

Lophatananon, A., Usher-Smith, J., Campbell, J., Warcaba, J., Silarova, B., Waters, E.A., Colditz, G.A. and Muir, K.R. (2017) *Development of a cancer risk prediction tool for use in the UK primary care and community settings*. *Cancer Prevention Research*, 10 (7). pp. 421-430.

## Downloaded from

<https://kar.kent.ac.uk/98060/> The University of Kent's Academic Repository KAR

## The version of record is available from

<https://doi.org/10.1158/1940-6207.CAPR-16-0288>

## This document version

Publisher pdf

## DOI for this version

## Licence for this version

CC BY (Attribution)

## Additional information

cited By 4

## Versions of research works

### Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

### Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal*, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

## Enquiries

If you have questions about this document contact [ResearchSupport@kent.ac.uk](mailto:ResearchSupport@kent.ac.uk). Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

# Development of a Cancer Risk Prediction Tool for Use in the UK Primary Care and Community Settings



Artitaya Lophatananon<sup>1</sup>, Juliet Usher-Smith<sup>2</sup>, Jackie Campbell<sup>3</sup>, Joanne Warcaba<sup>4</sup>, Barbora Silarova<sup>5</sup>, Erika A. Waters<sup>6</sup>, Graham A. Colditz<sup>6</sup>, and Kenneth R. Muir<sup>1</sup>

## Abstract

Several multivariable risk prediction models have been developed to assess an individual's risk of developing specific cancers. Such models can be used in a variety of settings for prevention, screening, and guiding investigations and treatments. Models aimed at predicting future disease risk that contains lifestyle factors may be of particular use for targeting health promotion activities at an individual level. This type of cancer risk prediction is not yet available in the UK. We have adopted the approach used by the well-established U.S.-derived "YourCancerRisk" model for use in the UK population, which allow users to quantify their individual risk of developing individual cancers relative to the population average risk. The UK version of "YourCancerRisk" computes 10-year cancer risk estimates for 11 cancers utiliz-

ing UK figures for prevalence of risk factors and cancer incidence. Because the prevalence of risk factors and the incidence rates for cancer are different between the U.S. and the UK population, this UK model provides more accurate estimates of risks for a UK population. Using an example of breast cancer and data from UK Biobank cohort, we demonstrate that the individual risk factor estimates are similar for the U.S. and UK populations. Assessment of the performance and validation of the multivariate model predictions based on a binary score confirm the model's applicability. The model can be used to estimate absolute and relative cancer risk for use in Primary Care and community settings and is being used in the community to guide lifestyle change. *Cancer Prev Res*; 10(7); 421–30. ©2017 AACR.

## Introduction

Over recent years, there has been a growth in the development of risk prediction models for cancer and other diseases (1–6). These models provide a range of estimates of future risk of developing disease for applications in

prevention, screening, diagnosis, and treatment. Most of them are disease specific. In general, risk algorithms include phenotypic information such as sex, age, and lifestyle factors. Some algorithms may also allow for the incorporation of emerging "omics"-based factors and other biomarkers (7).

Few cancer risk prediction models have included the modifiable risk components such as physical activity, diet, and smoking. One such model that has been developed for the United States is the "YourCancerRisk" model which has been used as a tool for education as well as providing an approach to quantifying the effects of changing key lifestyle exposures. This was subsequently expanded into "YourDiseaseRisk" (8), which extends the range of endpoints to include 12 of the commonest types of cancers in the United States and 6 other chronic diseases (e.g., chronic bronchitis, stroke, emphysema, heart disease, diabetes, and osteoporosis). The model was validated for ovarian, colon, and pancreatic cancers in the United States in Nurses' Health Study and the U.S. Health Professionals cohorts. The results show the model to be well calibrated for ovarian and colon cancer in women and pancreatic cancer in men and moderately calibrated for colon cancer

<sup>1</sup>Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom.

<sup>2</sup>The Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom. <sup>3</sup>Institute of Health and Wellbeing, The University of Northampton, Northampton, United Kingdom. <sup>4</sup>Moulton Surgery, Northampton Lane, Northampton, United Kingdom. <sup>5</sup>MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom. <sup>6</sup>Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, Missouri.

**Corresponding Author:** Artitaya Lophatananon, The University of Manchester, Manchester M139PL, UK. Phone: 441612757227; Fax: +441612751617; E-mail: artitaya.lophananon@manchester.ac.uk

**doi:** 10.1158/1940-6207.CAPR-16-0288

©2017 American Association for Cancer Research.

in men. Discriminatory accuracy for pancreatic cancer showed a concordance index of 0.72, and for colon cancer in men and women, concordance indices were 0.71 and 0.67, respectively (9).

The "YourDiseaseRisk" model aims to predict the risks for individuals (aged 40 and above) of developing the 12 cancers relative to the general population. Uniquely, the approach adopted to develop such models involved extensive systematic reviews of existing studies and finding a consensus of expert opinions to identify risk factors and to summarize the level of evidence as "definite," "probable," and "possible" causes of cancer. Risk points were then allocated according to the strength of the causal association and summed. Population average risk of cancer and cumulative 10-year risk were obtained from the U.S. SEER data (10). Finally, individual ranking relative to the population average was determined.

