Title: Predicting circulating CA125 levels among healthy premenopausal women

Authors and affiliations: Naoko Sasamoto[1,2], Ana Babic,[2,3] Bernard A. Rosner [2,4], Renée T. Fortner[5], Allison F. Vitonis[1], Hidemi Yamamoto[6], Raina N. Fichorova [2,6], Anne Tjønneland[7], Louise Hansen[7], Kim Overvad[8], Marina Kvaskoff[9,10], Agnès Fournier[9,10], Francesca Romana Mancini[9,10], Heiner Boeing[11], Antonia Trichopoulou[12,13], Eleni Peppa[12], Anna Karakatsani[12,14], Domenico Palli[15], Valeria Pala[16], Amalia Mattiello[17], Rosario Tumino[18], Chiara Grasso[19], N. Charlotte Onland-Moret[20], Elisabete Weiderpass[21-24], J. Ramón Quirós[25], Leila Lujan-Barroso[26], Miguel Rodríguez-Barranco[27,28], Sandra Colorado-Yohar[29,30], Aurelio Barricarte[31,32], Miren Dorronsoro[33], Annika Idahl[34,35], Eva Lundin[36], Hanna Sartor[37,38], Kay-Tee Khaw[39], Timothy J Key[40], David Muller[41], Elio Riboli[42], Marc Gunter[42], Laure Dossus[42], Rudolf Kaaks[5], Daniel W. Cramer[1,2], Shelley S. Tworoger[43], Kathryn L. Terry[1,2]

1. Obstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital, Boston, Massachusetts
2. Harvard Medical School, Boston, Massachusetts
3. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts
4. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts
5. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ) Heidelberg, Germany.
6. Laboratory of Genital Tract Biology, Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, Massachusetts
7. Diet, Genes and Environment, Danish Cancer Society Research Center, Copenhagen, Denmark
8. Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus, Denmark
9. CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine - UVSQ, INSERM, Université Paris-Saclay, Villejuif, France
10. Gustave Roussy, Villejuif, France
11. Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

12. Hellenic Health Foundation, Athens, Greece
13. WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Dept. of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece
14. 2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Greece
15. Head, Cancer Risk Factors and Life-Style Epidemiology Unit Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy
16. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milano, Italy
17. Dipartimento Di Medicina Clinica E Chirurgia, Federico II University, Naples, Italy
18. Cancer Registry and Histopathology Department, "Civic - M.P. Arezzo"Hospital, ASP Ragusa, Italy
19. Unit of Cancer Epidemiology– CeRMS, Department of Medical Sciences, University of Turin, Turin, Italy
20. Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
21. Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway
22. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
23. Genetic Epidemiology Group, Folkhälsan Research Center and Faculty of Medicine, Helsinki University, Helsinki, Finland
24. Department of Community Medicine, University of Tromsø , The Arctic University of Norway, Tromsø, Norway
25. Public Health Directorate, Asturias, Spain
26. Unit of Nutrition and Cancer. Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

27. Escuela Andaluza de Salud Pública. Instituto de Investigación Biosanitaria ibs.GRANADA. Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain
28. CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
29. Department of Epidemiology, Murcia Health Council, IMIB-Arrixaca, Spain.
30. Research Group on Demography and Health, National Faculty of Public Health, University of Antioquia, Medellín, Colombia
31. Navarra Public Health Institute, Pamplona, Spain Navarra Institute for Health Research (IdiSNA) Pamplona, Spain
32. CIBER Epidemiology and Public Health CIBERESP, Madrid. Spain
33. Public Health Direction and Biodonostia Research Institute and Ciberesp, Basque Regional Health Department, San Sebastian. Spain
34. Department of Clinical Sciences, Obstetrics and Gynecology, Umeå University, Umeå, Sweden
35. Department of Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden
36. Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden
37. Department of Medical Imaging and Physiology, Skåne University Hospital, Lund, Sweden
38. Department of Translational Medicine, Lund University, Sweden
39. Cancer Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom
40. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford
41. Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, United Kingdom
42. International Agency for Research on Cancer, Lyon, France
43. Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida

Running title: CA125 prediction model in premenopausal women

Keywords: CA125, prediction model, premenopausal, ovarian cancer, screening

Additional Information:

Financial support: This study was funded by National Institutes of Health R01 CA193965, R01 CA 158119, and R35 CA197605. The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF), Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); ERC-2009-AdG 232997 and Nordforsk, Nordic Centre of Excellence programme on Food, Nutrition and Health (Norway); Health Research Fund (FIS), PI13/00061 to Granada; , PI13/01162 to EPIC-Murcia), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten, The Cancer Research Foundation of Northern Sweden (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (United Kingdom).

Corresponding author:

Naoko Sasamoto
Obstetrics and Gynecology Epidemiology Center
Brigham and Women's Hospital
221 Longwood Avenue, Boston, MA 02115
Phone: 617-732-4895, Fax: 617-732-4899
Email: nsasmoto@bwh.harvard.edu

Conflict of interest: The authors declare no potential conflicts of interest.

Abstract word count: 250 words

Manuscript word count: 3,899 words

Number of figures and tables: 3 Tables, 3 Figures, 4 Supplementary Tables

Abstract

Background: CA125 is the most promising ovarian cancer screening biomarker to date. Multiple studies reported CA125 levels vary by personal characteristics, which could inform personalized CA125 thresholds. However, this has not been well described in premenopausal women.

Methods: We evaluated predictors of CA125 levels among 815 premenopausal women from the New England Case Control Study (NEC). We developed linear and dichotomous (≥ 35 U/ mL) CA125 prediction models and externally validated an abridged model restricting to available predictors among 473 premenopausal women in the European Prospective Investigation into Cancer and Nutrition Study (EPIC).

