

## Validation and Characterization of Human Kallikrein 11 as a Serum Marker for Diagnosis of Ovarian Carcinoma

Martin W. McIntosh,<sup>1</sup> Yan Liu,<sup>1</sup> Charles Drescher,<sup>1</sup> Nicole Urban,<sup>1</sup> and Eleftherios P. Diamandis<sup>2</sup>

**Abstract Purpose:** The serum tumor marker CA 125 is elevated in most clinically advanced ovarian carcinomas, and currently, one of the most promising early detection strategies for ovarian cancer uses CA 125 level in conjunction with imaging. However, CA 125 is elevated in only 50% of early-stage ovarian cancer and is often elevated in women with benign ovarian tumors and other gynecologic diseases. Additional markers may improve on its individual performance if they increase sensitivity and specificity and are less sensitive to other gynecologic conditions. The human kallikrein 11 (hK11) marker has been reported to have favorable predictive value for ovarian cancer, although, by itself, it may be inferior to CA 125.

**Experimental Design:** We here validate the performance of hK11 on an independent data set and further characterize its behavior in multiple types of controls. We also investigate its behavior when combined with CA 125 to form a composite marker. hK11 had not previously been evaluated on these serum samples. CA 125, hK11, and the composite marker were evaluated for their performance in identifying ovarian cancer and for temporal stability.

**Results:** hK11 significantly distinguished ovarian cancer cases from healthy controls and is less sensitive to benign ovarian disease than is CA 125.

**Conclusion:** We conclude that hK11 is a valuable new biomarker for ovarian cancer and its temporal stability implies that it may do even better when used in a longitudinal screening program for early detection.

Women who are diagnosed with ovarian cancer when the tumor is confined to the ovary have good prognoses; however, most ovarian cancers are diagnosed after the disease has spread throughout the peritoneal cavity, when prognosis is poor. More than 80% of these women with late-stage disease will die within 5 years (1, 2). One strategy to improve survival is to detect cancer early, when the disease is localized and potentially treatable by radical surgery.

Currently, one of the most promising early detection strategies for ovarian cancer uses the serum biomarker CA 125 in conjunction with imaging as a trigger for surgical intervention. CA 125 can identify 85% of clinically advanced ovarian carcinomas (3–5), and several studies have shown that

elevations in CA 125 may occur 18 months or more before clinical diagnosis (6, 7). Moreover, the ability of CA 125 to detect cancer early in a screening program is supported by the observation that individual women have temporally stable levels of the marker (8). This suggests that specially tailored screening algorithms could lead to disease detection based on very small serial elevations in CA 125 levels (9–12).

Although CA 125 may be among the best available single diagnostic ovarian cancer biomarkers, its sensitivity and specificity are imperfect. In particular, it is elevated above reference levels in only 50% of clinically detectable early-stage disease (3–5) and it is frequently elevated in patients with benign ovarian tumors and other gynecologic diseases (13–15). Adding one or several markers to CA 125 for use as a composite marker (CM) could improve diagnostic performance if sensitivity were improved with no loss in specificity.

Several research groups recently reported candidate biomarkers for ovarian cancer diagnosis and early detection. For example, carcinoembryonic antigen, placental alkaline phosphatase, various other carbohydrate antigens (e.g., CA15-3 and CA19-9), OVX1, matrix metalloproteinases, prostatic acid phosphatase, HE4 protein, mesothelin, members of the interleukin family, and inhibin have all been proposed as candidate ovarian cancer biomarkers (reviewed in ref. 16). In addition, the family of kallikreins, a group of serine proteases encoded by 15 genes that are localized in tandem on human chromosome 19q13.4, has also been shown to be candidate markers for ovarian and other cancers (17–20), including breast, testicular, and prostate cancer (reviewed in refs. 17, 20). Specifically for ovarian cancer, we have previously reported that serum levels of human kallikreins 5, 6, 8, 10, 11, and 14 are elevated in many patients

**Authors' Affiliations:** <sup>1</sup>Fred Hutchinson Cancer Research Center, Seattle, Washington and <sup>2</sup>Department of Pathology and Laboratory Medicine, Mount Sinai Hospital; Department of Clinical Biochemistry, University Health Network and Toronto Medical Laboratories; and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada  
Received 9/6/06; revised 3/5/07; accepted 5/30/07.

**Grant support:** National Cancer Institute grant P50 CA83636. Research on kallikreins in E.P. Diamandis' lab is supported by grants from the National Cancer Institute (grant R21CA093568), the Natural Sciences and Engineering Research Council of Canada, and IBEX Technologies, Inc., Montreal, Quebec, Canada. 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:** Eleftherios P. Diamandis, Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue Toronto, Ontario, Canada M5G 1X5. Phone: 416-586-8443; Fax: 416-586-8628; E-mail: ediamandis@mtsinai.on.ca.

