Pooling of Case Specimens to Create Standard Serum Sets for Screening Cancer Biomarkers

Steven J. Skates,[1] Nora K. Horick,[1] Joseph M. Moy,[2] Anna M. Minihan,[1] Michael V. Seiden,[3] Jeffrey R. Marks,[6] Patrick Sluss,[4] and Daniel W. Cramer[5]

[1]Biostatistics Center, [2]Reproductive Endocrine Unit, [3]Ovarian Cancer Biology Laboratory, Department of Medicine, and [4]Pathology Department, Massachusetts General Hospital; [5]Obstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; and [6]Duke University Medical Center, Durham, North Carolina

Abstract

Background: **Multiple identical sets of sera from cancer cases and controls would facilitate standardized testing of biomarkers. We describe the creation and use of standard serum sets developed from healthy donors and pooled sera from ovarian, breast, and endometrial cancer cases.**

Methods: **Two hundred seventy-five 0.3-mL aliquots of sera were created for each of the 95 healthy women, and residual serum was pooled to create 275 identical sets of 20 0.3-mL aliquots. Aliquots (1.0-1.5 mL) from 441 women were combined to create 12 breast and pelvic disease pools with at least 115 0.3-mL aliquots. Sets were assembled to contain aliquots from individual controls, replicates, and disease pools. Cancer antigens (CA), CA 125, CA 19.9, and CA 15.3, and carcinoembryonic antigen were measured in one set and in 217 women comprising six of the pelvic disease pools. Use of a set was illustrated for mesothelin (soluble mesothelin-related protein). Statistical output included concentration differences between pooled cases and controls (z values for single analytes; Mahalanobis distances for pairs), correlation between z values and sensitivities, coefficient of variations, and standardized biases.**

Results: **Marker concentrations in the six pelvic disease pools were generally within 0.25 SD of the actual average, and z values correlated well with sensitivities. CA 125 remains the best single marker for nonmucinous ovarian cancer, complemented by CA 15.3 or soluble mesothelin-related protein. There is no comparable breast cancer biomarker among the current analytes tested.**

Conclusion: **The potential value of standard serum sets for initial assessment of candidate biomarkers is illustrated. Sets are now available through the Early Detection Research Network to evaluate biomarkers for women's cancers.** (Cancer Epidemiol Biomarkers Prev 2007;16(2):334–41)

Introduction

A popular model for the development of biomarkers for the early detection of cancer divides the process into five phases (1). Phase I is the discovery of potential biomarkers by proteomic or genomic approaches. Phase II is the application of a working clinical assay in cancer case and control specimens to show that the marker can actually distinguish the two groups. Phase III uses retrospective banks of longitudinally collected specimens to establish whether the biomarker can detect preclinical disease. Phases IV and V prospectively evaluates the ability of the biomarker to reduce cancer morbidity and mortality in a cost-effective manner in field trials.

It is fair to say that the pace of discovery of potential new biomarkers far exceeds the current ability to move them forward to later phases of development. Even the seemingly easy task of moving a biomarker from phase I to II may pose a challenge. After a working serum assay has been developed, investigators must assemble cancer cases and controls. Blood from cases should be collected before any therapy that might affect biomarker levels. Demographics on cases and controls contributing sera should be available as well as comparative data for other tumor markers. In reality, biomarker developers often lack good clinical specimens for testing a potential marker and, as best they can, assemble convenience samples, often without good description of the demographics on subjects or normative data on other biomarkers. Ultimately, data may be published that have little possibility of allowing comparison with other biomarkers or generalizing to other populations.

One solution might be highly replicated ($\geq$100) sets assembled from large-volume serum samples split into multiple small-volume aliquots, with a set consisting of a single aliquot from each sample. Blood would be obtained from cases with the target cancer before treatment and from healthy matched controls. The large number of replicate sets would then provide a resource to standardize phase II screening of new candidate biomarkers for the target cancer. Investigators wishing to assess the performance of a new cancer biomarker could request a set, perform the assay in a blinded fashion, and submit the results for unbiased data analysis by an independent party who would return results to the investigator and make them available to the broader research community through an accessible Web site.

Practically, it is possible to prepare sets of multiple aliquots from blood electively donated by healthy controls; however, high-volume blood draws done preoperatively on cancer cases would not be safe or ethical. Pooling of small-volume case specimens is an alternate approach for obtaining multiple aliquots for cancer cases. Comparing assay results for a new biomarker in the pooled specimen with values from the individual control specimens would yield a simple test of the performance of a new marker compared with standard biomarkers already measured in the set. Multiple pools of cases with specific combinations of stage and histology could provide further insight into the characteristics of new markers. In this article, we describe the actual construction of standard serum sets for evaluating biomarkers for ovarian, breast, or endometrial cancer and discuss the inferences possible using pooled case specimens.

Received 7/27/06; revised 11/2/06; accepted 11/30/06.

Grant support: National Cancer Institute grant U01CA086381 and the Ovarian Cancer Education Awareness Network Foundation at Massachusetts General Hospital.

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.