The "YourDiseaseRisk" online tool has been available in the United States since 2000. It is offered as an educational tool, and in 2005, the site recorded 54 million hits with 6.2 million page views. It was first hosted at Harvard University and, in 2007, transitioned to the Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine (CITE Siteman/Wash U/BJH only; ref. 8).

The focus of this article is to describe the steps taken to adapt the "YourDiseaseRisk" models focused on cancers for the UK population for use in Primary Care and community-based settings. We also assess the utility of the approach by scoring the suggested risk factors in the UK Biobank cohort.

We have used the adapted UK version of the "YourDiseaseRisk" models (8) in a pilot study that used individual interviews to assess participants' understanding and preferences for how such information is offered (publication in press). These results will be presented elsewhere.

## Materials and Methods

### Cancer risk models development

We adapted the YourDiseaseRisk models for 11 cancers using the UK data. The 11 cancers chosen were lung, prostate, breast, kidney, bladder, colon, skin, stomach, pancreatic, uterine, and ovarian cancers. Cervical cancer was not included as it required participants to disclose information on sexual history which was considered too sensitive.

In general terms, the information required to develop prediction models are: (1) the identified list of risk factors for inclusion; (2) point estimates of the relative risk for each risk factor; and (3) population prevalence for each of the exposures. To be able to compare individual risk to the population, further information such as cancer incidence by 10-year age bands is required.

### Comparison of relative risk between the United States and the UK

To illustrate the comparative relative risks (RR) between the two populations, we used the UKBiobank national cohort and analyzed the RRs for breast cancer. The UKBiobank female cohort consists of 273,467 women with age ranged between 40 and 69 years when recruited. Participants were enrolled in the UK Biobank from April 2007 to July 2010, from 21 assessment centers across England, Wales, and Scotland using standardized procedures (11). The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee, and all participants provided written informed consent to participate in the UK Biobank study. To date, the cohort has been followed up for 6 years. We computed RRs adjusted only for age. The results are presented in Table 6.

### Model validation: An example

To demonstrate model validation, we selected breast cancer risk prediction as an example. The UKBiobank cohort was used for the validation exercise. For breast cancer cases, we used ICD10, ICD9, and self-reported codes (verified by the UKBiobank health professionals), and only incident cases were included in the analysis. For controls, we used two comparison groups. Firstly only those subjects with no cancer code recorded in ICD10, ICD9 and secondly the coded with no cancer and no other self-reported illness. The total number of incident breast cancer cases was 3,378, and the total number of noncancer controls was 235,603 or healthy controls was 59,731. We coded all variables (except Tamoxifen/Raloxifene usage as such data were not available) present in the breast cancer risk prediction model based on the presence or absence of the exposure for each individual as illustrated in Table 3. To demonstrate model validation, we calculated the area under the curve of the model based on including all factors scored as a binary variable and generated calibration plots -the observed and expected proportions compare within the groups formed by the Hosmer-Lemeshow test. All analyses were performed using STATA 14 (12).

## Results

To develop the UK version of the "YourDiseaseRisk" models, we have assumed that the risk factors for these cancers in the United States and UK are the same and have obtained the list of risk factors and point estimates of relative risk for each risk factor from the U.S. "YourDiseaseRisk" models. Since the "YourDiseaseRisk" model was first developed more than a decade ago and to ensure we use the most updated version, we started the process by extracting risk factors for each cancer from the "YourDiseaseRisk" online version (ref. 8; Step 1). Once we created a list of risk factors for each cancer, we assigned point estimates of RR for each factor (Step 2). These point estimates of RR's were extracted from the original

