Reliability and Validity of Commercially Available, Direct Radioimmunoassays for Measurement of Blood Androgens and Estrogens in Postmenopausal Women

Sabina Rinaldi, Henri Déchaud, Carine Biessy, Véronique Morin-Raverot, Paolo Toniolo, Anne Zeleniuch-Jacquotte, Arslan Akhmedkhanov, Roy E. Shore, Giorgio Secreto, Antonio Ciampi,[1] Elio Riboli, and Rudolf Kaaks[2]

Unit of Nutrition and Cancer, International Agency for Research on Cancer, 69372 Lyon, France [S. R., C. B., E. R., R. K.]; Central Laboratory for Biochemistry, Hôpital de l'Antiquaille, 69005 Lyon, France [H. D., V. M-R.]; Departments of Obstetrics and Gynecology [P. T., A. A.] and Environmental Medicine [P. T., A. Z-J., A. A., R. E. S.], New York University School of Medicine, New York, New York 10010; and Unit of Endocrinology, National Cancer Institute, 20133 Milan, Italy [G. S.], Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, H3A 1A2 Canada [A. C.]

Abstract

In large-scale epidemiological studies on endogenous sex steroids and cancer risk, direct immunoassays of circulating hormone levels have the advantage of being fast and comparatively inexpensive while requiring only small sample volumes. On the other hand, indirect assays after organic extraction and chromatographic prepurification have the advantage of reducing specific interferences and matrix effects and hence are thought to have better validity. We compared direct assays of testosterone (T, six different assays), Δ4-androstenedione (A, four assays), estrone (E_1, one assay), and 17β-estradiol (E_2, five assays) with measurements obtained by an indirect assay in a representative subset of 20 postmenopausal women who were part of a large prospective cohort study.

Within-batch reproducibilities of the subject rankings by relative hormone levels were good (intraclass correlations >0.89) for all direct assays tested. Between batches, reproducibilities generally were also acceptable ($r > 0.80$) to good ($r > 0.90$) in terms of Pearson's correlations. The between-batch reproducibility in terms of intraclass correlations was systematically lower in terms of Pearson's correlations, however, because of between-batch variations in the absolute scale of measurements. The relative validity of direct *versus* indirect assays in terms of the subjects' ranking by relative hormone levels was also high for most of the kits tested for T, A, and E_1 (Pearson's correlations between 0.70 and 0.89) but was high for only two kits of five tested for E_2 (correlations of 0.86 and 0.84). On an absolute scale, mean measurement values were generally higher for direct assays than for the indirect assay and, for each hormone, varied substantially, depending on the kit used.

Overall, the results of this study show that, with careful selection, commercial kits for direct radioimmunoassays of steroid hormones in postmenopausal serum can be found that may allow a reliable estimation of relative risks in epidemiological studies. However, standardization of the absolute scale of assays remains problematic.

Introduction

Blood levels of sex steroids in postmenopausal women have been associated with risk of various chronic diseases, including cancer (1–5), cardiovascular disease (6–8), and osteoporosis (9–11). It is increasingly recognized, however, that the accurate estimation of hormone-disease associations requires that blood hormone concentrations be measured with a maximum level of reliability (12–16). For the estimation of relative risks, the key prerequisite is that subjects should be ranked accurately by their long-term, average hormone levels. Random errors in the measurement of individuals' hormone levels attenuate relative risk estimates and decrease the power of statistical tests for hormone-disease associations (12, 13, 16). One possible source of measurement errors are variations over time in subjects' hormone levels. Additional sources of variation are inaccuracies in the laboratory assays (12, 14–23). Because postmenopausal women have low blood levels of sex steroids, and especially of estrogens, assays must have sufficient sensitivity to allow accurate measurements.

Besides accuracy, assays must meet a number of practical criteria to be applicable in large-scale epidemiological investigations. Because often only a limited amount of serum or plasma is available per person, the assays should require only small serum volumes. Furthermore, the assays should be reasonably fast and inexpensive. Some methods, including those based on sample extraction plus chromatography and mass spectrometry, are generally thought to be highly accurate but require comparatively large volumes of serum, are labor-intensive, and expensive. Other accurate methods, such as "indirect" radioimmunological assays after extraction and chromatographic separation of the sex steroids, that were once the standard technique in many clinical laboratories require smaller volumes but remain comparatively labor intensive and slow. These methods, therefore, cannot be easily applied for routine measurements in large epidemiological investigations but can

Received 12/8/00; revised 4/13/01; accepted 5/2/01.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
[1] This project was carried out during Dr. Ciampi's sabbatical year at the International Agency for Research on Cancer.
[2] To whom requests for reprints should be addressed, at Unit of Nutrition and Cancer, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon, France. Fax: 33-4-72-73-83-61.

Table 1 Details of commercial kits for direct assays

Hormones and Commercial name[a]	RIA	Declared sensitivity	Required amount of sample	Comments
T				
Immunotech Rif 1119	Coated tubes	2.5 ng/dl	50 μl	
Orion Spectria	Coated tubes	2.9 ng/dl	25 μl	Avoid highly lipemic specimens
Cis-BIO CT-2	Coated tubes	2.9 ng/dl	25 μl	Avoid lipemic specimens
DSL-4100	Double antibody	5 ng/dl	50 μl	Avoid hemolyzed and lipemic specimens
Sorin CTK P3093	Coated tubes	5 ng/dl	50 μl	Avoid hemolyzed specimens
Byk RIA-mat	Coated tubes	2.9 ng/dl	25 μl	Avoid clotted, lipemic, hemolyzed, icteric, or contaminated specimens
A				
Immunotech Rif. 1322	Coated tubes	10 ng/dl	50 μl	
DSL-4200	Double antibody	2 ng/dl	50 μl	Avoid hemolyzed and lipemic specimens
Sorin CA-1725	Coated tubes	3 ng/dl	50 μl	Avoid grossly emolyzed or grossly lipemic specimens
DPC Coat-A-Count	Coated tubes	4 ng/dl	100 μl	
E_2				
Immunotech Ref 1663	Coated tubes	3 pg/ml	100 μl	
Cis Bio ESTR-US-CT	Coated tubes	1.36 pg/ml	200 μl	Avoid citrate plasma samples
DSL-39100	Double antibody	0.6 pg/ml	200 μl	Avoid hemolyzed and lipemic specimens
Sorin Estradiol 2	Double antibody	5 pg/ml	50 μl	
Bio Source E2-RIA-CT	Coated tubes	4.8 ± 1.2 pg/ml	100 μl	
E_1				
DSL-8700	Double antibody	1.2 pg/ml	50 μl	Avoid hemolyzed and lipemic specimens

[a] Immunotech = Immunotech, Marseille, France; Orion = Orion Diagnostica, Espoo, Finland; Cis-Bio = Cis-Bio International; Gif-sur-Yvette, France; DSL, Diagnostic System Laboratories, Webster, Texas; Sorin = DiaSorin, Saluggia, Italy; Byk = Byk-Sangtec Diagnostica, Dietzenbach, Germany; Bio-Source = Bio Source Europe, Nivelles, Belgium; DPC = Diagnostica Products Corporation, Los Angeles, CA.

serve more suitably as measurements for comparison in validation studies.