Requests for reprints: Daniel W. Cramer, Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, MA 02115. Phone: 617-732-4895; Fax: 617-732-4899. E-mail: dcramer@partners.org

doi:10.1158/1055-9965.EPI-06-0681

Materials and Methods

Individual Controls. Controls were recruited, screened, and phlebotomized through a commercial blood donation laboratory, Research Blood Components, LLC, in Brighton, MA. The study was approved both by the ''Partners'' Institutional Review Board, jointly covering protocols at the Brigham and Women's Hospital and the Massachusetts General Hospital, and by the New England Institutional Review Board covering donation of blood for research purposes at Research Blood Components. Subjects were recruited from prior donors, newspaper ads, and the Craig's List Web site. Subjects had no personal history or strong family history of cancer (i.e., two or more first-degree relatives with the same type of cancer or a mother or sister with ovarian cancer at any age or breast cancer diagnosed before age 50 years) and no medical problems that would preclude blood donation. We sought to match the ages of the donors with the ages of the cases selected for the disease pools. Eligible subjects completed a self-administered questionnaire capturing relevant medical history, including age, pregnancy and birth control history, and hormone and tobacco use. Participants consented to the use of their donated specimen and de-identified medical history by investigators affiliated with the Early Detection Research Network and received $50 for participating in the study.

Between December 2004 to January 2006, 98 women who were eligible agreed to participate and had 300 to 350 mL whole blood collected in standard 10 mL red-stoppered plastic serum tubes with clot activator (Becton Dickinson, Franklin Lakes, NJ). The specimens were allowed to clot for 15 to 30 min, centrifuged at $\sim$2,500 $\times$ *g* for 15 min, and decanted to yield $\sim$100 mL of serum on each subject. The serum was stored at -80°C before transfer to the Tumor Marker Laboratory at Massachusetts General Hospital. After aliquoting, the serum on four of these subjects yielded less than the minimum 280 aliquots desired. One subject was redrawn; however, three others could not be redrawn and were excluded, leaving a total of 95 controls.

Construction of Disease Pools. Sera for the cancer and benign disease pools were collected under the ''Pre-Operative Pelvic Mass'' protocol at Partners and under the ''Blood and Tissue Bank for the Discovery and Validation of Circulating Breast Cancer Markers'' protocol at Duke. Under both protocols, blood was collected from subjects before surgical resection and included consent for sharing specimens with Early Detection Research Network investigators. Specimens for the pelvic disease pools (1-7 in the list below) were collected between January 2001 and November 2005 and for the breast disease pools (8-12 in the list below) between June 1999 and October 2005. All specimens were collected in red-stoppered tubes and processed generally within 4 h (but not >12 h) after collection in a manner similar to that described above and stored at -80°C. There were 12 benign disease and cancer groups as described below (number of subjects contributing to each pool shown in parentheses):

1. Premenopausal women with late-stage, nonmucinous ovarian cancer (35).
2. Postmenopausal women with late-stage, nonmucinous ovarian cancer (39).
3. Postmenopausal women with early-stage, nonmucinous ovarian cancer (35).
4. Premenopausal/postmenopausal women with mucinous ovarian cancer (35).
5. Premenopausal/postmenopausal women with endometrial cancer (12).
6. Premenopausal women with endometriosis (38).
7. Postmenopausal women with benign serous ovarian tumors (35).
8. Premenopausal women with invasive breast cancer (43).
9. Postmenopausal women with estrogen receptor–positive invasive breast cancer (36).
10. Premenopausal/postmenopausal women with ductal carcinoma *in situ* (43).
11. Premenopausal women with benign breast disease (45).
12. Postmenopausal women with benign breast disease (45).

For the nonmucinous ovarian cancers, the distribution of histologies were 61% serous invasive, 13% endometrioid, 11% serous borderline, 9% mixed/undifferentiated, and 6% clear cell. Insofar as possible, we used equal volumes for each individual contributing to a pool. One original (and not previously thawed) 1.0-mL aliquot from each breast disease case and 1.5 mL from each pelvic disease case (three aliquots for the endometrial cancer cases) were thawed at 4°C and pooled to create a total volume of $\sim$45 mL for each of the 12 disease pools, which was placed in 250 mL disposable plastic beakers and homogenized by mixing. For all of the pelvic disease pools except the endometrial cancer group, there were a sufficient number of aliquots remaining to allow measurement of standard biomarkers on each of the individuals contributing to the pools.

Creation of Aliquots and the Standard Serum Sets. We used an automated system to fill, seal, and barcode 0.3 mL plastic capillary ''straws'' (Cryo Bio System, L'Aigle, France) from the freshly created disease pools or individual controls. For the latter, the frozen bulk specimens were thawed at 4°C and transferred to 250 mL plastic beakers for mixing, one at a time, before aliquoting. The aliquots created for each individual control and each disease pool were counted and stored in separate cylindrical containers called ''goblets,'' holding up to 144 straws, and placed in liquid nitrogen before assembly into the sets. Residual serum from controls, beyond that necessary for filling 300 straws, was combined into one pool and supplemented with pooled female sera used for laboratory standardization (ProMedDx, Norton, MA) to make 2 liters of serum. From this, 300 sets of 20 0.3-mL identical replicates were created for calculating the coefficient of variation of an assay. Straws were assigned randomly generated barcoded IDs such that a set recipient would be unaware whether the aliquot came from an individual control, a disease pool, or the replicate set. All automated aliquoting and labeling was done at room temperature and completed within 20 to 60 min.

After inventorying the straws created, the number of straws that could be used to construct sets of equal numbers was determined. The minimum number of aliquots for controls was 280, providing enough for 275 control sets and allowing at least five additional 0.3-mL aliquots for each control to measure standard markers. The number of aliquots created for the disease pools ranged from 117 to 150, allowing a minimum of 115 sets that could contain the breast disease pools and 135 sets that could contain the pelvic disease pools plus at least two additional aliquots for measuring standard markers. Twenty of the identical ''control pool'' replicates were added to each set. The 0.3-mL aliquots from each control, disease pool, and set of replicates were then reassembled to create standard serum sets, each contained in a goblet labeled to describe the set. Set 1 contains 95 individual controls and 20 replicates; 140 were created. Set 2 contains the controls and replicates plus seven pelvic disease pools; 20 were created. Set 3 contains the controls, replicates, seven pelvic disease pools, and the five breast disease pools; 115 were created. The goblets were then packaged on dry ice and sent by air express (World Courier) to the National Cancer Institute Biorepository (Frederick, MD), where the (still frozen) goblets were promptly transferred into liquid nitrogen storage tanks.

