Risk prediction models for endometrial cancer: development and validation in an international consortium

Joy Shi [iD], PhD,[1,*] Peter Kraft [iD], PhD,[1,2] Bernard A. Rosner [iD], PhD,[2,3] Yolanda Benavente, MStat,[4,5] Amanda Black, PhD,[6] Louise A. Brinton, PhD,[6] Chu Chen, PhD,[7] Megan A. Clarke [iD], PhD,[6] Linda S. Cook, PhD,[8,9] Laura Costas [iD], MD, PhD,[4,5] Luigino Dal Maso [iD], PhD,[10] Jo L. Freudenheim, PhD,[11] Jon Frias-Gomez, MStat,[4,12] Christine M. Friedenreich, PhD,[9] Montserrat Garcia-Closas, MD, DrPH,[6] Marc T. Goodman, PhD,[13] Lisa Johnson, PhD,[7] Carlo La Vecchia, MD,[14] Fabio Levi, MD,[15] Jolanta Lissowska [iD], PhD,[16] Lingeng Lu [iD], MD, PhD,[17] Susan E. McCann [iD], PhD, RD,[18,†] Kirsten B. Moysich, PhD,[18] Eva Negri, ScD,[14,19] Kelli O'Connell, MSPHc,[20] Fabio Parazzini [iD], MD,[14] Stacey Petruzella, MPH,[20] Jerry Polesel, ScD,[10] Jeanette Ponte, MPH,[20] Timothy R. Rebbeck [iD], PhD,[1,21] Peggy Reynolds, PhD,[22] Fulvio Ricceri [iD], PhD,[23] Harvey A. Risch, MD, PhD,[17] Carlotta Sacerdote [iD], PhD,[24] Veronica W. Setiawan, PhD,[25] Xiao-Ou Shu [iD], MD, PhD,[26] Amanda B. Spurdle, PhD,[27,28] Britton Trabert [iD], PhD,[6,29] Penelope M. Webb, DPhil,[27] Nicolas Wentzensen [iD], MD, PhD,[6] Lynne R. Wilkens, DrPH,[30] Wang Hong Xu, MD,[31] Hannah P. Yang, PhD,[6] Herbert Yu, MD,[30] Mengmeng Du, ScD,[20,‡] Immaculata De Vivo, PhD[1,3,32,‡]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[4]Cancer Epidemiology Research Programme, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute, Barcelona, Spain
[5]Consortium for Biomedical Research in Epidemiology and Public Health (CIBER Epidemiología y Salud Pública, CIBERESP), Madrid, Spain
[6]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA
[7]Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA
[8]Department of Epidemiology, Colorado School of Public Heath, University of Colorado-Anschutz, Aurora, CO, USA
[9]Department of Cancer Epidemiology and Prevention Research, Alberta Health Services, Calgary, AB, Canada
[10]Cancer Epidemiology Unit, Centro di Riferimento Oncologico di Aviano (CRO), Aviano, Italy
[11]Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, The State University of New York at Buffalo, Buffalo, NY, USA
[12]Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain
[13]Community and Population Health Research Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[14]Department of Clinical Medicine and Community Health, Università degli Studi di Milano, Milan, Italy
[15]Department of Epidemiology and Health Services Research, Centre for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland
[16]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland
[17]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA
[18]Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA
[19]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
[20]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[21]Division of Population Science, Dana-Farber Cancer Institute, Boston, MA, USA
[22]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA
[23]Department of Clinical and Biological Sciences, University of Turin, Orbassano, Italy
[24]Unit of Cancer Epidemiology, Città della Salute e della Scienza University-Hospital and Center for Cancer Prevention (CPO), Turin, Italy
[25]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[26]Division of Epidemiology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
[27]Population Health Department, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia
[28]Genetics and Computational Biology Department, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia
[29]Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, UT, USA
[30]University of Hawaii Cancer Center, Honolulu, HI, USA
[31]Department of Epidemiology, Fudan University School of Public Health, Shanghai, China
[32]Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA, USA

*Correspondence to: Joy Shi, PhD, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA (e-mail: joyshi@hsph.harvard.edu).
†Deceased May 5, 2022.
‡Authors contributed equally to this work and are considered co-senior authors.

Abstract

Background: Endometrial cancer risk stratification may help target interventions, screening, or prophylactic hysterectomy to mitigate the rising burden of this cancer. However, existing prediction models have been developed in select cohorts and have not considered genetic factors.

Methods: We developed endometrial cancer risk prediction models using data on postmenopausal White women aged 45-85 years from 19 case-control studies in the Epidemiology of Endometrial Cancer Consortium (E2C2). Relative risk estimates for predictors were combined with age-specific endometrial cancer incidence rates and estimates for the underlying risk factor distribution. We externally validated the models in 3 cohorts: Nurses' Health Study (NHS), NHS II, and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial.

