

Reliability and Validity of Serum Sex Hormone Measurements

Lisa M. McShane,^{1, 2} Joanne F. Dorgan, Susan Greenhut, and James J. Damato

Biometric Research Branch, Division of Cancer Treatment, Diagnosis, and Centers [L. M. M.] and Cancer Prevention Studies Branch, Division of Cancer Prevention and Control [J. F. D.], National Cancer Institute, Bethesda, Maryland 20892, and McKesson BioServices, Rockville, Maryland 20850 [S. G., J. J. D.]

Abstract

The laboratory reliability and validity of sex hormone measurements were examined at multiple levels, including lower levels characteristic of children and postmenopausal women. Serum was drawn from four adult male and four adult female healthy volunteers. From each individual's serum pool, a medium- and a low-dilution pool were created. Biochemical analyses for total and non-sex hormone-binding globulin (SHBG)-bound estradiol, estrone, estrone sulfate, progesterone, and SHBG were performed on female samples. Male samples were analyzed for total and non-SHBG-bound testosterone, dihydrotestosterone, androstenedione, and dehydroepiandrosterone sulfate. Two aliquots from each pool were assayed twice in each of two labs. All assays except SHBG in one lab used RIA procedures. Reliability was assessed by variance components analyses and estimated coefficients of variation (CVs). Validity was assessed by comparing observed measurements versus expected values based on known dilution ratios.

For the testosterone and dihydrotestosterone assays, CVs were usually less than 10%. For estradiol and progesterone, CVs were usually less than 15%. Assays with larger estimated CVs included androstenedione, dehydroepiandrosterone sulfate, estrone, and estrone sulfate. Absolute levels differed markedly between labs for most assays. Observed measurements generally agreed with values expected from the dilution ratios. A notable exception was the estrone assay at the lowest dilution level, where observed measurements were 2-4 times those expected. A similar but less pronounced overestimation bias for the low levels of estradiol was also suggested. This intra- and interlaboratory variability and apparent low dilution overestimation should be accounted for in studies relating hormones to cancer risk, especially those involving children and postmenopausal women.

Introduction

Several recent studies of laboratory reproducibility of steroid hormone and SHBG³ assays have reported substantial variability both within and between laboratories (1-3). Such studies, as well as anecdotal evidence of wildly inaccurate hormone measurements (4), have brought into question the use of these hormones (4) measurements in epidemiological studies. In at least one case, the results from an entire study were challenged on the basis of the validity of the laboratory measurements (5-7). Most laboratory reproducibility studies for steroid sex hormones and related compounds have looked only at the higher levels typically found in adults, with the exception of some hormones found at lower concentrations in postmenopausal women.

There is increasing speculation among researchers that cancer risk could be associated with exposures earlier in life than previously thought. The DISC Hormone Ancillary Study is being conducted by the National Cancer Institute to investigate the relationship between childhood and adolescent diet and serum levels of sex hormones that may be associated with risk of cancer, particularly breast and prostate cancer, in later life. The parent study, DISC (8), is a multicenter randomized clinical trial being conducted by the National Heart, Lung, and Blood Institute to evaluate the efficacy of a fat-modified diet to reduce low density lipoprotein-cholesterol and the safety of this diet in children. Before conducting the DISC Hormone Ancillary Study, a pilot study was conducted to evaluate the reliability and validity of estrogen and androgen assays at two laboratories. We wanted to verify that the laboratories could reliably measure all of the hormones of interest to us at levels found over the age range (8-18 years) of male and female participants in the DISC Hormone Ancillary Study and maintain the relative ratios (a measure of internal validity) between high and low levels.

Materials and Methods

Each of four adult male (20-35 years old) and four adult female (20-37 years old) healthy volunteers donated up to 500 ml of plasma on a single occasion between 8 and 11 a.m. after a fast of at least 12 h. All plasma was drawn in November 1994. The women all had regular menstrual cycles and were not pregnant, lactating, or taking oral contraceptives; the plasma was drawn during the luteal phase (day 20-24 after the start of last menses). Plasma was defibrinated from each of the eight samples to create eight individual serum pools, which we will refer to as the high-level pools. To simulate the range of hormone levels in children, each of the original eight pools was mixed with charcoal-stripped serum that did not contain any steroid hormones to create two dilutions. For the female serum, medium

Received 4/8/96; revised 8/20/96; accepted 8/26/96.

