References
1. Pinell AE, Northam BE. New automated dye-binding method for serum albumin determination with bromcresol purple. Clin Chem 1978;24:80–6.
2. Webster D, Bignell AHC, Atwood EL. An assessment of the suitability of bromcresol green for the determination of serum albumin. Clin Chim Acta 1974;53:101–8.
3. Stern J, Lewis WHP. The colorimetric estimation of calcium in serum with *o*-cresolphthalein complexone. Clin Chim Acta 1957; 2:576–80.
4. Schmitz J, Klein L, Jain C, Shu F. Bichromatic determination of total calcium in serum with Arsenazo III [Abstract]. Clin Chem 1986;32:1198.
5. Cali JP, Bowers GN Jr, Young DS. A Reference Method for the determination of total calcium in serum. Clin Chem 1973;19:1208–12.
6. European Committee for Clinical Laboratory Standards. Guidelines for the evaluation of analysers in clinical chemistry, Vol. 3, No. 2. Berlin: Beuth Verlag GmbH.; 1986:15.
7. Maguire GA, Price CP. Bromocresol purple method for serum albumin gives falsely low values in patients with renal insufficiency. Clin Chim Acta 1986;155:83–8.
8. Bush V, Reed RG. Bromcresol purple dye-binding methods underestimate albumin that is carrying covalently bound bilirubin. Clin Chem 1987;33:821–3.
9. Gustafsson JEC. Improved specificity of serum albumin determination and estimation of "acute phase reactants" by use of the bromcresol green reaction. Clin Chem 1976;22:616–22.
10. Di Polo R, Requena R, Brinley FJ, Mullins LJ, Scarpa A, Tiffert T. Ionized calcium concentration in squid axons. J Gen Physiol 1976;67:433–67.
11. Freaney R, Egan T, McKenna MJ, Doolin MC, Muldowney FP. Determination of ionised calcium by ion specific electrode is not independent of albumin concentration. Clin Chim Acta 1986;158: 129–38.
12. Payne RB, Little AJ, Evans RT. Albumin-adjusted concentration in serum increases during normal pregnancy. Clin Chem 1990;36:142–4.

CLIN. CHEM. 38/7, 1345–1349 (1992)

Interassay Variability of Immunometric Methods for Thyrotropin in an External Quality Assessment Survey: Evidence That Functional Sensitivity Is Not Always Adequate for Clinical Decisions

A. Pilo,[1] G. C. Zucchelli,[1] R. Malvano,[2] A. Clerico,[1] G. Iervasi,[1] and C. Signorini[2]

We investigated the ability of current immunometric methods for thyrotropin (TSH; thyroid-stimulating hormone) to distinguish between low-normal and subnormal hormone concentrations by using the data from an external quality assessment (EQA) survey in 1990. We computed the interassay (between-run) precision profiles from results from 101 laboratories, which used the five most popular kits in the survey; during the control period (one year) each laboratory assayed 4 EQA pools distributed (as hidden replicates) in five occasions. The interassay CV was relatively low (9–13%) for three pools in the normal TSH range (>0.8 milli-int. unit/L) but markedly higher (30–40%, except for one more precise kit) in the subnormal range (0.2 milli-int. unit/L). We calculated the effect of the between-run variability on the diagnostic accuracy (discrimination between normal and subnormal values) for three representative TSH concentrations: 0.2, 0.4, and 0.5 milli-int. unit/L (0.3 milli-int. unit/L was considered the lower normal limit). The three concentrations were reasonably discriminated ($P \leq 5\%$), and only one kit showed a between-run CV <18% at 0.2 milli-int. unit/L. For the other four less-precise kits, only the higher TSH value (0.5 milli-int. unit/L) could be classified with an acceptable diagnostic reliability. With the most precise kit, one can distinguish two TSH concentrations in the 0.3–0.5 milli-int. unit/L range that differ by at least 30%; with the other kits, differences greater than 50–60% are needed for reliable discrimination. Thus many laboratories fail to achieve the functional sensitivity of a second-generation assay, even if they use immunometric methods. TSH assays with a better interassay precision in the low concentration range are needed.

