

## Assay Reproducibility of Hormone Measurements in Postmenopausal Women

Roni T. Falk,<sup>1</sup> Joanne F. Dorgan, Lisa Kahle, Nancy Potischman, and Christopher Longcope

Environmental Epidemiology Branch [R. T. F., N. P.] and Division of Cancer Prevention and Control [J. F. D.], National Cancer Institute, NIH, Bethesda, Maryland 20892-7372; Information Management Services, Inc., Silver Spring, Maryland 20904 [L. K.]; and Department of Obstetrics and Gynecology, University of Massachusetts, Worcester, Massachusetts 01655-0321 [C. L.]

### Abstract

As part of a breast cancer case-control study of serum hormones conducted in Columbia, MO, we included several replicate quality control samples to monitor the consistency of laboratory assays. Sera were obtained from three postmenopausal women; from each woman, three samples were placed randomly in each of nine batches with the laboratory unaware of which sample corresponded to whom. Laboratory assays for estrone ( $E_1$ ), estradiol ( $E_2$ ), testosterone, androstenedione (Adione),  $E_1SO_4$ , dehydroepiandrosterone sulfate (DHEAS), follicle-stimulating hormone (FSH), sex hormone binding globulin (SHBG), and percentages of free and albumin-bound  $E_2$  were done at a single academic facility. ANOVA results showed that hormone values varied considerably from one batch to the next. The overall coefficients of variation (CVs) estimated for  $E_2$ , percentage of unbound  $E_2$ , and percentage of albumin-bound  $E_2$  were higher than 15%, but of these, only percentage of unbound  $E_2$  had both inter- and intra-assay CVs greater than 10%. Intraclass correlations (ICC) for FSH, SHBG, and DHEAS were high, suggesting that these assays are suitable for population-based studies attempting to link hormone levels to disease risk. The ICC estimated for  $E_1SO_4$  was quite low due to aberrant values reported in a single batch. For the remaining hormones, the ICCs were fair (ranging from 47% for albumin-bound  $E_2$  to 67% for Adione), and studies using these assays would require a substantial increase in the sample size to detect small case-control differences.

### Introduction

Currently, RIAs for steroid hormones are not standardized across laboratories in the United States, and doubts about the accuracy and reproducibility of these assays raise concerns about their utility in epidemiological studies of serum hor-

mones (1-3). Unreliable assay results cause individuals to be misclassified and may falsely lead to null findings. In an effort to monitor ongoing hormone assays of sera from a case-control study of breast cancer, replicate quality control samples were included in each batch. We report herein on the reproducibility of several steroid hormones assayed in a single laboratory.

### Materials and Methods

**Study Methods.** To study the role of serum hormones and breast cancer, we conducted a nested case-control study using sera from a cohort of women who volunteered to donate blood to the Columbia, MO, Serum Bank between 1977 and 1987 and were followed for breast cancer through 1989. The study design and methods have been described previously (4). From this, sera from a sample of disease-free, postmenopausal women who were not taking replacement estrogens at the time of blood draw were identified as controls. Women were considered postmenopausal if they reported natural or surgical menopause, or radiation to the ovaries prior to blood draw and if FSH<sup>2</sup> levels were 35 mIU/ml or higher. Serum was drawn using standard procedures and then chilled and aliquotted into glass vials within 2 h of collection. Vials were shipped on dry ice to the Mayo Foundation repository and maintained at  $-70^\circ\text{C}$  until assayed.

To monitor the consistency of the laboratory assays, three nonfasting postmenopausal women whose reported date of last menses was at least a year earlier and not currently using exogenous estrogens volunteered to donate a unit of whole blood. Serum was separated, carefully mixed, and aliquotted into glass vials and labeled so as to be indistinguishable from control samples. Vials were then stored at  $-70^\circ\text{C}$ .

Aliquots from the three volunteers were randomized among the case-control study samples. For each volunteer, and in each of nine batches, three vials were placed randomly among the breast cancer control samples, with the laboratory being unaware of which sample corresponded to which woman. A total of 27 measurements of each hormone was made for each volunteer.