**Table 1.** Colon cancer risk factors, its prevalence, and reference information

<b>Risk factor</b>	<b>Definition</b>	<b>Male prevalence</b>	<b>Female prevalence</b>	<b>Source</b>	<b>Population</b>
1. Family history	Brother, sister, or parent had colon cancer (1 <sup>st</sup> -degree relatives affected with colon cancer)	6.0%	7.0%	Sandhu MS, Luben R, Khaw KT (2001) Prevalence and family history of colorectal cancer: implications for screening. <i>J Med Screen</i> 8(2): 69–72.	30,353 participants ages 45–74 were recruited from GP between 1993 and 1997 as part of the East Anglian component of the European Prospective Investigation into Cancer (EPIC-Norfolk).
2. Obesity	BMI $\geq 27$ kg/m <sup>2</sup>	52.0%	49.0%	Health England survey 2013 (Micro data). The Health Survey for England series was designed to monitor trends in the nation's health, to estimate the proportion of people in England who have specified health conditions, and to estimate the prevalence of certain risk factors and combinations of risk factors associated with these conditions.	UK population with a total of 2,362 males and 2,810 females aged $\geq 45$ .
3. Saturated fat	Milk or dairy products $\geq 3$ serving/day	9.0%	12.0%	Findings from the National Adult Nutrition Survey: Dairy Intakes and Compliance with Food Pyramid Recommendations among Irish Adults Aged 65 years and over	Irish population (database contains details data on servings, we selected high intake as $\geq 3.5$ servings/day).
4. Alcohol	More than 7 servings per week	59.0%	41.0%	Health England survey 2013 (Micro data).	UK population with a total of 2,362 males and 2,810 females aged $\geq 45$ . Variable used for the analysis: total unit of drinks per week.
5. Vegetables	3 or more servings per day	14.0%	14.0%	Health England survey 2013 (Micro data).	UK population with a total of 2,362 males and 2,810 females aged $\geq 45$ . Variable used for the analysis: total portion of vegetable per day (adults aged 45+).
6. Height	5 ft. 7 in. or taller	10.0%	7.0%	Health England survey 2013 (Micro data).	UK population with a total of 2,362 males and 2,810 females aged $\geq 45$ . Variable used for the analysis: valid height.
7. Physical activity	3 or more hours total leisure-time physical activity per week.	37.0%	23.0%	Physical statistic 2015. Report by the British Heart Foundation Centre on Population Approaches for Non-Communicable Disease Prevention, Nuffield Department of Population Health, University of Oxford.	UK population. Guidelines issued by the Chief Medical Officers (CMOs) of England, Scotland, Wales, and Northern Ireland in 2011. Over a week, activity should add up to at least 150 minutes (2½ hours) of moderate intensity activity in bouts of 10 minutes or more.
8. Red meat	Eating 3 or more servings a week. (U.S. meat one serving = 4 ounces or = 113.4 grams)	97.5%	91.0%	Parkin DM (2011) 5. Cancers attributable to dietary factors in the UK in 2010. II. Meat consumption. <i>Br J Cancer</i> 105 Suppl 2: S24–26.	Data on consumption of meat in the UK year 2000–2001 from the National Diet and Nutrition Survey (Food Standards Agency, 2002) as mean consumption, in grams of different types of meat per week, by age group and sex.

(Continued on the following page)

**Table 1.** Colon cancer risk factors, its prevalence, and reference information (Cont'd)

<b>Risk factor</b>	<b>Definition</b>	<b>Male prevalence</b>	<b>Female prevalence</b>	<b>Source</b>	<b>Population</b>
9. Use of birth control pills	5 or more years of use	N/A	57.0%	Farrow A, Hull MGR, Northstone K, Taylor H, Ford WCL, Golding J (2002) Prolonged use of oral contraception before a planned pregnancy is associated with a decreased risk of delayed conception. Human Reproduction 17(10): 2754–2761.	The Avon Longitudinal Study of Parents and Children (ALSPAC).
10. Use of postmenopausal hormones	5 or more years of use	N/A	8.0%	Benson VS, Kirichek O Fau - Beral V, Beral V Fau - Green J, Green J Menopausal hormone therapy and central nervous system tumor risk: large UK prospective study and meta-analysis. (1097-0215 (Electronic)).	UK population (General Practice Research Database (GPRD)).
11. Aspirin use	Use daily for 15 years or more.	22.0%	13.0%	Elwood P, Morgan G, White J, Dunstan F, Pickering J, Mitchell C, Fone D (2011) Aspirin taking in a south Wales county. The British Journal of Cardiology 18: 238-240.	Sample of adults residing in the south Wales county of Caerphilly, the study conducted a survey of a sample 9,551 adults resident in the county aged ≥18 years. UK population
12. Multivitamin (folate)	Folate intake reflected in regular multivitamin use (>15 yrs vs. no use)	5.0%	7.0%	Comparison of standardized dietary folate intake across ten countries participating in the European Prospective Investigation into Cancer and Nutrition. British Journal of Nutrition (2012), 108, 552–569.	UK population (General Practice Research Database (GPRD)).
13. Inflammatory bowel disease	Affected by the condition for 10 or more years.	1.0%	3.0%	Canavan C, Card T, West J (2014) The incidence of other gastroenterological disease following diagnosis of irritable bowel syndrome in the UK: a cohort study. PLoS One 9(9): e106478. <a href="http://www.telegraph.co.uk/news/health/news/11900171/Calcium-supplements-dont-work-say-experts.html">http://www.telegraph.co.uk/news/health/news/11900171/Calcium-supplements-dont-work-say-experts.html</a>	In UK population, up to 11% of British adults are estimated to take calcium supplements.
14. Calcium supplement	Regular use of calcium supplement everyday	11.0%	11.0%	Spiro A and Buttriss J.L. (2014) Vitamin D: An overview of vitamin D status and intake in Europe. Nutrition Bulletin Vol 39;4.	Irish population
16. Vitamin D supplement	Regular use of calcium supplement everyday	15.0%	15.0%		

**Table 2.** RRs for selected breast cancer risk factors extracted from the original publication by Colditz and colleagues (14) and from the UKBiobank data