Additional Keyphrases: *distinguishing low-normal from below-normal thyrotropin concentrations · thyroid status*

During the past five years, the traditional competitive RIA methods for thyrotropin (thyroid-stimulating hormone, TSH) determination have been increasingly replaced by the so-called ultrasensitive or supersensitive immunometric assays (IMAs) based on noncompetitive "two-site" technology.[3] Almost all of these methods use monoclonal antibodies to extract TSH from the serum samples (solid-phase "capture" antibody) and to quantify the bound analyte ("tracer" antibody). The label for the tracer antibody may be either ^{125}I or a nonradioactive molecule, e.g., enzymes, fluorescent dyes, and chemiluminescent compounds.

Commercially available kits based on IMA technology are generally presented as being ~10-fold more sensitive than the traditional competitive RIAs (detection limits of ≤0.1 milli-int. unit/L versus 0.6–1.0 milli-int. unit/L, calculated either from within-run replicate

[1] CNR Institute of Clinical Physiology, via Savi 8, 56100 Pisa, Italy.
[2] Institute of Chemistry, School of Medicine, University of Brescia, Italy.
[3] Nonstandard abbreviations: TSH, thyrotropin (thyroid-stimulating hormone); IMA, immunometric assay; EQA, external quality assessment; and ANOVA, analysis of variance.
Received October 15, 1991; February 13, 1992.

measurements of the zero standard or from within-run precision profiles). Because of this markedly increased analytical sensitivity, these methods are claimed as effective in discriminating between normal and subnormal TSH concentrations. However, analytical sensitivity per se, although necessary, is not sufficient to assure a reliable classification of the low TSH values. In fact, clinical practice relies on determinations obtained in different runs, and the continuity of clinical classification is therefore entrusted primarily to the stability of the assay over time (i.e., with respect to changes in reagent characteristics or operators' performance or both). Thus, the functional performance of an assay is more adequately assessed from the between-run precision in the range of clinical concern. In particular, it has been proposed (*1–3*) that functional sensitivity, rather than analytical sensitivity, be used to assess the ability of a TSH assay in discriminating low hormone concentrations. The functional sensitivity is defined as the TSH value at which interassay precision (CV) is 20%, and only the methods that exhibit a functional sensitivity <0.2 milli-int. unit/L have been defined as second-generation assays (*3*).

Following this line of reasoning, we used the data collected during an external quality assessment (EQA) for TSH assays (*4, 5*) to calculate the between-run precision profiles of the participants that used the five most popular kits in the survey. We evaluated functional sensitivities and the ability to distinguish suppressed TSH concentrations from TSH concentrations in euthyroidism.

Materials and Methods

EQA scheme. The EQA scheme for TSH assay of the CNR Institute for Clinical Physiology does not substantially differ from similar programs (*6, 7*); participating laboratories (~170), supplied monthly with three unknown samples, are asked to perform the assay routinely and to return results indicating the method/kit used. The data collected are computer-processed, and periodic and end-of-period reports containing statistical evaluations of results and estimates of the performance of laboratories and kits are prepared and distributed to participants. The approaches followed for the analysis of EQA results and the reports for participants are described in detail elsewhere (*8, 9*).

Control material. Pools P042, P049, and P050, distributed in the EQA survey considered in this study, were prepared by pooling normal sera. A low-concentration pool (P048) was obtained by removing some of the TSH content of normal sera by affinity chromatography with CNBr-activated Sepharose 4B, previously conjugated with a monoclonal anti-TSH antibody. Sodium thimerosal and neomycin sulfate were added as preservatives and the samples were lyophilized.

Data analysis. We sent each of the four pools (P048, P042, P049, P050) to participants in five different monthly dispatches as unidentified replicate samples during the 1990 EQA cycle. Thus, five determinations carried out on the same control material were collected over a period of at least seven months; consequently, the interassay precision of each laboratory computed from these results takes into account both the variations of experimental conditions of the assays and the variations in the batches of kit reagents. The interassay precision of the kit was estimated as the median of the precisions of all laboratories using that kit.

