Interlaboratory/Intermethod Differences in Functional Sensitivity of Immunometric Assays of Thyrotropin (TSH) and Impact on Reliability of Measurement of Subnormal Concentrations of TSH

Carole A. Spencer,1,4 Michael Takeuchi,1 Margarita Kazarosyan,1 Finlay MacKenzie,2 Geoffrey J. Beckett,3 and Ewan Wilkinson2

Clinically relevant interassay precision profiles for thyrotropin (thyroid-stimulating hormone; TSH) were constructed with human serum pools measured over 4–8 weeks by six immunometric assays, in at least two different reagent lots. Functional sensitivities (the concentration at which the interassay CV is ≤20%) were determined in four to eight clinical laboratories plus the respective manufacturer’s laboratory. These studies revealed that the manufacturer’s stated functional sensitivity limit is rarely duplicated in clinical practice. Loss of specificity (indicated by artificially high values) was seen with some methods when used to measure certain unrefrigerated low-TSH sera. Measurement of TSH in four human serum pools (TSH <0.05–0.25 mIU/L) by 16 different methods (each in at least eight UK or US laboratories) showed that some methods could not reliably distinguish subnormal from normal TSH values. Better pool rankings and fewer misclassifications of low-TSH sera as “normal” were seen with use of assays capable of “third-generation” functional sensitivity (0.01–0.02 mIU/L) than with assays with “second-generation” functional sensitivity (0.1–0.2 mIU/L). Because inter- and intramethod differences in functional sensitivity negatively impact the diagnostic accuracy and cost-effectiveness of a TSH-centered thyroid testing strategy, laboratories should independently establish an assay’s functional sensitivity by a clinically relevant protocol. Moreover, manufacturers should assess functional sensitivity more realistically and improve the robustness of assays to ensure that their performance potential is consistently met in clinical practice.

Indexing Terms: intermethod comparison/variation, source of/sample handling

Currently, most clinical measurements of thyrotropin (TSH) are made by immunometric assay (IMA) technology, which has a lower detection limit and faster turnaround time than the previous RIA methods (7). These TSH IMAs were called “sensitive” or “ultrasensitive” assays when they were first introduced in the mid-1980s (2). Traditionally, the “sensitivity” or limit of detection of an assay has been calculated from an intraassay measure reflecting imprecision of the zero calibrator (3).

The American Thyroid Association now considers such intraassay determinations of detection limit to be clinically irrelevant and to give an overly optimistic estimate of assay performance (4). Despite this position, most manufacturers in their package inserts still cite a lower assay limit based on intraassay measurements, and some laboratories still inappropriately use this unrealistic detection limit on their patients’ reports. Currently, the use of intraassay measures and descriptive terms such as ultrasensitive are being abandoned in favor of a new “generational” nomenclature system, based on a more clinically relevant interassay measure: functional sensitivity, i.e., the lowest TSH concentration that achieves an interassay CV of 20% (1, 5). “First-generation” TSH assays, such as the RIAs had a functional sensitivity limit between 1.0 and 2.0 mIU/L. Current TSH IMAs have second- or third-generation functional sensitivity potential (0.1–0.2 and 0.01–0.02 mIU/L, respectively).

Although the generation nomenclature system has a more scientific basis than descriptive terms (e.g., ultrasensitive), it too can be misleading. Some assays have a functional sensitivity intermediate between the above-mentioned “generational” tiers, and the values mentioned in published studies usually reflect assay performance under ideal laboratory conditions. For instance, interassay precision profiles are usually constructed over short periods, periods not typical of the clinical interval used in the setting of thyroid outpatients, and the nonhuman or modified TSH-free matrices used in the published assay evaluations may give better precision estimates than the human sera used in clinical practice (6). Additionally, because the order of the pools analyzed during precision studies is rarely

1 University of Southern California, Department of Medicine, 2025 Zonal Ave., Los Angeles, CA 90033. 2 UK NEQAS, Clinical Chemistry, Queens Elizabeth Hospital, Birmingham, UK. 3 University Department of Clinical Biochemistry, Royal Infirmary, Edinburgh EH3 9YW, Scotland, UK. 4 Author for correspondence. Fax 213-226-4969; E-mail cspencer@hsu.sc.edu. Presented in part at the 67th meeting of the American Thyroid Association, Tampa, FL, November 1993. Received September 22, 1994; accepted November 7, 1994.