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 research was conducted while L. M. M. was a Senior Staff Fellow in the Biometry Branch in the Division of Cancer Prevention and Control at the National Cancer Institute.

² To whom requests for reprints should be addressed, at National Cancer Institute, Biometric Research Branch, CTEP, DCTDC, Executive Plaza North, Room 739, 6130 Executive Boulevard MSC 7434, Bethesda, MD 20892-7434.

³ The abbreviations used are: SHBG, sex hormone-binding globulin; CV, coefficient of variation; DHEAS, dehydroepiandrosterone sulfate; DISC, Dietary Intervention Study in Children; O/E, observed/expected; GC-MS, gas chromatography-mass spectrometry.

dilutions were prepared using a dilution factor of 1:4, and low dilutions were prepared using a dilution factor of 1:12. The corresponding dilution factors for the male serum were 1:3 for the medium dilutions and 1:10 for the low dilutions. After dilution, the hormone levels covered the normal ranges generally expected for males and females 8–18 years of age. These normal ranges were those reported to us by the laboratories, and they may have differed somewhat between the two laboratories. Each of the resulting 24 pools (2 sexes \times 4 individuals/sex \times 3 dilutions/individual) was aliquoted into glass vials and stored at -70°C until the time of analysis. All samples were assayed in the period from December 1994 through January 1995.

Both laboratories used RIA techniques after extraction and chromatography to measure estradiol, estrone, testosterone, and dihydrotestosterone. Lab A extracted the samples with hexane: ethyl acetate and used LH20 microcolumn chromatography for estradiol and estrone and Al_2O_3 micro column chromatography for testosterone and dihydrotestosterone. Lab B used organic extraction and celite chromatography for all four hormones. Each lab measured estrone sulfate as estrone after enzymolysis of the estrone sulfate followed by their standard techniques for measuring estrone. Lab A measured both progesterone and androstenedione using RIA after hexane-ethyl acetate extraction and centrifugation. Lab B measured progesterone by RIA preceded by organic extraction and androstenedione by RIA after organic extraction and celite chromatography. Lab A measured DHEAS as dehydroepiandrosterone by RIA after enzymolysis of the DHEAS. Lab B measured DHEAS directly by RIA after serial dilution to eliminate most of the cross-reacting compounds. SHBG was measured in lab A by its binding capacity, using a displacement technique. Lab B measured SHBG using the Diagnostic Systems Laboratory SHBG RIA kit. Lab A reported SHBG in units (micrograms/deciliter) that we converted to nanomoles/liter by multiplying by 34.67. In both labs, the percentage of non-SHBG-bound estradiol and non-SHBG-bound testosterone was measured by ammonium sulfate precipitation, and absolute amounts were calculated as those percentages times the measured values of total estradiol and testosterone, respectively.

Because the laboratory will be aware of the sex and age of the participants during the conduct of the Hormone Ancillary Study, samples sent to the labs for this pilot study were identified with sex and age range. For females, low samples were identified as 9–11 years, medium samples were identified as 12–16 years, and high samples were identified as 18 years and over. For males, low samples were identified as 9–12 years, medium samples were identified as 13–16 years, and high samples were identified as 18 years and over. No further information was provided regarding the particular individual from whom the sample came.

At the time of this pilot study, we planned to measure androgens only in serum from boys in the Hormone Ancillary Study and estrogens only in serum from girls. Therefore, testosterone (total and non-SHBG-bound), dihydrotestosterone, androstenedione, and DHEAS were measured only in male samples in the pilot study, and estradiol (total and non-SHBG-bound), estrone, estrone sulfate, and progesterone were measured only in female samples. Additionally, for the youngest boys in the Hormone Ancillary Study, we planned to measure only testosterone and SHBG, and for the youngest girls, we planned to measure only estradiol, estrone, and SHBG. In the pilot study, only testosterone was measured in "low" male samples, and only estradiol and estrone were measured in "low" female samples. Because SHBG is not stripped from serum by charcoal, SHBG was measured only in "high" female samples,