The total variability (variability of all results produced by the laboratories assaying the same pool on five occasions) was broken down into the within-kit and between-kit components by an analysis of variance (ANOVA) technique previously described (*5, 10*).

The effect of the between-run variability on the diagnostic accuracy of the five kits considered was evaluated as reported by Bayer (*11*). In detail, the rate of false negatives was estimated from the tail of the distribution of the determinations obtained for an assigned subnormal TSH value [gaussianly distributed with the SD computed from linear interpolation of the four experimental points of the between-run precision profile (see Figure 2)] that lies beyond the lower limit of the normal range (assumed as 0.3 milli-int. unit/L). Conversely, we obtained the rate of false positives from the tail of the distribution for a normal value that falls below the lower limit of the normal range. The percentage of false results (either false negatives or false positives) was read on the table of the standard normal distribution (z score); the z-score value was computed as $z = (\text{lower limit} - x)/SD_x$, where x is the chosen TSH concentration (0.2 milli-int. unit/L being subnormal and 0.4 and 0.5 milli-int. unit/L being low-normal values), and SD_x is the between-run SD computed from the precision profile corresponding to the concentration x chosen (see Figure 3). The diagnostic accuracy is considered acceptable when the number of false results is <5%.

Following the same statistical approach, we used the between-run precision profile to estimate the ability to discriminate between two TSH values in the low concentration range. After a value x was assigned, the nearest y value ($y < x$) distinguishable from x was determined by setting the z score, $z = (x - y)/(SD_x^2 + SD_y^2)^{0.5}$, equal to 1.65. In other words, we ensured that the difference $(x - y)$ significantly differed from 0 at $P < 0.05$. These computations were carried out for three arbitrarily assigned TSH values (0.3, 0.4, and 0.5 milli-int. unit/L) and for the five kits considered.

Results

The overall means of the determinations (~700 data points) collected for each of the four pools analyzed were 0.77, 1.08, and 1.39 milli-int. unit/L, respectively, for the normal pools P042, P049, and P050, and 0.22 milli-int. unit/L for the subnormal pool P048. The consensus means of the EQA samples derived from each pool (assayed as five hidden replicates) showed a variability (CV) of 4.2%, 3.1%, 1.9%, and 2.2% for pools P048, P042, P049, and P050, respectively, demonstrating the stability of the control materials during the EQA period.

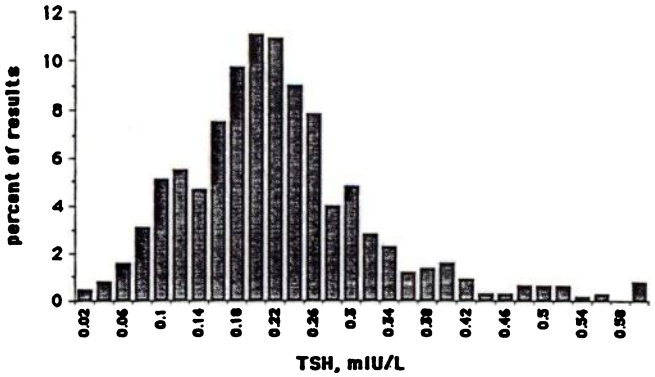

Fig. 1. Frequency distribution of the EQA results obtained for pool P048 (mean 0.22 milli-int. unit/L, n = 642)

The overall agreement (or total CV) of the results in the normal TSH range was relatively good in all cases (28.3% for P042, 27.6% for P049, 18.6% for P050), but was markedly worse for the low-concentration pool P048 (45.6%).

Figure 1 shows the histogram of all data collected for the pool P048. Despite some right-sided skewness, the distribution could be reasonably regarded as symmetrical around the mean (0.22 milli-int. unit/L). However, the spread of the data is so large that a consistent fraction of both clearly subnormal and normal values have been reported by participants for the same control material.

