Linking Medical Needs and Performance Goals: Clinical and Laboratory Perspectives on Thyroid Disease

Ian D. Hay¹ and George G. Klee²,³

Diagnosis and management of thyroid disease benefit substantially from close interactions between clinicians and laboratorians. Clinicians desire reliable tests for diagnosis and management of both hypothyroidism and hyperthyroidism. Traditionally, multiformed methods have been used, beginning with an index of free thyroxine concentration, which, if abnormal, has been followed by basal thyrotropin (TSH) measurements (for hypothyroidism) or thyrotropin-releasing hormone-stimulated TSH measurements (for hyperthyroidism). Improvements in analytical methods for TSH have made it practical to use basal TSH measurements in serum as the first-line test for thyroid disease. However, performance guidelines are needed, because not all TSH assays work well in this role. The performance guidelines developed by the American Thyroid Association are reviewed to illustrate how they help to assure consistent analytical testing. Further improvements in TSH assays (third- and fourth-generation assays), which support measurements down to 0.01 mIU/L and 0.001 mIU/L, may provide additional advantages for classifying thyroid patients and monitoring thyroxine-suppressive therapy in patients with thyroid cancer. The potential advantages of these newer assays are illustrated with case examples. Additional analytical performance monitors are proposed to help ensure that these next-generation TSH assays meet the expanded clinical needs.

Indexing Terms: thyroxine · thyrotropin · thyroid status · immunometric assays

Clinical Perspective

It has long been recognized that, after diabetes mellitus, the most frequently diagnosed endocrine disorders are those of the thyroid gland. More than 40% of healthy adults studied by real-time ultrasound imaging show thyroid nodular disease, and ~10% of the population at large possess circulating thyroid autoantibodies, important predictors of the autoimmune thyroid diseases that represent the commonest causes for thyroid dysfunction in iodine-sufficient areas (1). Recent estimates suggest that in US and European communities the prevalence rates for hyper- and hypothyroidism may be as high as 1% and 2%, respectively.

In past years the diagnosis of thyroid dysfunction might have required the use of multiple indirect tests of thyroid function and the in vivo administration of radioactive ¹³¹I for uptake studies (2). Particularly over the last 20 years, the thyroid-function testing armamentarium has grown considerably, but the choice of tests used by clinicians to diagnose thyroid dysfunction has fortunately been made easier as the available in vitro thyroid tests have become more direct, sensitive, and cost effective.

In the assessment of thyroid disease, the principal use for laboratory testing is to confirm or refute a clinical suspicion of either thyroid hypofunction or thyrotoxicosis (hyperthyroidism). It is the typical practice for clinicians to order a screening test and, if the result proves abnormal, to consider a second (or third) confirmatory test. The principal secreted thyroid hormone is thyroxine (T₄), and dogma dictates that only the free T₄ (FT₄) is responsible for hormone action and clinical thyroid status, so theoretically a measurement of FT₄ would represent an ideal screening test (3). However, the technical difficulties of "direct" FT₄ assays have led most clinicians in past years to depend more on either a total T₄ or an "indirect" FT₄ method (4).

Because the thyroid gland is controlled by secretion of thyrotropin (thyroid-stimulating hormone; TSH) from the anterior pituitary and, in the face of an intact pituitary "sensor," the serum TSH concentration is inversely related to the FT₄ concentration, an alternative method to screen for thyroid dysfunction is to use a "TSH-first" approach (5). With the advent of more-sensitive TSH immunometric assays, such an approach is proving more popular in the 1990s, and clinicians currently are in the process of considering a switch away from the time-proven T₄- or FT₄-based strategies.

Further details of the testing algorithms used during the past decade, and favored by the American Thyroid Association in its 1990 guidelines for clinicians (6), are outlined below. Consideration also is given to the newer TSH-based philosophy and the potential role that the latest generations of TSH assays may play in clarifying currently hazy areas of thyroid dysfunction and disease management (7).

Diagnosis of Hypothyroidism

Figure 1 outlines the typical laboratory evaluation of a patient clinically suspected to be hypothyroid. In this example, the primary (screening) test is an FT₄ estimate (FT₄E), here shown as the FT₄ index derived from the T₄

¹ Division of Endocrinology and Metabolism, Department of Internal Medicine, and ² Division of Metabolic and Hematologic Biochemistry, Department of Laboratory Medicine and Pathology, Mayo Clinic, and Mayo Foundation, Rochester, MN 55905.
² Address correspondence to this author, at: Department of Laboratory Medicine and Pathology, Mayo Clinic, 200 First St., SW, Rochester, MN 55905.
³ Received December 17, 1992; accepted March 18, 1993.