Risk factor	RR <sup>a</sup>	RR <sup>b</sup>	95% CI <sup>b</sup>	Score assigned in Colditz and colleagues	Score assigned in the UKBiobank study
Family history (mother and sister)	3.0	3.0	2.0-3.6	25	25
Family history (first-degree relative)	1.8	1.5	1.4-1.7	10	10
Height	1.3	1.3	1.3-1.5	5	5
Age of first period	0.8	0.9	0.9-1.0	-5	-5
Age at menopause	1.2	1.3	1.2-1.4	5	5
OC use	1.4	1.1	1.0-1.2	5	5
Estrogen replacement $\geq$ 5 years	1.7	1.3	1.2-1.4	10	5
Estrogen replacement < 5 years	1.1	1.2	1.0-1.3	5	5
Physical activity	0.6	0.9	0.8-0.9	10	5
Alcohol	1.4	1.1	1.0-1.2	5	5
Obesity (postmenopausal)	1.3	1.1	1.1-1.2	5	5
Obesity (premenopausal)	0.8	0.8	0.8-1.0	-5	-5
Multivitamin supplement	0.5	0.9	0.9-1.0	10	0
Number of births	1.1	1.23	1.13-1.24	5	5
Benign breast disease (MD diagnosed)	1.5	1.4	1.1-1.8	10	5
Birth weight	1.1	1.1	0.9-1.2	5	5

<sup>a</sup>RR from publication by Colditz and colleagues.

<sup>b</sup>RR from the UKBiobank study.

publication (13, 14); however if any factors did not have any cited value, we performed a literature search to obtain any missing estimates. To maintain consistency across RR values, we used figures from publications by the Colditz study group and those with a cohort study design (examples of references are depicted in Table 1). For example, as the multivitamin factor was not listed in the original article, we therefore chose and applied an RR of 0.7 from Zhang and colleagues (15).

As the original report applied RRs from compiling evidence from the U.S. cohort studies over a period of time, it is important to justify the use of the RRs published by Colditz and colleagues (14) in the UK models. We demonstrated the similarities/variations of risks between the two populations (US and UK). To do this, we have presented as an example the RRs for breast cancer derived from the UKBiobank study (Table 2). The majority of the point estimates are similar and convert to the same risk score when using values in Table 3. The two exceptions were for multivitamin use and physical activity where the protective effects for both of these factors were less pronounced in the UK as compared to the U.S. estimate.

To obtain the population prevalence (step 3) for each of the exposures, we then reviewed the literature on the UK prevalence of each factor for men and women for all 11 cancers. The criteria for publication selection included (1)

UK prevalence data from National surveys (16) or prevalence derived from large cohort studies representative of the general population in the UK, or (2) if no data were available from those sources, information/figures from cohort studies in European countries. As an example, the colon cancer risk factors, RR, and references chosen for prevalence are shown in Table 1.

### Computing risk scores

Once information on the UK prevalence of each risk factor was obtained, we then applied a score to each risk factor using the same scheme as presented in the original article (summary as shown in Table 3).

This risk score is used to compute two further scores—the population average risk score and an individual risk score relative to the population average. The population average score is calculated by multiplying the risk score of each factor by the population prevalence of that particular factor. To prevent negative scores, we chose the direction of each risk factor to make the population average score the highest possible. Taking physical activity, for example, the prevalence of carrying out 3 or more hours of total leisure-time physical activity per week in the UK population is 23%. This figure means that 77% of the population do not do physical activity regularly at this particular level. When we apply a prevalence of 77%, then the assigned score instead of being -10 (for those who are doing regular exercise) will be +10. This conversion allows us to demonstrate the change in the individual risk score following any change in factors that are modifiable. Summation of these scores produces the average population score.

The risk score for a given individual relative to the population average is based on the presence or absence of each factor. An example is illustrated in Table 4. The

**Table 3.** Risk score applied to level of RR

Relative risk	Risk score
0.9- $<$ 1.1	0
0.7- $<$ 0.9 or 1.1- $<$ 1.5	5
0.4- $<$ 0.7 or 1.5- $<$ 3.0	10
0.2- $<$ 0.4 or 3.0- $<$ 7.0	25
$<$ 0.2 or $\geq$ 7.0	50