Results: Area under the receiver operating characteristic curves for the epidemiologic model ranged from 0.64 (95% confidence interval [CI] = 0.62 to 0.67) to 0.69 (95% CI = 0.66 to 0.72). Improvements in discrimination from the addition of genetic factors were modest (no change in area under the receiver operating characteristic curves in NHS; PLCO = 0.64 to 0.66). The epidemiologic model was well calibrated in NHS II (overall expected-to-observed ratio [E/O] = 1.09, 95% CI = 0.98 to 1.22) and PLCO (overall E/O = 1.04, 95% CI = 0.95 to 1.13) but poorly calibrated in NHS (overall E/O = 0.55, 95% CI = 0.51 to 0.59).

Conclusions: Using data from the largest, most heterogeneous study population to date (to our knowledge), prediction models based on epidemiologic factors alone successfully identified women at high risk of endometrial cancer. Genetic factors offered limited improvements in discrimination. Further work is needed to refine this tool for clinical or public health practice and expand these models to multiethnic populations.

Endometrial cancer is the fourth-most commonly diagnosed cancer among women and the most common gynecological malignancy in the United States (1). Its incidence and mortality have been increasing since 2000 and are expected to continue rising with increasing prevalence of endometrial cancer risk factors (2). Clinical practice guidelines currently do not recommend screening for endometrial cancer in the general population. However, among individuals with high endometrial cancer risk (eg, those with Lynch syndrome), the benefits of screening may outweigh the discomfort and risks—driving existing recommendations for annual endometrial biopsy screening in Lynch-positive individuals (3). Building on this framework, a predictive model translated into research and eventual clinical settings could identify high-risk individuals for enrolment in screening and prevention trials and for whom benefits of risk-reducing interventions may outweigh risks.

Previously developed risk models for endometrial cancer have used pooled data from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial and the National Institutes of Health-American Association of Retired Persons Diet and Health Study (NIH-AARP) (4), or data from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort (5). Predictors included hormonal-related factors (eg, parity, oral contraceptive [OC] use, hormone therapy [HT] use), smoking, and body mass index (BMI) (4,5). However, these models were trained on selective study populations, which may limit generalizability. Moreover, contributions of genetic factors in improving endometrial cancer risk discrimination have not, to our knowledge, been assessed. We used training data from the largest, most heterogeneous study population to date (to our knowledge), and external testing data from 3 large longitudinal cohorts to achieve the following objectives: (1) develop and evaluate absolute risk models for endometrial cancer using epidemiologic questionnaire data that can serve as a suitable framework for the eventual development of a risk prediction tool for clinical or public health practice; and (2) assess the predictive contributions of genetic data to this prediction model.

Methods
Data for model development
Study population
Data from 19 case-control studies in the Epidemiology of Endometrial Cancer Consortium (E2C2) were used for model development (Supplementary Table 1, available online). E2C2 included participants from the United States, Canada, Europe, China, and Australia (6-8). The current analysis was restricted to postmenopausal White women aged 45-85 years. Informed consent was obtained from all study participants by the original studies as per the requirements of each study's institutional review board (IRB).

Data collection and case definition
Data from individual studies were received and harmonized at Memorial Sloan Kettering Cancer Center (6,7). Case-control data were collected based on a specific reference date, generally 6-12 months before the date of diagnosis for cases and at the date of the interview for controls. Cases were defined as incident cases of endometrial cancer.

We considered the following variables for inclusion in the model based on previously identified risk factors: education, smoking status, BMI, parity, age at first birth, age at menarche, any HT use, any estrogen-only HT use, duration of estrogen-only HT use, any combination of estrogen and progestin (E+P) HT use, duration of E+P HT use, any OC use, duration of OC use, history of diabetes, and history of hypertension. Data availability varied by study site (Supplementary Table 2, available online), and parameterization of these risk factors is described in the Supplementary Methods (available online).

Data for model validation
Study populations
Data from the Nurses' Health Study (NHS), NHS II, and PLCO were used for model validation (9,10). For NHS, 121 700 female registered nurses aged 30-55 years were enrolled in 1976. For NHS II, 116 430 female registered nurses aged 25-42 years were enrolled in 1989. In both cohorts, biennial questionnaires were administered to collect updated information on risk factors and incident health outcomes. The study protocol was approved by IRBs of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health and those of participating registries as required. For PLCO, 78 232 women aged 55-74 years were enrolled between 1993 and 2001 across 10 screening centers. A self-administered lifestyle questionnaire was completed by all participants at baseline. A supplemental questionnaire was completed by a subset (n = 32 434) of participants in 2006. Each institution obtained annual approval from its IRB, and all participants provided written informed consent. Inclusion and exclusion criteria for the current analysis are further described in the Supplementary Methods and Supplementary Figures 1-3 (available online).

Data collection

For NHS and NHS II, data on risk factors were obtained from the first biennial questionnaire completed after the participant met all eligibility criteria. In NHS, 32 826 blood samples and 29 684 buccal cell samples have been collected since 1989. Data from 12 different genome-wide association studies of various disease outcomes were pooled to form a subsample of NHS participants for whom genotyping data were available (11). Incident endometrial cancer cases were identified via questionnaire or death records. With permission of the participant or next of kin, medical records were obtained, and the cancer diagnosis was confirmed by study pathologists and/or physicians.