Direct immunoassays, which can be applied without preliminary treatment of plasma or serum samples, are generally very fast and require only small amounts of serum or plasma. However, commercially available direct assay kits are designed primarily for clinical use. In a clinical setting, the conditions under which blood samples are collected (*e.g.*, fasting conditions, to avoid lipemia) can be controlled. Furthermore, serum volumes are generally sufficient to carry out the assays, even with comparatively less sensitive methods, and the main interest is usually in identifying individuals with relatively extreme (pathological) hormone levels. In epidemiology, by contrast, the conditions under which samples are collected cannot always be fully controlled, available serum volumes are often smaller, and the aim is to classify subjects by relative hormone levels within the normal (*i.e.*, nonpathological) range. Therefore, the validity and precision of direct assays by commercial kits must be assessed before their use in large-scale epidemiological studies.

We report here the results of a study for the validation of a number of commercially available direct assays for T,[3] A, E_2, and E_1, in view of their possible use in prospective cohort studies on cancer risk. The direct assays were compared with indirect radioimmunoassays. By conducting this study on a representative subsample of women participating in an ongoing prospective cohort study, we could estimate meaningful measures of the accuracy of subject rankings by relative hormone levels.

Materials and Methods

General Outline. We measured the concentration of T, A, E_1, and E_2 by direct radioimmunoassays (Table 1) in serum samples from 20 postmenopausal women, and we compared these measurements with those obtained by radioimmunoassays after organic extraction and chromatographic prepurification on celite columns ("indirect method").

All direct assays were performed twice, on different days, and using kits with different lot numbers. Within each of these two batches, all 20 samples were assayed in duplicate, and the duplicates were averaged into a single measurement. The maximum time interval between the first and last batches of the same type of assay was ~6 months. The direct assays were done at the Unit of Nutrition and Cancer, IARC (Lyon, France). The indirect assays were also performed twice entirely on the same 20 serum samples at the Central Laboratory for Biochemistry, Hôpital de l'Antiquaille (Lyon, France). All indirect assays were done in a blinded fashion so that during the second (replicate) series of measurements, technicians had no knowledge about hormone concentrations measured during the first series.

Subjects and Blood Collection. Serum samples were taken from 20 postmenopausal women who participated in the New York University Women's Health Study, an ongoing prospective cohort study in New York. Details of blood collection and sample preparation in this cohort study have been described in detail elsewhere (24). The 20 study subjects were selected at random from a subset of about 2000 cohort participants who had donated blood to the original cohort at least four times

[3] The abbreviations used are: T, testosterone; A, Δ4-androstenedione; E_2, 17β-estradiol; E_1, estrone; GC-MS, gas chromatography-mass spectrometry; ICC, intraclass correlation coefficient; SHBG, sex hormone binding globulin; CI, confidence interval.

during the course of the study. For these 20 women, the 12 available 1-ml aliquots from each of the two most recent blood donations were retrieved from storage. These aliquots were then shipped to our laboratories on dry ice, thawed at room temperature, and pooled into a single serum sample of ~24 ml, from which new aliquots of ~1 ml were prepared and frozen, until the measurement of hormones by direct and indirect radioimmunoassays.

Direct Steroid Hormone Assays. Kits for the four steroids were chosen among those commercially available in France. Only direct immunoassays with ^{125}I tracer were chosen, because ^{125}I has a highly specific signal. All direct assays used polyclonal antibodies. An overview of the assays used is in Table 1. Assays were performed exactly according to the protocols from the manufacturers, using an automated liquid handling system with computer connections to a gamma counter.

Indirect Assays (Indirect Method). Our indirect assays of T, A, E_2, and E_1 were based on sample extraction by an organic solvent, partition chromatography on celite columns, collection of different elution fractions each containing one of the four steroids, and a duplicate RIA to quantify these steroids in each fraction. Corrections for incomplete recovery during the extraction and chromatography steps were made using ^{3}H-labeled internal standards. The androgens were measured in 1 ml of serum, whereas the estrogen measurements required a separate aliquot of 2 ml. Full details of our analytical procedures are in Appendix I. In the indirect assay, the detection limits for T and A were 1.5 and 3.5 ng/dl, and for E_1 and E_2 of 0.54 and 0.40 pg/ml, respectively.

The validity of the assays of T, A, E_2, and E_1 was evaluated using dilution and surcharge tests within the range of concentrations normally found for postmenopausal women. Dilutions were done with serum from which all sex steroids had first been removed by stripping with active charcoal. Dilution ranges were from 77.7 to 2.4 ng/dl for T, from 154.6 to 19.3 ng/dl for A, from 69 to 4 pg/ml for E_1, and from 43 to 1 pg/ml for E_2. Mean recoveries for dilution tests were 100.8% for T, 83.2% for A, 114.2% for E_1, and 104% for E_2. For the surcharge tests, a fixed volume of stripped serum was loaded with increasing, known quantities of each of the four sex steroids. Ranges of surcharge were between 6.3 and 100 ng/dl for T, between 6.2 and 200 ng/dl for A, 1.6 and 50 pg/ml for E_1, and between 6.6 and 50 pg/ml for E_2. Mean recoveries were 95.7% for T, 90% for A, 75% for E_1, and 90% for E_2.

Additional Assays of Standard Control Sera. For T and E_2, a complementary evaluation of the validity of direct and indirect assays was made by measurement of T and E_2 in a commercially available control serum PROBIOQUAL (Lyon, France; Reference RIA55). According to the manufacturer's indications, the reference values of these control sera had been obtained by GC-MS.

Statistical Analyses. All analyses were performed on log-transformed variables so as to approximately normalize their frequency distributions. Statistical analyses included the calculation of means and population SDs of the measurements and calculation of various correlation coefficients.