**Laboratory Methods.** The laboratory used RIA methods to measure circulating hormones as follows:  $E_1$  and Adione were measured using kits obtained from Diagnostic Systems Laboratories (Webster, TX);  $E_2$  and T were measured using kits obtained from Diagnostic Products Co. (Los Angeles, CA); DHEAS was measured using a kit from ICN Biomedical (Costa Mesa, CA). SHBG was measured using an immunoradiometric assay kit (Farnos Group Ltd., Oulunsalo, Finland). Percentages of unbound and albumin-bound  $E_2$  were measured using a

Received 8/23/96; revised 2/7/97; accepted 2/17/97.

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.

<sup>1</sup> To whom requests for reprints should be addressed, at the National Cancer Institute, 6130 Executive Boulevard, Rm. 443, Bethesda, MD 20892-7372. Phone: (301) 435-3982; Fax: (301) 402-0916; E-mail: falkr@epndce.nc.nih.gov.

<sup>2</sup> The abbreviations used are: FSH, follicle-stimulating hormone;  $E_1$ , estrone;  $E_2$ , estradiol; Adione, androstenedione; DHEAS, dehydroepiandrosterone sulfate; T, testosterone; SHBG, sex hormone binding hormone; CV, coefficient of variation; ICC, intraclass correlation.

Table 1 Mean hormone level by subject: Mayo study, 1996

Hormone <sup>a</sup>	Subject	Mean (SD)	Controls (n = 82) Mean (sd)
E <sub>1</sub> (pg/ml)	a	19.6 (6.4)	34.0 (17.9)
	b	31.2 (6.7)	
	c	27.7 (9.8)	
E <sub>2</sub> (pg/ml)	a	6.9 (4.5)	16.8 (14.9)
	b	9.8 (4.8)	
	c	15.3 (6.4)	
T (pg/ml)	a	131.1 (67.4)	212.4 (140.0)
	b	381.1 (103.8)	
	c	133.3 (165.5)	
Adione (pg/ml)	a	620.0 (111.3)	938.3 (420.8)
	b	1186.7 (262.9)	
	c	777.8 (367.4)	
FSH (mIU/ml)	a	73.5 (5.0)	76.6 (32.9)
	b	113.6 (14.2)	
	c	130.2 (14.8)	
SHBG (nmol/ml)	a	51.9 (6.2)	60.4 (32.2)
	b	48.3 (9.6)	
	c	73.9 (14.7)	
Albumin-bound E <sub>2</sub> (%)	a	40.7 (9.2)	46.1 (12.2)
	b	47.4 (10.0)	
	c	42.5 (7.7)	
Free E <sub>2</sub> (%)	a	1.4 (0.2)	1.4 (0.4)
	b	1.4 (0.3)	
	c	1.3 (0.2)	
E <sub>1</sub> SO <sub>4</sub> (pg/ml)	a	145.4 (76.9)	212.3 (122.1)
	b	119.9 (57.2)	
	c	210.0 (111.6)	
DHEAS (μg/ml)	a	0.8 (0.2)	0.9 (0.5)
	b	0.8 (0.3)	
	c	1.1 (0.2)	

<sup>a</sup> Values for E<sub>1</sub>, E<sub>2</sub>, T, Adione, FSH, DHEAS, and E<sub>1</sub>SO<sub>4</sub> are based on a natural log transformation; for percentages of free and albumin-bound E<sub>2</sub>, an arcsine transformation was used.

centrifugal ultrafiltration method as described (5, 6). E<sub>1</sub>SO<sub>4</sub> was measured as E<sub>1</sub> after solvent extraction removed the unconjugated steroids, followed by specific hydrolysis and celite column chromatography as described previously (7). FSH was measured by a microparticle enzyme immunoassay (Abbott Laboratories, Abbott Park, IL). Prior to their use, the kits were tested extensively by comparing column results from established procedures run at the same time as the assay kits.

