

Analytical Commentary

Are Routine Testosterone Assays Good Enough?

Sydney S Sacks

St John of God Pathology, PO Box 646, Wembley, Perth, WA 6913, Australia

For correspondence: Dr Sydney Sacks email: sydney.sacks@sjog.org.au

Burgeoning clinical demand has driven the automation of testosterone measurement. Original assays were onerous and labour intensive and involved extraction and chromatography steps followed by RIA. As more specific antibodies were developed, kits were manufactured that dispensed with first chromatography and then extraction; so-called “direct” assays. Today most routine testosterone measurements are performed using direct automated chemiluminescent methods run on multipurpose immunoassay analysers. Analysis has become cheap, convenient, and rapid and as a result easily available, to match the very high clinical demand. In our laboratory for example, requests for testosterone increased by 2.7 times over 3 years (1999 to 2002) and have stabilised at this level since then.

With the growth in testing has come a broadening of clinical expectations and applications of the test. Initially utilised to diagnose hypogonadism in males and frank hyperandrogenism in females (i.e. levels in the normal and low male range) these same routine assays are now expected to measure accurately well down into the normal adult female range and beyond. The vogue for testosterone supplementation in perimenopausal women now demands reliable measurement into the “androgen deficient” female range to justify androgen therapy. Detection of the onset of puberty or abnormal androgenisation of children, requires measurement into the paediatric range. Do the assays come up to expectations?

Recent papers and editorials in the *Journal of Clinical Endocrinology and Metabolism* (JCEM) and *Clinical Chemistry* address this issue. Both journals critically compare testosterone results using gold standard methods with routine RIA and automated methods. Both find the state of affairs parlous. Of note is the fact that the former journal, which has primarily a clinician readership, draws attention to the failure of manufacturers (and hence laboratories and the profession!) to provide assays fit for purpose.

In the first JCEM paper, Wang et al used samples from 62 eugonadal and 60 hypogonadal men and compared results to a gold standard liquid chromatography/tandem mass spectrometry (LC/MSMS) method of proven accuracy and precision.¹ Samples were also tested using 4 automated assays, namely Roche Elecsys, Vitros ECI, Bayer Centaur and DPC Immulite 2000, and also 2 RIAs, one a research assay and the other the DPC Coat-a-count RIA. Despite good correlations between all the assays and the reference method when comparing the total range of eugonadal and hypogonadal results, all of the assays apart from the Elecsys showed significant and often variable biases. Both RIAs and the Centaur had positive bias, the Vitros and Immulite negative bias. The Immulite, despite the negative bias overall, clearly had a significant positive bias in the hypogonadal range. The lack of bias of the Elecsys may be due to this assay being standardised against isotope dilution gas chromatography mass spectrometry (ID GCMS). The authors thus demonstrate there are significant differences between results using nearly all the routine assays and the reference method. The differences due to bias were also unrelated to the differences in reference intervals of the different assays. The outcome was that the classification of samples into eugonadal and hypogonadal categories varied markedly between assays, both when the reference method reference limit of 10.4 nmol/L, or the maker’s stated reference limit, were used. This leads to the conclusion that the assays are probably only useful for correctly classifying sera if laboratories set up proper reference intervals for the assays themselves. Moreover, the percentage differences between routine and reference results showed wide scatter particularly in the hypogonadal range, implying poor precision or non-specific interferences in the routine assays. At 8 nmol/L more than 40% of routine assay results were more than 20% different from the reference assay values and in the range of <3.47 nmol/L, over 60% of results were more than 20% different. The DPC RIA and the Elecsys gave the best agreement with the reference method. The authors conclude that none of the assays, including the

RIAs, should be used to measure testosterone in the normal female or paediatric ranges. The DPC RIA, the Elecsys and Vitros are probably, however, adequate for detecting abnormal elevations in testosterone in women and children (i.e. levels above 3.7 nmol/L) by virtue of their relative lack of bias and lesser differences from the reference results.

In an editorial in the same journal, Matsumoto and Bremner stress the criticality of accurate, precise and dependable assays to endocrinology as a specialty.² They are highly critical of the processes which have led to release and continuing acceptance of routine assays which give such different and frequently incorrect results. The original testosterone RIAs used pure testosterone standards, were well validated against reference methodology and consistently had a normal range of about 10 nmol/L-34 nmol/L across methods.

They point out that current automated testosterone immunoassays use testosterone analogues as standards and generally have not been properly validated and standardised against reference methodology. All that has been necessary for American FDA approval is the demonstration that results show agreement with a previously licensed method (which may itself be biased, e.g. the ACS 180 method). Furthermore, the Australian Therapeutic Goods Administration's focus for approval of in vitro testing kits has been on the biological risk and safety aspects rather than their analytical performance. In this way the proper standardisation of testosterone assays has not been maintained and the assays have lost accuracy. This is also demonstrated by reference range drift: e.g. the manufacturer's lower reference limit for males in one assay (Vitros ECI) is 4.6 nmol/L, less than half the level previously validated by reference methods. Matsumoto and Bremner are also critical of the EQA program run by the College of American Pathologists, as assay performance is assessed relative to other laboratories using the same method (as is also the case in Australia and elsewhere) and there is no attempt made to assess absolute accuracy of QAP results against reference methodology. While this view does not take into account the problems matrix effects often cause in EQA specimens, the same criticism could be applied to many biochemistry tests in use in routine laboratories today.