Table 1 Study design for testing 2 samples from each of 24 pools^a

Subject	Day 1				Day 2			
	A	L	M		L	M	H	H
B	L	M	H	H	L	M		
C	L	L		H		M	M	H
D		M	M	H	L	L		H
E	L	L		H		M	M	H
F		M		H	L	L	M	H
G	L	M	M		L			H
H	L	M		H	H	L	M	

^aL, low dilution; M, medium dilution; H, high (undiluted).

and non-SHBG-bound estradiol and testosterone are reported only in "high" female and male samples, respectively. In each laboratory, two 2.5-ml aliquots from each of the low and medium pools and two 5-ml aliquots from each of the high pools were measured, either twice in the same batch (same as day in this study) or once on each of two separate days, according to the design displayed in Table 1.

To quantify the laboratory variability, the hormone measurements were modeled by a variance components model. The model contained, in addition to an overall mean, random effects for subject, laboratory assay batch (same as day in this study), and within-assay error. Each random effect was assumed to be normally distributed with mean zero and some unknown variance. The variances of these random effects are termed variance components, and they were estimated using restricted maximum likelihood techniques (9) as implemented by SAS PROC MIXED (10). The batch variance component is commonly referred to as the interassay variance, and the within-assay error variance is referred to as the intraassay variance. For each dilution level, the interassay and intraassay CVs are computed separately as the square roots of the interassay and intraassay variances, respectively, divided by the mean hormone levels. The total laboratory variance is computed as the sum of the inter- and intraassay variances, and the total laboratory CV is the square root of the total variance divided by the mean hormone level. By dividing by group means in calculating our CV measures, we ignored subject-to-subject variability, and the actual CV may be higher for subjects with hormone levels at the lower end of the group range and lower for subjects at the higher end. CVs were computed on both the original data scale and the natural log-transformed scale. The log-transformed scale was more appropriate for those hormone measurements that exhibited variance increasing with mean level. Other researchers have suggested that for many steroid hormones, log-transforming the data values removes the dependence of the variance on the mean and makes the values more closely normally (Gaussian) distributed (1, 3, 11–12).

Ideally, we would like to compare the laboratory measurements obtained by RIA to measurements obtained using a gold standard procedure to assess validity, thus giving an estimate of bias in the RIA procedure. However, no true gold standard technique exists. GC-MS techniques have been recently developed and are considered the best approximation to a gold standard for most of these hormone assays, but GC-MS was unavailable to us at the time these samples were assayed. We made some assessment of the internal validity of the RIA procedures by comparing laboratory measurements obtained from undiluted pools with measurements from known dilutions of those same pools. We quantified the degree to which the RIA measurement process preserved dilution ratios by computing O/E ratios from the RIA measurements. The O/E ratios for each

Table 2 Mean^a hormone readings (O/E ratio) by subject and dilution for male samples

Hormone	Subject	Lab A			Lab B		
		High	Medium	Low	High	Medium	Low
Testosterone (total) (ng/dl)	A	522.5	187.5 (1.1)	57.5 (1.1)	446.5	185.0 (1.2)	57.0 (1.3)
	B	551.0	178.0 (1.0)	57.0 (1.0)	564.0	170.5 (0.9)	63.5 (1.1)
	C	447.0	160.0 (1.1)	53.0 (1.2)	396.5	149.0 (1.1)	50.5 (1.3)
	D	307.0	103.0 (1.0)	30.0 (1.0)	264.5	106.0 (1.2)	30.0 (1.1)
	Mean	456.9	157.1 (1.0)	49.4 (1.1)	417.9	152.6 (1.1)	50.3 (1.2)
Testosterone (non-SHBG-bound) (ng/dl)	A	324.0			195.3		
	B	325.0			197.6		
	C	252.5			148.7		
	D	119.5			76.7		
	Mean	255.3			154.6		
Dihydrotestosterone (ng/dl)	A	67.5	20.5 (0.9)		53.0	17.0 (1.0)	
	B	39.0	13.5 (1.0)		39.5	11.0 (0.8)	
	C	32.0	8.9 (0.8)		31.5	10.0 (1.0)	
	D	38.0	15.0 (1.2)		37.5	10.5 (0.8)	
	Mean	44.1	14.5 (1.0)		40.4	12.1 (0.9)	
Androstenedione (ng/dl)	A	122.0	37.5 (0.9)		84.0	27.5 (1.0)	
	B	125.0	33.0 (0.8)		72.5	30.5 (1.3)	
	C	111.5	40.0 (1.1)		84.0	31.0 (1.1)	
	D	102.0	41.5 (1.2)		73.5	25.5 (1.0)	
	Mean	115.1	38.0 (1.0)		78.5	28.6 (1.1)	
DHEAS (μ g/dl)	A	186.0	59.0 (1.0)		316.0	105.0 (1.0)	
	B	220.5	74.0 (1.0)		306.5	98.5 (1.0)	
	C	161.0	55.0 (1.0)		253.0	74.5 (0.9)	
	D	511.5	190.0 (1.1)		927.5	276.5 (0.9)	
	Mean	269.8	94.5 (1.0)		450.8	138.6 (0.9)	