For PLCO, risk factor data were obtained from the baseline questionnaire, except for HT use, which was collected on the supplemental questionnaire. Blood samples were collected from participants at enrollment and at annual screening visits. Endometrial cancer cases were identified through annual study updates or via death records and confirmed by medical records. In all validation cohorts, participants were followed up for 10 years or until they were lost to follow-up. In sensitivity analyses, we censored participants upon experiencing a competing event (other cancers, hysterectomy, or death).

Statistical analysis

Model development and validation

To develop absolute risk prediction models for endometrial cancer, the Individualized Coherent Absolute Risk Estimation package in R was used (12). We developed epidemiologic models, which included only questionnaire data, and epidemiologic plus genetic models, which additionally included 18 previously identified genome-wide statistically significant single nucleotide polymorphisms for endometrial cancer (Supplementary Table 3, available online) (13). Three data sources were used to build these models: (1) the log relative risk parameters for model predictors, which were approximated using log odds ratios from a group least absolute shrinkage and selection operator (LASSO) penalized logistic regression model applied to pooled data from E2C2; (2) marginal age-specific incidence rates for endometrial cancer and its competing risks (ie, other cancers, hysterectomy, and death; Supplementary Table 4, available online); and (3) data from the National Health and Nutrition Examination Survey (NHANES) to serve as a reference dataset to estimate the risk factor distribution for the underlying population (Supplementary Table 5, available online).

In the validation cohorts, we assessed discrimination of 10-year endometrial cancer risks (based on model predictors and marginal age-specific incidence rates) using area under the receiver operating characteristic curves (AUC) estimates. For calibration, we used goodness-of-fit tests to compare predicted and observed 10-year relative risks, and the expected-to-observed (E/O) ratio and the Hosmer-Lemeshow test to compare 10-year absolute risks. Additional details are available in the Supplementary Methods (available online).

Estimating risks among the general US population

To estimate absolute risks of endometrial cancer among a more current general US population of White women, we developed a prediction model that incorporated more current data on marginal age-specific incidence rates for endometrial cancer and its competing risks, and the underlying risk factor distribution. We assessed cumulative absolute risks by categories of risk percentiles. Additional details are available in the Supplementary Methods (available online).

Results

Our training data included 6665 endometrial cancer cases and 9062 controls (Supplementary Figure 4; Supplementary Table 6, available online). Our final LASSO model retained all main effects included in the initial model except for duration of E+P HT use; the only interaction terms that remained were those between any HT use and BMI (Supplementary Table 5, available online).

In our validation cohorts, 700 of the 68 150 NHS participants, 304 of the 56 076 NHS II participants, and 511 of the 39 996 PLCO participants developed endometrial cancer during follow-up (Supplementary Figures 1-3; Supplementary Table 7, available online). On average, cumulative risk of endometrial cancer between ages 45 and 85 years was 5.4% (Supplementary Figures 5-6, available online). However, women in the highest decile of risk were predicted to experience a cumulative risk of 13.7%-15.0%, whereas women in the lowest decile were predicted to experience a cumulative risk of 1.4%-1.8%.

AUCs for the 10-year risk of endometrial cancer based on the epidemiologic models were 0.65 (95% confidence interval [CI] = 0.63 to 0.67) for NHS, 0.69 (95% CI = 0.66 to 0.72) for NHS II, and 0.64 (95% CI = 0.62 to 0.67) for PLCO (Table 1). Among participants for whom genetic data were available, the addition of genetic factors did not change the AUC in NHS (AUC = 0.61, 95% CI = 0.57 to 0.66) but improved the AUC from 0.64 (95% CI = 0.61 to 0.67) to 0.66 (95% CI = 0.63 to 0.69) in PLCO. Age-specific AUCs (within 10-year age bands) were similar to overall AUCs (Supplementary Table 8, available online). The epidemiologic model underpredicted the number of events in the NHS cohort (E/O = 0.55, 95% CI = 0.51 to 0.59) but more accurately predicted the overall number of events in NHS II (E/O = 1.09, 95% CI = 0.98 to 1.22) and PLCO (E/O = 1.04, 95% CI = 0.95 to 1.13) (Figures 1 and 2; Supplementary Tables 9 and 10, available online). The epidemiologic plus genetic model was well-calibrated in the PLCO cohort (E/O = 0.94, 95% CI = 0.85 to 1.03) (Figure 2; Supplementary Table 10, available online). In sensitivity analyses, calibration of the models in which individuals were censored upon development of a competing event were similar to our main findings (Supplementary Tables 11 and 12, available online).

When the risk model was applied to a more recent and representative US population of White women, the model identified 2.5% of women with at least 20 percentage cumulative risk of endometrial cancer between ages 40 and 85 years (Figure 3).

Discussion

We developed risk prediction models for endometrial cancer using pooled data for White women from an international consortium and externally validated the models in 3 large United States-based cohorts. The epidemiologic model demonstrated moderate discriminatory accuracy (AUCs ranging 0.64 to 0.69) with only modest improvements when genetic factors were included. The epidemiologic model was well-calibrated in NHS II (E/O = 1.09) but poorly calibrated in NHS (E/O = 0.55) for 10-year risk of endometrial cancer. Both the epidemiologic and the epidemiologic plus genetic models were well-calibrated in PLCO (E/O = 1.04 and E/O = 0.94, respectively).