Nonstandard abbreviations: TSH, thyrotropin (thyroid-stimulating hormone); IMA, immunometric assay; UK NEQAS, National External Quality Assessment Scheme; ABB, Abbott IMx; ACC, Sanofi Access; BD, Becton Dickinson Simultrac; BIO, BioRad CoTube; BM, Boehringer Mannheim TSH; COR, Corning ACS 180; DAK, Dako Novocline; DEL, Wallac Delfia; DPC, Diagnostic Products Corp. Immulite; IDS, Immunodiagnostics Systems TSH; K-AM, Kodak Amerlite; K-CT, Kodak Coated Tube; K-30, Kodak TSH-30; NET, Netris IRMA; NIC, Nichols chemiluminescence; and SER, Serono Mialclone.
randomized (another departure from routine analysis), carryover problems may not be detected. Further, it is customary to probe the reliability of TSH measurement in the subnormal range by using overt Graves disease thyrotoxic sera, which have profoundly low TSH (<0.01 mIU/L), in preference to sera with milder degrees of TSH suppression, which are more often encountered in clinical practice (7).

Reliable measurement of subnormal range TSH concentrations is essential for a diagnostically accurate and cost-effective thyroid-function testing strategy. Currently, in recognition of the sensitivity potential of current TSH IMAs, the American Thyroid Association has recommended that, in ambulatory patients, TSH measurement is the optimal first-line thyroid test and can be used in preference to a panel approach (free thyroxine + TSH) if hypothalamic–pituitary function is intact and thyroid status is stable (8, 9). Compliance with this recommendation necessitates that reliable measurement of subnormal range TSH concentrations be available in all clinical laboratories. The variability in clinical performance of current TSH IMA methods prompted the present study, which was designed to (a) construct clinically relevant precision profiles for several different TSH IMA methods; (b) assess whether the functional sensitivity achieved by the manufacturer of a method is duplicated in clinical practice; and (c) assess the reliability of subnormal range measurement by a range of TSH IMAs commonly used in the UK and the US.

Materials and Methods

Serum Pools

**Precision profile study.** Nine serum pools were made by selectively pooling patients’ sera with TSH values ranging from subnormal (<0.4 mIU/L) to normal (0.4–4.6 mIU/L) (10). Eight of the pools with TSH between 0.02 and 3.5 mIU/L were constructed solely on the basis of the TSH value (determined by Nichols Labs. (San Juan Capistrano, CA) chemiluminescence (NIC) assay) without clinical information. One pool was constructed entirely of sera from clinically and biochemically hyperthyroid patients (TSH <0.01 mIU/L) (10). All pools were then aliquoted into small volumes for frozen shipment (−10°C) to participating laboratories for analysis.

One equine pool and one porcine pool (Gibco Laboratories, Grand Island, NY) were included in the precision profile study to evaluate the zero blank (noise level) of each method.

**UK NEQAS/US TSH pool study.** Four different low-TSH (<0.05 mIU/L) human serum pools (“A” pools) were prepared with sera either from triiodothyronine-treated volunteers (50–75 μg of liothyronine daily for 6 weeks) or from patients with subnormal serum TSH concentrations (as assessed by NIC). Two A pools were prepared by the UK National External Quality Assessment Scheme (NEQAS) for distribution in the UK. Two different A pools were prepared by the University of Southern California (Los Angeles, CA) for distribution in the US. Derivative pools B, C, and D were made from each A pool by addition of TSH (2nd International Reference Preparation 80/558; National Institute for Biological Standards and Control, South Mimms, Herts, UK), 0.1, 0.15, and 0.2 mIU/L, respectively. Sodium azide was added as a preservative to the UK material but not to the US pools.

