Recent developments in the field of mass spectrometry have provided the accuracy and sensitivity to evaluate very-low-abundance steroids such as testosterone in female and pediatric patients. In this issue of Clinical Chemistry, Taieb et al. (1) present the most comprehensive evaluation of automated testosterone immunoassays to date. They compared 10 commercially available immunoassays with isotope-dilution gas chromatography–mass spectrometry (ID-GC/MS) and reached the inescapable conclusion that testosterone immunoassay results for specimens from females are inaccurate. Similar data have been reported for individual testosterone immunoassays previously (2), but Taieb et al. (1) are the first to show that for every commercially available testosterone assay studied, the values are in error—by a factor of 2 on average and in some cases by a factor of almost 5. Are assays that miss target values by 200–500% meaningful? Guessing would be more accurate and additionally could provide cheaper and faster testosterone results for females—without even having to draw the patient’s blood.

By limiting all guesses to a narrow range, e.g., 2.04–2.44 nmol/L, the results would rarely be off by more than a factor of 3. Using a random number generator, we generated values close to the average female concentration measured by Taieb et al. (they were kind enough to share their data with us as an aid to writing this editorial). A Bland–Altman plot for guessed values vs ID-GC/MS values had a mean difference for the 55 female samples of 0 nmol/L with a SD of the differences of 1.2 nmol/L. This SD compares favorably with those presented by Taieb et al. (1) in Table 4. Although not intended to be a statistically rigorous proof that random numbers are better than measuring female testosterone values with immunoassays, guessing appears to be nearly as good as most commercially available immunoassays and clearly superior to some!

Because medical test decisions are not made in a vacuum, a patient’s appearance and presenting complaints would give the person guessing the serum testosterone concentration important information. Women with rapidly evolving signs and symptoms of viralization will have dramatically increased testosterone [>10.4 nmol/L (300 ng/dL)] (3), whereas women with late-onset 21-hydroxylase deficiency have moderately increased testosterone [~4.2 nmol/L (~120 ng/dL)] (4). Using this information while making an educated guess should give dramatically improved results. This would make educated guessing the better choice with the added benefits of rapid turnaround time and very low cost.

What are the implications of the results of the study by Taieb et al. (1) for epidemiologic research? A recent study by Dorgan et al. (5) designed to address this issue concluded “that although absolute concentrations may differ for some hormones, RIA and mass spectrometry can yield similar estimates of between subject differences in serum concentrations of most steroid sex hormones commonly measured in population studies”. The testosterone assay that Dorgan et al. were comparing with MS included an extraction and column purification. Many people believe that liquid-liquid extraction combined with column purification before RIA analysis provides accurate results for testosterone in specimens from females. However, we have previously demonstrated that RIAs that include extraction and column purification steps do not agree well with ID-GC/MS (6). An important limitation in the study by Dorgan et al. (5) is that for female specimens they tested only sample pools (low, mid, and high). Determining how the assay would work on individual patient samples is not possible when pooled samples are used. This is a critical flaw, because clinicians are concerned about the concentration of testosterone in an individual; in contrast, when pooled samples are analyzed, any cross-reacting substances in an individual sample are diluted in the rest of the pool. In Fig. 1 of their report, Taieb et al. (1) show that there is a wide degree of scatter when an extraction chromatography RIA is compared with ID-GC/MS for individual specimens. Although it does appear that extraction chromatography RIA is slightly more accurate than commercially available testosterone immunoassays, until an extraction chromatography RIA has been properly validated, results from epidemiologic studies based on these methodologies are also suspect.

How can assays that are grossly inaccurate gain approval for use in diagnosis and treatment of endocrine abnormalities? Several factors warrant consideration. In the US, the Food and Drug Administration approval process for a new diagnostic assay when there is an existing, approved diagnostic assay consists of demonstrating substantial equivalence to a predicate assay in a premarket notification 510(k) process. For testosterone, one of the predicate devices that is acceptable for demonstrating substantial equivalence is the Chiron ACS-180 testosterone assay. Several years ago, we compared the ACS-180 testosterone assay with ID-MS. The ACS-180 did not provide reliable results for female specimens (2). If the predicate device is not accurate, how can the newly designed assay hope to function properly in a clinical setting? This feature of the 510(k) process is one reason that our profession has made little progress in developing clinically acceptable testosterone immunoassays. From our clinical laboratory perspective, we suggest that predicate devices need to be validated by an independent chemical technique, preferable by a reference (or definitive) method (7, 8), before they are accepted as the standard to establish substantial equivalence. With the current regulatory environment, clinical chemistry is allowed, or perhaps even legislated, to perpetuate substandard levels of performance.

Recently, attention has focused on the need for better reporting of diagnostic accuracy of laboratory tests in peer-reviewed journals (9). Clearly, diagnostic accuracy is
different from analytical accuracy, but concepts from the STARD initiative can also be applied to improving the testing and reporting of analytical accuracy. As stated by Bruns, diagnostic accuracy compares the results of one or more tests “with a reference (‘gold’) standard in a group of patients suspected of having the condition of interest” (10). We suggest that tests of analytical accuracy should also include analysis of specimens from diseased individuals. In the case of testosterone, the immunoassays do not work in healthy females and fail miserably when used in potentially diseased females (1, 2, 6).

Laboratory professionals should not be associated with a test where an educated guess would provide an equivalent or better result.

References

David A. Herold*
Robert L. Fitzgerald
VA San Diego Healthcare System
and
University of California, San Diego
San Diego, CA 92161

*Author for correspondence. E-mail dherold@ucsd.edu.