Current status and performance goals for serum thyrotropin (TSH) assays

CAROLE A. SPENCER, Michael Takeuchi, and Margarita Kazarosyan

Current medical needs dictate that laboratories offer thyrotpin (TSH) assays that can reliably measure low TSH concentrations—a prerequisite for using the more cost-effective TSH-centered strategy currently recommended by the American Thyroid Association. This study reviews the functional performance of the TSH immunometric assay methods currently used in clinical practice. Methodological differences between methods, the rationale for using the 20% interassay CV as the functional sensitivity limit for patient reports, and both TSH-related and non-TSH-related specificity problems are reviewed. We recommend that manufacturers and clinical laboratories use a clinically relevant standard protocol for functional sensitivity assessments. In this protocol, human serum pools are analyzed in random order 10 or more times across a clinically realistic time span (which approximates to 6–8 weeks for TSH measurements used in an outpatient setting). Laboratories should independently establish and periodically check their functional sensitivity by the standard protocol and enlist the manufacturer’s help to accomplish this if necessary. Manufacturers should (a) develop promotional material that realistically projects the assay’s functional sensitivity expected in a clinical laboratory setting, (b) ensure that new clinical laboratory users achieve the projected functional sensitivity target when using the standard protocol, and (c) focus on the typical functional sensitivity obtained by clinical laboratory users rather than the assay’s “generation” achieved under ideal conditions. If manufacturers and laboratories collaborate to solve the sensitivity and specificity problems discussed here, clinical laboratories should be better able to consistently deliver reliable serum TSH measurements across the full range of TSH concentrations encountered in clinical practice.

INDEXING TERMS: functional sensitivity • immunoassays

__________

General Clinical Research Center 6602, Department of Medicine, University of Southern California, 2025 Zonal Ave., Los Angeles, CA 90033.

*Author for correspondence. Fax 213-226-4969; e-mail spencer@hsc.usc.edu. Received July 20, 1995; accepted October 23, 1995.

The sensitivity of thyrotpin (TSH) measurement has improved 100-fold over the last 20 years, from a sensitivity of 1 to 2 mU/L, typical of the first RIA methods developed in the early 1970s [1, 2], to between 0.01 and 0.02 mU/L, currently achieved by some nonisotopic immunometric assay (IMA) methods [3, 4]. Historically, TSH assay “quality” has been judged on the basis of clinical sensitivity—the ability to distinguish euthyroid concentrations (−0.4−4.0 mU/L) from the profoundly low TSH concentrations typical of overt Graves thyrotoxicosis (<0.01 mU/L) [5]. Although this clinical benchmark is critical, laboratories also need methodological indices of sensitivity with which to evaluate different methods. Over time, these indices have changed to improve their correlation with clinical performance. Specifically, the American Thyroid Association now considers that estimates of intraassay analytical sensitivity (limit of detection), based on the imprecision of the zero matrix [6], have little clinical relevance and recommends that analytical sensitivity estimates be replaced by an estimate of functional sensitivity (limit of quantification), based on low-end interassay precision [7].

Functional sensitivity is currently defined as the point of the interassay precision profile where the CV = 20% [8, 9]. The 20% CV definition was selected on the basis of both methodological and biological principles. In a previous study of the performance characteristics of a range of TSH IMA methods, the TSH concentration associated with the 20% interassay CV point was consistently above the experimentally determined analytical sensitivity limit [9]. Therefore, the selection of 20% CV TSH value as the lower assay limit ensures that any TSH concentration above this is indeed measurable. Despite a variety of strategies used to define desirable standards for laboratory tests, the broad consensus appears to be that optimal imprecision should be less than one-half of the within-subject biological variation [10]. Previous studies suggested that the 20% CV limit was compatible this principle. Specifically, an evaluation of batchwise serum TSH measurements made in specimens drawn from euthyroid subjects at 4-h intervals for 6 days (at clock times

1 Nonstandard abbreviations: TSH, thyrotpin; IMA, immunometric assay; T4, thyroxine; T3, triiodothyronine; TRH, thyrotoxicosis; ICMA, immunochnemiluminometric assay; and NTI, nonthyroidal illness.
between 0800 and 1800—an interval comparable with the outpatient setting) found within-subject biological variation averaged 31% (range 14–45%) [11].