4 Nonstandard abbreviations: T₄, thyroxine; FT₄, free thyroxine; FT₄E, free thyroxine estimate; T₃, triiodothyronine; TSH, thyrotropin; and TRH, thyrotropin-releasing hormone.
and the thyroid hormone binding ratio. However, an alternative might be to use one of the "two-step" sequential FT₄E methods based on the "labeled hormone backtitration" principle (1). If the FT₄E result proves to be subnormal or marginal, the "back-up" test is serum TSH, and in the majority of patients with thyroid deficiency (who have a primary thyroid disease) the TSH value would be increased (as in the extreme right of the flow diagram). If, however, despite a low FT₄E the TSH is either borderline high or inappropriately normal, consideration should be given to a possible central (hypothalamic-pituitary) cause for the hypothyroidism. In this rare situation other anterior pituitary functions should be assessed and a thyrotropin-releasing hormone (TRH) stimulation test may help prove the presence of a pituitary thyrotrropic deficiency state.

The association of a marginal (low normal) FT₄E and borderline increased serum TSH concentration may be considered either "compensated euthyroidism" or "subclinical hypothyroidism." Such a situation may often be seen after iatrogenic thyroid damage by sublethal doses of ¹³¹I or subtotal/partial thyroidectomy. In the absence of such a history, the possibility of chronic autoimmune thyroiditis of the Hashimoto type should be a consideration, and confirmatory testing would include assessment either by thyroid autoantibodies (against thyroglobulin or thyroid peroxidase) or, rarely, by a fine-needle aspiration biopsy demonstrating the presence of lymphocytic thyroiditis and follicular epithelial oxphilia.

Diagnosis of Thyrotoxicosis

Figure 2 demonstrates the laboratory evaluation conventionally performed in a clinically thyrotoxic patient. In almost all thyrotoxic patients the value for FT₄E will be above normal; in this circumstance (especially if the patient has typical clinical features of Graves disease), no further in vitro testing may be necessary. However, a sensitive TSH test result <0.1 IU/L would certainly exclude the rare possibility of a TSH-induced thyrotoxicosis caused by primary pituitary disease. If the FT₄E is either normal or borderline increased, assay of serum triiodothyronine (T₃), or free T₃, should be ordered to determine whether prefiamental T₃ secretion or T₃ ingestion may have resulted in so-called T₃ toxicsosis.

In former days, if the value for neither FT₄E nor T₃ was above normal, clinicians would resort to a TRH stimulation test to confirm or refute a thyrotoxic diagnosis. As more-sensitive TSH immunometric assays have become readily available, an accurate basal TSH has virtually replaced the TRH test in this algorithm. However, a suppressed (subnormal) "more-sensitive" TSH concentration associated with normal concentrations of T₃ and FT₄E may signify only subclinical hyperthyroidism, possibly due to autonomous thyroid function, as may be seen in adenomatous goiter or autoimmune thyroid disease. Such a circumstance may also denote the fact that the patient recently underwent a thyrotoxic episode, e.g., spontaneously resolving thyroidosis of subacute thyroiditis or, alternatively, the patient may be at risk of future hyperthyroidism, e.g., an early relapse of Graves disease or a potentially toxic autonomously functioning nodular (adenomatous) goiter. Identification of the specific etiologies for such a profile may be aided by the measurement of in vivo ¹³¹I-uptake or by obtaining a radionuclide scan of the thyroid.

"TSH-First" Testing Strategy

A scheme for using a more-sensitive TSH assay as the initial (screening) test is illustrated by Figure 3. In such a scheme, most ambulatory patients would be predicted to have no thyroid dysfunction and would not require further testing unless there was suspicion of hypothalamic-pituitary disease. Such a TSH-based strategy would have the attraction of detecting patients with subclinical hyper- or hypothyroidism in whom, by definition, the concentration of total or free thyroid hormones would be normal.

The potential advantage of the newer TSH assays, capable of measuring concentrations <0.01 mIU/L, would be to better distinguish those patients who in
earlier immunometric assays would have been classified as either subnormal or undetectable. Many of the currently available TSH assays cannot reliably differentiate between the mildly subnormal values (0.01–0.1 mIU/L) seen in hospitalized or T4-treated patients and the profoundly low values (<0.01 mIU/L) found in most thyrotoxic patients (4). The more general availability of “third- and fourth-generation” assays may permit an easier distinction between mildly subnormal TSH concentrations (such as may be seen with nonthyroidal illnesses or in therapy with pharmacological steroid doses) and frankly hyperthyroid values (9). However, it remains to be proven whether the quantification of various degrees of TSH suppression in subclinical hyperthyroidism will significantly advance the management of patients with thyroid disease (9).