Measuring Standard Cancer Biomarkers. Established cancer antigens (CA), CA 125, CA 15.3, and CA 19.9, and

carcinoembryonic antigen (CEA), were measured in singlet using a fully automated immunoanalyzer (Modular E170; Roche Diagnostics, Indianapolis, IN). Reagent kits, calibrators, and quality control materials were purchased from the manufacturer (Roche Diagnostics). Reagent handling, instrument operation, and specimen testing were done according to manufacturer's guidelines. Quality controls were done each operational day, and proficiency was monitored by participation in the College of American Pathologists program.

The CA 125 assay is a double-antibody chemiluminescence immunoassay based on two monoclonal antibodies (M 11 and OC 125; Fujirebio Diagnostics, Inc., Malvern, PA). The Roche CA 125 II test has been standardized against the Enzymun-Test CA 125 II method that, in turn, was standardized against the CA 125 II RIA from Fujirebio Diagnostics. The upper 95th percentile cutoff for healthy women is 35 units/mL. The CA 15.3 II assay is a double-antibody chemiluminescence immunoassay based on two monoclonal antibodies (115D8 and DF3; Fujirebio Diagnostics). The Roche CA 15.3 test has been standardized against the Roche 15.3 assay that, in turn, was standardized against the Enzymun-Test CA 15-3 and CA 15-3 RIA (Fujirebio Diagnostics). The upper 95th percentile cutoff for healthy premenopausal and postmenopausal women is 25 units/mL. The CA 19.9 assay is a double-antibody chemiluminescence immunoassay based on two monoclonal antibodies (1116 NS 19 9, Fujirebio Diagnostics, and a Roche monoclonal antibody). The Roche CA 19.9 test has been standardized against the Enzymun Test CA 19.9 test. The upper 97.5th percentile cutoff for healthy premenopausal and postmenopausal women is 35 units/mL. The CEA assay is a double-antibody electrochemiluminescence immunoassay based on two monoclonal antibodies (Roche Diagnostics). The Roche CEA test has been standardized against the first WHO International Reference Standard (WHO 73/601). The upper 95th percentile cutoff for healthy individuals is 3.4 ng/mL. The equivalent cutoff value for smokers is 4.3 ng/mL.

The method used to measure soluble mesothelin-related protein (SMRP) in serum is a double monoclonal ELISA (MESOMARK, Fujirebio Diagnostics). The assay was done according to the manufacturer's recommendations, and results are reported as nanomoles per liter based on the manufacturer's calibration. There is no internationally recognized SMRP reference standard available at this time. The analytic sensitivity of the test is 2 nmol/L. The dynamic range of the assay is 2 to 32 nmol/L, and the reportable range extends up to 320 nmol/L using a 10-fold dilution. Study specimens were tested in duplicate. The imprecision (percentage coefficient of variation) of the test, based on two concentrations of manufacturer-supplied quality control materials containing 5.04 and 13.54 nmol/L, was 3.1% and 2.5%, respectively.

Statistical Analysis. Single and multiple linear regression analyses were used to examine the influence of demographic and medical characteristics on marker concentration among controls after applying a logarithmic transformation on each marker. In univariate models, age and menopausal status affected the level of CA 125, CA 19.9, and CEA. In multivariate models, the effects of age and menopausal status were highly correlated. As menopausal status was a stronger predictor, age was dropped from the final multiple linear regression.

For the four standard markers in the six pelvic disease pools (1-4 and 6-7) in which we measured the markers in each individual case, we compared the concentration in the pool with the average from the individual measurements using a measure called the ''standardized bias.'' When pooling samples, the concentration of a biomarker in the pooled sample should be the average of the concentrations in the individual samples, assuming the same volume is contributed to the pool from each sample and the fact that concentrations combine linearly on the original concentration scale. Because the distribution of these markers are all strongly skewed to the right and cover orders of magnitude, a logarithmic transformation is required to achieve a distribution close to a normal distribution. Normal theory statistics, such as the mean and SD, are most appropriately calculated after transforming a variable to a scale for which it is more normally distributed than the original scale. Thus, the standardized bias is the number of SDs on the logarithmic scale that the pooled concentration is from the actual average.

Next, we compared the concentration of the markers in the pooled sample with the distribution in the appropriate control group. For the control groups, the median was always less than the average, and the distributions were skewed to the right similar to the cases. Log transformation on values measured in the control subjects resulted in distributions much closer to the normal distribution. When means and SDs are listed, they are always calculated on the logarithmic scale and then exponentiated back to the original biomarker scale for interpretation on a more familiar scale (although they are not precisely equivalent). To assess and rank the value of the biomarkers for detecting breast or pelvic disease, the number of SDs (in controls) in the pooled concentration that was above (or below) the mean of the appropriate control group was calculated. Such a quantity is often called a z value. The usual measure of a biomarker is its sensitivity at a high specificity, the estimation of which requires individual measurements on cases and control subjects. Because the reference set does not provide individual samples for cases, we assessed the accuracy of the z value as a surrogate for sensitivity. For the four standard markers in the six disease pools in which individuals were tested, we examined Pearson's correlation between the z value and the estimated sensitivity at 90% specificity. With fewer than 100 control subjects, this lower level was chosen because a higher specificity such as 95% to 98%, although desirable for ovarian cancer, would require around 500 or more control subjects to estimate accurately.