Results: The final linear CA125 prediction model included age, race, tubal ligation, endometriosis, menstrual phase at blood draw, and fibroids, which explained 7% of the total variance of CA125. The correlation between observed and predicted CA125 levels based on the abridged model (including age, race, and menstrual phase at blood draw) had similar correlation coefficients in NEC(r=0.22) and in EPIC(r=0.22). The dichotomous CA125 prediction model included age, tubal ligation, endometriosis, prior personal cancer diagnosis, family history of ovarian cancer, number of miscarriages, menstrual phase at blood draw and smoking status with AUC of 0.83. The abridged dichotomous model (including age, number of miscarriages, menstrual phase at blood draw, and smoking status) showed similar AUCs in NEC(0.73) and in EPIC(0.78).

Conclusions: We identified a combination of factors associated with CA125 levels in premenopausal women.

Impact: Our model could be valuable in identifying healthy women likely to have elevated CA125 and consequently improve its specificity for ovarian cancer screening.

Introduction

Ovarian cancer is the eighth leading cause of cancer death in 2012 with 151,900 deaths worldwide due, in part, to lack of specific symptoms leading to diagnosis at late stage when prognosis is poor(1,2). More than 80% of ovarian cancer patients have elevated cancer antigen 125 (CA125), a membrane bound glycosylated mucin (MUC16), which is used clinically as a prognostic biomarker and to monitor response to therapy(3,4). However, results from two large randomized screening trials in primarily postmenopausal women using transvaginal ultrasound and CA125 (either using 35 U/ml as a cutoff, or the risk of ovarian cancer algorithm (ROCA)) showed no clinically significant benefit(5,6). MUC16 is expressed on a variety of tissues, including the lung, pancreas, stomach, liver, endometrium, and breast, and levels vary between individuals based on demographic, reproductive and lifestyle characteristics(7-11). Therefore, identifying personal characteristics that are associated with CA125 levels could be used to create personalized thresholds for CA125 instead of a single 35 U/mL cutoff, thereby improving the interpretation of measured CA125 and its performance as a screening biomarker and ultimately leading to decreased ovarian cancer mortality.

However, prior studies examining factors associated with CA125 have focused on postmenopausal women(7,8). Thus, we evaluated factors associated with CA125 in premenopausal women and developed and validated CA125 prediction models (linear and dichotomous) among premenopausal women without ovarian cancer from the New England Case-Control Study and the European Prospective Investigation into Nutrition and Cancer study.

Materials and Methods

Study population

The New England Case Control Study (NEC) is a population-based case-control study of ovarian cancer with 2,100 population-based controls enrolled from New Hampshire and eastern Massachusetts over the three study phases (1992-97, 1998-2002, 2003-2008). Details on the study design have been described previously(12-14). Briefly, controls were identified using random-digit dialing, town book selection, and drivers' license lists and frequency matched on age and

state of residence. Approximately half (54%) of the eligible controls that were contacted agreed to participate. We restricted the study population to controls (n=2,100) and excluded women without CA125 measurements (n=96), women postmenopausal at time of blood draw (n=1,130), women who had hysterectomy due to unknown menopausal status (n=30), women who were pregnant or breastfeeding at time of blood draw (n=25), and women with extreme CA125 values ranging from 115.3U/ mL to 411.7 U/mL (n=4) identified based on the generalized extreme studentized deviated many-outlier detection approach applied to log-transformed values(15). In sum, our analysis included 815 premenopausal NEC controls.

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is a multicenter prospective cohort including participants from ten Western European countries developed to evaluate the association between nutrition and cancer. Briefly, 519,978 participants (366,521 women) were enrolled between 1991 and 1998 across 23 research centers. Details on the study design have been described previously(16). A nested case-control study of ovarian cancer was designed within the cohort(17). For each ovarian cancer case, up to four controls were randomly selected using incidence density sampling for a total of 1,939 controls(17). We excluded women without CA125 measurements (n=12), women who were either postmenopausal (n=1,416), or had a hysterectomy or unknown menopausal status (n=38). There were no outlying values in these EPIC controls. In sum, our analysis included a total of 473 premenopausal EPIC controls.

CA125 measurements

In NEC controls we measured CA125 using the CA125II radioimmmunoassay (Centocor, Malvern, PA) at the CER Lab at Boston Children's Hospitals. We assessed the reproducibility of the assay by including five blinded aliquots of a uniform quality control pool in each of the 46 assay batches. The average of the coefficients of variation (CV) was 1%. In EPIC controls and in a subset of NEC controls, we previously measured CA125 using the volume-effective highly sensitive multiplex platform (Meso Scale Discovery (MSD), Gaithersburg, MD) in the Genital Tract Biology Laboratory at Brigham and Women's Hospital(17). The average CV across the assay batches was 19%.

Candidate predictors

We selected factors that have been previously reported to be associated with CA125 in at least one prior study(7-9,11), ovarian cancer risk factors(18), as well as several factors which were biologically plausible to be associated with CA125(10). Those included age at blood draw, race, body mass index (BMI, kg/m^2), smoking status (never, former, current), pack-years calculated by number of packs of cigarettes per day multiplied by the number of years a person had smoked, age at menarche, oral contraceptive use and its duration (months), parity, self-reported endometriosis, tubal ligation, family history of ovarian cancer, prior personal cancer diagnosis, caffeine intake (mg), genital powder use, infertility, number of miscarriages, ectopic pregnancy, ever use of intrauterine device (IUD), fibroids, menstrual cycle regularity and days between last menstrual period (LMP) and blood draw(7-11). Furthermore, we evaluated additional variables related to menstrual characteristics and pregnancy timing: cycle length, days with menstrual bleeding, dysmenorrhea, age at first live birth, age at last live birth, and years since last live birth.