©2007 American Association for Cancer Research.  
doi:10.1158/1078-0432.CCR-06-2224

with ovarian cancer (21–26). Human kallikrein 11 (hK11), measured by a newly developed ELISA (19), was found to be elevated in the serum of ~70% of ovarian cancer patients (at 95% specificity) and has favorable prognostic value in ovarian cancer (26, 27).

Here, we report the performance of hK11 on its own in a set of sera collected from cohorts not previously evaluated for hK11, which constitutes a validation set for this marker. We characterize hK11 on its ability to distinguish ovarian cancer from healthy controls, from women with benign ovarian disease, and from women undergoing surgery who have histologically normal ovaries. We include the latter surgical normal control group to evaluate the potential for biases that may arise when different sample collection methods are used for cases undergoing surgery and for healthy controls; differences between surgical controls and normal controls may indicate biases in the ascertainment of the case and control samples or perhaps biomarker sensitivity to nonspecific conditions. As a first step toward evaluating the performance of hK11 as an early detection marker, we also evaluated the temporal stability of the marker over a 1-year period among healthy women. This temporal stability assesses whether the sensitivity and specificity of the marker will improve if used in a longitudinal screening program (11, 12, 28).

In addition, we evaluated the marker CA 125 in each subject to compare hK11 and CA 125 and also to investigate whether the two markers could complement each other when combined in a CM.

## Materials and Methods

### Serum specimens

Serum samples from women with ovarian cancer ( $n = 34$ ) and from women representing three different control groups were collected under human subjects approved protocols as part of National Cancer Institute–funded ovarian cancer research programs. Informed consent was obtained from all participants. Controls of several types were selected to help profile the performance of the biomarker. Controls were matched to age and menopausal status of the cases. Healthy controls ( $n = 36$ ), appropriate for evaluating the relevance of the marker for ovarian cancer early detection, were from asymptomatic women participating in a National Cancer Institute–funded ovarian cancer

screening trial. Benign controls ( $n = 21$ ) were used to evaluate marker sensitivity to malignant ovarian disease and were relevant for evaluating the use of the marker as a diagnostic test. Surgical normal controls ( $n = 24$ ) were from women undergoing noncancer gynecologic surgical procedures for pelvic inflammatory disease but who had histologically normal ovaries. The surgical normal controls were included to evaluate the sensitivity of the marker to nonovarian gynecologic conditions and to assess the potential bias due to the surgical collection of our cases compared with nonsurgical collection of the healthy controls. If no difference between healthy and surgical normal specimens is found, we can be confident that a collection bias is not present and that the marker is not sensitive to nonovarian gynecologic disorders.

Specimens from women with ovarian cancer, women with benign ovarian disease, and surgical normal controls were collected in surgery, following anesthesia, but before surgical intervention (i.e., removal of the ovaries). A pathologist examined fixed, paraffin-embedded specimens to confirm the histology for these tumors. The serum from healthy women came from a National Cancer Institute–funded ovarian cancer screening research trial and so represented healthy asymptomatic women (29, 30). Specimen collection and processing protocols were identical for all women regardless of case or control status; participants donated up to 50 mL of blood, which was processed into sera, plasma, and WBC and epithelial cell pellets.

We characterized all research participants at the time of specimen collection with respect to race, age, menopausal status, and use of hormone replacement therapy (HRT). Women were considered postmenopausal if they reported no menstruation for 6 months, used HRT, or were >50 years old and did not report menstrual history. All women using HRT had been taking it for  $\geq 1$  year. Stage and histology were recorded for all cancer cases. Table 1 summarizes these variables for the study population.

To reflect the behavior of the marker among the overall population of ovarian cancer cases, all cases were randomly selected from the specimen repository. We selected 34 cases, including 27 late-stage and 7 early-stage cancers. Of those 34 cases, 21 had serous histology and 2 were early-stage, serous cancers. These sample sizes for the specimens, used as part of a larger National Cancer Institute–funded ovarian cancer biomarker validation study, were determined to have a 70% chance of detecting a sensitivity of 30% or more at specificity of 98% when discriminating cases from combined healthy controls and benign controls using a Wilcoxon rank-sum test. Power also had better than even chance of detecting a sensitivity of 20% or more at 98% specificity. Power was determined by a conservative simulation, where the 70% false-negative cases were assigned marker behavior equivalent to true-negative controls and the 2% false-positive controls were assigned marker behavior equivalent to the true-positive cases. Power under

**Table 1.** Summary description of study participants donating specimen

	Total	Menopausal status		HRT use ever			Race					Stage			
		Pre	Post	Yes	No	NA	Asian	Black	Other*	White	NA	I	II	III	IV
Ovarian cancer	34	1	33	22	8	4	0	1	0	30	3	4	3	20	7
Serous	21	1	20	14	5	2	0	0	0	20	1	1	1	15	4
Mucinous	1	0	1	0	1	0	0	0	0	1	0	1	0	0	0
Endometrioid	4	0	4	2	1	1	0	0	0	3	1	2	1	0	1
Undifferentiated	3	0	3	3	3	0	0	1	0	2	0	0	0	3	0
Other	5	0	5	3	1	1	0	0	0	4	1	0	1	2	2
Benigns	21	2	19	13	7	1	3	0	1	16	1	NA	NA	NA	NA
Surgical controls	24	2	22	17	7	0	0	0	0	24	0	NA	NA	NA	NA
Healthy controls	36	4	32	0	0	36	0	0	0	34	1	NA	NA	NA	NA