**Table 4.** Illustration of population average score and individual risk score for breast cancer

<b>Risk factor</b>	<b>RR</b>	<b>Description</b>	<b>Score</b>	<b>UK prevalence women</b>	<b>Population average points for women</b>	<b>Individual factor profile</b>	<b>Individual score</b>
Family history (mother and sister)	3.0	Two first-degree relatives (mother and sister) affected with breast cancer before aged 65	25	0.0011	0.03	No	0
Family history (first-degree relative)	1.8	First-degree relative who has a history of breast cancer before age 65 vs. none	10	0.09	0.90	No	0
Height	1.3	5 feet 7 inch or taller for women	5	0.07	0.35	Yes	5
Age of first period	0.8	Age of first period (15 vs. 11)	-5	0.17	-0.85	Age 11	0
Age at menopause	1.2	Age at menopause (at the age of 55 or older)	5	0.07	0.34	Age 55	5
OC use	1.4	OC use (current use vs. none)	5	0.29	1.45	Yes	5
Estrogen replacement	1.7	Estrogen replacement $\geq$ 5 yrs	10	0.08	0.80	No	0
Estrogen replacement	1.1	Estrogen replacement < 5 yrs	5	0.08	0.40	No	0
Physical activity	0.6	3 or more hours total leisure-time physical activity per week	10	0.77	7.70	No	10
Jewish heritage	1.2	Jewish heritage	5	0.005	0.03	No	0
Alcohol	1.4	More than 1 drink per day vs. 0	5	0.41	2.05	Yes	5
Obesity (postmenopausal)	1.3	27 kg/m <sup>2</sup> or more	5	0.49	2.45	No	0
Obesity (premenopausal)	0.8	27 kg/m <sup>2</sup> or more	-5	0.36	-1.80	Yes	-5
Multivitamin supplement	0.7	Lack of use of multivitamin or B complex	5	0.73	7.3	Use vitamin	0
Number of births	1.1	Number of births (0 or 1 child)	5	0.24	1.20	No child	0
Benign breast disease (MD diagnosed)	1.5	Benign breast disease (MD diagnosed)	10	0.13	1.30	No	0
Tamoxifen or raloxifene	0.5	Tamoxifen or raloxifene for 5 years or more	10	0.30	3.0	No	10
Birth weight	1.1	Birth weight >3.9 kg or more	5	0.07	0.37	No	0
				Population average score	27	Individual risk score	35

summation of scores for each risk factor for an individual provides the total risk score for a particular person. That total risk score is then divided by the average population score to give an individual index that is relative to population score ( $35/27 = 1.3$ ). As mentioned earlier, participants can see how their risk changes if they adopt suggested behaviors, for example, by choosing to do regular exercise, the total score for the individual illustrated in Table 4 can be reduced by 10 points (from 35 to 25), making their index score relative to the population reduced from 1.3 to 0.8. This calculation aims to illustrate to each individual the effect of a particular lifestyle or behavior change leading to cancer risk reduction.

#### Conversion of individual index score to 5-category cancer level of risk

The index score for an individual can then be further transformed into a level of risk. A numeric factor indicative of the strength of the risk level is assigned to the individual index score (Table 5). This is done to give an average value of the range of individual index scores as a single numeric estimate that reflects the risk level.

For the individual in Table 3, for example, the individual index score is 1.3, which is equivalent to "above average risk" and gives a numeric factor of 1.5.

#### Ten-year estimated cancer risk

To enable estimation of an individual's 10-year estimated cancer risk, we calculated the average 10-year estimated risk for different ages and sexes of the UK population for all 11 cancers. We used the "Current Probability" method proposed by Esteve and colleagues in 1994 (17). This method uses a life-table approach for calculating the risk of developing cancer and takes into account the likelihood of dying from other causes. The method also requires information on deaths from all causes for each age group. This method provides estimate of the 10-year risk of cancer.

We obtained age- and sex-specific cancer incidence and mortality rates and numbers from Cancer Research UK (18) and age- and sex-specific data on all-cause mortality from the Office of National Statistic (ONS) which are available online (19). The following shows the specific data used for the calculation:

1. The annual number of cancer deaths (cancer mortality for males and females); data from CRUK 2010–2012.
2. The annual number of (registered) cancer cases for males and females; data from CRUK 2009–2011.
3. The annual number of deaths (all-cause mortality for males and females); data from ONS 2011.
4. The size of the mid-year population for males and females; data from ONS 2011.

From these data, we computed 10-year cancer risk for each cancer in males and females in 10-year age bands from birth to over 80 years.

**Table 5.** Conversion of individual index score to single numeric factor

Individual index score →	Level of risk →	Factor
<0	Very much below average risk	0.2
0, or < 0.5	Much below average risk	0.4
0.5 < 0.9	Below average risk	0.7
0.9 < 1.1	About average risk	1
1.1 < 2.0	Above average risk	1.5
2.0 < 5.0	Much above average risk	3
5.0 or more times the average score	Very much above average risk	5

#### Individual estimate 10-year cancer risk

Assuming the individual in Table 4 is a 40-year-old woman, the average 10-year risk of breast cancer (age 40–49) is 1.45% in the next 10 years (Table 6). Multiplying this figure by the numeric factor of 1.5 (Table 4) will give a risk of 2.2% (above the population average) or approximately 1 case in 45 women.

#### Example of breast cancer risk prediction model performance and validation based on binary scoring of risk factors

The AUC for breast cancer risk prediction model based on utilizing a binary score for each available risk factor in the UKBiobank dataset was 0.58 [95% confidence interval (CI), 0.57–0.60] for a comparison group of controls with no cancer and 0.64 (95% CI, 0.63–0.66) for a comparison group of controls with no cancer or other illnesses. Model calibration curves for both comparisons also suggested both models calibrated well (Fig. 1).

