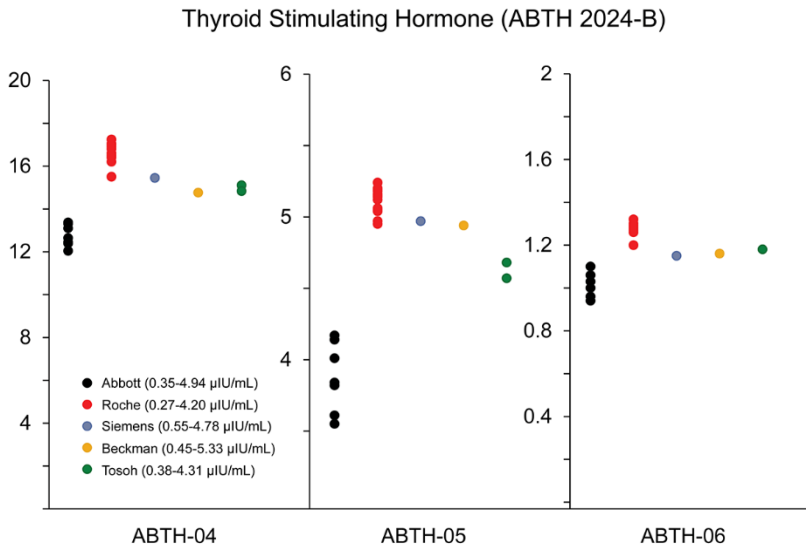
COLLEGE of AMERICAN PATHOLOGISTS

Educational Discussion

2024-B Harmonized Thyroid (ABTH)

We continue to observe biased results for thyroid stimulating hormone (TSH) across platforms. An earlier method comparison from the literature demonstrated that results from one platform were 14 times higher than another (1). Those discrepancies were published 20 years ago and their implications have recently been highlighted in the endocrinology literature as an important consideration in clinical care (2). The lack of concordance makes it impractical to design meaningful clinical practice guidelines for screening and therapeutic monitoring using TSH. One of the goals of the Accuracy-Based Proficiency Testing Program is to help elucidate the differences that exist between platforms in the hopes of establishing a baseline from which we could improve concordance and clinical care. These efforts rely on using actual human samples treated carefully to ensure commutability across platforms.

As we have mentioned before, creating matrix-matched commutable Survey materials for TSH and other thyroid function testing has been trickier than for other Surveys. For many years, we drew specimens from healthy donors, whose TSH values were almost always in the reference interval. By pooling donor samples with the lowest, highest, or mid-range TSH values, we were hoping to obtain materials with a wide range of values for each test in the ABTH Survey. Unfortunately, the range we observed was much narrower than we would have liked. It was a frustrating realization, considering the success we have had with steroid hormones, hemoglobin A1c, lipids, and vitamin D.
To get around the problem, we collaborated with an endocrinologist at Tufts Medical Center in Boston, Massachusetts, who identified patients in her practice with treated thyroid disease, from whom she obtained informed consent to donate several additional tubes of blood after a routine clinic visit. The samples were sent frozen to a central site, where they were pooled, homogenized, and separated into aliquots before being mailed out to Survey participants. The pooling strategy allowed us to achieve sufficient volumes of serum with a much broader range of TSH concentrations than our previous approach, with the highest TSH concentration of the three challenges in this mailing above 10 µIU/mL.



The figure illustrates the reported concentrations (in µIUmL) from the three matrix-matched samples included in the ABTH 2024-B survey. The 26 results were from 5 different manufacturer's instruments and the reference ranges are indicated for each.

The results are instructive. There is clearly significant bias that remains, even after 20 years. Since there is no reference method for TSH, it is not possible to define the *correct* result, but it is obvious that there is much work to be done if we ever hope to see consistent results from laboratory to laboratory and medical center to medical center. Efforts at harmonization must continue for this analyte so that reliable clinical cut-offs can be defined.

It is important to discuss the reference ranges listed in the package inserts distributed by the manufacturers. It would be expected that differences in reagents and calibration would lead to manufacturer-determined reference ranges that would classify patients similarly across platforms (even without cut-offs driven by the clinical guidance documents), but that is not the case. Instead, for example, the upper limit of the reference range for Abbott is higher than for Roche, even though their results are significantly lower (further, all of Abbott's results are lower than those from the other platforms). As a result, relying on manufacturers' reference ranges to normalize clinical practice and outcomes across platforms will not succeed.

Unfortunately, we have only one peer group reflected in the accompanying report. The CAP statisticians require **at least 3 participants** to report minimum and maximum values and **at least 10 participants** for additional statistics. As presented in the Figure above, we do have data from other platforms, but there are too few participants to draw any statistical conclusions. It would be extremely valuable to have more participants, so please spread the word about this Accuracy-Based Survey. We also hoped to achieve a range of concentrations for the other thyroid function tests in order to compare results across platforms in different disease states, but our pooling strategy appears to

have averaged out the concentrations and the three samples are each similar for all 4 analytes. However, there is still something to be learned. When we compare the results for each analyte individually, on average the difference between the highest and lowest platforms was larger than might be expected, at 27.5% for Total T3, 35.8% for Free T3, 37.8% for Free T4, and 42.6% for Total T4, which indicates poor harmonization for these analytes as well.

The take home message is that it is important for laboratories to verify manufacturers' proposed reference intervals in their own clinical population and to communicate with clinical colleagues to ensure that your reported values and reference intervals are meeting clinical needs at your institutions.

(1) Rawlins ML and Roberts WL. Performance characteristics of six third-generation assays for thyroid-stimulating hormone. Clinical Chemistry 2004;50:2338-44.

(2) Kalaria T, Sanders A, Fenn J, et al. The diagnosis and management of subclinical hypothyroidism is assay-dependent – Implications for clinical practice. Clinical Endocrinology 2021;94:1012-1016.

Andrew Hoofnagle, MD, PhD, FCAP, Chair
Accuracy-Based Programs Committee