NOTE: The "other" ovarian cancer histology cancer category includes 1 adenocarcinoma, 1 clear cell, and 3 unclassified epithelial tumors. Benign tumor conditions include 3 nonneoplastic, 2 mucinous, 1 Brenner, 10 serous, 1 NML, and 4 other categories.

\*Other does not include Hispanic or Native American.

**Table 2.** Summary of raw and standardized markers

Raw CA 125	n	Mean (SD)	25%	50%	75%	95%
Healthy	36	13.25 (6.75)	8.42	11.20	17.72	24.54
Surgically normal	24	30.63 (81.91)	8.56	13.20	16.79	51.50
All nonbenign controls	60	20.20 (52.11)	8.42	12.58	17.46	24.96
Benign	21	70.81 (196.36)	9.21	20.76	32.62	138.71
Cases all	34	414.45 (446.63)	60.77	219.39	648.21	1,148.90
Early stage	7	313.09 (236.49)	134.79	256.28	488.20	657.87
Late stage	27	440.72 (486.61)	49.25	214.47	735.35	1,383.07
Serous	21	397.55 (400.47)	67.43	224.31	619.23	1,148.90
Nonserous	13	441.74 (529.11)	31.49	202.16	657.87	980.10
Standardized (log) CA 125						
Healthy	36	0.000 (1.000)	-0.579	-0.058	0.781	1.376
Surgically normal	24	0.391 (1.654)	-0.501	0.298	0.741	2.801
All nonbenign controls	60	0.134 (1.293)	-0.579	1.155	0.754	1.406
Benign	21	1.302 (2.218)	-0.363	1.131	1.961	4.622
Cases all	34	5.102 (2.876)	3.102	5.464	7.455	8.507
Early stage	7	5.372 (2.076)	4.305	5.750	6.934	7.482
Late stage	27	5.032 (3.079)	2.685	5.423	7.687	8.848
Serous	21	5.283 (2.538)	3.296	5.505	7.371	8.507
Nonserous	13	4.810 (3.445)	1.896	5.314	7.482	8.215
Raw hK11						
Healthy	36	0.445 (0.115)	0.350	0.445	0.543	0.600
Surgically normal	24	0.491 (0.142)	0.378	0.460	0.568	0.700
All nonbenign controls	60	0.463 (0.127)	0.368	0.445	0.550	0.660
Benign	21	0.543 (0.228)	0.410	0.510	0.580	0.990
Cases all	34	0.793 (0.446)	0.525	0.600	0.978	1.500
Early stage	7	0.911 (0.391)	0.625	0.860	1.100	1.600
Late stage	27	0.762 (0.461)	0.480	0.590	0.880	1.500
Serous	21	0.749 (0.317)	0.540	0.600	0.910	1.300
Nonserous	13	0.865 (0.609)	0.440	0.660	1.100	1.600
Standardized (log) hK11						
Healthy	36	0.000 (1.000)	-0.723	0.131	0.835	1.193
Surgically normal	24	0.344 (0.993)	-0.454	0.241	0.994	1.741
All nonbenign controls	60	0.138 (1.003)	-0.549	0.131	0.884	1.532
Benign	21	0.616 (1.206)	-0.160	0.615	1.072	2.973
Cases all	34	1.763 (1.677)	0.718	1.193	2.925	4.449
Early stage	7	2.401 (1.523)	1.319	2.472	3.347	4.679
Late stage	27	1.598 (1.702)	0.388	1.133	2.552	4.449
Serous	21	1.699 (0.317)	0.818	1.193	2.673	3.941
Nonserous	13	1.867 (2.079)	0.091	1.532	3.347	4.679

more complex behavior of the marker may detect even less subtle sensitivities. This power was more than sufficient to detect or independently validate hK11, which in previous studies reported a sensitivity of 70% at 98% specificity (22).

As a first step toward evaluating the relevance of our marker when used in a screening study, we also did measurements on specimens collected serially (1 year apart) from a subsample ( $n = 20$ ) of the healthy controls. Temporal stability was measured by the correlation between the two time points. The performance of markers with high temporal stability may improve when measured over time to monitor deviations from baseline.