Assays were done in duplicate for all hormones and repeated if values differed by more than 15%. In addition, assays were repeated if the commercial standards differed by more than 15% from the mean. For each analyte, assays were conducted consecutively over a 2-month period between July and September 1994 and by the same technician.

Assay sensitivities are as follows: E<sub>1</sub>, 1.2 pg/ml; E<sub>2</sub>, 8 pg/ml; T, 0.04 ng/ml; Adione, 0.02 ng/ml; SHBG, 0.5 nmol/L; FSH, 0.2 mIU/ml; DHEAS, 16 ng/ml; and E<sub>1</sub>SO<sub>4</sub>, 50 pg/ml. For E<sub>2</sub> at levels of 3.6 ng/ml, 10% cross-reactivity with E<sub>1</sub> occurs; although not tested, this effect would be expected to be marginal at the low E<sub>2</sub> levels measured in this study. For E<sub>1</sub>, a 1.25% cross-reaction with E<sub>2</sub> was reported by the kit manufacturer. The assay for DHEAS had relatively high cross-reactivities with DHEA and Adione, ranging from 30 to 60%. However, because DHEAS circulates at levels at least 1000 times that of these androgens, little assay interference would be expected. For T, there were no notable cross-reactivities with any of the other analytes measured, and the antibody used in the

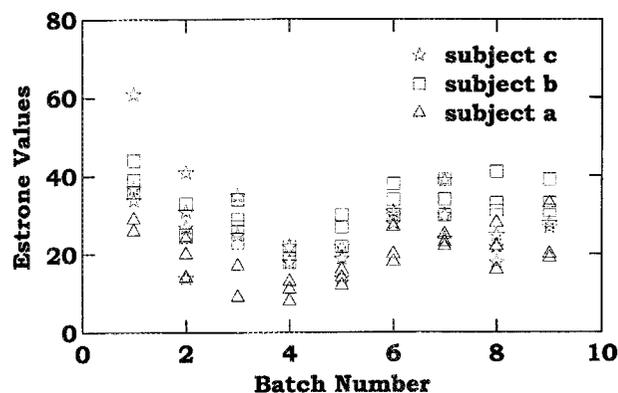


Fig. 1. Serum E<sub>1</sub> values (pg/ml) plotted according to batch, with three measurements for each of the three volunteers in each batch.

SHBG assay has no known reactions with other circulating proteins.

**Statistical Methods.** To test for hormone reproducibility between batches, and to calculate intra- and interassay CV, assay results from the three volunteers were used to calculate variance estimates with a nested ANOVA model. This model adjusts for the correlation that arises from assaying the same sample repeatedly. To satisfy the assumptions of the ANOVA model, hormones were transformed to the natural logarithm, except for percentage of free E<sub>2</sub> and percentage of albumin-bound E<sub>2</sub>; for these measures an arcsine transformation was used. These transformations normalize the data and remove the dependence of the variance on the mean. Between-person variance was estimated using the variance calculated from assays on 82 serum bank controls minus the batch-to-batch and within-batch variance estimates from the ANOVA model. The ICC was calculated by comparing the between-person variance to the total variance [*i.e.*, ICC = (between-person variance)/(between-person + batch-to-batch + within-batch laboratory variance)]. Statistical analyses were conducted using SYSTAT (8).

## Results

Hormone assay results are shown in Table 1. For each of the three volunteers, we computed the mean and SD of the 27 measurements (three samples assayed in each of nine batches). Statistics for the serum bank controls are also presented. Compared to the serum bank controls, the quality control subjects had lower values of E<sub>1</sub>, E<sub>2</sub>, and E<sub>1</sub>SO<sub>4</sub>, but values of the other analytes were similar.

Hormone values plotted over time showed marked variability, and the nested ANOVA results confirmed that significant batch effects existed for all of the assays studied. For descriptive purposes, the plot of E<sub>1</sub> values according to batch and subject is shown in Fig. 1. Although the measurements for each subject are fairly consistent within a batch, the drift from one batch to the next is notable and of similar magnitude for each of the three volunteers, suggesting problems with the internal standard.