The authors also argue that the diagnosis of hypogonadism requires at least two low testosterone levels as the hormone levels often fluctuate into the hypogonadal range in eugonadal men. The editorial concludes with a call to arms for the Endocrine Society to exert pressure on all concerned to improve the standardisation and dependability of all hormone measurements, including testosterone.

A similar paper in *Clinical Chemistry* by Taieb et al. compared 8 automated immunoassays (Vitros ECI, Abbott Architect, Bayer ACS 180, Immulite 2000, Vidas, Bayer Immuno 1, Bio Merieux Vidas and Auto Delfia) and 2 commercial RIAs (DPC Coat A Count and Immunotech) to ID GCMS in 50 males, 52 females and 11 children who were endocrine patients and had testosterone levels spanning high, normal and low levels.³ They too demonstrated significant bias in the routine assays, with the ACS and Auto Delfia having a positive bias over the male and female ranges, the Immulite 2000 and Immunotech RIA underestimating testosterone in the normal male range and overestimating in the female range, the Architect, Immuno 1 and DPC RIA were not biased in the male range but overestimated results in the female range, and the Elecsys showed negative bias in the male and female ranges. In the female range, testosterone were overestimated by up to 500% by some assays. The Immulite 2000 and Auto Delfia results in particular showed marked scatter in comparison to reference results in women, but none of the routine assays was regarded as satisfactory for use in the female (or paediatric) ranges. Misclassification of patients into normal and abnormal groups was also a problem with these assays whether the reference method's or the manufacturer's reference limits were applied.

In an editorial in the same issue of *Clinical Chemistry*, Herold and Fitzgerald are scathing in their criticism of the testosterone assays that were evaluated by Taieb, going so far as to demonstrate that guessing the testosterone in women was likely to be closer to the true value than measuring it with any of these assays.⁴ They conclude by recommending that laboratory professionals should not be associated with such methods!

Looking specifically at free testosterone measurement, an article by Miller et al in the same edition of *JCEM* focuses on the issue of androgen deficiency in women and its measurement.⁵ Total testosterone measurement alone is inadequate to diagnose androgen deficiency in women as most testosterone is bound to SHBG and albumin, the levels of SHBG are very variable and only the free hormone (and perhaps that bound to albumin as well, so-called bioavailable testosterone) is considered to be active. Miller et al. studied 147 women with normal and deficient androgen status and measured free testosterone using a gold standard method based on equilibrium dialysis. Results were compared to a commercial direct free testosterone RIA (Diagnostic Systems Laboratories) and two derived indices of free testosterone which use total testosterone and SHBG measurements for calculation; namely the free androgen index and the calculated free testosterone. The direct RIA method had high bias and random variability and did not correlate well with equilibrium dialysis. Calculated free testosterone

and free androgen index both correlated (equally) well with equilibrium dialysis provided reliable total testosterone (the authors used an extraction based RIA) and SHBG values were used to calculate them. Miller et al prefer calculated free testosterone as it has units that agree with those obtained by equilibrium dialysis, rather than free androgen index which is a unitless index. The direct free testosterone RIA is not recommended by them.

The editorial by Matsumoto and Bremner agrees with Miller's conclusions and quotes additional research which has demonstrated that in men, calculated free testosterone also agrees well with free testosterone by equilibrium dialysis, but free androgen index does not.²

What are the implications for our daily practice? In Australia Medicare will only fund testosterone therapy for hypogonadal men if serum levels are below 8 nmol/L. Clearly the lack of standardisation of assays means that different laboratories will give different results on the same specimen, despite the generally acceptable correlation coefficients between methods in the male range. Initiation of therapy and access to funding are thus being determined largely on the basis of analytical bias. With regard to testosterone measurement in women, the bias, precision and reference range problems outlined above will all cause frequent misclassification of patients and marked differences in classification between assays. We have taken the view that all of the routine testosterone assays evaluated above are wanting, but a few are less wanting than the rest; on the other hand some should be actively avoided. It is up to laboratories to choose the best method available, to use properly validated reference ranges, and to also report either the calculated free testosterone or free androgen index in women and the calculated free testosterone in men as estimates of free hormone levels. We must also encourage manufacturers to develop properly standardised assays with applicable reference intervals. When such assays do become available, we should do everything possible to use them and abandon methods which are an embarrassment to our profession.

References

1. Wang C, Catlin DH, Demers LM, Starcevic B, Swerdloff RS. Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. *J Clin Endocrinol Metab* 2004;89:534-43.
2. Matsumoto AM, Bremner WJ. Editorial: Serum testosterone assays - accuracy matters. *J Clin Endocrinol Metab* 2004;89:520-4.
3. Taieb J, Mathian B, Millot F, et al. Testosterone measured by ten immunoassays and by isotope dilution gas chromatography-mass spectrometry in sera from 116 men, women and children. *Clin Chem* 2003;49:1371-85.
4. Herald DA, Fitzgerald RL. Immunoassays for testosterone in women: better than a guess? *Clin Chem* 2003;49:1250-1.
5. Miller KK, Rosner W, Lee H, et al. Measurement of free testosterone in normal women and women with androgen deficiency: comparison of methods. *J Clin Endocrinol Metab* 2004;89:525-33.