Currently, there is no evidence to support endometrial cancer screening of asymptomatic average-risk menopausal women (3).

Table 1. Discrimination of epidemiologic and E+G prediction models for endometrial cancer risk[a]

Validation cohort	Population	Prediction model	No. of participants	No. of events	AUC (95% CI)
NHS	Full cohort	Epidemiologic	68 150	700	0.647 (0.626 to 0.667)
	Genetic cohort	Epidemiologic	11 365	166	0.613 (0.570 to 0.656)
	Genetic cohort	E+G	11 365	166	0.613 (0.570 to 0.656)
NHS II	Full cohort	Epidemiologic	56 076	304	0.693 (0.664 to 0.723)
PLCO	Full cohort	Epidemiologic	39 996	511	0.640 (0.615 to 0.665)
	Genetic cohort	Epidemiologic	30 102	401	0.635 (0.606 to 0.664)
	Genetic cohort	E+G	30 102	401	0.665 (0.636 to 0.693)

[a] AUC = area under the receiver operating characteristic curve; CI = confidence interval; E+G = epidemiologic and genetic; NHS = Nurses' Health Study; PLCO = Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.
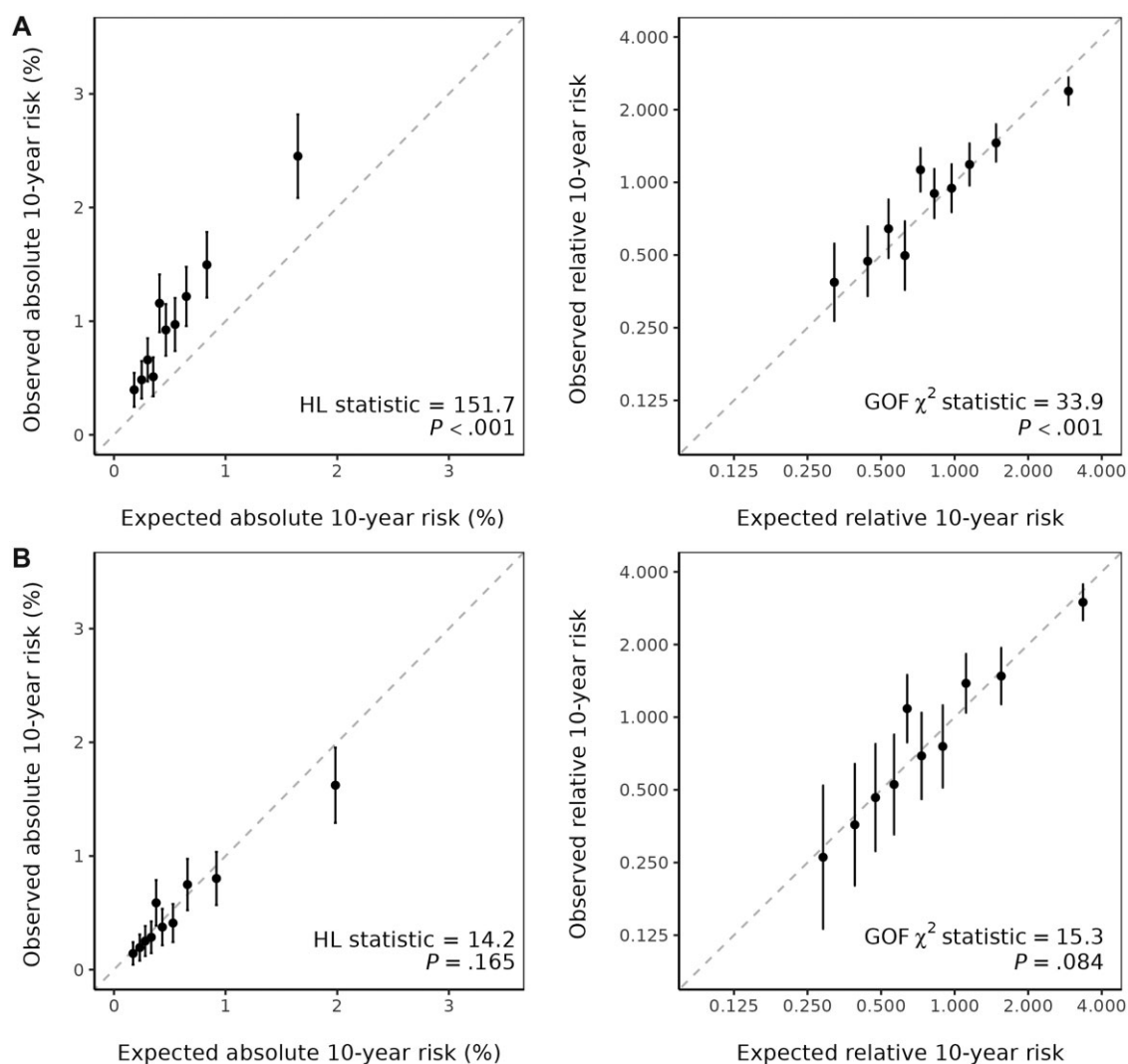
Figure 1. Absolute and relative 10-year risk calibration of the epidemiologic model in the (**A**) Nurses' Health Study and (**B**) Nurses' Health Study II. The **dots** represent the estimates for the decile-specific expected-to-observed ratios, and the **error bars** represent the 95% confidence intervals. Test statistics and P values correspond to results from Hosmer-Lemeshow (HL) and goodness-of-fit (GOF) tests.