Inter- and intra-assay CVs are presented on Table 2, along with overall CVs and ICCs. The intra-assay CV is an average of each subject's within-batch variability, and the inter-assay CV averages the batch-to-batch variability across subjects. For the overall CV, we used the square root of the sum of the squares of the inter- and intra-assay CV. Overall CVs tended to

Table 2 Reproducibility of serum hormones: Mayo study, 1996

Hormone <sup>a</sup>	Overall	CVs		ICC
		Interassay	Intra-assay	
E <sub>1</sub>	9.2	7.0	6.1	56.6
E <sub>2</sub>	20.2	12.2	16.1	63.9
T	8.9	6.2	6.4	58.1
Adione	4.1	2.0	3.5	66.9
FSH	2.6	1.7	2.0	95.5
SHBG	4.3	3.0	3.1	90.9
E <sub>1</sub> SO <sub>4</sub>	9.7	7.8	5.8	12.6
DHEAS	1.2	1.0	0.7	84.3
Free E <sub>2</sub> (%)	16.8	11.1	12.6	64.5
Albumin-bound E <sub>2</sub> (%)	19.2	8.9	17.1	47.3

<sup>a</sup> Values for E<sub>1</sub>, E<sub>2</sub>, T, Adione, FSH, DHEAS, and E<sub>1</sub>SO<sub>4</sub> are based on a natural log transformation; for percentages of free and albumin-bound E<sub>2</sub>, an arcsine transformation was used.

be high for E<sub>2</sub>, percentage of free E<sub>2</sub>, and percentage of albumin-bound E<sub>2</sub>, but of these, only E<sub>2</sub> and percentage of free E<sub>2</sub> had both inter- and intra-assay CVs greater than 10%.

The ICCs for FSH, SHBG, and DHEAS were greater than 80% (Table 2), indicating that variability due to hormonal differences between postmenopausal women is much larger than that due to the assays. Except for E<sub>1</sub>SO<sub>4</sub>, correlation coefficients for the remaining analytes were fair, ranging from 47 to 67%. The poor performance of E<sub>1</sub>SO<sub>4</sub> (ICC of 13%) was caused by extremely low values of this hormone in one volunteer, which occurred in a single batch. Had we ignored that batch in the analysis, the coefficient would have approached 70%.

For epidemiological purposes, assay problems can be overcome by increasing the size of the study using the intraclass correlation as an inflation factor; provided the coefficient is large, little power will be lost if no adjustment is made. As an example, in the absence of assay variability, a study of estrogens and breast cancer in postmenopausal women would require 175 cases and controls to detect a 10% case-control difference in E<sub>2</sub> with 80% power. On the basis of our ICC estimate of 64%, we would need to recruit 273 cases and controls to achieve this degree of statistical power; had the correlation been as high as 80%, only 219 would be necessary.

## Discussion

Assays for steroid hormones, especially E<sub>1</sub> and E<sub>2</sub>, are problematic because their concentrations in postmenopausal women are quite low (measured in pg/ml). In fact, levels may be very near the limit of sensitivity, at which the assays may be too imprecise to detect small hormone differences that may be biologically relevant to cancer. We found measurements of protein-bound and free E<sub>2</sub> were not consistent based on the CVs, but their ICCs suggest that assay precision may be adequate for epidemiological studies of these analytes, provided the study is of sufficient size. That conclusions based on the CV are at odds with those based on the ICC is not surprising, because the latter provides the percentage of the variability in a measurement due to biological differences between women, whereas the CV measures variability as a percentage of the sample mean, without regard to differences between women. When the intraclass correlation is high, biological differences between postmenopausal women far outweigh the assay variability, and sufficient statistical power can be obtained by increasing the size of the study sample using the coefficient as a sample weight. Of the analytes studied, FSH, SHBG, and

DHEAS measurements are reliable, and little will be gained from a larger case-control group. With the exception of E<sub>1</sub>SO<sub>4</sub>, the assays for the remaining hormones are only precise enough to reliably compare large groups when small differences are sought. E<sub>1</sub>SO<sub>4</sub> reproducibility was poor, with the ICC being very sensitive to problems with a single batch. With appropriate quality control efforts, however, such a flawed batch can be easily identified and corrected.