To simultaneously assess multiple markers, we calculated the multivariate version of the z value, namely the Mahalanobis distance, which is the squared distance between the multivariate concentration of the pooled sample and the multivariate mean of the controls normalized by the variance/covariance matrix (2). As above, biomarker concentrations are transformed to the logarithmic scale before calculating the Mahalanobis distance. Hence, for multiple markers, the square root of the Mahalanobis distance is analogous to the z value and is exactly equivalent for one marker. We use this measure to rank pairs of biomarkers and provide an assessment of the complementarity of an additional marker compared with the z value for the current best single marker (e.g., CA 125 for ovarian cancer).

Results

Coefficients of Variation and Effects of Control Characteristics on Standard Markers. Based on the 20 identical aliquots, the coefficients of variation for the four standard markers were as follows: CA 125 (1.9%), CA 19.9 (1.7%), CA 15.3 (2.5%), and CEA (2.6%). Several demographic/medical variables were found to affect the level of the standard markers measured in controls (Table 1). Postmenopausal women had significantly lower CA 125 levels but higher levels of CEA levels and CA 19.9 compared with premenopausal women. Whites had higher levels of CA 125 and CA 15.3 compared with non-Whites. Women who had ever smoked had higher levels of CEA compared with women who never smoked. Women who had ever been pregnant had higher levels of CEA compared with women who had never been pregnant in a univariate comparison; however, this effect diminished when accounting for the other variables in the multivariate model. There were

Table 1. Effect of key demographic and clinical characteristics on standard markers

Characteristic	*n* (%)	Mean marker levels (SD)				
		CA 125 (units/mL)	CA 19.9 (units/mL)	CA 15.3 (units/mL)	CEA (ng/mL)	SMRP (nmol/L)
Race						
White	79 (83%)	16.3 (1.8)	10.4 (3.2)	16.8 (1.5)	1.8 (1.7)	0.7 (1.6)
Black	11 (12%)	10.9 (1.7)	9.6 (4.4)	14.9 (1.6)	2.2 (1.8)	0.5 (1.7)
Other/unknown	5 (5%)	11.3 (1.2)	5.9 (5.1)	10.2 (1.5)	2.4 (2.1)	0.7 (2.0)
Univariate *P**		0.05	0.61	0.03	0.29	0.19
Multivariate *P*†		0.01	0.43	0.01	0.31	0.22
Menopausal status						
Premenopausal	39 (41%)	18.2 (2.0)	7.5 (4.1)	15.9 (1.6)	1.6 (1.7)	0.6 (1.5)
Postmenopausal	56 (59%)	13.5 (1.6)	12.2 (2.8)	16.3 (1.5)	2.1 (1.8)	0.8 (1.6)
Univariate *P**		0.01	0.06	0.81	0.02	<0.01
Multivariate *P*†		<0.01	0.06	0.84	0.01	0.01
Ever pregnant?						
Yes	68 (72%)	15.8 (1.9)	10.7 (3.4)	16.7 (1.5)	2.0 (1.7)	0.7 (1.6)
No	27 (28%)	13.9 (1.5)	8.2 (3.6)	14.8 (1.6)	1.5 (1.8)	0.6 (1.5)
Univariate *P**		0.35	0.34	0.18	0.02	0.12
Multivariate *P*†		0.07	0.36	0.06	0.09	0.16
Ever smoked?						
Yes	44 (46%)	14.8 (1.6)	10.2 (3.4)	16.0 (1.6)	2.3 (1.8)	0.7 (1.6)
No	51 (54%)	15.7 (2.0)	9.7 (3.5)	16.3 (1.5)	1.5 (1.6)	0.7 (1.6)
Univariate *P**		0.63	0.84	0.82	<0.01	0.56
Multivariate *P*†		0.62	0.98	0.65	<0.01	0.66
Overall	95	15.3 (1.8)	10.0 (3.4)	16.1 (1.5)	1.8 (1.8)	0.7 (1.6)

*P value from linear regression of individual characteristic on log-transformed marker values.
† P value from linear regression of all characteristics on log-transformed marker values.

no effects on any of the four standard markers due to ever use of birth control pills or menopausal hormone therapy (data not shown).

Compared with the racial distribution of controls of 83% Whites and 17% non-Whites, the racial distribution of the women contributing to the pelvic pools was 216 (94.3%) Whites and 13 (5.7%) non-Whites, indicating significantly fewer non-Whites in the pelvic disease pool. For women contributing to the breast disease pools, there were 171 (80.7%) Whites compared with 41 (19.3%) non-Whites, which did not differ significantly from controls. The proportions of premenopausal and postmenopausal women among the pelvic disease pools, 37.1% premenopausal and 62.9% postmenopausal, were nearly identical to the control distribution, 41% premenopausal and 59% postmenopausal, whereas there were fewer (but not significantly fewer) postmenopausal women among those contributing to the breast disease pools, 48.6% premenopausal and 51.4% postmenopausal. The pelvic disease cases did not differ significantly from controls in gravidity, birth control or hormonal therapy use, and smoking (data not shown). These variables were not available for the individuals contributing to the breast disease pools.