TSH assay nomenclature has evolved in parallel with the refinement in the measures of methodological sensitivity. When the IMA methods were first developed in the mid-1980s, they were distinguished from the less-sensitive RIA methods then still in use by being called “sensitive” or “ultrasensitive” assays [12]. When it became apparent that the new IMA methods themselves could differ in sensitivity by an order of magnitude, a “generational” nomenclature system was introduced [9, 13], in each generation represented a 10-fold improvement in functional sensitivity. “First-generation” assays had functional sensitivity between 1.0 and 2.0 mU/L, whereas “second-” or “third-generation” assays had functional sensitivities of between 0.1–0.2 and 0.01–0.02 mU/L, respectively. “Fourth-generation” research assays capable of functional sensitivity in the 0.001–0.002 mU/L range have now also been developed [5]. Unfortunately, the scientific value of the generation system has been eroded by commercial marketing practices and is now viewed as being as misleading as the descriptive “ultrasensitive” nomenclature it was designed to replace [4]. Further, methods can have functional sensitivity intermediate between “generations” and operate with an inferior sensitivity (sometimes 10-fold poorer) in clinical practice as compared with that achieved by the manufacturer’s laboratory [4].

Euthyroid TSH reference intervals have also progressively contracted—from early RIA estimates of 2.0–15.2 mU/L [2] to between 0.4–4.0 mU/L as reported by most current methods [Fig. 1] [4]. This refinement in the reference range reflects three factors: the recognition that euthyroid serum TSH concentrations are log-gaussian distributed; the use of sensitive thyroid antibody tests for excluding individuals with subclinical thyroid disease, whose inclusion previously skewed the reference inter-

val; and the use of a monoclonal capture antibody with specificity for the β-subunit of TSH, which has now eliminated cross-reactivity with the other pituitary glycoprotein hormones [14]. Although some monoclonal capture antibodies detect differences between serum isofoms and the pituitary isofoms present in TSH standards [15], the impact of epitope specificity differences is small and clinically insignificant. Unfortunately, not all specificity differences are minor or TSH-antibody related. Specificity loss can result from heterophile antibodies, which can interfere with any two-site IMA method [16, 17], and from other poorly defined non-TSH-related serum or preanalytical factors, which are especially impairing when measuring low TSH concentrations associated with a low signal [18].

**Current TSH IMA Methods**

Immunometric methods are noncompetitive assays that use an excess of TSH antibody, usually a monoclonal (s), bound to a solid support [19]. This capture antibody extracts TSH molecules from the serum specimen while a labeled TSH antibody (mono- or polyclonal) simultaneously binds the captured TSH molecules at a different epitope(s). The label can be one of a range of signal-producers—an isotope, usually 125I (in immunoradiometric assays, IRMAs); an enzyme (in immunoenzymometric assays, IEMAs); a fluorophor (in immunofluorometric assays, IFMAs) [20]; or a chemiluminescent molecule (in immunochemilumimetric assays, ICIMAs) [18]. After an incubation period of <0.5 to 2 h, the capture and labeled antibodies form a “sandwich” with serum TSH molecules: capture antibody—TSH—labeled antibody. After the unbound reactants are washed away, the label bound to the solid support is measured, and is directly proportional to the serum TSH concentration [21].

TSH IMAs have now replaced virtually all RIA methods used in clinical practice because they offer more sensitivity than RIA [22, 23] as well as the practical advantages of a shorter incubation time, a wider working range, a near-linear calibration curve (which facilitates data reduction), and a more stable labeled-antibody reagent. Many nonsotopic IMAs are now automated and use a one- or two-point calibration, which, although convenient, may not always produce optimal low-end precision [24]. Nonsotopic IMAs, especially ICIMAs, are increasingly dominating the market, because chemiluminescence is thought to offer the most sensitivity potential. One should recognize, however, that the use of a chemiluminescent signal per se does not guarantee the best sensitivity. Different ICIMAs use different chemiluminescent molecules that, when triggered, emit a flash type of signal or utilize an enzyme (e.g., alkaline phosphatase) that acts on a substrate to produce a sustained chemiluminescent glow as the signal [25]. The sensitivity of different ICIMAs is determined not only by the inherent chemistry, but also by the instrumentation. Although the “hook” effect potentially affects all high-dose IMA measurements [26], this does not appear to be a problem for TSH IMA measurements, presumably because the highest TSH concentrations encountered in clinical practice rarely exceed 100-times the upper reference limit [5], concentrations that are apparently within the binding capacity of the monoclonal capture antibody.