This large spread cannot be attributed to systematic differences between the means of different kits, as demonstrated by the relevant data shown in Table 1 for the five kits most frequently used in EQA (about 70% of results). Consequently, the major cause of variability should be the poor within-kit reproducibility. To quantify the relative percentages that systematic between-kit differences and the within-kit variability contribute to the total variability, we performed an ANOVA of the results from pools P048 and P050. The results (Figure 2) indicate that, for the pool at higher TSH concentration (P050), the total variability (CV 18.6%) is ~70% attributable to the within-kit component, with the remaining 30% being due to systematic between-kit differences; on the other hand, for the low-concentration pool P048, the within-kit component accounts for as much as 95% of the total variability of the data.

Clearly, total variability is enlarged to some extent by the fact that results are produced in different laboratories that use different kits. For this reason, we evaluated the variability of single laboratories to estimate whether their within-laboratory interassay precision was adequate to allow the discrimination of TSH subnormal values from euthyroidism. Table 1 shows the median and range of the interassay precision obtained for the four EQA pools by the laboratories that used the kit considered. These data are also presented in the interassay precision profiles depicted in Figure 3, profiles that can be considered as representative of the analytical performance of the different kits in the hands of the average-quality laboratory.

For the kits from Serono Maiaclone (ARS), Boehringer Enzymun (BHN), Abbott (IMX), and Byk-Sangtec (MAT), the interassay variability of the low-concentration pool (0.22 milli-int. unit/L) was very high (CV = 30–40%), implying that many users of these kits fail to achieve a "second-generation" functional performance; in particular, only 20–25% of the laboratories that used ARS, IMX, and MAT kits (but none of the BHN users) exhibited a second-generation performance. On the other hand, the Behring immunoradiometric assay (BEH) showed a CV of 17.6% for the low pool, and 75% of the users of this kit achieved a functional sensitivity better than 0.2 milli-int. unit/L. The precision at higher concentrations (>0.8 milli-int. unit/L) was relatively good (9–13%, on average) and adequate for all five kits.

The effect of the interassay precision on the diagnostic accuracy (discrimination between normal and subnormal values) was calculated by assuming 0.3 milli-int. unit/L as the lower normal limit. The diagnostic accuracy of the average-quality laboratory is reported in Table 2 for three representative TSH concentrations, one below the cut-off value and two in the lower part of the normal range. The data indicate that the three concentrations are reasonably discriminated ($P \leq 5\%$), with only one kit (BEH) showing a between-run precision <18% at 0.2 milli-int. unit/L. For the other kits, only the higher TSH concentration (0.5 milli-int. unit/L) could be classified with an acceptable diagnostic reliability.

Following the same statistical approach, for each kit we computed the TSH concentration interval (around the assumed cutoff of 0.3 milli-int. unit/L) in which the diagnostic accuracy was worse than 95%. Because of the poor between-run precision, the TSH values included within this indeterminate interval cannot be classified with a reasonable diagnostic accuracy (percent of false results ≥5%). The concentrations in the interval from ~0.2–0.5 milli-int. unit/L could not be reliably classified as normal or subnormal for four of five kits. In detail, the indeterminate intervals were: 0.19–0.5 milli-int. unit/L for ARS, 0.18–0.57 for BHN, 0.19–0.46 for IMX and 0.19–0.51 for MAT; for the BEH kit, which shows the best functional sensitivity, the indeterminate zone was only 0.23–0.39 milli-int. unit/L.

The effect of the interassay variability on the minimum TSH decrease that can be significantly detected was evaluated at three assigned concentrations (0.3, 0.4, and 0.5 milli-int. unit/L) for each of the five kits. This discrimination is typically requested in monitoring patients receiving suppressive therapy with thyroxin. With the kit that had the best between-run precision (BEH), we could distinguish two TSH concentrations in the range 0.3–0.5 milli-int. unit/L, if they differed by at least 30%; with the other, less precise, kits, differences of 50–60% or more were needed.