Pooled Marker Values Compared with Actual Means. Table 2 compares the pooled value for the standard markers CA 125, CA 15.3, CA 19.9, and CEA with the actual mean observed in the six pelvic disease pools for which individual aliquots were available using the standardized bias. In general, a standardized bias between plus or minus one quarter (0.25) of a SD is desirable; standardized biases in this range were observed for all of the 24 disease-by-marker groups, except for CA 125 in postmenopausal women with benign disease, CEA in premenopausal women with endometriosis, and CA 19.9 in women with mucinous tumors. That the three largest biases occurred in three different pools and for three different markers suggests that neither a laboratory error in pooling nor a systematic problem in one of the assays when measuring pooled samples are likely causes of the biases.

***Z* values and Relation to Sensitivity.** Table 3 shows the z values for the four standard markers in the 12 disease pools based on the pooled value alone. (We note that for the premenopausal control group, there was one very distinct outlier for CA 125 that was removed before calculation of the mean and SD for controls). In general, the greater the z value is above 3.0 (or below −3.0 in the event the candidate biomarker is decreased in cases), the greater the likelihood that the candidate biomarker significantly distinguishes cases from controls. It is not surprising that CA 125 was an especially strong marker for the pelvic diseases. None of the markers for breast cancer did as well as CA 125 for ovarian cancer, with the strongest being CEA for estrogen receptor–positive invasive breast cancer in postmenopausal women ($z = 1.2$).

Although the sensitivity of a marker cannot be determined when only the pooled value is available, the relationship between z value and sensitivity for the four standard markers can be examined in the six disease pools for which individual marker values were available. Figure 1 shows the correlation between z value and the corresponding sensitivity (at a fixed specificity of 90% in controls). There was a very good correlation ($\rho = 0.93$) between the z value and sensitivity, suggesting that the z value is a reasonable surrogate for ranking biomarkers and for sensitivity when only the pooled value is available for cases.

Assessing Pairs of Biomarker Candidates. A ranking of pairs of biomarkers for each disease pool is gained by calculating the Mahalanobis distance between the pairs of concentrations for a pool and the mean of the controls (2). The square roots of the Mahalanobis distances for bivariate combinations with CA 125 are shown in Table 4 to provide a measure of complementarity to CA 125 and are directly comparable with the z values for CA 125 alone. These values provide benchmarks for new biomarkers when combined in pairs. For example, if a new biomarker combined with CA 125 had a (square root) Mahalanobis distance >11.9 for postmenopausal women with advanced-stage, nonmucinous cancer, then we could conclude that the new biomarker complemented CA 125 better than CA 15.3 and would likely have a higher joint sensitivity with CA 125 than with CA 15.3.

Anticipated Output for a New Marker. The output anticipated for a marker to be evaluated through use of a standard serum set is illustrated for a prototypic ''new''

Table 2. Comparison of pooled value marker with actual average of in individual cases contributing to the pools

	Pelvic disease group					
Menopause	Pre	Post	Post	Pre/post	Pre	Post
Stage	III/IV	III/IV	I/II	I/II/III/IV	Benign	Benign
Histology	Nonmucinous	Nonmucinous	Nonmucinous	Mucinous	Endometriosis	Ovarian serous
CA 125						
Pool	2,349.0	2,246.0	374.5	114.0	113.1	87.1
Sample average	2,265.6	2,145.8	337.8	111.1	109.3	40.6
Standard bias*	0.02	0.03	0.07	0.02	0.04	0.86
CA 19.9						
Pool	30.0	29.0	199.9	2557.0	34.7	17.7
Sample average	29.3	28.1	231.9	766.7	33.6	18.9
Standard bias*	0.02	0.03	−0.08	0.50	0.02	−0.06
CA 15.3						
Pool	216.8	187.9	46.6	26.8	21.2	19.8
Sample average	216.3	187.3	46.7	27.6	20.8	19.0
Standard bias*	≈0	≈0	≈0	−0.04	0.04	0.09
CEA						
Pool	1.8	1.9	2.3	7.8	1.6	2.2
Sample average	1.6	1.6	2.2	7.4	1.4	2.0
Standard bias*	0.14	0.18	0.11	0.05	0.38	0.23

NOTE: The formula for standard bias follows from the sequential application of two principles. The first is that pooling of samples averages the individual concentrations for the cases by linearity of concentrations in mixing equal volumes. Therefore, the first calculation is ''mean (individual cases),'' and this value is compared with ''pool value.'' The second is that normal theory should be applied on a scale for which the distribution is closest to normality. Hence, the next step is to apply the transform, in this case ''log,'' to all measurements in the calculation, then determine the difference and the SD on the scale that is closest to normality.
*Standard bias = [log(pool value) − log(mean(individual cases))] / SD(log(individual cases)).

marker, SMRP. Table 1 reveals that SMRP is higher in postmenopausal subjects. Figure 2 reveals that z values for SMRP achieved elevated levels only for women with advanced-stage, nonmucinous tumors. However, z values were no better than the current best marker, CA 125. SMRP complemented CA 125 for women with late-stage, nonmucinous cancers of the ovary in whom it was ranked by the Mahalanobis distance nearly identical to CA 15.3; however, for other disease groups, it added little value (Table 4). The coefficient of variation of SMRP was 6.9%, calculated using the 20 replicate aliquots included in the set.

Discussion

Pooling of specimens is suggested as a cost-effective way to obtain population estimates for the frequency of infectious diseases (3) or even genetic mutations (4). In the context of evaluating biomarkers for disease, pooling reduces the expense of costly assays by creating multiple small pools from both the cases and controls by combining specimens from two, four, or more (but generally less than 10) individuals in each group (5). Simulated experiments show that this strategy produces reliable estimates of sensitivity and specificity for the biomarker compared with testing each individual in the sample. Our concern, however, is not the cost of the assay but, rather, generating multiple (hundreds) of identical sets of specimens to allow a comparative database on many candidate biomarkers to be developed. Especially for valuable (preoperative) case specimens, creation of cancer pools may be the only feasible option for developing a large number of multiple aliquots. For controls, pooling even small numbers of individuals is unnecessary when healthy blood donors can be recruited.