**Study Design**

**Precision profile study.** Six different TSH IMAs were each evaluated in three to seven clinical laboratories in the UK or the US as well as in the respective manufacturer’s laboratory. The clinical laboratories included hospital, reference, and doctor’s office laboratories. The methods evaluated were: Abbott IMx (ABB; Abbott Laboratories, Chicago, IL), four clinical laboratories; Bio-Rad CoTube (BIO; Bio-Rad, Hercules, CA), three; Corning ACS 180 (COR; CIBA-Corning, Medfield, MA), three; Diagnostic Products Immulite (DPC; Diagnostic Products Corp., Los Angeles, CA), seven; Kodak TSH-30 (K-30; Kodak Clinical Diagnostics, Amersham, UK), seven; and NIC, six. The nine human serum pools and the equine and porcine pools were each analyzed 10 times by each laboratory over a 4–8-week period, each run being made in a different blinded randomized order. Laboratories were instructed to (a) treat the pools as if they were routine clinical specimens, (b) report TSH values even if they were below the laboratory’s usual reporting limit, and (c) include two or more reagent lots and (or) instrument calibrations. The values reported for the nine human serum pools were used to construct interassay precision profiles for the participating laboratories utilizing each method.

**Pool study.** This study was designed to assess how well the various TSH IMA methods could distinguish subnormal from normal TSH values in clinical practice. Sixteen TSH methods were evaluated across a range of clinical laboratories in the UK and the US, each method being evaluated in at least eight laboratories; two (ABB and COR) were tested by laboratories in both countries. The methods evaluated were: Abbott IMX (ABB), 12 UK laboratories and 24 US laboratories; Sanofi Access (ACC; Sanofi, Chaska, MN), 17 US; Becton Dickinson Simultrac (BD; Becton Dickinson, Oxford, UK) 13 UK; BM TSH (BM; Boehringer Mannheim, Lewes, UK) 15 UK; COR, 5 UK and 28 US; Dako Novoclone (DAK; Dako, Cambridge, UK) 8 UK; DPC, 14 US; Wallac Delfia (DEL; Wallac, Milton Keynes, UK), 19 UK; Immunodiagnostics Systems (IDS; Immunodiagnostic Systems, Washington, UK), 14 UK; Kodak Amerlite (K-AM), 13 UK; Kodak Coated Tube (K-CT), 11 UK; K-30, 13 UK; Nertia IRMA (NET; Nertia, London, UK), 10 UK; NIC, 26 US; and Serono Maiacolone (SER; Serono, Woking, UK), 12 UK. The A pools and their derivative B–D pools were distributed to the participating laboratories with instructions that they be treated as routine clinical specimens. Most laboratories received two or more A pools and one or more of the derivative pools on two separate occasions 2 to 6.
months apart. The laboratories in the UK were all participants of UK NEQAS. Laboratories in the US were randomly selected from lists provided by each manufacturer. Specimens were frozen and then sent to the participating laboratories by overnight mail.

**Stability study.** This study was designed to assess the effects of transporting or storing a specimen at suboptimal temperature (>8°C). Eight aliquots of the two US A pools and their derivative B-D pools were exposed to ambient temperature (22 ± 2°C) for 0–7 days before being frozen and shipped on solid CO2 to the manufacturer's or an author's (C.A.S.) laboratory for batch-wise TSH measurement by one of six methods (ABB, ACC, BIO, COR, DPC, or NIC). Three methods (ACC, COR, and DPC) were evaluated with two sets of pools to test whether the temperature effects, if any, were serum-dependent. The ABB and BIO methods were tested only with the C and D pools from serum 1 respectively, these pools' values being in the working ranges of the two methods.

**Results**

**Precision Profile Study**

Figure 1 displays the interassay precision profiles (CV vs mean TSH) reported by the manufacturer and the participating clinical laboratories for each method. Each manufacturer's functional sensitivity point (CV = 20%) was below that of most of the clinical laboratories.

The spread of functional sensitivity values between laboratories that used automated methods (ABB, COR, DPC, and K-30) appeared to be similar to that of the manual methods (BIO and NIC). Poor precision resulting from instrument-related technical problems were detected retrospectively in two clinical laboratories that used automated methods (one DPC and one K-30).

Although the three methods for which "third-generation" functional sensitivity (0.01–0.02 mIU/L) was claimed (DPC, K-30, and NIC) met this claim, the manufacturer's laboratory, only five of seven clinical laboratories using DPC, none of seven using K-30, and one of six using NIC achieved this functional sensitivity.

The mean (± SE) TSH values for the equine, porcine, and human thyrotropic serum pools reported by each method are shown in Fig. 2. As expected, the three least-sensitive methods tested (ABB, BIO, and COR) reported a higher noise level and more variability with most matrices than the other assays. Surprisingly, the lowest blank value was not always obtained with the matrix used by the method's manufacturer.