The reference intervals for different TSH IMA methods are
now relatively comparable (Fig. 1); however, methods differ with respect to their practical execution (e.g., specimen volume, range, incubation times, automation, cost) as well as their reliability of low-range measurement [4]. Laboratory managers currently find it difficult to predict how a method will perform in the laboratory for two reasons: (a) variation in the method's functional sensitivity among the laboratories using that manufacturer's method, presumably because of instrument differences and other less well-defined variables, and (b) misleading information in the manufacturers' package inserts when performance is described in terms of analytical sensitivity or the "generation" achieved under optimal conditions.

SENSITIVITY
As described above, functional sensitivity has undergone redefinition designed to improve clinical relevance [8, 27, 28], currently being the point on the interassay precision profile at which CV = 20%. Clearly, the time span used for interassay precision assessment is critical if functional sensitivity estimates are to be clinically relevant. As shown in Fig. 2, interassay precision profiles erode substantially over time, even when a so-called third-generation assay is used. This erosion of functional sensitivity probably reflects cumulative effects of multiple expected sources of variation (reagent batch, instrument calibration, technical operator) as well as poorly defined (temperature, voltage, reagent aging, etc.) variables. The Fig. 2 data emphasizes the importance of determining interassay precision across a time span that relates to clinical practice (usually 6 to 8 weeks for TSH used in an outpatient setting). Furthermore, since the precision of measuring TSH in modified protein matrices may be overly optimistic compared with that of measurements in human serum pools [28], a clinically realistic precision estimate will be obtained only when unmodified human sera are tested. A further constraint needed to mimic interassay precision in clinical practice would be to randomly analyze the serum pools to reflect any carryover, which would have the greatest negative impact on low-range precision.

As one might expect, there is a relation between the functional sensitivity potential of a method and its ability to deliver reliable low-range measurement [4]. Fig. 1 displays the data from human serum pools having a target TSH value in the 0.02–0.04 mU/L range. These pools were measured as unknown clinical specimens by several clinical laboratories in the UK and US, using 1 of 16 different TSH IMA methods [4]. The methods that had demonstrated third-generation functional sensitivity potential produced the fewest discordantly high (>0.1 mU/L) TSH values for these sera with low TSH concentrations.

SPECIFICITY
TSH-related as well as TSH-unrelated specificity differences exist for most methods. TSH-related immunoassay differences stem from differences in epitope specificity [28] for the different isoforms present in different sources of TSH (i.e., serum, pituitary extract, or recombinant human TSH preparations) [14, 15, 29, 30]. These isoforms, the spectrum of which may reflect the underlying pathology, differ with respect to their glycosyl residues, which influence bioactivity [28, 30, 31, 32].

IMAs using monoclonal antibodies are more prone to non-TSH-related specificity loss, which is most often marked by inappropriately high serum TSH concentrations. Because low-range TSH concentrations have the lowest signal, samples with low TSH content are most susceptible to such interferences leading to a discordant serum TSH report (not necessarily increased but maybe merely inappropriately normal). Heterophile antibody interference, a well-recognized non-TSH cause of specificity loss [16, 17, 33], is generally detected from a discordance between the TSH value and either the clinical impression or the concentrations of ancillary tests such as free thyroxine (T₄). Although manufacturers incorporate mouse serum or IgG fractions into their assays to negate heterophile antibody interference, these measures are unlikely to be totally effective for all affected sera [17].

More difficult to detect is non-TSH-related loss of specificity from non-TSH serum components or from preanalytical factors, especially the specimen container used [34] or the length of exposure to ambient temperature [4]. This type of specificity loss is often subtle and may result in an artifactual signal elevation and inappropriately high TSH result or, in the case of hirudin contamination, an artifically low signal (unpublished observations). Either type of specificity loss can lead to inappropriate clinical action or inaction. The incidence of these types of specificity loss is unknown but the problem merits further study by manufacturers.

Clinical Utility
The enhanced sensitivity of the TSH IMA methods has dramatically expanded the clinical role of TSH measurement [35, 36]. The expansion in this assay's diagnostic potential is linked to the reliability of low-range measurement in clinical practice [35]. The most sensitive TSH assays, i.e., the third-generation assays

![Fig. 2. Erosion of TSH interassay precision profiles (10 or more analyses of human serum pools) across time. The 10-day interassay profile reflects the use of a single calibration and single reagent batch. Thereafter, there is a cumulative increase in numbers of reagent lots and instrument calibrations involved. The intraassay profile is shown for comparison.](image-url)
with functional sensitivity in the 0.01–0.02 mU/L range, have provided the following clinical insights:

1) The TSH response after stimulation with thyroliberin (TSH-releasing hormone; TRH) is predictably 8 to 10 times higher than the basal TSH concentration, and TRH testing does not add any clinical information to that provided by an accurate basal TSH measurement, provided that patients have intact hypothalamic–pituitary function [5].