Statistical analyses

We log-transformed CA125 values to achieve a normal distribution. With this transformation, the distribution of log-transformed CA125 was normally distributed with skewness of 0.35 and kurtosis of 0.34, with a bell-shaped histogram. .

Recalibration of CA125

Since the EPIC samples had CA125 measured using an alternate assay (MSD assay) with a different scale, we used recalibration to rescale these measurement results to be comparable to the CA125II assay values. We recalibrated the EPIC CA125 values based on 187 NEC premenopausal controls with CA125 measurements on both CA125II and MSD assays using the drift correction method(19). We regressed the log-transformed MSD assay values to the log-transformed CA125II assay values and used the intercept and effect estimates of the model to calculate the recalibrated CA125II assay values based on the measured MSD assay values for all premenopausal controls in EPIC and used the recalibrated values in our analyses.

Predictors of CA125 in premenopausal women

First, we evaluated the association between individual candidate predictors and CA125 using linear or logistic regression adjusted for continuous age. We used effect estimates of the linear regression for each predictor to calculate the percent change in CA125 levels, calculated as [exp (beta) - 1] x100 for a 1-unit change in the predictor. We determined the optimal modeling of continuous variables (age, BMI, age at menarche, duration of OC use, parity, and smoking pack-years) using restricted cubic splines to test for linearity(20). We used categorical variables for age, dichotomous variable for age at menarche, and piecewise linear spline with single knot for BMI since these variables were non-linearly associated with log-transformed CA125. We created composite categorical variables for OC use and duration and smoking status and pack-years, and compared nested models using likelihood ratio test and non-nested models using the Akaike information criterion and Vuong test(21). Based on these evaluations, candidate predictors were modeled as follows: age at blood draw (categorical, by 10 year intervals from age 30), race (white, non-white), BMI (piecewise linear spline model with single knot at 27), height (continuous, centered at 165), smoking status(categorical, never, former, current) and pack-years (continuous, never smokers, pack-years among former smokers, pack-years among current smokers), age at menarche (age 12 and under, above 12), duration of OC use (continuous, including never users), parity (continuous), endometriosis (no, yes), tubal ligation (no, yes), family history of ovarian cancer (no, yes), prior personal cancer diagnosis (no, yes), caffeine intake (quartiles), genital powder use (no use, body use, genital use), infertility (no, yes), number of miscarriages (0, 1, 2, 3 or more), ectopic pregnancy (no, yes), intrauterine device use (never, ever), fibroids (no, yes), menstrual cycle regularity (regular, irregular) and predicted phase of the menstrual cycle (early follicular, late follicular, peri-ovulatory, luteal, long cycle, irregular, missing) based on the number of days between the last menstrual period and blood draw.

Prediction modeling

Overall, we developed CA125 prediction models (linear and dichotomous) in NEC and conducted external validation in EPIC (Figure 1). We used cross-validation to conduct internal validation of the model developed in NEC. Since information on some of the predictors were not collected in

EPIC, we developed an abridged model restricted to variables available in EPIC from the final model, and then validated the abridged model in EPIC.

Linear model

First, we developed a linear CA125 prediction model of log-transformed CA125 in NEC. We used stepwise linear regression analysis using $p<0.15$ as significance level for entry and retention in the model. In our primary prediction modeling, we used missing indicators for menstrual phase at blood draw due to a proportion of missing values. For variables with a limited number of missing values (age at menarche (n=1), caffeine intake (n=23), menstrual cycle length (n=23)), women with missing values were excluded. Age was forced in the model and the r-squared was calculated for the final prediction model, adjusted for study phase (1992-1997, 1998-2003, 2003-2008) and center (Massachusetts, New Hampshire). In addition, we calculated a delta r-squared that excluded the variability explained by study phase and center as these were matching factors in NEC and were forced into the model(22). The predicted log-transformed CA125 values in NEC were calculated using the effect estimates from the final prediction model. We evaluated the performance of the model by calculating the Pearson correlation coefficient to assess how well the predicted and the observed CA125 values agreed (i.e. calibration). We used 5-fold cross-validation to assess for overfitting in NEC and calculated the average r-squared across all sampled datasets(23). To evaluate potential bias due to missing data of candidate predictors, we conducted a sensitivity analysis restricted to women who had no missing predictors. We also conducted multiple imputation by chained equations (MICE) to impute the missing variables conditional on all of the predictors and outcome(24). We allowed 100 iterations and generated 20 imputed datasets. We applied the final prediction model in the 20 imputed datasets using the methods described and pooled the results of the model estimates using the Rubin's rules(25).

For external validation, we sought to validate our linear CA125 prediction model in EPIC. However, some of our key predictors (endometriosis and fibroids) were not collected in EPIC or were missing in majority of women (tubal ligation), thus, we validated an abridged model restricted to variables available in EPIC. First, among the predictors selected in the final model developed in NEC, we identified predictors available both in NEC and EPIC. We next ran the abridged model in NEC restricting to those variables available in both NEC and EPIC. We used

the effect estimates from this model to calculate the predicted value of log-transformed CA125 in the EPIC samples. We calculated the Pearson correlation coefficient between the predicted and the observed log-transformed CA125 to assess agreement and compare to that in the discovery dataset. We plotted the predicted versus the observed log-transformed CA125 for visual assessment.

Dichotomous model

Next, we developed and validated a dichotomous prediction model of elevated CA125 defined as having CA125 ≥ 35 U/ mL following the same method used for developing the linear CA125 prediction model described above but using logistic stepwise regression analysis. We evaluated the performance of the model by calculating the area under the curve (AUC) in the NEC (discovery) and EPIC (validation).

All statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC), STATA version 12.1 (StataCorp, College Station, TX), and R version 3.4.3.

Results

Overall, the mean CA125 values were 17.3 U/ mL in 815 NEC controls and 14.9 U/ mL in 473 EPIC controls after recalibration. The baseline characteristics of NEC and EPIC premenopausal women were similar except age at blood draw, race, age at menarche, OC use, current hormone use, infertility, parity, and tubal ligation were significantly different. (Supplementary Table S1).

Recalibration of CA125

We recalibrated the CA125 values in EPIC using the recalibration model based on 187 NEC premenopausal controls with both CA125II and MSD assay measurements. These two measurements were highly correlated (r=0.96, 95%CI 0.94, 0.97). After recalibration, the measured and recalibrated CA125 values also showed high correlation (r=0.95, 95%CI 0.93, 0.96). The recalibration model showed high performance in general.

Predictors of CA125 in premenopausal women

Age at blood draw was non-linearly associated with CA125, with women younger than 30 or more than 50 years old having significantly lower CA125 than those aged 30-39 years (Table 1). In age-adjusted models, menstrual phase at blood draw was significantly associated with CA125 levels, with early follicular phase levels being 8 to 21% higher than in other menstrual phases. Endometriosis and fibroids were associated with significantly higher CA125 levels, with 21 and 13% difference, respectively, compared to those who did not have the condition. Current hormonal contraception use and tubal ligation were significantly associated with lower CA125 levels, with -16% and -11% difference, respectively. Cycle length, days with menstrual bleeding, dysmenorrhea, age at first live birth, age at last live birth, and years since last live birth were not significantly associated with CA125 levels in premenopausal women. Similar predictors were significantly associated with CA125 levels in the dichotomous model (Supplementary Table S2).

Linear CA125 prediction modeling

The final linear CA125 prediction model included age at blood draw, race, tubal ligation, endometriosis, menstrual phase at blood draw, and fibroids, with an r-squared of 0.07 (95%CI 0.02, 0.09) (Table 2). The association between individual predictors and CA125 were similar in univariate and multivariate adjusted models. The r-squared of this full linear model when conducting 5-fold cross-validation was 0.02. When we restricted the analysis to the 498 controls with complete information on all predictors and applied the final linear CA125 prediction model, the r-squared was 0.12 (95%CI 0.05, 0.15) (Supplementary Table S3). When restricting to women with complete information on all predictors, the same variables were retained in the final model. For all the models, the delta r-squared, which subtracts the variance attributable to study phase and center from the total variance, was similar to the r-squared reported above. When evaluating the final continuous model in multiple imputed datasets in NEC, the beta coefficients, standard errors, and the r-squared were similar to the original model (Supplementary Table S3). The small differences in the measures of association when running the final model in the dataset using missing indicators, dataset restricted to those with complete information on all potential predictors, and multiple imputed datasets suggest that the missingness of menstrual phase at blood draw do

not largely influence the results. We also observed similar performance of the model when including all significant predictors in the univariate analyses, suggesting that the final model included important key predictors. Predicted log-transformed CA125 calculated based on the final model and the observed log-transformed CA125 were weakly correlated with a Pearson correlation coefficient of 0.26 (95%CI 0.19, 0.33) (Figure 2A).

For external validation, we developed an abridged linear CA125 prediction model which included age at blood draw, race, and menstrual phase at blood draw with r-squared of 0.05 (95%CI 0.01, 0.07) in NEC (Table 2). Using the measures of association from this abridged model, we calculated the predicted log-transformed CA125 values in EPIC. The predicted log-transformed CA125 values had a similar correlation with the observed log-transformed CA125 values in EPIC (r=0.22 (95%CI 0.13, 0.31)) as in the NEC abridged linear model (r=0.22 (95%CI 0.15, 0.29)) (Figure 2B, 2C). The spread of the predicted CA125 values in Figure 2 are much smaller than the spread of the observed CA125 values because the linear prediction model only explains a small proportion of the total variance of the observed CA125 values.

Dichotomous CA125 prediction modeling

The final dichotomous prediction model to predict women with CA125 ≥ 35 U/ mL included age at blood draw, tubal ligation, endometriosis, prior personal cancer diagnosis, family history of ovarian cancer, number of miscarriages, menstrual phase at blood draw, and smoking status and duration with an AUC of 0.83 (95%CI 0.77, 0.89) (Table 3, Figure 3). For menstrual phase at blood draw, we collapsed the other phase and irregular menstruation categories because few individuals had CA125 ≥ 35 U/mL in these groups. The association between individual predictors and CA125 were similar in univariate and multivariate adjusted models. The AUC of this full dichotomous model when conducting 5-fold cross-validation was 0.67. When we restricted the analysis to the 498 controls with complete information on all predictors and applied the final dichotomous model, the AUC was 0.84 (95%CI 0.76, 0.93) (Supplementary Table S4). When we conducted variable selection process using stepwise regression among women with complete information on all predictors, similar predictors were retained except number of miscarriages and

smoking status, resulting with an AUC of 0.79 (95%CI 0.69, 0.89). When evaluating the model in the multiple imputed datasets in NEC, the odds ratios and the AUC were largely similar to the primary analysis (Supplementary Table S4). We also observed a similar performance of the model when including all significant predictors from the univariate analyses, suggesting that the final model included important key predictors. We also considered using 65 U/mL cutoff which has been proposed for premenopausal women(26), but were limited with five controls who had CA125 greater than 65 U/mL so were not able to investigate further.