Discussion

The use of the within-run precision profile and of the detection limit (both estimated from within-run repli-

Table 1. Means and Interassay Variability[a] Observed for EQA Pools Assayed with Five Kits

		P048		P042		P049		P050	
Kit[b]	No. of labs	TSH, milli-int. unit/L	CV (range), %	TSH, milli-int. unit/L	CV (range), %	TSH, milli-int. unit/L	CV (range), %	TSH, milli-int. unit/L	CV (range), %
ARS	22	0.22	35.4 (7–66)	0.81	12.2 (5–50)	1.24	10.2 (7–23)	1.53	6.7 (3–19)
BEH	16	0.22	17.6 (4–41)	0.85	7.3 (3–19)	1.25	5.0 (3–16)	1.57	5.4 (3–14)
BHN	28	0.20	42.5 (21–64)	0.85	18.6 (11–36)	1.11	9.8 (6–37)	1.40	9.2 (5–21)
IMX[c]	16	0.21	32.8 (7–64)	0.68	12.0 (5–31)	0.93	7.3 (3–16)	1.19	7.4 (3–22)
MAT	19	0.24	33.4 (17–58)	0.80	15.9 (7–28)	1.11	12.5 (6–37)	1.39	14.6 (4–28)

[a] The interassay variability is the median of the interassay CVs obtained by the laboratories; the range of CVs is also reported.

[b] Kit abbreviations: ARS, Serono Maiaclone immunoradiometric assay (Serono Labs., Milan, Italy); BEH, Behring immunoradiometric assay (Behringwerke, Marburg, FRG); BHN, Boehringer Enzymum test (Boehringer-Mannheim, Mannheim, FRG); IMX, Abbott IMx system (Abbott Labs., Abbott Park, IL); MAT, Byk-Sangtec immunoradiometric assay (Byk-Sangtec Diagnostica, Dietzenbach, FRG).

[c] These data were produced with a version of the IMX kit that was available in Italy during 1990. In January 1991 this kit was replaced with a new version, "IMX hTSH ultrasensitive," for which the manufacturer reports much better precision in the low-concentration range.

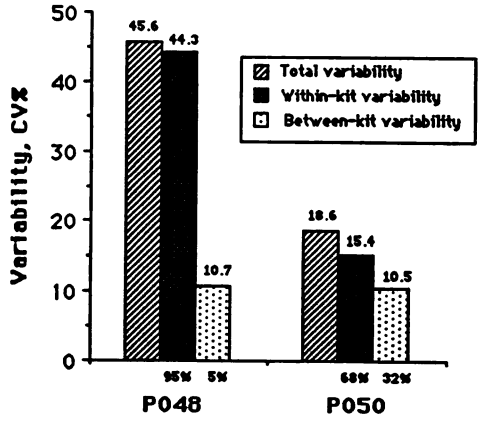

Fig. 2. Breakdown of the total variability observed for EQA pools P048 and P050 showing the between- and within-kit components

CV% is indicated *over the bars;* the percent contribution of the two components to the total variability is reported *under the bars*

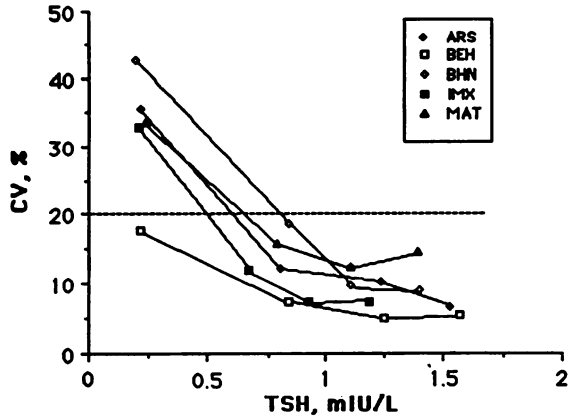

Fig. 3. Between-run precision profiles of the five kits considered in this study

Ordinate: median of CV achieved by the laboratories; *abscissa:* consensus means of the four pools; the profiles are calculated by linear interpolation. Functional sensitivity can be read as the TSH concentration corresponding to the intersection of the precision profile with the *broken line* indicating 20% between-run CV. For kit codes, see footnote to Table 1

cates) to evaluate a kit's analytical performance involves some risk of extrapolating to overall routine operation a judgment derived from a merely episodic observation. To prevent this risk, several investigators have proposed and adopted the use of between-run assessment (*1–3, 12*). This approach to the analysis of