Within- and between-batch reproducibilities of the direct assays were evaluated by computing ICCs from the following model:

$$y = \text{subject} + \text{batch (subject)} + \text{error} \qquad \text{(A)}$$

where y denotes the values obtained for each of the duplicates in a direct assay, and the terms on the right hand side are random effects. The parentheses in Eq. A indicate nested effects. This model corresponds to the variance (Var) decomposition:

$$\text{Var}[y] = \text{Var}[\text{subject}] + \text{Var}[\text{batch}] + \text{Var}[\text{error}]$$

where Var[y] represents the total variance of the assay duplicates within each of the two batches and Var[subject], Var[batch], and Var[error] are the variances attributable to subject, to batch (within a given subject), and to random errors between the duplicates (within the same subject and the same batch), respectively. These variances were estimated by maximizing the likelihood of the random effects model of Eq. A, as implemented by the SAS procedure VARCOMP. Because the overall result of a direct assay, x, is defined as the average over the duplicates, $y_{1(b(s))}$ and $y_{2(b(s))}$, we have Var[x] = Var[subject] + Var[batch] + ½Var[error], and thus the within- and between-batch ICCs are defined as (Var[subject] + Var[batch])/(Var[subject] + Var[batch] + ½Var[error]), and Var[subject]/(Var[subject] + Var[batch] + ½Var[error]), respectively. Confidence intervals for the within- and between-batch ICCs were calculated according to a formula derived by the δ-method (see "Appendix II"). In addition to the ICCs, we estimated the between-batch reproducibility by the Pearson's product moment correlation between the direct measurements of the first and second batches.

The validity of the direct assays, relative to the subjects' averages of their two indirect assays, was evaluated by calculating Pearson's coefficient of correlation. Because the direct assays had been performed twice, we calculated the mean Pearson's correlation coefficient, averaging the two batches of direct assays. To do so, we concatenated the measurements from the two batches (*i.e.*, creating a data set with $2 \times 20 = 40$ observations for direct plus indirect assays), and we then computed a partial coefficient of correlation of the direct measurements with the indirect (reference) measurements adjusted for "batch." Somewhat conservative confidence intervals for these average Pearson's correlations were computed by standard methods based on the Z-transform (25), assuming a total number of 20 (not 40) observations.

Results

Means and population CIs of indirect and direct assays in the 20 postmenopausal serum samples are in Table 2. For each steroid, mean values varied widely between the different kits used. For all four steroids, mean values by the direct assays were systematically higher than those obtained with the indirect method. For only two women did we obtain assay results below the detection limit, and both of these were for testosterone (IMMUNOTECH and SORIN kits), and in batch 2. For these women, we used the detection limit as a measurement value.

Variance components related to "subject," "batch," or "error" and ICCs for within- and between-batch reproducibility of the direct and indirect assays are in Table 3. Within batches, the reproducibility of subject rankings by the direct assays was generally high, with ICCs all >0.89 and more than half of these correlations >0.95. For the indirect assay, samples were not measured in duplicate in each batch, and therefore within-batch correlations could not be estimated. Between batches, Pearson's correlations were also high, both for the direct and indirect assays, all >0.80, and more than half of them >0.90. The between-batch ICCs, however, were more variable for the direct assays and ranged from 0.53 to 0.94 for all direct assays except one, the E_2 assay by SORIN, which had an ICC of only 0.28. For the indirect method, all between-batch ICCs were between 0.73 and 1.0.

The validity of ranking, as judged from Pearson's corre-

Table 2 Mean and confidence intervals (CIs) of testosterone, Δ4-androstenedione, estradiol, and estrone measurements by indirect method and direct assays in 20 serum samples of postmenopausal women

Hormone and method	Batch 1		Batch 2	
	Mean	CI	Mean	CI
Testosterone (ng/dl)				
CELITE	11.14	8.20–15.14	11.83	8.87–15.79
IMMUNOTECH	18.87	14.67–24.28	11.31	8.02–15.95
ORION	20.81	15.79–27.43	29.58	22.26–39.32
Cis-Bio	35.33	28.8–43.35	25.04	18.40–34.09
DSL	35.14	29.06–42.48	24.41	18.99–31.38
SORIN	21.20	16.50–27.24	18.08	13.47–24.28
BYK	23.68	17.86–31.40	27.65	19.83–38.55
Δ4-Androstenedione (ng/dl)				
CELITE	40.32	29.95–54.29	37.64	28.24–50.18
IMMUNOTECH	65.41	45.94–93.12	126.35	95.16–167.75
DSL	106.81	85.34–133.67	112.55	87.1–145.44
SORIN	93.85	74.55–118.14	107.26	87.22–131.91
DPC	73.73	54.24–100.23	89.38	65.10–122.71
Estradiol (pg/ml)				
CELITE	1.84	1.30–2.62	2.85	2.19–3.72
IMMUNOTECH	15.23	12.49–18.58	18.13	14.95–21.98
Cis-Bio	7.40	5.75–9.53	8.53	7.02–10.36
DSL	7.60	5.96–9.70	9.84	8.31–11.65
SORIN	43.84	40.40–47.57	31.84	27.96–36.25
BioSource	21.38	17.26–26.48	23.96	21.07–27.24
Estrone (pg/ml)				
CELITE	8.56	6.47–11.31	5.94	4.15–8.51
DSL	18.33	15.38–21.84	24.35	21.25–27.92

lations between direct and indirect assays, was good for all six kits for T ($r = 0.70–0.86$), for all four kits for A ($r = 0.82–0.89$), and for the assay of E_1 ($r = 0.81$; Table 4). For E_2, however, only two direct kits gave good correlations with the indirect assay (DSL, $r = 0.84$; SORIN, $r = 0.86$), whereas the assays by IMMUNOTECH, Cis-Bio, and BioSource kits all had correlations <0.65.

The low correlations for E_2 for the direct assays by the Immunotech ($r = 0.29$), BioSource ($r = 0.42$), and to some extent, Cis-Bio kits ($r = 0.65$) systematically appeared to be the result of errors for the same three individuals. This is illustrated in Fig. 1, with data from batch 2. After exclusion of these three subjects, the correlations between direct and indirect assays of E_2 improved, to 0.77 for IMMUNOTECH, 0.77 for BioSource, and 0.87 for Cis-Bio, whereas the correlations for DSL and SORIN did not change. These observations suggest a problem of strong matrix effects in the serum samples of these three women, attributable to specific interferences by other estradiol-like compounds or to nonspecific interferences by compounds that alter the antigen-antibody reaction (see also "Discussion").