2) The serum TSH/free T4 relationship is log/linear [3]. Because thyroid disease presents in differing degrees [5], a low TSH value may be the only biochemical indicator of subclinical hyperthyroidism—a common result of l-T4 overtreatment of primary hypothyroidism. Some patients benefit from the diagnosis and treatment of subclinical hypothryroidism [37, 38]; therefore, reliable low-range TSH measurement is necessary for cost-effective diagnosis and management of the full spectrum of thyroid conditions.

3) A substantial number of hospitalized patients in the acute stage of a nonthyroidal illness (NTI) will develop a transiently low TSH concentration in the 0.01–0.4 mU/L range [13, 39]. Only assays with functional sensitivity in the 0.01–0.02 mU/L range can be used to distinguish the mild degrees of TSH suppression in NTI from the profoundly low TSH concentrations (<0.02 mU/L) typical of sick thyrotoxic patients [3].

4) TSH suppression is relative, not absolute, as evident from the finding of low TSH concentrations (0.001–0.01 mU/L) in the sera of overtly thyrotoxic patients [5, 18]. Reliable low TSH measurements are essential when managing patients with differentiated thyroid cancer, who typically receive high-dose l-T4 therapy to suppress the trophic influence of TSH on thyroid tumor tissue. Typically, such patients have submaximally suppressed basal TSH and TRH-stimulated responses [5, 18]. As shown in Fig. 3a, it takes a considerable amount of time (weeks) for the basal TSH and TRH-stimulated TSH responses to become suppressed to the concentrations seen in overt Graves thyrotoxicosis. The lag in achieving maximal suppression presumably reflects the time needed to inhibit TSH biosynthesis and deplete thyrotrophic TSH content. As shown in Fig. 3b, a reciprocal response is seen during the restoration of a normal basal and TRH-stimulated TSH response after successful treatment of Graves thyrotoxicosis. This lag in normalization of TSH concentrations presumably reflects the time needed to reinstitute TSH biosynthetic processes and replete thyrotroph TSH content.

5) The TSH suppression pattern is remarkably predictable [18]. Within 48 h after the administration of a thyroid hormone load—300 μg of l-triiodothyronine (T3; Cytomel®) or 2 mg of l-T4—to an individual with normal pituitary function, serum TSH concentrations are predictably suppressed to <10% of control value [18, 40]. This response has been used as the basis for developing an outpatient T3 challenge test for diagnosing thyroid hormone resistance and TSH-producing pituitary tumors, as well as a means to obtain sera with concentrations below the lower reference limit to use for functional sensitivity studies or quality-control samples.

**Laboratory Recommendations**

Both performance (sensitivity and specificity) and practical factors (e.g., instrumentation, incubation time, cost, technical support) influence the selection of a TSH method. It is the laboratory's responsibility to the physician, and ultimately the patient, to report biochemical concentrations that generate an appropriate clinical response. This responsibility requires that laboratories minimize sources of random error, establish relevant reference intervals, and realistically identify the upper and lower limits of the measurement on the basis of experimentally determined confidence limits.

**Functional Sensitivity Assessments**

Functional sensitivity is the most important performance criterion that should influence the selection of a TSH method, given that current methods have comparable diagnostic utility for detecting above-normal serum TSH concentrations. Laboratories should independently and realistically establish the functional sensitivity limit of their method by the standard protocol described below and use the functional sensitivity as the lower reporting limit on patients' reports. A failure to do this increases the risk of reporting misleading results that will either prompt unnecessary confirmatory testing, as in the case of artifactualy abnormal concentrations, or lead to clinical inaction when inappropriately normal concentrations are reported for patients with low TSH concentrations.

A realistic assessment of functional sensitivity would dictate that laboratories assess low-end precision by a protocol based on the following constraints:

1) Use human sera in preference to modified sera or non-human-based protein matrices.

2) Select serum concentrations that cover the proposed assay range, including a value "≈.0.02 mU/L above the expected functional sensitivity limit.

3) Establish the interassay precision profile from 10 or more analyses of each sera performed in different runs.