**Pool Study**

Figure 3 shows the TSH values reported for pools A–D, relative to the manufacturers' normal reference and thyrotropic (<0.01 mIU/L) reference ranges. Small but significant differences were detected.
(by K-30 in UK and DPC in US) between the two different A pools prepared in the UK—0.02 ± 0.02 (SD)
vs 0.04 ± 0.04 mIU/L (P < 0.05)—for UK and the two US-prepared A pools—0.01 ± 0.01 vs 0.04 ± 0.01 mIU/L (P < 0.05).

Pool data compared with functional sensitivity results for methods ABB, BIO, COR, DPC, K-30, and NIC. Although the pools were shipped to the laboratories on the same day, the data reported were obtained
with a range of different reagent lots: ABB, 12 different lot numbers; ACC, 3 lots; BIO, 4 lots; COR, 13 lots;
DPC, 3 lots; and NIC, 13 lots. A relationship between functional sensitivity and the range and ranking of pool values was seen for most methods (see Figs. 1 and 3):

1) ABB: This method did not consistently meet second-generation functional sensitivity (Fig. 1). A wide
spectrum of TSH values, ranging from thyrotoxic (<0.01 mIU/L) to “normal,” was reported for all four
pools.

2) BIO: This method had similar precision profiles to ABB but appeared to give fewer “normal” misclassifications and a greater number of appropriate pool rankings.

3) COR: This method, tested in both countries, gave good pool rankings and no misclassification of pool values as normal in the six UK laboratories; in contrast, a wide spread of values and poor ranking was reported for the US pools.

4) DPC: This method most consistently achieved third-generation functional sensitivity (0.01–0.02 mIU/L) across clinical laboratories (Fig. 1). One laboratory, which exhibited a poor functional sensitivity (>0.2 mIU/L), reported technical problems retrospectively. The pool values and rankings shown in Fig. 3 were in general accordance with a good functional sensitivity, although two “normal” misclassifications were reported for the first US pool A distributed. One laboratory confirmed one of the A pool outlier values (0.09 mIU/L) in a subsequent run. It was not possible to determine whether temperature-dependent sensitivity

(see Fig. 4) or some other problem was responsible for these outliers.

5) K-30: Good reproducibility and pool ranking were seen across the laboratories that used this method, in accordance with the functional sensitivity assessments shown in Fig. 1.

6) NIC: No misclassifications and good ranking between pools were seen with this method, in accordance with the functional sensitivity assessments shown in Fig. 1.

Pool data unaccompanied by precision profiles. No precision profiles were obtained for methods ACC, BD, BM, DAK, DEL, IDS, K-AM, K-CT, NET, and SER, which were evaluated by pool study alone. The A pools were appropriately classified as having TSH <0.1 mIU/L and the derivative B-D pools were appropriately ranked by most of the laboratories that used the ACC, DAK, DEL, and K-AM methods. In contrast, poor pool ranking and more A pool values >0.1 mIU/L were reported for the BD, BM, IDS, K-CT, NET, and SER methods.

Stability Study

Figure 4 summarizes the effects of preanalytical exposure to ambient temperature (22 ± 2°C) on TSH measurement in the two US A pools and their derivative B-D pools by the ABB, ACC, BIO, COR, DPC, and NIC methods. Most methods reported appropriate pool values and rankings when sera were exposed for <48 h to ambient temperature; thereafter, specificity was lost and artefactually high TSH values were observed with some methods. The specificity loss was most pronounced at very low TSH values (A pools).

Three methods (ACC, COR, and DPC) were evaluated with two sets of pools (sera 1 and 2 in Figure 4). The ACC method displayed marked temperature sensitivity to serum 1 but not serum 2, whereas the COR method showed temperature sensitivity to serum 2 but not serum 1. DPC measurements of TSH in both sera appeared stable for 5 days but rose in one serum (serum 1) when exposed for 6 or 7 days.