Laboratory Perspective

The two major areas for laboratory support in the diagnosis and management of thyroid disease are tests for hypothyroidism and tests for thyrotoxicosis. The traditional approach to these disorders has been sequential testing: beginning with an FTLE, followed by TSH or TRH-stimulated TSH, followed by an ancillary test such as for anti-thyroid antibodies and thyroid biopsy or imaging studies. Improvements in the detection limit and precision of TSH assays have expanded the options for laboratory testing to make TSH-first testing practical. This new approach to thyroid testing, however, requires more stringent performance goals. Not all TSH assays labeled as “sensitive” or “ultrasensitive” perform equally well in these expanded roles. To help in the standardization of thyroid tests, the American Thyroid Association published in 1991 (4) an assessment of current free thyroid hormone and TSH measurements and defined analytical performance goals for future clinical assays. These guidelines, which were developed by a combined group of clinicians and laboratorians, provide a good example of the linkage of medical needs with laboratory performance goals.

The expanded role of TSH as a first-line thyroid test was driven by the analytical improvements, which provided reliable measurements down to 0.1 mIU/L (second-generation assay). Analytical improvements have continued, and some assays can now reliably measure concentrations down to 0.002 mIU/L. These assays may provide further information for the classification of hyperthyroidism and the monitoring of T4 suppression therapy for thyroid cancers. It is unlikely that this next level of analytical improvement will have as major an impact on thyroid testing as the initial improvement had, but this next level of application also will require more specific analytical goals to ensure that the assays will meet the clinical needs.

Justification for Expanded Role of Sensitive TSH Assays

Figure 4, which is reprinted from a 1971 article, shows that early TSH radioimmunoassays (RIAs) there was a demonstrable inverse relationship between circulating T4 and TSH concentrations (10). Low T4 concentrations were associated with increased TSH values; when the T4 concentration rose, the TSH concentrations in serum decreased. The earliest TSH RIAs could not measure TSH concentrations <3 mIU/L, but by extrapolating the measurable data it was postulated that TSH concentrations would continue to decrease when the T4 concentration exceeded the normal set point of an individual.

Before the development of sensitive TSH assays, the “gold standard” for confirming or refuting a clinical diagnosis of hyperthyroidism was an intravenous stimulation test with TRH and measurement of the serum TSH response. Euthyroid subjects had a stimulated TSH response of between 2 and 35 mIU/L, whereas hypothyroid subjects had a larger (exaggerated) response and hyperthyroid subjects typically had either no response or a response <1 mIU/L. Stimulated TSH responses between 1 and 2 mIU/L were considered suppressed and

![Figure 3. TSH-based strategy for the laboratory investigation of thyroid function](http://example.com/image3)

![Figure 4. Relation of TSH and total T4 in patients on various doses of T4 as measured with a "first-generation" TSH assay](http://example.com/image4)
the patients were presumed to be borderline hyperthyroid (5).

With a sensitive TSH assay, investigators found that basal TSH measurements were predictive of the peak (30-min) TSH response (5). Figure 5 shows a cross-classification of basal TSH vs TSH response to TRH stimulation for 150 patients with basal TSH values ≤1.0 mIU/L. No patients with a basal TSH ≤0.1 mIU/L responded to TRH stimulation. All patients with a basal value ≥0.40 mIU/L had at least a borderline response. Therefore, with a sensitive TSH assay a large proportion of patients with basal TSH values <0.1 mIU/L or >0.4 mIU/L do not need TRH-stimulation tests because the TSH response can be accurately predicted from the basal TSH measurements.

Guidelines for Sensitive TSH Assays

The Committee on Nomenclature of the American Thyroid Association developed performance guidelines for free thyroid hormone and TSH assays to help ensure consistency in laboratory measurements. The guidelines for TSH assays were intended to ensure that the measurements function well for diagnosing both hypothyroidism and hyperthyroidism. The complete guidelines can be found in the published report (6), but the highlights are summarized below:

1) The manufacturer should define the functional detection limit, i.e., the lowest concentration that can be measured with an intra-assay CV of ≤20% (11).
2) The cross-reactivity with lutropin, follitropin, and human chorionic gonadotropin should be <0.1%.
3) Values for dilutions should be linear within ±10%.
4) Analytical recovery should be within ±10% of added value.
5) Accuracy of average results should be within ±5%.
6) Interassay precision (CV) should be <20% down to 0.1 mIU/L.