**ELISA assay for hK11**

The ELISA assay for quantifying hK11 in serum has been previously published and validated (22). In short, this assay has a detection limit of 0.1 µg/L, and the dynamic range extends to 50 µg/L. The assay has no cross-reactivity from other tissue kallikreins and varies within the measurement range by <10%. All samples were analyzed undiluted in duplicate. More recently, we improved the detection limit to 0.02 µg/L without changing any other assay characteristics.<sup>3</sup> We evaluated CA 125 using the CA 125II sandwich RIA kit from Fujirebio Diagnostics

according to the manufacturer's directions. The intraassay and inter-assay coefficients of variation were <10%. All samples were blinded to the technologists running the assays, and the code was broken to the statisticians after the database was constructed.

**Statistical analysis**

*Quantifying the diagnostic ability of a marker.* Receiver operating characteristic (ROC) curve methods were used to quantify marker performance for hK11 and CA 125. ROC curves associate the sensitivity of a diagnostic test to the entire range of the possible false-positive rate. The false-positive rate is equal to one minus test specificity. The area under the ROC curve (AUC) indicates the average sensitivity of a marker over the entire ROC curve. We also computed the sensitivity of each marker at 95% specificity, a value more relevant to diagnosis and early detection than the overall average sensitivity measured by the AUC. Establishing statistical significance of a single marker is done by the Wilcoxon rank-sum test, which evaluates the significance of the entire ROC curve.

*Standardizing markers.* To aid interpretation of our data when comparing two markers, we first transformed all markers with the natural log so their behavior among healthy women more accurately reflected a normal distribution. We then standardized the markers so they had a mean of 0 and unit SD in the sample of healthy controls (28). Standardization of the markers, which leaves the ROC curves and temporal stability unchanged, facilitates the comparison of two

<sup>3</sup> Unpublished data.

different markers because their units of measurement are now similar (the number of SDs above the average normal subject) as illustrated below.

**Combining markers.** We evaluated a CM of hK11 with CA 125 as a linear combination, or weighting, of the standardized CA 125 and hK11, where logistic regression is used to estimate the weights. Logistic regression has several theoretical properties that make it convenient for applied biomarker research (31), including its capacity to estimate the optimal marker combination.

We estimated our CM by predicting ovarian cancer cases from among all noncancer controls, including healthy, benign, and surgical controls. We limited our attention to a linear combination (i.e., a logistic regression linear link) to facilitate ease of interpretation, although other more complex rules are possible.

**Evaluating temporal stability.** We measured the temporal stability in healthy subjects by computing the Pearson correlation from two time points in the 20 healthy women for whom yearly specimens were available. Markers with high Pearson correlation yielded improved performance in a longitudinal algorithm (8, 11). A high correlation, in particular one exceeding 0.5, implies that monitoring markers for their deviation from historical levels using the parametric empirical Bayes screening rule will yield earlier detection than a simpler diagnostic rule that ignores screening history (11, 32), although we cannot conclude that one marker is better than another, or a marker panel is better than a simple marker, based solely on its temporal stability.

## Results

Table 2 summarizes raw and standardized levels of hK11 and CA 125 for the controls and ovarian cancer cases. Within each subgroup, we provide the number of women, the mean level of the marker, and its SD. We also provide the quantiles of the marker within those groups. The quantiles help summarize the distribution of the marker and provide reference ranges. For example, 95% of all healthy women have hK11 levels below 0.60  $\mu\text{g/L}$  on the raw scale (see Table 2, row 22, column 7) or, equivalently, below 1.193 SDs on the standardized scale

(see Table 2, row 32, column 7). Moreover, the median (50th percentile) level of hK11 in all cases is 0.60  $\mu\text{g/L}$  on the raw scale (see Table 2, row 26, column 5) or 1.193 SDs on the standardized scale (see Table 2, row 36, column 5). This means hK11 has 50% or greater sensitivity at 95% specificity.

Although raw values are helpful when examining a single marker in different samples, they are not useful for comparing different markers because their scales differ. Converting the raw scores to a standardized scale enables us to compare two or more markers across samples. For instance, the median raw scores in our case samples are 219.39 for CA 125 and 0.600 for hK11, but we cannot compare the relative elevation of the two markers; the greater number for CA 125 does not imply that it is hundreds of times more elevated than hK11 in cases. However, numbers on the standardized scale are more comparable, 5.464 for CA 125 and 1.133 for hK11, which implies that CA 125 elevates, on average,  $\sim 4.822$  times more than hK11 in a typical case ( $5.464/1.133 = 4.822$ ).