However, women with or at risk for Lynch syndrome have a lifetime endometrial cancer risk of 22%-50% and are recommended to undergo annual screening with an endometrial biopsy (14-17). When the risk model was applied to a more recent and representative US population of White women, those in the 97th percentile of risk had a predicted lifetime risk of at least 20% for endometrial cancer—a lifetime risk within range of that for women with Lynch syndrome. Although not prospectively evaluated for calibration, this model suggests that epidemiologic factors alone are sufficient to identify high-risk women for enrollment in prevention or screening trials of endometrial cancer. Moreover, understanding individual risks can help personalize clinical management of endometrial cancer.

To our knowledge, our study is the first to assess the utility of genetic factors in endometrial cancer risk prediction. In contrast to other cancers (eg, breast) (18), we observed only slight
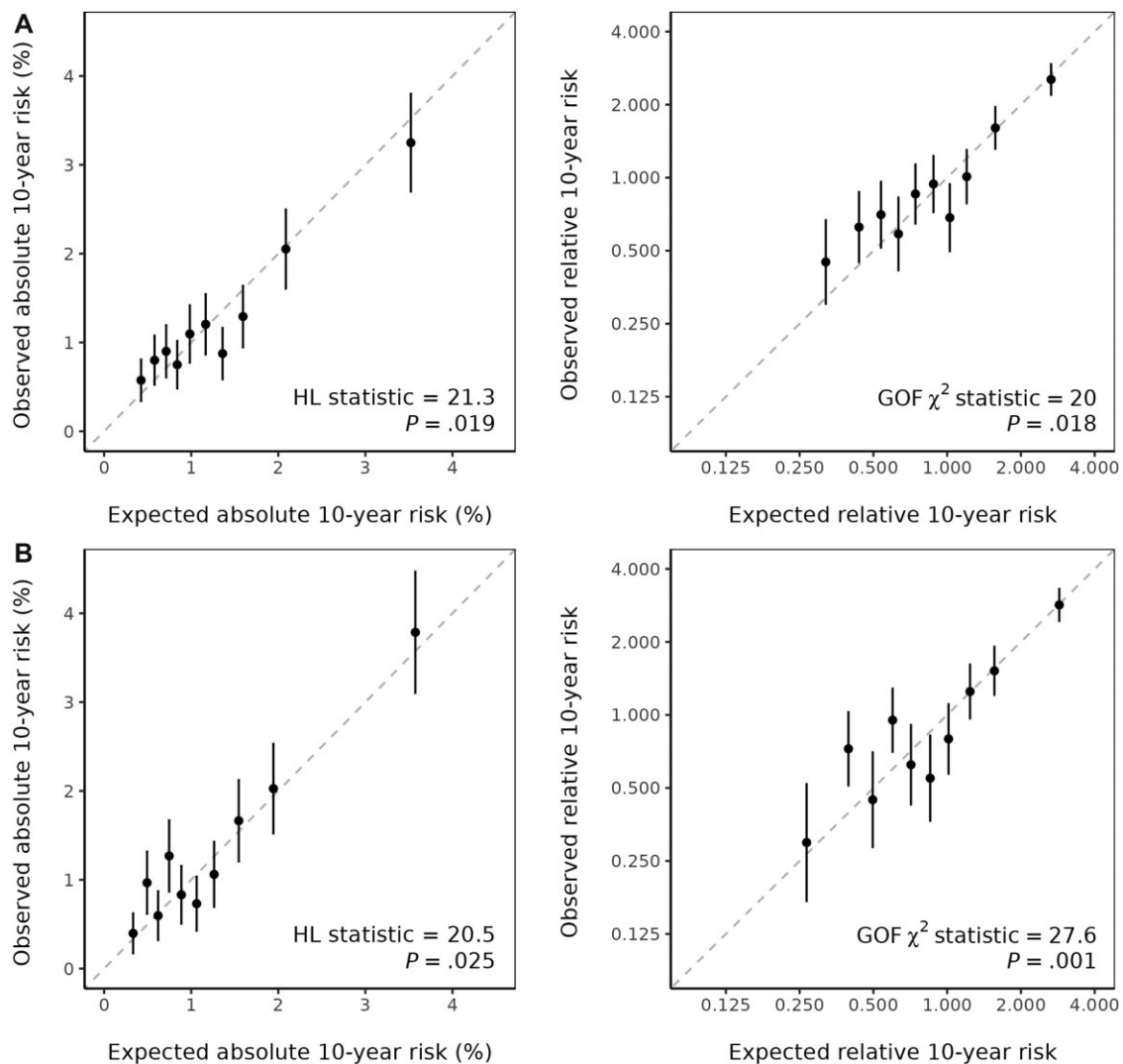
Figure 2. Absolute and relative 10-year risk calibration of the (**A**) epidemiologic model and the (**B**) epidemiologic plus genetic model in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. The **dots** represent the estimates for the decile-specific expected-to-observed (E/O) ratios, and the **error bars** represent the 95% confidence intervals. Test statistics and *P* values correspond to results from Hosmer-Lemeshow (HL) and goodness-of-fit (GOF) tests.