4) Make a random, not ordered, analysis of these sera to reflect any carryover effect on low concentrations.
5) Use more than one batch of reagents, and more than one instrument calibration, when assessing interassay precision.

6) Runs used for interassay precision assessments should be made across a clinically relevant time span, i.e., for TSH in an outpatient setting, ~6 to 8 weeks.

The requirement that human sera be used for establishing functional sensitivity is a problem for many laboratories. Obtaining suitable sera requires a time commitment that may be impractical for a smaller laboratories that perform a low number of thyroid tests. Further, when the laboratory’s current TSH method is unreliable in the low range, it cannot be used to identify sera with low concentrations. Laboratories unable to generate their own low-TSH serum pools from subjects with low TSH concentrations have two options: Engage manufacturers to provide suitable sera for precision studies (manufacturers have the capability to obtain defined serum specimens from a variety of sources), or, when practical, generate a low-TSH pool from euthyroid volunteers (as described above); this pool can also be used to spot-check the status of the precision profile in the low range relative to manufacturers’ claims.

**SPECIFICITY ASSESSMENTS**

TSH-related specificity arising from monoclonal antibody specificity for different TSH isoforms does not particularly affect the clinical utility of TSH IMA measurements. In contrast, non-TSH-related loss of specificity can be serious. There are no quality assurance checks for heterophile antibody or other nonspecific interferences that may lead to the reporting of an inappropriate TSH value. Although manual or computer-based checks of concordance between estimates of TSH and free T4 concentrations may be a proactive approach, the trend towards a TSH-centered strategy [35, 36, 38] will make this less feasible. Accordingly, nonspecific interferences will usually be first detected by the physician who notices a discordance between the TSH result and the clinical status of the patient. Laboratories and physicians can use the following approaches to identify a nonspecific interference in a suspect serum:

1) Repeat the test after checking specimen identity.
2) Check for parallelism in dilutions.
3) Check concordance with the result of a different TSH method.
4) Check a new specimen from the patient.
5) Check that the serum TSH concentration measured is modulated appropriately by physiological factors such as TRH stimulation or thyroid hormone suppression.

**Recommendations for Manufacturers.**

The manufacturer’s responsibility is to realistically define the performance characteristics of the method and ensure that this performance can be reproduced by a wide range of laboratories in clinical practice. Manufacturers need to determine which factors are responsible for the suboptimal functional sensitivity currently experienced in the field and improve the ruggedness of their methods so that clinical laboratories can consistently achieve optimal functional sensitivity [4].

**Use of the term “third generation.”** The term third generation cannot be used interchangeably with descriptive terms such as ultrasensitive. Third generation has a scientific definition based on functional sensitivity: i.e., TSH in the 0.01–0.02 mU/L range can be measured with 20% interassay CV. Manufacturers and clinical laboratories should use the same protocol for establishing functional sensitivity (described above). To prevent the term third generation from becoming merely marketing jargon and retain some utility as a performance characteristic for laboratories to consider when selecting a method, all manufacturers should play by the same rules.

**Kit package inserts.** Manufacturers have the responsibility to produce kit package inserts that realistically depict the expected performance of their product in a clinical laboratory setting. Manufacturers should no longer cite analytical sensitivity alone in kit package inserts, because this can mislead laboratories to adopt an inappropriate sensitivity limit on patients’ reports. Manufacturers should place less emphasis on the assay’s “generation” and more on providing a realistic estimate of expected functional sensitivity.

Manufacturers should display interassay precision profiles obtained in a range of different clinical laboratories. They should also provide defined sera and help their laboratory users perform performance checks before, and periodically after, a method has been placed in clinical use.

Unfortunately, suboptimal sensitivity and specificity of many current TSH IMA methods decrease their diagnostic accuracy for measurement of low serum TSH concentrations. These methodological problems limit the widespread adoption of the more cost-effective TSH-centered thyroid-testing strategy currently recommended by the American Thyroid Association [35, 36, 38]. The differences in functional sensitivity between methods, and even across laboratories using the same manufacturer’s method, are disturbing [4]. Until functional sensitivity can be standardized across methods, and manufacturers can guarantee consistent performance in clinical practice, serum TSH measurement will not reach its maximum diagnostic potential.

This work was supported in part by NIH Grants DK-11727 and NCRR General Clinical Research Center Grant M01-RR-43.

**References**

17. Lauberg P. Persistent problems with the specificity of immunoassay TSH. Thyroid 1993;4:279–83.