Results of direct and indirect assays of the PROBIOQUAL control serum showed that, for E_2, the reference value of 47.7 pg/ml was measured accurately by the direct kits from DSL (46.9 pg/ml) and SORIN (51.8 pg/ml), whereas all other assays resulted in substantially lower values (indirect assay, 30.5 pg/ml; IMMUNOTECH, 20.9 pg/ml; Cis-Bio, 21.9 pg/ml). For T, values close to the reference (136 ng/dl) were obtained by the DSL kit (127 ng/dl), whereas the indirect assay values were slightly underestimated (111 ng/dl) and the direct assay values by the CisBio kit were slightly overestimated (164 ng/dl). The other values obtained for T were further away from the reference value given by the manufacturer; these included ORION (97.0 ng/dl), BYK (77.8 ng/dl), SORIN (77.9 ng/dl), and IMMUNOTECH (31.9 ng/dl).

Discussion

In this study, we evaluated the reproducibility and relative validity of measurements of sex steroids by direct radioimmunoassays using commercially available kits, comparing with indirect assays after sample extraction and chromatographic prepurification.

Study Design. To our knowledge, this study is the first to compare direct with indirect assays as an approximate reference, while at the same time being conducted on a representative sample of women from a well-defined epidemiological study population. Several previous studies (13, 15, 16, 19) have addressed the reproducibility of hormonal assays within and between laboratories but generally only included small numbers of subjects (from 4 to 15, including both pre- and postmenopausal women), who generally were not selected as a representative sample from a well-defined epidemiological study population. Other studies (26–30) have estimated the correlations between replicate hormone assays by the same technique in blood samples collected at different points in time from a given series of study subjects. The latter studies were aimed primarily at evaluating whether biological variation in hormone levels over time was a major possible source of random misclassification by long-term hormone status.

By conducting our study on a representative sample of postmenopausal women in the New York University Women's Health Study, we could estimate not only the variances of measurement errors within and between batches but also the representative between-subject variations in hormone levels within this cohort. We could thus estimate coefficients of correlation that indicate the extent to which laboratory errors affected the reproducibility and validity of the ranking of subjects by relative hormone levels. Because our study was conducted on representative population sample, these correlation coefficients can be translated into an expected magnitude of attenuation bias in relative risk estimates within the New York University Women's Health Study cohort caused by random assay errors (31). This study was not designed, however, to examine the reproducibility of serum hormone concentrations over time, which constitutes an additional source of random error and misclassification.

Measurements obtained by mass spectrometric identification and quantification after organic extraction and chromatographic prepurification are generally considered the "gold standard" for steroid hormone assays. However, despite improvements during recent years, such methods have not reached a degree of sensitivity that is sufficient to measure E_2 at postmenopausal concentrations. By contrast, our indirect radioimmunoassays had sufficiently low detection limits and hence could be applied on reasonably small serum volumes. The indirect immunoassays are an interesting comparison method against which direct assays can be validated, because they strongly reduce the cross-reaction of assay antibodies with other hormone-like substances, and because they eliminate nonspecific interferences by substances that can modify the rates of reaction between antigen (the hormone to be measured) and antibodies. These so-called "matrix effects" are probably the main source of random and systematic errors in direct assays (17, 18, 32, 33).

Reproducibility. All direct assays tested in this study showed a high degree of reproducibility. Reproducibility of the classification of subjects by hormone level was calculated in terms of both Pearson's product moment correlation and ICCs. In this study of 20 subjects, which were few enough to be all analyzed together within a single analytical batch of direct immunoas-

Table 3 Subject variance, batch variance, error variance, ICC within and between batches of T, A, E_2, and E_1 measurements by direct and indirect assays and their CIs

Hormone and method	Var (subject)	Var (batch)	Var (error)	Correlation		
				Within batch (intraclass)	Between batch	
					Intraclass	Pearson
T						
CELITE	0.458	$3\ 10^{-5}$	/	/	1.00	0.99
					/	(0.97–1.00)
IMMUNOTECH	0.349	0.177	0.023	0.98	0.65	0.91
				(0.97–0.99)	(0.47–1.00)	(0.79–0.97)
ORION	0.318	0.077	0.073	0.92	0.74	0.87
				(0.87–0.96)	(0.58–1.00)	(0.69–0.95)
Cis-Bio	0.247	0.131	0.008	0.99	0.65	0.84
				(0.99–1.00)	(0.46–1.00)	(0.63–0.93)
DSL	0.136	0.119	0.002	1.00	0.53	0.80
				(0.99–1.00)	(0.33–1.00)	(0.53–0.93)
SORIN	0.348	0.012	0.057	0.93	0.89	0.93
				(0.89–0.97)	(0.82–0.99)	(0.83–0.097)
BYK	0.393	0.083	0.029	0.97	0.80	0.80
				(0.95–0.99)	(0.67–1.00)	(0.56–0.92)
A						
CELITE	0.438	0.004	/	/	0.99	0.99
					/	(0.99–1.00)
IMMUNOTECH	0.401	0.225	0.031	0.98	0.62	0.97
				(0.96–0.99)	(0.44–1.00)	(0.94–0.99)
DSL	0.272	0.024	0.010	0.98	0.90	0.90
				(0.98–0.99)	(0.83–1.00)	(0.76–0.96)
SORIN	0.226	0.023	0.009	0.98	0.89	0.92
				(0.97–0.99)	(0.81–0.99)	(0.81–0.97)
DPC	0.486	0.020	0.019	0.98	0.94	0.98
				(0.97–0.99)	(0.90–0.99)	(0.94–0.99)
E_2						
CELITE	0.402	0.143	/	/	0.73	0.92
					/	(0.81–0.97)
IMMUNOTECH	0.186	0.016	0.010	0.98	0.90	0.97
				(0.97–0.99)	(0.82–0.99)	(0.92–0.99)
Cis-Bio	0.246	0.014	0.018	0.97	0.91	0.98
				(0.95–0.98)	(0.85–0.99)	(0.95–0.99)
DSL	0.188	0.053	0.007	0.99	0.77	0.95
				(0.92–0.99)	(0.62–1.00)	(0.89–0.98)
SORIN	0.025	0.059	0.006	0.97	0.28	0.90
				(0.96–0.97)	(0.12–1.00)	(0.77–0.96)
BioSource	0.132	0.023	0.017	0.95	0.81	0.94
				(0.93–0.97)	(0.68–1.00)	(0.86–0.98)
E_1						
CELITE	0.472	0.094	/	/	0.83	0.97
					/	(0.92–0.99)
DSL	0.093	0.038	0.034	0.89	0.63	0.91
				(0.85–0.92)	(0.44–1.00)	(0.77–0.96)

Table 4 Pearson correlations between direct assays and the indirect method

	DSL	IMMUNOTECH	SORIN	Cis-Bio	BYK	ORION	BioSource	DPC
T	0.76	0.86	0.76	0.70	0.78	0.79		
A	0.82	0.89	0.86					0.85
E_2	0.84	0.29	0.86	0.65			0.42	
E_1	0.81							

says, Pearson's product moment correlation reflects agreement between the subjects' relative differences in hormone level on a linear scale. The ICC, in addition, reflects also the tendencies of the absolute scales of two or more series of measurements to agree in terms of their means and SDs. It thus indicates the Pearson's correlation for reproducibility (unadjusted for batch) that one would expect for a much larger series of serum samples that can be analyzed only by multiple analytical batches of immunoassays.