In this article, we describe the creation of a standard serum set for the initial assessment of cancer biomarkers in women and explore its potential value as a tool for screening new candidate biomarkers for ovarian, breast, and endometrial cancer. We believe that our analyses show that valid inferences can be made about the potential of a biomarker to distinguish diseased cases from controls. By measuring four standard biomarkers in six disease sets, we showed that the concentration of a biomarker in a pooled specimen is generally within one quarter of a SD, which could be derived if the individuals contributing to the pool had each been tested.

We then calculated z values to measure how distant the marker concentration in the pool was above (or below) the mean of the control group (relative to the control SD). Z values are calculated after transforming the original biomarker values to a scale for which the distribution in the control group was closer to the normal distribution. For the four standard biomarkers and SMRP, this transformation was the logarithm, but other transforms for new biomarkers are entirely possible (e.g., the Box-Cox family of transformations; ref. 6). Using

Table 3. *Z* scores for disease pools compared with control distribution—premenopausal control with outlying CA 125 value excluded

Group	Z score*			
	CA 125	CA 19.9	CA 15.3	CEA
Pelvic diseases				
Premenopausal, late stage, nonmucinous	11.0	1.1	6.0	0.2
Postmenopausal, late stage, nonmucinous	11.0	0.8	6.1	−0.2
Postmenopausal, early stage, nonmucinous	7.2	2.7	2.6	0.2
Premenopausal/ postmenopausal, mucinous	4.4	4.7	1.2	2.5
Premenopausal/postmenopausal, endometrial cancer	7.3	2.6	2.9	5.6
Premenopausal, endometriosis	4.2	1.2	0.7	≈ 0
Postmenopausal, benign serous	4.0	0.4	0.5	0.1
Breast diseases				
Premenopausal, invasive	1.0	0.8	0.6	0.3
Postmenopausal, ER+ invasive	0.8	0.1	1.0	1.2
Premenopausal/ postmenopausal, DCIS	0.4	0.4	0.9	0.5
Premenopausal, benign	0.3	0.6	≈ 0	0.5
Postmenopausal, benign	0.4	0.5	0.4	1.1

Abbreviations: DCIS, ductal carcinoma in situ; ER+, estrogen receptor positive.
*Z score = [log(value in pool) − mean(log(values in appropriate controls))] / SD(log(values in appropriate controls)).

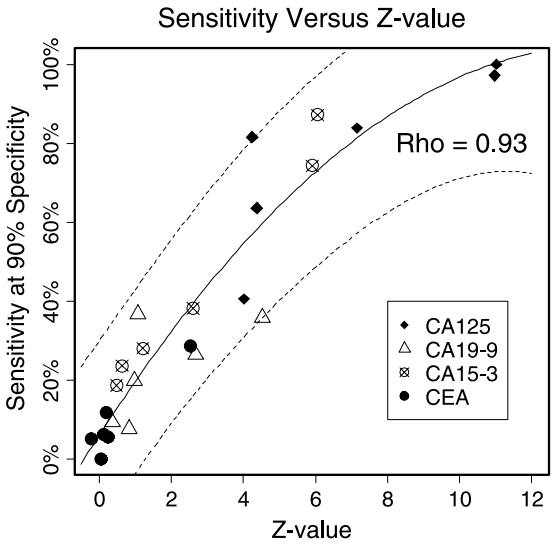


Figure 1. Relationship between z value and sensitivity (at 90% specificity) derived from the distributions in controls and individuals contributing to the case pools, with 95% prediction lines based on a quadratic regression.

disease sets for which we had measured the standard markers in the individuals contributing to the pools, we showed that there was a good correlation between sensitivity and the z value. Z values of ≥ 6 were generally associated with sensitivities of $\geq 80\%$ at 90% specificity in our set of cases and controls.

It is important to note that the z value is not a perfect surrogate for the sensitivity at a given specificity. In the standard biomarkers tested here, the shape of the distribution in the cases was similar to the distribution in the controls (normal on the logarithmic scale), except for a substantially increased mean (in many groups) and an increased SD. If, for a new biomarker, the investigator believes the distribution in the (unmeasured) cases would differ substantially in shape from the distribution in the controls, and not just an increase in the mean and SD, then the z value would be a less reliable surrogate than has been found here for the standard biomarkers. In such situations, the decision for the next stage of biomarker development would rely less on the results of this standard specimen set than in situations in which the biomarker distribution in cases relative to controls behaves similarly to the standard biomarkers. In any case, the pooled nature of the cases for this standard specimen set implies the decision for the next step would be of a moderate nature, such as further refinement of the assay or recommendation of testing in individual cases, rather than, for example, recommending immediate evaluation in a prospective clinical trial of early detection.

An extension of the z value for a combination of two markers is the square-root Mahalonobis distance. This allows an estimate of the complementarity of two markers even when based only on pooled values. Theoretically, the Mahalanobis distance could be extended to any number of markers. However, these reference sets are designed for initial triaging; extending this approach to panels with more than two markers requires estimation of an increasing number of covariances from a fixed moderate number of control subjects, resulting in estimates that may not be accurate.