## Discussion

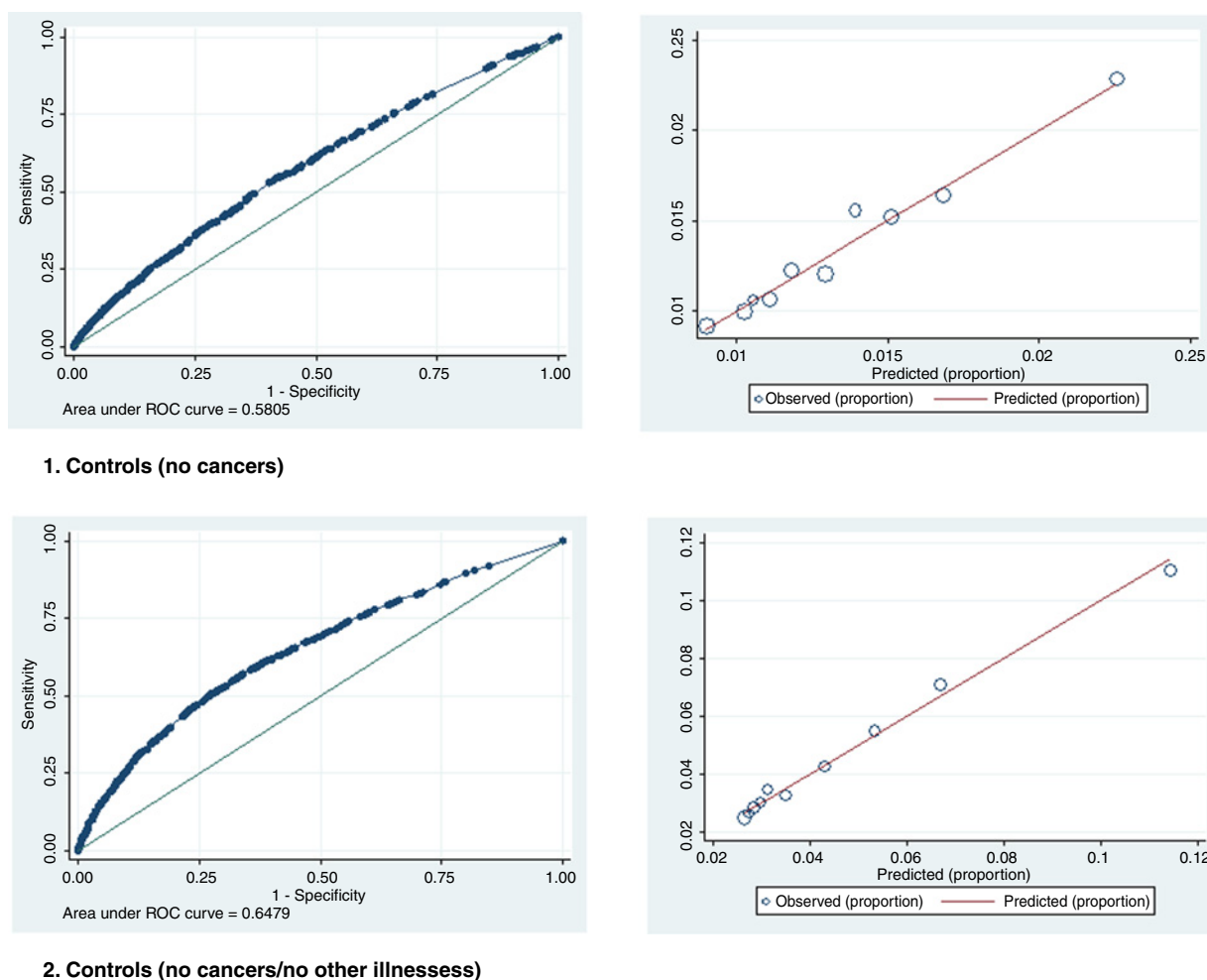
The "YourDiseaseRisk" model has been developed using three key data including an estimate of relative risk of each risk factor, the prevalence figure of exposure in the population, and the 10-year estimated cancer risk. These figures can be acquired through literature review, National data archives, or other organizations that compile these data publicly (20).

**Table 6.** 10-year estimated breast cancer risk in UK and U.S. female population

Age group	10-year estimated risk	
	UK females	U.S. females <sup>a</sup>
0–9	0.000%	0.00%
10–19	0.001%	0.00%
20–29	0.049%	0.06%
30–39	0.442%	0.44%
40–49	1.445%	1.44%
50–59	2.594%	2.28%
60–69	3.336%	3.46%
70–79	2.759%	3.89%
80+	2.944%	3.02%

<sup>a</sup>Estimates from U.S. SEER data.





**Figure 1.** Model performance of the breast cancer subcomponent of the model scored using binary factor categorization: AUC and calibration curves.

The "YourDiseaseRisk" site has been launched since 2000 and update/review of information is an ongoing process. The model is an educational tool and can be used for cancer prevention in clinics, community settings, or by individuals simply seeking information on their own. The popularity of "YourDiseaseRisk" is reflected by the large number of hits or page views.

We have adopted their methodology and applied it to build a UK version and to demonstrate the applicability of the approach for other populations as well. The main strength of the approach is the use of large population-based studies to estimate the parameters needed to generate the model.