Within-batch ICCs generally were >0.95 and always >0.89. The between-batch reproducibilities in terms of Pearson's correlations was also acceptable ($r > 0.80$) to good ($r > 0.90$). The between-batch ICCs, however, were generally lower than Pearson's correlation coefficients, and for a few assays (T assay by DSL, estradiol assay by SORIN) this difference was substantial. A low between-batch ICC in presence of a high Pearson's correlation indicates a good agreement between assays in terms of the subjects' rankings within each batch but poor agreement in terms of the absolute scale of measurements across batches. In practice, between-batch scale variations will tend to be larger when the kits used have been purchased in different time periods and when they have different lot numbers. A difference between lot numbers indicates that at least

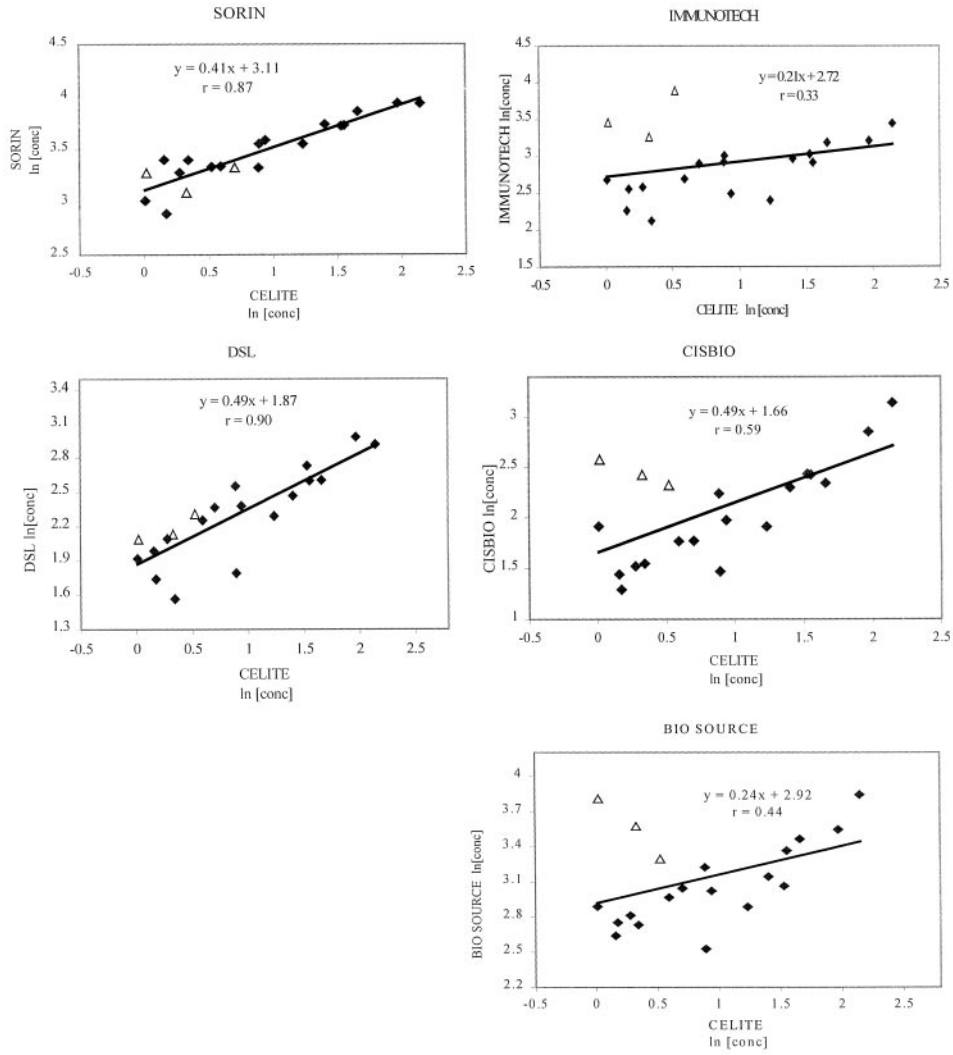

Fig. 1. Scatter plots for estradiol (pg/ml) as measured by direct assays (batch 2) and by the indirect method. All values were log normal transformed. △, the three outliers that explain the low correlations of direct assays by IMMUNOTECH, Cis-Bio, and BioSource.

part of the kit's reagents (*e.g.*, radioactively labeled tracers, antibodies, and standards) has been prepared as a separate lot by the manufacturer. The contrast between elevated reproducibility in terms of Pearson's correlations or in terms of (between-batch) ICCs implies that, to minimize misclassifications, the sera from cases with disease and control subjects should preferably be analyzed together within the same analytical batch (while matching the cases and controls for age and other possible confounding factors).

Validity. Apart from a high reproducibility within and between batches, most of the direct assays for T, A, and E_1 showed also good correlations (Pearson's $r > 0.70$) with the indirect assays. This suggests that matrix effects, which in theory may vary substantially between individuals, caused comparatively little random error and subject misclassification in the direct assays. For E_2, however, which has very low serum concentrations in postmenopausal women, only two kits (DSL and SORIN) of five tested showed good correlations with the indirect assays (>0.84). Interestingly, these two kits both used a second antibody for the separation of free and bound steroids (by immunoprecipitation of the complexes that the antigen forms with the first antibody), in contrast to the other three, which used only a single antibody on coated tubes. The correlations of each of the latter three assays with our indirect method improved substantially after the exclusion of three women from the statistical analyses. This suggests the presence of strong matrix effects especially affecting the assays of these three subjects.

Although accurate ranking of individuals by relative hormone levels is a primary requisite for hormone assays to be useful in epidemiology, an important secondary selection criterion is that measurements should be obtained on a valid absolute scale. Correct scaling of measurements enhances comparability of results between studies and, combining with SHBG measurements, may be also important for the calculation of approximate levels of bioavailable T or E_2 not bound to SHBG (34, 35).