For external validation, we developed an abridged model, which included age at blood draw, number of miscarriages, menstrual phase (collapsing those on hormones, blood draw at other phase, and having irregular menstruation due to power), and smoking status with an AUC of 0.73 (95%CI 0.65, 0.81) in NEC (Table 3, Figure3). When we applied this model to EPIC using recalibrated CA125 value of 35 U/ mL as cutoff, the AUC was 0.78 (95%CI 0.67, 0.89) (Figure 3).

Discussion

This is the largest population-based study to develop and validate CA125 prediction models among healthy premenopausal women considering both continuous levels as well as those over current clinical threshold of 35 U/ mL. Although, the model predicting continuous CA125 only explained a small percent of the total variability, the model did show comparable correlations between predicted and observed levels in EPIC, suggesting the validity of the model. Conversely, the AUC for predicting elevated CA125 (≥ 35 U/ mL) was relatively high in NEC and validated in EPIC.

Age was non-linearly associated with CA125 in our study, which is consistent with our prior study in EPIC in which we observed an inverse U-shaped association between age and CA125 levels among premenopausal women(9). Similarly, non-white race was associated with significantly lower CA125, which was consistent with prior studies in postmenopausal women(7,8), suggesting the need for different thresholds for minority populations. Unfortunately, we were underpowered

to evaluate differences in prediction models between racial subgroups, though others have described differences in CA125 levels between Black and Asian women(7,8).

Factors related to menstruation were strongly related to CA125. Specifically, an early follicular phase blood draw was significantly associated with higher CA125 levels and strong predictor of CA125 in our final model, which was consistent with previous reports(27). This association is likely driven by MUC16 expression on the endometrium and endometrial shedding during early follicular phase which may lead to higher circulating CA125 levels(10). This could explain the increased CA125 levels in women with fibroids, since fibroids are known to increase menstrual bleeding(28). In contrast, MUC16 expression on the endometrium may explain lower CA125 levels among women with a tubal ligation as this procedure would prevent retrograde menstruation, which occurs in approximately 85% of women during menstruation(29), leading to systemic exposure to the antigen. Factors related to infertility, particularly endometriosis, were also related to substantially higher CA125 levels, consistent with prior studies(30,31). A similar mechanism is likely responsible as endometriosis leads to ectopic endometrial tissue usually in the peritoneal cavity.

Our linear CA125 prediction model explained 7% of the variability in CA125 but showed moderate validation in EPIC, whereas our dichotomous CA125 prediction model had better predictive ability with good validation. These results suggest that the variability of CA125 may be small in general but change dramatically by certain factors such as menstrual phase and endometriosis, and therefore the dichotomous prediction model performed better. We decided to use a standard log-linear model for developing the linear CA125 prediction model because the distribution of log-transformed CA125 was normally distributed with low kurtosis and skewness. When we included all significant predictors in the univariate analyses, both linear and dichotomous models showed similar performance compared to our final model having fewer predictors, suggesting some predictors were correlated.

Interestingly, some factors, such as fibroids and race were only significantly associated with continuous CA125 and some factors, such as prior personal cancer diagnosis and family history of ovarian cancer were only significantly associated with elevated CA125 (above 35 U/ mL). We suspect more predictors were selected in the final dichotomous CA125 prediction model because the association between exposures and CA125 were non-linear.

The major strength of our study is that we had two large independent population-based studies with detailed information on candidate predictors of CA125 to develop and validate CA125 prediction models among premenopausal women. However, there are several limitations to our study. First, we had missing data on several variables. While we used missing indicators for our main analysis, our sensitivity analyses restricting to those with complete information on all predictors and using multiple imputation showed similar results, suggesting that the method for handling missing data did not influence overall results. In addition, we evaluated the performance of our prediction models using cross-validation and conducting external validation in an independent dataset, in which all the results were similar, suggesting a parsimonious model. Secondly, we were not able to validate the full prediction models in the independent dataset. Although we were only able to validate an abridged model in EPIC lacking tubal ligation and endometriosis, we expect the model performance to be better and closer to what we would have observed in NEC if we had information on all predictors. Thirdly, our model could be missing unknown predictors of CA125 since we restricted the candidate predictors to those previously described, which were mostly conducted among postmenopausal women. The relatively low r-squared of the final linear CA125 prediction model suggest that other candidate predictors may exist, such as genetic factors, common medications, and dietary factors, opening new opportunities for future studies. While hysterectomy has been previously described as a predictor of CA125, only few participants in NEC had hysterectomy. Given their ambiguous menopausal status we excluded them from current analysis of premenopausal women. Lastly, the model performance in EPIC could be underestimated because NEC and EPIC used different assays to measure CA125. However, the CA125 values of the two assays were highly correlated (r=0.96) and the predicted CA125 values calculated using the recalibration model were also very highly correlated with the observed CA125II assay values (r =0.95).

In summary, we developed and validated CA125 prediction models among premenopausal women in two independent studies that further our understanding of factors that influence CA125 levels and can therefore be used to optimize ovarian cancer screening with CA125. While performance of population-level screening for ovarian cancer in premenopausal women may be limited due to the lower incidence of ovarian cancer in this age range, approximately 30% of ovarian cancers are diagnosed before age 55. Furthermore, the impact of ovarian cancer in younger women results in potentially greater social, emotional, and economic impact. Further studies are needed to identify new predictors of CA125 to improve the model and to understand the predictors of changes in CA125 over time based on personal characteristics.

Acknowledgements

The authors would like to acknowledge the participants and staff of the New England Case Control Study and the European Prospective Investigation into Cancer and Nutrition study.