Table 2. Effect of Interassay Variability on Diagnostic Accuracy

TSH, *x* milli-int. unit/L	CV, %	TSH SD, milli-int. unit/L	*z* = (0.3 − *x*)/SD[a]	% False results[b]
ARS				
0.2	36.2	0.072	1.38	8.4 (FN)
0.4	28.3	0.11	−0.88	18.9 (FP)
0.5	24.4	0.12	−1.64	5.0 (FP)[c]
BEH				
0.2	17.9	0.036	2.79	0.3 (FN)[c]
0.4	14.7	0.059	−1.71	4.5 (FP)[c]
0.5	13.0	0.065	−3.07	<0.1 (FP)[c]
BHN				
0.2	42.5	0.085	1.18	11.9 (FN)
0.4	35.2	0.14	−0.71	24.2 (FP)
0.5	31.5	0.16	−1.27	10.2 (FP)
IMX				
0.2	33.2	0.066	1.50	6.7 (FN)
0.4	24.4	0.10	−1.02	15.4 (FP)
0.5	20.0	0.10	−2.00	2.3 (FP)[c]
MAT				
0.2	34.7	0.069	1.44	7.5 (FN)
0.4	28.4	0.11	−0.88	18.9 (FP)
0.5	25.3	0.13	−1.58	5.7 (FP)

[a] For computing the *z* score, 0.3 milli-int. unit/L was assumed as the lower limit of normal range.

[b] FN, false negative; FP, false positive.

[c] Diagnostic accuracy better than 95% (false results <5%) is indicated.

For kit codes, see footnote to Table 1.

data collected from EQA programs looks most effective: the between-run precision relative to the same EQA samples assayed in successive occasions allows, for each individual method, consistent information to be obtained not only based on a long-term evaluation but also involving a large number of laboratories. The state-of-the-art performance, as actually emerging from routine conditions, is thus evidenced.

We tried to evaluate the implications of the observed between-run variability for clinical practice. In fact, clinicians make their decisions on the basis of measurements that are produced by the same laboratory on different occasions and are therefore affected by the same variability as observed in the multicenter trials. Two typical situations were considered: classification of

a patient with respect to the lower limit of the normal range, and assessment of variations of TSH concentration in patients taking thyroid-hormone-suppressive therapy.

Analysis of the EQA results demonstrates that the between-run precision in the low concentration range is not generally adequate to ensure a discrimination between subnormal values and euthyroid values, even when immunometric kits are adopted. In fact, with only one of the five kits considered (i.e., BEH) did the majority of the users (75%) attain a functional sensitivity (<0.2 milli-int. unit/L) sufficient to meet this clinical purpose; only 25% or fewer of the participants achieved second-generation performance with the other kits.

Note that we used a single cutoff value (0.3 milli-int. unit/L) to estimate the diagnostic accuracy (Table 2), despite the large differences indicated by kit manufacturers for the lower normal limit (0.10–0.43 milli-int. unit/L). Changing the cutoff value, however, although modifying the rates of false positive and false negative results, affects only minimally the overall width of the indeterminate zone.

The lack of discrimination observed for these kits does not affect the classification of overt hyperthyroidism, for which the immunometric assays were developed. Instead, their still insufficient between-run precision makes it difficult to distinguish among two TSH measurements in the lower range and hence to evaluate over time the effects of thyroid-hormone therapy (*13*).