Our indirect assays showed to be reasonably valid for all four steroids. Surcharge and dilution tests within the postmenopausal range showed an acceptable validity for the absolute scale of the measurements, with recoveries close to 100% for all four steroids. Only in the surcharge test with E_1 was the recovery less complete (75%), suggesting some degree of underestimation of E_1 levels. On the other hand, comparisons with the PROBIOQUAL standard showed ~35% underestimation of E_2 levels by indirect method and a 18% underestimation of T.

It is unclear what could have been the cause of this relative underestimation compared with the PROBIOQUAL GC-MS reference values given by the manufacturer. One possible source of error is incomplete recovery of sex steroids from the extraction of serum samples or from chromatography, but this could then also affect the GC-MS assays. It is noteworthy that the PROBIOQUAL standard had higher concentrations, within the premenopausal range, than the levels we tested in the surcharge and dilution tests. Unfortunately, control sera where sex steroid levels in the postmenopausal range have been measured by GC-MS are not available.

For the direct assays, mean values were generally within what is considered the biologically plausible range (36, 37) but could vary substantially for a given hormone, depending on the kit used. Up to 20-fold differences were observed for mean levels of E_2 and up to almost 3-fold differences for mean levels of A (33, 38). Similarly large differences in absolute values were also noted in other studies, where aliquots of the same serum samples were sent to different laboratories and/or analyzed by different direct or indirect RIA methods (15, 16, 19, 33, 38). In comparison with our indirect assays, most of the direct assays resulted in approximately 2.0- to 2.5-fold higher mean measurements of T and A, 2-fold higher levels of E_1, and between 3.5- and almost 20-fold higher levels of E_2. In comparison to the PROBIOQUAL standard, however, some of the direct assays (especially Cis-Bio and DSL) had a smaller difference with the reference values than our indirect assay, which showed an average underestimation of E_2 values by ~35%.

Causes for such variations between mean direct assay levels can be multiple. A main source of error may be specific interferences by other hormone-like substances. For instance, the use of exogenous estrogen analogues for the treatment of menopausal symptoms has been shown to systematically affect in mean serum E_2 measurements (39). Besides specific interferences by other hormones, bias in mean group-level measurements may occur when there are systematic differences between the matrices of natural serum or plasma samples, and the matrix of the standards from commercial assay kits, which are generally reconstituted, artificial sera. Substances that have been implicated in causing matrix effects include serum lipids (40), hemoglobin, and bilirubin, and it is generally recommended that direct assay methods should not be used on lactescent, hemolyzed, or icteric sera. Water-soluble substances, such as uric acid (39), or other, unidentified substances (41) may also disturb certain immunoassays. Finally, direct assays of T and E_2 may be disturbed by extreme serum concentrations of SHBG, which binds T and E_2 with high specificity (14, 17, 18, 20).

In conclusion, direct immunoassays have the advantage over indirect assays of requiring smaller volumes of serum or plasma. A further advantage of direct immunoassays is that, because of their relative simplicity, they are amenable to automation. With a semiautomated system of a liquid handling robot with direct computer connections to a radioactivity counter, one laboratory technician can process up to 5 or 6 direct assay kits per day, which is equivalent to measuring six different sex steroids in duplicate for up to 40 subjects. A high speed of analysis has obvious advantages in terms of labor cost but also allows samples to be thawed and analyzed for several hormones within a single day, and even for large studies allows laboratory measurements to be completed within only weeks using as much as possible kits that carry the same lot number. Our results show that, with careful selection, and depending on the specific study population and samples to be analyzed, commercial kits can provide accurate results in terms of relative ranking by hormone level. Over longer time periods, however, continuous monitoring of the accuracy may be needed (*e.g.*, by comparison with an indirect assay), because the reagents of kits of given brands may change.

Acknowledgments

We thank Adalberto Cavalleri and Francine Claustrat for expert advice and constructive comments; David Achaintre and Béatrice Vozar for technical assistance with the direct immunoassays; and Jennie Dehedin for secretarial help.

Appendix I

Our indirect assays consisted of a sample extraction by an organic solvent, a chromatographic prepurification, and a RIA. Corrections for incomplete recovery were made using ^{3}H-labeled internal standards.

Extraction. For T and A assays, 100 μl of a phosphate albumin buffer solution containing ~6000 dpm (~8.0 pg) of [^{3}H]T and 6000 dpm (~11.0 pg) of [^{3}H]A were added as internal standards to 1 ml of serum to determine the analytical recovery of T and A from extraction plus chromatography. For E_1 and E_2 assays, 100 μl of a phosphate albumin buffer solution containing ~2000 dpm (~2.4 pg) of [^{3}H]E_1 and 2000 dpm (~3.0 pg) of [^{3}H]E_2 were added to 2 ml of serum to determine analytical recovery of E_1 and E_2. All serum samples were equilibrated for 15 min with the labeled steroids at 37°C. Steroids were then extracted using 5.5 ml of diethyl ether. The aqueous phase was frozen, and the organic phase was dried under nitrogen gas.

Partition Chromatography. The dried extracts were resuspended in 1 ml of iso-octane and subjected to celite column partition chromatography (stationary phase: 1 ml/2 g with ethylene glycol/1,2-propanediol 50/50 for T and A, and 1 ml/2 g with ethylene glycol for E_1 and E_2). Samples for androgen analysis were eluted successively with 4 ml of iso-octane (fraction 1), 4 ml of ethylacetate/iso-octane (13:100 by vol; fraction 2), and 5.5 ml of ethylacetate/iso-octane (20:100 by vol; fraction 3). A and T were eluted in fractions 2 and 3, respectively, and dried by evaporation. Samples for estrogen analysis were eluted successively with 4 ml of ethylacetate/iso-octane (5:100 by vol; fraction 1), 4 ml of ethylacetate/iso-octane (15:100 by vol; fraction 2), and 5 ml of ethylacetate/iso-octane (40:100 by vol; fraction 3). E_1 and E_2 were also eluted in fractions 2 and 3, respectively, and dried by evaporation.

Estimation of Recovery. All of the dried fractions were redissolved in a phosphate albumin buffer (0.5 ml for the fractions containing T and A; 0.3 ml for E_1 and E_2). The recovery of the extraction and chromatographic steps was calculated by counting for 10 min the radioactivity of the [^{3}H]T and [^{3}H]A internal standards in 200-μl aliquots of the redissolved androgen fractions and of [^{3}H]E_1 and [^{3}H]E_2 standards in the 80-μl aliquot of the redissolved estrogen fractions.