We also showed the type of output we can expect when a new marker (exemplified by SMRP in this article) is evaluated using only the pooled values. Excellent z values were observed for advanced-stage nonmucinous ovarian cancers (although not better than CA 125). Based on these findings, we would certainly not discourage further studies of SMRP, now using individual cases; however, further refinements to the assay or automation might be necessary to bring it up to the level of CA 125 or to improve its coefficient of variation from the 6.9% we observed. However, based on our results, we would also expect SMRP to have little value in detecting mucinous types of ovarian cancer or breast cancer.

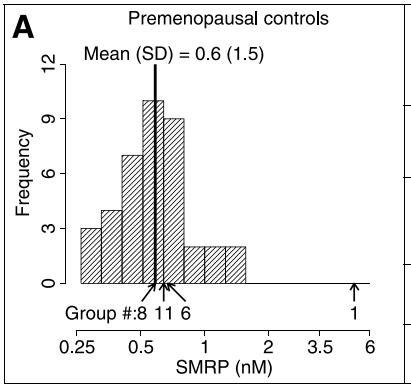
Additional information about a biomarker that can be derived from use of this standard serum set comes from analysis of demographic or medical characteristics that might affect the level of a biomarker. For the standard markers, race, menopausal status, and smoking were found to affect marker levels. Although we did not have a large number of control subjects, it is reassuring that we confirmed several associations previously identified, including lower levels of CA 125 in postmenopausal women (7) and higher levels of CEA in smokers (8).

This discussion of (confounding) variables that might affect marker levels highlights one of the major limitations of using

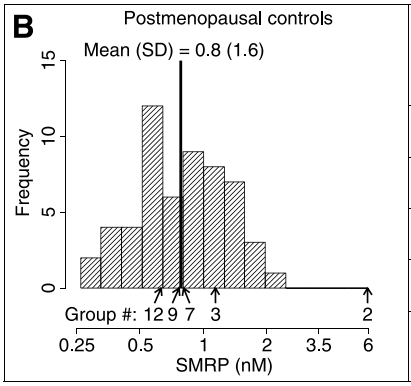
Table 4. Mahalanobis distances for bivariate combinations with CA 125

	Pelvic disease group					
Menopause	Pre	Post	Post	Pre/post	Pre	Post
Stage	III/IV	III/IV	I/II	I/II/III/IV	Benign	Benign
Histology	Nonmucinous	Nonmucinous	Nonmucinous	Mucinous	Endometriosis	Ovarian serous
Z value CA 125	6.8	11.0	7.2	3.4	2.6	4.0
Marker pair	Square-root Mahalanobis distance					
CA 125 and CA 15.3	8.0	11.9	7.3	3.4	2.6	4.0
CA 125 and CA 19.9	6.8	11.0	7.6	5.4	2.6	4.0
CA 125 and CEA	6.8	11.0	7.2	4.2	2.6	4.0
CA 125 and SMRP	8.4	11.8	7.2	3.4	2.6	4.0

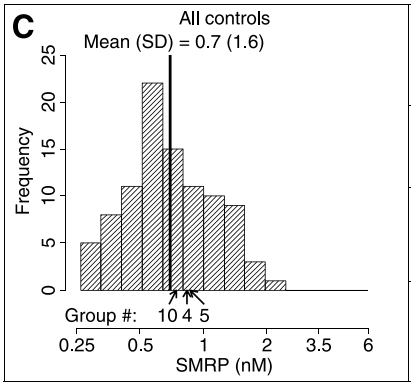
NOTE: Calculation of Mahalanobis distances: denote by x_1 = CA 125 value in pool, by x_2 = CA 19.9/CA 15.3/CEA value in pool, and let $X = [x_1 x_2]$. Denote by y_1 = vector of CA 125 values from appropriate control group, and y_2 = vector of CA 19.9/CA 15.3/CEA values from appropriate control group, and bivariately by $Y = [y_1 y_2]$. Then, the mean on the logarithmic scale is given by $m_1 = \text{mean}(\log(y_1))$ and $m_2 = \text{mean}(\log(y_2))$, bivariately denoted by $M = [m_1 m_2]$. The difference between the measurement in the pooled sample and the mean of the individual samples, on the logarithmic scale, is $D = \log(X) - M$. The variance-covariance matrix is calculated from the individual observations for each biomarker, Σ = variance-covariance matrix of $\log(Y)$. The variance-covariance matrix controls for any correlation between the biomarkers; thus, for example, when two biomarkers are elevated in the presence of ovarian cancer, a negative correlation induces greater complementarity than either a zero or positive correlation, the latter inducing a degree of redundancy. The Mahalanobis distance is given by $D^{\mathrm{T}}\Sigma^{-1}D$, and the square root of the Mahalanobis distance is analogous to the z value in one dimension.



Group #	Group Description	Z-score SMRP	Z-score Current Best (Marker)
1	Stage III/IV, non-mucinous ovarian cancer	5.3	11.0 (CA 125)
6	Endometriosis	0.3	4.2 (CA 125)
8	Invasive breast cancer	≈ 0	1.0 (CA 125)
11	Benign breast cancer	0.2	0.6 (CA 19.9)



Group #	Group Description	Z-score SMRP	Z-score Current Best (Marker)
2	Stage III/IV, non-mucinous ovarian cancer	4.2	11.0 (CA 125)
3	Stage I/II, non-mucinous ovarian cancer	0.8	7.2 (CA 125)
7	Benign serous ovarian tumor	0.1	4.0 (CA 125)
9	ER+ invasive breast cancer	≈ 0	1.2 (CEA)
12	Benign breast disease	-0.4	1.1 (CEA)