![Figure 5](image-url)

**Boots-Celitech basal TSH, mIU/L**

- **Responders** indicate a TSH response ≤0.2 mIU/L.
- **Borderline** indicates a TSH response of 0.2-0.4 mIU/L.
- **Suppressed** indicates a TSH response >0.4 mIU/L.

7) Sera from thyrotoxic subjects should be within ±5% of the original for the assay zero.
8) There should be no significant interference from heterophile antibodies.
9) The standard curve should not “hook” up to 300 mIU/L.
10) Less than 5% (preferably <1%) overlap in serum TSH values should exist between distribution of hyperthyroid and euthyroid individuals.

The Association's report stated that assays meeting these criteria should work well in ambulatory patients, but may not reliably classify hospitalized or T4-treated patients. It was predicted that newer assays with improved analytical sensitivity may provide further information in patients with subclinical hyperthyroidism.

Development of Third- and Fourth-Generation TSH Assays

Nicoloff and Spencer (7) coined the term “assay generation” to describe the analytical sensitivity of TSH assays. Each generation of the assay has a 10-fold increase in analytical sensitivity. Thus, a first-generation assay can measure down to 1.0 mIU/L with a 20% interassay CV; a second-generation assay can measure down to 0.1 mIU/L; a third-generation assay can measure to 0.01 mIU/L; and a fourth-generation assay, to 0.001 mIU/L.

We (12) recently developed a TSH assay that can measure concentrations as low as 0.007 mIU/L, with a 20% CV. The assay utilizes a preincubation step before the analysis on the ACS-180 instrument from Ciba Corning Diagnostics. The development and validation of this assay identified measurement questions that become even more important when assays are optimized for maximum analytical sensitivity. Some of these questions are highlighted here as potential issues for evaluating third- and fourth-generation assays.

- The matrix used for assay standards can significantly alter the definition of “zero” TSH. The International Federation for Clinical Chemistry has recently investigated this question and concluded that, although the matrix substantially affects assays, standardization of the matrix does not assure uniformity across assays because of the differences in the reactions of various assays with patients' specimens (13).
- The choice of the World Health Organization (WHO) standard can alter TSH values. Assays standardized with WHO 1st IRP (88/38) give values 15% to 20% higher than assays standardized with the WHO 2nd IRP (80/568 or 81/565).
- Cross-reactivity with related hormones must be exquisitely low, preferably <0.001%. If cross-reactivity is 0.1%, as recommended by the American Thyroid Association, a cross-reacting concentration equivalent to 100 mIU/L could produce TSH changes as large as 0.1 mIU/L, which would obscure the significance of low-end measurements.
- Small titers of heterophile antibodies can significantly alter the very low measurements. We found that both nonimmune mouse IgG and nonimmune sheep IgG were needed to neutralize these heterophile antibodies.
• The specifications for lot-to-lot differences in reagent preparations must be carefully controlled, because small differences can significantly alter the low-concentration measurements.

• The accuracy of instrument calibrations is critical. Replicate measurements of the calibrators may be needed to assure that the instrument is matched to the standard curve within tight tolerance limits. The accuracy of the standard curve also is critical.

Clinical Utility of Third- and Fourth-Generation TSH Assays

Spencer et al. (8) have shown that basal TSH measurements down to 0.005 mIU/L continue to correlate with TRH-stimulated TSH responses over the range of 0.01 to 1.0 mIU/L. They also have provided preliminary data that these low-level measurements are better for evaluating the thyroid status of hospitalized patients. Further studies in other medical centers will be needed to validate these findings.

Three potential roles for the latest TSH assays are (a) to provide further assurance in the detection of hyperthyroidism based on a 0.1 mIU/L decision point, (b) to permit subclassification of hyperthyroid patients according to the degree of TSH suppression, and (c) to more accurately monitor thyroid hormone therapy in thyroid-cancer patients, who may require nearly complete suppression of TSH. Definitive studies are not available for these potential roles, but we present three anecdotal cases to illustrate these potential uses:

Case I. This 51-year-old woman was losing weight and had “suppressed” TSH (reported TSH, <0.05 mIU/L; enhanced TSH = 0.046 mIU/L). A repeat routine test ordered by the clinician 4 days later gave 0.12 mIU/L.

Because the initial test did not match the clinical presentation, the clinician in this case decided to repeat the TSH assay, even though the initial value was <0.1 mIU/L. The more quantitative (enhanced) test result showed that the value was, in fact, just below the measurement threshold of the second-generation assays. Knowing that the value was close to the decision threshold made it easier for the clinician to recognize that the patient's TSH was not fully suppressed, and therefore this patient was unlikely to be clinically thyrotoxic.