improvements with the addition of published genetic factors. This likely reflects the limited number of endometrial cancer susceptibility loci to date compared with cancers like breast, which has hundreds (19). Our results suggest collecting genetic data in clinical settings for endometrial cancer risk stratification is unwarranted until additional susceptibility loci are discovered. However, increasingly widespread direct-to-consumer genetic testing should motivate future studies of genetic factors in risk prediction as more endometrial cancer loci are uncovered.

Rather than develop a single absolute risk model, we developed and evaluated multiple absolute risk models to allow for flexibility in calibrating to the target population of interest. This allowed us to account for real-world differences between populations and changes over time. Model discrimination and relative risk calibration were similar across the validation cohorts, suggesting that our relative risk estimates could be implemented in future risk prediction tools for clinical or public health application. In contrast, absolute risk calibration varied, which may be due to the choice of reference dataset. The reference dataset is used to estimate the underlying risk factor distributions, which is needed to obtain accurate estimates of the baseline endometrial

cancer incidence rates from the marginal incidence rates inputted into the model. Risk factor distributions in NHANES may not have appropriately reflected the underlying risk factor distribution for NHS. The median year for start of follow-up was 1986 in NHS. In contrast, the earliest NHANES cycle that collected all relevant endometrial cancer risk factors was 1999-2000, which was unlikely to reflect the true underlying distribution of risk factors for women in 1986. For NHS II and PLCO, the median year for start of follow-up (2007 for NHS II and 1997 for PLCO) was more congruent with the timing of their corresponding NHANES cycles (2007-2008 and 1999-2000, respectively). Our findings underlie the importance of using an appropriate reference dataset when calibrating absolute risks.

Two absolute risk prediction models for endometrial cancer were previously developed (4,5). Our risk prediction model differs from previous models in 3 aspects. First, we used LASSO for variable selection; in contrast, previously developed models used stepwise approaches, which favor more parsimonious models (20,21). As such, our model included additional risk factors, such as education, E+P HT use, history of diabetes, and history of hypertension. Second, participants in our validation
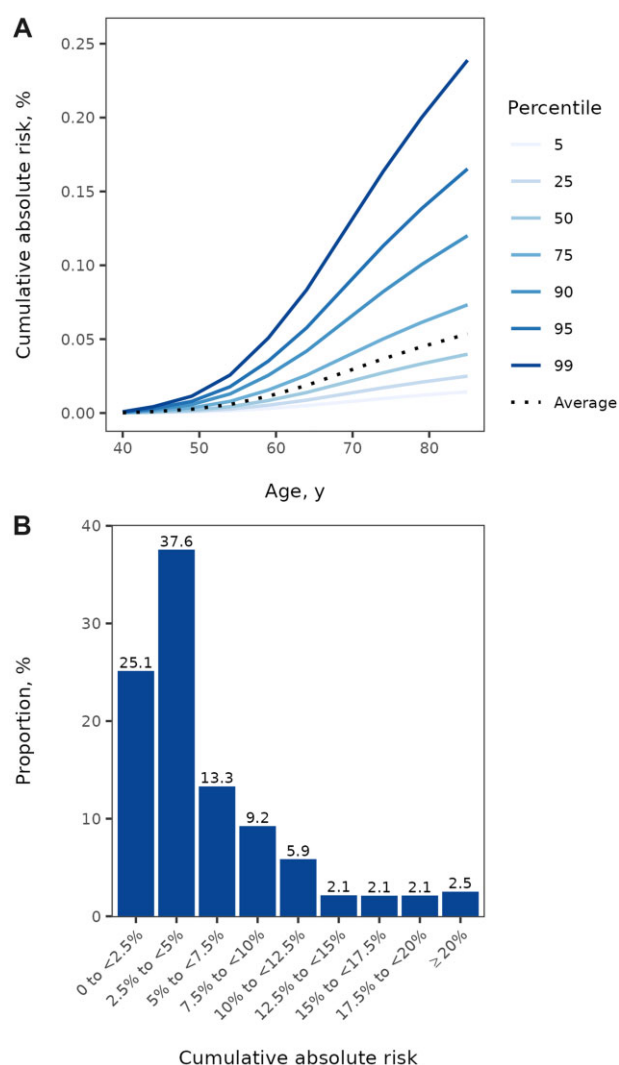
Figure 3. A) Estimated cumulative absolute risk stratified by risk deciles and (**B**) distribution of cumulative absolute risks in current US population of White women.