Group #	Group Description	Z-score SMRP	Z-score Current Best (Marker)
4	Mucinous ovarian cancer	0.4	4.7 (CA 19.9)
5	Endometrial cancer	0.5	7.3 (CA 125)
10	DCIS	0.1	0.9 (CA 15.3)

Figure 2. Anticipated output for a new marker, illustrated using SMRP (mesothelin).

pooled specimens, namely, the inability to adjust for these variables in examining marker performance. When marker values in individual cases and controls are available together with the demographic information, then adjustment for characteristics can be carried out in analyses of the marker performance. With pooled specimens, such control is only possible at the stage of the original construction of the set in being certain that the controls selected for study matched the demographic characteristics of the individual cases selected for the disease pools. Although the match was reasonable in this set with regard to menopausal status, some disease subsets may differ from controls for other variables. Thus, it is important to understand which demographic variables might affect the markers. For example, in evaluating SMRP, postmenopausal women had higher levels (a mean of 0.6 nmol/L for premenopausal women compared with 0.8 nmol/L for postmenopausal women). Although statistically significant, this difference is small compared with the pooled values between 5 and 6 nmol/L for women with nonmucinous ovarian cancer.

Other potential weaknesses of the set include processing artifacts, heterogeneity of cases, and freezer degradation. Although we aimed to process case bloods within 4 h of draw, the fact that controls were drawn in the same location as the blood was processed undoubtedly led to much shorter times between draw and processing for controls. Although we think this is unlikely to affect robust immune-based assays, our set would not be ideal for techniques sensitive to processing differences such as global proteomic searches. Freezer degradation is possible but expected to be minimal in liquid nitrogen and can be periodically assessed by rerunning the standard markers for the replicate aliquots. Diversity of cancer types could also obscure the value of markers that vary substantially by cancer subtypes. Our separation into mucinous and nonmucinous types of ovarian cancers was intended to address major known contributors to marker variation in ovarian cancer but would not be satisfactory in pinpointing a marker, say for clear cell cancer of the ovary that comprised only 6% or serous borderline tumors comprising 11% of the nonmucinous cases in this series. New pools can be created as knowledge of the molecular profiles of cancer increase.

We believe that the true value of these serum sets will be in standardizing an evaluative process for candidate biomarkers

for breast, ovarian, and endometrial cancer. The fact that the analyses we described recapitulate established knowledge concerning the standard biomarkers examined provides reassurance that evaluation of new candidates with these sets will give reasonable guidance for ranking and determining which candidates warrant further expenditures of resources. The specific elements of our plan that will facilitate these goals include the (*a*) ability to evaluate many dozens of biomarkers on identical sets of serum stored in liquid nitrogen, (*b*) blinded performance of the assays, (*c*) statistical analysis disconnected from the assay site, (*d*) rapid Web accessibility of the results without waiting for publication, and (*e*) ease of comparison with all previous markers tested. A further advantage of our set is that the specimens are totally anonymized. Pooled specimens are inherently anonymous, and the identity of the blood donors was never shared with the Partners' investigators. The ID we used to designate the specimen was not shared with the Blood Bank. Accordingly, the sets have been declared Human Subjects Exempt by the National Cancer Institute.

The sets are now available from the National Cancer Institute for disbursement to biomarker developers. Researchers requesting one of these sets must show that they have a workable assay that may be of value for breast, ovarian, or endometrial cancer. Researchers wishing to assess biomarkers for other cancers in women may request one of the sets consisting of controls and replicates only and collect their own case specimens. Possibly, other disease pools could be added to these control-only sets to begin a database for other cancers in women. Recipients must agree to allow the results to be posted on the Early Detection Research Network Web site. The application process is described in an accompanying announcement with this article. It is hoped that over time, the sets will be useful for prioritizing those markers that early on reveal better performance than and/or complementarity to the best of previous biomarkers measured.

Acknowledgments

We thank Donna Kemp (to whom this work is dedicated) at Research Blood Components, John Moy at the Massachusetts General Hospital Reproductive Endocrine Unit, Allison Vitonis at the Brigham and Women's Hospital Ob/Gyn Epidemiology Center, Karen Drew and Judith Franke at the National Cancer Institute Frederick Facility, the participants in this study, Dr. Mark Thornquist and Jackie Dahlgren of the Fred Hutchinson Cancer Research Center, and Dr. Padma Maruvada of the Cancer Biomarkers Research Group of the National Cancer Institute for their support and encouragement.

References

1. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054–61.
2. Rao R. Linear statistical inference and its application. New York: Wiley Interscience; 2002.
3. Peeling RW, Toye B, Jessamine P, Gemmill I. Pooling of urine specimens for PCR testing: a cost saving strategy for *Chlamydia trachomatis* control programmes. Sex Transm Infect 1998;74:66–70.
4. Amos CI, Frazier ML, Wang W. DNA pooling in mutation detection with reference to sequence analysis. Am J Hum Genet 2000;66:1689–92.
5. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. Stat Med 2003;22:2515–27.
6. Box G, Cox D. An analysis of transformations. J R Stat Soc Ser B 1964;26:211–46.
7. Bon GG, Kenemans P, Verstraeten R, van Kamp GJ, Hilgers J. Serum tumor marker immunoassays in gynecologic oncology: establishment of reference values. Am J Obstet Gynecol 1996;174:107–14.
8. Behbehani AI, Mathew A, Farghaly M, van Dalen A. Reference levels of the tumor markers carcinoembryonic antigen, the carbohydrate antigens 19-9 and 72-4, and cytokeratin fragment 19 using the Elecsys Relecsys 1010 analyzer in a normal population in Kuwait. The importance of the determination of local reference levels. Int J Biol Markers 2002;17:67–70.