^aThe individual subject means and O/E ratios are computed from the two replicates/subject.

dilution were defined as the mean RIA reading obtained for the dilution multiplied by the dilution factor and divided by the mean measurement obtained for the undiluted pool. An O/E ratio of 1.0 indicates perfect preservation of the dilution ratio.

Before conducting this study, each laboratory reported their estimates of the intra- and interassay CVs to us. Lab A reported most of its intraassay CVs to be in the 3–9% range and most of its interassay CVs to be in the 6–14% range, depending on the assay and true level of the analyte being measured. In each assay run, lab A included known standards, multiple-level control pools, and randomly repeated samples. Standards were run at the beginning of each assay. The control pool results were compared against two SD control charts to detect aberrant assays. Lab B reported estimated intra- and interassay CVs similar to those from lab A, with the exception of the progesterone, androstenedione, and DHEAS assays in which lab B anticipated the CVs to be nearly double those reported by lab A. For example, for the progesterone assay, lab A estimated its intraassay CV to be in the 5.5–8.3% range and its interassay CV to be in the 2–5.8% range. Lab B's estimates were 9.3–12.5% for the intraassay CV and 9.6–17.9% for the interassay CV. Lab B's quality control procedures included running two SD control charts on control pool samples and running standard curves at the beginning and end of each assay. Both labs performed the RIA portion of their RIA-based assays in duplicate and reported the mean of those two values; however, if the duplicates differed by more than 15%, then the entire assay procedure was repeated for that sample.

Results

Subject-specific and group mean hormone values for the male subject samples are presented in Table 2. The two labs were in general agreement with regard to total testosterone measurements. The range of values for the high pools was similar, the

group means differed by less than 10%, and no pair of individual means differed by more than 17%. The agreement was even better for the low and medium pools. The degree of agreement for dihydrotestosterone values was similar to that for total testosterone. Larger differences were evident in the remaining three androgen assays. Non-SHBG-bound testosterone and androstenedione readings from lab A averaged 50–60% higher than those from lab B. For DHEAS, lab B values averaged over 60% higher than those from lab A. In all of the assays except for androstenedione, however, the two labs ranked the four subjects' levels in the same or nearly the same order (reversing only one pair of consecutively ordered values), despite their different absolute levels. The O/E ratios indicated no strong evidence against internal validity. It should be noted that because the subject-specific values in this table are actually the means of measurements made on two different aliquots, the variability inherent in single measurements is somewhat greater than the variability demonstrated in this table.