Our results are consistent with earlier findings of problems with assay reproducibility for some steroid sex hormones (1-3). Measurements varied considerably from one batch to the next for all of the analytes, which was not due to changes in the technicians performing the assay, methods of analysis, or antibodies used. This finding recommends that sera from matched cases and controls be analyzed in the same batch. For unmatched study designs, the lack of reliable measurements from one batch to another is a concern, and at a minimum, statistical adjustment should be considered. Quality control samples placed in each batch will uncover drift in the assays and help identify flawed batches for which repeated assaying may be required.

Using assay results from the serum bank controls allowed us to estimate the between-subject component of variance more precisely than had we used only the sera from the three volunteers for this purpose. On average, hormone values among the volunteers were similar to the serum bank controls for all of the analytes, making it unlikely that this introduced systematic error into the calculation of the ICC. We note that the ICC is influenced by all sources of variance in hormone measurements, some of which were not measured in this study. Most important, biological hormone fluctuations were not assessed. Although some have attempted to separate out the contributions of laboratory error and biological fluctuations to hormone variability in populations, not all of the relevant components of variance were considered (9, 10). To do so, multiple samples over time with replicate aliquots from a large number of randomly selected postmenopausal women would be required; such a study is methodologically complex and costly.

This study supports the growing body of evidence (1-3) pointing to the need for a national program for standardization of steroid hormone assays much like that developed for cholesterol measurement (11). An external quality control scheme would provide a basis with which to judge assay reproducibility and to stimulate the production of reliable commercial assay

kits. Until such time, the study of case-control differences in large numbers of postmenopausal women may be feasible only when the biological variability of the analyte is high. Care must be taken to evaluate assay methods prior to the analysis of the serum hormones and to establish a rigorous quality control scheme to monitor ongoing analyses.

### References

1. Cauley, J. A., Gutai, J. P., Kuller, L. H., and Powell, J. G. Reliability and interrelations among sex hormones in postmenopausal women. *Am. J. Epidemiol.*, *133*: 50-57, 1991.
2. 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., Biomarkers & Prev.*, *3*: 51-56, 1994.
3. 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.
4. Dorgan, J. F., Longcope, C., Stephenson, H. E., Jr., Falk, R. T., Miller, R., Francz, C., Kahle, L., Campbell, W. S., Tangrea, J. A., and Schatzkin, A. Relationship of prediagnostic serum estrogen and androgen levels to breast cancer risk. *Cancer Epidemiol., Biomarkers & Prev.*, *5*: 533-539, 1996.
5. Hammond, G. L., Nisker, J. A., Jones, L. A., and Siiteri, P. K. Estimation of the percentage of free steroid in undiluted serum by centrifugal ultrafiltration-dialysis. *J. Biol. Chem.*, *224*: 5023-5026, 1980.
6. Longcope, C., Hui, S. L., and Johnston, C. C. Free estradiol, free testosterone and sex hormone binding globulin in peri-menopausal women. *J. Clin. Endocrinol. Metab.*, *64*: 513-519, 1987.
7. Francz, C., Watson, D., and Longcope, C. Estrone sulfate and dehydroepiandrosterone sulfate concentrations in normal subjects and men with cirrhosis. *Steroids*, *34*: 563-573, 1979.
8. Wilkinson, L. SYSTAT: The System for Statistics. Evanston, IL: SYSTAT, Inc., 1986.
9. 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., Biomarkers & Prev.*, *3*: 47-50, 1994.
10. 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., Biomarkers & Prev.*, *4*: 649-654, 1995.
11. Schaffer, R., Bowers, G. N., Jr., and Melville, R. S. History of NIST's contributions to the development of standard reference materials and reference and definitive methods for clinical chemistry. *Clin. Chem.*, *41*: 1306-1312, 1995.