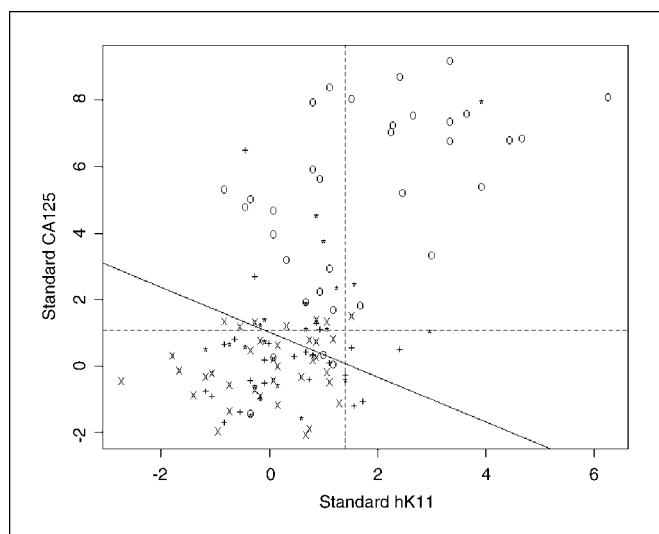
Among healthy subjects alone, we found no statistically significant differences in CA 125 and hK11 marker concentration with respect to age, HRT use, or race. Nearly all cases and controls were postmenopausal, but data for CA 125 were consistent with it being elevated among women who are premenopausal and consistent with previous studies for these markers.

**Results for CA 125.** Table 3 shows that CA 125 is a significant predictor of ovarian cancer from among healthy women and that it did as expected ( $P < 0.0005$ ; AUC = 0.936; sensitivity of 88.2% at 95% specificity). Moreover, CA 125 significantly differentiates between ovarian cancer cases and surgical normal controls ( $P < 0.0005$ ; AUC = 0.902; sensitivity of 76%) and to a similar degree as it does between cases and healthy controls. There is no statistically significant difference between surgical normal and healthy normal controls ( $P = 0.840$ , bottom row of Table 3), implying that the marker

**Table 3.** AUC, Wilcoxon  $P$  value, and sensitivity at near 95% specificity

	hK11						CA 125					
	Wilcoxon			Confidence interval			Wilcoxon			Confidence interval		
	AUC	$P$	Sensitivity	SE	Lower	Upper	AUC	$P$	Sensitivity	SE	Lower	Upper
Cases ( $n = 34$ ) vs healthy ( $n = 36$ )	0.815	<0.001	0.471	0.0856	0.303	0.639	0.936	<0.001	0.882	0.0553	0.774	0.990
Cases vs surgical normal ( $n = 24$ )	0.770	0.001	0.412	0.0844	0.247	0.577	0.902	<0.001	0.765	0.0727	0.622	0.908
Cases vs all nonbenign controls ( $n = 60$ )	0.797	<0.001	0.441	0.0852	0.274	0.608	0.923	<0.001	0.882	0.0553	0.774	0.990
Cases vs benigns ( $n = 21$ )	0.716	0.009	0.265	0.0757	0.117	0.413	0.840	<0.001	0.647	0.0820	0.486	0.808
Cases vs all noncancer controls ( $n = 81$ )	0.776	<0.001	0.441	0.0852	0.274	0.608	0.901	<0.001	0.765	0.0727	0.622	0.908
Benigns vs healthy	0.631	0.145	0.238	0.0929	0.0559	0.420	0.675	0.029	0.333	0.103	0.131	0.535
Benigns vs surgical normals	0.604	0.569	0.095	0.0640	-0.0304	0.220	0.671	0.060	0.143	0.0764	-0.00673	0.293
Benigns vs all nonbenign controls	0.604	0.221	0.143	0.0764	-0.00673	0.293	0.671	0.021	0.333	0.103	0.131	0.535
Surgical normal vs healthy	0.436	0.342	0.083	0.0563	-0.0274	0.193	0.484	0.840	0.083	0.0563	-0.0274	0.193

NOTE: AUC, Wilcoxon  $P$  value, and sensitivity at near 95% specificity are used to evaluate the capabilities of all the markers.



**Fig. 1.** Graphical display of association between standardized CA 125 and standardized hK11, including 34 cases (o), 36 healthy controls (x), 24 surgical normal controls (+), and 21 benign ovarian controls (\*). Line, CM defined by  $CM = 1.126 \times \text{standard CA 125} + 0.761 \times \text{standard hK11}$ . The line represented here gives the classification rule for 95% specificity.

is not affected by surgical collection. However, as expected, CA 125 is significantly different between healthy normal controls and women with benign ovarian disease ( $P = 0.029$ ), and data also suggest differences between benign ovarian disease and surgical normal controls ( $P = 0.060$ ). This shows the well-known sensitivity of CA 125 to nonmalignant ovarian tumors (13–15).