Table 3 presents the subject-specific and group mean hormone values for the female subjects. For non-SHBG-bound estradiol and estrone sulfate, lab A's measurements were higher than lab B's for all subjects, averaging 56% higher for estradiol and 64% higher for estrone sulfate. Progesterone values were higher in lab B than in lab A, with the mean 23% higher. For total estradiol, lab B reported higher measurements than lab A for all samples at the medium and low dilutions with the higher degree of difference at the low dilution, in which the group mean for lab B was 32% higher than in lab A. The relative rankings of the four subjects' levels also differed in all of the assays except SHBG. One of the most striking features of Table 3 is the O/E ratios for the estrone assay. In both labs, the ratios for the low values are higher than 1.8 for all subjects, with a maximum value of 4.2 for subject E. Errors in the dilution level could be ruled out because the same dilution pool was used for

Table 3 Mean^a hormone readings (O/E ratio) by subject and dilution for female

Hormone	Subject	Lab A			Lab B		
		High	Medium	Low	High	Medium	Low
Estradiol (total) (ng/dl)	E	10.5	2.7 (1.0)	0.9 (1.0)	9.5	3.6 (1.5)	1.3 (1.6)
	F	4.4	1.2 (1.1)	0.5 (1.2)	4.9	1.4 (1.1)	0.6 (1.4)
	G	8.2	2.3 (1.1)	0.7 (1.0)	13.0	2.5 (0.8)	1.1 (1.0)
	H	13.5	3.0 (0.9)	1.0 (0.8)	12.5	3.2 (1.0)	1.4 (1.3)
	Mean	9.1	2.3 (1.0)	0.8 (1.0)	10.0	2.6 (1.0)	1.1 (1.3)
Estradiol (non-SHBG-bound) (ng/dl)	E	4.3			1.9		
	F	3.7			2.7		
	G	4.9			4.6		
	H	6.6			3.4		
	Mean	4.9			3.1		
Estrone (ng/dl)	E	6.2	2.5 (1.6)	1.7 (3.3)	7.5	2.0 (1.1)	2.6 (4.2)
	F	4.2	1.9 (1.8)	1.3 (3.6)	4.1	1.9 (1.9)	0.7 (2.1)
	G	6.6	1.5 (0.9)	1.4 (2.5)	5.3	4.1 (3.1)	1.4 (3.2)
	H	6.2	2.3 (1.5)	1.6 (3.0)	7.7	2.7 (1.4)	1.2 (1.8)
	Mean	5.8	2.0 (1.4)	1.5 (3.1)	6.1	2.7 (1.9)	1.5 (2.8)
Estrone sulfate (ng/dl)	E	210.0	58.5 (1.1)		125.4	34.2 (1.1)	
	F	240.0	54.0 (0.9)		137.9	35.0 (1.0)	
	G	299.0	61.5 (0.8)		184.9	43.9 (0.9)	
	H	218.0	68.0 (1.2)		140.0	36.1 (1.0)	
	Mean	241.8	60.5 (1.0)		147.0	37.3 (1.0)	
Progesterone (ng/dl)	E	889.0	197.0 (0.9)		1034.0	254.0 (1.0)	
	F	241.0	70.0 (1.2)		309.0	76.0 (1.0)	
	G	431.5	137.0 (1.3)		672.5	156.5 (0.9)	
	H	553.5	141.5 (1.0)		577.5	145.0 (1.0)	
	Mean	528.8	136.4 (1.1)		648.3	157.9 (1.0)	
SHBG (nmol/L)	E	90.1			65.5		
	F	15.6			7.5		
	G	24.3			25.0		
	H	72.8			50.5		
	Mean	50.7			37.1		

^a The individual subject means and O/E ratios are computed from the two replicates/subject.

all of the assays, and none of the other assays showed O/E ratios markedly higher than 1.0. We considered the possibility that the charcoal-stripped serum used for the dilution was somehow contaminated with estrone, but when the remaining pool of stripped serum was analyzed for the presence of estrone, none was found. The low dilution levels were at or below the limits of sensitivity reported by the two labs, and it is conceivable that the assays failed at those low levels due to cross-reactivity with other serum components or contaminants (7). The low dilution measurements of total estradiol in lab B also gave a hint of a possible problem, although the pattern of $O/E > 1$ is less convincing than in the case of estrone, and all of the measurements seem to be above lab B's lower limit of sensitivity (reported as 0.2 ng/dl for estradiol). As was the case for Table 2, the subject-specific values in this table are the means of measurements made on two different aliquots, and the variability inherent in single measurements is somewhat greater than the variability demonstrated in this table.