References

1. Howlader N NA, Krapcho M, Miller D, Bishop K, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2014. National Cancer Institute. Bethesda, MD, 2017.
2. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. *Cancer Epidemiol Biomarkers Prev.* 2017;26(4):444-457.
3. Duffy MJ, Bonfrer JM, Kulpa J, Rustin GJ, Soletormos G, Torre GC, et al. CA125 in ovarian cancer: European Group on Tumor Markers guidelines for clinical use. *Int J Gynecol Cancer.* 2005;15(5):679-691.
4. Bast RC, Jr., Klug TL, St John E, Jenison E, Niloff JM, Lazarus H, et al. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *N Engl J Med.* 1983;309(15):883-887.
5. Buys SS, Partridge E, Black A, Johnson CC, Lamerato L, Isaacs C, et al. Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. *JAMA.* 2011;305(22):2295-2303.
6. Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer

Screening (UKCTOCS): a randomised controlled trial. *Lancet.* 2016;387(10022):945-956.
7. Pauler DK, Menon U, McIntosh M, Symecko HL, Skates SJ, Jacobs IJ. Factors influencing serum CA125II levels in healthy postmenopausal women. *Cancer Epidemiol Biomarkers Prev.* 2001;10(5):489-493.
8. Johnson CC, Kessel B, Riley TL, Ragard LR, Williams CR, Xu JL, et al. The epidemiology of CA-125 in women without evidence of ovarian cancer in the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial. *Gynecol Oncol.* 2008;110(3):383-389.
9. Fortner RT, Vitonis AF, Schock H, Husing A, Johnson T, Fichorova RN, et al. Correlates of circulating ovarian cancer early detection markers and their contribution to discrimination of early detection models: results from the EPIC cohort. *J Ovarian Res.* 2017;10(1):20.
10. Haridas D, Ponnusamy MP, Chugh S, Lakshmanan I, Seshacharyulu P, Batra SK. MUC16: molecular analysis and its functional implications in benign and malignant conditions. *FASEB J.* 2014;28(10):4183-4199.
11. Westhoff C, Gollub E, Patel J, Rivera H, Bast R, Jr. CA 125 levels in menopausal women. *Obstet Gynecol.* 1990;76(3 Pt 1):428-431.
12. Vitonis AF, Titus-Ernstoff L, Cramer DW. Assessing ovarian cancer risk when considering elective oophorectomy at the time of hysterectomy. *Obstet Gynecol.* 2011;117(5):1042-1050.
13. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol.* 1985;122(1):51-65.
14. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am J Epidemiol.* 1992;135(10):1114-1126; discussion 1127-1136.
15. Rosner B. Percentage points for a generalized ESD many-outlier procedure. . *Technomoetrics* 1983;25:165-172.
16. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* 2002;5(6B):1113-1124.
17. Terry KL, Schock H, Fortner RT, Husing A, Fichorova RN, Yamamoto HS, et al. A Prospective Evaluation of Early Detection Biomarkers for Ovarian Cancer in the European EPIC Cohort. *Clin Cancer Res.* 2016;22(18):4664-4675.
18. Wentzensen N, Poole EM, Trabert B, White E, Arslan AA, Patel AV, et al. Ovarian Cancer Risk Factors by Histologic Subtype: An Analysis From the Ovarian Cancer Cohort Consortium. *J Clin Oncol.* 2016;34(24):2888-2898.
19. Eliassen AH, Hendrickson SJ, Brinton LA, Buring JE, Campos H, Dai Q, et al. Circulating carotenoids and risk of breast cancer: pooled analysis of eight prospective studies. *J Natl Cancer Inst.* 2012;104(24):1905-1916.
20. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med.* 1989;8(5):551-561.
21. Clarke KA. Testing Nonnested Models of International Relations: Reevaluating Realism. *American Journal of Political Science.* 2001;45(3):724-744.

22. Rosner B. *Fundamentals of Biostatistics.* Eighth Edition ed: CENGAGE Learning; 2016.
23. Tworoger SS, Zhang X, Eliassen AH, Qian J, Colditz GA, Willett WC, et al. Inclusion of endogenous hormone levels in risk prediction models of postmenopausal breast cancer. *J Clin Oncol.* 2014;32(28):3111-3117.
24. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.* 2007;16(3):219-242.
25. DB R. *Multiple Imputation for Nonresponse in Surveys.*: John Wiley and Sons: New York; 1987.
26. Eltabbakh GH, Belinson JL, Kennedy AW, Gupta M, Webster K, Blumenson LE. Serum CA-125 measurements > 65 U/mL. Clinical value. *J Reprod Med.* 1997;42(10):617-624.
27. Bon GG, Kenemans P, Dekker JJ, Hompes PG, Verstraeten RA, van Kamp GJ, et al. Fluctuations in CA 125 and CA 15-3 serum concentrations during spontaneous ovulatory cycles. *Hum Reprod.* 1999;14(2):566-570.
28. Bischof P, Galfetti MA, Seydoux J, von Hospenthal JU, Campana A. Peripheral CA 125 levels in patients with uterine fibroids. *Hum Reprod.* 1992;7(1):35-38.
29. D'Hooghe TM, Bambra CS, Raeymaekers BM, Koninckx PR. Increased prevalence and recurrence of retrograde menstruation in baboons with spontaneous endometriosis. *Hum Reprod.* 1996;11(9):2022-2025.
30. Muyldermans M, Cornillie FJ, Koninckx PR. CA125 and endometriosis. *Hum Reprod Update.* 1995;1(2):173-187.
31. Mol BW, Bayram N, Lijmer JG, Wiegerinck MA, Bongers MY, van der Veen F, et al. The performance of CA-125 measurement in the detection of endometriosis: a meta-analysis. *Fertil Steril.* 1998;70(6):1101-1108.