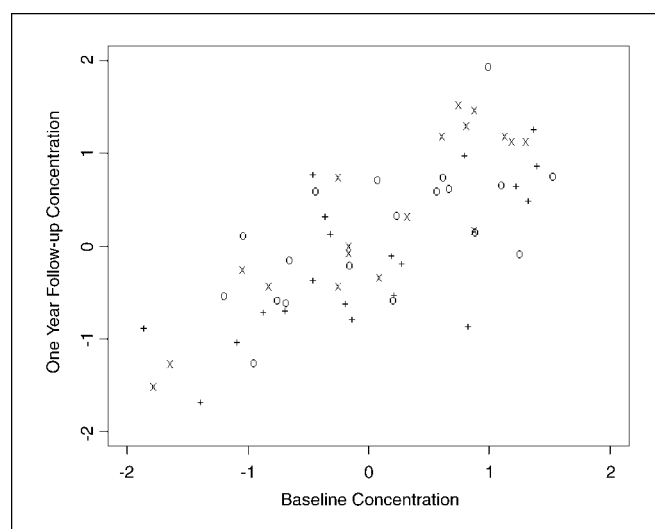
**Results for hK11.** Table 3 shows that hK11 is also a significant classifier of ovarian cancer compared with healthy controls, although it did not do as well as CA 125 ( $P < 0.0005$ ; AUC = 0.77; and sensitivity of 41% for 95% specificity). hK11 also distinguishes between benign controls and ovarian cancer cases ( $P = 0.009$ ; AUC = 0.716; and sensitivity of 26.5%). However, there is no detected difference between hK11 in benign ovarian disease and healthy controls ( $P = 0.145$ ), between benign disease and surgical normal controls ( $P = 0.569$ ), or between surgical normal controls and healthy controls ( $P = 0.342$ ). So, although hK11 cannot distinguish cases from healthy controls as well as CA 125, it may be less sensitive than CA 125 to benign ovarian disease. Moreover, because of the comparable performance between surgical normal and healthy controls, we can be reasonably confident that this conclusion is not related to ascertainment bias.

**Marker combinations.** We investigated the ability of hK11 and CA 125 to form a CM with a better ROC curve than either individual marker. The initial logistic regression controlled for menopausal status and indicator of serous or nonserous tumor. Because neither the interaction terms nor the main effects were significant, they were dropped from the regression. Our resulting CM weights standardized CA 125 with a coefficient of 1.126 ( $P < 0.001$ ) and standardized hK11 with a coefficient of 0.761 ( $P = 0.094$ ). Because the markers have been standardized, these weights imply that CA 125 conveys 60% [ $1.126 / (1.126 + 0.761) \times 100$ ] of the total information in the panel compared with 40% of hK11. The CM is thus defined by  $CM = 1.126 \times (\text{standard CA 125}) + 0.761 \times (\text{standard hK11})$ .

In the marker panel, the coefficient of standard CA 125 is highly significant, but the coefficient for standard hK11 is only modestly significant perhaps due to sample size limitations. Thus, we could not be certain that the two markers combine to form an effective diagnostic panel. Addressing whether they combine together to diagnose cancer early will require the availability of preclinically collected specimens.

The joint behavior of standardized hK11 and CA 125 among cases and all control groups is displayed in Fig. 1. The horizontal axis represents standardized hK11, and the vertical axis represents standardized CA 125. Note that the healthy controls, denoted with “x,” have mean 0 and variance 1 for each marker. The cases, denoted by “o” in the figure, tend to have higher standardized values for each marker than do the various controls, which means a classification rule that uses only one or the other marker has some ability to separate the cases from controls. The role of a CM is to separate cases from controls using two dimensions instead of one. The diagonal line shown in Fig. 1 represents the 95% specificity classification rule estimated from a linear combination of the markers; points above the line are classified as cases and those below are classified as controls. Lines with more (or less) specificity can be represented by lines parallel to that given but higher (or lower) than that shown.

The results above suggest that the diagnostic ROC curves of the CM may not be greater than that of CA 125 alone. Another potential explanation for this result is the lack of power our study has for evaluating marker combinations due to the small sample size of the cases. Careful consideration of the requirements of finding a marker complementary to CA 125 suggests that such a task could be difficult because CA 125 does so well on its own (its overall sensitivity for ovarian cancer is 80%). In our sample, only four individual cases had levels of CA 125 under the usual reference range, and so there are only four cases where hK11 can improve over CA 125 using the usual ROC criterion at high specificity. Definitive conclusions about the complementarities of hK11 and CA 125 for diagnostic testing



**Fig. 2.** Temporal stability as measured from two samples collected 1 y apart for 20 pairs of women. x, standardized hK11; +, CA 125; o, CM. Correlations in plot are 0.85 (95% confidence interval, 0.64–0.94; hK11), 0.72 (95% confidence interval, 0.38–0.87; CA 125), and 0.64 (95% confidence interval, 0.25–0.84; CM).



can only be ascertained with larger studies of the type we have undertaken here.

**Longitudinal marker behavior.** The performance of markers with temporal stability can be improved by accounting for marker history in a longitudinal screening program (8, 33), and several longitudinal algorithms have been proposed (9, 11, 12). In particular, the parametric empirical Bayes algorithm, intended for application of novel markers, makes use of the simple Pearson correlation of a marker measured at two different time points to generate a screening rule (11). Because the CM summarizes both markers into a single numerical score, the parametric empirical Bayes rule can also be applied to the CM.