Estimated coefficients of variation are presented in Table 4 for the male samples and Table 5 for the female samples. The total CV and intraassay CVs are reported. It should be noted that due to the small sample sizes, none of these CV estimates are very precise. The interassay CVs were estimated particularly imprecisely due to the fact that the assays were run on only two different days. For this reason, we do not present them. In some cases, this small number of assays (days) resulted in finding no statistically measurable interassay variability, and the intraassay CV was estimated to be equal to the total CV. This is likely an artifact of the small sample sizes rather than a true absence of variability between assays. The total CVs pre-

sented here are consistent with the range of values reported by others (1-3). The total CVs in the two labs ranged from 2.4-26.5% for the androgen assays (Table 4). Excluding the estrone assay, which we are considering a failed assay at least at the lower levels, the estimated total CVs for the assays performed on the female samples ranged from 3.1-33.2% (Table 5). For the estrone assay, lab A identified three of the eight measurements made at the low dilution (one sample each from subject F, G, and H) as resulting from "poor duplicates," and lab B flagged those values as falling below their limit of sensitivity for estrone. Ordinarily, the labs would have reanalyzed those samples if they had been provided with adequate sample volumes, so the CVs for estrone should be interpreted with that in mind. In this pilot study, we supplied only the sample volumes that would be available during the actual study, so these unacceptable data values would have become missing data. In Tables 4 and 5, we also present CV estimates based on the natural log-transformed data. This was motivated by our observation that for many of the assays, the intraassay and total variances for the untransformed data tended to increase with the mean hormone levels, whereas the CVs remained relatively constant across dilution levels. Typically, the CVs on the log scale are much lower, with the exception of the estrone and total estradiol assays at the low and medium dilutions. This is a result of the behavior of the CV for data with small mean.

Discussion

This study adds more weight to the body of evidence suggesting that laboratory variability in some steroid sex hormone meas-

Table 4 Intraassay and total CVs (%) by laboratory on original and log-transformed scales for hormones measured on male samples

Hormone		Original Scale		Log Scale		
		Lab A	Lab B	Lab A	Lab B	
		Testosterone (total)	Low	Intraassay	6.0	7.4
		Total	7.2	7.4	1.6	1.8
	Medium	Intraassay	6.6	5.2	1.2	1.2
		Total	6.6	5.2	1.2	1.2
	High	Intraassay	5.3	16.5	0.7	2.8
		Total	5.8	16.5	1.0	2.8
Testosterone (non-SHBG-bound)	High	Intraassay	6.3	18.6	0.8	3.2
		Total	8.9	26.5	2.0	7.7
Dihydrotestosterone	Medium	Intraassay	5.9	5.9	2.2	2.7
		Total	6.1	7.0	2.2	3.1
	High	Intraassay	2.2	4.6	0.8	1.3
		Total	2.4	4.6	0.8	1.3
Androstenedione	Medium	Intraassay	19.7	4.1	6.2	1.2
		Total	22.0	4.1	7.1	1.2
	High	Intraassay	7.4	10.6	1.6	2.6
		Total	7.4	10.6	1.6	2.6
DHEAS	Medium	Intraassay	3.7	7.2	0.8	2.0
		Total	3.7	7.2	1.0	2.0
	High	Intraassay	11.6	2.9	0.8	1.1
		Total	20.3	8.5	2.4	1.8

Table 5 Intraassay and total CVs (%) by laboratory on original and log-transformed scales for hormones measured on female samples

Hormone		Original Scale		Log Scale		
		Lab A	Lab B	Lab A	Lab B	
		Estradiol (total)	Low	Intraassay	6.5	11.1
		Total	15.8	11.1	53.0	27.0
	Medium	Intraassay	6.6	7.5	6.8	7.1
		Total	7.5	7.5	7.7	7.1
	High	Intraassay	5.5	12.9	2.0	4.6
		Total	5.5	12.9	2.0	4.6
Estradiol (non-SHBG-bound)	High	Intraassay	9.4	21.7	2.6	4.7
		Total	11.1	21.7	3.1	4.7
Estrone	Low	Intraassay	7.6	15.8	23.2	74.8
		Total	20.3	62.6	55.9	369.3
	Medium	Intraassay	5.8	8.3	8.1	11.1
		Total	5.8	37.6	8.1	51.2
	High	Intraassay	9.3	13.6	4.8	6.3
		Total	11.2	29.1	6.1	16.8
Estrone sulfate	Medium	Intraassay	14.9	11.3	3.6	3.2
		Total	14.9	11.3	3.6	3.2
	High	Intraassay	14.5	4.9	2.7	1.1
		Total	33.2	5.0	6.5	1.4
Progesterone	Medium	Intraassay	4.6	12.2	0.9	2.2
		Total	4.6	12.2	1.2	2.2
	High	Intraassay	3.1	8.1	0.6	1.7
		Total	3.1	8.1	0.6	1.8
SHBG	High	Intraassay	5.9	5.9	0.5	2.0
		Total	13.0	5.9	3.7	2.0