cohorts were less variable in baseline age (45-65 years for almost all NHS and NHS II participants, and 55-75 years for PLCO participants). In contrast, the 2.5th and 97.5th percentiles for participant age in EPIC ranged between 27 and 68 years. Given that age is one of the strongest risk factors for endometrial cancer, the discriminatory performance of a model can be driven largely by variability in participant age. This may explain why our AUCs, ranging from 0.64 to 0.69, are lower than those reported for the EPIC model in which age alone (AUC = 0.71) contributed to much of the discriminatory ability of their full model (AUC = 0.77) (5). Third, we used external data from a more generalizable population of White women (ie, NHANES) to estimate the underlying risk factor distribution. In contrast, previous models estimated these distributions using data on selective study populations. The PLCO/NIH-AARP model (AUC = 0.68) had similar discriminatory ability as our epidemiologic model but overpredicted the number of cases among NHS participants (E/O = 1.20). In comparison, our model underpredicted the number of cases among NHS participants (E/O = 0.55) and was not better calibrated than the PLCO/NIH-AARP model in this specific cohort (4). The PLCO/NIH-AARP model overpredicted the number of cases among EPIC participants (E/O = 2.4), which the

authors attributed to geographic differences between the cohorts (5). The EPIC model was internally validated using five-fold cross-validation but was not externally validated (5).

There are several limitations of our study to consider. First, availability of measured risk factors was of concern, especially in the reference dataset (NHANES). For example, we could not include family history of endometrial cancer in our model because these data were not collected in NHANES. Second, recall bias may differentially affect the measurement of risk factors in our case-control studies. However, previous analyses of E2C2 data have reported similar estimates for endometrial cancer risk factors between cohort and case-control studies, which mitigates some of this concern (22). Third, we had limited genetic data available for NHS participants. In the NHS, genetic data were pooled from 12 different case-control genome-wide association studies of various disease outcomes. As such, participants with available genetic data may not be representative of the broader NHS cohort, and matching on certain factors (eg, HT use) between cases and controls may explain the lower AUC observed in the genetic vs full cohort. Fourth, although our relative risk estimates could be adapted in future prediction tools, our proposed models will likely need to be recalibrated to obtain accurate absolute risks for endometrial cancer risk in new geographic, temporal, or population settings. Last, most endometrial cancer cases in our study were of the endometrioid subtype, which precluded us from conducting histologic subtype-specific analyses for rarer subtypes. In addition, the lack of racial diversity among participants in our training data (93% White) and validation cohorts (93%-95% White) precluded our ability to reliably evaluate the performance of our model or to generate and validate race-specific estimates in non-White women. As such, our models are most applicable to White individuals, who make up most (72%) of endometrial cancers in the United States (24). A crucial next step is to develop and validate risk prediction models in other racial groups, including Black women, who experience the highest endometrial cancer-related mortality (25). These efforts will need to account for differences between White and non-White women with respect to (1) the magnitudes of association between certain risk factors (eg, parity) and endometrial cancer (2,22) marginal incidence rates for endometrial cancer and its competing risks, and (3) underlying risk factors distributions. This will be a focus for future work in E2C2 as we strive to add multiethnic studies in our network.

Our study has many strengths. First, we used the largest, most heterogeneous population to date to develop an endometrial cancer risk prediction model, with data spanning 19 studies across different continents. The consortium provided individual-level data to standardize risk factor definitions. Second, we externally validated our model in 3 large cohorts. Specifically, we evaluated calibration by deciles of risk, which is relevant for risk-based prevention and screening of highest-risk individuals. Third, we developed the model using population-based estimates for the marginal incidence rates for endometrial cancer and for competing risks, and the underlying risk factor distributions. As such, our model was well-calibrated to both NHS II and PLCO—2 studies with very different source populations—and is likely generalizable to the general US population of White women. Last, our study evaluated the contributions of genetic data to the performance of these prediction models.

In conclusion, we developed and validated absolute risk models for endometrial cancer among postmenopausal women aged at least 45 years in the largest, most heterogeneous study population to date. Our model including epidemiologic factors alone

was sufficient to identify women with risks comparable with that in Lynch patients, who are recommended annual screening; this may provide a basis for the eventual development of a risk prediction tool for clinical or public health practice. Future work is necessary to evaluate the accuracy of our models in multiethnic populations and specific histological subtypes.

Notes

Results (SEER) Program. Central registries may also be supported by state agencies, universities, and cancer centers. Participating central cancer registries include the following: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Indiana, Iowa, Kentucky, Louisiana, Massachusetts, Maine, Maryland, Michigan, Mississippi, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Puerto Rico, Rhode Island, Seattle SEER Registry, South Carolina, Tennessee, Texas, Utah, Virginia, West Virginia, Wyoming. The authors assume full responsibility for analyses and interpretation of these data.

Data availability

Data that support the findings of this study are not publicly available to protect participants' privacy and confidentiality. Further information including the procedures to obtain and access data from the Epidemiology of Endometrial Cancer Consortium is described at https://epi.grants.cancer.gov/eecc/membership.html. The analytic code is available at https://github.com/joy-shi1/endometrial-cancer-prediction.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;70(1):7-30. doi:10.3322/caac.21590.

2. Clarke MA, Devesa SS, Harvey SV, Wentzensen N. Hysterectomy-corrected uterine corpus cancer incidence trends and differences in relative survival reveal racial disparities and rising rates of nonendometrioid cancers. *J Clin Oncol*. 2019;37(22):1895-1908. doi:10.1200/JCO.19.00151.