To demonstrate the relative validity of the approach, we chose breast cancer as an exemplar. We have demonstrated that the quantitative estimates of the individual risk factors apply to current UK population using data from the UKBiobank study. Furthermore, we performed a model validation of the combined risk factors using a binary scoring system, and the results suggest such a model is

reasonably calibrated and has moderate discriminatory power [0.58 for a comparison group of controls with no cancer and 0.64 (95% CI, 0.63–0.66) for a comparison group of controls with no cancer or other illnesses]. The prediction performance of a fully specified model using the point estimates of each risk factor rather than a binary value is likely to be higher as it will be derived from more precise estimates of their individual effect. Directly comparative data are not yet available in the UKBiobank for all parameters to undertake such analyses.

There are many breast cancer risk prediction models, but only a few that contained epidemiologic factors and most have similar predictive capabilities (21–30). The majority of these models are extended versions of the Gail model (23). They have been shown to have good calibration but moderate discrimination ranging from 0.56 to 0.89 with few having been assessed in external validation analysis (23, 28, 31, 32). Colditz and colleagues further reported an AUC of 0.64 (95% CI, 0.62–0.66) in their extended validation study using data from the Nurses' Health Study

(32). The validation exercise, based on breast cancer example presented here, supports the conclusion that the model can be used in the UK population when the prevalence of risk factors is substituted with the UK figures.

There are, however, potential limitations to the approach that need to be taken into consideration. Firstly, although we were systematic in our literature reviews to obtain data on the point estimate RRs for risk factors missing from the original published "YourDiseaseRisk" model and the prevalence of all risk factors in the UK population, the studies we have chosen may not provide the most accurate estimates. Secondly, we have assumed that the prevalence of each risk factor does not change with age, that the risk associated with each risk factor is the same across all ages and both sexes, and that the risk factors do not interact with each other. These assumptions were also made in the development of the "YourDiseaseRisk" models but, as noted in that report (14), may result in misclassification of risk for exposures across large age ranges and underestimate possible synergistic effects of exposures such as alcohol and smoking. As for the "YourDiseaseRisk" models, therefore, these models should be considered as a guide for assessing an individual's risk of cancer in the UK rather than a precise estimate.

Risk prediction models are widely used for many diseases. With the concept that some cancers can be preventable, the "Your Disease risk" model developed by Colditz and colleagues (14) provides a good platform because their model is based on modifiable factors that have been scrutinized and carefully selected by a panel of experts. The approach has wide utility in allowing for the rapid development of models for educational purposes.

In this article, we have described how we have adapted that model to produce a UK version for 11 cancers using data from the UK population. We are now using the model in the community in the UK. Going forward, we will evaluate how the tool affects perceptions of risk, how to best present the risk, and how the public understands their individual risk. This information will then inform future studies in the community exploring the potential for the use of this model to promote lifestyle change. Finally, we will extend and further validate this risk prediction in the

UKBiobank cohort as more precise individual level data and longer term follow-up data become available.

### Disclosure of Potential Conflicts of Interest

J. Warcaba is Macmillan GP. No potential conflicts of interest were disclosed by the other authors.

### Authors' Contributions

**Conception and design:** A. Lophatananon, J. Usher-Smith, J. Campbell, J. Warcaba, G.A. Colditz, K.R. Muir

**Development of methodology:** A. Lophatananon, J. Usher-Smith, J. Campbell, J. Warcaba, G.A. Colditz, K.R. Muir

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** J. Warcaba

**Analysis and interpretation of data (e.g., statistical analysis, bio-statistics, computational analysis):** A. Lophatananon, J. Usher-Smith

**Writing, review, and/or revision of the manuscript:** A. Lophatananon, J. Usher-Smith, J. Campbell, J. Warcaba, B. Silarova, E.A. Waters, G.A. Colditz, K.R. Muir

**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** A. Lophatananon

**Study supervision:** J. Usher-Smith, J. Campbell, K.R. Muir

### Acknowledgments

We would like to thank Kawthar-Al-Ajmi, PhD student at the University of Manchester, for her help with preparation the UKBiobank data. This research has been done using the UK Biobank resources application number 5791.

### Grant Support

A. Lophatananon, K.R. Muir, J. Usher-Smith, J. Campbell, J. Warcaba, and B. Silarova were funded by an Innovation Grant from the Cancer Research UK – BUPA Foundation Fund (C55650/A20818). A. Lophatananon and K.R. Muir were also funded by ICEP ("This work was also supported by CRUK [grant number C18281/A19169]"). J. Usher-Smith was also supported by the National Institute for Health Research Clinical Lectureship, and B. Silarova was supported by the Medical Research Council (MC\_UU\_12015/4). E.A. Waters and G.A. Colditz are funded in part by the Foundation for Barnes-Jewish Hospital, St Louis, MO.

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.

Received November 11, 2016; revised April 18, 2017; accepted May 22, 2017; published OnlineFirst May 30, 2017.