measurements and related compounds such as SHBG may be substantial between and/or within labs. The sample sizes in our study were small and limited our ability to perform formal statistical inference on the magnitude of the between- and within-laboratory variability, but the data suggested the possibility of substantial variation in absolute levels reported by different labs for several of the assays. The within-laboratory variability that we observed in our data set, if viewed in terms of the CV, was satisfactory for most of the assays. For the testosterone and dihydrotestosterone assays, the within-laboratory variability as measured by total CV was less than 10%, and the O/E ratios for all of the androgens were generally close to 1.0. On the log scale, the CVs were typically less than 5%. The estrogen and progesterone assays generally exhibited greater variability than the androgen assays, although CVs were still frequently below 15%. Our estimated CVs for the estrone, estrone sulfate, androstenedione, and DHEAS assays were rather large and warrant further investigation to determine if these were just chance occurrences in our small data set or represent true large variances associated with these assays. One notable problem with the internal validity was the substantial overestimation of estrone concentration at low levels.

When samples are analyzed for the DISC Hormone Ancillary Study, quality control pool samples will be run in each assay batch and may help to identify aberrant batches that could be reanalyzed, so the within-lab variability may be further reduced in practice. Continued monitoring of laboratory performance is clearly advisable. It should also be kept in mind that the two laboratories used in our pilot study were carefully selected on the basis of extensive written proposals and their many years of experience in analyzing steroid hormones in pediatric serum samples. Two randomly chosen labs may not perform as well.

The impact of the variability in these assays on epidemiological studies must be carefully considered. The between-laboratory variability is problematic when one wishes to compare results between studies or even within the same study when more than one laboratory is used, and laboratory differ-

ences hamper the ability to translate research study findings into public health recommendations and screening guidelines. The within-laboratory variability has many implications for study design and analysis. If the hormone values are used as outcome variables in a study, the increased variability means increased sample sizes are required to attain desired power levels. If hormone values are used as explanatory variables in a study, one must be concerned with the effect of the measurement error on inferences regarding observed associations, for example, possible attenuation of relative risk estimates and reduced power.

Our data also suggested problems with measuring estrone and possibly estradiol at low levels. The results from our dilution pools, however, must be interpreted with some caution because our low levels were created artificially, and the levels of the other background serum components may not be fully representative of those in children's serum samples with naturally occurring low levels of the hormones of interest. We used the dilution method because, for a feasibility study such as this, we could ethically draw serum only from adults and not from children. Also, without a true gold standard (even GC-MS may not be more reliable at low levels), direct assessment of bias was impossible. Our results are consistent with the suggestions of others (7, 13) that laboratory methods for estrone and estradiol using RIA may be highly unreliable or greatly overestimate the typically low values found in children and postmenopausal women (normal ranges reported by our labs for postmenopausal women overlapped with the levels in our "medium-low" female dilutions). Although both of our labs did flag several of the low estrone measurements as being unreliable they could not provide information on the direction or magnitude of potential bias without access to the dilution information. In practice, an investigator would have no dilution data and would have to choose between omitting these data values completely

or using them despite their potential unreliability. Either alternative could bias study results.