Tables

Table 1. Association between predictors and CA125 levels in premenopausal women without ovarian cancer in the New England Case Control Study (n=815)

Variables	N (%)	Mean CA125 (95%CI)[a]	Differences in CA125 levels[b]	p-value[b]
Age, years				
< 30	69 (8)	13 (11,14)	-20%	0.001
30- 39	215 (26)	16 (15, 17)	ref	
40- 49	374 (46)	15 (15, 16)	-5%	0.20
50+	157 (19)	15 (13, 16)	-10%	0.05
P-trend				0.65
Race				
white	778 (95)	15 (15, 16)	ref	
non-white	37 (5)	14 (12, 16)	-10%	0.20
BMI, kg/m^2				
< 20	76 (9)	15 (14, 17)	1%	0.90
20- < 25	414 (51)	15 (14, 16)	ref	
25- < 30	206 (25)	15 (14, 16)	-1%	0.77
30- < 35	84 (10)	16 (14, 17)	3%	0.6
35+	35 (4)	17 (14, 20)	13%	0.16
P-trend				0.30
Height, cm				
< 160	182 (22)	16 (14, 17)	ref	
160- < 165	218 (27)	14 (13, 15)	-8%	0.08
165- < 170	231 (28)	16 (15, 17)	1%	0.85
170- < 175	128 (16)	15 (14, 17)	-1%	0.84
175+	56 (7)	14 (12, 16)	-8%	0.25
P-trend				0.88
Smoking				
never	421 (52)	15 (14, 16)	ref	
former	269 (33)	16 (15, 17)	5%	0.19
current	125 (15)	14 (13, 15)	-7%	0.16
Smoking status and duration				
never smokers	421 (52)	15 (14, 16)	ref	
< 5 packyears among former	140 (17)	16 (15, 17)	5%	0.30
5- < 15 packyears among former	79 (10)	16 (15, 18)	8%	0.20
15+ packyears among former	50 (6)	15 (13, 17)	1%	0.92
< 5 packyears among current	23 (3)	15 (12, 19)	1%	0.93
5- < 15 packyears among current	46 (6)	14 (12, 16)	-6%	0.47
15+ packyears among current	56 (7)	13 (12, 15)	-11%	0.11
Age at menarche, years				

≤ 12	386 (47)	16 (15, 16)	ref	
13+	428 (53)	15 (14, 15)	-6%	0.09
Oral contraceptive use				
Never	180 (22)	15 (14, 16)	ref	
Ever	635 (78)	15 (15, 16)	5%	0.25
Duration of oral contraceptive use among ever users, years				
< 2	140 (22)	17 (16, 19)	ref	
2 - 3	133 (21)	15 (14, 17)	-10%	0.07
4 - 5	103 (16)	15 (14, 17)	-11%	0.06
6 - 9	125 (20)	15 (13, 16)	-14%	0.01
10+	134 (21)	14 (13, 16)	-16%	0.005
P-trend				0.01
Current hormonal contraception use[c]				
no	427 (82)	15 (15, 16)	ref	
yes	94 (18)	13 (12, 14)	-16%	0.003
Menstrual phase at time of blood draw (days)				
early follicular (0 - 7)	126 (15)	17 (15, 18)	ref	
late follicular (8 - 11)	56 (7)	16 (14, 18)	-8%	0.31
peri-ovulatory (12 - 16)	73 (9)	15 (13, 16)	-13%	0.06
luteal (17 - 35)	155 (19)	14 (13, 16)	-14%	0.01
long cycle (36+)	63 (8)	13 (12, 15)	-21%	0.002
irregular	65 (8)	13 (12, 15)	-21%	0.002
missing	277 (34)	16 (15, 17)	-8%	0.15
Cause of infertility				
none	649 (80)	15 (15, 16)	ref	
male factor	20 (2)	17 (14, 21)	12%	0.32
tubal	13 (2)	15 (11, 20)	-1%	0.92
endometriosis	13 (2)	23 (17, 30)	50%	0.004
ovulatory	14 (2)	14 (11, 18)	-7%	0.59
unknown	106 (13)	14 (13, 15)	-9%	0.08
Number of miscarriages				
0	625 (77)	15 (15, 16)	ref	
1	137 (17)	15 (13, 16)	-5%	0.29
2+	53 (7)	14 (13, 17)	-6%	0.41
Ectopic pregnancy				
no	795 (98)	15 (15, 16)	ref	
yes	20 (2)	16 (13, 20)	5%	0.67
Parity				
0	214 (26)	15 (14, 16)	ref	
1	140 (17)	15 (14, 16)	1%	0.82
2	264 (32)	15 (14, 16)	3%	0.53

3+	197 (24)	15 (14, 16)	2%	0.66
Tubal ligation				
no	674 (83)	15 (15, 16)	ref	
yes	141 (17)	14 (13, 15)	-11%	0.01
Intrauterine device use				
never	689 (85)	15 (14, 16)	ref	
ever	126 (15)	16 (15, 17)	6%	0.26
Unilateral oophorectomy				
no	808 (99)	15 (15, 16)	ref	
yes	7 (1)	19 (13, 27)	23%	0.28
Endometriosis				
no	764 (94)	15 (14, 15)	ref	
yes	51 (6)	18 (16, 21)	21%	0.01
Fibroids				
no	737 (90)	15 (14,16)	ref	
yes	78 (10)	17 (15,19)	13%	0.04
Prior personal cancer diagnosis				
no	778(95)	15 (15, 16)	ref	
yes	37(4)	17 (15, 20)	15%	0.10
Family history of ovarian cancer				
no	795 (98)	15 (15, 16)	ref	
yes	20 (2)	17 (13, 21)	10%	0.40
Genital powder use				
no use	484 (59)	15 (15, 16)	ref	
body use	157 (19)	15 (14, 16)	-5%	0.30
genital use	174 (21)	15 (14, 16)	-3%	0.57
Caffeine intake, mg				
< 70.1	198 (25)	16 (14, 17)	ref	
70.1- < 169.6	198 (25)	14 (13, 15)	-9%	0.07
169.6- < 348.7	198 (25)	15 (14, 17)	0%	0.97
348.7+	198 (25)	16 (15, 17)	1%	0.90
P-trend				0.44