Discussion

Constructing TSH IMA precision profiles over a clinically relevant timespan (4–8 weeks) based on human sera analyzed randomly with more than one lot of reagents resulted in a wide range of functional sensitivity values reported by different clinical laboratories using the same method. Typically, the best (lowest) functional sensitivity is achieved by the manufacturer; the same assay can give a 10-fold higher functional sensitivity limit (one “generation” poorer) in some clinical laboratories. No striking difference in functional sensitivity was apparent across laboratories using the automated methods (ABB, COR, DPC, and K-30) as compared with those using the manual methods (BIO and NIC). Fluctuation in a method’s functional sensitivity across laboratories was thought to reflect inherent instrument–instrument differences, different calibration practices, and fluctuations in such
local variables as temperature and voltage. Uncharacteristically poor precision was considered an indicator of instrument problems, suggesting that another benefit of ongoing assessments of precision may be the detection of problems in automated "black box" methods. Because each method was assessed in only a few laboratories, it was not possible to definitively determine the effect, if any, on the functional sensitivity of an assay by type of clinical laboratory (hospital, reference, or doctor's office); however, the poorest precision was associated with a doctor's office laboratory in two methods (ABB and NIC).

The precision profile data (Fig. 1) and assay noise (Fig. 2) were related to the accuracy, reproducibility, and ranking of low-TSH values across clinical laboratories (Fig. 3). The two methods that achieved only second-generation functional sensitivity (0.1–0.2 mIU/L) in the manufacturer's laboratory (ABB, BIO) reported more subnormal-TSH pools misclassified as "normal" and gave poorer rankings between the pools than did the three methods (DPC, K-30, and NIC) that met third-generation functional sensitivity requirements (0.01–0.02 mIU/L) in the manufacturer's laboratory. Of the 10 methods that were evaluated by the pool study alone, four (ACC, DAK, DEL, and K-AM) had good pool rankings and reported few A pool values >0.1 mIU/L, suggesting good functional sensitivity potential. The other six methods (BD, BM, IDS, K-CT, NET, and SER) reported poor rankings and a number of values >0.1 mIU/L in the A pools. Because most

Fig. 4. Temperature sensitivity of different TSH IMAs: apparent changes in serum TSH concentration after exposure of the two US A pools (C) and their corresponding derivative pools, B (+), C (C), and D (Δ), to ambient temperature (22 ± 2°C) for 0 to 7 days. Serum 1 was the first US set of pools distributed (C, Fig. 3); serum 2, the second set of US pools distributed (+, Fig. 3).
laboratories would interpret a TSH value >0.1 mIU/L as excluding hyperthyroidism, methods that report a significant number of values above >0.1 mIU/L in the A pools (e.g., ABB, BD, BIO, BM, IDS, K-CT, NET, and SER) would be considered unsuitable for detecting subnormal concentrations of TSH, inaccurate measurement of which might lead to an inappropriate lack of clinical action in a TSH-centered diagnostic strategy. Although A pool values >0.1 mIU/L were also seen with the ACC, COR, and DPC methods when measuring one of the two A sera, we concluded that these aberrant values more probably reflected preanalytical temperature effects or specimen misidentification and not inherent method imprecision (see Figs. 1 and 4).

In response to the observation of outliers in the A pool reported by some methods with generally excellent functional sensitivity, we evaluated the effect of preanalytical factors such as temperature exposure on specificity loss for assaying sera with subnormal concentrations of TSH. Although manufacturers recommend that serum specimens be refrigerated at 4°C before analysis and stored at –20°C if analysis is delayed, sera are usually exposed to ambient temperature at some period before and during analysis. Typically, specimens are brought to ambient temperature before assay and incubated at ambient or warmer temperatures as part of the assay procedure. The discrepancy in the UK vs US data for the COR method (pool set 1) did not appear to reflect preanalytical temperature effects (Fig. 4, serum 1) but rather might be related to the addition of azide to UK pools but not US pools. However, marked temperature sensitivity was seen with the COR method and the second set of US pools (Fig. 4, serum 2), as well as with the ACC method and the first set of US pools (Fig. 4, serum 1).

We were not able to determine the component(s) of the sera responsible for the temperature-dependent specificity loss. Given that the second set of US pools was derived entirely from sera from a single triiodothyronine-suppressed healthy subject, it was unlikely that drug contamination of the pool was responsible for the specificity loss. Indeed, because the temperature effects with ACC and COR were serum-dependent, different methods may be sensitive to different serum components. It is unlikely that the artifactual TSH increase generated with temperature exposure reflects changes in the immunoactivity of the TSH molecule itself, and we conclude that preanalytical temperature effects may be able to influence the specificity of measurements of a wide spectrum of other non-TSH analytes measured by IMA technology. The rapidity of onset of the temperature effect seen in serum 2 measured by COR has significant implications for the need to standardize specimen-transport and laboratory handling procedures.