Figure 2 plots the values of standardized CA 125, hK11, and the CM for the participants who provided repeat measurements. Both CA 125 and hK11 show high temporal stability, with correlations equaling 0.72 and 0.85, respectively. The high temporal stability of CA 125 has been well established (8), but the temporal stability of hK11 is reported here for the first time. The heterogeneity of hK11 and CA 125 imply that the baseline levels of these markers can be highly informative and, if used for early detection in a longitudinal algorithm, that controls for baseline could achieve levels of performance that exceed the performance of the cross-sectional studies presented here. For example, McIntosh et al. (11, 12) show that, when used in a longitudinal study, a correlation of 0.85 for hK11 could mean detecting elevations that are 38% smaller than seen in this cross-sectional study while maintaining the same specificity. One cannot conclude how much earlier cancer would be detected using longitudinal samples without preclinical samples from women who eventually went on to develop cancer; this calculation only establishes the possibility of further gain by controlling baseline, whereas small correlations may rule out such gains.

## Discussion

Recent studies have shown that multiple kallikreins are among the most promising biomarkers for diagnosis and prognosis of ovarian cancer (19–27). For example, hK5, hK6,

hK8, hK10, hK11, and hK14 (21–26) have shown promise as diagnostic serologic markers, and some kallikreins were also found to have prognostic value. We have previously speculated that kallikreins might represent an enzymatic cascade pathway that is activated in ovarian cancer (34, 35). We also have indications that this proteolytic system may cross-talk to other proteolytic systems, such as the metalloproteinases and the urokinase plasminogen activator system (17).

Here, we have independently validated hK11 as a serologic diagnostic biomarker for ovarian cancer. Our results are independent because we evaluated the concentrations of the marker in cases and in controls not previously characterized by hK11. In addition, we evaluated the temporal stability of hK11 and have shown (because the marker is highly heterogeneous in the population) that the marker could do even better if used in a longitudinal algorithm for early disease detection.

We also combined CA 125 and hK11 to evaluate the potential increase in diagnostic efficacy of the resulting CM. In the future, a multiparametric marker panel may be developed by combining other kallikreins that have shown diagnostic value for ovarian cancer (as mentioned above) as well as other newly discovered biomarkers. Future investigation of the combination of CA 125 and hK11 needs to be confirmed in studies specifically powered to evaluate such combinations (36).

The high correlation of hK11, comparable with that of CA 125 (Fig. 2), implies that its AUC and sensitivity reported in Table 3 could be improved if longitudinal algorithms are used to monitor for abnormal marker levels. The potential improvement in the detectable limit of a biomarker when using the parametric empirical Bayes rule can be quantified by computing [square root (1 - correlation)], which gives the relative size of a revised reference range for markers when controlling for screening history. Estimating this for each of the three markers while controlling for screening history would allow us to detect deviations of approximately half the magnitude with the same specificity and sensitivity. In particular for CA 125, hK11, and the CM, we could detect deviations of 44%, 47%, or 59% the size of those detected without controlling for history, respectively.

## References

1. Ozols RF, Rubin SC, Thomas GM, Robboy SJ. Epithelial ovarian cancer. In: Hoskins WJ, Perez CA, Young RC, editors. *Principals and practice of gynecologic oncology*. 3rd ed. Philadelphia: Lippincott Williams and Wilkins; 2000. p. 981–1058.
2. Cannistra SA. Cancer of the ovary. *N Engl J Med* 2004;351:2519–29.
3. Einhorn N, Bast RC, Jr., Knapp RC, Tjernberg B, Zurawski VR, Jr. Preoperative evaluation of serum CA 125 levels in patients with primary epithelial ovarian cancer. *Obstet Gynecol* 1986;67:414–6.
4. Bast RC, Jr., Badgwell D, Lu Z, et al. New tumor markers: CA125 and beyond. *Int J Gynecol Cancer* 2005;15:274–81.
5. Bast RC, Jr., Lilja H, Urban N, et al. Translational crossroads for biomarkers. *Clin Cancer Res* 2005;11:6103–8.
6. Zurawski VR, Knapp RC, Einhorn N, Kenemans P. An initial analysis of preoperative serum CA 125 levels in patients with early stage ovarian carcinoma. *Gynecol Oncol* 1988;30:7–14.
7. Jacobs I, Skates S, MacDonald N, et al. Screening for ovarian cancer: a pilot randomised controlled trial. *Lancet* 1999;353:1207–10.
8. Skates SJ, Singer DE. Quantifying the potential benefit of CA 125 screening for ovarian cancer. *J Clin Epidemiol* 1991;44:365–80.
9. Skates C.J., Pauler DK, Jacobs I. Screening based on the risk of cancer calculation from Bayesian hierarchical change-point models of longitudinal markers. *J Am Stat Assoc* 2001;96:429–39.
10. Crump KC, McIntosh MW, Urban N, Anderson G, Karlan BY. Ovarian cancer tumor marker behavior in asymptomatic healthy women: implications for screening. *Cancer Epidemiol Biomarkers Prev* 2000;9:1107–11.
11. McIntosh M, Urban N, Karlan B. Generating longitudinal screening algorithms using novel biomarkers for disease. *Cancer Epidemiol Biomarkers Prev* 2002;11:159–66.
12. McIntosh MW, Urban N. A parametric empirical Bayes method for screening using longitudinal observations of a biomarker. *Biostatistics* 2003;4:27–40.
13. Jacobs I, Bast RC, Jr. The CA125 tumour-associated antigen: a review of the literature. *Hum Reprod* 1989;4:1–12.
14. Bast RC, Jr., Xu FJ, Yu YH, Barnhill S, Zhang Z, Mills GB. CA125: the past and the future. *Int J Biol Markers* 1998;13:179–87.
15. DiBaise JK, Donovan JP. Markedly elevated CA125 in hepatic cirrhosis: two cases illustrations and the review of the literature. *J Clin Gastroenterol* 1999;28:159–61.
16. Bast RC, Jr. Status of tumor markers in ovarian cancer screening. *J Clin Oncol* 2003;21:200–5.
17. Diamandis EP, Yousef GM. Human tissue kallikreins: a family of new cancer biomarkers. *Clin Chem* 2002;48:1198–205.
18. Yousef GM, Diamandis EP. The new human tissue kallikrein gene family: structure, function, and association to disease. *Endocr Rev* 2001;22:184–204.
19. Borgono C, Michael I, Diamandis E. Human tissue kallikreins: physiologic roles and applications in cancer. *Mol Cancer Res* 2004;2:257–80.
20. Borgono C, Diamandis E. The emerging roles of