Radioimmunoassays. Quantification of T, A, E_1, and E_2 by RIA was done in duplicate. For the assays of T and A, we incubated for 2 h at 4°C 100 μl of each of the redissolved steroid fractions (or of a standard sample) with 200 μl of antiserum plus 100 μl of ^{3}H-labeled standards (25,000 dpm) for each of the two steroids. For the assays of estrogens in postmenopausal women, we used sequential incubation to lower the detection limits. We incubated 100 μl of the redissolved steroid fractions (or of standard sample) with 100 μl of antiserum for 24 h at 4°C. After 20 h of incubation, we added 100 μl of [^{3}H]E_1 or [^{3}H]E_2 (15,000 dpm) for 45 min for E_1 and for 30 min for E_2. After incubation, we added 0.5 ml of a charcoal-dextran suspension to all androgen and estrogen fractions and

centrifuged for 10 min at 3000 × *g* and at 4°C. We placed 0.5 ml of the resulting supernatant into a scintillation vial, added 3 ml of scintillation fluid, and counted the radioactivity of the supernatant. We corrected the calculated concentration for sample volume and for the percentage of recovery from extraction plus chromatography.

Appendix II

According to the delta method, if a vector parameter θ is estimated by the maximum likelihood estimator θ_{mle}, then a differentiable real-valued function $f(\theta_{mle})$ is also asymptotically normal with variance given by the δ-method formula:

$$\mathrm{Var}(f) = \mathrm{grad}(f)^{\mathrm{T}} \times B(\theta_{mle}) \times \mathrm{grad}(f)$$

where $B(\theta_{mle})$ is the variance-covariance matrix of θ_{mle}, obtained as output of any maximum likelihood estimation procedure, and grad(*f*) denotes the gradient of *f*, *i.e.*, the vector of the partial derivatives of *f* with respect to the components of θ. Because taking the inverse of ICC improves approximation to the Normal distribution, we used the δ-method to obtain $\mathrm{Var}(\mathrm{ICC}^{-1})$ and, correspondingly, the 95% CI: $\mathrm{ICC}^{-1} \pm 1.96\mathrm{Var}(\mathrm{ICC}^{-1})$, from which we obtained the 95% CI for ICC:

$$[(\mathrm{ICC}^{-1} + 1.96\mathrm{Var}(\mathrm{ICC})^{-1}) - 1,$$

$$(\mathrm{ICC}^{-1} - 1.96\mathrm{Var}(\mathrm{ICC}^{-1}) - 1]$$

Writing: θ = (Var[subject], Var[batch], Var[error])$^{\mathrm{T}}$, the expression for grad$(\mathrm{ICC}_{\mathrm{Witheen}}{}^{-1})^{\mathrm{T}}$ is:

$$(1/\mathrm{Var}[\mathrm{subject}]\{1 - \mathrm{Var}[x]/\mathrm{Var}[\mathrm{subject}]\},$$

$$1/\mathrm{Var}[\mathrm{subject}], 1/\mathrm{Var}[\mathrm{subject}])$$

and similarly, we have for grad$(\mathrm{ICC}_{\mathrm{Between}}{}^{-1})^{\mathrm{T}}$:

$$(1/\mathrm{Var}[\mathrm{batch}], 1/\mathrm{Var}[\mathrm{batch}]\{1 - \mathrm{Var}[x]/\mathrm{Var}[\mathrm{batch}]\},$$

$$1/\mathrm{Var}[\mathrm{batch}])$$

References

1. Weiderpass, E., Baron, J. A., Adami, H. O., Magnusson, C., Lindgren, A., Bergstrom, R., Correia, N., and Persson, I. Low-potency oestrogen and risk of endometrial cancer: a case-control study. Lancet, *353:* 1824–1828, 1999.

2. Thomas, H. V., Reeves, G. K., and Key, T. J. Endogenous estrogen and postmenopausal breast cancer: a quantitative review. Cancer Causes Control, *8:* 922–928, 1997.

3. Colditz, G. A., Hankinson, S. E., Hunter, D. J., Willett, W. C., Manson, J. E., Stampfer, M. J., Hennekens, C., Rosner, B., and Speizer, F. E. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. N. Engl. J. Med., *332:* 1589–1593, 1995.

4. Bernstein, L., and Ross, R. K. Endogenous hormones and breast cancer risk. Epidemiol. Rev., *15:* 48–65, 1993.

5. Secreto, G., Toniolo, P., Berrino, F., Recchione, C., Cavalleri, A., Pisani, P., Totis, A., Fariselli, G., and Di Pietro, S. Serum and urinary androgens and risk of breast cancer in postmenopausal women. Cancer Res., *51:* 2572–2576, 1991.

6. Kardos, A., and Casadei, B. Hormone replacement therapy and ischaemic heart disease among postmenopausal women. J. Cardiovasc. Risk, *6:* 105–112, 1999.

7. Rosano, G. M., and Panina, G. Oestrogens and the heart. Therapie, *54:* 381–385, 1999.

8. Sarrel, P. M. Cardiovascular aspects of androgens in women. Semin. Reprod. Endocrinol., *16:* 121–128, 1998.

9. Vestergaard, P., Hermann, A. P., Orskov, H., and Mosekilde, L. Effect of sex hormone replacement on the insulin-like growth factor system and bone mineral: a cross-sectional and longitudinal study in 595 perimenopausal women participating in the Danish Osteoporosis Prevention Study. J. Clin. Endocrinol. Metab., *84:* 2286–2290, 1999.

10. Stone, K., Bauer, D. C., Black, D. M., Sklarin, P., Ensrud, K. E., and Cummings, S. R. Hormonal predictors of bone loss in elderly women: a prospective study. The Study of Osteoporotic Fractures Research Group. J. Bone Miner. Res., *13:* 1167–1174, 1998.

11. Riggs, B. L., Khosla, S., and Melton, L. J., III. A unitary model for involutional osteoporosis: estrogen deficiency causes both type I and type II osteoporosis in postmenopausal women and contributes to bone loss in aging men. J. Bone Miner Res., *13:* 763–773, 1998.

12. Potischman, N., Falk, R. T., Laiming, V. A., Siiteri, P. K., and Hoover, R. N. Reproducibility of laboratory assays for steroid hormones and sex hormone-binding globulin. Cancer Res., *54:* 5363–5367, 1994.

13. Falk, R. T., Dorgan, J. F., Kahle, L., Potischman, N., and Longcope, C. Assay reproducibility of hormone measurements in postmenopausal women. Cancer Epidemiol. Biomark. Prev., *6:* 429–432, 1997.

14. Boots, L. R., Potter, S., Potter, D., and Azziz, R. Measurement of total serum testosterone levels using commercially available kits: high degree of between-kit variability. Fertil. Steril., *69:* 286–292, 1998.