[a]geometric mean adjusted for age

[b]age-adjusted

[c] includes oral contraceptives and injections

Table 2. Linear CA125 prediction model in premenopausal women using stepwise regression in the New England Case Control Study (NEC)

Selected Predictors	Development of model in NEC (n=768)[a b]			Abridged model in NEC (n=768)[a b]		
	differences in CA125 levels	SE	*p*-value	differences in CA125 levels	SE	*p*-value
Age, years						
< 30	-18%	0.07	0.01	-19%	0.07	0.004
30- < 40	ref			ref		
40- < 50	-8%	0.04	0.06	-8%	0.04	0.05
50+	-10%	0.06	0.06	-11%	0.06	0.04
Non-white race	-13%	0.09	0.10	-14%	0.09	0.08
Tubal ligation	-10%	0.05	0.03			
Endometriosis	21%	0.07	0.01			
Menstrual phase at time of blood draw						
early follicular phase	ref			ref		
on hormonal contraceptives[c]	-28%	0.07	<.0001	-27%	0.07	<.0001
other cycle phase	-17%	0.06	0.002	-17%	0.06	0.001
irregular cycles	-30%	0.11	0.001	-30%	0.11	0.001
missing	-17%	0.10	0.05	-18%	0.10	0.04
Fibroids	13%	0.06	0.04			
R-square[d]	0.07			0.05		

[a]adjusted for study phase and center

[b]exclude missing age at menarche(n=1), caffeine intake(n=23), and menstrual cycle length(n=23)

[c] includes oral contraceptives and injections

[d]delta r-square, subtracting the effect of center and study phase

612

Table 3. Dichotomous CA125 prediction model of high CA125 (≥ 35 U/ mL) in premenopausal women using stepwise regression in the New England Case Control Study (NEC)

Selected predictors	Development of model in NEC(n=768) [a b]		Abridged model in NEC(n=768)[a b]	
	Odds ratio	95%CI	Odds ratio	95%CI
Age, years				
< 30	0.34	0.07, 1.61	0.29	0.06, 1.30
30- < 40	ref		ref	
40- < 50	0.34	0.15, 0.74	0.40	0.20, 0.82
50+	0.37	0.13, 1.04	0.44	0.17, 1.12
Tubal ligation	0.16	0.03, 0.73		
Endometriosis	3.08	1.09, 8.72		
Prior personal cancer diagnosis	5.41	1.70, 17.18		
Family history of ovarian cancer	13.32	3.42, 51.93		
Number of miscarriages				
0	ref		ref	
1	0.23	0.06, 0.89	0.36	0.11, 1.21
2+	0.18	0.02, 1.56	0.29	0.04, 2.18
Menstrual phase at time of blood draw				
regular menstruation + blood draw at early follicular phase	ref		ref	
on hormonal contraceptives [c]	0.07	0.01, 0.63		
regular menstruation + blood draw at other phase /irregular menstruation	0.49	0.19, 1.25	0.35	0.15, 0.82
missing time at blood draw	0.62	0.08, 4.70	0.35	0.05, 2.45
Smoking status				
Never	ref		ref	
Former	2.15	0.97, 4.79	1.85	0.96, 3.57
Current	0.12	0.01, 1.51	0.47	0.13, 1.64
Pack-years among former smokers	0.99	0.94, 1.05		
Pack-years among current smokers	1.09	1.00, 1.18		
AUC (95%CI)	0.83	0.77, 0.89	0.73	0.65, 0.81

[a]adjusted for study phase and center

[b]exclude missing age at menarche(n=1), caffeine intake(n=23), and menstrual cycle length(n=23)

[c] includes oral contraceptives and injections

613

Figure Legend

Figure 1. Study design of the development and validation of the CA125 prediction model using the New England Case Control Study (NEC) and European Prospective Investigation into Cancer and Nutrition Study (EPIC)

We developed the CA125 prediction models (linear and dichotomous) using the New England Case Control Study (NEC) and conducted external validation using the European Prospective Investigation into Cancer and Nutrition Study (EPIC).

Figure 2. Development and validation of linear CA125 prediction model in the New England Case Control Study (NEC) and European Prospective Investigation into Cancer and Nutrition Study (EPIC)

The predicted versus the observed log-transformed CA125 values were plotted and Pearson correlation coefficient (r) was calculated to assess the performance of the linear CA125 prediction model in the New England Case Control Study (NEC) and European Prospective Investigation into Cancer and Nutrition Study (EPIC). **A**, Linear CA125 prediction model performance in NEC. **B**, Abridged linear CA125 prediction model performance in NEC. **C**, Abridged linear CA125 prediction model performance in EPIC.

Figure 3. Development and validation of dichotomous CA125 prediction model the New England Case Control Study (NEC) and European Prospective Investigation into Cancer and Nutrition Study (EPIC)

Receiver Operating Characteristic (ROC) curves were plotted and the Area Under the Curve (AUC) was calculated to assess the discriminatory performance of the dichotomous CA125 prediction model in the New England Case Control Study (NEC) and European Prospective Investigation into Cancer and Nutrition Study (EPIC). Dichotomous CA125 prediction model performance in NEC (solid line), abridged dichotomous CA125 prediction model performance in

NEC (dashed line), abridged dichotomous CA125 prediction model performance in EPIC (dotted line).