TSH IMA sensitivity is determined by the specific assay components selected, including the type of signal, the capture and signal antibodies, and the TSH-free matrix used for diluting standards. All components affect the ultimate signal-to-noise ratio, which determines the functional sensitivity potential of an assay. Although this study is concordant with previous reports (13) in showing that nonisotopic, especially chemiluminescent, signals have the greatest sensitivity potential (DPC, K-30, and NIC in Fig. 1), not all assays based on chemiluminescence are capable of third-generation functional sensitivity (see COR in Figs. 1 and 3), and assays based on other nonisotopic signals such as enzymes and time-resolved fluorescence clearly offer good functional sensitivity (DAK and DEL in Fig. 3). The three methods that achieved third-generation functional sensitivity in the manufacturer’s laboratory also displayed the lowest the assay noise with the animal and human serum matrices (Fig. 2). We point out, however, that the porcine-based (ABB) and equine-based (BIO and NIC) assays did not always exhibit the lowest noise values for their respective zero matrices. This reflects differences not only between the sera from different animal species, but also between batches of sera from the same species (C.A.S., unpublished observations). The method with the lowest blank value (DPC) appeared to have engineered a zero point, which negated detection of the acute increase in serum TSH (from 0.005 to 0.02 mIU/L, by NIC) after initiation of anti-thyroid drug treatment for hyperthyroidism (C.A.S., unpublished observation).

When laboratories execute with good technique, most current TSH IMAs have the potential for second-generation functional sensitivity (interassay CV = 20%) at 0.1–0.2 mIU/L (1), the minimum sensitivity currently found acceptable by the American Thyroid Association (4). This should allow diagnosis of overt Graves thyrotoxicosis, characterized by TSH <0.01 mIU/L (1, 10, 12). Our current study shows that TSH IMA functional sensitivity in clinical practice is usually suboptimal in comparison with that reported by the manufacturer. This loss of functional sensitivity, together with loss of specificity because of the effects of preanalytical factors such as temperature, can lead to the misclassification of hyperthyroid patients as euthyroid. These sensitivity and specificity problems directly affect the diagnostic accuracy and cost-effectiveness of the TSH-centered single-test strategy currently being promoted by the American Thyroid Association (8, 9). Compliance with new standards of care guidelines (14) necessitates that all clinical TSH IMAs provide reliable measurement of TSH in the subnormal range. This is especially important when diagnosing mild degrees of hyperthyroidism, which are often iatrogenic (7) and have cardiac sequelae (15). Accurate and precise subnormal range measurement is also important when optimizing suppressive therapy with L-thyroxine for differentiated thyroid cancer (16) or when diagnosing hyperthyroidism in hospitalized patients with nonthyroidal illness (10).

In conclusion, clinical laboratories must independently establish their own functional sensitivity for an assay by a clinically relevant protocol, and must use
this functional sensitivity to define their lower reporting limit, irrespective of a manufacturer's "generation" claims. Indeed, the wide variability in functional sensitivity seen for the same method across different clinical laboratories suggests that claims of third-generation are just as misleading as descriptive terms such as ultrasensitive and should be abandoned in favor a clinically relevant assessment of functional sensitivity. Manufacturers need to address the issues of sensitivity and specificity raised in this study, to assess functional sensitivity realistically, and to improve the sensitivity, specificity, and robustness of current assays so that performance potential is consistently met in clinical practice.

We thank Richard O'Kell for valuable discussions concerning this study design and data interpretation. Although this study was supported entirely from research funding (NIH Grants DK-11727 and NCCR General Clinical Research Center Grant MO1-RR-43), we thank all the manufacturers of the TSH methods evaluated for their cooperation and interest in this study, particularly Abbott Laboratories, Bio-Rad Laboratories, Ciba Corning Diagnostics Corp., Diagnostic Products Corp., Kodak Clinical Diagnostics, the Nichols Institute, and Sanofi Diagnostics Pasteur. Some manufacturers provided participating laboratories credit for kits used to generate precision profiles.

References