15. Gail, M. H., Fears, T. R., Hoover, R. N., Chandler, D. W., Donaldson, J. L., Hyer, M. B., Pee, D., Ricker, W. V., Siiteri, P. K., Stanczyk, F. Z., Vaught, J. B., and Ziegler, R. G. Reproducibility studies and interlaboratory concordance for assays of serum hormone levels: estrone, estradiol, estrone sulfate, and progesterone. Cancer Epidemiol. Biomark. Prev., *5:* 835–844, 1996.

16. Hankinson, S. E., Manson, J. E., London, S. J., Willett, W. C., and Speizer, F. E. Laboratory reproducibility of endogenous hormone levels in postmenopausal women. Cancer Epidemiol. Biomark. Prev., *3:* 51–56, 1994.

17. Slaats, E. H., Kennedy, J. C., and Kruijswijk, H. Interference of sex-hormone binding globulin in the "Coat-A-Count" testosterone no-extraction radioimmunoassay. Clin. Chem., *33:* 300–302, 1987.

18. Masters, A. M., and Hahnel, R. Investigation of sex-hormone binding globulin interference in direct radioimmunoassays for testosterone and estradiol. Clin. Chem., *35:* 979–984, 1989.

19. McShane, L. M., Dorgan, J. F., Greenhut, S., and Damato, J. J. Reliability and validity of serum sex hormone measurements. Cancer Epidemiol. Biomark. Prev., *5:* 923–928, 1996.

20. Cook, N. J., and Read, G. F. Oestradiol measurement in women on oral hormone replacement therapy: the validity of commercial test kits. Br. J. Biomed. Sci., *52:* 97–101, 1995.

21. Diver, M. J., and Nisbet, J. A. Warning on plasma oestradiol measurement. Lancet, *2:* 1097, 1987.

22. Nisbet, J. A., and Jomain, P. A. Discrepancies in plasma estradiol values obtained with commercial kits. Clin. Chem., *33:* 1672, 1987.

23. Lee, C. S., Smith, N. M., and Kahn, S. N. Analytic variability and clinical significance of different assays for serum estradiol. J. Reprod. Med., *36:* 156–160, 1991.

24. Toniolo, P. G., Levitz, M., Zeleniuch-Jacquotte, A., Banerjee, S., Koenig, K. L., Shore, R. E., Strax, P., and Pasternack, B. S. A prospective study of endogenous estrogens and breast cancer in postmenopausal women. J. Natl. Cancer Inst., *87:* 190–197, 1995.

25. Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. Applied Regression Analysis and Other Multivariable Methods. Belmont, California: Duxbury Press, 1987.

26. Cauley, J. A., Gutai, J. P., Kuller, L. H., and Powell, J. G. Reliability and interrelations among serum sex hormones in postmenopausal women. Am. J. Epidemiol., *133:* 50–57, 1991.

27. Micheli, A., Muti, P., Pisani, P., Secreto, G., Recchione, C., Totis, A., Fissi, R., Cavalleri, A., Panico, S., and Berrino, F. Repeated serum and urinary androgen measurements in premenopausal and postmenopausal women. J. Clin. Epidemiol., *44:* 1055–1061, 1991.

28. Toniolo, P., Koenig, K. L., Pasternack, B. S., Banerjee, S., Rosenberg, C., Shore, R. E., Strax, P., and Levitz, M. Reliability of measurements of total, protein-bound, and unbound estradiol in serum. Cancer Epidemiol. Biomark. Prev., *3:* 47–50, 1994.

29. Hankinson, S. E., Manson, J. E., Spiegelman, D., Willett, W. C., Longcope, C., and Speizer, F. E. Reproducibility of plasma hormone levels in postmenopausal women over a 2–3-year period. Cancer Epidemiol. Biomark. Prev., *4:* 649–654, 1995.

30. Muti, P., Trevisan, M., Micheli, A., Krogh, V., Bolelli, G., Sciajno, R., and Berrino, F. Reliability of serum hormones in premenopausal and postmenopausal women over a one-year period. Cancer Epidemiol. Biomark. Prev., *5:* 917–922, 1996.

31. de Klerk, N. H., English, D. R., and Armstrong, B. K. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. Int. J. Epidemiol., *18:* 705–712, 1989.

32. Bolelli, G. F., Franceschetti, F., Malvano, R., Mimmi, P., Pilo, A., Rota, G., and Zucchelli, G. C. An interlaboratory study of lipid effects on steroid radioimmunoassay. J. Nucl. Med. Allied Sci., *30:* 209–213, 1986.

33. Schioler, V., and Thode, J. Six direct radioimmunoassays of estradiol evaluated. Clin. Chem., *34:* 949–952, 1988.

34. Vermeulen, A., Verdonck, L., and Kaufman, J. M. A critical evaluation of simple methods for the estimation of free testosterone in serum. J. Clin. Endocrinol. Metab., *84:* 3666–3672, 1999.

35. Belgorosky, A., Escobar, M. E., and Rivarola, M. A. Validity of the calculation of non-sex hormone-binding globulin-bound estradiol from total testosterone, total estradiol and sex hormone-binding globulin concentrations in human serum. J. Steroid Biochem., *28:* 429–432, 1987.

36. Clinical Assays. *In:* L. Speroff, G. H. Glass, and N. G. Kase (eds.), Clinical Gynecologic Endocrinology and Infertility, pp. 967–989. Baltimore, MD: Williams and Wilkins, 1994.

37. Shoupe, D., Brenner, P. F., and Mishell, D. R. J. Menopause. *In:* R. A. Lobo, D. R. Mishell, Jr., R. J. Paulson, and D. Shoupe (eds.), Infertility, Contraception, and Reproductive Endocrinology, pp. 415–448. Cambridge, MA: Blackwell Science, 1997.

38. Patricot, M. C., Badonnel, Y., Bugugnani, M. J., Collignon, I., Delvigne, L., Lacroix, I., and Mathian, B. Validity of immunochemical assays for blood estradiol. Study conducted in 1994. Ann. Biol. Clin., *53:* 399–406, 1995.

39. Levesque, A., Letellier, M., Dillon, P. W., and Grant, A. Analytical performance of Bayer Immuno 1 estradiol and progesterone assays. Clin. Chem., *43:* 1601–1609, 1997.

40. Wheeler, M. J., D'Souza, A., Matadeen, J., and Croos, P. Ciba Corning ACS:180 testosterone assay evaluated. Clin. Chem., *42:* 1445–1449, 1996.

41. Leung, Y. S., Dees, K., Cyr, R., Schloegel, I., and Kao, P. C. Falsely increased serum estradiol results reported in direct estradiol assays. Clin. Chem., *43:* 1250–1251, 1997.