### References

1. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk prediction models for colorectal cancer: a systematic review. *Cancer Prev Res (Phila)* 2016;9:13–26.
2. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS Med* 2012;9:e1001344.
3. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. *Ann Intern Med* 2013;158:596–603.
4. Engel C, Fischer C. Breast cancer risks and risk prediction models. *Breast Care (Basel)* 2015;10:7–12.
5. Gray EP, Teare MD, Stevens J, Archer R. Risk prediction models for lung cancer: a systematic review. *Clin Lung Cancer* 2016;17:95–106.
6. Usher-Smith J, Emery J, Hamilton W, Griffin SJ, Walter FM. Risk prediction tools for cancer in primary care. *Br J Cancer* 2015;113:1645–50.
7. Bruzelius M, Bottai M, Sabater-Lleal M, Strawbridge RJ, Bergendal A, Silveira A, et al. Predicting venous thrombosis in women using a combination of genetic markers and clinical risk factors. *J Thromb Haemost* 2015;13:219–27.
8. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine. Your Disease Risk. 2013 [cited 2016 Jul 20]. Available from: <http://www.yourdiseaserisk.wustl>.

- edu/YDRDefault.aspx?ScreenControl=YDRGeneral&ScreenName=YDRAbout.
9. Kim DJ, Rockhill B, Colditz GA. Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk. *J Clin Epidemiol* 2004;57:332–40.
  10. National Cancer Institute (US). Cancer Statistics [database on the Internet]. 2016 [cited 2016 Jul 26]. Available from: [www.seer.cancer.gov](http://www.seer.cancer.gov).
  11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779.
  12. StataCorp. Stata Statistical Software: Release 14 College Station, Texas: StataCorp LP; 2015.
  13. Tomeo CA, Colditz GA, Willett WC, Giovannucci E, Platz E, Rockhill B, et al. Harvard report on cancer prevention. Volume 3: prevention of colon cancer in the United States. *Cancer Causes Control* 1999;10:167–80.
  14. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, et al. Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. *Cancer Causes Control* 2000;11:477–88.
  15. Zhang S, Hunter DJ, Hankinson SE, Giovannucci EL, Rosner BA, Colditz GA, et al. A prospective study of folate intake and the risk of breast cancer. *JAMA* 1999;281:1632–7.
  16. Office for National Statistics. Health Survey for England, 2013. [data collection] [database on the Internet]. 2015 [cited 2015 Jul 26]. <http://dx.doi.org/10.5255/UKDA-SN-7649-1>.
  17. Esteve J, Benhamou E, Raymond L. Statistical methods in cancer research. Volume IV. Descriptive epidemiology. *IARC Sci Publ* 1994;1–302.
  18. Cancer Research UK. Cancer Statistics for the UK [database on the Internet]. 2016 [cited 2016 Jul 26]. Available from: <http://www.cancerresearchuk.org/health-professional/cancer-statistics>.
  19. Office for National Statistics. All Cause mortality [database on the Internet]. 2016 [cited 2016 Jul 26]. Available from: <https://www.ons.gov.uk/>.
  20. National Cancer Institute (US). Data & Software for Researchers. 2017 [cited 2017 Feb 23]. Available from: <https://seer.cancer.gov/resources/>.
  21. Banegas MP, Gail MH, LaCroix A, Thompson B, Martinez ME, Wactawski-Wende J, et al. Evaluating breast cancer risk projections for Hispanic women. *Breast Cancer Res Treat* 2012;132:347–53.
  22. Boyle P, Mezzetti M, La Vecchia C, Franceschi S, Decarli A, Robertson C. Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. *Eur J Cancer Prev* 2004;13:183–91.
  23. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, et al. Projecting individualized absolute invasive breast cancer risk in African American women. *J Nat Cancer Inst* 2007;99:1782–92.
  24. Lee C, Lee JC, Park B, Bae J, Lim MH, Kang D, et al. Computational discrimination of breast cancer for Korean women based on epidemiologic data only. *J Korean Med Sci* 2015;30:1025–34.
  25. Lee E-O, Ahn S-H, You C, Lee D-S, Han W, Choe K-J, et al. Determining the main risk factors and high-risk groups of breast cancer using a predictive model for breast cancer risk assessment in South Korea. *Cancer Nursing* 2004;27:400–6.
  26. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. *J Nat Cancer Inst* 2011;103:951–61.
  27. Novotny J, Pecan L, Petruzella L, Svobodnik A, Dusek L, Danes J, et al. Breast cancer risk assessment in the Czech female population—an adjustment of the original Gail model. *Breast Cancer Res Treat* 2006;95:29–35.
  28. Park B, Ma SH, Shin A, Chang M-C, Choi J-Y, Kim S, et al. Korean risk assessment model for breast cancer risk prediction. *PLoS One* 2013;8:e76736.
  29. Ueda K, Tsukuma H, Tanaka H, Ajiki W, Oshima A. Estimation of individualized probabilities of developing breast cancer for Japanese women. *Breast Cancer* 2003;10:54–62.
  30. Wang S, Ogundiran T, Ademola A, Olayiwola OA, Adeoye A, Adeniji-Sofoluwe A, et al. Abstract 2590: development and validation of a breast cancer risk prediction model for black women: findings from the Nigerian breast cancer study. *Cancer Res* 2016;76:2590.
  31. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. *J Nat Cancer Inst* 1996;88:359–64.
  32. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. *Am J Epidemiol* 2000;152:950–64.