Investigations to characterize serum hormone measures now need to proceed beyond studies of laboratory variability alone. As important as laboratory variability is, statistical power and potential bias in relative risk estimates depend on much more than laboratory variability. Factors that are equally important in determining power and degree of bias in relative risks are the distribution of true long-term average subject values in the population, the temporal fluctuation of each subject's level about their true long-term average values, and the magnitude of difference in level associated with different levels of risk. A few studies have examined temporal fluctuation in hormone measurements (3, 12-14), and not all of them separate the contribution of laboratory error from temporal (biological) fluctuation. For studies involving menstruating women, temporal fluctuations have both systematic (monthly cycles) and random components. In our study, all women's plasma samples were drawn at the same time in the menstrual cycle (day 20-24), so we controlled for the systematic component. When we refer to distribution of true subject values, we need to know the functional form of the distribution (*e.g.*, Gaussian, log-Gaussian, etc.) as well as its parameters such as the mean and variance. For the temporal fluctuation and laboratory variability, both of which can be considered measurement error with respect to a subject's true long-term average hormone level, we also need to know the form of the distributions and particularly whether the degree of variability depends on the subject's true underlying level. It is only then that we can put into perspective the relative importance of the laboratory measurement error. It is not just a matter of whether the laboratory CV looks high or low. We demonstrated for our data that the laboratory CVs on the log scale were generally much smaller than the corresponding ones on the untransformed scale, which might make it seem as though the laboratory variability problem has disappeared. However, the log-transform may also reduce the absolute level of the subject-to-subject variability, the magnitude of difference in hormone level associated with different levels of risk, and the within-subject temporal variability. The net change in power for statistical comparisons is unclear.

Studies of the total variability must be specific to individual laboratory techniques. Each laboratory's procedure for measuring a particular hormone may use different antibodies with differing specificities for the analyte of interest. This may result in different absolute levels of the hormone and differences in the subject-to-subject variability and within-subject temporal variability. Whether our strategy is to reduce the measurement error (due to both laboratory variability and within-person temporal variability), adjust for it, or both, we must correctly characterize it. Measurement error adjustments may be sensitive to misspecification of the measurement error proc-

ess. If our goal is to reduce the variability, we must first fully understand where the greatest proportion of the variability problem lies - subject-to-subject, within subject temporal fluctuations, or laboratory error. Then we will know how to appropriately concentrate replication efforts or use other design or analysis strategies such as stratification, laboratory batching or matched cases and controls, or covariate adjustment to reduce the effects of the variability and attain reasonable power to detect potential associations between certain hormones and cancer.

Acknowledgments

We thank C. Brown for the design of the laboratory assay schedule.

References

1. 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.
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. 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.
4. Holzman, D. Elusive estrogens may hold key to some cancer risk. *J. Natl. Cancer Inst.* 87: 1207-1209, 1995.
5. Kuller, L. H., and Gutai, J. P. Re: "A prospective study of endogenous estrogens and breast cancer in postmenopausal women" (Letter). *J. Natl. Cancer Inst.* 87: 1414, 1995.
6. Levitz, M., Banerjee, S., Koenig, K., Shore, R. E., Toniolo, P., and Zeleniuch-Jacquotte, A. Response to Kuller and Gutai (Letter). *J. Natl. Cancer Inst.* 87: 1414-1415, 1995.
7. Siiteri, P. K. Response to Kuller and Gutai (Letter). *J. Natl. Cancer Inst.* 87: 1415, 1995.
8. DISC Collaborative Research Group. Dietary Intervention Study in Children (DISC) with elevated low-density-lipoprotein cholesterol: design and baseline characteristics. *Ann. Epidemiol.* 3: 393-402, 1993.
9. Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*, pp. 249-255. New York: John Wiley & Sons, Inc., 1992.
10. SAS Institute, Inc. *The Mixed Procedure*. SAS Technical Report P-229. SAS/STAT Software Changes and Enhancements, pp. 287-366. Cary, NC: SAS Institute, 1992.
11. Ross, R., Bernstein, L., Judd, H., Hanisch, R., Pike, M., and Henderson, B. Serum testosterone levels in healthy young black and white men. *J. Natl. Cancer Inst.* 76: 45-48, 1986.
12. 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.
13. 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.
14. Phillips, G. B. The variability of the serum estradiol level in men: effect of stress (college examinations), cigarette smoking, and coffee drinking on the serum sex hormones and other hormone levels. *Steroids* 57: 135-141, 1992.