to include SHBG because, in conjunction with testosterone and albumin, it provides a reasonably good way to calculate Free Testosterone, an analyte that has stirred much controversy. The “reference method” (equilibrium dialysis) is difficult to do well and beyond the capability of most clinical laboratories. Some laboratories use an analogue RIA method instead. Values from that method are many-fold higher than the equilibrium dialysis method,⁴ so it is not clear what they are actually measuring. We hope to have more to say on this matter in the future.

For grading purposes, we again used the criterion from our regular Survey, in this case within 3 SD of the all method mean. **With this criterion, almost 100% of the participating laboratories submitted ACCEPTABLE results.**

**TSH**

TSH provided some interesting data as well. Of note, there is no reference method for this analyte, so we cannot determine the TRUE value. However, we can make some comments on harmonization, the degree to which different methods yield the same result, because of the minimal processing of these samples.
In contrast to the data from calcium and perhaps even cortisol, the data from TSH seems to suggest that there are some peer group differences, especially as the concentration increases into the high normal/borderline abnormal range. On all 4 samples, one peer group seems to run a bit higher, which may be of little consequence. However, on ABS-04, the difference of 4.8 versus 3.6 and 3.9 could well represent a clinical issue, with 4.8 representing an abnormal value according to that manufacturer’s package insert (reference interval 0.3-4.2), whereas the 3.6 and 3.9 values represent normal values according to their respective manufacturers’ package inserts.

Once again, in the absence of a reference method, we decided to use as our grading criterion what we use on our regular proficiency testing Survey, within 3SD of the peer group mean. *With this criterion, almost 100% of participating laboratories submitted ACCEPTABLE results.*

References

1. Desirable specifications for total error, imprecision, and bias, derived from biologic variation [online] [accessed August 17, 2010]. Available from URL: http://www.westgard.com/biodatabase1.htm


Gary L. Horowitz, MD, Chair  
William L. Roberts, MD, PhD, Vice Chair  
Chemistry Resource Committee