- human tissue kallikreins in cancer. *Nat Rev Cancer* 2004;4:876–90.
21. Yousef GM, Polymeris ME, Grass L, et al. Human kallikrein 5: a potential novel serum biomarker for breast and ovarian cancer. *Cancer Res* 2003;63:3958–65.
22. Diamandis EP, Okui A, Mitsui S, et al. Human kallikrein 11: a new biomarker of prostate and ovarian carcinoma. *Cancer Res* 2002;62:295–300.
23. Diamandis E, Scorilas A, Fracchioli S, et al. Human kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma. *J Clin Oncol* 2003;21:1035–43.
24. Luo LY, Katsaros D, Scorilas A, et al. The serum concentration of human kallikrein 10 represents a novel biomarker for ovarian cancer diagnosis and prognosis. *Cancer Res* 2003;63:807–11.
25. Kishi T, Grass L, Soosaipillai A, et al. Human kallikrein 8, a novel biomarker for ovarian carcinoma. *Cancer Res* 2003;63:2771–4.
26. Borgono CA, Fracchioli S, Yousef GM, et al. Favorable prognostic value of tissue human kallikrein 11 (hK11) in patients with ovarian carcinoma. *Int J Cancer* 2003;106:605–10.
27. Diamandis EP, Borgono CA, Scorilas A, Harbeck N, Dorn J, Schmitt M. Human kallikrein 11: an indicator of favorable prognosis in ovarian cancer patients. *Clin Biochem* 2004;37:823–9.
28. McIntosh M, Drescher C, Karlan B, Scholler N, Hellstrom K, Hellstrom I. Combining CA 125 and SMR serum markers for diagnosis and early detection of ovarian carcinoma. *Gynecol Oncol* 2004;95:9–15.
29. Drescher C, Holt SK, Andersen MR, Anderson G, Urban N. Reported ovarian cancer screening among a population-based sample in Washington State. *Obstet Gynecol* 2000;96:70–4.
30. Drescher CW, Nelson J, Peacock S, Andersen MR, McIntosh MW, Urban N. Compliance of average- and intermediate-risk women to semiannual ovarian cancer screening. *Cancer Epidemiol Biomarkers Prev* 2004;13:600–6.
31. McIntosh MW, Pepe M. Combining several screening tests: optimality of the risk score. *Biometrics* 2002;58:657–64.
32. Sato A, Anderson GL, Urban N, McIntosh M. Comparing adaptive and non-adaptive algorithms for cancer early detection with novel biomarkers. *Cancer Biomarkers* 2006;2:151–62.
33. Slate EH, Cronin KA. Change-point modeling of longitudinal PSA as a biomarker for prostate cancer. In: Gatsonis C, Kass RE, McCulloch R, Rossi P, Singpurwall ND, editors. *Case studies in Bayesian statistics III*, vol. III. New York: SpringerVerlag;1997. p. 444–56.
34. Yousef GM, Diamandis EP. Kallikreins, steroid hormones and ovarian cancer: is there a link? *Minerva Endocrinol* 2002;27:157–66.
35. Yousef GM, Diamandis EP. Human tissue kallikreins: a new enzymatic cascade pathway? *Biol Chem* 2002;383:1045–57.
36. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMC Med Res Methodol* 2002;2:4.