Case II. This 66-year-old woman had a history of primary hypothyroidism on replacement with levothyroxine, 0.1 mg/day. Her test results were: $T_4 = 98 \mu g/L$, $TSH <0.05$ mIU/L, enhanced TSH $<0.002$ mIU/L.

There is not yet universal agreement that $T_4$ therapy dosage in primary hypothyroidism should be adjusted to provide normal TSH concentrations. However, having the knowledge that the serum TSH is suppressed to $<0.002$ mIU/L may influence the clinician to reevaluate the potential risks of prolonged overreplacement, and probably decrease the daily dose of $T_4$ to permit a TSH value within the euthyroid range.

Case III. This 71-year-old man with aggressively metastatic follicular thyroid cancer was on TSH-suppressive therapy with 0.175 mg of levothyroxine per day. His test results were: $T_4 = 85 \mu g/L$, $TSH <0.05$ mIU/L, enhanced TSH $= 0.032$ mIU/L.

Patients with a history of thyroid cancer often are maintained on TSH-suppressive doses of $T_4$. Although this patient had a suppressed TSH value by the second-generation assay, measurement with the more-sensitive assay showed a small but measurable TSH. The quantitative TSH measurement provides the clinician with the option of adjusting the $T_4$ dose. The degree of TSH suppression should probably be judged from weighing the risk of tumor recurrence against concerns regarding the possible cardiac, hepatic, and bone effects that may result from iatrogenic subclinical hyperthyroidism. In this case, because of the high risk of cancer mortality, the clinician elected to increase the $T_4$ dose to suppress the basal TSH concentration in sera to the $<0.01$ mIU/L range.

The laboratory support for thyroid-disease management has undergone major changes in the past few years. The newer “sensitive” TSH assays are used not only to confirm primary hypothyroidism, but also as frontline tests for both hypothyroidism and thyrotoxicosis. Not all TSH assays work well in this expanded role, so analytical performance guidelines are needed. The American Thyroid Association helped fill this need by recommending accurate performance guidelines, which include both clinical and analytical benchmarks. TSH assays are continuing to improve, and third- and fourth-generation TSH assays provide additional potential advantages, especially in distinguishing mildly subnormal TSH concentrations from frankly hyperthyroid values. These potential advantages have been illustrated here with specific case studies, and additional laboratory performance monitors for these further roles have been outlined.

References
10. Cotton GE, Gorman CA, Mayberry WE. Suppression of thyrotropin (h-TSH) in sera of patients with myxedema of varying

Discussion

Robert Rej: What is the difference between short-term bias and day-to-day imprecision?

Ian Hay: I think yours is a semantic question. What I define as short-term bias is a shift that can occur in a method after the laboratory has set up the procedure and collected the data on which the decision points are based. So, short-term bias is the difference between the true value as defined when the method was established, and what it is now, whereas the analytical variability is a distribution of negative and positive components, which cancel each other in the long run. With short-term bias, the canceling effect is absent; therefore, this bias can have a major effect on the decision value. I think short-term bias needs to be split out. I don’t think one can lump it all into variability.

Neal Dawson: I commend you for the work you are doing. I think the task you are taking is an excellent one. My question has to do with the fact that most of the prototypical curves we present when we look at decision-making have prevalences of 50–50. The curves are presented as having the same area under them, when, in fact, the prevalence of the diseases we look for is minuscule with respect to the population that could be tested. You talked about looking at specificity because that is what causes commotion among clinicians: They tend to notice when they get a lot of false positives. However, if you drift in the other direction, the diminishment of sensitivity may be much harder to detect, although basically, the heart of why you are doing the test at all is to detect the disease, if any. What are your thoughts on that?

George Klee: False-positive rates or specificity rates are independent of prevalence; however, when one combines sensitivity and specificity, prevalence becomes important. One problem with analyzing sensitivity is that many laboratory tests are used for multiple diseases or multiple purposes clinically. So our analysis has to include all different types of settings. This fact becomes a major multiplier in the amount of work that has to be done—probably 10-fold more to define the sensitivity issue than to get at the specificity issue. The other side of that is that specificity appears to be much more sensitive to small changes in the setting of the decision point. Decision limits are generally set in the tail of the distribution of the population of patients with the disease. When it is shifted slightly, we change from picking up 80% to maybe 75% of the diseased patients, but we will also change from having 3–5% false-positives up to 7–8%. The relative change, the relative magnitude of those two things, means that moving the decision limit doubles the number of false positives but is generally decreasing the sensitivity by only a few percentage points.