These conclusions obviously refer to the laboratory of average quality, with the understanding that better-performing laboratories can fully discriminate, even when using less-precise kits, and that some users of the more precise kits cannot achieve these diagnostic goals. In any case, our findings depict a much more pessimistic situation than that emerging from the sensitivity and precision data provided with kit instructions and based on the most favorable experimental conditions. Moreover, our conclusions on the clinical performances of the kits take into account only the between-run variability estimated from replicates of control materials; other factors, e.g., the presence of interfering substances in patients' samples, can further deteriorate diagnostic accuracy.

A new class of kits, having enhanced precision at lower concentrations—either from greater sensitivity or from improved stability over time of the reagents and experimental conditions—should be developed to replace the second-generation methods currently used in the laboratory. The first third-generation TSH immunoassay with a functional sensitivity of about 0.02 milli-int. unit/L (10-fold greater than that of the second-generation IMAs) has been recently described (*12–14*) and appears to be a promising tool to refine the assessment of the thyroid status.

References

1. Bayer MF. Performance criteria for appropriate characterization of "(highly) sensitive" thyrotropin assays [Letter]. Clin Chem 1987;33:630–1.
2. Chubb SAP, Davis JA, Wood AT. Comparing the "sensitivity" of thyrotropin assays by using between-run precision profiles [Letter]. Clin Chem 1990;36:1855.
3. Spencer CA. Clinical utility and cost-effectiveness of sensitive thyrotropin assays in ambulatory and hospitalized patients. Mayo Clin Proc 1988;63:1214–22.
4. Pilo A, Zucchelli GC, Piro MA, Chiesa MR. Interlaboratory survey for T3, T4, and TSH assays in Italy. In: Radioimmunoassay and related procedures in medicine. Int Atomic Energy Agency, Vienna: 1982: 621–9.
5. Pilo A, Zucchelli GC, Chiesa MR, Bolelli GF, Albertini A. Components of variance analysis of data produced in a national quality-control survey of radioimmunoassay of thyroxin, triiodothyronine, thyrotropin, prolactin, and progesterone. Clin Chem 1986;32:171–5.
6. Bacon RR, Hunter WM, Mackenzie L. The U.K. external quality assessment schemes for peptide hormones: objectives and strategy. In: Hunter WM, Corrie JET, eds., Immunoassay for clinical chemistry. Edinburgh: Churchill Livingstone, 1983:669–79.
7. Seth J, Manning L, Bacon RR, Hunter WM. Progress and problems in immunoassay for serum pituitary gonadotropins: evidence from the U.K. external quality assessment scheme (EQAS), 1980–1988. Clin Chim Acta 1989;186:67–82.
8. Zucchelli GC, Pilo A, Chiesa MR, Piro MA. Progress report on a national quality-control survey of triiodothyronine and thyroxin assay. Clin Chem 1984;30:395–8.
9. Pilo A, Zucchelli GC, Chiesa MR, Masini S, Clerico A. The CNR external quality assessment program for immunoassay: statistical analysis and report for participants. Ann Ist Super Sanità 1991; 27:469–78.
10. Munson PJ, Rodbard D. An elementary component of variance analysis for multicentre quality control. In: Radioimmunoassay and related procedures in medicine, Vol. 2. Vienna: Int Atomic Energy Agency, 1977:106–13.
11. Bayer MF. Effect of interassay precision on diagnostic accuracy [Letter]. Clin Chem 1988;34:2605–6.
12. Nicoloff JT, Spencer CA. The use and misuse of the sensitive thyrotropin assays. J Clin Endocrinol Metab 1990;71:553–9.
13. Ross DS, Ardisson LJ, Meskell MJ. Measurement of thyrotropin in clinical and subclinical hyperthyroidism using a new chemiluminescent assay. J Clin Endocrinol Metab 1989;69:684–8.
14. Spencer CA, LoPresti JS, Patel A, et al. Applications of a new chemiluminometric thyrotropin assay to subnormal measurement. J Clin Endocrinol Metab 1990;70:453–60.