3. Smith RA, von Eschenbach AC, Wender R, et al. American Cancer Society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers: also: update 2001—testing for early lung cancer detection. *CA Cancer J Clin*. 2001;51(1):38-75.

4. Pfeiffer RM, Park Y, Kreimer AR, et al. Risk prediction for breast, endometrial, and ovarian cancer in White women aged 50 y or older: derivation and validation from population-based cohort studies. *PLoS Med*. 2013;10(7):e1001492. doi:10.1371/journal.pmed.1001492.

5. Hüsing A, Dossus L, Ferrari P, et al. An epidemiological model for prediction of endometrial cancer risk in Europe. *Eur J Epidemiol*. 2016;31(1):51-60. doi:10.1007/s10654-015-0030-9.

6. Setiawan VW, Yang HP, Pike MC, et al. Type I and II endometrial cancers: have they different risk factors? *J Clin Oncol*. 2013;31(20):2607-2618. doi:10.1200/JCO.2012.48.2596.

7. Jordan SJ, Na R, Weiderpass E, et al. Pregnancy outcomes and risk of endometrial cancer: A pooled analysis of individual participant data in the Epidemiology of Endometrial Cancer Consortium. *Int J Cancer*. 2021;148(9):2068-2078. doi:10.1002/ijc.33360.

8. Olson SH, Chen C, De Vivo I, et al. Maximizing resources to study an uncommon cancer: E2C2–Epidemiology of Endometrial Cancer Consortium. *Cancer Causes Control*. 2009;20(4):491-496. doi:10.1007/s10552-008-9290-y.

9. Bao Y, Bertoia ML, Lenart EB, et al. Origin, methods, and evolution of the three Nurses' Health Studies. *Am J Public Health*. 2016;106(9):1573-1581. doi:10.2105/AJPH.2016.303338.

10. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Control Clin Trials*. 2000;21(suppl 6):273S-309S. http://www.ncbi.nlm.nih.gov/pubmed/11189684. Accessed October 17, 2018.

11. Lindström S, Loomis S, Turman C, et al. A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts. *PLoS One*. 2017;12(3):e0173997. doi:10.1371/journal.pone.0173997.

12. Pal Choudhury P, Maas P, Wilcox A, et al. iCARE: An R package to build, validate and apply absolute risk models. *PLoS One*. 2020;15(2):e0228198. doi:10.1371/journal.pone.0228198.

13. O'Mara TA, Glubb DM, Amant F, et al. Identification of nine new susceptibility loci for endometrial cancer. *Nat Commun*. 2018;9(1):3166. doi:10.1038/s41467-018-05427-7.

14. Watson P, Vasen HFA, Mecklin JP, Järvinen H, Lynch HT. The risk of endometrial cancer in hereditary nonpolyposis colorectal cancer. *Am J Med*. 1994;96(6):516-520. doi:10.1016/0002-9343(94)90091-4.

15. Aarnio M, Mecklin J -P, Aaltonen LA, Nyström-Lahti M, Järvinen HJ. Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *Int J Cancer*. 1995;64(6):430-433. doi:10.1002/ijc.2910640613.

16. Aarnio M, Sankila R, Pukkala E, et al. Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer*. 1999;81(2):214-218. doi:10.1002/(sici)1097-0215(19990412)81:2<214::aid-ijc8>3.0.co;2-l.

17. Win AK, Jenkins MA, Dowty JG, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 2017;26(3):404-412. doi:10.1158/1055-9965.EPI-16-0693.

18. Maas P, Barrdahl M, Joshi AD, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among White women in the United States. *JAMA Oncol*. 2016;2(10):1295-1302. doi:10.1001/JAMAONCOL.2016.1025.

19. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*. 2020;52(6):572-581. doi:10.1038/S41588-020-0609-2.

20. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results [published online ahead of print October 22, 2015]. *J Clin Epidemiol*. 2016;71:76-85. doi:10.1016/j.jclinepi.2015.10.002.

21. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. *Int J Med Inform*. 2018;116:10-17. doi:10.1016/j.ijmedinf.2018.05.006.

22. Cote ML, Alhajj T, Ruterbusch JJ, et al. Risk factors for endometrial cancer in Black and White women: a pooled analysis from the Epidemiology of Endometrial Cancer Consortium (E2C2). *Cancer Causes Control*. 2015;26(2):287-296. doi:10.1007/s10552-014-0510-3.

23. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in quantitative-trait association studies. *Genet Epidemiol*. 2011;35(3):133-138. doi:10.1002/gepi.20551.

24. Clarke MA, Devesa SS, Hammer A, Wentzensen N. Racial and ethnic differences in hysterectomy-corrected uterine corpus cancer mortality by stage and histologic subtype. *JAMA Oncol*. 2022;8(6):895-903. doi:10.1001/jamaoncol.2022.0009.

25. Allard JE, Maxwell GL. Race disparities between Black and White women in the incidence, treatment, and prognosis of endometrial cancer. *Cancer Control*. 2009;16(1):53-56. doi:10.1177/107327480901600108.